Site Pattern Probabilities Under the Multispecies Coalescent and a Relaxed Molecular Clock: Theory and Applications

A. Richards and L. Kubatko

August 11, 2021

1 Abstract

The first step in statistical inference of the evolutionary histories of species is developing a probability model that describes the mutation process as accurately and realistically as possible. A major complication of this inference is that different loci on the genome can have histories that diverge from the common species history and each other. The multispecies coalescent process is commonly used to model one source of this divergence, incomplete lineage sorting, or ILS. Chifman and Kubatko (2015) computed the site pattern probabilities for four taxa under a full probability model based on the Jukes-Cantor substitution model when the molecular clock holds. This paper generalizes that work to a relaxed clock model, allowing for mutation rates to differ among species. This will enable better phylogenetic inference in cases where the molecular clock does not hold.

2 Introduction

We begin with a collection of species as demonstrated in figure 1. The thick outer lines represent the evolutionary history of a set of six mammalian species. We can see that the chimpanzee is our closest relative, the mouse is most closely related to the rabbit, and so on. The inner blue lines represents the history of one locus on the genomes. At this particular locus, the mouse is more closely related to the (chimp,human) pair than it is to the rabbit. Different lineage histories could be depicted for this set of six

Figure 1: The multispecies coalescent problem: how to infer the species tree topology (black outer lines) when there may be mismatch with the gene tree (inner blue lines).

species for each common locus on their respective genomes.

We describe the evolutionary histories of the species via a species tree:

Definition 1. A *species tree* is an acyclic graph $S = (V(S), E(S), \tau_S)$ where $V(S)$ is the vertex set of S , $E(S)$ is the edge set of S , and τ_S is a vector of node times.

The external nodes on the species tree represent extant species while the internal nodes represent speciation events. Internal branches represent ancestral species. If we know the common ancestor of all the species under consideration, then the tree is *rooted*, and the tree becomes a directed graph from the root outward. In the case that the location of the common ancestor is unknown, then we have an unrooted tree. Then we only know the direction of the branches that connect to external nodes. As we will see, an unrooted tree is sometimes also used when knowledge of the root location is unimportant in order to ease the computational burden.

Species tree inference is complicated by the possibility for divergence between the evolutionary history

Figure 2: Species trees with four taxa under the coalescent process. The green lines show example gene trees evolving within the underlying species tree.

of species and individual elements of their genomes. Causes for this divergence include incomplete lineage sorting (ILS), gene duplication and loss (GDL) and horizontal gene transfer (HGT). ILS is commonly modeled by the coalescent process [8]. The history of individual loci on the genome is represented by a gene tree.

Definition 2. A *gene tree* is a acyclic graph $G = (V(G), E(G), t_G)$ where $V(G)$ is the vertex set of G , $E(G)$ is the edge set of G , and t_G is a vector of node times. These node times are subject to the constraint that the nodes of the gene tree must occur prior to the speciation time of the species in question.

In the sequel, we will assume there is one individual sampled from each species and there is no missing data. Then, both trees have the same leaf set L . When we need to distinguish the leaves of G from the leaves of S we use capital letters $A, B, C, \dots$ for the gene tree and lower case for the species tree: $a, b, c, \dots$. Internal nodes on the gene tree represent coalescent events, which identify the most recent common ancestor of two gene lineages. So the symmetric tree in figure 2a has a species tree $((a, b), (c, d))$ and gene tree $(D, (A, (B, C)))$.

A number of different approaches have been taken with regard to species tree inference in the presence of ILS. The first is essentially to ignore the problem: perform gene tree inference on concatenated data using methods such as RAxML [28] or FastTree [16], treating all sites as if they share a single, common evolutionary history. This can be fast and accurate for estimating S . But, there are some concerns: concatenation has been shown to be statistically inconsistent for some values of (S, τ) [4, 20], and speciation time estimates are biased since the coalescent event must naturally occur before the speciation time. Another approach is the use of summary statistics that first estimate the gene trees independently for each gene, and then use

the gene tree estimates as inputs for species tree estimates. Examples of this approach include STEM [9], ASTRAL [13, 14, 36], and MP-EST [12]. These methods can be computationally efficient, however they depend on the accuracy of the gene tree estimates that are used as inputs as well as proper delineation of recombination-free segments of the genome [7, 27]. A third approach uses the full data to coestimate the species tree and each of the gene trees, generally using Markov chain Monte Carlo (MCMC) methods. Examples of this approach include BEST [11], *BEAST [5, 15], and BPP [35]. These methods can be quite accurate but are very computationally intensive when the number of loci and/or leaf set is large. Assessment of convergence is also a challenge, especially due to the multi-modal nature of the likelihood in the tree space [22].

A fourth approach, and the one we take in this paper, is to treat the gene trees as a nuisance parameter that can be integrated over. Previous examples of this approach include SVDQuartets [2] and Lily-Q [19]. SVDQuartets uses a methodology based in linear algebra to infer the proper unrooted species tree for each quartet of species under consideration, and then uses an assembly algorithm to infer the final n-taxon species tree estimate taking the set of unrooted quartet trees as input. Lily-Q instead takes the site pattern probabilities $P(D|(S, \tau))$ and then assumes a distribution on τ and S to find a posterior probability of S given the data D . This was shown to be fast and highly accurate, but the probabilities $P(D|(S, \tau))$ from [3] were calculated assuming a constant mutation rate over time (or more precisely, if λ represents the mutation rate there is a common $\lambda(t)$ for all branches at time t), an assumption known as the *strict molecular clock*. Inference using Lily-Q was very sensitive to this assumption, prompting the need to generalize the work of [3] to cases for which the clock does not hold.

3 Model

Our model is as follows. We assume that there is no missing data and that each site is properly aligned across species and has an independent evolutionary history from every other site given the species tree. In other words, each site has its own gene tree which is a random draw from the distribution of gene trees induced by the coalescent process, $P(G = g|S)$. We also assume neutral selection, which implies that the substitution process and coalescent process are independent [33]. Let $\delta_{x_A x_B x_C x_D} = P(i_A = x_A, i_B = x_B, i_C = x_C, i_D = x_D)$, $X_j \in \{A, C, G, T\}$ for $j \in \{A, B, C, D\}$. Then δ_{TCCG} , for example, represents the probability that at a certain site the first species has T, the next two have C, and the fourth species has G,

i.e., $\delta_{TCCG} = P(i_A = T, i_B = C, i_C = C, i_D = G)$ where i_x is the nucleotide for species x at any given site. Under the Jukes-Cantor substitution model, all four nucleotides have the same limiting probability of 1/4 and the same mutation rate between them, so $\delta_{TCCG} = \delta_{AGGT}$, for example, and we can instead refer to both of these probabilities as δ_{YXXZ} . In fact, there are fifteen different site patterns under the Jukes-Cantor model, whose frequencies can be summarized by the 15×1 vector $\mathbf{D} = \{d_{XXXX}, d_{XXXY}, \dots, d_{XYZW}\}$ which follows a multinomial distribution with probabilities $\boldsymbol{\delta} = \{\delta_{XXXX}, \delta_{XXXY}, \dots, \delta_{XYZW}\}$ [19]. We parameterize the species and gene tree as follows. Let $\boldsymbol{\tau}$ represent the speciation times (the times associated with the nodes of the species tree) looking back from the present. Similarly, $\mathbf{t}$ represents coalescent times measured back from the present. Time is measured in *mutation units*, so one unit of time is the expected time for one mutation in the common ancestor of all species (the mutation rate on the branch above the root of the tree).

Initially, we assume that there is a constant mutation rate on each branch of the tree, and parameterize these rates relative to the mutation rate above the root with the vector of relative rates given by $\boldsymbol{\gamma}$. For the symmetric tree, $\boldsymbol{\gamma} = \{\gamma_A, \gamma_B, \gamma_C, \gamma_D, \gamma_{AB}, \gamma_{CD}\}$ and for the asymmetric tree $\boldsymbol{\gamma} = \{\gamma_A, \gamma_B, \gamma_C, \gamma_D, \gamma_{CD}, \gamma_{BCD}\}$, where the subscript indicates the set of species descending from the branch. Thus, we do not need to make an assumption about the absolute mutation rate to calculate probabilities, since we use mutation units for time and reference all relative rates to the rate above the root. For the pendant branches, whether or not there is a constant rate on the branch is irrelevant, but for internal branches we will later generalize our model in section 4 to allow for changing mutation rates over time.

The probability of the data is also dependent on the effective population size N_e through the population parameter $\theta = 4N_e\mu$, where μ is the mutation rate per site per generation. So, our goal is to determine:

$$P(\mathbf{D} = \mathbf{d} | (S, \boldsymbol{\tau}, \boldsymbol{\gamma}, \theta))$$

which due to the independence of the coalescent and substitution processes can be broken into

$$\int_{(G, \mathbf{t})} P(\mathbf{D} = \mathbf{d} | (G, (\mathbf{t}, \boldsymbol{\gamma})) f((G, \mathbf{t}) | (S, \boldsymbol{\tau}, \theta))$$

where the first term describes the substitution process and the second the coalescent process. Since we view the gene trees as a nuisance parameter, we integrate over all gene trees consistent with the species tree, $\mathcal{G}_S$:

$$\delta | (S, \boldsymbol{\tau}, \boldsymbol{\gamma}, \theta) = \sum_{G \in \mathcal{G}_S} \int_{\mathbf{t}} \delta | (G, \mathbf{t}, \boldsymbol{\gamma}) f((G, \mathbf{t}) | (S, \boldsymbol{\tau}, \theta)) d\mathbf{t} \quad (1)$$

We will discuss the derivation of both of these terms beginning with the gene tree density given the species tree.

