*Biometrika* (2024), **111**, 1, *pp.* 171–193 https://doi.org/10.1093/biomet/asad025
Advance Access publication 26 April 2023

# Statistical summaries of unlabelled evolutionary trees

By RAJANALA SAMYAK and JULIA A. PALACIOS

*Department of Statistics, Stanford University,*
*390 Jane Stanford Way, Stanford, California 94305, U.S.A.*
samyak@stanford.edu juliapr@stanford.edu

## Summary

Rooted and ranked phylogenetic trees are mathematical objects that are useful in modelling hierarchical data and evolutionary relationships with applications to many fields such as evolutionary biology and genetic epidemiology. Bayesian phylogenetic inference usually explores the posterior distribution of trees via Markov chain Monte Carlo methods. However, assessing uncertainty and summarizing distributions remains challenging for these types of structures. While labelled phylogenetic trees have been extensively studied, relatively less literature exists for unlabelled trees that are increasingly useful, for example when one seeks to summarize samples of trees obtained with different methods, or from different samples and environments, and wishes to assess the stability and generalizability of these summaries. In our paper, we exploit recently proposed distance metrics of unlabelled ranked binary trees and unlabelled ranked genealogies, or trees equipped with branch lengths, to define the Fréchet mean, variance and interquartile sets as summaries of these tree distributions. We provide an efficient combinatorial optimization algorithm for computing the Fréchet mean of a sample or of distributions on unlabelled ranked tree shapes and unlabelled ranked genealogies. We show the applicability of our summary statistics for studying popular tree distributions and for comparing the SARS-CoV-2 evolutionary trees across different locations during the COVID-19 epidemic in 2020. Our current implementations are publicly available at `https://github.com/RSamyak/fmatrix`.

*Some key words*: Binary tree; Combinatorial optimization; Evolutionary tree; Fréchet mean; Summarizing tree; Unlabelled tree.

## 1. Introduction

Phylogenetic trees are used to represent the ancestral relationships of individuals from a sample of molecular sequences or phenotypic traits, from a population. These individuals can be viral sequences from infected hosts as in viral phylodynamics, species as in phylogenetics, individuals from a single species as in population genetics or cells such as in cancer evolution. The estimated tree is of interest because it provides information about whether genes are under selection (Yang et al., 2018), and about the past evolutionary dynamics of the sample's population. For example, in the context of viral phylodynamics, the tree provides information about the past transmission history and pathogenesis (Volz et al., 2013).

Distance-based summaries of labelled tree structures have been extensively studied in the last few decades (Hillis et al., 2005; Chakerian & Holmes, 2012; Benner & Bačák, 2014; Willis & Bell, 2018; Brown & Owen, 2020). These summaries rely on metrics of labelled

trees such as the Billera–Holmes–Vogtmann or BHV distance (Billera et al., 2001) and it is usually assumed that all sampled trees have the same set of leaves, i.e., labels. Kuhner & Yamato (2014) presented a comparison of different distances on such space.

In this article, we are interested in summarizing different tree structures, namely those that are ranked and unlabelled. These trees are useful in the study of the ancestral relationships of a sample of objects that are exchangeable. One potential area of application is in the study of cancer evolution where the tree represents the evolutionary history of many cells in a patient's tumour. We may want to summarize many such trees, each tree inferred from a different patient, in order to find a representative tree and to quantify how much variation or heterogeneity is present across patients with the same type of cancer or across different types of cancer. Similar types of questions have been studied assuming a coarser type of tree structure (Govek et al., 2018).

While summarizing real-valued parameters is straightforward, summarizing a sample of unlabelled discrete structures is more challenging. Recently proposed distance metrics on the space of unlabelled ranked evolutionary trees enable quantitative comparisons of evolutionary trees of different sets of organisms living at different geographic locations and different time periods. These metrics have been used for visual comparison of empirical posterior distributions via multi-dimensional scaling, for summarizing empirical distributions via medoids, and for a two-sample permutation test of equality in distribution (Kim et al., 2020). However, the calculation of medoids and the permutation test are computationally expensive and their statistical properties are unknown. At present, we are not aware of any other method applicable for summarizing distributions of unlabelled ranked trees.

In this article, we use the previously defined distance metrics on unlabelled ranked evolutionary trees to understand distributional properties of some popular tree models and to summarize samples of trees. Tree samples can either be obtained from the posterior distribution for a given sample of molecular sequences such as those obtained with BEAST (Suchard et al., 2018), or trees independently obtained in different studies. To summarize samples and populations of unlabelled ranked evolutionary trees, we define the sample and population Fréchet means, variances and interquartile sets, in terms of the recently proposed distances. We compare these summaries to other measures of centrality and dispersion on the same space. In particular, the fact that the Fréchet mean is not restricted to the sample provides a clear advantage over the medoid, for example when summarizing theoretical distributions.

One of the main contributions of the present work is to establish the bijection between the space of unlabelled ranked tree shapes and the space of triangular matrices of integer values that satisfy a set of linear constraints. This allows us to formulate the problem of finding the Fréchet mean as an integer programming problem and to better understand the properties of the metric space. We show that the computational complexity of the Fréchet mean is independent of $N$, the number of trees, and instead depends only on $n$, the number of leaves of the trees. For large $n$, we rely on stochastic combinatorial optimization and propose a simulated annealing algorithm for estimating the Fréchet mean. A key aspect of the algorithm is the definition of a novel Markov chain that efficiently explores the space. Our algorithms are applicable to the discrete space of tree topologies only, as in Fig. 1(a), and to the mixed space of tree topologies and branch lengths, as in Fig. 1(b). They are also applicable to trees whose samples are all obtained at the same time-point and to trees with time-stamped leaves, as in Fig. 1(c).

We apply our methods to the standard Kingman coalescent and the Blum–François family of tree distributions. As a real data example, we analyse the posterior distributions of

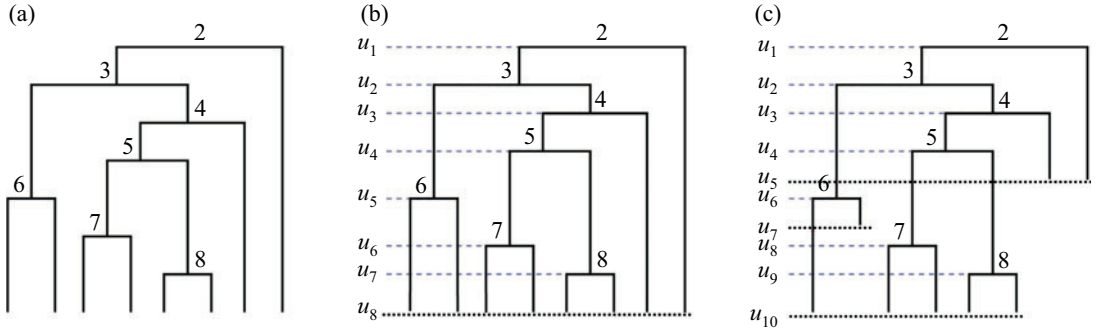


Fig. 1. Tree examples: (a) isochronous ranked tree shape, or topology only; (b) isochronous ranked genealogies, or ranked tree shape with branching times; and (c) heterochronous ranked genealogy with different sampling times (dotted lines).

evolutionary trees inferred from SARS-CoV-2 molecular sequences from the states of California, Texas, Florida and Washington. We obtain Fréchet mean trees for different samples and display multi-dimensional scaling plots to visualize intrastate and interstate variability.

We have developed an `R` package, `fmatrix`, available at https://github.com/RSamyak/fmatrix, which implements the various methods discussed in this paper. The package is compatible with `phylodyn` (Karcher et al., 2017), an `R` package for phylodynamic simulation and inference, and `ape` (Paradis & Schliep, 2019), an `R` package to handle phylogenetic trees (R Development Core Team, 2024).

All omitted proofs are given in the Supplementary Material.

## 2. Preliminaries

Ranked unlabelled trees or ranked tree shapes are rooted binary trees, with an increasing ordering of the interior nodes, as in Fig. 1(a). They are unlabelled in the sense that the external nodes, i.e., leaves, are unlabelled. However, we rank the internal nodes, starting at the root with label 2. A ranked unlabelled tree, additionally equipped with the vector of branching times is called a ranked genealogy; see Fig. 1(b). We use the two terms coalescent times and branching event times interchangeably. We call a tree isochronous if all the leaves are sampled at the same time, usually assumed to be sampled at time 0, such as in Fig. 1(a)–(b). In applications of rapidly evolving pathogens such as the influenza A virus, molecular sequences, which are the leaves of the tree, are sampled at different times, as in Fig. 1(c), and these trees are called heterochronous.

Kim et al. (2020) showed that an isochronous ranked tree shape with $n$ tips can be uniquely encoded as a triangular matrix of integers, called an $\mathbb{F}$-matrix, as follows. Let $u_{i-1}$ denote the time of the branching event at node $i$, let $I_i = (u_{i-1}, u_i)$ denote the time interval between the two subsequent nodes $i$ and $i+1$, and let $u_n = 0$ at the leaves. The corresponding $\mathbb{F}$-matrix is an $(n-1) \times (n-1)$ triangular matrix of nonnegative integers. The diagonal elements of the $\mathbb{F}$-matrix indicate the number of branches at each time interval. The off-diagonal element $F_{i,j}$, $2 \leqslant j < i \leqslant n-1$, represents the number of branches extant at $I_j = (u_{j+1}, u_j)$ that do not bifurcate during the interval $(u_{i+1}, u_j)$. Figure 2 shows all ranked tree shapes with five leaves (first row) and their corresponding $\mathbb{F}$-matrix encodings (second row). Kim et al. (2020) constructed an injective map from the space of trees to the space of $\mathbb{F}$-matrices. The following theorem establishes the bijective relationship and defines the space of $\mathbb{F}$-matrices as the space of triangular matrices of integers subject to a set of specific linear constraints.

