A Metric Space of Ranked Tree Shapes and Ranked Genealogies

Jaehee Kim[1], Noah A. Rosenberg[1], and Julia A. Palacios[2,3]

[1] *Department of Biology, Stanford University, Stanford, CA 94305, USA*
[2] *Department of Statistics, Stanford University, Stanford, CA 94305, USA*
[3] *Department of Biomedical Data Science, Stanford Medicine, Stanford, CA 94305, USA*
[3] *Corresponding author. Email: juliapr@stanford.edu*

December 23, 2019

Abstract

Genealogical tree modeling is essential for estimating evolutionary parameters in population genetics and phylogenetics. Recent mathematical results concerning ranked genealogies without leaf labels enable new opportunities in the analysis of evolutionary trees. In particular, comparisons between ranked genealogies facilitate the study of evolutionary processes for organisms sampled in multiple time periods. We propose a metric space on ranked genealogies for lineages sampled from both isochronous and time-stamped heterochronous sampling. Our new tree metrics make it possible to conduct statistical analyses of ranked tree shapes and timed ranked tree shapes, or ranked genealogies. Such analyses allow us to assess differences in tree distributions, quantify estimation uncertainty, and summarize tree distributions. We show the utility of our metrics via simulations and an application in infectious diseases.

Introduction

Gene genealogies are rooted binary trees that describe the ancestral relationships of a sample of molecular sequences at a locus. The properties of these genealogies are influenced by the nature of the evolutionary forces experienced by the sample's ancestry. Hence, assessing differences between gene genealogies can provide information about differences in these forces. In this manuscript, we propose a distance on genealogies that enables biologically meaningful comparisons between genealogies of non-overlapping samples. Our proposed distance, in its more general form, is a distance on *ranked genealogies*.

Ranked genealogies are rooted binary and unlabeled trees with branch lengths and ordered internal nodes (Figure 1(A)). Ranked genealogies are also known as Tajima's genealogies (Sainudiin et al., 2015; Palacios et al., 2019) as they were examined by Tajima (1983). Ranked genealogies are coarser than labeled topologies with branch lengths specified but finer than unranked unlabeled genealogies in which neither branch order nor taxon labels are specified (Figure 1). Recently, there has been increasing interest in modeling ranked genealogies for studying evolutionary dynamics (Ford et al., 2009; Lambert and Stadler, 2013; Sainudiin et al.,

2015). A new method for inferring evolutionary parameters from molecular data is based on the Tajima coalescent of ranked genealogies (Palacios et al., 2019): modeling of ranked unlabeled genealogies, as opposed to ranked labeled genealogies (Kingman coalescent), reduces the dimensionality of the inference problem. In studying macroevolution, Maliet et al. (2018) proposed a model on ranked unlabeled tree topologies to investigate factors influencing nonrandom extinction and the loss of phylogenetic diversity.

Many metrics on labeled binary trees—such as the Robinson-Foulds (RF) metric (Robinson and Foulds, 1981), the Billera-Holmes-Vogtmann (BHV) metric (Billera et al., 2001), and the Kendall-Colijn metric (Kendall and Colijn, 2016)—have been proposed. These metrics make possible comparisons between labeled genealogies. They have been used for summarizing posterior distributions and bootstrap distributions of trees on the same set of taxa (Chakerian and Holmes, 2012; Brown and Owen, 2019), for comparing estimated genealogies of the same organisms with different procedures, and for quantifying uncertainty (Willis and Bell, 2018). A metric on the lower resolution space of tree shapes—unlabeled unranked binary trees without branch lengths—has recently been proposed by Kendall and Colijn (2016). To our knowledge, no other metrics on ranked unlabeled genealogies have been proposed to date.

A tree metric on the space of ranked unlabeled genealogies facilitates evaluations of the quality of an estimation procedure, by enabling measurements of the distance between an estimated ranked genealogy and the true ranked genealogy. It can assist in comparing different estimators from different procedures, and in comparing estimated ranked genealogies of different organisms at different geographical locations and different time periods. Moreover, a useful tree metric not only provides a quantitative measure for ranked genealogy comparison, it can also discriminate between the key aspects of different evolutionary processes (Holmes, 2003). We show that our metrics separate samples of genealogies originating from different distributions of ranked tree topologies and ranked genealogies. Our distances enable the computation of summary statistics, such as the mean and the variance, from samples of ranked genealogies. When ranked genealogy samples are obtained from posterior distributions of genealogies, such as those obtained from `BEAST` (Suchard et al., 2018), our tree metrics enable the construction of credible sets.

We first define a metric on ranked unlabeled tree shapes. Our metric relies on an integer-valued triangular matrix representation of ranked tree shapes. This matrix representation allows us to use metrics on matrices, such as the Frobenius norm, to define metrics on ranked tree shapes. The choice of the Frobenius norm produces computational benefits, as computations of the metrics are quadratic in the number of leaves. Our metrics on ranked unlabeled tree shapes retain more information than metrics based on unlabeled unranked tree shapes alone.

We expand our metric definition to ranked genealogies including branch lengths by weighting the matrix representation of ranked tree shapes by branch lengths. We first define a distance on ranked unlabeled isochronous trees, with all samples obtained at the same point in time. We then later define a metric space on heterochronous ranked tree shapes and heterochronous genealogies, in which samples are time-stamped. While modeling isochronous genealogies is the standard practice for slowly evolving organisms, such as humans and other types of animals, modeling heterochronous genealogies is the standard approach for rapidly evolving organisms, such as viruses and other pathogens. Heterochronous genealogies are also increasingly relevant in the study of ancient DNA samples.

We analyze different properties of our proposed metrics and compare them to other tree-valued metrics—such as BHV (Billera et al., 2001) and Kendall-Colijn (Kendall and Colijn, 2016)—by projecting these other metrics into the space of ranked tree shapes and ranked genealogies. We then demonstrate the performance of our metrics on simulations from various tree topology distributions and demographic scenarios, and we use them to compare posterior samples of genealogies of human influenza A virus from different geographic regions.