3.1 Gene tree densities given the species tree

3.1.1 Background

Looking at figure 2, we can see that not all combinations (G, t) are possible given (S, τ) . For example, if the species tree is $((a, b), (c, d))$, then $(D, (A, (B, C)))$ can only occur with the constraint that $t_1 > \tau_3$ because b and c only share a common ancestor before τ_3 . In fact, 25 different gene tree histories are possible in the symmetric case and 34 are possible in the asymmetric case as laid out in tables 4 to 6.

The gene tree density for each of these histories is given by [18] and [3]. Given j lineages, the time to the next coalescent event t_j follows an exponential distribution with rate $\binom{j}{2} \frac{2}{\theta}$. For any interior branch of the species tree, this process is censored, and so with j lineages entering a branch of length τ_b the probability of no coalescent events on the branch is given by $\exp[-\frac{j(j-1)}{\theta} \tau_b]$. More generally, the density of the vector of coalescent times where j lineages enter an interior branch and k leave is given by:

$$f(t) = \prod_{k+1}^j \left(\frac{2}{\theta} \exp[-\frac{i(i-1)}{2} (t_i - t_{i-1})] \right) \exp[-\frac{k(k-1)}{\theta} (\tau_b - t_{k-j})]$$

where $t_0 = 0$ and time is measured from the start of the branch.

3.1.2 Example

Figure 3 shows one example species tree, in this case the symmetric tree $((a, b), (c, d))$ with the asymmetric gene tree $(A, (B, (C, D)))$ embedded within. In this section, we will describe the derivation of that gene tree density, and in the next we will derive the mutation probabilities along the gene tree.

There is no coalescent event on the internal branch from the root to the divergence of (a, b) . This occurs with probability $e^{-\frac{2}{\theta}(\tau_3 - \tau_2)}$. The coalescent event between C and D happens at rate $\frac{2}{\theta}$, and can only occur between τ_1 and τ_3 in this gene tree history. The coalescent event between B and (C, D) occurs above the root and with a rate $\frac{6}{\theta}$, while the final coalescent event occurs after t_2 and with a rate $\frac{2}{\theta}$. Finally, since there are three possibilities for the first coalescent event after the root, we divide the total probability by $\frac{1}{3}$, giving

Figure 3: Example species tree and gene tree history to illustrate probability calculations

the density:

$$f(t_1, t_2, t_3) = \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_1 - \tau_1) - \frac{2}{\theta}(\tau_3 - \tau_2) - \frac{6}{\theta}(t_2 - \tau_3) - \frac{2}{\theta}(t_3 - t_2)\right]$$

We will show in the next section that the site pattern probabilities given the gene tree are all linear combinations of terms that have the form $g_0 e^{-g_1 t_1 - g_2 t_2 - g_3 t_3}$. Then, we integrate over (t_1, t_2, t_3) in the following manner:

$$\begin{aligned} & \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_1}^{\tau_3} g_0 \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_1 - \tau_1) - \frac{2}{\theta}(\tau_3 - \tau_2) - \frac{6}{\theta}(t_2 - \tau_3) - \frac{2}{\theta}(t_3 - t_2)\right] e^{-g_1 t_1 - g_2 t_2 - g_3 t_3} dt_1 dt_2 dt_3 \\ &= g_0 \left(\frac{2}{\theta}\right)^3 \exp\left[\frac{-2}{\theta}(\tau_3 - \tau_2) + \frac{2\tau_1}{\theta} + \frac{6\tau_3}{\theta}\right] \int_{\tau_3}^{\infty} e^{-t_3(g_3 + \frac{2}{\theta})} \left(\int_{\tau_3}^{t_3} e^{-t_2(g_2 + \frac{4}{\theta})} \left(\int_{\tau_1}^{\tau_3} e^{-t_1(g_1 + \frac{2}{\theta})} dt_1 \right) dt_2 \right) dt_3 \end{aligned}$$

Let

$$\zeta = g_0 \left(\frac{2}{\theta}\right)^3 \exp\left[\frac{-2}{\theta}(\tau_3 - \tau_2) + \frac{2\tau_1}{\theta} + \frac{6\tau_3}{\theta}\right]$$

Then the above triple integral becomes:

$$\begin{aligned} & \frac{\zeta \theta}{2 + g_1 \theta} (e^{\tau_1(g_1 + \frac{2}{\theta})} - e^{\tau_3(g_1 + \frac{2}{\theta})}) \int_{\tau_3}^{\infty} e^{-t_3(g_3 + \frac{2}{\theta})} \left(\int_{\tau_3}^{t_3} e^{-t_2(g_2 + \frac{4}{\theta})} dt_2 \right) dt_3 \\ &= \zeta \frac{\theta}{2 + g_1 \theta} \frac{\theta}{4 + g_2 \theta} (e^{\tau_1(g_1 + \frac{2}{\theta})} - e^{\tau_3(g_1 + \frac{2}{\theta})}) \int_{\tau_3}^{\infty} e^{-t_3(g_3 + \frac{2}{\theta}) - \tau_3(g_2 + g_3 + \frac{4}{\theta})} - e^{-t_3(g_2 + g_3 + \frac{6}{\theta})} dt_3 \\ &= \zeta \frac{\theta}{2 + g_1 \theta} \frac{\theta}{4 + g_2 \theta} (e^{\tau_1(g_1 + \frac{2}{\theta})} - e^{\tau_3(g_1 + \frac{2}{\theta})}) \left(\frac{\theta}{2 + g_3 \theta} e^{-\tau_3(g_2 + g_3 + \frac{6}{\theta})} - \frac{\theta}{6 + g_2 + g_3 \theta} e^{-\tau_3(g_2 + g_3 + \frac{6}{\theta})} \right) \end{aligned}$$

Figure 4: Branch length notation for 4-taxon (unrooted) gene tree

Canceling terms with ζ gives:

$$\begin{aligned}
 &= \left(\frac{2}{2+g_1\theta}\right)\left(\frac{2}{4+g_2\theta}\right)\left(\frac{2}{2+g_3\theta} - \frac{2}{6+g_2\theta+g_3\theta}\right)e^{\frac{-2}{\theta}(\tau_3-\tau_2)}e^{-\tau_3(g_2+g_3)}(e^{-g_1\tau_1} - \exp[-\frac{2}{\theta}(\tau_3-\tau_1) - g_1\tau_3]) \\
 &= g_0 \frac{8}{(2+g_1\theta)(2+g_3\theta)(6+g_2\theta+g_3\theta)}e^{-\frac{2(\tau_3-\tau_2)}{\theta}}e^{-\tau_3(g_2+g_3)}(e^{-g_1\tau_1} - \exp[-\frac{2}{\theta}(\tau_3-\tau_1) - g_1\tau_3]) \quad (2)
 \end{aligned}$$

after algebraic simplification. This one example corresponds to history #10 from table 4 and integral form B. Similar calculations give the integral results for all 25 symmetric and 34 asymmetric gene tree histories, as listed in the appendix.

3.2 Site pattern probabilities given gene trees:

We have from the Jukes-Cantor model that when time is measured in mutation units the transition probabilities on branch l with length d_l are given by:

$$p_{ii}(d_l) = (1/4)(1 + 3e^{-d_l})$$

$$p_{ij}(d_l) = (1/4)(1 - e^{-d_l})$$

where $i, j \in \{A, C, G, T\}$.

Then, the probability of the observed data i_A, i_B, i_C, i_D is given by:

$$\delta_{i_A i_B i_C i_D} = \sum_{i_E \in \{A, C, G, T\}} \pi_E p_{i_E i_A}(d_1) p_{i_E i_B}(d_2) \left(\sum_{i_F \in \{A, C, G, T\}} p_{i_E i_F}(d_5) p_{i_F i_C}(d_3) p_{i_F i_D}(d_4) \right) \quad (3)$$

From the assumptions of the JC69 model, $\pi_E = 1/4$ for all nucleotides. We treat the gene tree as an unrooted tree in order to have one fewer branch substitution probability to concern ourselves with, and the nucleotide at the root can be safely ignored due to the Markov property. Although the Felsenstein pruning algorithm could be used here, the full enumeration of states is small enough to be accomplished readily by computer. The probability of any one particular state of the tree $i_A, i_B, i_C, i_D, i_E, i_F$ is given by:

$$\frac{1}{4} p_{i_E i_A}(d_1) * p_{i_E i_B}(d_2) * p_{i_F i_C}(d_3) * p_{i_F i_D}(d_4) * p_{i_E i_F}(d_5)$$

After the enumeration of states, the states that fit each site pattern were identified and common terms summed, with the results given both by [3] and repeated in the supplementary material. We now note that each of the terms $p_{i_E i_A}(d_1) * p_{i_E i_B}(d_2) * p_{i_F i_C}(d_3) * p_{i_F i_D}(d_4) * p_{i_E i_F}(d_5)$ is a linear combination of 32 terms of the form:

$$\left(\frac{1}{4}\right)^5 \kappa_1 \kappa_2 \kappa_3 \kappa_4 \kappa_5 e^{-(f_1 d_1 + f_2 d_2 + f_3 d_3 + f_4 d_4 + f_5 d_5)} \quad (4)$$

where each of the $\kappa_m \in \{-1, 3\}$ and $f_m \in \{0, 1\}$.