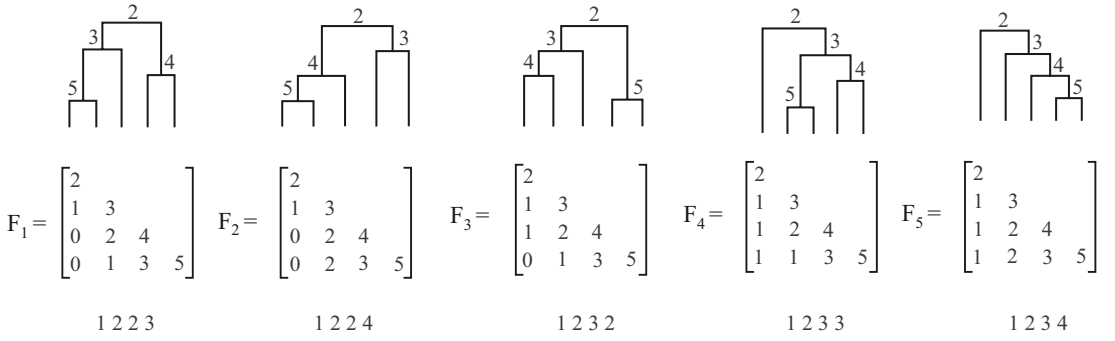


Fig. 2. All ranked tree shapes with $n = 5$ leaves. The second row shows the corresponding $\mathbb{F}$-matrices representation of the ranked tree shape of the first row, and the third row shows their corresponding functional code representations; see the Supplementary Material.

THEOREM 1. *The space of ranked tree shapes with n leaves $\mathcal{T}_n$ is in bijection with the space $\mathcal{F}_n$ of $(n-1) \times (n-1)$ $\mathbb{F}$-matrices, which are lower triangular square matrices of nonnegative integers that obey the following constraints.*

(i) *The diagonal elements are $F_{i,i} = i+1$ for $i = 1, \ldots, n-1$ and the subdiagonal elements are $F_{i+1,i} = i$ for $i = 1, \ldots, n-2$.*
(ii) *The elements $F_{i,1}$, $i = 3, \ldots, n-1$, in the first column satisfy* $\max\{0, F_{i-1,1}-1\} \leqslant F_{i,1} \leqslant F_{i-1,1}$.
(iii) *All other elements $F_{i,k}$, $i = 4, \ldots, n-1$ and $k = 2, \ldots, i-2$, satisfy the inequalities*

$$\max\{0, F_{i,k-1}\} \leqslant F_{i,k},$$
$$F_{i-1,k} - 1 \leqslant F_{i,k} \leqslant F_{i-1,k},$$
$$F_{i,k-1} + F_{i-1,k} - F_{i-1,k-1} - 1 \leqslant F_{i,k} \leqslant F_{i,k-1} + F_{i-1,k} - F_{i-1,k-1}.$$

The definition of the $\mathbb{F}$-matrix in Theorem 1 allows us to enumerate in a constructive way all elements in the space. For example, for $n = 5$, the first two diagonals of the $\mathbb{F}$-matrices are fixed, see Theorem 1(i), and the rest of the elements satisfy $F_{3,1} \in \{0, 1\}$, $\max\{0, F_{3,1} - 1\} \leqslant F_{4,1} \leqslant F_{3,1}$ and $\max\{0, F_{4,1}, F_{4,1} - F_{3,1} + 1\} \leqslant F_{4,2} \leqslant \min\{2, F_{4,1} - F_{3,1} + 2\}$. The second row of Fig. 2 shows all possible $\mathbb{F}$-matrices for $n = 5$ following these constraints.

Using the $\mathbb{F}$-matrix encoding, Kim et al. (2020) proposed to compute the distance between two trees as the distance between two matrices. In particular, we use the distance

$$d(y_1, y_2) := d(F_1, F_2) = \left[ \sum_{i,j} \{(F_1)_{ij} - (F_2)_{ij}\}^2 \right]^{1/2} \tag{1}$$

for $y_1$ and $y_2 \in \mathcal{T}_n$, with corresponding $\mathbb{F}$-matrices $F_1$ and $F_2$, and the distance on ranked genealogies

$$d(G_1, G_2) := \left[ \sum_{i,j} \{(F_1)_{ij}(W_1)_{ij} - (F_2)_{ij}(W_2)_{ij}\}^2 \right]^{1/2} \tag{2}$$

for $G_1$ and $G_2 \in \mathcal{G}_n$, with corresponding $\mathbb{F}$-matrices $F_1$ and $F_2$, where $W_1$ and $W_2$ are weight matrices constructed using the respective branching event times $u_k = (u_{k,n}, u_{k,n-1}, \ldots, u_{k,1})$ of the $k$th tree, $k = 1, 2$, with $(W_k)_{ij} := |u_{k,j} - u_{k,i+1}|$; see Fig. 1(b).

For defining a distance on the space of heterochronous ranked tree shapes or genealogies, see Fig. 1(c), Kim et al. (2020) proposed supplementing the $\mathbb{F}$-matrix with additional rows for sampling events. The heterochronous distances are then computed analogously to the isochronous distances, as the Euclidean distances between the extended $\mathbb{F}$-matrices of the same size, as detailed in § 3 of the Supplementary Material of Kim et al. (2020). For improving computational efficiency when computing pairwise distances among a large number of trees, we modify the distance slightly and consider all trees together when adding additional rows to the $\mathbb{F}$-matrix. Further details can be found in the Supplementary Material.

## 3. Central summaries

### 3.1. *Set-up*

Let $y_1, \ldots, y_m \in \mathcal{T}_n$ be $m$ ranked tree shapes with $n$ leaves independently drawn from a common probability distribution. We are interested in summarizing such samples and identifying a representative tree of the sample. Similarly, given a probability distribution over the space of ranked tree shapes, we are interested in knowing what is the expected tree of that distribution.

Current practices for summarizing labelled trees include reporting a majority-rule consensus tree, a maximum clade credibility tree and a median tree based on metrics on labelled trees (Benner & Bačák, 2014; Brown & Owen, 2020). The majority-rule consensus tree is obtained by choosing partitions with probability greater than 0.5 from the list of observed partitions and it is usually annotated with marginal probabilities of each partition as a measure of uncertainty (Cranston & Rannala, 2007). The maximum clade credibility tree is the tree with the maximum product of clade probabilities and it is arguably the most used central summary of labelled trees.

The concept of consensus partition is not applicable for ranked tree shapes since they do not have labels. Instead, we can rely on the proposed distance in the same line as the median tree for labelled trees. In fact, we can use the distance to define a loss function and approach the problem of finding a representative tree as the one that minimizes the expected squared loss. This decision theoretic solution corresponds to finding the Fréchet mean. The Fréchet mean is the tree in the space that has the minimum average or expected squared distance to the sample and, hence, it provides a natural notion of a central tree. We extend this notion to ranked genealogies, including heterochronous genealogies.

Kim et al. (2020) used the same distance function to find the medoid, that is, the set of trees in the sample that has the minimum distance to all trees in the sample. The Fréchet mean instead is not restricted to be in the sample. For population distributions of trees, the medoid and the Fréchet mean are equivalent. We discuss their differences further in this section and in § 6.

### 3.2. *The Fréchet mean*

We first consider the metric spaces $(\mathcal{T}_n, d)$, where $d$ is given in (1), and let $\mu$ denote a finite probability mass function on $\mathcal{T}_n$; then the barycentre of $\mu$, also called the Fréchet mean tree (Fréchet, 1948), is any element $\bar{T} \in \mathcal{T}_n$ such that

$$\bar{T} \in \arg\min_{x \in \mathcal{T}_n} \sum_{y \in \mathcal{T}_n} d(x, y)^2 \mu(y). \tag{3}$$

Since $\mathcal{T}_n$ is finite, we immediately have the existence of the minimizer in (3). Uniqueness may not be guaranteed, as, though the objective function is convex, the space is discrete. As we will show in § 6, this is the case when $\mu$ corresponds to the coalescent model on ranked tree shapes. In this case, all Fréchet means are close to each other.

For the metric spaces $(\mathcal{G}_n, d)$ with $d$ given in (2), and $\nu$ a probability measure on $\mathcal{G}_n$ such that

$$\int_{\mathcal{G}_n} d(x,y)^2 \, \mathrm{d}\nu(y) < \infty,$$

the Fréchet mean genealogy is any element $\bar{G} \in \mathcal{G}_n$ such that

$$\bar{G} \in \operatorname*{arg\,min}_{G \in \mathcal{G}_n} \int_{H \in \mathcal{G}_n} d(G,H)^2 \, \mathrm{d}\nu(H). \tag{4}$$

In the Supplementary Material we show that in the special case when isochronous genealogies $G = (F, \mathbb{u})$ have densities of the form $\mathrm{d}\nu(G) = \mu(F) \prod_{j=1}^{n-1} f(u_j \mid u_{j+1}) d(\mathbb{u})$, that is, the tree topology and the branching event times are independent, the Fréchet mean $\bar{G} = (\bar{F}, \bar{\mathbb{u}})$ can be obtained by setting $\bar{\mathbb{u}} = \mathbb{E}[\mathbb{u}]$ and then finding the tree topology that satisfies (4). In evolutionary biology applications, this assumption corresponds to neutral evolution in a closed population (Wakeley, 2008). This is the case, for example, in the standard coalescent with variable population size (Slatkin & Hudson, 1991). This appealing computational property is not only due to the independence assumption, but is also due to the nature of the Euclidean distance, which allows for such separation.

The empirical Fréchet mean of a given sample $y_1, \ldots, y_m$ from the metric space $(\mathcal{T}_n, d)$ is obtained by taking $\mu$ in (3) to be the empirical measure. The mean of a sample $h_1, \ldots, h_m$ from $(\mathcal{G}_n, d)$ is obtained analogously, where $\nu$ in (4) is taken to be the empirical measure.

The cardinality of the space $\mathcal{T}_n$ grows superexponentially with $n$, $|\mathcal{T}|_n \sim 2(2/\pi)^{n+1} \cdot n!$, hence, finding the Fréchet mean is computationally challenging for large $n$. An alternative summary of centrality is the in-sample version of the empirical Fréchet mean, also called the medoid or restricted Fréchet mean:

$$\bar{T}^{\text{in sample}} \in \operatorname*{arg\,min}_{x \in \{y_1, \ldots, y_m\}} \sum_{j=1}^{m} d(x, y_j)^2,$$

$$\bar{G}^{\text{in sample}} \in \operatorname*{arg\,min}_{g \in \{h_1, \ldots, h_m\}} \sum_{j=1}^{m} d(g, h_j)^2.$$

This may be reasonable for spaces with a large number of leaves when direct computation of the Fréchet mean is not possible. However, constraining ourselves to stay only within the sample may be undesirable; see § 6 for further discussion.

