CRP-Tree: A phylogenetic association test for binary traits

Julie Zhang[1], Gabriel A. Preising[2], Molly Schumer[2], and Julia A. Palacios[1,3]

[1]Department of Statistics, Stanford University
[2]Department of Biology, Stanford University
[3]Department of Biomedical Data Science, Stanford University

February 17, 2023

Abstract

An important problem in evolutionary genomics is to investigate whether a certain trait measured on each sample is associated with the sample phylogenetic tree. The phylogenetic tree represents the shared evolutionary history of the samples and it is usually estimated from molecular sequence data at a locus or from other type of genetic data. We propose a model for trait evolution inspired by the Chinese Restaurant Process that includes a parameter that controls the degree of preferential attachment, that is, the tendency of nodes in the tree to subtend from nodes of the same type. This model with no preferential attachment is equivalent to a structured coalescent model with simultaneous migration and coalescence events and serves as a null model. We derive a test for phylogenetic binary trait association with linear computational complexity and empirically demonstrate that it is more powerful than some other methods. We apply our test to study the phylogenetic association of some traits in swordtail fish, breast cancer, yellow fever virus and influenza A H1N1 virus. R package implementation of our methods is available at `https://github.com/jyzhang27/CRPTree`.

Keywords: Phylogenetic comparative methods, Phylogenetic mapping, Coalescent, Chinese Restaurant Process.

1   Introduction

Understanding the genetic basis of phenotypic traits is a fundamental goal in evolutionary biology and molecular epidemiology of infectious diseases. In particular, an important question is whether a certain observed trait is associated with the phylogenetic tree structure at a certain locus. Traits include geographic location, disease susceptibility, physical characteristics, and behavioral traits. To give a concrete example in infectious diseases, we can consider the phylogenetic tree (Figure 15(b)) reconstructed from yellow fever virus (YFV) molecular sequences obtained from infected humans (blue tips) and nonhuman primates (red tips) in South America. It is of interest to investigate whether the virus has been spreading only within each population, that is, whether human samples are more closely related to other human samples than to nonhuman primate samples. The answer to this question would provide insight about the recent outbreak of YFV in South America (Faria et al., 2018). This relation is known as **phylogenetic trait association** or **phylogenetic signal**. That is, the tendency of related organisms to share some trait characteristics more than organisms drawn at random from the same tree. Phylogenetic trait association is also sometimes assessed prior to doing a comparative analysis between traits or phenotypes. For example, one can use phylogenetically independent contrasts (Felsenstein, 1985; Garland Jr et al., 1992) to compare the association between two phenotypes given the phylogeny.

To illustrate the problem in phylogenetic trait-association, consider the following two trees in Figure 1. They have the same tree topology, but the tips have different trait values. In Figure 1(a), we see that all the nodes of the same type are in one subtree, and so clearly the trait is associated with the tree structure. In Figure 1(b), it is not so obvious whether there is a relation because the types are scattered throughout the tree topology. Later, we will illustrate that there is no phlyogenetic trait association in tree in Figure 1(b) according to our test.
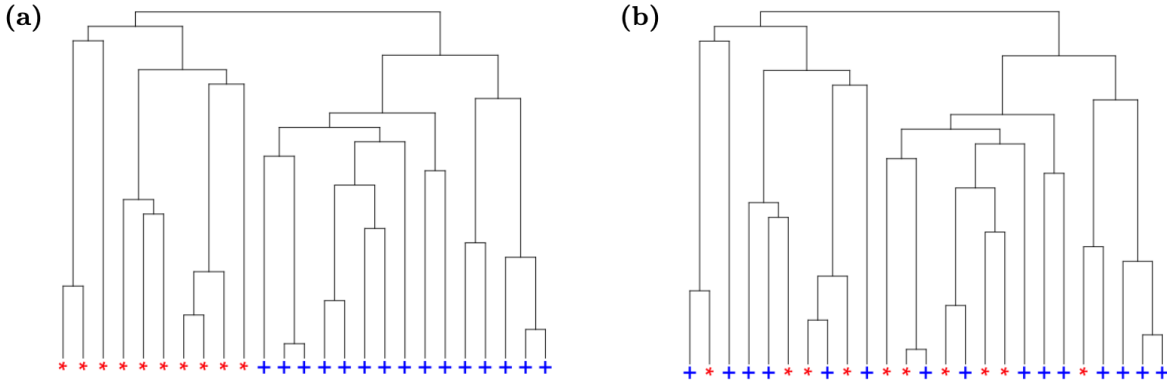

Figure 1: **Examples of phylogenetic trait associations.** Two trees with 25 tips, 10 of type 1 and 15 of type 2. **(a)** Clear association between the trait values (colors at the tips) and the tree topology. **(b)** We cannot tell visually if there is any association. Our proposed test does not detect association.

Many methods for discrete phylogenetic trait association have been developed in the fields of phylogenetics and molecular epidemiology. A large class of these methods consists in calculating a single summary statistic that conveys information about the trait-phylogeny association. Significance is then assessed through a permutation p-value obtained by permuting the leaf labels. The parsimony score (PS) is one of the most commonly used statistic. It counts the minimum number of trait state changes needed in the phylogeny in order to reconstruct character states at ancestral nodes (Fitch, 1971; Hartigan, 1973; Slatkin and Maddison, 1989). It is closely related to the Maximum Parsimony tree, which is the tree that minimizes the PS statistic. Wang et al. (2001) propose the association index (AI) statistic that measures the imbalance of the internal phylogeny nodes, and Borges et al. (2019) propose a measure based on Shannon entropy per node.

Phylogenetic diversity (Faith, 1992), nearest taxa index and nearest relatedness index (Webb, 2000; Webb et al., 2002), and UniFrac (Lozupone and Knight, 2005) are statistics that utilize branch length information and tree topology to capture trait-phylogeny association. However, a major drawback of the aforementioned class of methods is that they ignore phylogenetic uncertainty. Phylogenetic trees are usually estimated from molecular data with large uncertainty. In order to solve this problem, Parker et al. (2008) propose a method called BaTS (Bayesian Tip-association Significant testing), that incorporates phylogenetic uncertainty in a Bayesian Markov chain Monte Carlo framework. The authors use the posterior median of the association statistic to assess significance by comparing its value to the approximated null distribution of the median association statistic obtained by random permutation of the label set.

Another class of methods is based on change point detection along the phylogeny (Ansari and Didelot, 2016; Behr et al., 2020). In Behr et al. (2020), binary trait values at the tips of node i are assumed to be independent Bernoulli random variables with success probability p_i. The goal is then to detect the internal nodes at which the success probability changes via likelihood ratio statistics. The authors suggest their method can be extended to categorical and continuous traits. However, these methods also assume that the phylogeny is known without uncertainty. Phylogenetic trait-association methods for continuous traits usually model trait states via tree-valued Gaussian processes. Münkemüller et al. (2012) provides an extensive review of these methods and are not considered here.

In this article, we develop a phylogenetic trait association test for binary traits. Our method is applicable to fixed phylogenies and to latent phylogenies within a Bayesian Markov chain Monte Carlo framework. We propose a model for trait evolution inspired by the Chinese Restaurant Process (Aldous, 1985) that depends on a single parameter α. The model provides a tree-generating process in which the likelihood of lineages to descend from lineages of the same type is controlled by α. In this model, the number of same-type attachments is a sufficient statistic for α, therefore, our test statistic uses this information. Having a general model of trait evolution allows us to empirically test the power of the test under a large family of alternatives.

We structure our article as follows. First, in Section 2 we will introduce some terminology and definitions of the different tree topologies to be analyzed later, together with some novel and known enumerative results. We propose a coalescent model on partially labeled phylogenetic trees in Section 3. This model serves as the null model for our testing problem. In Section 4, we introduce the CRP-Tree model inspired by the Chinese Restaurant Process and state several results. We provide our test statistic and discuss how to assess significance in both a fixed tree case and a Bayesian framework in Section 5. In Sections 6 and 7, we analyze the performance of our test in simulated and real data applications. Finally, in Section 8 we summarize our contributions and discuss future directions.

2   Preliminaries

We first describe the four types of phylogenetics trees considered in this manuscript and provide some enumerative properties that will be used later. All trees are rooted and binary, and we only consider their tree topology ignoring branch length information.

1. A **ranked tree shape** T_N is a rooted, binary tree shape with N unlabeled tips and a total ordering of the internal nodes. The number of such trees is given by the Euler zig-zag numbers $e(N)$, defined via the recurrence relation (Murtagh, 1984).

$$e(N) = \frac{1}{2} \sum_{k=0}^{N-2} \binom{N-2}{k} e(k+1)e(n-k-1). \tag{1}$$

The base cases are $e(1) = e(2) = 1$. To see this, let T^L, T^R denote the left and right subtrees of T_N, and suppose T^L has k internal nodes and $k+1$ tips, while T^R has $n-k-2$ internal nodes and $n-k-1$

tips where $0 \leq k \leq n-2$. Then there are $\binom{n-2}{k}$ ways to arrange the internal nodes of both subtrees in order. In addition, there are a total of $e(k+1)$ possibilities for T^L and $e(n-k-1)$ possibilities for T^R. Accounting for the symmetry of T^L and T^R gives the final formula Equation (1). The first elements of the sequence are $1, 1, 1, 2, 5, 16, 61$.

2. A **ranked planar tree shape** $\tilde{T}_N$ is a ranked tree shape with N unlabeled tips where the left and right child nodes of an internal node are distinguished. The number of such trees is the number of permutations of $\{1, 2, ..., N-1\}$, that is $(N-1)!$ (Cleary et al., 2015). To see this, note that the order of appearance of internal node labels from left to right defines a ranked planar tree shape. For example, if $N = 4$, there are $(4-1)! = 6$ ranked planar tree shapes given by the 6 orderings of $\{1, 2, 3\}$. Figure 2 shows the 6 trees and their internal node orderings.
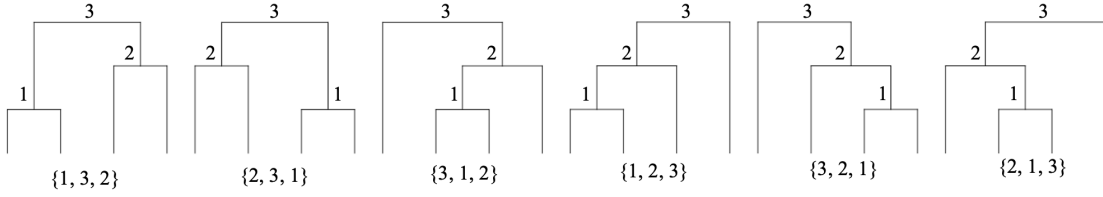

Figure 2: **Example of all 6 ranked planar tree shapes with 4 tips.** Each tree is identified by the order of its internal nodes from left to right.

3. A **ranked partially labeled tree** $T_{N,B}^{\ell}$ with N tips is a ranked tree shape whose tip labels form a multiset. In particular, we will focus on multisets with two unique elements, which we denote by {Blue, Red} for the rest of this manuscript. We can count the number of such trees $R(N, B)$ by a recursion similar to the number of ranked tree shapes.

$$R(N, B) = \frac{1}{2} \sum_{n=1}^{N-1} \sum_{b=0}^{\min(B,n)} R(n, b) R(N-n, B-b) \binom{N-2}{n-1}. \tag{2}$$

The base cases are $R(1, 1) = R(1, 0) = 1, R(2, 2) = R(2, 0) = R(2, 1) = 1$. The derivation is parallel to that of the Euler zig-zig numbers. Let the left subtree of $T_{N,B}^{\ell}$ have n tips. Then there are $\binom{N-2}{n-1}$ ways to arrange the internal nodes of both subtrees in order ($n-1$ in the left, and $N-n-1$ in the right). The term $R(n, b) R(N-n, B-b)$ counts the number of trees in which the left subtree of $T_{N,B}^{\ell}$ has b blue tips and n total tips, and the right subtree of $T_{N,B}^{\ell}$ has $B-b$ blue tips and $N-n$ total tips. Summing over all possible n and accounting for the symmetry of the left and right subtrees gives the final Equation (2). Note also that $R(N, 0) = R(N, N) = e(N)$. Starting from $N = 3$, the first few values of $R(N, \lfloor N/2 \rfloor)$ are $2, 7, 27, 152, 935$.

4. A **ranked planar partially labeled tree** $\tilde{T}_{N,B}^{\ell}$ is a ranked planar tree shape with N leaves, and B tips labeled blue, $N - B$ tips labeled red. There are $(N-1)! \binom{N}{B}$ such trees because there are $\binom{N}{B}$ possible labelings on every ranked planar tree.

We use the superscript ℓ in $T_{N,B}^{\ell}$ to indicate the partial labeling, and the tile in $\tilde{T}^{\ell}$ to indicate the tree is planar. In addition, for any tree T (regardless of resolution), we will use $C(T)$ to denote the number of cherries of T, that is, the number of subtrees with exactly two tips. We will use $C_S(T^{\ell})$ to denote the number of cherries of T^{ℓ} with the same label (regardless of resolution). Then for a given ranked tree shape T_N, there are $2^{N-1-C(T_N)}$ ranked planar trees $\tilde{T}_N$. This is because there are $N - 1 - C(T_N)$ nodes with distinct left and right subtrees that can be swapped to generate new planar trees. Similarly, for a given ranked partially labeled tree $T_{N,B}^{\ell}$, there are $2^{N-1-C_S(T_N)}$ ranked planar partially labeled trees $\tilde{T}_{N,B}^{\ell}$. Therefore, for a given ranked tree shape T_N, there are $2^{N-1-C(T_N)} \times \binom{N}{B}$ ranked, partially labeled, planar trees. Figure 3 displays
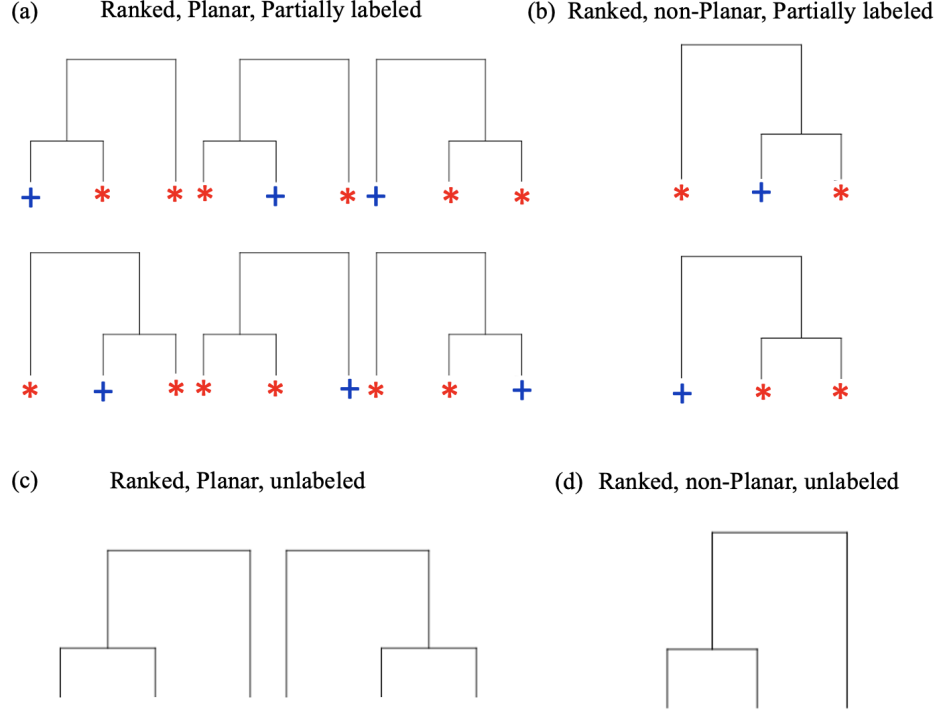
5

Figure 3: **Four different tree resolutions for trees with three tips**. The colors (and shapes for clarity) on the tips represent the tip labels, with $B = 1$. We see that there are 6 ranked planar partially labeled trees $\tilde{T}^{\ell}_{N,B}$, 2 ranked planar tree shapes $\tilde{T}_N$, 2 ranked partially labeled tree shapes $T^{\ell}_{N_B}$, and 1 ranked tree shape T_N.

the trees with $N = 3, B = 1$ in the four resolutions.

A ranked partially labeled tree with N tips and B blue, that has a monophyletic clade with respect to at least one color (blue or red) will be called a **perfect tree shape** and denoted by $T^{\ell,P}_{N,B}$. We say a tree is **exactly perfect** $T^{\ell,EP}_{N,B}$ if there is monophyly with respect to both colors. That is, all red tips and all blue tips are contained respectively in the two subtrees descending from the root. For example, Figure 1(a) is an exactly perfect tree shape, while a caterpillar tree with the tips of the cherry being blue and all other tips being red is a perfect tree shape. We extend these definitions to $\tilde{T}^{\ell,P}_{N,B}, \tilde{T}^{\ell,EP}_{N,B}$ if in addition, the tree is planar. Perfect trees are in a sense the most extreme partial labeling possible that separates the two types on a ranked tree shape.

Proposition 1. *The number of exactly perfect tree shapes with N tips and B blue is*

$$EP(N,B) = e(B) \times e(N - B) \times \binom{N-2}{B-1}.$$

The number of perfect tree shapes given N and B is defined recursively as

$$P(N,B) = \left(\sum_{i=0}^{N-B-1} e(N - B - i)P(B + i, B)\binom{N-2}{B+i-1} \right) +$$
$$\left(\sum_{i=0}^{B-1} e(B - i)P(N - B + i, N - B)\binom{N-2}{N-B+i-1} \right) - e(B) \times e(N - B) \times \binom{N-2}{B-1}.$$

6

The base cases are $P(n,n) = e(n), P(n+1,n) = e(n)$.

Proof. To prove the first statement note that there are $\binom{N-2}{B-1}$ ways to interleave the internal nodes of the two monophyletic subtrees together, each with $e(B)$ and $e(N-B)$ ranked tree shapes.