Taken together, each of the fifteen different site pattern probabilities in the equations from the supplementary material is a linear combination of the terms in equation 4. Most of the terms vanish due to cancellations, with the result being another linear combination (where the coefficients are given by table 1):

$$\begin{aligned} \delta|(G, \mathbf{t}, \gamma) = & \frac{1}{256} (1 + c_1 e^{-(d_1+d_2)} + c_2 e^{-(d_3+d_4)} + c_3 e^{-(d_1+d_3+d_5)} + c_4 e^{-(d_1+d_4+d_5)} + c_5 e^{-(d_2+d_3+d_5)} \\ & + c_6 e^{-(d_2+d_4+d_5)} + c_7 e^{-(d_1+d_2+d_3+d_4)} + c_8 e^{-(d_1+d_2+d_3+d_5)} + c_9 e^{-(d_1+d_2+d_4+d_5)} \\ & + c_{10} e^{-(d_1+d_3+d_4+d_5)} + c_{11} e^{-(d_2+d_3+d_4+d_5)} + c_{12} e^{-(d_1+d_2+d_3+d_4+d_5)}) \end{aligned} \quad (5)$$

3.2.1 Example

To find the branch lengths in the unrooted tree, we note that the branch lengths will be a function of the coalescent times and the mutation rate on each branch of the species tree. Refer back to our example in figure 3 and we can see how the gene tree in figure 4 arises within the species tree. Lineages C and D meet at time t_1 , but due to the different rates

$$d_3 = \alpha(\gamma_C \tau_1 + \gamma_{CD}(t_1 - \tau_1))$$

Site pattern	c_0	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	c_{11}	c_{12}
XXXX	1	3	3	3	3	3	3	9	6	6	6	6	12
XXXY	1	3	-1	3	-1	3	-1	-3	6	-2	-2	-2	-4
XXYX	1	3	-1	-1	3	-1	3	-3	-2	6	-2	-2	-4
XYXX	1	-1	3	3	3	-1	-1	-3	-2	-2	6	-2	-4
YXXX	1	-1	3	-1	-1	3	3	-3	-2	-2	-2	6	-4
XXYY	1	3	3	-1	-1	-1	-1	9	-2	-2	-2	-2	-4
XYXY	1	-1	-1	3	-1	-1	3	1	-2	-2	-2	-2	4
XYXX	1	-1	-1	-1	3	3	-1	1	-2	-2	-2	-2	4
XXYZ	1	3	-1	-1	-1	-1	-1	-3	-2	-2	2	2	4
XYXZ	1	-1	-1	3	-1	-1	-1	1	-2	2	-2	2	0
XYZX	1	-1	-1	-1	3	-1	-1	1	2	-2	-2	2	0
YXXZ	1	-1	-1	-1	-1	3	-1	1	-2	2	2	-2	0
YXZX	1	-1	-1	-1	-1	-1	3	1	2	-2	2	-2	0
YZXX	1	-1	3	-1	-1	-1	-1	-3	2	2	-2	-2	4
XYZW	1	-1	-1	-1	-1	-1	-1	1	2	2	2	2	-4

Table 1: Coefficients for each gene tree site pattern probability

$$d_4 = \alpha(\gamma_D\tau_1 + \gamma_{CD}(t_1 - \tau_1)).$$

The term α arises from conversion to mutation units and for the Jukes-Cantor model is equal to $4/3$. Likewise, it can be shown that:

$$d_1 = \alpha(\gamma_A\tau_2 + \gamma_{AB}(\tau_3 - \tau_2) + (t_3 - \tau_3) + (t_3 - t_2))$$

$$d_2 = \alpha(\gamma_B\tau_2 + \gamma_{AB}(\tau_3 - \tau_2) + (t_2 - \tau_3))$$

$$d_5 = \alpha(\gamma_{CD}(\tau_3 - t_1) + (t_2 - t_1))$$

The additional term in d_1 come from ignoring the root, so the path for d_1 has to go “up” to the root and then back “down” to t_2 . Note we do not need a rate multiplier for the terms above the root of the species tree because we have defined the relative rates in reference to the rate of unity above the root. (Also, under the strict molecular clock all $\gamma_k = 1$ so $d_1 = 2t_3 - t_2$, $d_2 = t_2$, $d_3 = d_4 = t_1$, and $d_5 = t_2 - t_1$.)

We will continue the example by considering one term of equation 5, showing how this term is found, and the results of entering it as an input into the integral result equation 2. The term we will expand is $c_7 e^{-(d_1+d_2+d_3+d_4)}$. Then we have (ignoring the constant c_7):

$$\begin{aligned} & e^{-(d_1+d_2+d_3+d_4)} \\ &= e^{(\alpha(\gamma_A\tau_2+\gamma_{AB}(\tau_3-\tau_2)+(t_3-\tau_3)+(t_3-t_2)+\gamma_B\tau_2+\gamma_{AB}(\tau_3-\tau_2)+(t_2-\tau_3)+\gamma_C\tau_1+\gamma_{CD}(t_1-\tau_1))+\gamma_D\tau_1+\gamma_{CD}(t_1-\tau_1))} \\ &= e^{\alpha(\tau_1(\gamma_C+\gamma_D-2\gamma_{CD})+\tau_2(\gamma_A-\gamma_B-2\gamma_{AB})+2\tau_3(\gamma_{AB}-1)+2t_1\gamma_{CD}+2t_3)} \\ &= e^{\alpha(\tau_1(\gamma_C+\gamma_D-2\gamma_{CD})+\tau_2(\gamma_A-\gamma_B-2\gamma_{AB})+2\tau_3(\gamma_{AB}-1))} e^{2\alpha t_1\gamma_{CD}} e^{2\alpha t_3} \end{aligned}$$

It can be seen then that the terms of equation 5 can be expressed by means of exponential functions of the coalescent times multiplied by a constant:

$$g_0 e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)}$$

which can then be used in the integrals from the previous section to find the probability integrating over the gene tree histories. For the term in our example, $g_0 = e^{\alpha(\tau_1(\gamma_C+\gamma_D-2\gamma_{CD})+\tau_2(\gamma_A-\gamma_B-2\gamma_{AB})+2\tau_3(\gamma_{AB}-1))}$, $g_1 = 2\alpha\gamma_{CD}$, $g_2 = 0$, $g_3 = 2\alpha$. This can be repeated for all the terms of equation 5 for all histories, and the results are given in the supplementary material.

Plugging these results into equation 2 then gives the following:

$$\frac{e^{\alpha(\tau_1(\gamma_C+\gamma_D-2\gamma_{CD})+\tau_2(\gamma_A-\gamma_B-2\gamma_{AB})+2\tau_3(\gamma_{AB}-1))}}{(1+\alpha\gamma_{CD}\theta)(1+\alpha\theta)(3+\alpha\theta)} e^{-\frac{2(\tau_3-\tau_2)}{\theta}} e^{-2\alpha\tau_3} (e^{-2\alpha\gamma_{CD}\tau_1} - \exp[-\frac{2}{\theta}(\tau_3-\tau_1)-2\alpha\gamma_{CD}\tau_3])$$

Repeating for all gene tree histories and terms in equation 5 gives the final results in the supplementary material. This example calculation corresponds to the k_{114} term, where we can see $k_{114} = 2c_7$ because identical computations are performed for gene history 14.

If the unrooted gene tree is not of the topology $(A, B|C, D)$ we need to permute the data. Figure 5 shows an example of how this can be done for one site pattern probability. Let G_U and G_L describe the upper and lower gene trees. Then given these phylogenies,

$$P(XXXY|G_U) = P(XXYX|G_L).$$

Table 2 gives the permutation matrix for when the gene tree is not $(A, B|C, D)$. For example, let the gene tree of concern be $(C, A|B, D)$ and we want to find the coefficients for equation 5 for site pattern XYYX.

Figure 5: Site pattern probabilities given topologies other than $(A, B|C, D)$ can be found by permuting as needed. G_U – upper tree. G_L – lower tree

We go down the $(C, A|B, D)$ column of table 2 until we find XYYX and then over to the $(A, B|C, D)$ column and find this corresponds to XXYY. Then from table 1 we use the coefficients from the XXYY row in equation 5.