### 3.3. *Mixed integer programming*

The Fréchet mean defined in (3) is the minimizer of a simple quadratic objective function subject to linear constraints over integer variables; hence, this problem can be framed as a mixed integer programming problem. Moreover, in the case when $\mu$ is a probability measure

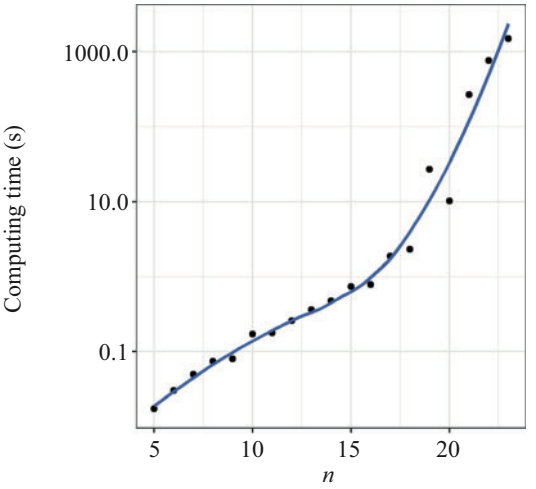


Fig. 3. Running time. Exact computation for the Fréchet mean under the Yule model using `gurobi`, plotted against the dimension of the $\mathbb{F}$-matrices. Computations were performed on a laptop with an Intel® i7 processor.

on $\mathcal{T}_n$, or, equivalently, on $\mathcal{F}_n$, then the Fréchet mean $\bar{F}$ is given by

$$\begin{aligned}\bar{F} &\in \operatorname*{arg\,min}_{F\in\mathcal{F}_n} \sum_{H\in\mathcal{F}_n}\sum_{k,l}(F_{kl}-H_{kl})^2\mu(H)\\ &= \operatorname*{arg\,min}_{F\in\mathcal{F}_n} \sum_{H\in\mathcal{F}_n}\sum_{k,l}(F_{kl}^2-2F_{kl}H_{kl})\mu(H)\\ &= \operatorname*{arg\,min}_{F\in\mathcal{F}_n} \sum_{k,l}(F_{kl}^2-2F_{kl}M_{kl}),\end{aligned}$$

where $M_{kl}=\sum_{H\in\mathcal{F}_n}H_{kl}\cdot\mu(H)$. That is, once the means $M_{kl}$ are computed, the rest of the problem no longer involves the $m$ samples. That is, the problem scales with the number of leaves $n$, but not with the number of samples $m$. This is particularly important when summarizing samples obtained through Markov chain Monte Carlo, since $m$ is usually of high order. We use `gurobi` (Gurobi Optimization, 2020), a standard mixed integer programming solver, to directly perform the optimization. The implementation of the code is available in the `R` package `fmatrix` (R Development Core Team, 2024).

This method works well for a small number of leaves $n$, such as $n=20$, but it quickly becomes impractical with larger $n$. Figure 3 shows how the computation time grows exponentially in $n$. For larger $n$, we resort to stochastic combinatorial optimization algorithms that scale well at the expense of solution guarantees, as discussed in § 3.4.

### 3.4. *Simulated annealing algorithm*

When the number of leaves is large, the mixed integer programming solution is computationally demanding and often unfeasible. A simple technique that works well is simulated annealing (Kirkpatrick et al., 1983), which is a general-purpose stochastic algorithm for optimizing an objective function over a potentially large discrete set. Simulated annealing explores the ranked tree shape space via a Metropolis–Hastings algorithm. We trade the guarantee of an exact solution for computational tractability.

In order to implement the simulated annealing algorithm, we define two Markov chains on the space of ranked tree shapes, one for isochronous and one for heterochronous trees. The details of the Markov chains can be found in § 3.5. The two Markov chains are then used as proposal distributions in the Metropolis–Hastings step of the simulated annealing algorithm.

In the case of the Fréchet mean, simulated annealing aims to minimize the energy function $E(x) = \sum_{i=1}^{m} d(x, y_i)^2$ for a sample of trees $\{y_i\}_{i=1}^{m}$ over $x \in \mathcal{T}_n$. This problem is equivalent to finding the maximum of $\exp\{-E(x)/R\}$ at any given temperature $R > 0$. We then define a sequence of monotone decreasing temperatures $\{R_k\}$ such that $\lim_{k\to\infty} R_k = 0$. For example, $R_k = \alpha^k R_0$ for some high initial temperature $R_0$, and $\alpha < 1$. This is called the exponential cooling schedule. Then, at each temperature, the simulated annealing algorithm consists of Metropolis–Hastings steps that target $\pi_k(x) \propto \exp\{-E(x)/R_k\}$ as the stationary distribution. As the number of steps increases, $\pi_k(x)$ puts more and more of its probability mass in the set of global maxima. Simulated annealing differs from descent algorithms by allowing transitions to higher-energy states at higher temperatures, in order to avoid being stuck at local maxima.

In our implementations, Algorithm 1 below, for isochronous and heterochronous Fréchet means, the Metropolis–Hastings transition kernels are symmetric and, hence, the Metropolis–Hastings acceptance probability of moving from $x_{k-1}$ to $x_k$ is given by

$$a_k = \exp\left\{ -\frac{E(x_k)}{R_k} + \frac{E(x_{k-1})}{R_{k-1}} \right\} \wedge 1.$$

*Algorithm* 1. Fréchet mean of a sample of ranked unlabelled trees via simulated annealing.

Require: $T_1, \ldots, T_m$ sample of ranked unlabelled trees or, equivalently, $M = (1/m)\sum_{i=1}^{m} T_i$, starting position $T^{(0)}$, initial temperature $R_0 > 0$, decay parameter $\alpha \in (0, 1)$.
Define the energy function $E(T) = \sum_{i=1}^{m} d(T, T_i)^2$; with $d$ a metric defined in § 2.
$k \leftarrow 0$
repeat
  $S \leftarrow$ random neighbour of $T^{(k)}$. Generate the proposal using Definition 2 or 4 below for isochronous and heterochronous trees, respectively.
  if `runif(1)` $< \exp[-\{E(S) - E(T^{(k)})\}/R_k]$ then
    $T^{(k+1)} \leftarrow S$ (accept)
  else
    $T^{(k+1)} \leftarrow T^{(k)}$ (reject)
  end if
  $R_{k+1} \leftarrow \alpha R_k$ (reduce temperature)
  $k \leftarrow k + 1$
until convergence of $T^{(k)}$

The temperature schedule in simulated annealing needs to be specified and affects the time taken for convergence of the algorithm. Theoretical convergence guarantees exist for the logarithmic cooling schedule $R_k = R_0\{1 + \alpha \log(1 + k)\}^{-1}$ with sufficiently high initial temperature and appropriately chosen $\alpha$; see Chapter 3 of Aarts & Korst (1988). However, this schedule is prohibitively slow for most problems. In practice, we observe that the exponential cooling schedule with $\alpha$ chosen very close to 1 performs reasonably well; see the

Supplementary Material. The benefits of simulated annealing are its easy implementation and the design of the algorithm that allows getting out of local optima.

Using the result shown in § 3.3, we can replace the energy function $E(T) = \sum_{i=1}^{m} d(T, T_i)^2$ by $E(T) = \|F - M\|^2$, where $M$ is the Euclidean mean $\mathbb{F}$-matrix.

For both isochronous and heterochronous genealogies, and motivated by the result in § C of the Supplementary Material, our algorithm first finds the average branching, or coalescent, event times and then finds the tree topology via Algorithm 1. Calculating the average branching event times is not trivial in the case of heterochronous genealogies. We rely on augmenting the $F$ and $W$ matrices, as explained in the Supplementary Material. In the case of heterochronous genealogies, the Markov chain used is conditioned on a fixed set of coalescent and sampling times. We analyse the computational performance of the simulated annealing algorithm in the Supplementary Material.

### 3.5. *A Markov chain on the space of ranked tree shapes*

In this section, we describe the Markov chains used in our simulated annealing algorithm. We now drop the $\mathbb{F}$-matrix representation of ranked tree shapes and instead use two string representations of the spaces of isochronous and heterochronous ranked tree shapes. We use the string representations to define two Markov chains on the corresponding spaces.

An isochronous ranked tree shape is encoded as a string of $n-1$ integers $t = (t_1, t_2, \ldots, t_{n-1})$, where $t_k$ indicates the parent node of the internal node with ranking $k+1$, $k \in \{1, \ldots, n-1\}$. It is assumed that the first integer $t_1$ of the string representation is 1, for the parent of the root node. Figure 2 shows the string encodings of each of the five ranked tree shapes at the bottom. The string representation was introduced earlier as the functional code for binary increasing trees (Donaghey, 1975). The set of all string representations of $n-1$ elements is in bijection with the space of isochronous ranked tree shapes of $n$ leaves and the space of binary increasing trees of $n-1$ nodes (Stanley, 1999).

To recover the tree $T$ from the encoding $t$, we can proceed in a generative fashion: we start at the root that has label 2, and proceed by bifurcating the leaves in the order determined by $t$. The space of strings $\mathfrak{T}_n$ is the set of all $t$ strings of length $n-1$ defined as follows.

Definition 1 (Isochronous string representation). *A string $t$ of nonnegative integers that encodes an isochronous ranked tree shape has the following defining properties:*

(i) $t_1 = 1$;
(ii) *for* $i > 1$, $2 \leqslant t_i \leqslant i$;
(iii) *no entry of $t$ can appear more than twice.*

Definition 2 (Markov chain on isochronous strings). *Let $t \in \mathfrak{T}_n$ be a string encoding an isochronous ranked tree shape as described in Definition 1. We define a Markov chain on $\mathfrak{T}_n$ as follows.*

(i) *Pick an element $i \in 2, \ldots, n$ uniformly at random.*
(ii) *Pick the value of $t_i$ uniformly at random from the allowable choices in $2, \ldots, i$, i.e., from those choices that do not already appear twice among $t_{-i}$.*

An example transition of the Markov chain of Definition 2 is depicted in Fig. 4.

Proposition 1. *The Markov chain on isochronous strings, Definition 2, is ergodic with uniform stationary distribution on the space of strings of length $n-1$, or, equivalently, on the space of ranked tree shapes with $n$ leaves.*

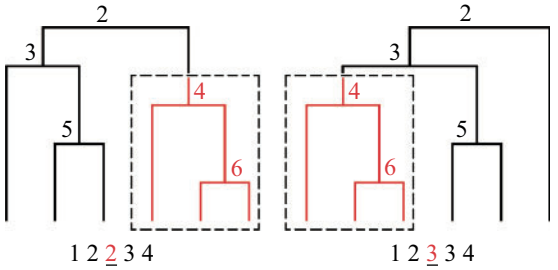


Fig. 4. An example transition under the Markov chain of Definition 2 from $(1,2,2,3,4)$ to $(1,2,3,3,4)$. The subtree of node 4 is plucked from under node 2 and planted under node 3.