Preliminaries

Definitions of tree topologies and genealogies

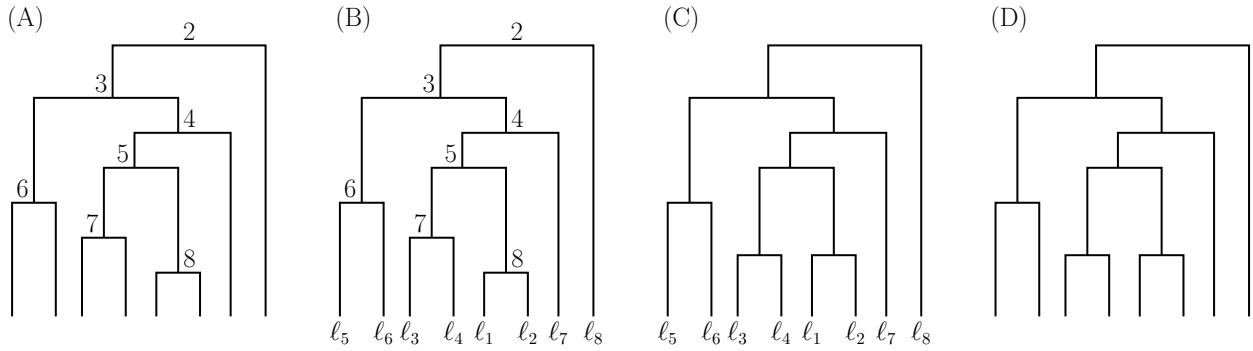


Figure 1: **Types of tree topology**. (A) Ranked tree shape (T^R). (B) Labeled ranked tree shape (T^{LR}). (C) Labeled unranked tree shape (T^L). (D) Tree shape (T^S). Genealogies corresponding to each topology include branch lengths.

All the trees we consider are rooted and binary. We first assume that trees are isochronously sampled, that is, all tips of the trees start at the same time. A *ranked tree shape* with n leaves is a rooted unlabeled binary tree with an increasing ordering of the $n-1$ interior nodes, starting at the root with label 2 (Figure 1(A)). We use the symbol T^R to denote a ranked tree shape. A *ranked genealogy* is a ranked tree shape equipped with branch lengths. We use the symbol $\mathbf{g}^R$ to denote a ranked genealogy. Although we mainly focus on ranked tree shapes and ranked genealogies, we will compare our metrics to metrics defined on other rooted tree spaces. A *labeled ranked tree shape* (Figure 1(B)) and a *labeled ranked genealogy* are the corresponding labeled counterparts of unlabeled ranked trees, and they are denoted by T^{LR} and $\mathbf{g}^{LR}$ respectively. A *labeled unranked tree shape* (T^L) (Figure 1(C)) and *labeled unranked genealogy* ($\mathbf{g}^L$) are a rooted labeled tree shape and a rooted labeled genealogy without ranking of internal nodes, respectively. A *tree shape* (Figure 1(D)) is an unlabeled unranked rooted bifurcating tree denoted by T^S and the corresponding *genealogy* is denoted by $\mathbf{g}^S$.

Probability models on tree topologies and genealogies

Generative probability models on *isochronous* binary tree topologies with n leaves are Markov jump processes that either start at the root with two branches and proceed (forward in time) by choosing one of the extant branches to bifurcate at random until the desired number of leaves is obtained, or start at the bottom with

n leaves and proceed (backward in time) by choosing two extant lineages to merge into a single lineage until there is a single lineage. The first class corresponds to branching processes (Aldous, 1996; Ford, 2005; Steel, 2016) and the second class corresponds to coalescent jump processes (Kingman, 1982; Wakeley, 2009). The corresponding *heterochronous* Markov jump forward and backward in time processes are the birth-death process and the heterochronous coalescent process, respectively. In the heterochornous coalescent, sampling times (death of lineages) are usually assumed fixed (Felsenstein and Rodrigo, 1999). In each case, the distribution of branching times or coalescent times is independent of the tree topology and depends only on the number of extant branches.

In this manuscript, we simulate isochronous ranked tree shapes from two different models: the one-parameter beta-splitting model (Aldous, 1996) and a two-parameter model which we term the alpha-beta splitting model (Maliet et al., 2018). To generate isochronous ranked genealogies and heterochronous ranked tree shapes and ranked genealogies, we use the Tajima coalescent formulation of the coalescent (Sainudiin et al., 2015; Palacios et al., 2019). In the following, we briefly describe these models. Extensive accounts of probability models of evolutionary trees can be found in (Mooers and Heard, 1997; Ford, 2005; Blum and François, 2006; Steel, 2016; Sainudiin and Véber, 2016).

Beta-splitting model on labeled tree shapes

We first consider the single-parameter *beta-splitting model* on labeled tree shapes (Aldous, 1996). For a parent clade of size n, the model chooses its child clade size to be i on the left branch and $n-i$ on the right branch with probability

$$q_n(i) = \frac{1}{a_n(\beta)} \frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)},$$

where $a_n(\beta)$ is a normalizing constant and $i \in \{1, 2, \ldots, n-1\}$. The splitting is repeated recursively in each branch independently until the tree is fully resolved. The parameter $\beta \in [-2, \infty)$ controls the degree of balance of the generated trees. With $\beta = -2$, the model generates the perfect unbalanced tree (caterpillar tree) with probability one, and with $\beta = \infty$, the model generates the perfect balanced tree with probability one. In particular, we consider the following special cases: the Yule model (Yule, 1925; Harding, 1971) ($\beta = 0$), the proportional-to-distinguishable-arrangements (PDA) model (Blum and François, 2006) ($\beta = -1.5$), and the Aldous' branching (AB) model (Aldous, 2001) ($\beta = -1$). Additionally, we include $\beta = -1.9$ and $\beta = 100$, to which we refer as "unbalanced" and "balanced", respectively. We note that Ford's alpha-model Ford (2005) is another single parameter family of models on the same class of trees as the beta-splitting model and it is not considered in this manuscript.

Alpha-beta splitting model and beta-splitting model on ranked tree shapes

The *alpha-beta model* of Maliet et al. (2018) generates labeled ranked tree shapes according to a size-biased distribution with a stick-breaking construction. The algorithm first generates n independent draws $u_1, \ldots, u_n$ from a Unif$(0, 1)$ distribution. The u_i's correspond to the n leaves of the tree. The first partition of the n leaves (at the root) is determined by drawing a random number $R_1 \sim \text{Beta}(\beta+1, \beta+1)$. All the $n_1 = \sum_{i=1}^{n} 1(u_i < R_1)$ are placed on the left side of the tree and the rest on the other side. Then, if $n_1 > 1$, the left branch is chosen to bifurcate with probability proportional to n_1^α. The algorithm continues

generating beta-distributed values to bi-partition the leaves and chooses the order (ranking) proportional to their number of descendants until the interval $(0, 1)$ is partitioned into n intervals, each corresponding to an u_i number. The pseudocode is in Section S1. Labels are then removed to generate a ranked tree shape. The $\beta \in [-2, \infty)$ parameter determines the balance of the tree as in the beta-splitting model, and the $\alpha \in (-\infty, \infty)$ parameter regulates the relationship between subtree family size (number of descendants) and closeness to the root. More specifically, when $\alpha < 0$, subtrees with small family sizes are closer to the root and when $\alpha > 0$, subtrees with small family sizes are closer to the tips. When $\alpha = 1$, the alpha-beta model becomes a *Beta-splitting model on ranked tree shapes*.