3.3 Implementation

Four programs, all written in C++ are provided at <https://github.com/richards-1227/noClock>. The first performs the calculations that produced equation 3. The second conducts the linear algebra that results in table 4. The third produces the matrices of k values for the final result shown in the supplementary material. In addition, to prevent copying and editing errors, it outputs as strings the exact code used to instantiate the k matrix arrays used in the fourth program. The fourth program computes the final probabilities, and is included both as raw C++ code and a Mac and Unix executables. To run the program, enter:

```
./NoClock 1  $\tau_1$   $\tau_2$   $\tau_3$   $\gamma_A$   $\gamma_B$   $\gamma_C$   $\gamma_D$   $\gamma_{CD}$   $\gamma_{AB}$   $\theta$ 
```

for a symmetric tree and

$A, B C, D$	$A, C B, D$	$A, D B, C$	$B, A C, D$	$B, C A, D$	$B, D A, C$	$C, A B, D$	$C, B A, D$	$C, D A, B$	$D, A B, C$	$D, B A, C$	$D, C A, B$
XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX	XXXX
XXXY	XXXY	XYXX	XXXY	XXXY	XYXX	XXXY	XXXY	XYXX	YXXX	YXXX	YXXX
XXYX	XYXX	XXXY	XXYX	XYXX	XXXY	YXXX	YXXX	YXXX	XXXY	XXXY	XYXX
XYXX	XXYX	XXYX	YXXX	YXXX	YXXX	XXYX	XYXX	XXXY	XXYX	XYXX	XXXY
YXXX	YXXX	YXXX	XYXX	XXYX	XXYX	XYXX	XXYX	XXYX	XYXX	XXYX	XXYX
XXYY	XYXY	XYXY	XXYY	XYXY	XYXY	XYXY	XYXY	XXYY	XYXY	XYXY	XXYY
XYXY	XXYY	XYXY	XYXY	XYXY	XXYY	XXYY	XYXY	XYXY	XYXY	XXYY	XYXY
XYXY	XYXY	XXYY	XYXY	XXYY	XYXY	XYXY	XXYY	XYXY	XXYY	XYXY	XYXY
XXYZ	XYXZ	XYXZ	XXYZ	XYXZ	XYXZ	YXXZ	YXXZ	YXXZ	YXXZ	YXXZ	YXXZ
XYXZ	XXYZ	XYXZ	YXXZ	YXXZ	YXXZ	XXYZ	XYXZ	XYXZ	YXXZ	YXXZ	YXXZ
XYZX	XYZX	XXYZ	YXXZ	YXXZ	YXXZ	YXXZ	YXXZ	YXXZ	XXYZ	XYXZ	XYXZ
YXXZ	YXXZ	YXXZ	XYXZ	XXYZ	XYXZ	XYXZ	XXYZ	XYXZ	YXXZ	YXXZ	YXXZ
YXZX	YXZX	YXXZ	XYZX	XYZX	XXYZ	YXXZ	YXZX	YXZX	XYXZ	XXYZ	XYZX
YXZX	YXZX	YXZX	YXZX	YXZX	YXZX	XYZX	XYZX	XXYZ	XYZX	XYZX	XXYZ
XYZW	XYZW	XYZW	XYZW	XYZW	XYZW	XYZW	XYZW	XYZW	XYZW	XYZW	XYZW

Table 2: Site pattern permutation matrix

`./NoClock 0  $\tau_1$   $\tau_2$   $\tau_3$   $\gamma_A$   $\gamma_B$   $\gamma_C$   $\gamma_D$   $\gamma_{CD}$   $\gamma_{BCD}$   $\theta$`

for an asymmetric tree. Unlike the calculations in this paper, τ is parameterized in coalescent units, which are mutation units scaled by a factor of $\theta/2$. The program outputs the vector of site pattern probabilities in the order given by table 4.

To check the accuracy of the implementation, we first generated site pattern probabilities for different branch lengths with γ equal to a vector of ones, representing a strict molecular clock. We compared that output to existing code which implements the results of [3] and confirmed that the results were identical.

Tian and Kubatko (2020) developed code that simulates the coalescent process for a 4-taxon species tree and then adjusts the gene tree lengths to allow for a relaxed clock. We simulated 1000 different species trees with various relative mutation rates, and then took the gene trees as inputs for simulating sequence evolution using seq-gen [17]. We then performed a Chi-squared goodness-of-fit test comparing the expected to actual counts of each site pattern and plotted the p-values (see figure 6). The results are consistent with having determined the correct site pattern probabilities, as it is improbable that both our implementation and the Tian and Kubatko implementation made simultaneous parallel errors resulting in the same probability distribution.

Figure 6: Distribution of 1000 goodness-of-fit test of our site pattern probabilities against simulated data generated by [32] and [17].

4 Varying mutation rates within branches

We next examine the case where the relative mutation rate, rather than being constant on a branch, is allowed to vary over time. For pendant branches, this distinction is unimportant, as coalescent events cannot happen on pendant branches of the species tree when there is only one individual per species. Thus, if the time-varying mutation rate on the pendant branch leading to species d is given by $\gamma_D(u)$, the total expected mutation is $\int_0^{\tau_1} \gamma_D(u) du$, which leads to the same probability distribution as a constant rate of $\gamma_D = \frac{1}{\tau_1} \int_0^{\tau_1} \gamma_D(u) du$.

For internal branches, it may be important to carefully consider differing relative rates at differing times. We could again try to replace the time-varying rates with an average rate as with pendant branches, for example $\gamma_{CD} = \frac{1}{(\tau_2 - \tau_1)} \int_{\tau_1}^{\tau_2} \gamma_{CD}(u) du$. To see why this would fail to produce the correct site pattern probabilities, consider an asymmetric species tree $(a, (b, (c, d)))$ where instead of a constant rate on the branch cd , the rate jumps halfway between τ_1 and τ_2 (see figure 7). We can let the two rates be γ_{CD1} and γ_{CD2} looking back from the speciation time between species c and d toward the root. Intuitively, if $\gamma_{CD1} > \gamma_{CD2}$ there will be more mutation before the two lineages coalesce and *ceteris paribus* we will expect more XXXY and XXYX sites and fewer XXYY sites. Conversely, if $\gamma_{CD1} < \gamma_{CD2}$ we have more

Figure 7: Effect of changing rates in internal branches. $\gamma_{CD1} > \gamma_{CD2}$ will tend to produce more sites with the XXXY and XXYX patterns than taking a weighted average of γ_{CD1} and γ_{CD2} . Conversely, $\gamma_{CD1} < \gamma_{CD2}$ will tend to produce more XXYY sites.

XXYY and fewer XXXY and XXYX sites than if we treated the branch as having a single rate γ_{CD} .

Unfortunately, we cannot examine the full range of functional forms for $\gamma_{CD}(u)$ and other relative rates on the internal branches. Even a relatively simple linear form such as $\gamma_{CD}(u) = u - \tau_1$ founders on the fact that after integrating over u we are left with a $e^{-t_1^2}$ term, which famously has no known antiderivative. We thus limit our exploration of the possible relative rates on internal branches to step functions. This allows us to generalize our work from section 3 to find the site pattern probabilities, which also serve as a limiting approximation for any time-varying rate function that is continuous almost everywhere.

Unlike in the single rate case, we do not have a single algebraic formula that we can present for the site pattern probabilities. We will, however, show how the calculation steps of section 3 can be generalized, and discuss in detail our implementation program that calculates the site pattern probabilities given an input file of relative rates at different parts of the tree.

We use the following notation. Let ψ represent the series of breakpoints between different rates, where for example ψ_{CDi} is the i^{th} breakpoint on the branch above the speciation point between species c and d . Then, if there are I different rates on that branch, we have $\psi_{CD0} = \tau_1, \psi_{CD1}, \dots, \psi_{CDI} = \tau_2$. Then $\gamma_{j,i}$ describes the relative mutation rate on branch j between $\psi_{j,i-1}$ and $\psi_{j,i}$.

4.1 Generalizing the gene tree density integrals

Looking at the example in section 3 and the densities in the appendix, we see that each density is the product of two terms: a function of θ and τ that is a constant with respect to the coalescent times t , and a function

of t of the form $\exp[-(a_1 t_1 + a_2 t_2 + a_3 t_3)]$. Now, though, instead of 25 or 34 different gene tree histories (for symmetric and asymmetric species trees) we have additional histories depending on which breakpoints the coalescent events fall between. In the example in figure 3 if we had two different rates between τ_1 and τ_3 separated by ψ_{CD1} , histories 1, 2 10, and 14 would need to be divided into separate cases for t_1 before or after ψ_{CD1} . Examining the limits of integration, we only need to consider four general cases (using $\dot{=}$ to indicate two times fall in the same relative rate category and $t_1 \dot{<} t_2$ to indicate that a breakpoint falls between t_1 and t_2):

- $t_1 \dot{<} t_2 \dot{<} t_3$
- $t_1 \dot{<} t_2 \dot{=} t_3$
- $t_1 \dot{=} t_2 \dot{<} t_3$
- $t_1 \dot{=} t_2 \dot{=} t_3$

As before, we include the $g_0 e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)}$ term from the the mutation process to get an integral of the form $\int \int g_0 f(\theta, \tau) \exp[-(t_1(a_1 + g_1)) - (t_2(a_2 + g_2)) - (t_3(a_3 + g_3))]$. Full details are given in the appendix.

As a final note, for a symmetric gene tree, we also need to consider cases when $t_2 \dot{<} t_1$ and when $t_1 \dot{=} t_2$ and $t_2 < t_1$. Since, by the assumptions of the coalescent model, the two cases are exchangeable and occur with equal probability, we can simply replace t_1 with t_2 and vice versa in the evaluation of the integral.