*Proof.* Let $t$ be an arbitrary element of $\mathfrak{T}_n$. We show that $t$ is path connected to $t^* = (1,2,\ldots,n-1)$. This string corresponds to the most unbalanced tree, also called the caterpillar or the comb tree. Since the Markov chain is symmetric, $t^*$ is path connected to every element of $\mathfrak{T}_n$ as well and, hence, the chain is irreducible. The following path has all transitions with positive probability:

$$\begin{aligned} t = t^{(0)} &= (1, t_2, t_3, \ldots, t_{n-2}, t_{n-1}),\\ t^{(1)} &= (1, t_2, t_3, \ldots, t_{n-2}, n-1),\\ t^{(2)} &= (1, t_2, t_3, \ldots, n-2, n-1),\\ &\vdots\\ t^{(n-3)} &= (1, t_2, 3, \ldots, n-2, n-1),\\ t^{(n-2)} &= (1, 2, 3, \ldots, n-2, n-1) = t^*. \end{aligned}$$

Note that $t^{(i)} \mapsto t^{(i+1)}$ is always a valid transition due to Definition 1(iii). □

This representation can be extended to heterochronous trees as well, with additional entries indicating the sampling events. We define these strings in the following way.

DEFINITION 3 (HETEROCHRONOUS STRING REPRESENTATION). *A heterochronous ranked tree shape with $n$ leaves is encoded as a pair of strings $(t,\sigma)$, each of length $2n-1$. As before, $t$ is a string of nonnegative integers that indicates the parent nodes of internal nodes, coalescent events; however, $t$ now also includes the parent nodes of all leaves. The sequence order is given by the time they are created and $\sigma$ is a* 0-1 *string that indicates whether the corresponding node is internal (*1*) or a leaf (*0*). These strings have the following defining properties:*

(i) $t_1 = 1$, $\sigma_1 = 1$;
(ii) $|\{i\colon \sigma_i = 1\}| = n-1$;
(iii) $|\{i\colon \sigma_i = 0\}| = n$;
(iv) *each element of* $\{2,\ldots,n\}$ *occurs exactly twice in* $t$;
(v) *for each* $i > 1$, $2 \leqslant t_i \leqslant 1 + \sum_{j=1}^{i-1} \sigma_j$.

For example, the string representation of the ranked tree shape of Fig. 1(c) is ($t$ = 123442365567788, $\sigma$ = 111100101100000). The string $t_\sigma = (t_i\colon \sigma_i = 1)$ is a valid string

encoding of an isochronous ranked tree shape. In addition, this representation can also admit extensions to multifurcating trees, which we leave for future study.

DEFINITION 4 (MARKOV CHAIN ON HETEROCHRONOUS STRINGS). *Let $(t, \sigma)$ be a pair of strings encoding a heterochronous ranked tree shape as described in Definition 3. We define a Markov chain on the space of such strings, conditional on $\sigma$ a fixed sequence of sampling and coalescent events, with transitions as follows.*

(i) *Pick two distinct element $i, j \in 2, \ldots, 2n-1$ uniformly at random.*
(ii) *Swap $t_i$ and $t_j$. If the result is a valid heterochronous string, accept the move; otherwise, reject the move.*

PROPOSITION 2. *The Markov chain on heterochronous strings, Definition 4, with a given sequence of sampling and coalescence events is symmetric, aperiodic and irreducible.*

*Proof.* Let $(t, \sigma)$ be the encoding of an arbitrary heterochronous ranked tree shape.

We first define $(t^*, \sigma)$ that is the analogue of the caterpillar or the most unbalanced tree, but with the given $\sigma$. Let $t^*_\sigma = (t^*_i : \sigma_i = 1)$ be equal to $(1, 2, \ldots, n-1)$, and let $t^*_{-\sigma} = (t^*_i : \sigma_i = 0)$ be equal to $(2, 3, \ldots, n-1, n, n)$. We can follow a similar method as in the isochronous case and show that $t$ is path connected to $t^*$ by a sequence of steps with positive probability. Note that $t$ has length $2n-1$.

Let $t^{(0)} = t$. We obtain $t^{(i)}$ by swapping two terms of $t^{(i-1)}$ such that the last $i$ terms of $t^{(i)}$ and $t^*$ are equal. Explicitly, let $j$ be the largest element in $2, \ldots, 2n-i$ such that $t^{(i-1)}_j = t^*_{2n-i}$. We can swap $t^{(i-1)}_j$ and $t^{(i-1)}_{2n-i}$, since $t^*_{2n-i}$ is the maximum allowable entry at position $2n-i$, and hence $t^{(i-1)}_{2n-i} \leqslant t^{(i-1)}_j$, which satisfies Definition 3(v). The remaining conditions under Definition 3 are not affected by the transitions. Iteratively, we obtain $t^{(2n-1)} = t^*$.

Since the Markov chain is symmetric, we see that $t^*$ will be path connected to every $t$ as well, and the Markov chain is irreducible. The chain is aperiodic since we can pick a pair with the same label with positive probability. □

## 4. TOTAL ORDER IN THE SPACE OF RANKED TREE SHAPES

The histogram is another important summary of distributions. However, in order to have a meaningful comparison of histograms across distributions, we need to define a total order in the space of trees. Such an ordering roughly corresponds to a one-dimensional projection of the space. As we will show in the next section, this ordering is also useful for summarizing credible and interquartile balls.

The total order we propose is based on the distance to a reference ranked tree shape, for example, the Fréchet mean $T_K$ of the Kingman model (Kingman, 1982), which is a commonly used neutral model for evolution, together with a lexicographic order in the $\mathbb{F}$-matrix representation. We construct our ordering in such a way that the most unbalanced tree $T_{\text{unb}}$, also called the caterpillar tree, and the most balanced tree $T_{\text{bal}}$ are two poles of the order, and the Fréchet mean $T_K$ lies somewhere in between those two poles.

We have three main reasons for choosing $T_{\text{unb}}$ and $T_{\text{bal}}$ to be the two poles of the ordering. (i) The two unique extremes in the lexicographic order correspond to precisely these two trees. (ii) The maximum $d$ distance between two ranked tree shapes is uniquely achieved between these two trees. (iii) The notion of tree balance is important in phylogenetics and evolutionary biology (Mooers & Heard, 1997; Yang et al., 2018; Lemant et al., 2022). Several tree balance statistics have been proposed (Fischer et al., 2021) and used in several ways,

including the testing for natural selection and for model fit (Kirkpatrick & Slatkin, 1993; Yang et al., 2018).

PROPOSITION 3. *The ranked tree shape at maximum $d$ distance to the unbalanced tree $T_{\text{unb}} \in \mathcal{T}_n$ is $T_{\text{bal}}$ with the $\mathbb{F}$-matrix encoding given in (5). That is, $F^{(\text{bal})}_{i,j} = \max\{0, 2j - i + 1\}$ for $i = 1, \ldots, n-1$ and $j = 1, \ldots, i$, and $F^{(\text{bal})}_{i,j} = 0$ for $i = 1, \ldots, n-1$ and $j = i, \ldots, n-1$:*

$$F^{(\text{bal})} = \begin{bmatrix} 2 & & & & & & \\ 1 & 3 & & & & & \\ 0 & 2 & 4 & & & & \\ 0 & 1 & 3 & 5 & & & \\ 0 & 0 & 2 & 4 & 6 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ 0 & 0 & 0 & 0 & \cdots & n-2 & n \end{bmatrix} \quad \textit{and} \quad F^{(\text{unb})} = \begin{bmatrix} 2 & & & & & & \\ 1 & 3 & & & & & \\ 1 & 2 & 4 & & & & \\ 1 & 2 & 3 & 5 & & & \\ 1 & 2 & 3 & 4 & 6 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ 1 & 2 & 3 & 4 & \cdots & n-2 & n \end{bmatrix}. \quad (5)$$

*Proof.* First, the most unbalanced tree has the $\mathbb{F}$-matrix encoding given by $F^{(\text{unb})}$ in (5), that is, $F^{(\text{unb})}_{i,j} = j$ for $i = 2, \ldots, n-1$ and $j = 1, \ldots, i-1$, $F^{(\text{unb})}_{i,j} = 0$ for $i = 1, \ldots, n-1$ and $j = i+1, \ldots, n-1$ (upper triangle) and $F^{(\text{unb})}_{i,i} = i+1$ for $i = 1, \ldots, n-1$ (diagonal). For any $F \in \mathcal{F}_n$, the values in each row are nondecreasing to the right, i.e., $F_{i,j} \leqslant F_{i,j+1}$ (Theorem 1(iii)), and the values in each column are nonincreasing, i.e., $F_{i,j} \geqslant F_{i+1,j}$ (Theorem 1(ii) and (iii)). Second, $F^{(\text{unb})}_{i,j} \geqslant F_{i,j}$ for all $F \in \mathcal{F}_n$ and $i, j \leqslant n-1$, that is, $F^{(\text{unb})}$ has the largest $d_1$ and $d_2$ norms. Third, $F^{(\text{bal})}_{i,j} \leqslant F_{i,j}$ for all $F \in \mathcal{F}_n$ and $i, j \leqslant n-1$, that is, $F^{(\text{bal})}$ has the smallest $d_1$ and $d_2$ norms. Moreover, $F^{(\text{bal})}_{i,j} \leqslant F^{(\text{unb})}_{i,j}$ for all $i, j \leqslant n-1$ and the pair: $T_{\text{unb}}$ and $T_{\text{bal}}$ have the largest $d_1$ and $d_2$ distances among all pairwise distances in $\mathcal{T}_n$. □

While the caterpillar tree, the tree with one cherry and one internal node that subtends two leaves, is the unique tree widely recognized as the most unbalanced tree, there is no consensus notion of a unique most balanced tree (Fischer et al., 2021). The ranked tree shape corresponding to $F^{(\text{bal})}$ is here called the most balanced ranked tree shape for ease of interpretation. However, there may be arguably many more similarly balanced trees in the population.

We now define the signed distance as the distance to a reference ranked tree shape $\bar{T}$ with an additional sign depending on whether the tree is closer to the most unbalanced or to the most balanced tree. The distance to a reference ranked tree shape alone induces a partial order. Moreover, Kim et al. (2020) showed that balanced trees are closer to other balanced trees than to unbalanced trees and, similarly, that unbalanced trees are closer to other unbalanced trees than to balanced trees in the $d$ distance. This makes our proposed signed distance to the mean a natural tree balance index (Fischer et al., 2021).