Tajima coalescent on ranked genealogies

The Tajima coalescent is a model on ranked genealogies (Figure 1(A)). It is a Markov lumping of Kingman's n-coalescent on labeled and ranked genealogies (Sainudiin et al., 2015; Palacios et al., 2019). The Tajima coalescent is a pure death process that starts with n unlabeled leaves at time $t_{n+1} = 0$ and proceeds backward in time, merging pairs of branches to create interior nodes labeled by their order of appearance. The merging of two branches is a coalescent event. In the Tajima coalescent, the distribution of coalescence times is the same as in the Kingman coalescent and the probability of a topology, ranked tree shape, is given by

$$P(T^R) = \frac{2^{n-c-1}}{(n-1)!}, \tag{1}$$

where n is the number of leaves, and c is the number of cherries—the number of pairs of leaves that subtend from a shared interior node. The Tajima coalescent on ranked tree shapes without times corresponds to the beta-splitting model on ranked tree shapes with $\beta = 0$, also called the Yule model. A full description of the Tajima coalescent process can be found in Cappello and Palacios (2019).

Distributions on branching or coalescent times

As mentioned before, in this manuscript we consider tree models of neutral evolution in which branching or coalescent waiting times depend on the number of extant branches but are independent of the particular topology configuration or clade composition. A popular model in phylogenetics is to assume that, when there are k branches, the branching waiting time forward in time from root to tips is exponentially distributed with rate $k\lambda$ with $\lambda > 0$ (Mooers and Heard, 1997). We will refer to this model as λ-*branching*. In the standard coalescent model, when there are k lineages, the coalescent time t_k backward in time from tips to root (see Figure 3(A)) is exponentially distributed with rate $\frac{\binom{k}{2}}{\lambda}$, where λ is referred to as the effective population size, which is assumed to be constant over time. We will refer to this model as the λ-*coalescent*. In the coalescent with variable population size (Slatkin and Hudson, 1991), the first coalescent time when there are n lineages has marginal density

$$f(t_n) = \frac{\binom{n}{2}}{\lambda(t_n)} e^{-\binom{n}{2}\int_0^{t_n} \frac{dt}{\lambda(t)}},$$

and the conditional density of t_k when there are k branches is

$$f(t_k \mid t_{k+1}) = \frac{\binom{k}{2}}{\lambda(t_k)} e^{-\binom{k}{2}\int_{t_{k+1}}^{t_k} \frac{dt}{\lambda(t)}}, \text{ for } k = n-1, \ldots, 2,$$

where $\lambda(t) > 0$ is the effective population size at time t. We will refer to the variable population size distribution on coalescent times as the $\lambda(t)$-*coalescent*.

Coalescent distribution on heterochronous tree topologies and genealogies

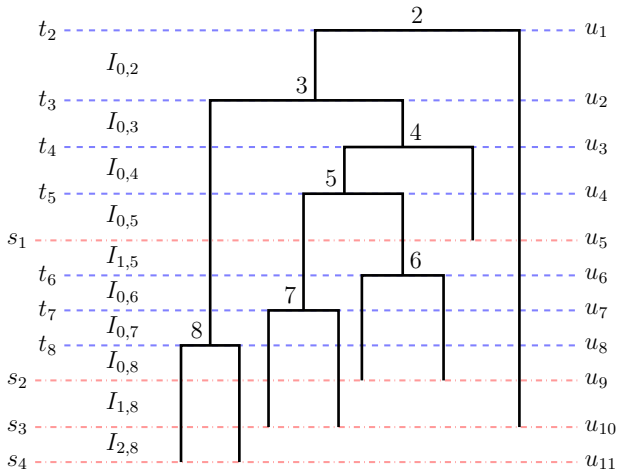


Figure 2: **Heterochronous genealogy.** Example of a ranked genealogy with heterochronous sampling. $s_4, \ldots, s_1$ and $t_8, \ldots, t_2$ indicate sampling times (red dotted lines) and coalescent times (blue dotted lines), respectively. $u_{11}, \ldots, u_1$ are the ordered times of change points where the number of lineages changes due to either a sampling event or a coalescent event. The time increases backwards in time starting with $u_{11} = 0$ as the present time. $I_{0,k}$ is the interval that ends with a coalescent event at t_k. $I_{i,k}$ $(i > 0)$ represents the interval that ends with a sampling time within the interval (t_{k+1}, t_k). For $k = n$, we adopt the convention $t_{n+1} = 0$.

The ranked tree shape and ranked genealogy of time-stamped samples are termed *heterochronous ranked tree shapes* and *heterochronous genealogies*, respectively (Figure 2). Here, we assume that samples are collected at times $s_m, s_{m-1}, \ldots, s_1$, with $s_m = t_{n+1} = 0$ (the present), and $s_j < s_{j-1}$ for $j = m, \ldots, 2$. At time s_j, n_j lineages are sampled, and $\sum_{j=1}^{m} n_j = n$.

The *$\lambda(t)$-heterochronous-coalescent* (Felsenstein and Rodrigo, 1999) describes the distribution of coalescent times conditioned on collecting $n_m, n_{m-1}, \ldots, n_1$ samples at times $s_m, s_{m-1}, \ldots, s_1$. As before, $t_{n+1}, t_n, ..., t_2$ denote the coalescent times, except that the subindex no longer indicates the number of lineages. Instead, the subindex indicates the rank order of the coalescent events going forward in time from the root at t_2. Define $(u_{n+m-1}, u_{n+m-2}, \ldots, u_1)$ as the vector of change points (coalescent or sampling times), with $0 = u_{n+m-1} < u_{n+m-2} < \cdots < u_1 = t_2$. To state the density of coalescent times according to the λ(t)-heterochronous-coalescent, going backwards in time, we denote the intervals that end with a coalescent event at t_k by $I_{0,k}$, and the intervals that end with a sampling time within the interval (t_{k+1}, t_k) as $I_{i,k}$, where i

is an index tracking the sampling events in (t_{k+1}, t_k). More specifically,

$$\begin{aligned} I_{0,k} &= (\max\{t_{k+1}, s_j\}, t_k] \text{ for } s_j < t_k \\ I_{i,k} &= (\max\{t_{k+1}, s_{j+i}\}, s_{j+i-1}] \text{ for } s_{j+i-1} > t_{k+1} \text{ and } s_j < t_k, \end{aligned} \tag{2}$$

with $k = 2, ..., n$ and i ranges from 1 to the number of sampling events in (t_{k+1}, t_k). An example of the annotated time intervals using $I_{0,k}$ and $I_{i,k}$ is shown in Figure 2.