To find the number of perfect tree shapes, first suppose there is a blue monophyletic clade. Let the subtree containing the blue monophyletic clade have a total of $B+i$ tips, with $i = 0, ..., N-B-1$. There are a total of $P(B+i, B)$ such subtrees. The other subtree has $N-B-i$ red tips with $e(N-B-i)$ possible ranked tree shapes. Since there are $\binom{N-2}{B+i-1}$ ways to interleave both subtrees, we then get the first term in the summation. The second summand is obtained equivalently considering the red monphyletic clade. To get the final answer, we must subtract the number of exactly perfect trees to correct for double counting. $\square$

Remark: For any $T^\ell_{N,B}$, we would ideally like to know how "extreme" an observed partial labeling on a ranked tree shape T_N is with respect to the uniform distribution on label assignments. Though we can define a "most extreme" partial labeling, there is not a clear way to define an ordering among possible partial labelings, and therein lies the difficulty of phylogenetic trait association.

3 A null coalescent model

We describe a coalescent model on ranked partially labeled tree shapes with N tips, such that B leaves are blue and $R = N - B$ leaves are red. The model can be described as a bottom-up Markov chain in which every pair of lineages have equal probability of merging. In the tree, every internal lineage is labeled according to the order it is created. The state space can be described as (r_t, b_t, S_t) that records the number of red lineages r_t, the number of blue lineages b_t, and S_t denotes the set of internal lineages. The full realization $\{(r_t, b_t, S_t)\}_{t=0}^{N-1}$ uniquely encodes a ranked partially labeled tree shape. The initial state at the bottom of the tree is $(R, B, \emptyset)$ and the absorbing state at the root is $(0, 0, \{N-1\})$.

The initial state (at the tips) has no internal lineages, only blue and red nodes. It then transitions as follows

$$(R, B, \emptyset) \to \begin{cases} (R-2, B, \{1\}) & \text{w.p. } \frac{\binom{R}{2}}{\binom{N}{2}} \\ (R, B-2, \{1\}) & \text{w.p. } \frac{\binom{B}{2}}{\binom{N}{2}} \\ (R-1, B-1, \{1\}) & \text{w.p. } \frac{RB}{\binom{N}{2}} \end{cases} \tag{3}$$

After t steps, the state (r_t, b_t, S_t) indicates the tree has $r_t + b_t + |S_t| = N - t$ extant lineages, of which r_t lineages subtend red leaves, b_t lineages subtend blues leaves, and $|S_t|$ subtend internal nodes. Let $k = |S_t|$ be the number of current lineages subtending internal nodes and $s_i, s_j \le t$ denote internal nodes that are to be removed (because they will be merged). Then the $(t+1)$th transition for $t > 1$ has the following transition probabilities.

$$(r_t, b_t, S_t) \to \begin{cases} (r_t - 2, b_t, S_t \cup \{t+1\}) & \text{w.p. } \frac{\binom{r_t}{2}}{\binom{r_t+b_t+k}{2}} \\ (r_t, b_t - 2, S_t \cup \{t+1\}) & \text{w.p. } \frac{\binom{b_t}{2}}{\binom{r_t+b_t+k}{2}} \\ (r_t - 1, b_t - 1, S_t \cup \{t+1\}) & \text{w.p. } \frac{r_t b_t}{\binom{r_t+b_t+k}{2}} \\ (r_t, b_t - 1, S_t \cup \{t+1\}\backslash\{s_i\}) & \text{w.p. } \frac{b_t}{\binom{r_t+b_t+k}{2}} \\ (r_t - 1, b_t, S_t \cup \{t+1\}\backslash\{s_i\}) & \text{w.p. } \frac{r_t}{\binom{r_t+b_t+k}{2}} \\ (r_t, b_t, S_t \cup \{t+1\}\backslash\{s_i, s_j\}) & \text{w.p. } \frac{1}{\binom{r_t+b_t+k}{2}} \end{cases} \tag{4}$$

Another way to intuitively understand this jump chain is via an urn process. We start off with B blue balls and $N-B$ red balls in an urn. At the t-th iteration, we draw two balls without replacement and add a

ball with label t back into the urn. This numbered ball represents the internal node that was created, while the two balls we removed denote the lineages that were merged. The urn process ends when there is one ball left in the urn, namely ball $N-1$. Figure 4 pictorially demonstrates an example of a full realization of the jump chain starting with 3 blue balls and 2 red balls.
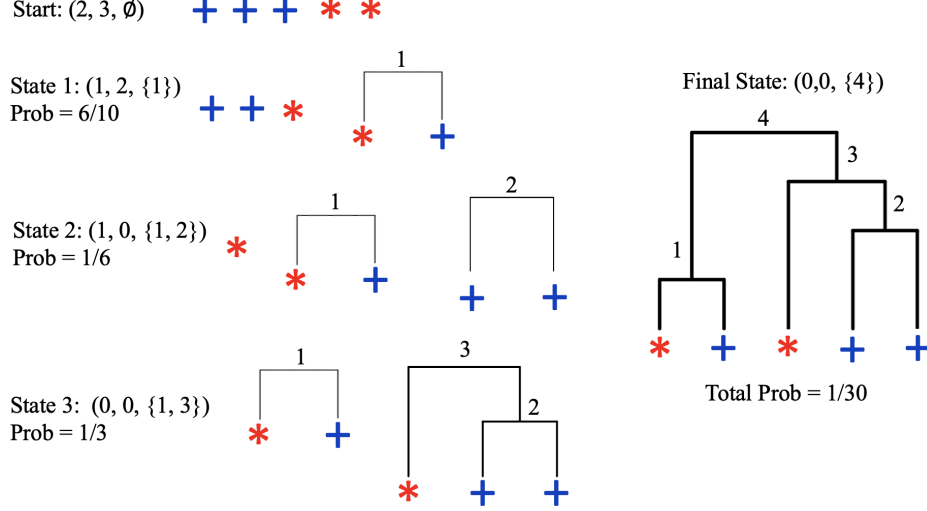

Figure 4: **One realization of the Markov jump chain starting with 2 red and 3 blue leaves**. At each step, we detail the states, the probability of the transition, and the corresponding step in the tree. The final probability of the tree matches the closed form expression from Theorem 2: $\frac{1}{30} = \frac{2^{5-1-1}}{24 \cdot 10}$.

Theorem 2. *The probability of observing $T_{N,B}^{\ell}$, a ranked partially labeled tree under the null coalescent model is*

$$\mathbb{P}(T_{N,B}^{\ell}) = \frac{2^{N-C_S(T_{N,B}^{\ell})-1}}{(N-1)!\binom{N}{B}},$$

where $C_S(T_{N,B}^{\ell})$ is the number of cherries of $T_{N,B}^{\ell}$ with the same label.

Proof. First, the denominator resulting from the product of all transitions probabilities in Equations (3) and (4) is

$$\binom{N}{2}\binom{N-1}{2}\cdots\binom{3}{2}\binom{2}{2} = \frac{N!(N-1)!}{2^{N-1}}.$$

The only transitions that invoke the $\frac{1}{2}$ factor are the coalescent events of two leaves with the same color, hence there is a term $\frac{1}{2^{C_S}}$ in the product of transition probabilities. If we merge two internal nodes, the numerator is multiplied by 1. When a red (blue) leaf is involved in a merger, the numerator of the transition probability is proportional to the number of red (blue) leaves. We then get the factor $B!(N-B)!$. Multiplying all these factors gives us the desired result. $\square$

As a special case, consider all samples being of the same type. In this case, the Markov chain corresponds to the Tajima coalescent (Sainudiin et al., 2015; Palacios et al., 2019), and the probability of a specific ranked tree shape is

$$\mathbb{P}(T_N) = \frac{2^{N-C(T_N)-1}}{(N-1)!},$$

where $C(T_N)$ is the number of cherries of T_N. Indeed, if we sum the probabilities of all the possible partial labelings on a specific ranked tree shape, we get the same probability expression.

Note that we can extend this null coalescent model to more than two categories of tip labels. Suppose there are m unique tip labels (i.e. m colors), with $n_1, ..., n_m$ number of labels of color $1, ..., m$ respectively, which we denote by $[n_i]_m$ for shorthand. Then the analogous extension of Theorem 2 is

$$\mathbb{P}(T^\ell_{N,[n_i]_m}) = \frac{2^{N-C_S(T^\ell_{N,[n_i]_m})-1}}{(N-1)!\binom{N}{n_1,...,n_m}}, \tag{5}$$

where $C_S(T^\ell_{N,[n_i]_m})$ is the number of cherries of $T^\ell_{N,[n_i]_m}$ with the same label and $\binom{N}{n_1,...,n_m}$ is the multinomial coefficient. In what follows, we will assume only binary labelings.

Proposition 3. *Fix a given ranked tree shape T_N, and $B \leq N$. The conditional probability of a specific partial labeling on T_N is*

$$\mathbb{P}(T^\ell_{N,B} \mid T_N) = \begin{cases} 0 & \text{if } T^\ell_{N,B} \not\prec T_N \\ \frac{2^{C(T^\ell_{N,B})-C_S(T^\ell_{N,B})}}{\binom{N}{B}} & \text{if } T^\ell_{N,B} \prec T_N, \end{cases}$$

where $T^\ell_{N,B} \prec T_N$ indicates that T_N is obtained from $T^\ell_{N,B}$ by removing the leaf labels in $T^\ell_{N,B}$.

Proof. We can prove this using our previous result and noting $C(T^\ell_{N,B}) = C(T_N)$.

$$\mathbb{P}(T^\ell_{N,B} \mid T_N) = \frac{P(T^\ell_{N,B} \cap T_N)}{P(T_N)} = \frac{P(T^\ell_{N,B})}{P(T_N)} \mathbb{1}\{T^\ell_{N,B} \prec T_N\}$$

$$= \frac{2^{C(T^\ell_{N,B})-C_S(T^\ell_{N,B})}}{\binom{N}{B}} \mathbb{1}\{T^\ell_{N,B} \prec T_N\}.$$

We directly see that the probability of observing a particular labeling given a ranked tree shape according to the coalescent null model (4) is not uniform. Yet, we can generate labeled trees with this probability by random permutation of the leaf labels. □

The model described is a lumping of the standard coalescent that models completely labeled ranked tree shapes (Kingman, 1982). However, this model is finer than the Tajima coalescent that models unlabeled ranked tree shapes (Sainudiin et al., 2015). The model is a modified structured coalescent model without explicit migration, but rather a simultaneous migration and coalescent event can happen in one transition (Notohara, 1990; Müller et al., 2017). Currently, we are ignoring branch lengths, so this is a discrete jump process, but one can easily incorporate exponential waiting times as per the standard continuous-time Markov Chain theory.

4 The CRP-Tree model

We wish to test whether an observed ranked partially labeled tree is a typical realization from the proposed null model and to evaluate the power of our test against competing hypotheses. For this reason, we now propose our alternative model based on the Chinese Restaurant Process (CRP).

The CRP is a discrete stochastic process used to generate a θ−biased random partition of $\{1, 2, ..., n\}$ (Aldous, 1985). It is used in many Bayesian nonparametric methods, with applications in topic modeling and population genetics (Griffiths et al., 2003; Qin, 2006). Imagine a restaurant with infinitely many tables and n customers who are lined up outside the door in order, with customer 1 first in line. The customers enter the restaurant one at a time. The kth customer chooses with probability $\frac{\theta}{k-1+\theta}$ to sit at a new table, and with probability $\frac{1}{k-1+\theta}$ to sit to the left of a particular person already seated. After all the customers have been seated, each non-empty table defines a cycle and the collection of all non-empty tables defines

a θ-biased random partition of $\{1, ..., n\}$. You would expect more small cycles when θ is large, and larger cycles when θ is small.

Suppose we are given N samples, B samples of one type (blue), and $N - B$ samples of the other type (red). The CRP-Tree model will generate a random ranked planar partially labeled tree. The parameter in our tree-generating model $\alpha \geq 1$ controls how likely are lineages to descend from a node of the same type. We will construct the tree forward in time, starting at the root.

1. Randomly order the B blues and $N - B$ reds into $C = (C_1, ..., C_N)$, where $C_i \in \{B, R\}$ is the color label of the ith node to be added.

2. Form the vector $(w_k = \sum_{i=1}^{k-1} \mathbb{1}(C_i = C_k) : k = 3, ..., N)$. Each w_k counts the number of nodes that precede node k that have the same color label as node k.

3. Form a binary tree with two tips, with the left, right tips labeled C_2, C_1 respectively.

4. For $k = 3, ..., N$: Let $(U_1, ..., U_{w_k})$ denote the leaves currently in the tree with same color as node k. Let $(V_1, ..., V_{k-1-w_k})$ denote the leaves with the opposite color as node k.

   (a) Generate a Bernoulli RV Z with success probability

   $$p = \frac{\alpha w_k}{(k - 1 - w_k) + \alpha w_k}$$

   (b) If $Z = 1$, uniformly select leaf U_i from $(U_1, ..., U_{w_k})$ to become the parent node of two leaves. Assign label C_k to the left leaf and the label of U_i to the right leaf.

   (c) If $Z = 0$, uniformly select leaf V_i from $(V_1, ..., V_{k-1-w_k})$ to become the parent node of two leaves. Assign label C_k to the left leaf and the label of V_i to the right leaf.

5. After all N tips are added, set the branch lengths so that the length between every consecutive internal node is 1 and all tips are equal distance to the root.

Ordering the tip labels $\{C_1, ..., C_N\}$ is equivalent to selecting the sequence of the node colors being added at each step. If $\alpha = 1$, then the probability of attaching to any color label is equal. If $\alpha > 1$, then the probability of attaching to a node of the same color label is larger. Notice the CRP-Tree model is Markovian because at each stage the transition probabilities only depend on the previous stage.

The analogy to the CRP is as follows. Suppose we have N customers in line and they each have blue or red business cards with the corresponding place in line, such that B customers have blue business cards. The first two customers 1 and 2 walk into the restaurant. If their business cards have the same color, they sit together at the same table, with customer 2 to the left of customer 1. Else, customers 1 and 2 sit at distinct tables. Next, customer k counts w_k customers who have the same color business card as them. With probability $\frac{\alpha}{k-1-w_k+\alpha w_k}$, customer k chooses to sit to the left of a person with the same color business card. With probability $\frac{1}{k-1-w_k+\alpha w_k}$, customer k select a person of the opposite color to ask for their business card. Then customer k moves to a new table, and places the business card to their right. Each customer will sit at only one table, but can have as many business cards at other tables. At the end, each non-empty table with a customer forms an ordered list, and the collection of non-empty tables defines our tree. The order of the tables is irrelevant. Each customer is a tip in our tree, and generating a new table represents a "mixing" event of the two colors because two tips of opposite labels are attached together. We give a small example in Figure 5 showing the tree and corresponding table representation. There will be a smaller number of tables for larger values of α, which implies more attachments of the same color.
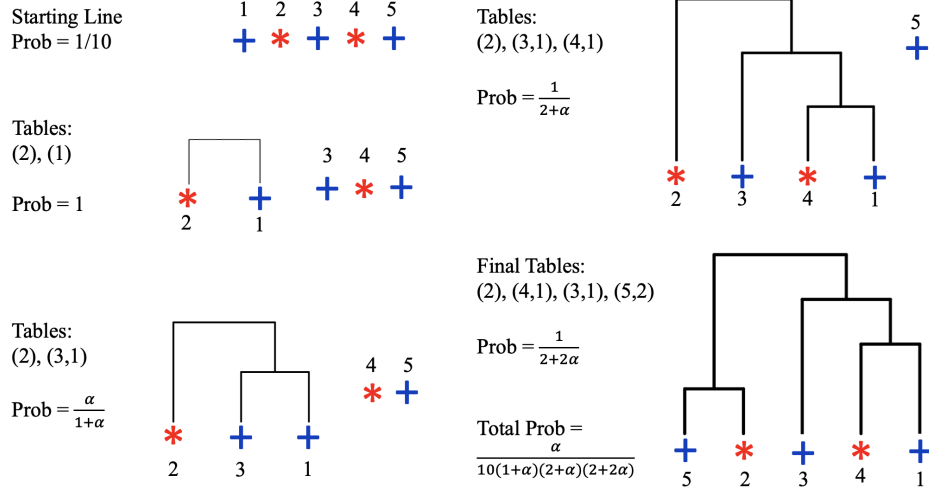
Figure 5: **An example of a tree generated using the CRP-Tree model with** $N = 5, B = 3$. Each time we add a node, we show the tree shape, corresponding table configuration, and probability of that attachment. We start with 2 tables because tips 1 and 2 are different colors, and end with 4. There were two same attachments in this example. The rankings at internal nodes are not shown for ease of visualization.

4.1 CRP-Tree as a 2-urn model

An intuitive way to understand this generative model is via a 2-urn model. Suppose we start with two urns: Urn 1 has B blue balls and $N - B$ red balls and Urn 2 is empty.

1. Select two balls without replacement in order from Urn 1 and place them into Urn 2. Also mark them as Ball 1 and Ball 2. This corresponds to creating a tree with two tips with tip labels (C_2, C_1), colors of Ball $2, 1$ respectively.

2. For $k = 3, ..., N$:

   (a) Select 1 ball from Urn 1 and mark it Ball k and note its color C_k.

   (b) In Urn 2, assign weight α to balls of the same color as Ball K and weight 1 to the rest of the balls.

   (c) Remove a ball from Urn 2 with probability proportional to its weight, call its number A_k and return it to Urn 2.

   (d) In the tree, make A_k the parent node of left leaf with label C_k (the color of Ball k) and right leaf with label the color of Ball A_k.

Notice that this implies W_k, the number of balls in the first $k - 1$ that are the same color as ball k, does not depend on α. Figure 6 shows an example of the 2-urn process for $N = 11, B = 5, \alpha = 2$, after 3 iterations (top panel) and after 4 iterations (bottom panel) when Ball $k = 5$ is attached to Ball 2.