DEFINITION 5 (SIGNED-DISTANCE FUNCTION TO $\bar{T}$). *Let $f(x) \colon \mathcal{T}_n \to \mathbb{R}^+$ and $\bar{T} \in \mathcal{T}_n$ be a reference ranked tree shape such that*

$$f(x) = \begin{cases} -d(x, \bar{T}) & \text{if } d(x, T_{\text{unb}}) \leqslant d(x, T_{\text{bal}}), \\ d(x, \bar{T}) & \text{if } d(x, T_{\text{unb}}) > d(x, T_{\text{bal}}). \end{cases}$$

The signed distance induces a partial order on $\mathcal{T}_n$; however, since many ranked tree shapes can have the same signed distance to $\bar{T}$, many pairs of trees will be incomparable. We say that $T_1 \sim T_2$ belongs to the same equivalence class if $f(T_1) = f(T_2)$. When a set of ranked

tree shapes belong to the same equivalence class, we order the ranked tree shapes in the equivalence class according to their lexicographic order using a vectorized $F$ representation as follows.

DEFINITION 6 (LEXICOGRAPHIC ORDER). *Let*

$$F^{(1)} = (F^{(1)}_{1,1}, F^{(1)}_{2,1}, \ldots, F^{(1)}_{1,n-1}, F^{(1)}_{2,2}, F^{(1)}_{2,3}, \ldots, F^{(1)}_{n-1,n-1})$$

*be the column-vectorized representation of $T_1 \in \mathcal{T}_n$, and let*

$$F^{(2)} = (F^{(2)}_{1,1}, F^{(2)}_{2,1}, \ldots, F^{(2)}_{1,n-1}, F^{(2)}_{2,2}, F^{(2)}_{2,3}, \ldots, F^{(2)}_{n-1,n-1})$$

*be the column-vectorized representation of $T_2 \in \mathcal{T}_n$. We say that $T_1 \preceq_{\text{lex}} T_2$ if $F^{(1)} = F^{(2)}$ or the first nonvanishing difference $F^{(1)}_i - F^{(2)}_i$ is positive for $i = 1, \ldots, m$, $m = n(n-1)/2$.*

For example, $T^{(\text{unb})} \preceq_{\text{lex}} T^{(\text{bal})}$ and $T^{(\text{unb})} \preceq_{\text{lex}} T \preceq_{\text{lex}} T^{(\text{bal})}$ for any $T \in \mathcal{T}_n$. We note that $\preceq_{\text{lex}}$ is not the only possible lexicographic order; for example, a row-vectorized representation of $T \in \mathcal{T}_n$ can be replaced in Definition 6 to generate another ordering. Any such order provides a consistent way for comparing histograms across different tree models on the same space; see, for example, the third row of Fig. 6 below. However, the column-vectorized representation has some biological meaning, as earlier columns correspond to descendants from early in the branching process. For example, consider the first column of a $4 \times 4$ $\mathbb{F}$-matrix. The lexicographic order, only considering the first column, will be $(2,1,0,0,\ldots) \preceq_{\text{lex}} (2,1,1,0,\ldots) \preceq_{\text{lex}} (2,1,1,1,\ldots)$. This corresponds to the following biological relationship: trees where the earliest born branches split early are placed before trees where those branches split later.

Although the lexicographic order is a total order, we propose to order all ranked tree shapes in the space according to their signed distance to $T_K$ and to only use the lexicographic order within equivalence classes as follows.

DEFINITION 7. *We say that $T_1 \preceq T_2$ if $f(T_1) < f(T_2)$ or if $f(T_1) = f(T_2)$ and $T_1 \preceq_{\text{lex}} T_2$.*

PROPOSITION 4. *The order induced by $\preceq$ of Definition 7 on $\mathcal{T}_n$ is a total order.*

*Proof.* To show antisymmetry, the only way that $T_1 \preceq T_2$ and $T_2 \preceq T_1$ is that $f(T_1) = f(T_2)$ and $T_1 \preceq_{\text{lex}} T_2$ and $T_2 \preceq_{\text{lex}} T_1$. This occurs only if $F^{(1)} = F^{(2)}$. The bijection of Theorem 1 then implies that $T_1 = T_2$. Transitivity and convexity follow directly from the transitivity and convexity of $<$ and $\preceq_{\text{lex}}$. □

## 5. MEASURES OF DISPERSION

In this section we define and discuss three notions for quantifying uncertainty or dispersion in a distribution or a sample of ranked tree shapes, or ranked genealogies.

The Fréchet variance is a natural measure of dispersion for arbitrary probability metric spaces. It measures the concentration around the Fréchet mean. The Fréchet variance of $y \sim \mu$ on $\mathcal{T}_n$ with respect to the metric $d$ is defined as

$$V = \sum_{y \in \mathcal{T}_n} d(y, \bar{T})^2 \cdot \mu(y), \quad \text{where} \quad \bar{T} = \arg\min_{x \in \mathcal{T}_n} \sum_{y \in \mathcal{T}_n} d(x,y)^2 \cdot \mu(y).$$

For a random sample $y_1, \ldots, y_m \in \mathcal{T}_n$, the sample Fréchet variance is given by

$$V_m = \frac{1}{m}\sum_{i=1}^{m} d(y_i, \bar{T})^2, \quad \text{where} \quad \bar{T} = \arg\min_{x \in \mathcal{T}_n} \sum_{i=1}^{m} d(x, y_i)^2.$$

Similarly, the Fréchet variance of $G \sim dv$ can be obtained by integrating over the probability space of branching event times and ranked tree shapes.

Another scalar measure of dispersion is entropy (Mezard & Montanari, 2009):

$$H = -\sum_{y \in \mathcal{T}_n} \mu(y) \cdot \log[\mu(y)].$$

Entropy is a function of the probability measure only and it does not depend on the metric $d$. A measure with zero entropy is concentrated on a single point, and a large entropy indicates greater uncertainty in the position of a random variable with the underlying measure. In Fig. 7 below we compare entropy with Fréchet variance for a particular class of probability models on ranked tree shapes. In this case, Fréchet variance shows more heterogeneous behaviour across different models. We compare those two measures in more detail in § 6.

In many applications, a single mean value and the variance are not enough for summarizing the distribution. Interquartiles and credible intervals are typically used to inform about the concentration of the distribution around the central value for real-valued distributions. The analogues in the space of ranked tree shapes are defined as follows.

A central interquartile ball of ranked tree shapes of level $1 - \alpha$, $\alpha \in [0, 1]$, is the set

$$B_\varepsilon(\bar{T}) \stackrel{\mathrm{D}}{=} \{y \in \mathcal{T}_n \colon d(y, \bar{T}) \leqslant \varepsilon\},$$

where $\varepsilon$ is the smallest $\epsilon \geqslant 0$ such that $P\{B_\epsilon(\bar{T})\} \geqslant 1 - \alpha$, where $\bar{T}$ is a point estimate. Similarly, a level $1 - \alpha$ credible ball is the set $B_\varepsilon(\bar{T})$ where $\varepsilon$ is the smallest $\epsilon \geqslant 0$ such that $P\{B_\epsilon(\bar{T}) \mid \mathcal{D}\} \geqslant 1 - \alpha$.

Although credible and interquartile balls can be defined in a meaningful way in terms of the $d$ distance to the mean value, summarizing boundaries of such sets remains challenging. One attempt to define boundaries for credible sets and interquartile sets is through a total ordering on $\mathcal{T}_n$. The boundaries of the set can then be taken to be the extreme points of the set with respect to the ordering.

Having established the $\preceq$ order, we summarize credible balls and interquartile balls by at most four ranked tree shapes and by at least two ranked tree shapes. Let $B_\varepsilon(\bar{T})$ denote the interquartile or credible ball and $\bar{T}$ the Fréchet mean of the distribution. Then the set of ranked tree shapes at the boundary of $B_\varepsilon(\bar{T})$ will be partitioned into two sets, one with positive signed distance to $\bar{T}$ and one with negative signed distance to $\bar{T}$. If the cardinality of the sets is greater than one, we then summarize each set by the smallest and the largest ranked tree shapes in each set according to the $\preceq$ order. We show an example in § 6.

## 6. Results

### 6.1. *Statistical summaries of Blum–François distributions on ranked tree shapes*

We present and analyse summaries of a large family of ranked tree shape models called the Blum–François $\beta$-splitting model (Sainudiin & Véber, 2016). We start with the model on ranked unlabelled planar trees; there is a distinction between the left and right offspring.

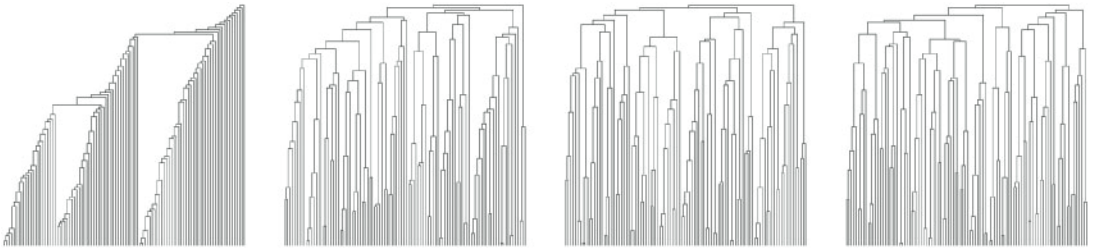

Fig. 5. Approximated Fréchet means. Fréchet means are found via simulated annealing from a sample of $N = 1000$ trees with $n = 100$ leaves from the $\beta$-splitting distribution. Left to right: $\beta = -0.99, -0.5, 0, 10$. Simulated annealing with exponential cooling schedule and decay parameter 0.9995, initial temperature 1000.

Let $n_i^L$ and $n_i^R$ denote the numbers of internal nodes in the left and right subtrees below node $i$. In particular, if node $i$ is a cherry, i.e., subtends two leaves, then $n_i^L = n_i^R = 0$. Then, a ranked unlabelled planar tree with $n$ leaves has a probability mass function given by

$$P(T_{\text{planar}}) = \prod_{i=1}^{n-1} \frac{B(n_i^L + \beta + 1, n_i^R + \beta + 1)}{B(\beta + 1, \beta + 1)},$$

where $B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$ is the beta function and $\beta \in [-1, \infty)$. A ranked tree shape $T$ obtained by ignoring the distinction between left and right subtrees then has probability mass function

$$P(T) = 2^{n-1-c} \prod_{i=1}^{n-1} \frac{B(n_i^L + \beta + 1, n_i^R + \beta + 1)}{B(\beta + 1, \beta + 1)},$$

where $c$ is the number of cherries in $T$. The $\beta$ parameter controls the level of balancedness of the distribution. In particular, when $\beta = 0$, the corresponding distribution $P(T) = 2^{n-1-c}/(n-1)!$ is the coalescent distribution on ranked tree shapes, also known as the Yule distribution.