The conditional density of t_{k-1} is the product of the conditional density of $t_{k-1} \in I_{0,k}$ and the probability of not having a coalescent event during the period of time spanned by intervals $I_{1,k}, ..., I_{m,k}$. That is, for $k = 2, ..., n$,

$$P[t_{k-1}|t_k, \mathbf{s}, \mathbf{n}, N_e(t)] = \frac{C_{0,k-1}}{N_e(t_{k-1})} \exp\left[-\left\{ \int_{I_{0,k-1}} \frac{C_{0,k-1}dt}{N_e(t)} + \sum_{i=1}^{m} \int_{I_{i,k-1}} \frac{C_{i,k-1}dt}{N_e(t)} \right\} \right], \tag{3}$$

where $C_{i,k} = \binom{n_{i,k}}{2}$.

Tree metrics for comparative analysis

Metrics on labeled trees. A large number of tree metrics have been proposed for phylogenetics. Billera et al. (2001) introduced a metric space of trees for labeled genealogies now commonly known as BHV space. Owen and Provan (2011) and Chakerian and Holmes (2012) provided polynomial algorithms and implementations for calculating the geodesic distance metric proposed by Billera et al. (2001). Recently, Kendall and Colijn (2016) proposed a new metric on labeled unranked trees and labeled genealogies representing each tree as a convex combination of two vectors—one vector encoding number of edges from the root to the most recent common ancestor of every pair of tips in the tree, and the other vector encoding the corresponding path lengths. The most popular metric that can be computed in polynomial time is the symmetric difference of Robinson and Foulds (RF) on labeled unranked tree shapes (Robinson and Foulds, 1981) and the branch-length measure RFL on labeled genealogies (Robinson and Foulds, 1979). Other metrics on labeled tree shapes are based on the number of rearrangement steps needed to transform one tree into the other, including the nearest neighbor interchange distance (Li and Zhang, 1999) and the metrics based on pruning and regrafting or tree bisection and reconnection (Allen and Steel, 2001; Bordewich and Semple, 2005).

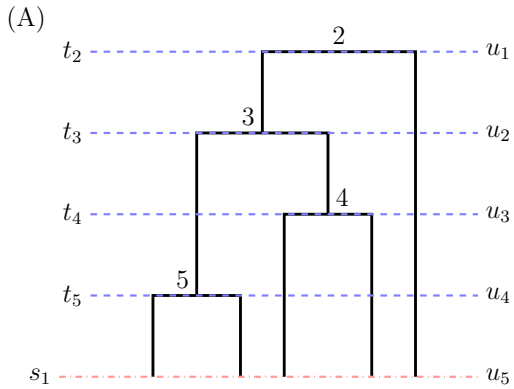
Metrics on unlabeled trees. A metric on unlabeled trees is desirable because it allows comparison between trees from different samples. However, not many such metrics are available. Colijn and Plazzotta (2018) proposed a metric on tree shapes which is the Euclidean norm of the difference between two integer vectors that uniquely describe the two trees. Poon et al. (2013) developed a kernel function that measures the similarity between two unlabeled genealogies by accounting for differences in branch lengths and matching number of descendants over all nodes for both trees. Lewitus and Morlon (2015) proposed the Jensen-Shannon distance between the spectral density profiles of the corresponding modified graph Laplacian of the unlabeled genealogies. The modified graph Laplacian of an unlabeled genealogy is constructed as the difference between its degree matrix—a diagonal matrix with i-th diagonal element being the sum of the branch lengths from node i to all the other nodes—and its distance matrix whose (i, j)-element is the branch length between nodes i and j.

To our knowledge, we are introducing the first metric for unlabeled ranked tree shapes and unlabeled ranked genealogies. In order to define our metric on ranked tree shapes, we need to introduce a unique encoding of a ranked tree shape as an integer-valued triangular matrix defined in the next section.

New Approaches

Unique encoding of ranked tree shapes and F-matrix

Let T^R be a ranked tree shape with n leaves sampled at time $0 = u_n$. Let $(u_{n-1}, \ldots, u_1)$ be the $n-1$ coalescent times. Here, time increases into the past, $u_{n-1} < \cdots < u_1$, and u_n and u_1 correspond to the most recent sampling time and the time to the most recent common ancestor (root), respectively. An **F**-matrix that encodes T^R is an $(n-1) \times (n-1)$ lower triangular matrix of integers with elements $F_{i,j} = 0$ for all $i < j$, and for $1 \leq j \leq i$, $F_{i,j}$ is the number of extant lineages in (u_{j+1}, u_j) that do not bifurcate during the entire time interval (u_{i+1}, u_j).



(B)

$$\mathbf{F} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 2 & 4 & 0 \\ 1 & 1 & 3 & 5 \end{pmatrix}$$

Figure 3: **Unique encoding of isochronous ranked tree shapes and F-matrices.** (A) Example of a ranked genealogy with isochronous sampling. (B) The corresponding **F**-matrix that encodes the ranked tree shape information of the tree in (A).

In Figure 3(B), we show the corresponding **F**-matrix to the ranked tree shape depicted in Figure 3(A). In the interval (u_2, u_1), there are two lineages so $F_{1,1} = 2$. One of the two lineages extant at time (u_2, u_1) branches at time u_2 while the other lineage does not branch throughout the entire time interval (u_5, u_1). This gives the first column of the **F**-matrix: $F_{2,1} = F_{3,1} = F_{4,1} = 1$. For the second column, we start with three lineages in the interval (u_3, u_2), $F_{2,2} = 3$. Of the three lineages extant at (u_3, u_2), one branches at u_3 ($F_{3,2} = 2$), and one branches at u_4 ($F_{4,2} = 1$). We construct the third column by starting from four lineages in (u_4, u_3), $F_{3,3} = 4$. One lineage branches at u_4, and thus, $F_{4,3} = 3$. Finally, in the interval (u_5, u_4), there are five lineages, which gives $F_{4,4} = 5$.

If T^R is a heterochronous ranked tree shape with n leaves and m sampling events, the corresponding **F**-matrix representation is an $(n+m-2) \times (n+m-2)$ lower triangular matrix of integers, where $F_{i,j}$ for $1 \leq j \leq i$ is the number of extant lineages in (u_{j+1}, u_j) that do not bifurcate or become extinct during the entire time interval (u_{i+1}, u_j) traversing forward in time. Here, $(u_{n+m-1}, u_{n+m-2}, \ldots, u_1)$, such that

$0 = u_{n+m-1} < u_{n+m-2} < \cdots < u_1$, are the $n+m-1$ ordered change points of T^R, at each of which the number of branches changes either by a coalescent event or by a sampling event. We show an example of a heterochronous ranked tree shape and its **F**-matrix encoding in Figure S2(C-D).