4.2 Properties

Although the CRP-Tree is well-defined for $B = 0, 1, N - 1, N$, we will not investigate phylogenetic trait association in these cases. Going forward we will assume that $B \in \{2, 3, ..., N - 2\}$. Let X_k be the indicator of the event that node k attaches to a node of the same color label. Let S be the number of attachments of the same color, that is, $S = \sum_{k=3}^{N} X_k$. Then the likelihood of the ranked planar partially labeled tree $\tilde{T}^\ell_{N,B}$
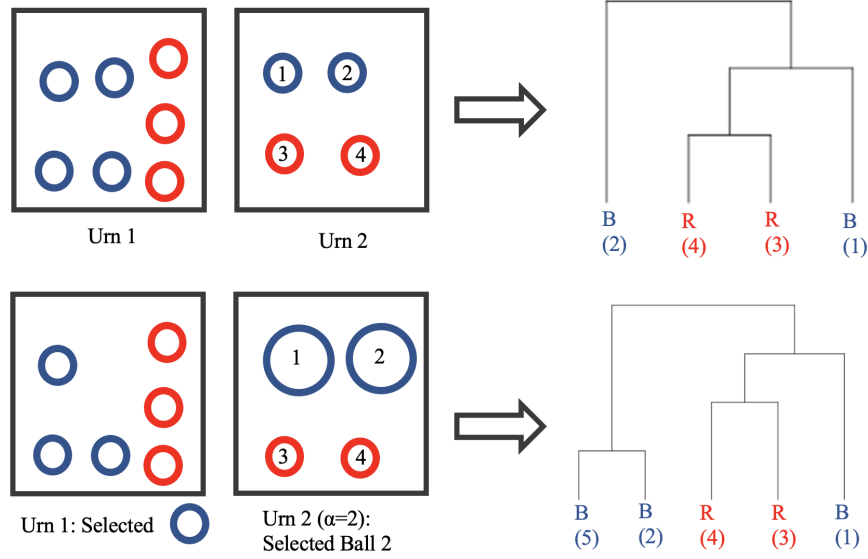
11

Figure 6: **An example of two possible states in the urn process corresponding to the CRP-Tree model.** In the top panel, a tree with 4 leaves has already been constructed, with 7 balls in Urn 1 and 4 labeled balls in Urn 2. In the bottom panel, we select one blue ball, increase the size of the blue balls in Urn 2, and ultimately select Ball 2 for Ball 5 to attach to. In the final tree, only the tip colors will be preserved, because the tip numbers can be recovered from the ranking of the internal nodes and the planarity. The rankings at internal nodes are not shown for ease of visualization.

under the CRP-Tree model is

$$L(\tilde{T}^\ell_{N,B}) = \frac{1}{\binom{N}{B}} \cdot \prod_{k=3}^{N} \frac{\alpha^{X_k}}{k - 1 - w_k + \alpha w_k}$$

$$= \frac{1}{\binom{N}{B}} \cdot \alpha^S \cdot \left(\prod_{k=3}^{N} (k - 1 - w_k + \alpha w_k) \right)^{-1}.$$

The log-likelihood is

$$\ell(\tilde{T}^\ell_{N,B}) = -\log\left(\binom{N}{B}\right) + S\log(\alpha) - \sum_{k=3}^{n} \log(k - 1 - w_k + \alpha w_k).$$

We see that $S, \{W_k : k = 3, ..., N\}$ are the sufficient statistics for α by Fisher–Neyman factorization theorem. Using the CRP-table representation, the number of same tyoe attachments S is given by

$$S = N - 2 - \# \text{ of new tables,}$$

where we start out with either one table $(2, 1)$ in the case C_1 and C_2 have the same color, or two tables $(2), (1)$ otherwise.

Proposition 4. *If $\alpha = 1$, then the probability of any ranked planar partially labeled tree under the CRP-Tree model is*

$$\mathbb{P}(\tilde{T}^\ell_{N,B}) = \frac{1}{(N-1)!} \cdot \frac{1}{\binom{N}{B}}.$$

12

Proof. We can directly see this result from the likelihood when $\alpha = 1$. Alternatively, the probability of any fixed initial ordering is $\frac{1}{\binom{N}{B}}$. Given the initial order, then at step k, we uniformly pick a branch to attach to with probability $\frac{1}{k-1}$. After all steps, we get $\frac{1}{(N-1)!}$. Notice this implies the CRP-Tree process uniformly generates ranked planar partially labeled tree shapes under $\alpha = 1$ (see Section 2). $\square$

Proposition 5. *If $\alpha = 1$, then the probability of a ranked (non-planar) partially labeled tree under the CRP-Tree model is*

$$\mathbb{P}(T^\ell_{N,B}) = \frac{2^{N - C_S(T^\ell_{N,B}) - 1}}{\binom{N}{B}(N-1)!}.$$

Proof. There are a total of $2^{N - C_S(T^\ell_{N,B}) - 1}$ ways to interchanging the left and right subtrees of an internal node without changing the ranked partially labeled tree shape. Combining with Proposition 4 gives the result. $\square$

An important fact implied by Proposition 5 is that the probability of observing $T^\ell_{N,B}$ under the CRP-Tree model with $\alpha = 1$, is equal to the probability of observing the same tree under the null coalescent model of Section 3. As α increases, the model will generate trees with more same-type attachments. In fact, we will show that as α goes to infinity, the probability of observing a perfect tree (trees with one color completely contained in a monophyletic clade) goes to one. This property will be made formal in the following results.

Lemma 6. $\tilde{T}^{\ell,EP}_{N,B}$ *is an exact perfect planar tree if and only if $S = N - 2$. $\tilde{T}^{\ell,P}_{N,B}$ is perfect but not exactly perfect if and only if $S = N - 3$. In addition, under the CRP-Tree model,*

$$\lim_{\alpha \to \infty} \mathbb{P}(\tilde{T}^{\ell,EP}_{N,B}) = \frac{1}{\binom{N}{B}} \times \frac{1}{(B-1)!(N-B-1)!},$$

$$\lim_{\alpha \to \infty} \mathbb{P}(\tilde{T}^{\ell,P}_{N,B}) = \frac{1}{(r-1)\binom{N}{B}} \times \frac{1}{(B-1)!(N-B-1)!},$$

where $w_r = 0$ for a unique $r \in \{3, ..., N\}$.

Proof. In order to form an exactly perfect tree, each attachment must be an attachment of the same color, which means $S = N - 2$ and we must start with either B, R or R, B in the order of attachments. Hence $w_k \neq 0$ for all $k = 3, ..., N$, and

$$\mathbb{P}(\tilde{T}^{\ell,EP}_{N,B}) = \frac{1}{\binom{N}{B}} \cdot \prod_{k=3}^{N} \frac{\alpha}{k - 1 - w_k + \alpha w_k} = \frac{1}{\binom{N}{B}} \cdot \prod_{k=3}^{N} \frac{1}{w_k + (k - 1 - w_k)/\alpha}.$$

Taking the limit gives $\lim_{\alpha \to \infty} \mathbb{P}(\tilde{T}^{\ell,EP}_{N,B}) = \frac{1}{\binom{N}{B}} \cdot \prod_{k=3}^{N} \frac{1}{w_k}$. Moreover, $\{w_k : k = 3, ..., N\}$ is an interleaving of $w^B = \{1, 2, ..., B-1\}$ and $w^R = \{1, 2, ..., N - B - 1\}$, so we have $\prod_{k=3}^{N} w_k = (B-1)!(N-B-1)!$ and

$$\lim_{\alpha \to \infty} \mathbb{P}(\tilde{T}^{\ell,EP}_{N,B}) = \frac{1}{\binom{N}{B}} \cdot \frac{1}{(B-1)!(N-B-1)!}.$$

For perfect but not exactly perfect trees, we must have $S = N - 3$ because the root of the monophyletic clade that contains all tips of one color, is attached to an opposite color. This can only happen if the initial order is B, B or R, R. Without loss of generality, let us suppose the ordering starts with B, B and that the first R appears at element r ($C_r = R$), which would imply $w_r = 0$, and this index is unique. In addition, for

all $k < r$, we have $w_k = k - 1$. Hence, for a perfect $\tilde{T}^\ell_{N,B}$ and initial ordering starting with B, B, we have

$$\lim_{\alpha \to \infty} \mathbb{P}(\tilde{T}^{\ell,P}_{N,B}) = \lim_{\alpha \to \infty} \frac{1}{\binom{N}{B}} \cdot \frac{\alpha^S}{\prod_{k=3}^N (k - 1 - w_k + \alpha w_k)}$$

$$= \lim_{\alpha \to \infty} \frac{1}{\binom{N}{B}} \cdot \prod_{k=3}^{r-1} \frac{\alpha}{\alpha w_k} \cdot \left(\frac{1}{r - 1 - w_r + \alpha w_r} \right) \cdot \prod_{k=r+1}^N \frac{\alpha}{(k - 1 - w_k + w_k \alpha)}$$

$$= \lim_{\alpha \to \infty} \frac{1}{\binom{N}{B}} \cdot \prod_{k=3}^{r-1} \frac{1}{w_k} \cdot \left(\frac{1}{r - 1} \right) \cdot \prod_{k=r+1}^N \frac{\alpha}{(k - 1 - w_k + w_k \alpha)}$$

$$= \frac{1}{(r - 1)\binom{N}{B}} \prod_{k=3, k \neq r}^N w_k.$$

Now, $\prod_{k=3, k \neq r}^N w_k = (B - 1)!(N - B - 1)!$ because $\{w_k : k = 3, ..., r - 1, r + 1, ...N\}$ is an interleaving of $w^B = \{2, ..., B - 1\}, w^R = \{1, 2, ..., N - B - 1\}$, where $w_r = 0$ is not counted in the product. Hence, $\lim_{\alpha \to \infty} \mathbb{P}(\tilde{T}^{\ell,P}_{N,B}) = \frac{1}{(r-1)\binom{N}{B}} \cdot \frac{1}{(B-1)!(N-B-1)!}$. $\qquad \square$

Theorem 7. *Under the CRP-Tree model with $\alpha \geq 1$, we have*

$$\lim_{\alpha \to \infty} \mathbb{P}(\{T^\ell_{N,B} : T^\ell_{N,B} \text{ is perfect}\}) = 1,$$

$$\lim_{\alpha \to \infty} \mathbb{P}(\{T^\ell_{N,B} : T^\ell_{N,B} \text{ is exactly perfect}\}) = \frac{2B(N - B)}{N(N - 1)},$$

$$\lim_{\alpha \to \infty} \mathbb{P}(\{T^\ell_{N,B} : T^\ell_{N,B} \text{ is perfect but not exactly perfect}\}) = 1 - \frac{2B(N - B)}{N(N - 1)}.$$

Proof. We will first consider a ranked planar partially labeled tree $\tilde{T}^\ell_{N,B}$ with $S < N - 3$. By the previous lemma, we know that $\tilde{T}^\ell_{N,B}$ is not a perfect tree. We will show that the probability of observing such a tree goes to 0 as $\alpha \to \infty$, and therefore, the probability of observing a planar perfect tree goes to 1 as $\alpha \to \infty$. Now, if the initial color ordering starts with B, B or R, R, let $r \in \{3, ..., N\}$ be such that $w_r = 0$. If the initial color ordering starts with B, R or R, B, then let $r = 2$. We then have

$$\lim_{\alpha \to \infty} \mathbb{P}(\tilde{T}^\ell_{N,B}) = \lim_{\alpha \to \infty} \frac{1}{\binom{N}{B}} \cdot \frac{\alpha^S}{\prod_{k=3}^N (k - 1 - w_k + \alpha w_k)}$$

$$= \lim_{\alpha \to \infty} \frac{1}{(r - 1)\binom{N}{B}} \cdot \frac{\alpha^S}{\prod_{k=3, k \neq r}^N (k - 1 - w_k + \alpha w_k)} = 0.$$

because the denominator of the second term has leading term α^{N-3} while the numerator has leading term α^S.

Any exactly perfect planar tree must be generated with initial color ordering B, R or R, B. Given the initial ordering, the number of exactly perfect planar trees that can be formed is $(B - 1)!(N - B - 1)!$ because at the kth step, node k has a choice of w_k nodes to attach to. Moreover, there are $2\binom{N-2}{B-1}$ initial orderings that start with B, R or R, B. Therefore by Lemma 6,

$$\lim_{\alpha \to \infty} \mathbb{P}(\{\tilde{T}^\ell_{N,B} : \tilde{T}^\ell_{N,B} \text{ is exactly perfect}\}) = \frac{2\binom{N-2}{B-1}}{\binom{N}{B}} = \frac{2B(N - B)}{N(N - 1)}, \text{ and}$$

$$\lim_{\alpha \to \infty} \mathbb{P}(\{\tilde{T}^\ell_{N,B} : \tilde{T}^\ell_{N,B} \text{ is perfect but not exactly perfect}\}) = 1 - \frac{2B(N - B)}{N(N - 1)}.$$

Finally, note $\mathbb{P}(\{\tilde{T}^\ell_{N,B} : \tilde{T}^\ell_{N,B} \text{ is perfect}\}) = \mathbb{P}(\{T^\ell_{N,B} : T^\ell_{N,B} \text{ is perfect}\})$ and the same holds for exactly perfect trees. Therefore the three results hold. $\qquad \square$

14

Theorem 8 (Expected Value of S).
If $\alpha = 1$, then

$$\mathbb{E}[S] = \frac{(N-2)\big(B(B-1) + (N-B)(N-B-1)\big)}{N(N-1)} = (N-2) - \frac{2B(N-B)(N-2)}{N(N-1)}.$$

If $\alpha > 1$, then

$$\mathbb{E}[S] = \frac{B}{N} \times \sum_{i=0}^{k-1} \frac{\alpha i}{(k-1-i) + \alpha i} \frac{\binom{k-1}{i}\binom{N-k}{B-(i+1)}}{\binom{N-1}{B-1}} + \frac{N-B}{N} \times \sum_{i=0}^{k-1} \frac{\alpha i}{(k-1-i) + \alpha i} \frac{\binom{k-1}{i}\binom{N-k}{N-B-(i+1)}}{\binom{N-1}{N-B-1}},$$

with the convention $\binom{a}{b} = 0$ if $a < b$.

Proof. Intuitively, S is the sum of linear combinations of Hypergeometric random variables. See Appendix Section 9.1 for the full proof. $\qquad\square$

4.3 Three equivalent representations

We now list three ways in which we can represent the information of a ranked planar partially labeled tree shape $\tilde{T}^\ell_{N,B}$: the tree form, a sequence of attachments together with an initial color order C, and in terms a collection of tables via the CRP. Figure 7 shows the three representations for $\tilde{T}^\ell_{15,6}$. Each representation has its benefits: the tree shape is what is usually given, the sequence of attachments and initial color order allow us to calculate the sufficient statistics $S, \{w_k : k = 3, ..., N\}$, while the collection of tables is a representation free of any color information. We will show that these representations are all bijective, and describe algorithms to reconstruct each representation from the other. Algorithm 1 allows us to recover the sequence of attachments and initial color order from $\tilde{T}^\ell_{N,B}$. Finally, we will define the set of conditions needed for a collection of tables to encode a ranked planar partially labeled tree, as well as a constructive proof to find the initial color order and the sequence of attachments from the collection of tables.
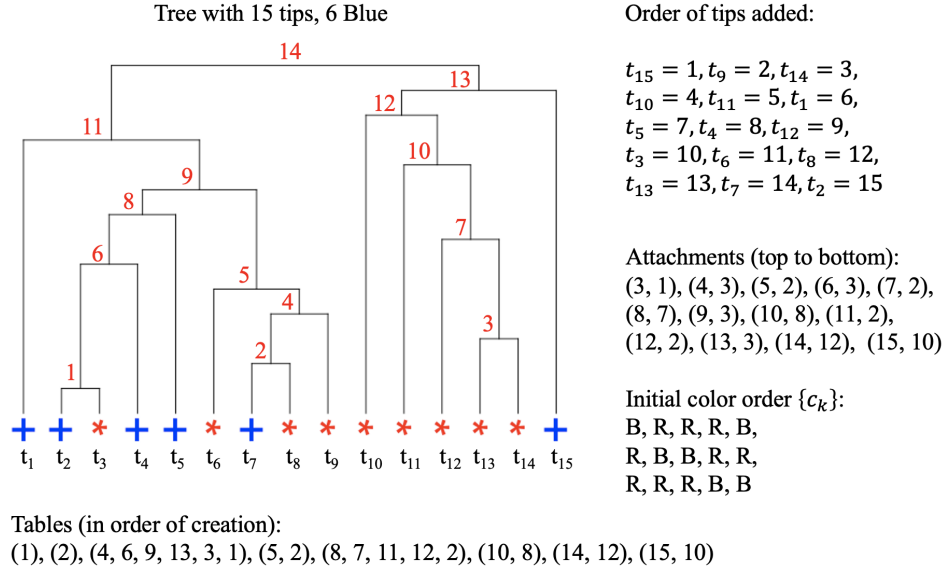

Tree with 15 tips, 6 Blue

Order of tips added:

$t_{15} = 1, t_9 = 2, t_{14} = 3,$
$t_{10} = 4, t_{11} = 5, t_1 = 6,$
$t_5 = 7, t_4 = 8, t_{12} = 9,$
$t_3 = 10, t_6 = 11, t_8 = 12,$
$t_{13} = 13, t_7 = 14, t_2 = 15$

Attachments (top to bottom):
(3, 1), (4, 3), (5, 2), (6, 3), (7, 2),
(8, 7), (9, 3), (10, 8), (11, 2),
(12, 2), (13, 3), (14, 12), (15, 10)

Initial color order $\{c_k\}$:
B, R, R, R, B,
R, B, B, R, R,
R, R, R, B, B

Tables (in order of creation):
(1), (2), (4, 6, 9, 13, 3, 1), (5, 2), (8, 7, 11, 12, 2), (10, 8), (14, 12), (15, 10)

Figure 7: **The three representations of $\tilde{T}^\ell_{15,6}$.** First, we label the tips $t_1, .., t_{15}$ from left to right and internal nodes from 1 to 14 from bottom to top (in red). We can determine the order in which the tips were added to the tree and the sequence of attachments from the tree. Similarly, we can derive this same information from the set of tables and vice versa.

15

4.3.1 Planar ranked tree to sequence of attachments, initial color order, and order of tips added

Given $\tilde{T}^\ell_{N,B}$, Algorithm 1 in Appendix Section 9.2 can uniquely determine $C = (C_1, ..., C_N)$, the initial ordering of the colors of leaf labels that generated the tree, and the sequence of attachments $\{(k, z_k) : k = 3, ..., N\}$. That is, a ranked planar partially labeled tree has a bijective correspondence with a color ordering $(C_1, ..., C_N)$ and a sequence of attachments.

The key for being able to recover C and the sequence of attachments from $\tilde{T}^\ell_{N,B}$, is our careful planar construction of the tree: new tips are always attached to the left of an existing tip. We first label the internal nodes ranked from bottom to top by 1 to $N - 1$, the numbers in red in Figure 7. This allows us to uniquely backtrack (bottom to top) which node was being added and which node was begin attached to. We will use the general notation of $t_1, ..., t_N$ to label the tips from left to right, where we will find $t_k \in \{1, ..., N\}$, the order in which tips were added.