The first row of Fig. 6 below shows the exact Fréchet means of the ranked tree shape distributions with nine leaves under the Blum–François distribution with $\beta \in \{-0.99, -0.5, 0, 10\}$. For small values of $\beta$, the distribution generates unbalanced trees and, for large values of $\beta$, the distribution generates balanced trees. When $\beta = 0$, there are two means very close to each other, which are depicted in the Supplementary Material. In this case, the distance between the two means is $d = 1.41$, the second smallest value possible. For ranked tree distributions with $n = 100$ leaves, we simulated $N = 1000$ ranked tree shapes and found the Fréchet mean via the simulated annealing of § 3.4. The resulting means are shown in Fig. 5 for $\beta \in \{-0.99, -0.5, 0, 10\}$.

To show the utility of summarizing central interquartile balls, we calculated the 95% interquartile ball of the Blum–François distribution with $\beta = 0$ and $n = 9$ tips. The ball is centred in one of the Fréchet means: the one on the right of Fig. 1 in the Supplementary Material. The 95% quartile of the distance to the mean is 16, and 41 ranked tree shapes are at a distance of 16 to the mean in the boundary of the ball. Panels (B)–(C) of Fig. 2 in the Supplementary Material show the two trees at the signed distance of $-16$ to the mean with minimum and maximum lexicographic order, respectively. Panels (D)–(E) of Fig. 2 in the

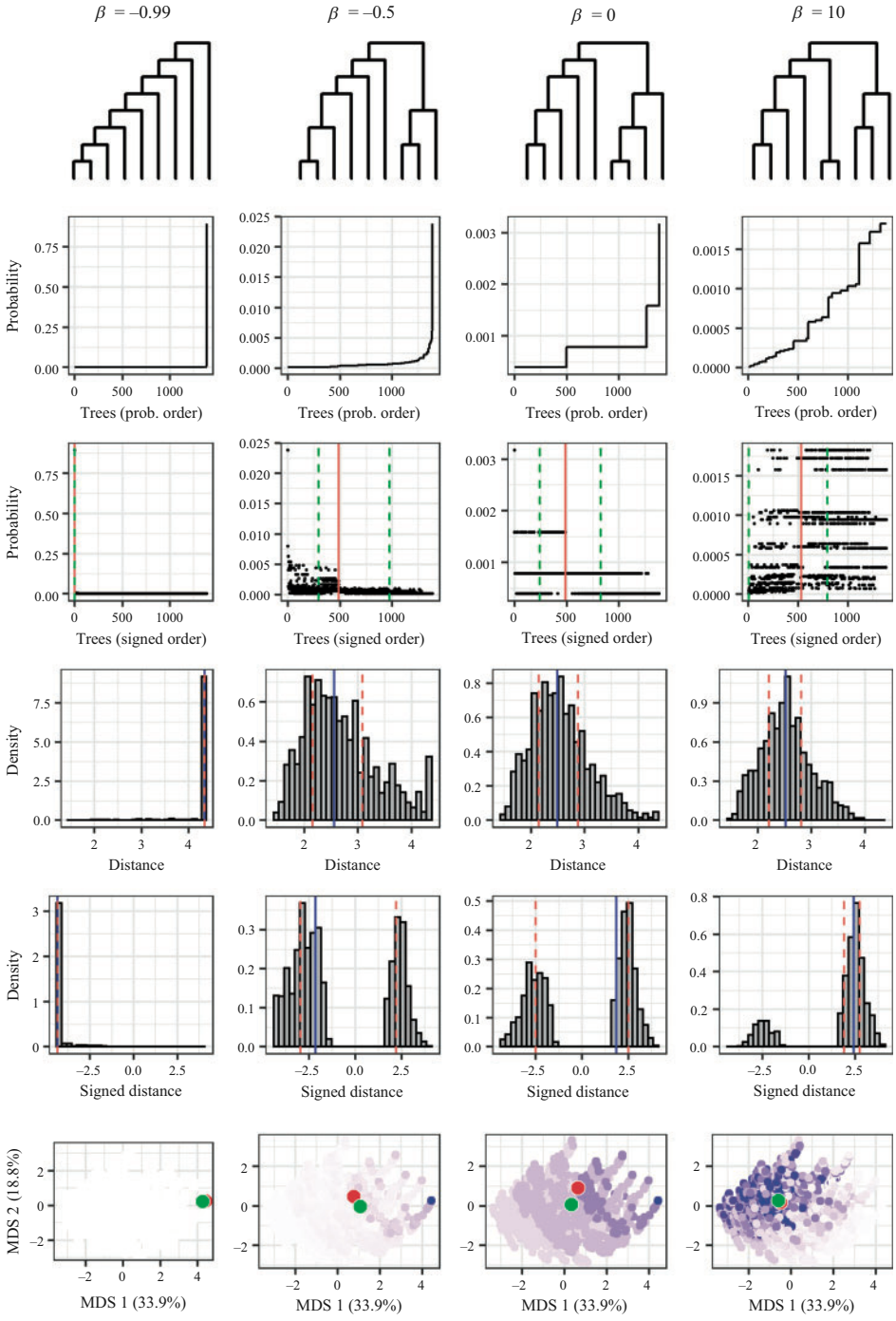


Fig. 6. Summarizing Blum–François distributions on ranked tree shapes. Blum–François distributions on ranked tree shapes with $n = 9$ leaves. The columns correspond to $\beta = -0.99, -0.5, 0$ (coalescent), 10, respectively. Row 1: Fréchet mean of the distribution. Row 2: probability mass function of trees, arranged in increasing order of probability. Row 3: probability mass function of trees, arranged in the signed-distance order of Definition 7, with Fréchet mean (solid line) of the distribution and interquartiles (dashed lines). Row 4: histogram of the distance to the Kingman Fréchet mean, with the median (solid line) and interquartiles (dashed lines) of the distance to the mean. Row 5: histogram of the signed distance to the Kingman Fréchet mean, with the median (solid line) and interquartiles (dashed lines) of the signed distance. Row 6: multi-dimensional scale (MDS) of the tree distribution, where each dot represents a tree coloured by its probability mass, with Fréchet mean (red dot) and expected value (Euclidean mean, green dot).

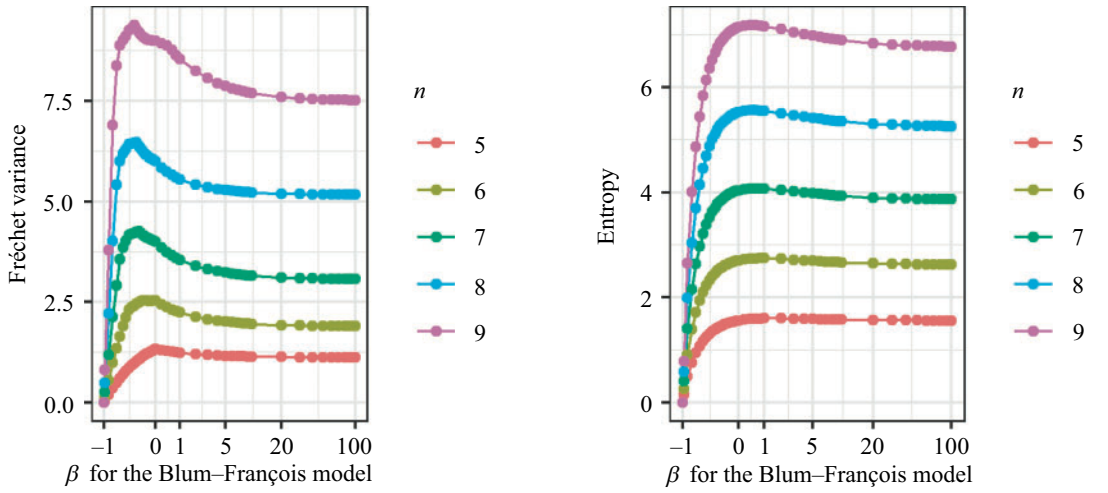


Fig. 7. Measures of dispersion. Fréchet variance and entropy for small $n$ under the Blum–François $\beta$-splitting model. Although both statistics show similar trajectories across models, the Fréchet variance is more differentiated across models than the entropy.

Supplementary Material show the two trees at the signed distance of 16 to the mean with minimum and maximum lexicographic order, respectively.

Figure 6 shows different summaries of four Blum–François distributions on ranked tree shapes with $n = 9$ leaves. The second row shows the probability mass function with trees arranged in increasing order of probability. When analysing these plots, it is impossible to assess whether the $\beta = -0.5$ distribution puts more probability mass on unbalanced trees than balanced trees when compared to the $\beta = 10$ distribution since the $x$ axes are not comparable. The third row shows the probability mass functions with the $x$ axes arranging trees in the signed-distance total order. Here, all $x$ axes correspond to the same tree arrangements. It is now clear that the $\beta = -0.5$ distribution assigns more probability mass to unbalanced trees and the $\beta = 10$ distribution assigns more mass to balanced trees. This is confirmed in the fifth row of Fig. 6. The histogram of the signed distance to the mean is skewed to the right (balanced) when $\beta = 10$. Row four of Fig. 6 shows the histograms of the distance to the mean. This one-dimensional summary of the tree distributions hinders whether some distributions put more probability mass to different types of trees. For example, the last three histograms of the fourth column look very similar. Finally, while the multi-dimensional scale plots in the last row of Fig. 6 only explain about 53% of all pairwise distances, the last panel shows the distinctions between the four probability mass distributions. Here, green dots correspond to the points whose $\mathbb{F}$-matrix is $\mathbb{E}(F)$ and do not lie in tree space and red dots are the Fréchet means.

Figure 7 shows the exact Fréchet variance and entropy for $n = 5, \ldots, 9$ and $\beta$ values spaced out in $[-1, \infty]$. For $\beta > 1$, the variance and entropy remain relatively constant as functions of $\beta$. The largest variance and entropy are obtained when $\beta \leqslant 0$. Variance and entropy show similar trajectories across models; however, the variance is more differentiated than the entropy for small values of $\beta$.