Although the branch lengths and the actual values of coalescent and sampling times are irrelevant for the specification of the ranked tree shape, we rely on the u_i to identify the order and type of change points, coalescence or sampling, in T^R and **F**.

Theorem 1. (Unique Encoding of Ranked Tree Shapes). The map by which ranked tree shapes with n samples and m sampling events are encoded as **F**-matrices of size $(n+m-2) \times (n+m-2)$ uniquely associates a ranked tree shape with an **F**-matrix.

In other words, given an **F**-matrix, if it encodes a ranked tree shape, it encodes exactly one ranked tree shape. The proof can be found in Section S2. In the next section, we will leverage the **F**-matrix representation of ranked tree shapes to define a distance between ranked tree shapes. From this point onward, we will assume a matrix **F** represents a ranked tree shape.

Metric spaces on ranked tree shapes and ranked genealogies

We define two distance functions d_1 and d_2 on the space of ranked tree shapes with n leaves as follows. For a pair of ranked tree shapes T_1^R and T_2^R and their corresponding **F**-matrix representations $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ of size $r \times r$,

$$d_1(T_1^R, T_2^R) = \sum_{i=1}^{r} \sum_{j=1}^{i} \left| F_{i,j}^{(1)} - F_{i,j}^{(2)} \right|,$$

$$d_2(T_1^R, T_2^R) = \sqrt{\sum_{i=1}^{r} \sum_{j=1}^{i} \left(F_{i,j}^{(1)} - F_{i,j}^{(2)} \right)^2}.$$

The two distances are metrics since they inherit properties of the L_1-norm (Manhattan distance) and L_2-norm (Frobenius norm). The definitions are valid for both isochronous and heterochronous ranked tree shapes as both types of trees have unique matrix encodings.

Figure 4 shows all isochronous ranked tree shapes of $n = 5$ leaves and their pairwise d_1 distances. The d_1 metric shows the following desirable features: the largest distances occur between trees with different tree shapes and between the most balanced and unbalanced trees. There are 3 different values of d_1 among the 10 possible pairs of trees: $1, 2$, and 3. The d_2 distance exhibits the same features. The three different values of d_2 pairwise distances are $1, \sqrt{2}$ and $\sqrt{3}$, and the relative comparisons remain the same as for d_1. In the following definition, we include branch lengths to define distances between ranked genealogies.

We next define two weighted distance functions d_1^w and d_2^w on the space of ranked genealogies with n samples and m sampling events. We first define the weight matrix **W** of a ranked genealogy $\mathbf{g}^R$ as a lower triangular matrix of size $(n+m-2) \times (n+m-2)$ with entries $W_{i,j} = u_j - u_{i+1}$ for $j \leq i$ and $W_{i,j} = 0$ otherwise. Here, $(u_{n+m-1}, u_{n+m-2}, \ldots, u_1)$, such that $0 = u_{n+m-1} < u_{n+m-2} < \cdots < u_1$, is the vector of ordered

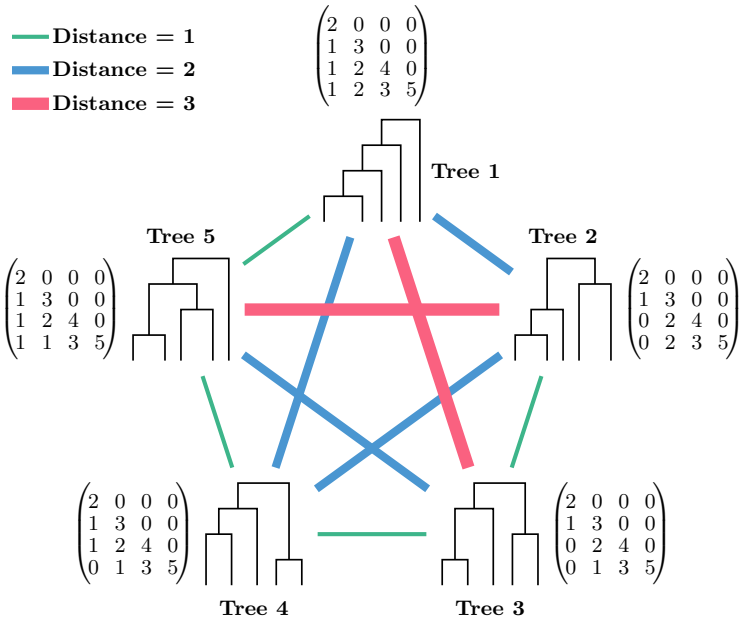


Figure 4: d_1 **distance between all pairs of ranked tree shapes of $n = 5$ leaves**. Rankings of internal nodes are removed for ease of visualization. There are three different distance values among the 10 pairs of distances. The maximum distance of 3 is between trees 2 and 5 (both trees have different shapes) and between trees 1 and 3 (most unbalanced tree and most balanced tree). Trees 2, 3, and 4 share the same shape but with different internal node rankings. Among those three trees, Trees 2 and 4 differ by two ranking moves, whereas the other tree pairs differ by one ranking move. Indeed, trees 2 and 4 have distance 2, and the other two pairs have distance 1 under d_1 metric.

$n-1$ coalescent times t_k and m sampling times s_ℓ with time increasing into the past. For a pair of ranked genealogies $\mathbf{g}_1^R$ and $\mathbf{g}_2^R$, and their corresponding $\mathbf{F}$-matrix representations $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$,

$$d_1^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = \sum_{i=1}^{r}\sum_{j=1}^{i}\left|F_{i,j}^{(1)}W_{i,j}^{(1)} - F_{i,j}^{(2)}W_{i,j}^{(2)}\right|,$$

$$d_2^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = \sqrt{\sum_{i=1}^{r}\sum_{j=1}^{i}\left(F_{i,j}^{(1)}W_{i,j}^{(1)} - F_{i,j}^{(2)}W_{i,j}^{(2)}\right)^2},$$

where $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are the weight matrices associated with $\mathbf{g}_1^R$ and $\mathbf{g}_2^R$, respectively. Figure S3 shows the weight matrix $\mathbf{W}$ associated with the example heterochronous ranked genealogy and its $\mathbf{F}$-matrix in Figures S2(C) and (D).

Proposition 2. The weighted distances d_1^w and d_2^w are metrics.

The proof can be found in Section S3. Our distances on ranked tree shapes and ranked genealogies are distances between trees with the same number of leaves and, in the heterochronous case, the same number of leaves and the same number of sampling events. The extension to cases in which the numbers of sampling events differ but the total numbers of leaves remain the same is described in the Materials and Methods section.

In the following section, we propose sample summary statistics based on our metrics d_1 and d_2 for ranked tree shapes and d_1^w and d_2^w for ranked genealogies.