4.2 Generalizing the mutation probability term $g_0 e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)}$

We now take a second look at the distance terms from the example in section 3. Taking a look at d_1 we have

$$\begin{aligned} d_1 &= \alpha(\gamma_A \tau_2 + \gamma_{AB}(\tau_3 - \tau_2) + (t_3 - \tau_3) + (t_3 - t_2)) \\ &= \alpha(\tau_2(\gamma_A - \gamma_{AB}) + \tau_3(\gamma_{AB} - 1) + 2t_3 - t_2) \end{aligned}$$

The first part of this result, $\alpha(\tau_2(\gamma_A - \gamma_{AB}) + \tau_3(\gamma_{AB} - 1))$, which is the contribution of this term to g_0 , is α times each of the break points times the difference in the relative mutation rates before and after the break point, or $\psi_{j,i}(\gamma_{j,i-1} - \gamma_{j,i})$. (In the event that $i = 0$ we can define $\gamma_{j,i-1}$ to be the final rate of the branch below j .) The second part of the result, $\alpha(2t_3 - t_2)$, gives the contribution to g_2 and g_3 from the d_1 term.

Taken together, we have

$$\delta = \sum_{G \in G_S} \sum_{t_3 \in \mathcal{T}_3} \sum_{t_2 \in \mathcal{T}_2} \sum_{t_1 \in \mathcal{T}_1} \int \int \int g_0 f(\theta, \tau) \exp[-(t_1(a_1 + g_1)) - (t_2(a_2 + g_2)) - (t_3(a_3 + g_3))] \quad (6)$$

where $\mathcal{T}_i$ represents the sets of relative rate locations t_i can occur in (for example, in figure 3 t_1 can occur in the cd branch or above the root, but not on the ab branch).

4.3 Implementing the step change model

We built a C++ program implementing equation 6 and located at <https://github.com/richards-1227/noClock>. We have included Mac and Unix executables as well as the C++ source code. Interested readers are encouraged to follow the source code along with the discussion in this section to see how the mathematical formulae are implemented.

Execution is via the command

```
./NoClockSC symInd fileName  $\theta$ 
```

where symInd = 0 for an asymmetric tree and 1 for a symmetric tree. “fileName” is the name of a text file containing the sets of rates and breakpoints in the following format:

```
branch start_time end_time rate
```

For example, for a case where the ab branch has a single rate and the cd and root branches have two different rates each:

A	0	τ_2	γ_A
B	0	τ_2	γ_B
C	0	τ_1	γ_C
D	0	τ_1	γ_D
AB	τ_2	τ_3	1.0
CD	τ_1	2.5	2.0
CD	2.5	τ_3	1.5
root	τ_3	5.0	0.5
root	5.0	1000.0	1.0

The final time above the root must be chosen such that $e^\psi \approx 0$. We have used 1000.0.

Pseudocode for implementing the calculations is shown below.

Algorithm 1 Site pattern probability calculation procedure

---

for Each gene tree topology **do**

for $t_1 \in \mathcal{T}_1$ **do**

for $t_2 \in \mathcal{T}_2$ **do**

for $t_3 \in \mathcal{T}_3$ **do**

                Get limits of integration for each of t_1 , t_2 , and t_3 ;

                Find correct integration form (e.g., $t_1 < t_2 < t_3$);

                Calculate d_1 , d_2 , d_3 , d_4 , and d_5 ;

for Each term in equation 5 **do**

                    Find g_0 , g_1 , g_2 , and g_3 from d_1 , d_2 , etc.;

                    Input into integral, apply coefficient multiplier from table 1 as modified by permutation matrix from table 2;

end for

end for

end for

end for

end for

Again, we confirmed the accuracy of the code. First, we used a single rate category for each internal branch and verified that the program produced the same results as the program in section 3. Second, we modified the simulation code of Tian and Kubatko (2017) to have varying rates along the internal branches and performed goodness of fit tests to check that the simulation site pattern frequencies fit our calculated site pattern probabilities within the bounds of normal sampling error.

5 Quantifying impact of varying γ on site pattern probabilities

To begin to see the impact of different relative mutation rates, we start with the asymmetric species tree ($a : 1.0, (b : 0.6, (c : 0.3, d : 0.3) : 0.3) : 0.4$). We compare the site pattern probabilities under the strict molecular clock to the case where $\gamma_D = 7.5$. For simplicity, in this section, we will assume constant rates

Site pattern	XXXX	XXXY	XXYX	XYXX	YXXX	XXYY	XYXY	XXYY	XXYZ	XYXZ	XYZX	YXXZ	YXZX	YZXX	XYZW
Clock	0.990	1.56e-3	1.56e-3	2.01e-3	3.20e-3	8.15e-4	3.69e-4	3.69e-4	2.92e-6	2.73e-6	2.73e-6	4.27e-6	4.27e-6	6.74e-6	1.06e-8
Non-clock	0.987	4.45e-3	1.56e-3	2.01e-3	3.19e-3	8.14e-4	3.70e-4	3.71e-4	7.55e-6	7.35e-6	2.73e-6	1.12e-5	4.27e-6	6.73e-6	2.39e-8
Change	-2.89e-3	2.89e-3	-3.77e-6	-5.51e-6	-8.98e-6	-8.54e-7	8.83e-7	2.04e-6	4.63e-6	4.63e-6	2.77e-9	6.94e-6	-3.25e-9	1.29e-8	1.33e-8

Table 3: Effect of a change in γ_D from 1.0 to 7.5 on the site pattern probabilities for the tree $(a : 1.0, (b : 0.6, (c : 0.3, d : 0.3) : 0.3) : 0.4)$.

on internal branches as in section 3. The results can be seen in table 3. We can see that δ_{XXXX} decreases by $-2.89e - 3$ and δ_{XXXY} increases by the same amount. Similarly, the increase in δ_{XXYZ} is equal to the sum of decreases in δ_{XXYX} and δ_{XXYY} . These changes can be explained by an increase in the number of mutations of the nucleotide of species d , which makes sense given the increases in γ_D lengthens the pendant branch leading to species d . These changes in site pattern probabilities can be summarized by:

$$\begin{aligned}
 \delta_{XXXX} &\Rightarrow \delta_{XXXY} \\
 \delta_{XXYX}, \delta_{XXYY} &\Rightarrow \delta_{XXYZ} \\
 \delta_{XYXX} &\Rightarrow \delta_{XYXY}, \delta_{XYXZ} \\
 \delta_{YXXX} &\Rightarrow \delta_{XYXX}, \delta_{YXXZ} \\
 \delta_{YXZX}, \delta_{YZXX} &\Rightarrow \delta_{XYZX}, \delta_{XYZW}
 \end{aligned}$$

Next, we look at changes to γ_D along a spectrum of values for the species tree $((a : 0.6, b : 0.6) : 0.4, (c : 0.3, d : 0.3) : 0.7)$ as shown in figure 8. Again, we see less of the XXXX pattern and more of XYZW, which makes sense as increasing γ_D increases the total size of the tree. Patterns XXYY, XYXY, and XYYX appear to be largely unaffected. In the XXXY, XXYX, XYXX, and YXXX group, we see an increase to where XXXY overtakes XXYX/YXXX. We see a similar change in rank ordering in the other group of patterns. Because we only see a major change in two of the groupings, we focus on those in the remaining plots.

Figure 9 shows the effect of changing γ_D for different lengths of species d 's pendant branch. We see that the longer the pendant branch, the greater the impact of γ_D changes on the rank ordering of site pattern probabilities.

Figures 10 and 11 show the effect of changing the relative rates on different branches of the tree. For the symmetric tree $((a : 0.6, b : 0.6) : 0.4, (c : 0.3, d : 0.3) : 0.7)$, there is a slightly greater change to the site

Figure 8: Effects of changing γ_D on species tree $((a : 0.6, b : 0.6) : 0.4, (c : 0.3, d : 0.3) : 0.7)$. Site patterns are grouped by similar magnitudes of the site pattern probabilities.

pattern probabilities for a change in γ_{CD} than for γ_D . This tree does have a much longer internal vs. pendant branch (0.7 vs. 0.3 coalescent units) so that explains much of this difference. We also see that changing γ_{CD} increases both δ_{XXYX} and δ_{XXXY} whereas γ_D does not increase δ_{XXYX} . For the asymmetric tree $(a : 1.0, (b : 0.6, (c : 0.3, d : 0.3) : 0.3) : 0.4)$, changing γ_B has a much greater impact on δ_{XYXX} than changing γ_D has on δ_{XXXY} . It is somewhat interesting that changing the branch on the internal rates γ_{CD} or γ_{BCD} causes as quadratic rather than a linear response on some of the site pattern probabilities such as δ_{XXYZ} , δ_{YZXX} . The deviation from linearity is small and we have no intuition for why the response is different, but it is worth noting.

5.1 Applications

In Richards and Kubatko (2021), it was shown that their method, Lily-T, would systematically infer the incorrect tree if the relative rates a branch were too high, with greater sensitivity to changes on the pendant branches. We considered another way in which clock violations could interfere with inference. Methods such as ABBA-BABA [6] and HyDe [10] use the relative frequencies of the site patterns XXYX, XYYX,

Figure 9: Interaction of branch length with γ_D .