We start with the youngest internal node (internal node 1) and look at its two immediate offspring, the left tip is the last tip added, with color label (C_N) and tip number label N, and the right tip is the tip being added to. In Figure 7, this corresponds to tip t_2 being added and labeled $t_2 = 15$, t_3 being attached to, and $C_N = B$. Proceeding in this manner, the right-most tip in the left subtree of internal node k will be the tip added, while the right-most tip in the right subtree of the internal node k will be the tip added to. By nature of the planarity, the tip added would not have been added previously. Returning back to our example, at internal node 2, the tip being added is t_7, so it is labeled $t_7 = 14$ and $C_{14} = B$, and t_8 being attached to. For the final step at the root (internal node $N - 1$), the right-most tip in the left subtree of the root will be the second tip added, while the right-most tip in the right subtree of the root will the the first tip added. These will be the two remaining tips that have not been added. In Figure 7, at the root, the right-most tip in the left subtree is t_9 and the right-most tip in the right subtree is t_{15}, so $t_9 = 2, t_{15} = 1$. One can check that indeed tip t_9 had not been added yet. The attachments made are given by the pairs $\{(A_i, D_i) = (N + 1 - i, t_{r_i}) : i = 1, ..., N - 1\}$, with A_i denoting the tip being added, and D_i, the tip being attached to, at step i. For example, the last attachment created was $(t_2, t_3) = (15, 10)$.

4.3.2 Ranked planar tree to collection of tables and vice versa

Every ranked planar partially labeled tree $\tilde{T}^\ell_{N,B}$ has a one-to-one correspondence to a collection of tables $\{E^1, E^2, ..., E^T\}$. These tables obey a set of conditions stated in Appendix Section 9.3. An implication of these assumptions is that color information is not needed in the table representation in order to reconstruct the tree.

To obtain the sequence of attachments from the collection of tables, first note that the number of attachments represented in a table is the number of elements minus one. For example, the attachments in table $E^t = (E^t_1, ..., E^t_n)$, can be denoted by $(E^t_1, E^t_{a_1}), ...(E^t_{n-1}, E^t_{a_{n-1}})$, where a_j is the smallest index satisfying $a_j > j$ and $E^t_j > E^t_{a_j}$. By construction, there will be a total of $N - 2$ attachments made across all the tables. We can rearrange all attachments to be in order: $(3, A_3), (4, A_4), ..., (N, A_N)$, including possibly $(2, 1)$, with $A_j < j$ for all $j = 3, ..., N$. For example, the collection of tables $\{E_1, E_2, E_3\} = \{(7, 6, 3, 1), (5, 1), (4, 2)\}$ corresponds to attachments $(7, 6), (6, 3), (3, 1), (5, 1), (4, 2)$.

Next, we determine the colors by explicitly stating which attachments must be of the same color. First, if $(2, 1)$ appears, then tips $1, 2$ must be of the same color, otherwise one is blue and the other is red. If $(2, 1)$ does not exist, then the pairs of the form (j, A_j) with $A_j = 1$ or $A_j = 2$, and j the smallest element that is attached to A_j, will be of opposite colors if (A_j) exists as a table in $\{E_1, ..., E_T\}$. Otherwise, j and A_j will be of opposite colors. Going back to our example, without loss of generality, we take tip 1 to be Blue and tip 2 to be Red. The smallest tip that is attached to 1 is 3 in pair $(3, 1)$, and (1) does not exist as a table, so tip 3 is also Blue. The smallest element that is attached to 2 is $(4, 2)$ and (2) does not exist as a table,

so tip 4 must be Red.

To determine the remaining attachment types, let E^{t_j} be the table that contains attachment (j, A_j). If $|E^{t_j}| = 2$, then j and A_j are of opposite colors. If $|E^{t_j}| \geq 3$, and there is an element smaller than j in the same table, it implies a table was already created when j was added, therefore j and A_j are of the same color. Otherwise, the table is newly created and j and A_j are of different colors. To finish off the example, the table $(5, 1)$ has two elements, so tip 5 must be Red. Finally, the attachments $(7, 6)$ and $(6, 3)$ are attachments made on pre-existing tables, so tips $3, 6, 7$ are all the same color. Therefore, our final set of attachments and tip colors is $(7B, 6B), (6B, 3B), (3B, 1B), (5R, 1B), (4R, 2R)$.

4.4 Resulting Tree Topology

In Section 4.2, we showed that the CRP-Tree model with $\alpha = 1$ generates ranked partially labeled trees with the same probability law as the null coalescent model of Section 3. We further showed that if we remove the color labels, we obtain unlabeled ranked tree shapes with the same law as in the Tajima coalescent. We empirically verify that α does not greatly affect the probability law of the ranked tree shapes by comparing the averages of various tree statistics to the expectation under the Tajima coalescent. Two popular statistics are the number of cherries (2-tip subtrees), and the number of pitchforks (3-tip subtrees), with expected values $N/3$ and $N/6$ respectively under the standard coalescent (McKenzie and Steel, 2000; Choi et al., 2020). Figure 17 in the Appendix shows that for each α, the number of cherries and pitchforks is concentrated around the expected value.

Another simulation to examine the distribution of ranked tree shape topologies is motivated by Kim et al. (2020). The authors proposed a distance on ranked tree shapes and use it to visualize tree distribution in 2 dimensions via multidimensional scaling (MDS). Figure 18 in the Appendix does not exhibit clustering of the trees per distribution, suggesting similarity among the three ranked tree shape distributions. More details on both these simulations can be found in Appendix Section 9.4.

4.5 Discussion about planarity

The output of our model is a ranked planar partially labeled tree, which is a largely unexplored tree resolution. In this case, planarity and the tip colors allows us to label the internal nodes blue or red and to know the sequence of node attachments. In many biological situations, there is a differentiation between the two children of a node, and therefore a way to distinguish them is by keeping track of left and right nodes. For example, these include speciation, transmission trees in epidemiology, and cell lineage diagrams,, where in each case, the left and right subtrees represent a biologically important distinction (Stewart et al., 2005; Hagen et al., 2015; Sainudiin and Welch, 2016).

Other methodologies are also built under the assumption of planarity. Behr et al. (2020) assume a fixed planar representation, and they show via simulation that their method is mostly robust to changes in this representation (i.e. their statistic does not change much with a change in the planarity). Ford et al. (2009) also work on the resolution of ranked planar trees. Their test statistic is calculated on the planar tree, and then planarity is marginalized out when calculating the final p-value. Sainudiin and Véber (2016) derive a Beta-splitting model at a variety of tree resolutions: ranked and planar, unranked and planar, ranked and non-planar, unranked and non-planar.

5 CRP-Tree phylogenetic association test

We will first consider the setting in which a ranked partially labeled phylogenetic tree is available, for example obtained via Maximum Likelihood estimation from molecular sequence data. We ignore any phylogenetic uncertainty in this case, which may not be ideal in many situations. In addition, we will assume

that a binary trait is completely observed at the tips of the phylogeny, that is, we observe $T_{N,B}^{\ell,0}$. We use the superscript 0 to denote that the tree is observed and fixed. Given $T_{N,B}^{\ell,0}$, we wish to test whether there is a phylogenetic association in the binary trait. In terms of the CRP-Tree model, the hypothesis of no phylogenetic association is equivalent to $H_0 : \alpha = 1$.

A natural test statistic is S, the number of same type attachments. However, this statistic is not directly observed since it depends on the initial color ordering and the sequence of type of attachments. Instead, our proposed test statistic is $\mu = \mathbb{E}[S|T_{N,B}^{\ell}]$. In practice for large N, μ is replaced by

$$\hat{\mu} = \frac{1}{M} \sum_{i=1}^{M} S_i,$$

where S_i is the number of same-type attachments in $\tilde{T}_{N,B}^{\ell,i} \sim \mathbb{P}_{\alpha=1}(\tilde{T}_{N,B}^{\ell,i} \mid T_{N,B}^{\ell})$.

5.1 Testing by permutation

To assess significance, we estimate the null distribution of our test statistic $\hat{\mu}$, conditional on the observed ranked tree shape $T_N : T_{N,B}^{\ell,0} \prec T_N$, and (N, B). We estimate the null distribution by random permutation of the leaf labels, that is, we generate $\{T_{N,B}^{\ell,i}\}_{i=1}^{K}$ by randomly relabeling the tips of T_N with B blues and $N - B$ reds, K times. To calculate $\hat{\mu}_0, \hat{\mu}_1, \ldots, \hat{\mu}_K$, we sample M ranked planar partially labeled tree shapes uniformly conditional on each $T_{N,B}^{\ell,i}$, for $i = 0, \ldots, K$ by picking an internal node uniformly among those that do not subtend a same type cherry, and then permuting its left and right subtrees. The p-value is then

$$p_S = \frac{1 + \sum_{i=1}^{K} \mathbb{1}(\hat{\mu}_i \geq \hat{\mu}_0)}{1 + K}.$$

When K and M are larger than the possible number of permutations, we simply generate all permutations to compute the exact p-value. Figure 8(a, c) show a schematic of the simulations needed to calculate p_S.

We propose a second test in which our test statistic is a sample of $\{S_i^{obs}\}_{i=1}^{M}$ from $\mathbb{P}_{\alpha=1}(S \mid T_{N,B}^{\ell,0})$ generated by sampling planar representations uniformly for the observed $T_{N,B}^{\ell,0}$, and computing the number of same type attachments. The cardinality of the space of all ranked planar partially labeled trees $\tilde{T}_{N,B}^{\ell}$ that are compatible with T_N is $2^{N-1-C(T_{N,B}^{\ell})} \times \binom{N}{B}$. Therefore, our null distribution is generated by randomly sampling K partial labelings on T_N together with a random planar representation. Let $\{S_j\}_{j=1}^{K}$ denote the empirical null distribution. Then, our p-value is defined as

$$p_T = \frac{1}{M} \sum_{i=1}^{M} \left(\frac{1 + \sum_{j=1}^{K} \mathbb{1}(S_j \geq S_i^{obs})}{1 + K} \right)$$

Figure 8(a, b) show a schematic of the simulations needed to calculate p_T.

Lemma 9. *Let* $T_{N,B}^{\ell,0}, \ldots, T_{N,B}^{\ell,K} \overset{iid}{\sim} \mathbb{P}(T_{N,B}^{\ell}|T_N, B)$, *where* $\mathbb{P}(T_{N,B}^{\ell}|T_N)$ *is the distribution of ranked and partially labeled tree shapes* $T_{N,B}^{\ell} \prec T_N$ *derived in Proposition 3. Let* $\mu_i = \mu(T_{N,B}^{\ell,i}) = \mathbb{E}[S \mid T_{N,B}^{\ell,i}]$ *for* $i = 0, \ldots, K$, *and let*

$$P_S = \frac{\sum_{i=0}^{K} \mathbb{1}(\mu_i \geq \mu_0)}{1 + K},$$

then $\mathbb{P}_{\alpha=1}(P_S \leq \alpha) \leq \alpha$ *for all* $\alpha \in [0, 1]$.
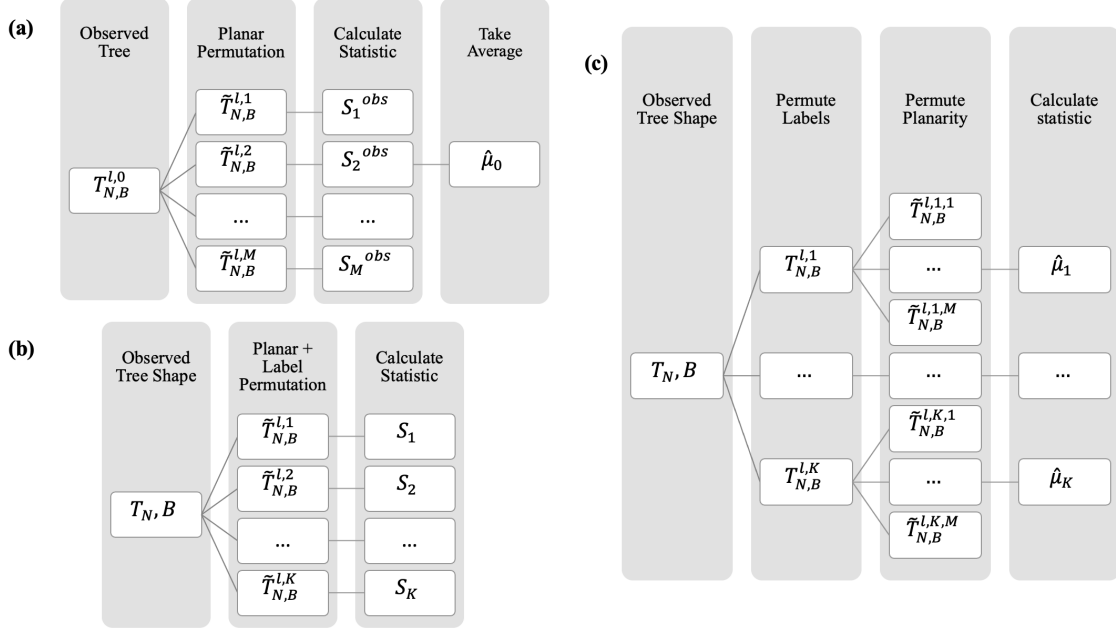
Figure 8: **A schematic for calculation of p-values.** In (a), we obtain $\hat{\mu}_0$ as the average of $\{S_i^{obs}\}_{i=1}^M$ statistics obtained by permutation of the planarity of $T_{N,B}^{\ell,0}$. In (b), we obtain an empirical null distribution of S given T_N, B by permuting planarity and tip labels. In (c), we generate a sample of $\{\hat{\mu}_i\}_{i=1}^K$ from the null distribution given T_N, B. The sample in (a) is used in both computation of p_S, p_T, the samples in (b) are needed to compute p_T, and the samples in (c) are needed to compute p_S.

Proof. First we condition on μ_0, T_N.

$$\mathbb{P}_{\alpha=1}(P_S \leq \alpha \mid \mu_0, T_N) = \mathbb{E}_{\alpha=1}\left[\mathbb{1}\left\{\frac{1}{K+1}\sum_{i=0}^K \mathbb{1}(\mu_i \geq \mu_0) \leq \alpha\right\} \mid \mu_0, T_N\right]$$

$$\leq \mathbb{E}_{\alpha=1}\left[\mathbb{1}\left\{\sum_{T_{N,B}^{\ell,i}} \mathbb{P}(T_{N,B}^{\ell,i}|T_N)\ \mathbb{1}\left(\mu(T_{N,B}^{\ell,i}) \geq \mu_0\right) \leq \alpha\right\} \mid \mu_0, T_N\right],$$

since the sample proportion converges to the population proportion by the Law of Large Numbers as $K \to \infty$, the last inequality results from Portmanteau's Lemma (Van der Vaart, 2000). Then

$$\mathbb{P}_{\alpha=1}(P_S \leq \alpha \mid T_N) \leq \mathbb{E}_{\alpha=1}\left[\mathbb{E}_{\alpha=1}\left\{\mathbb{1}\left(\sum_{T_{N,B}^{\ell,i}} \mathbb{P}(T_{N,B}^{\ell,i}|T_N)\ \mathbb{1}\left(\mu(T_{N,B}^{\ell,i}) \geq \mu_0\right) \leq \alpha\right) \mid \mu_0, T_N\right\}\right]$$

$$= \sum_{T_{N,B}^{\ell,j}} \mathbb{P}(T_{N,B}^{\ell,j}|T_N)\ \mathbb{1}\left(\sum_{T_{N,B}^{\ell,i}} \mathbb{P}(T_{N,B}^{\ell,i}|T_N)\ \mathbb{1}\left(\mu(T_{N,B}^{\ell,i}) \geq \mu(T_{N,B}^{\ell,j})\right) \leq \alpha\right)$$

Lemma A1 of Harrison (2012) states for all $t_0, ..., t_n \in [-\infty, \infty], \alpha, w_0, ..., w_n \in [0, \infty]$, then

$$\sum_{k=0}^n w_k \mathbb{1}\left(\sum_{i=0}^n w_i \mathbb{1}(t_i \geq t_k) \geq \alpha\right) \leq \alpha.$$

From this result, we deduce $\mathbb{P}_{\alpha=1}(P_S \leq \alpha \mid T_N) \leq \alpha$ and therefore $\mathbb{P}_{\alpha=1}(P_S \leq \alpha) \leq \alpha$. $\qquad \square$

Theorem 10. *p_S and p_T are asymptotically valid p-values.*

Proof. To show the result for p_S, notice that $\hat{\mu}_i \xrightarrow{p} \mu_i = \mu(T_{N,B}^{\ell,i})$ by the Law of Large Numbers as $M \to \infty$. Therefore, by the Continuous Mapping Theorem, $\hat{\mu}_i - \hat{\mu}_0 \xrightarrow{p} \mu_i - \mu_0$. By Lemma 11 in Appendix Section 9.5, for all $i = 1, ..., K$,

$$\mathbb{1}(\hat{\mu}_i - \hat{\mu}_0 \geq 0) \xrightarrow{p} \mathbb{1}(\mu_i - \mu_0 \geq 0),$$

Again applying Continuous Mapping Theorem, $p_S \xrightarrow{p} P_S$ as $M \to \infty$. Hence, $\lim_{M \to \infty} \mathbb{P}_{\alpha=1}(p_S \leq \alpha) = \mathbb{P}_{\alpha=1}(P_S \leq \alpha) \leq \alpha$ for all $\alpha \in [0,1]$, by Lemma 9.

To show the result for p_T, first note that with probability 1,

$$p_T \geq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{K} \sum_{j=1}^{K} \mathbb{1}(S_j \geq S_i^{obs}) := p_T'$$

and therefore it suffices to show p_T' is a valid p-value. Switching the order of summation gives $p_T' = \frac{1}{K} \sum_{j=1}^{K} \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(S_j \geq S_i^{obs})$. First consider S_j as fixed. As M increases

$$\frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(S_j \geq S_i^{obs}) \xrightarrow{a.s.} \mathbb{P}_{\alpha=1}(S_j \geq S^0 \mid T_{N,B}^{\ell,0}) := f(S_j)$$

where $S^0 \sim \mathbb{P}_{\alpha=1}(S \mid T_{N,B}^{\ell,0})$. Now $\frac{1}{K} \sum_{j=1}^{K} f(S_j)$ is a permutation p-value, which implies p_T is an asymptotically valid p-value. $\qquad \square$

In our simulations, we take $K = M$ between 200 and 500. We choose to use a permutation test instead of a likelihood-ratio test because we do not have an analytical expression for the likelihood of a ranked (non-planar) partially labeled tree shape under the alternative.