Ranked tree shape summary statistics

We use our proposed distances to define a notion of mean value and dispersion value from a finite sample $\{T_1^R, T_2^R, \ldots, T_s^R\}$ of ranked tree shapes with n leaves.

Our proposed measures of centrality are the L_2-medoid sets defined as:

$$\overline{T}_i := \underset{T \in \{T_1^R, T_2^R, \ldots, T_s^R\}}{\operatorname{argmin}} \sum_{j=1}^{s} d_i^2(T, T_j^R). \tag{4}$$

for $i = 1, 2$. We note that our definition of the L_2-medoid set corresponds to the ranked tree shapes in the sample that minimizes the sum of squared distances as opposed to minimize the sum of distances. In addition, when the sample is replaced by the complete population of ranked tree shapes with n leaves or when we allow the L_2-medoid to belong to the population of trees but not necessarily to be a sampled tree, Equation 4 corresponds to the *Fréchet mean* or *barycenter* under uniform sampling probabilities (Bacák, 2014).

We use the following as a measure of dispersion around the medoid for $i = 1, 2$:

$$\sigma_i^2 := \frac{1}{s} \sum_{j=1}^{s} d_i^2(\overline{T}_i, T_j^R). \tag{5}$$

Similarly, given a finite sample of ranked genealogies $\mathbf{g}_1^R, \ldots, \mathbf{g}_s^R$ with n leaves, the L_2-medoid set is defined as:

$$\overline{\mathbf{g}_i} := \underset{\mathbf{g} \in \{\mathbf{g}_1^R, \ldots, \mathbf{g}_s^R\}}{\operatorname{argmin}} \sum_{j=1}^{s} \left[d_i^w(\mathbf{g}, \mathbf{g}_j^R) \right]^2$$

for $i = 1, 2$. Similarly, our empirical measure of dispersion for ranked genealogies is the average sum of squared distances to the medoid.

Adapting other tree metrics to ranked tree shapes and ranked genealogies

Other distances on ranked tree shapes. Although there are no other metrics or distances on ranked tree shapes, we can adapt other tree distances to the space of ranked tree shapes and compare them to our metric. We start with two adaptations of metrics that are originally defined on the space of labeled genealogies: the BHV distance and the KC distance.

The Billera-Holmes-Vogtmann metric (BHV) metric. The BHV space (Billera et al., 2001) is obtained by representing each labeled genealogy $\mathbf{g}^L$ with edge set $\mathcal{E}$ by a vector in the Euclidean orthant $\mathbb{R}_+^{2n-1}$, whose coordinates correspond to edge lengths. The BHV space is the union of $(2n-3)!!$ orthants. The BHV distance (d_{BHV}) between two labeled genealogies is defined as a geodesic, the shortest path connecting two points that lies inside the BHV space. Note that unranked and labeled genealogies with positive intervals between coalescent and sample times are effectively ranked and labeled genealogies. To adapt the BHV distance to the space of ranked tree shapes, we define a modified BHV metric, $d_{\text{BHV-RTS}}$ as follows:

$$d_{\text{BHV-RTS}}(T_1^R, T_2^R) = d_{\text{BHV}}(\psi(T_1^R), \psi(T_2^R)),$$

where ψ maps a ranked tree shape to its corresponding ranked labeled genealogy by assigning a uniquely defined label to each leaf and assigning a unit length to each change point time interval (u_i, u_{i-1}).

The unique assignment of the leaf labels $\ell_1, \ldots, \ell_n$ on a ranked tree shape consists in assigning labels in increasing index order starting with leaves subtending from nodes closer to the bottom and ending with leaves subtending closer to the root. Detailed description of the unique label assignment can be found in Materials and Methods.

We could alternatively define another BHV-based distance $d_{\text{BHV-RTS}^*}$ on ranked tree shapes as follows:

$$d_{\text{BHV-RTS}^*}(T_1^R, T_2^R) = \min_{\pi_i, \pi_j \in S_n} \left\{ d_{\text{BHV}}\big((\pi_i \circ \psi)(T_1^R), (\pi_j \circ \psi)(T_2^R)\big) \right\}.$$

Here, ψ assigns an initial labeling to a ranked tree shape and assigns unit branch lengths. S_n is the set of all permutations of the set of leaf labels $\{\ell_1, \ldots, \ell_n\}$. Although $d_{\text{BHV-RTS}^*}$ is a valid distance on ranked tree shapes and perhaps more natural than $d_{\text{BHV-RTS}}$, it requires computing BHV distances between all possible pairs of permutations of leaf labels of the two ranked tree shapes. The number of such possible pairs is exponential in the number of leaves and hence it becomes computationally intractable for large n. In our results, we only analyze $d_{\text{BHV-RTS}^*}$ for the $n = 5$ case.

The Kendall-Colijn (KC) metric. The KC metric is another metric on labeled genealogies (Kendall and Colijn, 2016). A labeled genealogy $\mathbf{g}_n^L$ with n leaves is represented by an $\frac{n(n+1)}{2}$-dimensional vector $v_\lambda(\mathbf{g}_n^L)$ that is a convex combination of two vectors: $(1-\lambda)m(\mathbf{g}_n^L) + \lambda M(\mathbf{g}_n^L)$, $\lambda \in [0, 1]$. $m(\mathbf{g}^L)$ is a vector that concatenates n repetitions of one and a vector whose entry corresponds to the number of edges between the most recent common ancestor of a pair of leaves and the root; $M(\mathbf{g}_n^L)$ is a vector that concatenates the vector of leaf branch lengths and the branch length between the most recent common ancestor of a pair of tips and the root.

The KC distance $d_{\text{KC},\lambda}$ with $\lambda > 0$ between two labeled genealogies is the Euclidean norm of the difference between the two KC vector representations of the labeled genealogies. When $\lambda = 0$, the KC distance becomes a distance on the space of labeled unranked topologies: $d_{\text{KC},0}$.

To adapt the KC distance to the space of ranked tree shapes, we propose two distances. We first define a KC-based distance on ranked tree shapes, $d_{\text{KC-RTS}}$, as follows

$$d_{\text{KC-RTS}}(T_1^R, T_2^R) = d_{\text{KC},0}(\eta(T_1^R), \eta(T_2^R)),$$

where η maps a ranked tree shape to a labeled unranked tree shape by removing internal node rankings and uniquely labeling leaves following the procedure described for $d_{\text{BHV-RTS}}$.