### 6.2. *Characterization of the mean Kingman tree*

As stated in § 3.3, under $d$, the population Fréchet mean is given by

$$\bar{F}_2 = \underset{F \in \mathcal{F}_n}{\arg\min} \sum_{k,l} \{F_{kl}^2 - 2F_{kl}M_{kl}\},$$

where $M_{kl} = \mathbb{E}(F_{k,l})$. If we know matrix $M$, we only need to search for ranked tree shapes that are in a neighbourhood of $M$, see, for example, the large dots representing $M$ and the Fréchet mean in the last row of Fig. 6. In fact, the only data input needed using `gurobi` or simulated annealing is $M$. Although there is no explicit formula for the Fréchet mean for the distributions analysed here, there is an explicit formula for $M$ for the Kingman/Yule coalescent distribution; Blum–François with $\beta = 0$. In Fig. 6, we visualize $M$ and the Fréchet mean in a multi-dimensional scale plot of the entire space.

THEOREM 2. *Let $F \in \mathcal{F}_n$ be an $\mathbb{F}$-matrix distributed according to the Blum–François model with $\beta = 0$, i.e., according to the Kingman/Yule coalescent distribution; then*

(i) *the distribution of the ith row of $F$ is independent of $n$;*
(ii) $\mathbb{E}[F_{ij}] = j(j+1)/i$;
(iii) *we have*

$$\operatorname{var}[F_{ij}] = \frac{j^3 2(j+1)^2}{i^2(i-1)} + \frac{j(j+1)(i-2j-1)}{i(i-1)};$$

(iv) *we have*

$$\operatorname{cov}[F_{i_1j_1}, F_{i_2j_2}] = \begin{cases} \dfrac{j_1(j_1+1)[j_2(j_2+2)+(i_1+1)(i_1-2j_2-2)]}{i_1^2(i_1-1)} & \text{when } i_1 = i_2, j_1 < j_2, \\ \dfrac{j_2(j_2+1)(j_2-i_2)(j_2-i_2+1)}{i_1 i_2 (i_2-1)} & \text{when } i_1 > i_2, j_1 = j_2, \\ \dfrac{j_2(j_2+1)[(j_1+1)(j_1+2)+(i_1+1)(i_1-2j_1-2)]}{i_1^2(i_1-1)} & \\ \quad + \dfrac{(i_1-i_2)j_1j_2(j_2+1)}{i_1 i_2}\left[\dfrac{j_1-1}{i_2-1} - \dfrac{j_1+1}{i_2+1}\right] & \text{when } i_1 > i_2, j_1 > j_2, \\ \dfrac{j_1(j_1+1)[(j_2+1)(j_2+2)+(i_1+1)(i_1-2j_2-2)]}{i_1^2(i_1-1)} & \\ \quad + \dfrac{(i_1-i_2)j_1(j_1+1)j_2}{i_1 i_2}\left[\dfrac{j_2-1}{i_2-1} - \dfrac{j_2+1}{i_2+1}\right] & \text{when } i_1 > i_2, j_1 < j_2. \end{cases}$$

The relevance of Theorem 2 is that, for the standard coalescent, one of the most popular models in population genetics (Wakeley, 2008), the Fréchet mean can be obtained for any $n$, without the need for simulating a sample from the distribution as is done in Fig. 5. Moreover, given a sample of $\mathbb{F}$-matrices, the sample average converges almost surely to $M$ by the law of large numbers. Theorem 2, together with the multivariate central limit theorem (Hogg et al., 2019, Theorem 5.4.4) and Hotelling's T-squared test, can be used to test whether a random sample of ranked tree shapes follows the standard coalescent distribution such as the Kingman/Yule model.

THEOREM 3 (CENTRAL LIMIT THEOREM FOR $\mathbb{F}$-MATRICES). *Let $F^1, \ldots, F^m \in \mathcal{F}_n$ be an independent and identically distributed sample of $\mathbb{F}$-matrices drawn from some distribution $P$. Let $\bar{F}_m \in \mathbb{R}^{(n-1)\times(n-1)}$ be the matrix whose entries correspond to the sample average of $F^1, \ldots, F^m$. Then $m^{1/2}(\bar{F}_m - M) \xrightarrow{\mathrm{D}} N(0, \Sigma)$, where the mean $M \in \mathbb{R}^{(n-1)\times(n-1)}$ is given by $M_{ij} = \mathbb{E}_P[F_{ij}]$ and the covariance tensor $\Sigma \in \mathbb{R}^{(n-1)^2\times(n-1)^2}$ is given by $\Sigma_{ij,kl} = \operatorname{cov}_P[F_{ij}, F_{kl}]$.*

*Proof.* The proof follows directly from the multivariate central limit theorem, considering the $\mathbb{F}$-matrices as elements of $\mathbb{R}^{(n-1)\times(n-1)}$. Since each entry of the $\mathbb{F}$-matrices is bounded in $[0, n]$, all expectations are finite. □

COROLLARY 1. *Consider the setting of Theorem 3, and assume that* $\Sigma$ *is invertible. Let* $\hat{\Sigma} \in \mathbb{R}^{(n-1)^2\times(n-1)^2}$ *be the empirical covariance tensor. We have*

$$\hat{\Sigma}^{-1/2} m^{1/2} (\bar{F}_m - M) \xrightarrow{\mathrm{D}} N(0, I),$$

*where* $I \in \mathbb{R}^{(n-1)^2\times(n-1)^2}$ *is the identity tensor given by*

$$I_{ij,kl} = 1_{i=k, j=l}.$$

The proof follows by the multivariate version of Slutsky's theorem, using consistency of the empirical covariance, and the fact that the function $\Sigma \to \Sigma^{-1/2}$ is continuous when $\Sigma$ is an invertible covariance matrix.

Although classic tests for neutrality under the coalescent rely on summary statistics computed from observed molecular data (Tajima, 1989; Ferretti et al., 2010), it is well understood that these tests can be obscured by other processes such as demography that affect branch length distributions. To reduce this impact, tests based on phylogenetic tree topology have been proposed (Drummond & Suchard, 2008; Yang et al., 2018). These tests rely on one-dimensional tree summary statistics such as external tree length, assuming unit interbranching events, and the number of cherries. Here, we use Hotelling's T-squared test for assessing $H_0 \colon E[F] = M$, where $M$ corresponds to the expected $F$ under the standard coalescent. To show the applicability of the results in this section, we simulated 200 samples of 1000 ranked tree shapes with $n = 25$ tips from the Blum–François distribution, each with parameter $\beta \in (-1, 10]$. Figure 8 shows the power of Hotelling's T-squared test. We observe that the test is valid at level 0.05 and also has good power away from $\beta = 0$.

### 6.3. *Summaries of coalescent ranked genealogical distributions*

To show the applicability of our simulated annealing algorithm for summarizing genealogies, we simulated genealogies according to the neutral isochronous coalescent model with variable population size. Here, the tree topology and the branching times are independent. The ranked tree shape is distributed according to the Kingman/Yule/Blum–François model with $\beta = 0$, and the branching event times have the conditional density

$$f\{u_{i-1} \mid u_i, N_e(t)\} = \frac{\binom{i}{2}}{N_e(u_{i-1})} \exp\left\{ -\binom{i}{2} \int_{u_i}^{u_{i-1}} \frac{\mathrm{d}u}{N_e(u)} \right\}$$

with $u_n = 0$ and $N_e(t)$ a nonnegative function that denotes the effective population size (Slatkin & Hudson, 1991).

We simulated 1000 ranked genealogies according to the neutral coalescent model with three different $N_e(t)$ trajectories described in the Supplementary Material. We depict the three distributions in a multi-dimensional scaling plot and show the Fréchet means and medoids of each sample and the six central summaries are in the Supplementary Material, together with the three posterior distributions of tree height. For the three distributions, the medoid and the Fréchet mean are different. The main difference lies in the branch lengths: Fréchet means use posterior mean tree heights, while medoids show more atypical heights; in particular, the medoid tree under constant population size is much shorter than average.

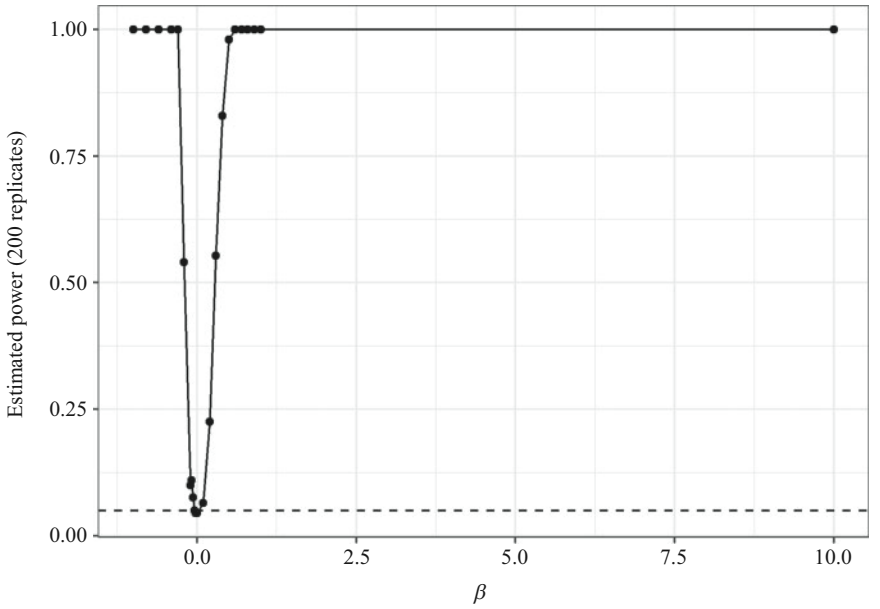


Fig. 8. Estimated power of Hotelling's T-squared test. Power is estimated using 200 replicates each of 1000 trees with $n = 25$ sampled from the Blum–François distribution with parameter $\beta$. We use Hotelling's T-squared test statistic to test the null $H_0$: $E[F] = M$, where $M$ corresponds to the expected $F$ under the standard coalescent.