5.2 Testing in the Bayesian framework

In the previous section we assumed that a ranked and partially labeled phylogeny was observed, however, phylogenies are typically not directly observed. Here we consider the case when one would use BEAST, or other Bayesian inference implementation, to generate a posterior distribution of the trees given molecular sequence data (Suchard et al., 2018; Ronquist et al., 2012). The tip label information (i.e. colors) is not used to generate these posterior trees. To account for phylogenetic uncertainty, we propose to simply estimate the posterior distribution of p-values p_T or p_S, and reject the null hypothesis according to whether there is posterior evidence of the p-values being smaller than the significance value.

We compare our method to BaTS (Bayesian Tip-association Significant testing) proposed by Parker et al. (2008), where a test statistic is obtained for each tree in the posterior sample, and the posterior median m_0 is used as the test statistic. Next, n random permutations of the color labels $\sigma_1, \dots, \sigma_n$ are generated such that all trees in the posterior distribution are relabeled according to the same permutation. From each permutation, a median statistic is obtained to generate a null posterior distribution of the median test statistic. The p-value is obtained by calculating the proportion of m_i values that are more extreme than the observed m_0. Notice that BaTS is effectively ignoring the Bayesian uncertainty by using the posterior median as the test statistic of interest, and therefore may yield small credible intervals. We will investigate this further in Section 6.

5.3 Power of the test by MCMC

To estimate the power of the test under different alternatives, we approximate the distribution of $\hat{\mu}$ or S under our model with $\alpha \neq 1$, conditional on the observed ranked tree shape T_N. We approximate this distribution via Metropolis-Hastings (Hastings, 1970). Given current state $\tilde{T}_{N,B}^{\ell,y}$, our proposal distribution generates $\tilde{T}_{N,B}^{\ell,x}$ by first assigning tip labels uniformly and then a node is chosen uniformly at random to swap its left and right subtrees, among those that are not cherries of the same type. This proposal is symmetric and therefore, our acceptance probability is simply $r = \frac{\mathbb{P}_\alpha(\tilde{T}_{N,B}^{\ell,x})}{\mathbb{P}_\alpha(\tilde{T}_{N,B}^{\ell,y})}$. Uniformly sampling the labels is more efficient than a local label move given that we only have two unique tip labels. Since all proposals are generated conditioning on a given tree shape, the stationary distribution of the Markov chain is $\mathbb{P}_\alpha(\tilde{T}_{N,B}^{\ell,i} \mid T_N, B)$. In practice, we found that generating M Metropolis-Hastings steps of planarity swaps per one step of relabeling of the tips, improves the mixing of the chain considerably.

6 Simulation Results

We first use our method to test for phylogenetic trait association for the tree in Figure 1(b) with a test statistic value of $\hat{\mu}_0 = 10.27$. Here we assumed a number of $M = K = 500$ of planar and label permutations to obtain the two p-values $p_S = 0.719, p_T = 0.746$. The two plots in Figure 9 show the approximate null distributions of $\hat{\mu}_0$ and S conditioned on T_N and B. In this case, we have strong support for no phylogenetic association, that is, we would not reject the null hypothesis of no association at the 5% significance level.
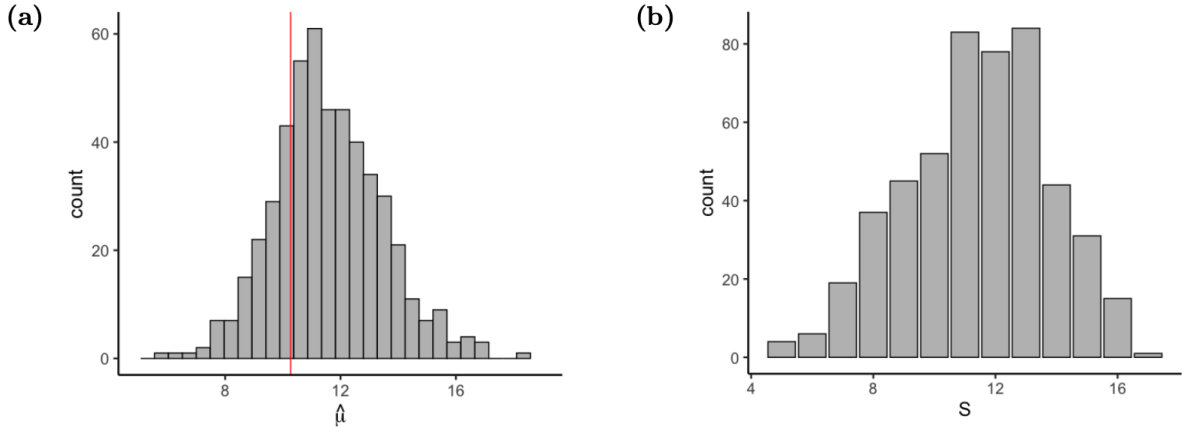

Figure 9: **Analysis of the fixed tree $T_{25,15}^\ell$ in Figure 1(b).** (a): Approximate null distribution of $\hat{\mu}$, with observed value in red. (b): Approximate null distribution of S.

In order to compare the performance of our two proposed tests, we simulated 200 random ranked partially labeled tree shapes according to the CRP-Tree with the following parameter values: $N \in \{20, 50, 100, 200, 500\}$, $\alpha \in \{1, 2, 5, 10, 25\}$, and fraction of one the types $B/N \in \{0.1, 0.25, 0.5\}$. For the results depicted in Table 1 we assume a significance level of 5%, $M = K = 200$ planar and label permutations. In general, both methods concur (97.7% of the time when $\alpha = 1$, and 87.2% of the time when $\alpha > 1$). When the trees are simulated under $\alpha = 1$ (left table), p_S is correctly above the significance threshold only few more times than p_T (99.8% vs 97.5%). When the trees are simulated under $\alpha > 1$, p_T is correctly below the significance 75.4% of the time, and p_S 62.5% of the time. We conclude that p_S is more conservative than p_T and that both p_T and p_S control the Type 1 error rate.

$\alpha = 1$		p_S	
		Do not reject	Reject
p_T	Do not reject	2925	0
	Reject	69	6

$\alpha > 1$		p_S	
		Do not reject	Reject
p_T	Do not reject	2951	0
	Reject	1540	7509

Table 1: **Contingency tables comparing p_T and p_S.** Each entry represents the number of times the two methods rejected or not when the true value is $\alpha = 1$ or $\alpha > 1$. Simulations carried out across a range of N and B values.

6.1 Power analyses

We investigate the power of our tests and compare them to that of Parsimony score (PS), Association Index (AI) and treeSeg via simulation. The Parsimony Score (PS) (Fitch, 1971) counts the minimum number of state changes in the phylogeny in order to reconstruct the states at the parent nodes. Here, the states at the tips are binary and so we use the Fitch algorithm for the score computation. Small values imply strong phylogenetic trait association. The Association Index (AI) (Wang et al., 2001) is defined by $AI = \sum_{i=1}^{N-1} \frac{1-f_i}{2^{m_i-1}}$, where m_i is the number of tips subtended by internal node i and f_i is the frequency of the most common trait value among the tips subtended. Note that smaller value of AI implies stronger phylogenetic trait association, because the numerator $1 - f_i$ is smaller for larger f_i. We also apply the changepoint detection method implemented in treeSeg (Behr et al., 2020) to the sample of trees. In this case, detection of at least one changepoint corresponds to rejecting the null hypothesis.

We first generated the five different ranked tree shapes depicted in Figures 19-23. Three of these tree shapes were generated uniformly at random ($N = 25, 50, 100$) while the other two tree shapes are the most balanced and most unbalanced trees ($N = 100$) according to the criteria defined in Rajanala and Palacios (2021). For each ranked tree shape, we approximated the power of the test under 4 different frequencies of the two types: $B/N \in \{0.1, 0.25, 0.4, 0.5\}$, and under 4 alternatives: $\alpha \in \{2, 5, 10, 20\}$. We set $M = 300$ for calculating $\hat{\mu}_0$ and $\{S_i^{obs}\}_{i=1}^{M}$ and generated 500 planar and label MCMC steps.

The power approximations are displayed in the tables of Figures 19-23 in the Appendix. We see that our two methods have much better power than any of the pre-existing statistics AI and PS, as well as treeSeg in all cases. We generally observe increasing power as B/N and α increase for each fixed tree. We note that we do not require a very large tree in order to be able to detect phylogenetic trait association. However, we do notice that less balanced tree shapes and an imbalance of label types may result in lower power.

We extend our power study to 100 randomly simulated ranked tree shapes with $N = 50$. The boxplots for the power of each test under $B/N \in \{0.1, 0.5\}$ and $\alpha \in \{2, 5, 10, 20\}$ are depicted in Figure 10. We confirm that our methods consistently perform better than AI and PS, with the $\hat{\mu}$ statistic achieving the highest power overall. We do not compare to treeSeg due to its high computational time.

6.2 Posterior Validation of p-values

As a validation check in the Bayesian setting, we first simulated two phylogenies ($N = 50$, $B = 20$) from the CRP-Tree model with $\alpha = 1$ and $\alpha = 10$ respectively, and simulated DNA sequences at the tips of each phylogeny. We then used BEAST (Suchard et al., 2018) to estimate the two posterior distributions and tested the null hypothesis of $\alpha = 1$ in both cases. Details on simulation experiment can be found in Appendix Section 9.7. We compared the posterior distributions of the p-values obtained with our test to the p-values obtained with BaTS. When $\alpha = 1$, the posterior mean p-value obtained with our method is 0.86, and the posterior median is 0.866. When $\alpha = 10$, the posterior mean p-value is 0.0021 and the posterior median is 0.0021. The posterior distributions of the p-values are depicted in Figure 11. In both cases, the user would have correctly concluded the true association. However, the BaTS p-values are both 0, which
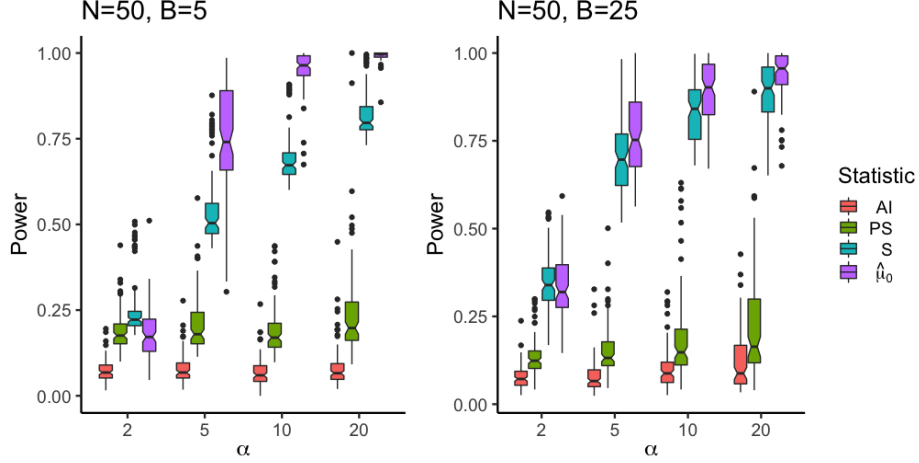
Figure 10: **Power simulations for 100 tree shapes with $N = 50$.** The left plot shows the power for $B = 5$ and the right plot shows the power for $B = 25$.

implies it would incorrectly reject the first case of $\alpha = 1$. We will show in the next section that we usually obtain concordant conclusions from the posterior distribution of p-values and BaTS p-values.
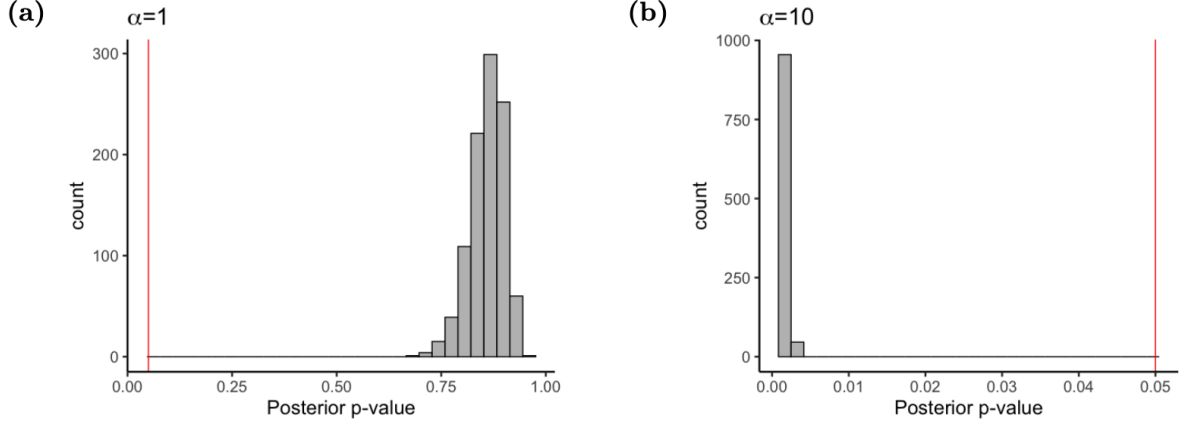

Figure 11: **Posterior distributions of p-values.** Posterior distribution of p_T for the simulation with (a) $\alpha = 1$, and (b) $\alpha = 10$. Red line is marks p-value of 0.05.

7 Case studies

We first apply our tests to two real data studies in which the ranked partially labeled tree shapes are available (without known uncertainty). We then apply our test to two studies in which the posterior distribution of trees is estimated via MCMC from molecular sequence data.

7.1 A breast cancer study

We re-analyze a publicly available breast cancer gene expression study from 98 patients (van't Veer et al., 2002) in which more than five thousand genes were found to be significantly associated to breast cancer,

23

out of a pool of approximately 25 thousand genes. An additional six clinical responses were collected: BRCA mutation, estrogen receptor expression, histological grade, lymphocytic infiltration, angioinvasion, and development of distant metastasis within 5 years, although this last variable had missing data and is excluded in this study. Behr et al. (2020) apply a hierarchical clustering algorithm using a similarity metric on these regulatory genes to create a tree. The results of our tests for tree association to each of these clinical responses are shown in Figure 12. The only trait that is not rejected with our methods is the association of the angioinvasion trait with the tree structure. These results are consistent with Behr et al. (2020).
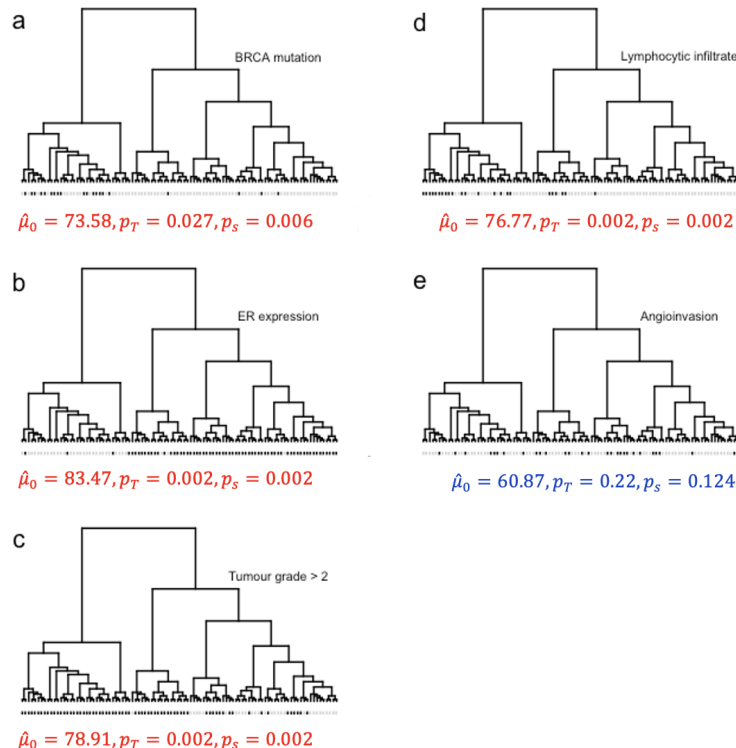

Figure 12: **Phylogenetic association tests of five clinical responses associated to breast cancer** ($N = 98$). The tick marks indicate whether the subject had the trait or not. The angioinvasion trait in (e) is the only case where we would not reject the null hypothesis with our tests.

7.2 Sexually attractive traits in Swordtail fish

In evolutionary biology, swordtail fish (*Xiphophorus*) are a classic model for studying sexual selection (Darwin, 1871). Decades of research have shown that females have preferences for large male body size (Ryan and Wagner Jr, 1987; Rosenthal and Evans, 1998; Cummings and Mollaghan, 2006) and several sexually selected ornaments, including the "sword" ornament for which the genus is named (Rosenthal et al., 2001; Basolo and Trainor, 2002). Preising et al. (2022) collected wild-caught individuals and used whole genome sequencing to infer phylogenetic relationships and examine the co-evolution of certain traits within the *Xiphophorus* clade. Their phylogenetic tree, constructed via maximum likelihood, is shown in Figure 13. Here, our interest is to test for phylogenetic association of the two main traits: presence/absence of the sword, and whether the size of the body is larger than 26.5 inches.