Alternatively, we can adapt the KC distance on labeled genealogies to define $d_{\text{KC-RTS}^*}$ by uniquely labeling the tips of the ranked tree shape, artificially assigning change point intervals length 1 as in $d_{\text{BHV-RTS}}$, and $\lambda = 0.5$ to account for rankings, as follows:

$$d_{\text{KC-RTS}^*}(T_1^R, T_2^R) = d_{\text{KC},0.5}(\psi(T_1^R), \psi(T_2^R)), \tag{6}$$

where ψ is the mapping previously defined for $d_{\text{BHV-RTS}}$.

The Colijn-Plazzotta (CP) metric. The CP metric is defined on tree shapes (Colijn and Plazzotta, 2018). The CP metric d_{CP} is defined as the Euclidean norm (L_2-norm) of the difference between two vectors that uniquely describe the two tree shapes. Each node of a tree is labeled by an integer recursively from tips to the root. The ith entry of the CP vector representing a tree shape records the frequency of the tree nodes labeled with integer i. We define a modified CP distance $d_{\text{CP-RTS}}$ on ranked tree shapes as

$$d_{\text{CP-RTS}}(T_1^R, T_2^R) = d_{\text{CP}}(\phi(T_1^R), \phi(T_2^R)),$$

where ϕ returns the corresponding tree shape of a ranked tree shape by removing the labels of its internal nodes. We note that $d_{\text{CP-RTS}}$ is not a metric but a pseudometric: all pairs of different ranked tree shapes with the same shape will have $d_{\text{CP-RTS}} = 0$. In addition, $d_{\text{CP-RTS}}$ does not account for heterochronous sampling, so we exclude $d_{\text{CP-RTS}}$ from our analyses on heterochronous ranked tree shapes.

The Robinson-Foulds (RF) distance. The RF distance on labeled and unranked tree shapes (Robinson and Foulds, 1981) is also a popular metric that can be adapted to ranked tree shapes by labeling the leaves with our unique labeling scheme: $d_{\text{RF-RTS}}(T_1^R, T_2^R) = d_{\text{RF}}(\eta(T_1^R), \eta(T_2^R))$, where η is the map defined for $d_{\text{KC-RTS}}$.

Other distances on ranked genealogies. We now present the modified BHV and KC distances so that they can be compared to our proposed distances on ranked genealogies.

$$d_{\text{BHV-RG}}(\mathbf{g}_1^R, \mathbf{g}_2^R) = d_{\text{BHV}}(\eta_2(\mathbf{g}_1^R), \eta_2(\mathbf{g}_2^R)),$$

where η_2 maps a ranked genealogy to a labeled ranked genealogy by uniquely labeling leaves as described for $d_{\text{BHV-RTS}}$. Similarly,

$$d_{\text{KC-RG}}(\mathbf{g}_1^R, \mathbf{g}_2^R) = d_{\text{KC},0.5}(\eta_2(\mathbf{g}_1^R), \eta_2(\mathbf{g}_2^R)).$$

Results

Having introduced the metrics on the space of ranked tree shapes and the space of ranked genealogies, we examine the behavior of our metrics in a schematic example for $n = 5$, simulated data under various models, and in real human influenza A virus data. The details of simulation and computation steps can be found in the Materials and Methods section.

Interpreting proposed distances between ranked tree shapes of $n = 5$ leaves

We first compare our distances on ranked tree shapes of $n = 5$ leaves with our adaptations of other metrics. In Figure 4, we show all 5 possible ranked tree shapes and their corresponding pairwise d_1 distances. Trees T_2^R, T_3^R and T_4^R are trees with the same tree shape but different ranked tree shapes. The pairs (T_3^R, T_4^R) and (T_2^R, T_3^R) differ by one ranking switch between two cherries—ranking 2 by 3 and ranking 3 by 4, respectively—and their pairwise distance is $d_1 = 1$. The d_1 distance between the pair (T_2^R, T_4^R) is 2. Indeed, to go from T_2^R to T_4^R, we need two ranking switches: ranking 2 by 3 and then ranking 3 by 4. All pairwise distances are shown in Figure 5. Qualitative behaviors of distances $d_{\text{KC-RTS}}$ and $d_{\text{BHV-RTS}^*}$ between the trees T_2^R, T_3^R

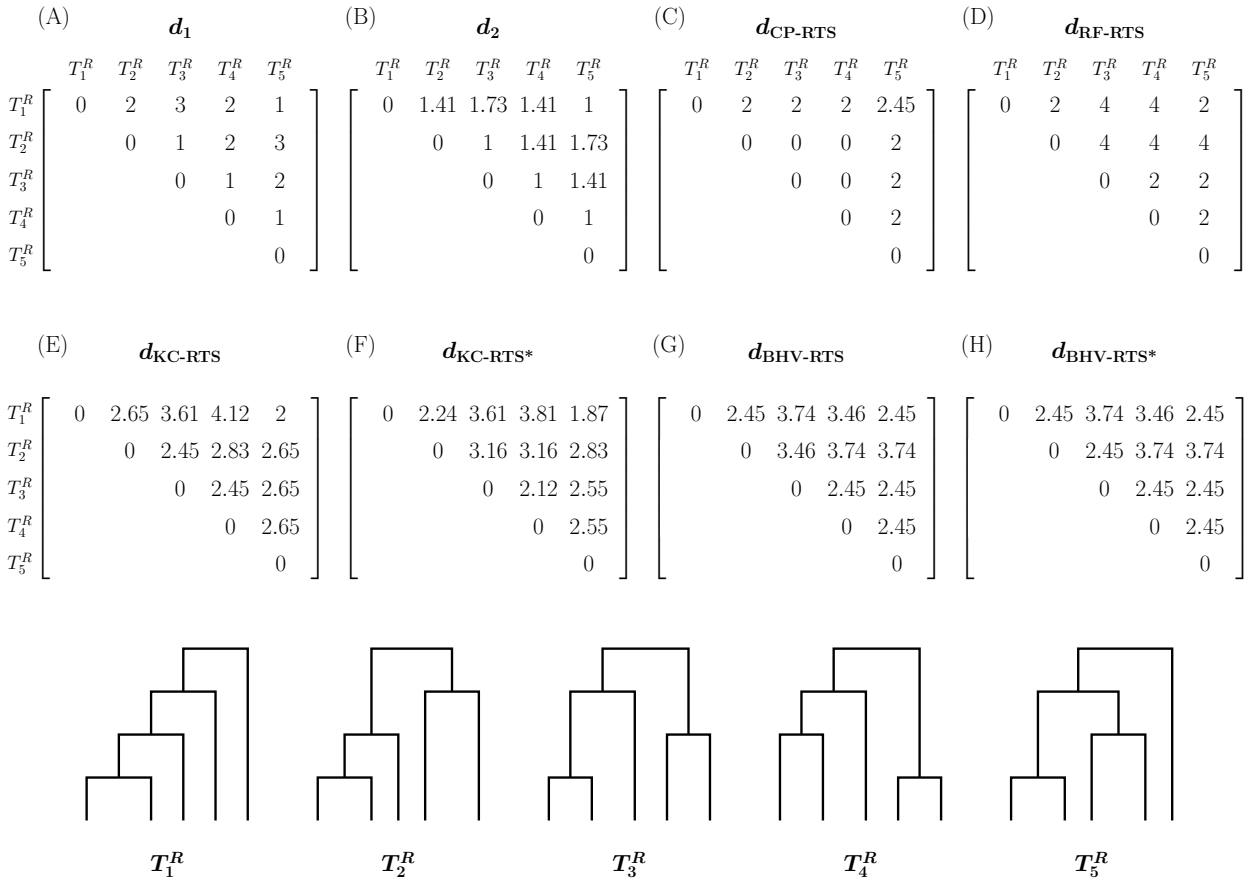


Figure 5: **Comparisons of metrics applied to isochronous ranked tree shapes with $n = 5$.**

and T_4^R mirror d_1.