## 7. Analysis of SARS-CoV-2

We first use our method to find the posterior sample Fréchet mean genealogy of 100 SARS-CoV-2 molecular sequences publicly available in the GISAID EpiCov database (Shu & McCauley, 2017) from the state of California, USA for the period of February 2020 to September 2020. The posterior distribution of genealogies is estimated with BEAST (Suchard et al., 2018). Details of parameters and prior distributions selected for BEAST analyses and data access acknowledgments can be found in the Supplementary Material. We show the multi-dimensional scaling plot of the posterior samples and different central summaries in the Supplementary Material. Both the Fréchet mean and the in-sample medoid are designed to be central with respect to the $d$ metric using $\mathbb{F}$-matrices, whereas the maximum clade credibility tree is not. Hence, it is not surprising that the maximum clade credibility tree is further away from the centre as compared to the other point summaries. The Fréchet mean is closer to the centre than the in-sample medoid.

In this study, we selected 100 sequences uniformly at random from the pool of available sequences in GISAID for the state of California during the first nine months of the pandemic. Given that molecular sampling efforts increased during that time, our uniform sampling would reflect the effects of such an effort. However, if sampling efforts were centred towards local outbreaks, genealogical posterior distributions from different uniform subsamples could be very different and this would raise concerns about the representativeness of our results to the population. Here, we propose to compare several subposterior genealogical distributions corresponding to different subsamples of molecular sequences drawn uniformly at random in order to assess the stability of their sample Fréchet means.

To analyse the stability of the posterior Fréchet means, we selected three samples of 100 sequences uniformly at random from GISAID and compared their subposterior distributions. We estimated the three subposteriors with BEAST for each of the four states:

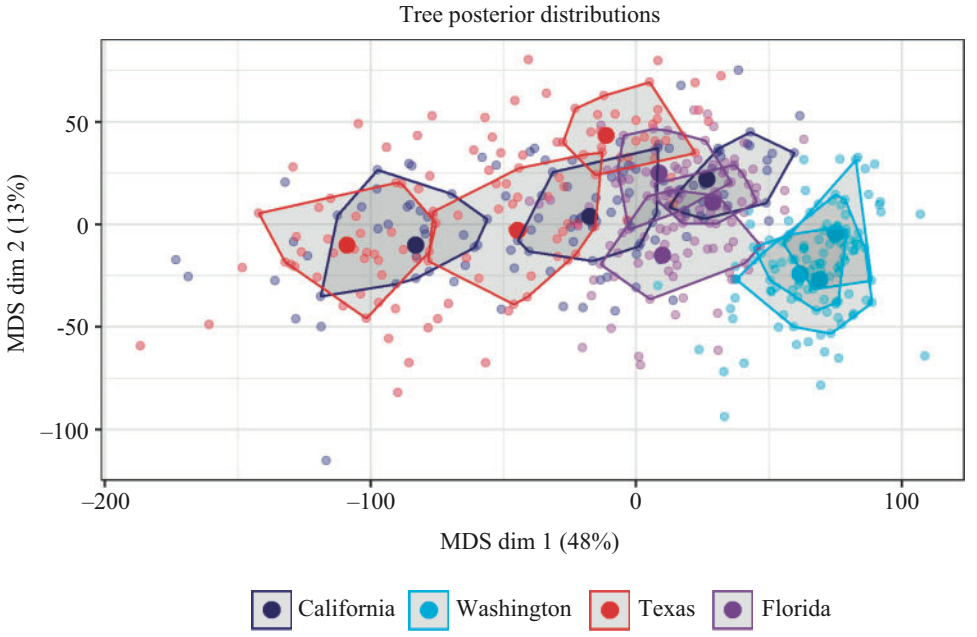


Fig. 9. Multi-dimensional scaling plot of multiple samples from California, Washington, Florida and Texas. Three samples of 20 trees of $n = 100$ samples randomly chosen among GISAID sequences in February–May 2020 per location. The Fréchet means are calculated using average coalescent times and marked as large dots. The shaded region corresponds to 50% credible convex hulls around the Fréchet means.

California, Washington, Texas and Florida. Details of parameters chosen for BEAST analyses and data access acknowledgments can be found in the Supplementary Material. We thinned subposterior samples to 20 trees, each with 100 tips, per subposterior.

Figure 9 shows the multi-dimensional scaling plot of the three subposteriors from each state. The subposteriors of Washington state, light blue dots in Fig. 9, are more concentrated than any of the other three states. In this case, the posterior sample Fréchet means are very close to each other and their posterior convex hulls overlap substantially, indicating stability of the posterior sample Fréchet mean. The posterior distributions of Florida, marked in purple, are the second more concentrated: their three Fréchet means are close to each other and their convex hulls overlap. In contrast, the posterior genealogical distributions of California are very different: their posterior sample Fréchet means are far from each other and their convex hulls do not overlap. A similar pattern is observed in Texas. This large heterogeneity observed in California and Texas may be the result of local outbreak sequencing efforts in the area. In these two cases, simple random subsampling from GISAID may not lead to representative results.

## 8. Discussion

For discrete tree topologies, the Fréchet mean ranked tree shape may not be unique. However, in our experience, we found the Fréchet means to be very close to each other. We conjecture that the set of Fréchet means has a very small diameter, which will be explored in future research. While the nonuniqueness of the Fréchet mean can be potentially problematic for hypotheses testing, we remark that the expected $\mathbb{F}$-matrix, here denoted by $M$, is unique and the limit of the sample mean $\bar{F}_m$ by the strong law of large numbers. In this manuscript, we provided a central limit theorem result for $\bar{F}_m$ and analytical expressions

for $M$ and the variance of the ranked tree shape distribution under the standard Kingman coalescent. Theoretical results of this kind for other distributions is left to future work. Similarly, analyses of several test statistics based on the distances analysed here and Fréchet means such as in Dubey & Müller (2019) are the subject of future research.

## Acknowledgement

We thank Paromita Dubey for useful discussions. Palacios was supported by the National Institutes of Health, the Alfred P. Sloan Foundation and a National Science Foundation Career Award. Palacios is also affiliated with the Department of Biomedical Data Science, Stanford Medicine.

## Supplementary material

The Supplementary Material includes additional details of distance calculations and simulated annealing, as well as proofs omitted from the main text.

## References

Aarts, E. & Korst, J. (1988). *Simulated Annealing and Boltzmann Machines*. Chichester: John Wiley and Sons.
Benner, P. & Bačák, M. (2014). Point estimates in phylogenetic reconstructions. *Bioinformatics* **30**, i534–40.
Billera, L. J., Holmes, S. P. & Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* **27**, 733–67.
Brown, D. G. & Owen, M. (2020). Mean and variance of phylogenetic trees. *Syst. Biol.* **69**, 139–54.
Chakerian, J. & Holmes, S. (2012). Computational tools for evaluating phylogenetic and hierarchical clustering trees. *J. Comp. Graph. Statist.* **21**, 581–99.
Cranston, K. A. & Rannala, B. (2007). Summarizing a posterior distribution of trees using agreement subtrees. *Syst. Biol.* **56**, 578–90.
Donaghey, R. (1975). Alternating permutations and binary increasing trees. *J. Combin. Theory* **18**, 141–8.
Drummond, A. J. & Suchard, M. A. (2008). Fully Bayesian tests of neutrality using genealogical summary statistics. *BMC Genet.* **9**, 1–12.
Dubey, P. & Müller, H.-G. (2019). Fréchet analysis of variance for random objects. *Biometrika* **106**, 803–21.
Ferretti, L., Perez-Enciso, M. & Ramos-Onsins, S. (2010). Optimal neutrality tests based on the frequency spectrum. *Genetics* **186**, 353–65.
Fischer, M., Herbst, L., Kersting, S., Kühn, L. & Wicke, K. (2021). Tree balance indices: a comprehensive survey. *arXiv:* 2109.12281v1.
Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré* **10**, 215–310.
Govek, K., Sikes, C. & Oesper, L. (2018). A consensus approach to infer tumor evolutionary histories. In *Proc. 2018 ACM Int. Conf. Bioinformatics, Comp. Biol., Health Informatics*, pp. 63–72. New York: Association for Computing Machinery.
Gurobi Optimization, L. (2020). *Gurobi Optimizer Reference Manual*.
Hillis, D. M., Heath, T. A. & John, K. S. (2005). Analysis and visualization of tree space. *Syst. Biol.* **54**, 471–82.
Hogg, R. V., McKean, J. & Craig, A. T. (2019). *Introduction to Mathematical Statistics*. Boston: Pearson.
Karcher, M. D., Palacios, J. A., Lan, S. & Minin, V. N. (2017). phylodyn: an R package for phylodynamic simulation and inference. *Mol. Ecol. Resour.* **17**, 96–100.
Kim, J., Rosenberg, N. A. & Palacios, J. A. (2020). Distance metrics for ranked evolutionary trees. *Proc. Nat. Acad. Sci.* **117**, 28876–86.
Kingman, J. (1982). The coalescent. *Stoch. Proces. Appl.* **13**, 235–48.
Kirkpatrick, M. & Slatkin, M. (1993). Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* **47**, 1171–81.
Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* **220**, 671–80.
Kuhner, M. K. & Yamato, J. (2014). Practical performance of tree comparison metrics. *Syst. Biol.* **64**, 205–14.
Lemant, J., Le Sueur, C., Manojlović, V. & Noble, R. (2022). Robust, universal tree balance indices. *Syst. Biol.* **71**, 1210–24.

Mezard, M. & Montanari, A. (2009). *Information, Physics, and Computation*. New York: Oxford University Press.
Mooers, A. O. & Heard, S. B. (1997). Inferring evolutionary process from phylogenetic tree shape. *Quart. Rev. Biol.* **72**, 31–54.
Paradis, E. & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–8.
R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Sainudiin, R. & Véber, A. (2016). A beta-splitting model for evolutionary trees. *R. Soc. Open Sci.* **3**, 160016.
Shu, Y. & McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22**, doi: 10.2807/1560-7917.ES.2017.22.13.30494.
Slatkin, M. & Hudson, R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129**, 555–62.
Stanley, R. P. (1999). *Enumerative Combinatorics*. Cambridge: Cambridge University Press.
Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J. & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, doi: 10.1093/ve/vey016.
Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–95.
Volz, E. M., Koelle, K. & Bedford, T. (2013). Viral phylodynamics. *PLoS Comput. Biol.* **9**, e1002947.
Wakeley, J. (2008). *Coalescent Theory: An Introduction*. Greenwood Village, CO: Roberts and Company.
Willis, A. & Bell, R. (2018). Confidence sets for phylogenetic trees. *J. Comp. Graph. Statist.* **27**, 542–52.
Yang, Z., Li, J., Wiehe, T. & Li, H. (2018). Detecting recent positive selection with a single locus test bipartitioning the coalescent tree. *Genetics* **208**, 791–805.

[*Received on* 1 *June* 2021. *Editorial decision on* 8 *March* 2023]