Figure 10: Effect of changing the relative rate on a pendant vs. an internal branch for the symmetric species tree: $((a : 0.6, b : 0.6) : 0.4, (c : 0.3, d : 0.3) : 0.7)$

Figure 11: Effect of changing the relative rate on various branches for the asymmetric species tree: (a : 1.0, (b : 0.6, (c : 0.3, d : 0.3) : 0.3) : 0.4)

and XXXX to infer whether or not hybridization occurs. The tests compare the hypotheses

$$H_0 : \text{No hybridization } (\xi = 0)$$

and

$$H_1 : c \text{ is a hybrid taxon of species } b \text{ and } d \ (\xi > 0)$$

using test statistics that follow the standard normal distribution. Surprisingly, changing the relative rate of any branch had little effect on inference. Even with the tree $(a : 1.0, (b : 0.6, (c : 0.5, d : 0.5) : 0.1) : 0.4)$, $\gamma_D = 7.5$ and 240,000 sites, the expected value of the test statistic only moved from 0 to 0.03. In fact, we had to change the tree to $(a : 10.0, (b : 9.5, (c : 9.0, d : 9.0) : 0.5) : 0.5)$, $\gamma_D = 10.0$, and 2.4 million sites before we would expect to falsely reject the null hypothesis of no hybridization.

So why is the Lily procedure so sensitive to violations of the clock where the ABBA-BABA and HyDe tests are relatively robust? The difference lies in the fact that the relative frequencies of $\{XXXXY, XXYX, XYXX, YXXX\}$ have a large contribution to the likelihood calculations Lily uses to make species tree inference. By contrast, as we saw in figure 8 and table 3 the relative frequencies of $\{XXYY, XYXY, XYYX\}$ are less sensitive to violations of the strict molecular clock, so methods that only depend on the relative frequencies of these three site patterns are more robust against clock violations.

6 Conclusions and future work

In this paper we provide complete site pattern probabilities under the multispecies coalescent model and the Jukes-Cantor substitution model, allowing for differing mutation rates in different branches of the species tree. Our work opens up two new avenues of research. The first is in expanding the probability calculations performed here to higher numbers of taxa, either with or without the molecular clock. Although the theoretical basis for this extension is well-known, the extensive computation required has prevented this in practice. Our framework for breaking down the site pattern probabilities arising from equation 3 can be easily expanded to higher numbers of taxa, and we expect that when the gene tree histories are enumerated we would again see a small enough number of integral forms and combinations of inputs as in table 4 to enable direct calculation of the probabilities.

This would open up improvements on existing assembly methods for species tree inference. Assembly methods such as [1], [21], [24], [25], [29], [34], [23], [26], [31], and [30] begin with a triplet or quartet of

species, infer the topology of these quartets, and use them as inputs to an assembly algorithm. These inputs are treated as independent, although for an n -taxon tree there is statistical dependence due to any branches the quartets share in common. Expanding the exact site pattern probabilities to 5 or 6 taxa would allow at least pairwise statistical dependence between triplets to be determined, which should improve both point estimation and uncertainty quantification.

It is worth noting that the invariants we see with gene trees do not occur with species trees under the multivariate coalescent. For example, for the gene tree $((A, B), (C, D))$ the two sides of the tree are independent such that $P(i_A = i_B \cap i_C = i_D) = P(i_A = i_B) \times P(i_C = i_D)$. This does not hold for the species tree $((a, b), (c, d))$ even under the molecular clock. (This can be shown from the site pattern probabilities in the supplemental material of [3] by marginalizing over the taxa). This is perhaps counterintuitive, but arises out of the fact that for any finite τ_3 , there is some possibility that the gene tree and species tree don't match, and therefore there are some sites where the characters on either side of the root of the species tree are no longer independent (because they arise from gene trees whose tips are no longer independent).

The results of this paper should also enable improving quartet inference. Lily-Q [19] is a procedure that takes $P(D = d | (S, \tau))$ and assuming a distribution on (S, τ) , produces posterior probabilities on S for groups of 4 taxa. Given either a distribution on γ or a means of estimating $\hat{\gamma}$, we should be able to develop an extension on Lily-Q to the relaxed-clock case, providing a means for fast, accurate species tree quartet inference in cases where the molecular clock does not hold.

7 Appendix

7.1 Gene tree histories and integrals – symmetric case

Form A

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_1 - \tau_1) - \frac{2}{\theta}(t_2 - \tau_2) - \frac{2}{\theta}(t_3 - \tau_3)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_2}^{\tau_3} \int_{\tau_1}^{\tau_3} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \left(\frac{2}{2+g_1\theta}\right) \left(\frac{2}{2+g_2\theta}\right) \left(\frac{2}{2+g_3\theta}\right) e^{-g_3\tau_3} (e^{-g_1\tau_1} - \exp[-\frac{2}{\theta}(\tau_3 - \tau_1) - g_1\tau_3]) (e^{-g_2\tau_2} - \exp[\frac{2}{\theta}(\tau_3 - \tau_2) - g_2\tau_3])$$

Form B

Index	Gene Tree History Topology	Time constraints	Integral Form
1	$((A, B), (C, D))$	$t_1 < \tau_3, t_2 < \tau_3$	A
2	$((A, B), (C, D))$	$t_1 < \tau_3 < t_2$	B
3	$((A, B), (C, D))$	$t_2 < \tau_3 < t_1$	C
4	$((A, B), (C, D))$	$\tau_3 < t_1 < t_2$	D
5	$((A, B), (C, D))$	$\tau_3 < t_2 < t_1$	E
6	$((A, C), (B, D))$	$\tau_3 < t_1 < t_2$	D
7	$((A, C), (B, D))$	$\tau_3 < t_2 < t_1$	E
8	$((A, D), (B, C))$	$\tau_3 < t_1 < t_2$	D
9	$((A, D), (B, C))$	$\tau_3 < t_2 < t_1$	E
10	$(A, (B, (C, D)))$	$t_1 < \tau_3 < t_2$	B
11	$(A, (B, (C, D)))$	$\tau_3 < t_1 < t_2$	D
12	$(A, (C, (B, D)))$	$\tau_3 < t_1 < t_2$	D
13	$(A, (D, (B, C)))$	$\tau_3 < t_1 < t_2$	D
14	$(B, (A, (C, D)))$	$t_1 < \tau_3 < t_2$	B
15	$(B, (A, (C, D)))$	$\tau_3 < t_1 < t_2$	D
16	$(B, (C, (A, D)))$	$\tau_3 < t_1 < t_2$	D
17	$(B, (D, (A, C)))$	$\tau_3 < t_1 < t_2$	D
18	$(C, (A, (B, D)))$	$\tau_3 < t_1 < t_2$	D
19	$(C, (B, (A, D)))$	$\tau_3 < t_1 < t_2$	D
20	$(C, (D, (A, B)))$	$t_1 < \tau_3 < t_2$	F
21	$(C, (D, (A, B)))$	$\tau_3 < t_1 < t_2$	D
22	$(D, (A, (B, C)))$	$\tau_3 < t_1 < t_2$	D
23	$(D, (B, (A, C)))$	$\tau_3 < t_1 < t_2$	D
24	$(D, (C, (A, B)))$	$t_1 < \tau_3 < t_2$	F
25	$(D, (C, (A, B)))$	$\tau_3 < t_1 < t_2$	D

Table 4: Symmetric species tree gene histories

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_1 - \tau_1) - \frac{2}{\theta}(\tau_3 - \tau_2) - \frac{6}{\theta}(t_2 - \tau_3) - \frac{2}{\theta}(t_3 - t_2)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_1}^{\tau_3} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \frac{8}{(2+g_1\theta)(2+g_3\theta)(6+g_2\theta+g_3\theta)} e^{-\frac{2(\tau_3-\tau_2)}{\theta}} e^{-\tau_3(g_2+g_3)} (e^{-g_1\tau_1} - \exp[-\frac{2}{\theta}(\tau_3 - \tau_1) - g_1\tau_3])$$

Form C

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_2 - \tau_2) - \frac{2}{\theta}(\tau_3 - \tau_1) - \frac{6}{\theta}(t_1 - \tau_3) - \frac{2}{\theta}(t_3 - t_1)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_2}^{\tau_3} \int_{\tau_3}^{t_3} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \frac{8}{(2+g_2\theta)(2+g_3\theta)(6+g_1\theta+g_3\theta)} e^{-\frac{2(\tau_3-\tau_1)}{\theta}} e^{-\tau_3(g_1+g_3)} (e^{-g_2\tau_2} - \exp[-\frac{2}{\theta}(\tau_3 - \tau_2) - g_2\tau_3])$$

Form D

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(\tau_3 - \tau_1) - \frac{2}{\theta}(\tau_3 - \tau_2) - \frac{12}{\theta}(t_1 - \tau_3) - \frac{6}{\theta}(t_2 - t_1) - \frac{2}{\theta}(t_3 - t_2)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_3}^{t_2} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \frac{8e^{-\tau_3(g_1+g_2+g_3)} e^{-\frac{2(\tau_3-\tau_1)}{\theta}} e^{-\frac{2(\tau_3-\tau_2)}{\theta}}}{(2+g_3\theta)(6+g_2\theta+g_3\theta)(12+g_1\theta+g_2\theta+g_3\theta)}$$

Form E

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(\tau_3 - \tau_1) - \frac{2}{\theta}(\tau_3 - \tau_2) - \frac{12}{\theta}(t_2 - \tau_3) - \frac{6}{\theta}(t_1 - t_2) - \frac{2}{\theta}(t_3 - t_1)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_3}^{t_1} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_2 dt_1 dt_3$$

$$\text{Result: } g_0 \frac{8e^{-\tau_3(g_1+g_2+g_3)} e^{-\frac{2(\tau_3-\tau_1)}{\theta}} e^{-\frac{2(\tau_3-\tau_2)}{\theta}}}{(2+g_3\theta)(6+g_1\theta+g_3\theta)(12+g_1\theta+g_2\theta+g_3\theta)}$$