We generated 500 planar permutations and 500 label permutations to approximate the null distribution of $\hat{\mu}$ for the two traits (Figure 14). For the size trait, we obtained $\hat{\mu}_0 = 13.3$, $p_T = 0.714$, and $p_S = 0.686$, and
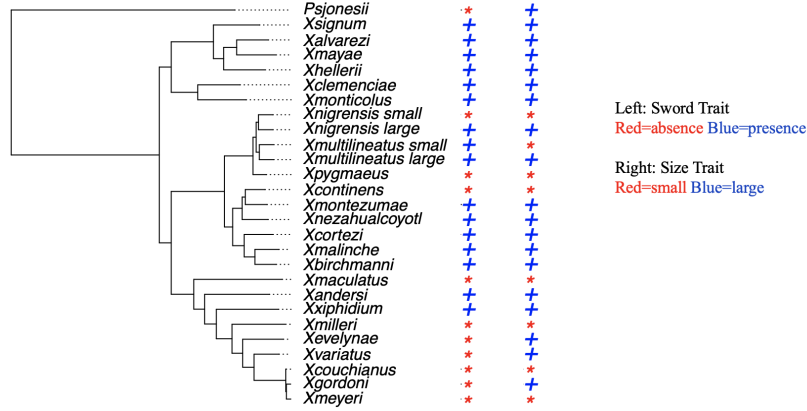
Figure 13: **A phylogeny of 27 swordtail fish species and the two traits of interest.** For the sword trait, $B = 16$ and for the size trait, $B = 19$.

so we conclude that the size of the fish is not associated with the phylogeny. This is somewhat expected since the phylogeny is based on the whole genome and body size is associated with a small number of polymorphic sites (Lampert et al., 2010). For the sword trait, we obtained $\hat{\mu}_0 = 17.05, p_T = 0.0706, p_S = 0.018$ and so the presence/absence of the sword appears to be associated with the tree topology.
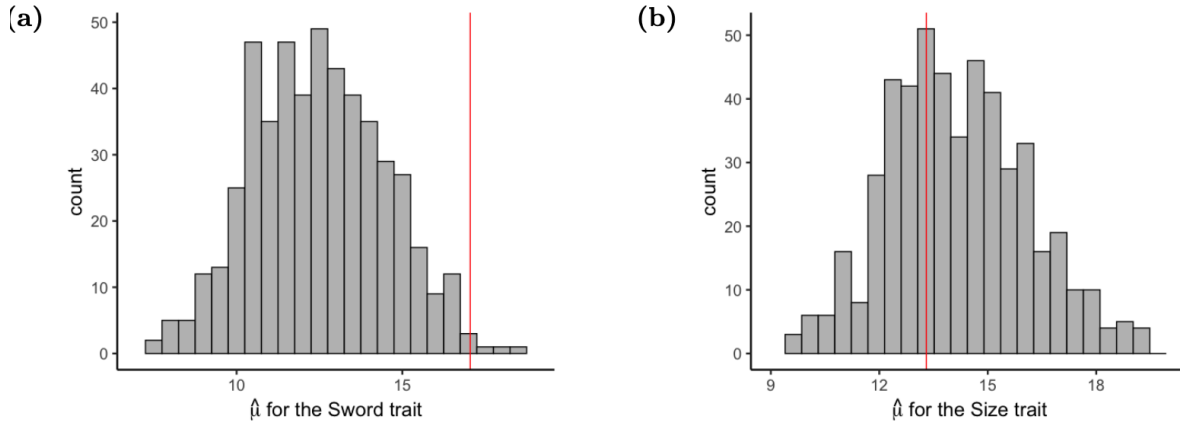

Figure 14: **Analysis of the swordtail fish sequences.** Null Distribution of $\hat{\mu}_0$ for the presence/absence of Sword trait (a) and large/small Size trait (b). Red line is the observed value.

7.3 Transmission cycles of Brazilian yellow fever virus

Brazil recently experienced yellow fever virus (YFV) outbreaks in multiple states causing more than 700 deaths between 2016 and 2018. YFV is the most severe mosquito-borne infection in South America. Nonhuman primates are usually infected by *Haemagogus* spp. and *Sabethes* spp., while humans are infected by *Aedes* spp. In (Faria et al., 2018), the authors generated 65 complete viral genomes collected from 33 infected humans and 32 non-human primates across several states in Brazil during 2016-2017. In order to investigate whether the virus is spreading between human and nonhuman primates, the authors estimated the movement of YFV lineages between humans and nonhuman primates using a structured coalescent model. The authors estimated a variable rate of transmission from nonhuman primates to humans rising from zero around

November 2016 and reaching a peak in February 2017. Here, we reanalyze the same sequences in order to investigate whether the virus is spreading between humans and nonhuman primates. If the virus was spreading between the two populations, there should not be a phylogenetic association with the human-nonhuman trait.

To incorporate phylogenetic uncertainty, we obtained a sample of 5,000 phylogenetic trees (after thinning every 100,000 iterations) from the posterior distribution generated with the software BEAST (Suchard et al., 2018). The posterior distribution of phylogenies is completely agnostic to the human-nonhuman label. Details of model assumptions can be found Appendix Section 9.7. We calculate $\hat{\mu}_0$ and p-values p_T for each tree in the posterior sample. We generated 500 permutations of the planarity to generate each $\hat{\mu}_0$ and 500 permutations of the labels to simulate each null distribution conditional on each ranked tree shape in the posterior. The posterior distribution of the p-values shown in Figure 15(a) has a mean of 0.04 and median of 0.031, with 70% of the values below the 0.05 significance level. This result suggests that the virus spreads mainly within each population. This new result contradicts the original finding of variable migration from nonhuman to human populations. However, the authors in the original study alert caution that hypotheses of human-to-human transmission should not be tested directly using phylogenetic data alone, due to large undersampling of NHP infections.
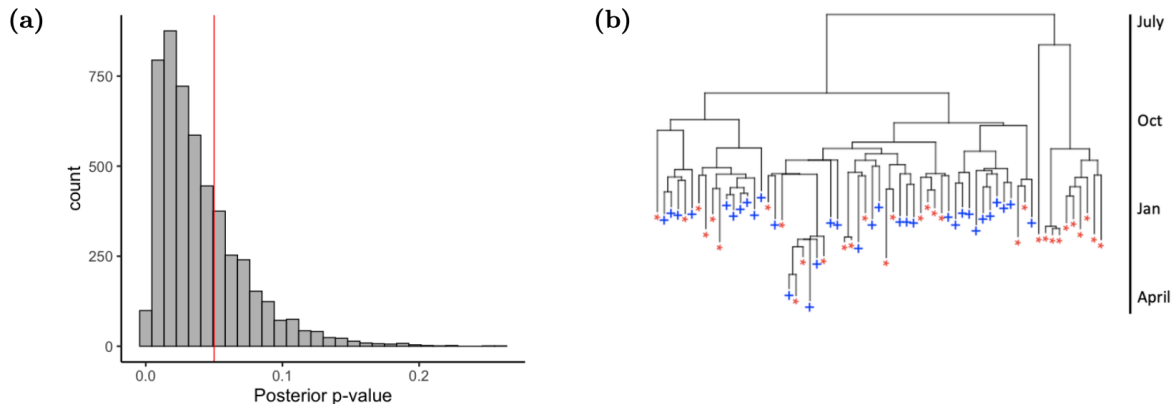

Figure 15: **Analysis of Yellow Fever Virus**. (a): Posterior distribution of p-values p_T for the Yellow fever virus study in Brazil. Red line is placed at p-value of 0.05. (b): Maximum-clade credibility tree, with branch lengths indicative of the date of collected samples in 2017. Blue tips correspond to human samples and red to nonhuman primate samples.

The observed posterior median value of $\hat{\mu}_0$ is 39.15. When we apply BaTS to our test statistic with 500 permutations of the label set, the 95% credible interval obtained for the medians of $\hat{\mu}_0$ under the null is $[30.49, 31.48]$ which allows us to obtain the same conclusion with our method: observed data suggests some level of preferential attachment.

Finally, we obtained a single tree from the posterior distribution that corresponds to the maximum clade credibility tree in Figure 15(b) and performed our test on this tree ignoring phylogenetic uncertainty. In this case, $\hat{\mu}_0 = 42.62, p_T = 0.0066$, in agreement with our previous result.

7.4 Population structure in H1N1 Transmission

In early 2009, the swine-origin influenza A (H1N1) virus originated as a novel combination of influenza genes. In just one year, the estimated number of H1N1 cases arose to more than 60 million in 2010 (Centers for Disease Control and Prevention, 2019). Smith et al. (2009) study the origins and evolutionary genomics of the H1N1 pandemic using phylogenetic analyses on related virus genomes, such as H3N2, classical swine H1N1, and North American avian. Suchard et al. (2018) analyze 50 H1N1 viral genome sequences, a subset

of the original study to estimate the origin date of the pandemic, the growth, and basic reproductive number. A sample of 1,000 posterior trees was generated in BEAST (thinning every 10,000 states) without using the geographical location of the sequences. We use our test to determine if there is population structure in the transmission of H1N1, a question that was not investigated in the original study.

Since geographic location is not binary, we choose to split our data into USA (22 sequences) and non-USA (28 sequences). Using the same procedures as the previous example, we generated $M = 500$ permutations of the planarity to generate each $\hat{\mu}_0$ and $K = 500$ permutations of the labels to simulate each null distribution conditional on each ranked tree shape in the posterior. The posterior distribution of the p-values (Figure 16(a)) has a mean of 0.0525 and median of 0.042, with close to 60% of the values below the 0.05 significance level. Therefore, there seems to exist population structure in the spread of influenza between strands in the USA and otherwise. The observed posterior median value of $\hat{\mu}_0$ is 31.32. Using BaTS with 500 label permutations, the 95% credible interval obtained for the medians of $\hat{\mu}_0$ under the null is $[23.43, 24.33]$ which allows us to obtain the same conclusion with our method: observed data suggests some level of preferential attachment.
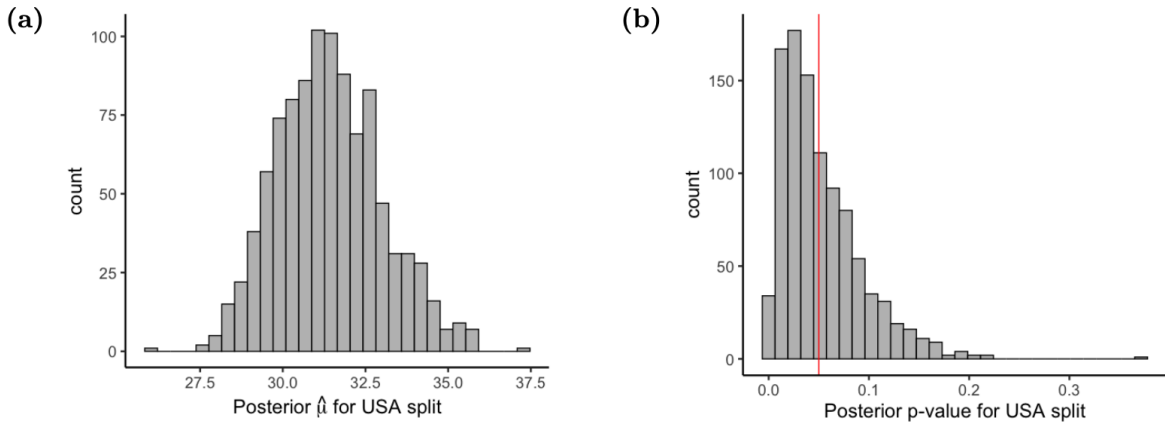
(a)

(b)

Figure 16: **Analysis of H1N1 Virus with USA/non-USA trait split**. Posterior distribution of $\hat{\mu}$ (a) and posterior p-values p_T (b) calculated on a posterior sample of 1000 trees. Red line is 0.05.

8 Discussion and Extensions

We propose a nonparametric permutation test of phylogenetic association with a binary trait. We empirically demonstrate that our test is more powerful than the parsimony score, the association index and treeSeg. We also showed that our test is computationally efficient. In average, testing takes about 7 seconds for a tree with 100 tips and 40 seconds for a tree with 500 tips. In particular, our test is much faster than treeSeg, which could not produce results for larger trees. In addition, we extend our test to the setting when the phylogeny is not directly observed and instead, a posterior distribution of phylogenetic trees is available.

The proposed test statistic, the mean number of same type attachments in a tree, is a sufficient statistic for the single parameter of the CRP-Tree model proposed in this manuscript. In the CRP-Tree model, this parameter controls the likelihood of lineages to attach to lineages of the same type and hence, the test statistic arises in a natural and interpretable way. However, our proposed CRP-Tree model cannot be used directly in a likelihood ratio test formulation. The difficulty lies in the fact that the CRP-Tree model is a model on planar and ranked trees, a finer tree resolution than the one usually observed (non-planar and ranked trees). Nevertheless, our proposed statistic is amenable to a permutation test, and in fact, the permutation p-value corresponds to the Monte Carlo p-value obtained by sampling planar ranked and partially labeled

tree shapes with the same ranked tree shape as the observed tree. Although our permutation p-value is a special case to the ones described in (Hemerik and Goeman, 2018; Ramdas et al., 2022), the validity of our p-values relies on the fact that randomly permuting leaf labels and planarity, effectively generates i.i.d. samples from the conditional CRP-Tree model with $\alpha = 1$ (null model). We note that our test statistic, the average number of same type attachments, is closely related to the parsimony score, the minimum possible value of different-type attachments. However, our statistic showed better power and overall performance in all scenarios considered.

8.1 Possible extensions

Our method is applicable to categorical traits with more than two possible values. Under the CPR-Tree model, a node attaches to a node of the same type or to a node of a different type, according to the same probabilities derived for the binary case. Therefore, the test statistic remains the same. However, in our experience, there usually is a large loss of power when the number of categories increases and therefore, we do not pursue it here.

Another possible extension is to allow for missing tip information. An effective way would consist in replacing our statistic, to the expected number of same type attachments conditional only on the partial labels, effectively, integrating out uncertainty in tip labels. Although we do not actively pursue this extension here, we do not anticipate any difficulty in its implementation.

We can also consider an extension of the CRP-Tree model to continuous trait values. The notion of preferentially attaching to a node of the same color is replaced by attaching to tips with smaller absolute difference. When $\alpha \to \infty$, new tips would always attach to the tip with the smallest absolute difference. Letting Y_i denote the trait value for the ith tip being added, we define the test statistic to be $S = \sum_{i=1}^{N} |Y_i - Y_{d_i}|$, where (i, d_i) is an attachment. If there is preferential attachment, then S will be small. Notice that taking $Y_i \in \{0, 1\}$ for binary traits gives S equal to the number of same attachments. Again, permutation tests can be utilized for assessing whether there is any association or not. This bypasses the need to assume any parametric model of trait evolution, such as Brownian Motion (Pagel, 1999; Blomberg et al., 2003), or the Ornstein–Uhlenbeck process (Felsenstein, 1988; Butler and King, 2004).

We are acutely aware that our test ignores branch length information. We could extend our model by allowing α to vary with time, where larger α would imply shorter branch lengths for an attachment of the same type. If α were allowed to vary with time, say according to a Poisson process, then we could test hypotheses on the subtrees sequentially to discover the location of these changepoints. That is, a subtree can be considered as a CRP-Tree realization with smaller values of N, B. In addition, we could place priors on $\alpha(t)$ and use Bayesian nonparametric methods for inferring $\alpha(t)$.

Finally, we note that method is not suitable for complex traits in which multiple genes are involved. An extension of our method for networks could be possible and subject of future research.

References

D. J. Aldous. Exchangeability and related topics. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, pages 1–198, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg. ISBN 978-3-540-39316-0.

M. A. Ansari and X. Didelot. Bayesian inference of the evolution of a phenotype distribution on a phylogenetic tree. *Nature*, 204:89–98, 2016.

A. L. Basolo and B. C. Trainor. The conformation of a female preference for a composite male trait in green swordtails. *Animal Behaviour*, 63(3):469–474, 2002.

M. Behr, M. A. Ansari, A. Munk, and C. Holmes. Testing for dependence on tree structures. *PNAS*, 117 (18):9787–9792, 2020.

S. Blomberg, T. Garland, and A. Ives. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution*, 57:717–745, 2003.

R. Borges, J. P. Machado, C. Gomes, A. P. Rocha, and A. Antunes. Measuring phylogenetic signal between categorical traits and phylogenies. *Bioinformatics*, 35(11):1862–1869, 2019.

M. A. Butler and A. A. King. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, 164(6):683–695, 2004.

Centers for Disease Control and Prevention. 2009 H1N1 pandemic (H1N1PDM09 virus), Jun 2019. URL https://www.cdc.gov/flu/pandemic-resources/2009-h1n1-pandemic.html.

K. P. Choi, A. Thompson, and T. Wu. On cherry and pitchfork distributions of random rooted and unrooted phylogenetic trees. *arXiv*, arXiv:2002.12643v1 [math.PR], 2020.

S. Cleary, M. Fischer, R. C. Griffiths, and R. Sainudiin. Some distributions on finite rooted binary tres. Technical report, UCDMS Research Report No. UCDMS2015/2, School Of Mathematics and Statistics, University of Canterbury, Christchurch, NZ, 2015.

M. Cummings and D. Mollaghan. Repeatability and consistency of female preference behaviours in a northern swordtail, Xiphophorus nigrensis. *Animal Behaviour*, 72(1):217–224, 2006.

C. Darwin. The descent of man. *New York: D. Appleton*, 1871.

D. Faith. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61:1–10, 1992.

N. R. Faria, M. U. G. Kraemer, S. C. Hill, J. G. de Jesus, R. S. Aguiar, F. C. M. Iani, J. Xavier, J. Quick, L. du Plessis, S. Dellicour, J. Thézé, R. D. O. Carvalho, G. Baele, C.-H. Wu, P. P. Silveira, M. B. Arruda, M. A. Pereira, G. C. Pereira, J. Lourenço, U. Obolski, L. Abade, T. I. Vasylyeva, M. Giovanetti, D. Yi, D. J. Weiss, G. R. W. Wint, F. M. Shearer, S. Funk, B. Nikolay, V. Fonseca, T. E. R. Adelino, M. A. A. Oliveira, M. V. F. Silva, L. Sacchetto, P. O. Figueiredo, I. M. Rezende, E. M. Mello, R. F. C. Said, D. A. Santos, M. L. Ferraz, M. G. Brito, L. F. Santana, M. T. Menezes, R. M. Brindeiro, A. Tanuri, F. C. P. dos Santos, M. S. Cunha, J. S. Nogueira, I. M. Rocco, A. C. da Costa, S. C. V. Komninakis, V. Azevedo, A. O. Chieppe, E. S. M. Araujo, M. C. L. Mendonça, C. C. dos Santos, C. D. dos Santos, A. M. Mares-Guia, R. M. R. Nogueira, P. C. Sequeira, R. G. Abreu, M. H. O. Garcia, A. L. Abreu, O. Okumoto, E. G. Kroon, C. F. C. de Albuquerque, K. Lewandowski, S. T. Pullan, M. Carroll, T. de Oliveira, E. C. Sabino, R. P. Souza, M. A. Suchard, P. Lemey, G. S. Trindade, B. P. Drumond, A. M. B. Filippis, N. J. Loman, S. Cauchemez, L. C. J. Alcantara, and O. G. Pybus. Genomic and epidemiological monitoring of Yellow Fever virus transmission potential. *Science*, 361(6405):894–899, 2018.