Form F

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_1 - \tau_2) - \frac{2}{\theta}(\tau_3 - \tau_1) - \frac{6}{\theta}(t_2 - \tau_3) - \frac{2}{\theta}(t_3 - t_2)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_2}^{\tau_3} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \frac{8}{(2+g_1\theta)(2+g_3\theta)(6+g_2\theta+g_3\theta)} e^{-\frac{2(\tau_3-\tau_1)}{\theta}} e^{-\tau_3(g_2+g_3)} (e^{-g_1\tau_2} - \exp[-\frac{2}{\theta}(\tau_3 - \tau_2) - g_1\tau_3])$$

7.2 Gene tree histories and integrals – asymmetric case

Form A

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_1 - \tau_1) - \frac{2}{\theta}(\tau_3 - \tau_2) - \frac{6}{\theta}(t_2 - \tau_3) - \frac{2}{\theta}(t_3 - t_2)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_1}^{\tau_2} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

Index	Gene History Topology	Time constraints	Integral Form
1	$((A, B), (C, D))$	$t_1 < \tau_2 < \tau_3 < t_2$	A
2	$((A, B), (C, D))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
3	$((A, B), (C, D))$	$\tau_3 < t_1 < t_2$	C
4	$((A, B), (C, D))$	$\tau_3 < t_2 < t_1$	D
5	$((A, C), (B, D))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
6	$((A, C), (B, D))$	$\tau_3 < t_1 < t_2$	C
7	$((A, C), (B, D))$	$\tau_3 < t_2 < t_1$	D
8	$((A, D), (B, C))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
9	$((A, D), (B, C))$	$\tau_3 < t_1 < t_2$	C
10	$((A, D), (B, C))$	$\tau_3 < t_2 < t_1$	D
11	$(A, (B, (C, D)))$	$t_1 < \tau_2 < t_2 < \tau_3$	E
12	$(A, (B, (C, D)))$	$t_1 < \tau_2 < \tau_3 < t_2$	A
13	$(A, (B, (C, D)))$	$\tau_2 < t_1 < t_2 < \tau_3$	F
14	$(A, (B, (C, D)))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
15	$(A, (B, (C, D)))$	$\tau_3 < t_1$	C
16	$(A, (C, (B, D)))$	$\tau_2 < t_1 < t_2 < \tau_3$	F
17	$(A, (C, (B, D)))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
18	$(A, (C, (B, D)))$	$\tau_3 < t_1$	C
19	$(A, (D, (B, C)))$	$\tau_2 < t_1 < t_2 < \tau_3$	F
20	$(A, (D, (B, C)))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
21	$(A, (D, (B, C)))$	$\tau_3 < t_1$	C
22	$(B, (A, (C, D)))$	$t_1 < \tau_2 < \tau_3 < t_2$	A
23	$(B, (A, (C, D)))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
24	$(B, (A, (C, D)))$	$\tau_3 < t_1$	C
25	$(B, (C, (A, D)))$	$\tau_3 < t_1$	C
26	$(B, (D, (A, C)))$	$\tau_3 < t_1$	C

Table 5: Asymmetric species tree gene histories, pt.1

Index	Topology	Time constraints	Integral Form
27	$(C, (A, (B, D)))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
28	$(C, (A, (B, D)))$	$\tau_3 < t_1$	C
29	$(C, (B, (A, D)))$	$\tau_3 < t_1$	C
30	$(C, (D, (A, B)))$	$\tau_3 < t_1$	C
31	$(D, (A, (B, C)))$	$\tau_2 < t_1 < \tau_3 < t_2$	B
32	$(D, (A, (B, C)))$	$\tau_3 < t_1$	C
33	$(D, (B, (A, C)))$	$\tau_3 < t_1$	C
34	$(D, (C, (A, B)))$	$\tau_3 < t_1$	C

Table 6: Asymmetric species tree gene histories, pt.2

$$\text{Result: } g_0 \frac{8e^{-\tau_3(g_2+g_3)} e^{-\frac{2(\tau_3-\tau_2)}{\theta}} (e^{-g_1\tau_1} - \exp[-\frac{2(\tau_2-\tau_1)}{\theta} - g_1\tau_2])}{(2+g_1\theta)(2+g_3\theta)(6+g_2\theta+g_3\theta)}$$

Form B

$$\text{Density: } (\frac{2}{\theta})^3 \exp[-\frac{2}{\theta}(\tau_2 - \tau_1) - \frac{6}{\theta}(t_1 - \tau_2) - \frac{2}{\theta}(\tau_3 - t_1) - \frac{6}{\theta}(t_2 - \tau_3) - \frac{2}{\theta}(t_3 - t_2)]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_2}^{\tau_3} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \frac{8e^{-\frac{2(\tau_2-\tau_1)}{\theta}}}{(4+g_1\theta)(2+g_3\theta)(6+g_2\theta+g_3\theta)} (e^{-\frac{2(\tau_3-\tau_2)}{\theta}} e^{-\tau_3(g_2+g_3)} e^{-g_1\tau_2} - e^{-\frac{6(\tau_3-\tau_2)}{\theta}} e^{-\tau_3(g_1+g_2+g_3)})$$

Form C

$$\text{Density: } (\frac{2}{\theta})^3 \exp[-\frac{2}{\theta}(\tau_2 - \tau_1) - \frac{6}{\theta}(\tau_3 - \tau_2) - \frac{12}{\theta}(t_1 - \tau_3) - \frac{6}{\theta}(t_2 - t_1) - \frac{2}{\theta}(t_3 - t_2)]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_3}^{t_2} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \frac{8e^{-(g_1+g_2+g_3)\tau_3} \exp[-\frac{2}{\theta}(\tau_2-\tau_1) - \frac{6}{\theta}(\tau_3-\tau_2)]}{(2+g_3\theta)(6+g_2\theta+g_3\theta)(12+g_1\theta+g_2\theta+g_3\theta)}$$

Form D

$$\text{Density: } (\frac{2}{\theta})^3 \exp[-\frac{2}{\theta}(\tau_2 - \tau_1) - \frac{6}{\theta}(\tau_3 - \tau_2) - \frac{12}{\theta}(t_2 - \tau_3) - \frac{6}{\theta}(t_1 - t_2) - \frac{2}{\theta}(t_3 - t_1)]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_3}^{t_3} \int_{\tau_3}^{t_1} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_2 dt_1 dt_3$$

$$\text{Result: } g_0 \frac{8e^{-(g_1+g_2+g_3)\tau_3} \exp[-\frac{2}{\theta}(\tau_2-\tau_1) - \frac{6}{\theta}(\tau_3-\tau_2)]}{(2+g_3\theta)(6+g_1\theta+g_3\theta)(12+g_1\theta+g_2\theta+g_3\theta)}$$

Form E

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(t_1 - \tau_1) - \frac{2}{\theta}(t_2 - \tau_2) - \frac{2}{\theta}(t_3 - \tau_3)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_2}^{\tau_3} \int_{\tau_1}^{\tau_2} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \left(\frac{2}{2+g_1\theta}\right) \left(\frac{2}{2+g_2\theta}\right) \left(\frac{2}{2+g_3\theta}\right) e^{-g_3\tau_3} (e^{-g_1\tau_1} - \exp[-\frac{2}{\theta}(\tau_2 - \tau_1) - g_1\tau_2]) (e^{-g_2\tau_2} - \exp[\frac{2}{\theta}(\tau_3 - \tau_2) - g_2\tau_3])$$

Form F

$$\text{Density: } \left(\frac{2}{\theta}\right)^3 \exp\left[-\frac{2}{\theta}(\tau_2 - \tau_1) - \frac{6}{\theta}(t_1 - \tau_2) - \frac{2}{\theta}(t_2 - t_1) - \frac{2}{\theta}(t_3 - \tau_3)\right]$$

$$\text{Integral: } \int_{\tau_3}^{\infty} \int_{\tau_2}^{\tau_3} \int_{\tau_2}^{t_2} g_0 f(t_1, t_2, t_3) e^{-(g_1 t_1 + g_2 t_2 + g_3 t_3)} dt_1 dt_2 dt_3$$

$$\text{Result: } g_0 \frac{8e^{-\frac{2(\tau_2 - \tau_1)}{\theta}} e^{-g_3\tau_3}}{(2+g_3\theta)(4+g_1\theta)} \left(\frac{e^{-\tau_2(g_1+g_2)} - e^{-\frac{2(\tau_3 - \tau_2)}{\theta}} e^{-g_2\tau_3} e^{-g_1\tau_2}}{2+g_2\theta} - \frac{e^{-\tau_2(g_1+g_2)} - e^{-\frac{6(\tau_3 - \tau_2)}{\theta}} e^{-\tau_3(g_1+g_2)}}{6+g_1\theta+g_2\theta} \right)$$

7.3 General gene density integrals

7.3.1 $t_1 < t_2 < t_3$

$$\begin{aligned} & \int_{\psi_{3a}}^{\psi_{3b}} \int_{\psi_{2a}}^{\psi_{2b}} \int_{\psi_{1a}}^{\psi_{1b}} g_0 f(\theta, \tau) \exp[-(t_1(a_1 + g_1)) - (t_2(a_2 + g_2)) - (t_3(a_3 + g_3))] dt_1 dt_2 dt_3 \\ &= g_0 f(\theta, \tau) \prod_{i=1}^3 \frac{e^{-\psi_{ia}(a_i+g_i)} - e^{-\psi_{ib}(a_i+g_i)}}{a_i + g_i} \end{aligned}$$

7.3.2 $t_1 \doteq t_2 < t_3$

$$\begin{aligned} & \int_{\psi_{3a}}^{\psi_{3b}} \int_{\psi_{2a}}^{\psi_{2b}} \int_{\psi_{1a}}^{t_2} g_0 f(\theta, \tau) \exp[-(t_1(a_1 + g_1)) - (t_2(a_2 + g_2)) - (t_3(a_3 + g_3))] dt_1 dt_2 dt_3 \\ &= g_0 f(\theta, \tau) \frac{e^{-\psi_{3a}(a_3+g_3)} - e^{-\psi_{3b}(a_3+g_3)}}{(a_1 + g_1)(a_3 + g_3)} \\ & * \left(\frac{e^{-\psi_{2a}(a_1+a_2+g_1+g_2)} - e^{-\psi_{2a}(a_1+g_1)-\psi_{2b}(a_2+g_2)}}{a_2 + g_2} - \frac{e^{-\psi_{2a}(a_1+a_2+g_1+g_2)} - e^{-\psi_{2b}(a_1+a_2+g_1+g_2)}}{a_1 + a_2 + g_1 + g_2} \right) \end{aligned}$$

7.3.3 $t_1 < t_2 \doteq t_3$

$$\begin{aligned} & \int_{\psi_{3a}}^{\psi_{3b}} \int_{\psi_{2a}}^{t_3} \int_{\psi_{1a}}^{\psi_{1b}} g_0 f(\theta, \tau) \exp[-(t_1(a_1 + g_1)) - (t_2(a_2 + g_2)) - (t_3(a_3 + g_3))] dt_1 dt_2 dt_3 \\ &= g_0 f(\theta, \tau) \frac{e^{-\psi_{1a}(a_1+g_1)} - e^{-\psi_{1b}(a_1+g_1)}}{(a_1 + g_1)(a_2 + g_2)} \\ & * \left(\frac{e^{-\psi_{3a}(a_2+a_3+g_2+g_3)} - e^{-\psi_{3a}(a_2+g_2)-\psi_{3b}(a_3+g_3)}}{a_3 + g_3} - \frac{e^{-\psi_{3a}(a_2+a_3+g_2+g_3)} - e^{-\psi_{3b}(a_2+a_3+g_2+g_3)}}{a_2 + a_3 + g_2 + g_3} \right) \end{aligned}$$

7.3.4 $t_1 \doteq t_2 \doteq t_3$

$$\begin{aligned} & \int_{\psi_{3a}}^{\psi_{3b}} \int_{\psi_{2a}}^{t_3} \int_{\psi_{1a}}^{t_2} g_0 f(\theta, \tau) \exp[-(t_1(a_1 + g_1)) - (t_2(a_2 + g_2)) - (t_3(a_3 + g_3))] dt_1 dt_2 dt_3 \\ &= \frac{g_0 f(\theta, \tau)}{a_1 + g_1} \\ & * \left(\frac{e^{-\psi_{3a}(a_1+a_2+a_3+g_1+g_2+g_3)} - e^{-\psi_{3a}(a_1+a_2+g_1+g_2)-\psi_{3b}(a_3+g_3)}}{(a_2 + g_2)(a_3 + g_3)} \right. \\ & - \frac{e^{-\psi_{3a}(a_1+a_2+a_3+g_1+g_2+g_3)} - e^{-\psi_{3a}(a_1+g_1)-\psi_{3b}(a_2+a_3+g_2+g_3)}}{(a_2 + g_2)(a_2 + a_3 + g_2 + g_3)} \\ & - \frac{e^{-\psi_{3a}(a_1+a_2+a_3+g_1+g_2+g_3)} - e^{-\psi_{3a}(a_1+a_2+g_1+g_2)-\psi_{3b}(a_3+g_3)}}{(a_1 + a_2 + g_1 + g_2)(a_3 + g_3)} \\ & \left. + \frac{e^{-\psi_{3a}(a_1+a_2+a_3+g_1+g_2+g_3)} - e^{-\psi_{3b}(a_1+a_2+a_3+g_1+g_2+g_3)}}{(a_1 + a_2 + g_1 + g_2)(a_1 + a_2 + a_3 + g_1 + g_2 + g_3)} \right) \end{aligned}$$

References

- [1] W.A. Akanni, M. Wilkerson, C.J. Creevy, P.G. Foster, and D. Pisani. Implementing and testing Bayesian and maximum-likelihood supertree methods in phylogenetics. *Royal Society Open Science* 2, 2015. <https://doi.org/10.1098/rsos.140436>.
- [2] J. Chifman and L. Kubatko. Quartet Inference from SNP Data Under the Coalescent Model. *Bioinformatics*, 30(23):3317–3324, 2014.
- [3] J. Chifman and 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(1):35–47, 2015.

- [4] J.H. Degnan and L.A. Salter. Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37, 2005.
- [5] A.J. Drummond and A. Rambaut. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, 7(214), 2007. <https://doi.org/10.1186/1471-2148-7-214>.
- [6] E.Y. Durand, N. Patterson, D. Reich, and M. Slatkin. Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, 28(8):2239–2252, 2011.
- [7] J. Gatesy and M.S. Springer. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, 80:231–266, 2014.
- [8] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.
- [9] L.S. Kubatko, B.C. Carstens, and L.L. Knowles. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, 25(7):971–973, 2009.
- [10] L.S. Kubatko and J. Chifman. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evolutionary Biology*, 19:1–13, 2019.
- [11] L. Liu and D.K. Pearl. Species Trees from Gene Trees: Reconstructing Bayesian Posterior Distributions of a Species Phylogeny Using Estimated Gene Tree Distributions. *Systematic Biology*, 56(3):504–514, 2007.
- [12] L. Liu, L. Yu, and S.V. Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(302), 2010. doi:10.1186/1471-2148-10-302.
- [13] S. Mirarab, R. Reaz, M.S. Bayzid, T. Zimmermann, M.S. Swenson, and T. Warnow. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548, 2014.
- [14] S. Mirarab and T. Warnow. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics*, 31(12):i44–i52, 2015.
- [15] H.A. Ogilvie, R. R. Bouckaert, and A.J. Drummond. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. *Molecular Biology and Evolution*, 34(8):2101–2114, 2017.

- [16] M.N. Price, P.S. Dehal, and A.P. Arkin. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. *Molecular Biology and Evolution*, 26:1641–1650, 2009.
- [17] A. Rambaut and N.C. Grassly. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, 13(3):235–238, 1997.
- [18] B. Rannala and Z. Yang. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, 164(4):1645–1656, 2003.
- [19] A. Richards and L. Kubatko. Bayesian Weighted Triplet and Quartet Methods for Species Tree Inference. 2020. <https://arxiv.org/abs/2010.06063>.
- [20] S. Roch and M. Steel. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100c:56–62, 2015.
- [21] F. Ronquist, J.P. Huelsenbeck, and T. Britton. Bayesian Supertrees. In O.R.P. Bininda-Emonds, editor, *Phylogenetic Supertrees*, pages 193–224. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2004.
- [22] L. Salter. Complexity of the likelihood surface for a large DNA data set. *Systematic Biology*, 50(6):970–978, 2001.
- [23] E. Sayyari and S. Mirarab. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. *Molecular Biology and Evolution*, 33(7):1654–1668, 2016.
- [24] C. Semple and M. Steel. A supertree method for rooted trees. *Discrete Applied Mathematics*, 105:147–158, 2000.
- [25] G. Sevillya, Z. Frenkel, and S. Snir. TripletMaxCut: a new toolkit for rooted supertree. *Methods in Ecology and Evolution*, 7(11):1359–1365, 2016.
- [26] S. Snir and S. Rao. Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Molecular Phylogenetics and Evolution*, 62(22):1–8, 2012.
- [27] M.S. Springer and J. Gatesy. Phylogenetic analysis at deep timescales: Unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, 94:1–33, 2016.

- [28] A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 01 2014.
- [29] M. Steel and A. Rodrigo. Maximum Likelihood Supertrees. *Systematic Biology*, 57(2):243–250, 2008.
- [30] K. Strimmer, N. Goldman, and A. von Haeseler. Bayesian Probabilities and Quartet Puzzling. *Molecular Biology and Evolution*, 14(2):210, 1997.
- [31] K. Strimmer and A. von Haeseler. Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution*, 13(7):964–969, 1996.
- [32] Y. Tian and L. Kubatko. Rooting phylogenetic trees under the coalescent model using site pattern probabilities. *BMC Evolutionary Biology*, 17:1–11, 2017.
- [33] J. Wakeley. *Coalescent Theory: An Introduction*. Roberts Company Publishers, 2009.
- [34] T.J. Warnow. Supertree Construction: Opportunities and Challenges. 2018. arXiv:1805.03530.
- [35] Z. Yang. The BPP program for species tree estimation and species delimitation. *Current Zoology*, 61(5):854–865, 2015.
- [36] C. Zhang, M. Rabiee, E. Sayyari, and S. Mirarab. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(Supp 6):15–30, 2018.