J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, 125:17–15, 1985.

J. Felsenstein. Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19(1): 445–471, 1988.

W. Fitch. Toward defining the course of evolution: minimal change for a specific tree topology. *Systematic Zoology*, 20:406–416, 1971.

D. Ford, F. A. Matsen, and T. Stadler. A method for investigating relative timing information on phylogenetic trees. *Systematic Biology*, 58(2):167–183, 2009.

T. Garland Jr, P. H. Harvey, and A. R. Ives. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Systematic biology*, 41(1):18–32, 1992.

T. Griffiths, M. Jordan, J. Tenenbaum, and D. Blei. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16, 2003.

O. Hagen, K. Hartmann, M. Steel, and T. Stadler. Age-dependent speciation can explain the shape of empirical phylogenies. *Systematic Biology*, 64(3):432–440, May 2015.

M. T. Harrison. Conservative hypothesis tests and confidence intervals using importance sampling. *Biometrika*, 99(1):57–69, 2012.

J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, pages 53–65, 1973.

W. K. Hastings. Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109, 1970.

J. Hemerik and J. Goeman. Exact testing with random permutations. *Test*, 27(4):811–825, 2018.

T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

M. D. Karcher, J. A. Palacios, S. Lan, and V. N. Minin. phylodyn: an R package for phylodynamic simulation and inference. *Molecular ecology resources*, 17(1):96–100, 2017.

J. Kim, N. A. Rosenberg, and J. A. Palacios. Distance metrics for ranked evolutionary trees. *Proceedings of the National Academy of Sciences*, 117(46):28876–28886, 2020.

J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

K. P. Lampert, C. Schmidt, P. Fischer, J.-N. Volff, C. Hoffmann, J. Muck, M. J. Lohse, M. J. Ryan, and M. Schartl. Determination of onset of sexual maturation and mating behavior by melanocortin receptor 4 polymorphisms. *Current Biology*, 20(19):1729–1734, 2010.

C. Lozupone and R. Knight. Unifrac: a new method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005.

A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164 (1):81–92, 2000.

V. N. Minin, E. W. Bloomquist, and M. A. Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471, 2008.

F. Murtagh. Counting dendograms. *Discrete Applied Mathematics*, 7(2):191–199, 1984.

N. F. Müller, D. A. Rasmussen, and T. Stadler. The structured coalescent and its approximations. *Molecular Biology and Evolution*, 34(11):2970–2981, 06 2017. ISSN 0737-4038. doi: 10.1093/molbev/msx186.

T. Münkemüller, S. Lavergne, B. Bzeznik, S. Dray, T. Jombart, K. Schiffers, and W. Thuiller. How to measure and test phylogenetic signal. *Methods in Ecology and Evolution*, 3:743–756, 2012.

M. Notohara. The coalescent and the genealogical process in geographically structured population. *Journal of mathematical biology*, 29(1):59–75, 1990.

M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401:877–884, 1999.

J. A. Palacios, A. Véber, L. Cappello, Z. Wang, J. Wakeley, and S. Ramachandran. Bayesian estimation of population size changes by sampling Tajima's trees. *Genetics*, 213(3):967–986, 2019.

J. Parker, A. Rambaut, and O. G. Pybus. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics, and Evolution*, 8:239–246, 2008.

G. A. Preising, T. Gunn, J. J. Baczenas, A. Pollock, D. L. Powell, T. O. Dodge, J. A. Machin Kairuz, M. Savage, Y. Lu, M. Fitschen-Brown, M. Cummings, S. Thakur, M. Tobler, O. Ríos-Cardenas, M. Morris, and M. Schumer. Recurrent evolution of small body size and loss of the sword ornament in northern swordtail fish. *bioRxiv*, 2022. doi: 10.1101/2022.12.24.521833.

Z. S. Qin. Clustering microarray gene expression data using weighted Chinese restaurant process. *Bioinformatics*, 22(16):1988–1997, 2006.

S. Rajanala and J. A. Palacios. Statistical summaries of unlabelled evolutionary trees and ranked hierarchical clustering trees. *arXiv preprint arXiv:2106.02724*, 2021.

A. Rambaut and N. C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, 1997.

A. Ramdas, R. Foygel Barber, E. J. Candes, and R. J. Tibshirani. Permutation tests using arbitrary permutation distributions. *arXiv e-prints*, pages arXiv–2204, 2022.

F. Ronquist, M. Teslenko, P. Van Der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. Mrbayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542, 2012.

G. G. Rosenthal and C. S. Evans. Female preference for swords in Xiphophorus helleri reflects a bias for large apparent size. *Proceedings of the National Academy of Sciences*, 95(8):4431–4436, 1998.

G. G. Rosenthal, T. Y. Flores Martinez, F. J. García de León, and M. J. Ryan. Shared preferences by predators and females for male ornaments in swordtails. *The American Naturalist*, 158(2):146–154, 2001.

M. J. Ryan and W. E. Wagner Jr. Asymmetries in mating preferences between species: female swordtails prefer heterospecific males. *Science*, 236(4801):595–597, 1987.

R. Sainudiin and A. Véber. A Beta-splitting model for evolutionary trees. *Royal Society Open Science*, 3 (5):160016, 2016.

R. Sainudiin and D. Welch. The transmission process: A combinatorial stochastic process for the evolution of transmission trees over networks. Technical report, UCDMS Research Report No. UCDMS2016/1, School of Mathematics and Statistics, University of Canterbury, Christchurch, NZ, 2016.

R. Sainudiin, T. Stadler, and A. Veber. Finding the best resolution for the Kingman–Tajima coalescent: theory and applications. *Journal of Mathematical Biology*, 70:1207–1247, 2015.

M. Slatkin and W. Maddison. A cladistic measure of gene flow measured from the phylogenies of alleles. *Genetics*, 123(3):603–613, 1989.

G. J. D. Smith, D. Vijaykrishna, J. Bahl, S. J. Lycett, M. Worobey, O. G. Pybus, S. K. Ma, C. L. Cheung, J. Raghwani, S. Bhatt, J. S. M. Peiris, Y. Guan, and A. Rambaut. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. *Nature*, 459:1122–1125, 2009.

E. J. Stewart, R. Madden, G. Paul, and F. Taddei. Aging and death in an organism that reproduces by morphologically symmetric division. *PLOS Biology*, 3(2):e45, 2005.

M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus evolution*, 4(1):vey016, 2018.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

L. van't Veer, H. Dai, M. van de Vijver, and et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–546, 2002.

T. Wang, Y. Donaldson, R. Brettle, J. Bell, and P. Simmonds. Identification of shared populations of Human Immunodeficiency Virus Type 1 infecting microglia and tissue macrophages outside the central nervous system. *Journal of Virology*, 75(23):11686–11699, 2001.

C. Webb. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, 156(2):145–155, 2000.

C. Webb, D. Ackerly, M. McPeek, and M. Donoghue. Phylogenies and community ecology. *Annual Review of Ecology, Evolution, and Systematics*, 33:475–505, 2002.

9   Appendix

9.1   Expected Value of S

Let X_k be the indicator of the event that node k attaches a node of the same color label, and let $\mathcal{C} = \{C_1, \dots, C_N\}$ be an ordering of attachments. Then, the sequence $\{W_k = w_k\}$ is known. We know $X_k \mid \mathcal{C} \sim Bern(\frac{\alpha w_k}{(k-1-w_k)+\alpha w_k})$ and X_k, X_j are independent for $k \neq j$. Therefore,

$$\mathbb{E}[S|\mathcal{C}] = \sum_{k=3}^{N} \frac{\alpha w_k}{(k-1-w_k)+\alpha w_k},$$

$$\mathbb{E}[S] = \mathbb{E}\big[\mathbb{E}[S|\mathcal{C}]\big] = \sum_{k=3}^{N} \mathbb{E}\left[\frac{\alpha W_k}{(k-1-W_k)+\alpha W_k}\right].$$

In order to calculate $\mathbb{E}[S]$, we only need the distribution of W_k, which does not depend on α. If we choose k balls from Urn 1 without replacement, the number of balls that match the color of the kth ball out of the first $k-1$ chosen is W_k. Let $Y_k^{(B,N)}$ be the number of blue balls drawn after drawing k balls without replacement from an urn with B blue balls and $N-B$ red balls. Then $Y_k^{(B,N)} \sim \text{Hypergeometric}(B, N, k)$, that is, for $i \in \{\max(0, k-(N-B)), \dots, \min(B, k)\}$:

$$\mathbb{P}(Y_k^{(B,N)} = i) = \frac{\binom{B}{i}\binom{N-B}{k-i}}{\binom{N}{k}}.$$

Now, given that k balls have already been drawn from our pool of B blue and $N-B$ red balls, let $Z_k^{(B,N)}$ be the indicator that the kth drawn ball is blue. Then,

$$
\begin{aligned}
\mathbb{P}(W_k = i) &= \mathbb{P}(Y_{k-1} = i, Z_k = 1) + \mathbb{P}(Y_{k-1} = k-1-i, Z_k = 0) \\
&= \mathbb{P}(Z_k = 1|Y_{k-1} = i)\mathbb{P}(Y_{k-1} = i) + \mathbb{P}(Z_k = 0|Y_{k-1} = k-1-i)\mathbb{P}(Y_{k-1} = k-1-i) \\
&= \frac{B-i}{N-(k-1)} \frac{\binom{B}{i}\binom{N-B}{k-1-i}}{\binom{N}{k-1}} + \frac{N-B-i}{N-(k-1)} \frac{\binom{B}{k-1-i}\binom{N-B}{i}}{\binom{N}{k-1}} \\
&= \frac{B}{N} \times \frac{\binom{B-1}{i}\binom{N-B}{k-1-i}}{\binom{N-1}{k-1}} + \frac{N-B}{N} \times \frac{\binom{B}{k-1-i}\binom{N-B-1}{i}}{\binom{N-1}{k-1}}.
\end{aligned}
$$

We recognize that this probability is

$$\mathbb{P}(W_k = i) = \frac{B}{N} \times \mathbb{P}\left(Y_{k-1}^{(B-1,N-1)} = i\right) + \frac{N-B}{N} \times \mathbb{P}\left(Y_{k-1}^{(N-B-1,N-1)} = i\right). \tag{6}$$

The interpretation of the two Hypergeometric RVs is that $Y_{k-1}^{(B-1,N-1)}$ denotes the number of blue balls in $k-1$ drawn balls from $N-1$ total balls and $B-1$ blues and $Y_{k-1}^{(N-B-1,N-1)}$ denotes the number of red balls in choosing $k-1$ balls from $N-1$ total balls and $N-B-1$ red. Therefore, W_k is a convex combination of two independent Hypergeometric random variables.

Expanding out the previous expression for $\mathbb{P}(W_k = i)$, we get

$$\mathbb{P}(W_k = i) = \frac{B!(N-B)!(k-1)!(N-k)!}{i!(B-i-1)!(k-1-i)!(N-B-(k-1-i))!N!} +$$
$$\frac{B!(N-B)!(k-1)!(N-k)!}{i!(N-B-i-1)!(k-1-i)!(B-(k-1-i))!N!}$$
$$= \left[\frac{1}{(B-i-1)!(N-B-(k-1-i))!} + \frac{1}{(N-B-i-1)!(B-(k-1-i))!} \right] \times$$
$$\frac{B!(N-B)!(k-1)!(N-k)!}{i!(k-1-i)!N!}$$
$$= \frac{\binom{k-1}{i}}{\binom{N}{B}} \left[\binom{N-k}{B-(i+1)} + \binom{N-k}{N-B-(i+1)} \right]. \tag{7}$$

Here is the interpretation for this expression: Suppose we are just ordering the N balls from 1 to N, there are a total of $\binom{N}{B}$ possible ways (the denominator). The event $\{W_k = i\}$ means there are i balls in the first $k-1$ balls the same color as the ith ball. The term $\binom{k-1}{i}$ is the number of ways to choose the locations of these i balls from $k-1$ spots. Next, the remaining $N-k$ spots have a total of $B-(i+1)$ blue balls if the kth ball is blue, and $N-B-(i+1)$ red balls of the kth ball is red, hence the result.

Using Equation (6) and the expected value of Hypergeometric random variables, we have

$$\mathbb{E}[W_k] = \frac{B}{N} \times \frac{(k-1)(B-1)}{(N-1)} + \frac{N-B}{N} \times \frac{(k-1)(N-B-1)}{(N-1)}.$$

Then for $\alpha = 1$,

$$\mathbb{E}[S] = \sum_{k=3}^{N} \mathbb{E}\left[\frac{W_k}{k-1} \right]$$
$$= \frac{(N-2)\big(B(B-1) + (N-B)(N-B-1)\big)}{N(N-1)}$$
$$= (N-2) - \frac{2B(N-B)(N-2)}{N(N-1)}.$$

For $\alpha > 1$, we can use our alternative formulation Equation (7).

$$\mathbb{E}\left[\frac{\alpha W_k}{(k-1-W_k) + \alpha W_k} \right] = \sum_{i=I_{\min}}^{I_{\max}} \frac{\alpha i}{(k-1-i) + \alpha i} \frac{\binom{k-1}{i}}{\binom{N}{B}} \left[\binom{N-k}{B-(i+1)} + \binom{N-k}{N-B-(i+1)} \right]$$
$$= \frac{B}{N} \times \sum_{i=I_{\min}}^{I_{\max}} \frac{\alpha i}{(k-1-i) + \alpha i} \frac{\binom{k-1}{i}\binom{N-k}{B-(i+1)}}{\binom{N-1}{B-1}} +$$
$$\frac{N-B}{N} \times \sum_{i=I_{\min}}^{I_{\max}} \frac{\alpha i}{(k-1-i) + \alpha i} \frac{\binom{k-1}{i}\binom{N-k}{N-B-(i+1)}}{\binom{N-1}{N-B-1}}.$$

Here, $I_{\min}, I_{\max}$ just denote the support. With the convention $\binom{a}{b} = 0$ if $a < b$, we can take $I_{\min} = 0, I_{\max} = k-1$ for each k. This can be calculated numerically.

9.2 Algorithm to reconstruct the sequence of attachments, initial color order, and the order of tips added.

Algorithm 1: Algorithm to reconstruct C and the sequence of attachments from $\tilde{T}_{N,B}^{\ell}$

Data:

Color labels $L = \{L_{t_1}, ..., L_{t_N}\}$: colors of the tips $t_1, ..., t_N$ on the planar tree from left to right.

$i = 1, ..., N-1$: ranks of internal nodes from bottom to top.

Result:

$A = [\ ]$: an ordered list of the tips added.

$C = [\ ]$: an ordered list of the colors added.

$D = [\ ]$: an ordered list of the tips that have been attached to.

Integer S: the number of same attachments.

Array $\{w_k : k = 3, ..., N\}$.

Initialize $S = 0$ and $t_k = 0$ for $k = 1, ..., N$.

for $i = 1, ..., N-1$ **do**

    Compute $l_i \in \{1, ..., N\}$: the right-most leaf of the left subtree of internal node i.

    Compute $r_i \in \{1, ..., N\}$: the right-most leaf of the right subtree of internal node i

    Set $t_{l_i} = N + 1 - i$ and $C_{N+1-i} = L_{l_i}$.

    Set $A_i = t_{l_i}$: this is the tip being added.

    Set $D_i = t_{r_i}$: this is the tip being attached to.

    if $L_{l_i} = L_{r_i}$ **then**

        $S \leftarrow S + 1$

    end

end

Compute $w_k = \sum_{i=1}^{k-1} \mathbb{1}(C_i = C_k)$ for each $k = 3, ..., N$.

9.3 Conditions for a list of tables to be valid

Let $\{E^1, E^2, ..., E^T\}$ be the list of tables under consideration. To check it corresponds to a ranked planar partially labeled tree with N tips, it must satisfy the following conditions. As per standard terminology, let $|E^t|$ denote the cardinality of E^t, which is the length of the table t.

1. The total number of lists T satisfies $1 \leq T \leq N$.

2. The total length of the lists $|E| = \sum_{t=1}^{T} |E^t|$ is either $|E| = N + T - 1$ or $|E| = N + T - 2$.

3. Each element $\{1, ..., N\}$ must appear at least once in $\{E^1, E^2, ..., E^T\}$.

4. There are no repeat elements in E^t for all $t = 1, ..., T$.

5. If E^t has length 1, then E^t can be only (1) or (2).

6. For $E^t = (E_1^t, ..., E_n^t)$ where $n \geq 2$: for all $i = 1, ..., n-1$, there exists j such that $i < j$ and $E_i^t > E_j^t$. That is, there exists a number smaller than E_i^t to the right of E_i^t. For example, if $|E^t| = 2$, then $E_1^t > E_2^t$.

7. The element 1 cannot appear to the left of any element in any list.

8. The element 2 can appear to the left of 1 only if $E^t = (1)$ does not exist.

9. Elements $3, ..., N$ appear to the left of some elements in exactly one list.

9.4   Details of tree topology simulation

We give the details of the two simulations conducted in Section 4.4. To simulate the distribution of cherries and pitchforks, we simulated 200 trees per (N, B, α) combination with $N \in \{4, 5, 6, 7, 8, 9, 10, 20, 50, 100, 200\}$, all values of $B \in [0, N/2]$, $\alpha \in \{1, 6, 11, 16, 21, 50, 200\}$. In Figure 17(a), we see the number of cherries/N is concentrated around $1/3$. In Figure 17(b), the number of pitchforks/N is concentrated around $1/6$. For the second simulation, we generated 500 trees per (N, B, α) combination with $\alpha \in \{1, 5, 25\}$ and $N = 10, B = 4$ or $N = 100, B = 10$. Then for each N, we apply MDS using distances on the F-matrices of these trees Kim et al. (2020). In Figure 18, we see that there is no clustering by alpha for either $N = 10$ or $N = 100$.

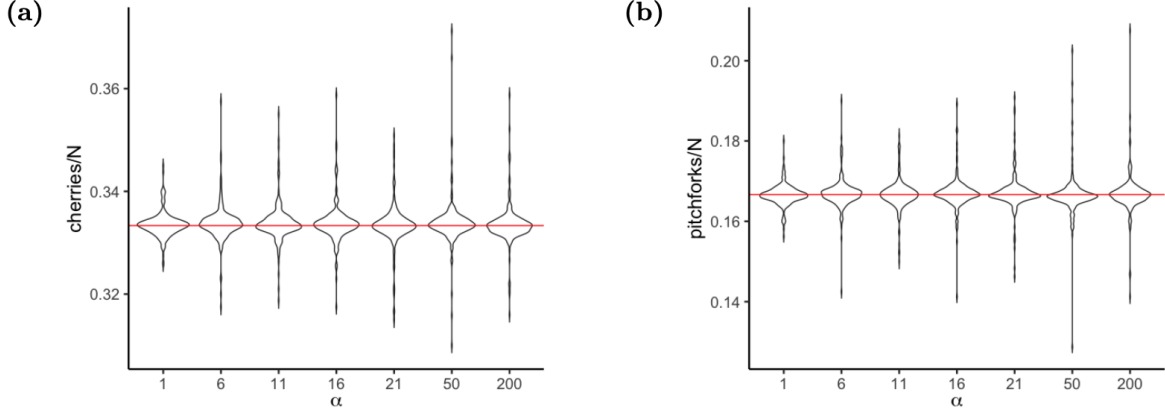Figure 17: **Violin plots of the simulated number of cherries and pitchforks.** For each α, the values are concentrated around the theoretical value. In (a), the red line is $1/3$, which is the expected value of # cherries/N. In (b), the red line is $1/6$, which is the expected value of # pitchforks/N.
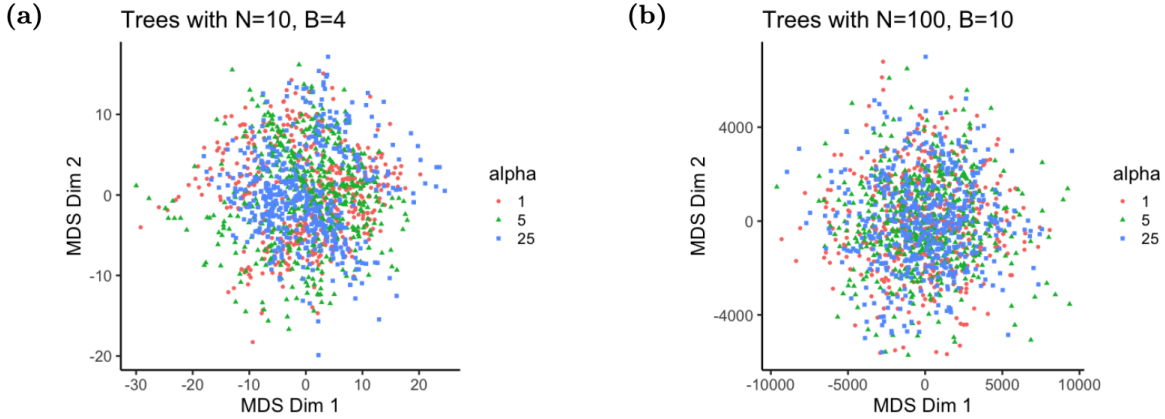

Figure 18: Plots of the first 2 MDS axes for (a) $N = 10, B = 4$ and (b) $N = 100, B = 10$

9.5 Validity of p-values

Lemma 11. *If $X_n \xrightarrow{p} X$ and c is a fixed constant, then $\mathbb{1}(X_n \geq c) \xrightarrow{p} \mathbb{1}(X \geq c)$.*

Proof. We only need to consider $\epsilon \in [0, 1]$ because we are working with indicators.

$$
\begin{aligned}
\mathbb{P}(|\mathbb{1}(X_n \geq c) - \mathbb{1}(X \geq c)| \geq \epsilon) &= \mathbb{P}\Big(\mathbb{1}(\{X_n \geq c > X\} \cup \{X \geq c > X_n\}) \geq \epsilon\Big) \\
&= \mathbb{P}\Big(\{X_n \geq c > X\} \cup \{X \geq c > X_n\}\Big) \\
&\leq \mathbb{P}(X_n - X \geq \epsilon_c) + \mathbb{P}(X - X_n \geq \epsilon_c) \text{ for some } \epsilon_c > 0 \\
&\to 0
\end{aligned}
$$

Therefore, we have convergence in probability. $\qquad\square$

9.6 Power Analyses on Specific Trees

The following figures show the power approximations calculated on five fixed ranked tree shapes detailed in Section 6.1.

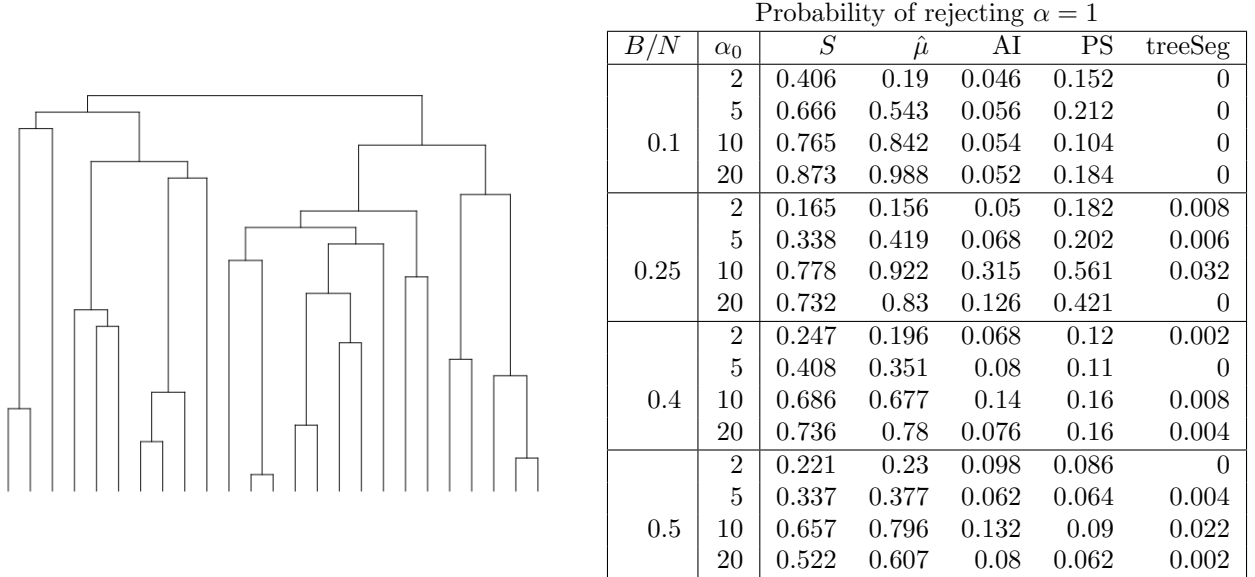|		Probability of rejecting $\alpha = 1$				
B/N	α_0	S	$\hat{\mu}$	AI	PS	treeSeg
	2	0.406	0.19	0.046	0.152	0
	5	0.666	0.543	0.056	0.212	0
0.1	10	0.765	0.842	0.054	0.104	0
	20	0.873	0.988	0.052	0.184	0
	2	0.165	0.156	0.05	0.182	0.008
	5	0.338	0.419	0.068	0.202	0.006
0.25	10	0.778	0.922	0.315	0.561	0.032
	20	0.732	0.83	0.126	0.421	0
	2	0.247	0.196	0.068	0.12	0.002
	5	0.408	0.351	0.08	0.11	0
0.4	10	0.686	0.677	0.14	0.16	0.008
	20	0.736	0.78	0.076	0.16	0.004
	2	0.221	0.23	0.098	0.086	0
	5	0.337	0.377	0.062	0.064	0.004
0.5	10	0.657	0.796	0.132	0.09	0.022
	20	0.522	0.607	0.08	0.062	0.002

Figure 19: Power Calculation for testing $H_0 : \alpha = 1$ vs $H_1 : \alpha = \alpha_0$ for the random tree shape with $N = 25$ and various values of B.

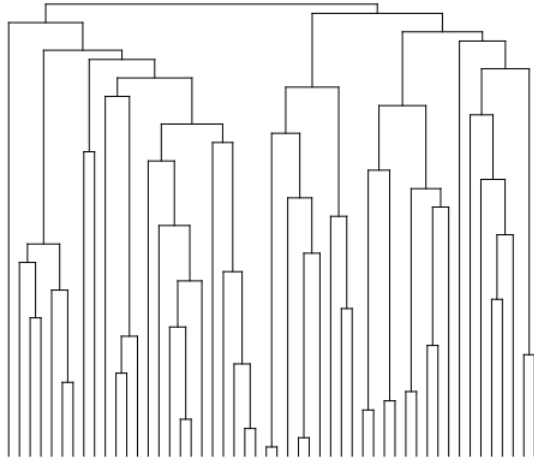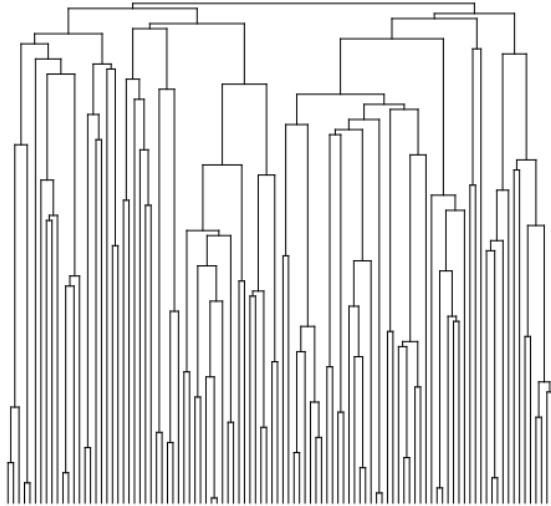|		Probability of rejecting $\alpha = 1$				
B/N	α_0	S	$\hat{\mu}$	AI	PS	treeSeg
	2	0.201	0.046	0.03	0.124	0
	5	0.516	0.379	0.084	0.17	0
0.1	10	0.736	0.858	0.11	0.291	0
	20	0.818	0.952	0.082	0.186	0
	2	0.2	0.226	0.04	0.164	0.002
	5	0.7	0.91	0.132	0.335	0
0.25	10	0.829	0.994	0.156	0.279	0.004
	20	0.929	1	0.152	0.307	0.018
	2	0.281	0.285	0.076	0.098	0
	5	0.635	0.735	0.076	0.098	0
0.4	10	0.957	0.998	0.267	0.266	0.006
	20	0.917	0.982	0.094	0.142	0.006
	2	0.403	0.379	0.124	0.162	0
	5	0.803	0.874	0.098	0.226	0.004
0.5	10	0.749	0.782	0.044	0.108	0
	20	0.945	0.972	0.096	0.225	0.014

Figure 20: Power Calculation for testing $H_0 : \alpha = 1$ vs $H_1 : \alpha = \alpha_0$ for the random tree shape with $N = 50$ and various values of B.
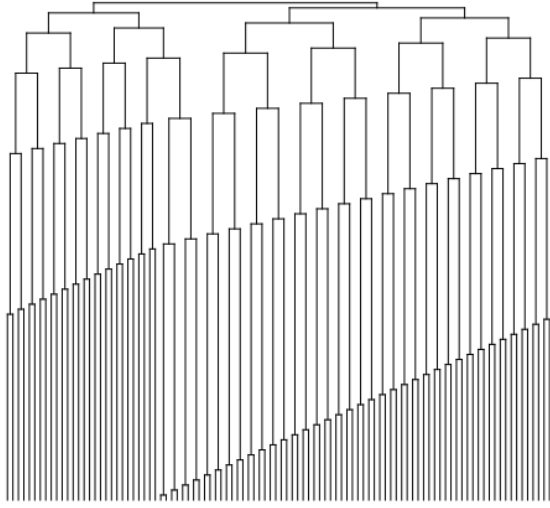

		Probability of rejecting $\alpha = 1$				
B/N	α_0	S	$\hat{\mu}$	AI	PS	treeSeg
	2	0.315	0.391	0.074	0.451	0
	5	0.755	0.978	0.066	0.477	0.002
0.1	10	0.862	1	0.066	0.419	0.002
	20	0.955	1	0.144	0.405	0
	2	0.518	0.661	0.058	0.214	0.002
	5	0.91	1	0.07	0.188	0
0.25	10	0.953	1	0.092	0.204	0.004
	20	0.964	1	0.08	0.242	0
	2	0.531	0.577	0.098	0.088	0.008
	5	0.983	1	0.202	0.228	0.004
0.4	10	0.931	0.998	0.06	0.09	0.006
	20	0.937	1	0.098	0.11	0.001
	2	0.53	0.635	0.082	0.134	0
	5	0.907	0.99	0.096	0.202	0
0.5	10	0.929	0.992	0.094	0.174	0
	20	0.93	0.998	0.072	0.14	0.002

Figure 21: Power Calculation for testing $H_0 : \alpha = 1$ vs $H_1 : \alpha = \alpha_0$ for the random tree shape with $N = 100$ and various values of B.
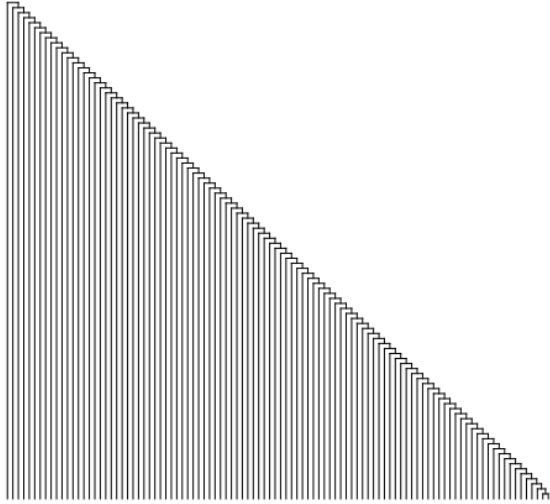
		Probability of rejecting $\alpha = 1$				
B/N	α_0	S	$\hat{\mu}$	AI	PS	treeSeg
	2	0.324	0.615	0.176	0.18	0
	5	0.731	1	0.106	0.106	0
0.1	10	0.884	1	0.108	0.108	0
	20	0.897	1	0.126	0.126	0.002
	2	0.322	0.491	0.082	0.056	0.002
	5	0.804	0.982	0.1	0.068	0.002
0.25	10	0.97	1	0.222	0.17	0
	20	0.918	1	0.1	0.084	0.008
	2	0.321	0.445	0.058	0.144	0.006
	5	0.941	1	0.23	0.321	0
0.4	10	0.936	1	0.112	0.255	0.018
	20	0.951	1	0.13	0.291	0.008
	2	0.398	0.531	0.036	0.056	0.006
	5	0.878	0.99	0.064	0.092	0.014
0.5	10	0.959	1	0.116	0.14	0
	20	0.893	0.994	0.042	0.064	0.004

Figure 22: Power Calculation for testing $H_0 : \alpha = 1$ vs $H_1 : \alpha = \alpha_0$ for the most-balanced tree shape with $N = 100$ and various values of B.

		Probability of rejecting $\alpha = 1$				
B/N	α_0	S	$\hat{\mu}$	AI	PS	treeSeg
	2	0.032	0.004	0.048	0.08	0
	5	0.062	0.09	0.036	0.08	0
0.1	10	0.141	0.202	0.056	0.098	0
	20	0.277	0.331	0.088	0.066	0.002
	2	0.079	0.016	0.072	0.052	0.002
	5	0.308	0.156	0.084	0.038	0
0.25	10	0.495	0.327	0.08	0.038	0
	20	0.584	0.421	0.062	0.058	0.006
	2	0.502	0.439	0.1	0.084	0.004
	5	0.891	0.926	0.108	0.086	0
0.4	10	0.944	0.978	0.094	0.068	0.008
	20	0.951	0.992	0.078	0.078	0.004
	2	0.958	0.984	0.098	0.064	0.002
	5	0.995	1	0.076	0.062	0.012
0.5	10	0.995	1	0.068	0.044	0
	20	0.996	1	0.08	0.044	0.002

Figure 23: Power Calculation for testing $H_0 : \alpha = 1$ vs $H_1 : \alpha = \alpha_0$ for the most-unbalanced tree shape with $N = 100$ and various values of B.

9.7 DNA data simulation to compare BaTS and Posterior p-values

We present the details of the simulation described in Section 6.2. We simulated two partially labeled ranked tree shapes with $N = 50$ and $B = 20$ from the CRPTree model with $\alpha = 1$ and $\alpha = 10$ respectively. We used the R-package phylodyn to simulate the branch lengths of the phylogenies according to the coalescent with exponentially growing effective population size (Karcher et al., 2017). We then used seqgen to simulate the 50 molecular sequences of 100 nucleotides at the tips of each phylogeny according to the Jukes Cantor mutation model (Jukes and Cantor, 1969; Rambaut and Grass, 1997).

To estimate the two posterior phylogenetic distributions we used BEAST assuming the Jukes-Cantor mutation model with fixed mutation rate, a coalescent prior on the phylogenies, and a Gaussian Markov random field prior on $N_e(t)$ (Minin et al., 2008). We generated 100 billion iterations and thinned every 100 thousand iterations to obtain a posterior sample of 1000 ranked and partially labeled trees.

To compute the 1000 p-values with our method for each analysis, we used 500 label and planar permutations per tree. For BaTS analyses, we generated 500 samples by permuting the labels and compared the distribution of the posterior median statistics with the observed median statistic. The posterior distribution of p-values p_T is shown in Figure 11. For $\alpha = 1$, the median value of $\hat{\mu} = 20.31$ with a 95% credible interval of $[19, 21.838]$. Using BaTS the 95% credible interval for the posterior median is $[24.042, 24.582]$. For $\alpha = 10$, the median value of $\hat{\mu} = 39.36$ with a 95% credible interval of $[39, 39.92]$. For BaTS, the 95% credible interval for the posterior median is $[23.17, 25.54]$. It is clear that our method correctly rejects the case where $\alpha = 10$, and fails to reject the case of $\alpha = 1$, while BaTS would reject in both cases.