The pairs (T_1^R, T_3^R) and (T_2^R, T_5^R) have the maximum d_1 distance. In both cases, we need a ranking switch and a change of tree shapes to move from one tree to the other in each pair. The pairs of trees with the maximum $d_{\text{BHV-RTS}}$ or $d_{\text{BHV-RTS*}}$ distance are the same two pairs with the maximum d_1 distance. However, an additional pair, (T_2^R, T_4^R), also has the same $d_{\text{BHV-RTS}}$ and $d_{\text{BHV-RTS*}}$ maximum value even when T_2^R and T_4^R differ by two ranking change events but have the same tree shape. Unlike $d_{\text{BHV-RTS}}$ or $d_{\text{BHV-RTS*}}$, our d_1 distance penalizes ranking changes and tree shape changes differently. The distances between pairs (T_3^R, T_5^R) and (T_1^R, T_2^R) are penalized more with the d_1 distance than with the $d_{\text{BHV-RTS*}}$ since it involves changes in tree shapes.

In the analysis of the $n = 5$ case, our d_1 distance is more similar to $d_{\text{BHV-RTS*}}$ than to any other of the distances considered. The second most similar distance to d_1 is $d_{\text{RF-RTS}}$. In general, we notice that $d_{\text{BHV-RTS*}}$ penalizes changes in internal node rankings and changes in tree shapes equally, whereas our d_1 distance penalizes tree shapes changes more than ranking changes. As mentioned previously, computation of $d_{\text{BHV-RTS*}}$ is expensive; however, $d_{\text{BHV-RTS}}$ and $d_{\text{BHV-RTS*}}$ produce the same pairwise distances except for the (T_2^R, T_3^R) pair.

In the following results sections, we use multidimensional scaling (MDS) (Mardia, 1978) to embed our distance metrics into Euclidean spaces of two dimensions for ease of visualization and interpretation. We set the two axes of each two-dimensional MDS plot to have the same scale for correct interpretations (Holmes and Huber, 2019). The MDS representation of trees is useful for exploring large sets of trees, but it has also been recognized to have some limitations (Hillis et al., 2005; Chakerian and Holmes, 2012). For this reason, in addition to MDS plots, we include additional summaries to assess the behavior of our metrics. Initially, we assume that our trees are isochronous, that is, all samples are obtained at time 0.

Separation between ranked tree shapes with different distributions

To assess how our proposed metrics differentiate ranked tree shapes sampled from distributions with different degrees of balance, we simulated 1000 ranked tree shapes with $n = 100$ leaves under the beta-splitting model for each β-value in $\{-1.9, -1.5, -1, 0, 100\}$ representing a sequence from unbalanced to balanced. From the resulting 5000 total simulated ranked tree shapes, we computed the 5000×5000 pairwise distance matrices with d_1, d_2, $d_{\text{BHV-RTS}}$, $d_{\text{CP-RTS}}$ and $d_{\text{KC-RTS}}$ distances. The MDS plots are depicted in Figure 6. Each dot corresponds to a tree, and each color corresponds to the sampling distribution for a specified β value. Figure 6(C) shows the L_2-medoid trees using our d_1 distance. The total distance explained by the first two MDS dimensions are 91.4% using d_1 (Figure 6(A)) and 94.4% using d_2 (Figure 6(B)). The 2-dimensional MDS mappings using the other distances explain less than 80%. Our distances visually discriminate trees with different balance distributions to a greater extent than $d_{\text{BHV-RTS}}$ and $d_{\text{KC-RTS}}$; $d_{\text{CP-RTS}}$ shows similar performance.

In Figure S5, we show the confusion matrices for the trees plotted in Figure 6. For $i, j \in \{1, \ldots, 5\}$, each (i, j)-th entry of the matrix represents the percentage of trees simulated from the i-th distribution that are closest to the L_2-medoid of the j-th distribution, where closeness is measured by the distance function. As an indication of overall separation, we average the diagonal entries of the confusion matrices. A larger value indicates better separation of the corresponding distance. Matrices in Figure S5 confirm the greater discrimination for d_1 and d_2: about 83% of the trees are closest to the L_2-medoid of their originating distribution with d_1 and d_2 distances, 70.5% with $d_{\text{KC-RTS}}$, 75.0% with $d_{\text{CP-RTS}}$, and only 20.3% with $d_{\text{BHV-RTS}}$.

Table S1(A) shows the dispersion statistic (Section S4.1) for each tree distribution and for each distance function. The group of unbalanced trees (green in Figure 6) is the group with the least dispersion according to d_1, d_2, and $d_{\text{BHV-RTS}}$ but not according to the other three metrics. The MDS figures (Figure 6) accord with Table S1: both distances $d_{\text{KC-RTS}}$ and $d_{\text{CP-RTS}}$ show a linear trend that increases dispersion with degrees of unbalancedness. This pattern is particularly evident in Figure 6(E).

We next simulated trees from the alpha-beta splitting model (Maliet et al., 2018). We generated 1000 random trees with $n = 100$ for each α in $\{-2, -1, 0, 1, 2\}$, producing 5000 total trees. By varying the value of α while keeping the balance parameter (β) fixed, we test the performance of our metrics in distinguishing between ranked tree shapes with small family sizes closer to the root ($\alpha < 0$) and ranked tree shapes with small family sizes closer to the tips ($\alpha > 0$). Our two-dimensional MDS representations are shown in Figure 7 and their corresponding confusion matrices in Figure S6. Our metrics, d_1 (Figure 7(A)) and d_2 (Figure 7(B)), when compared to the other three metrics, distinguish to a greater extent between trees with positive and negative

α parameters. More than 75% of the trees are closest to the L_2-medoid of their originating distributions with d_1 and d_2, and less than 35% have this property with the other distances.

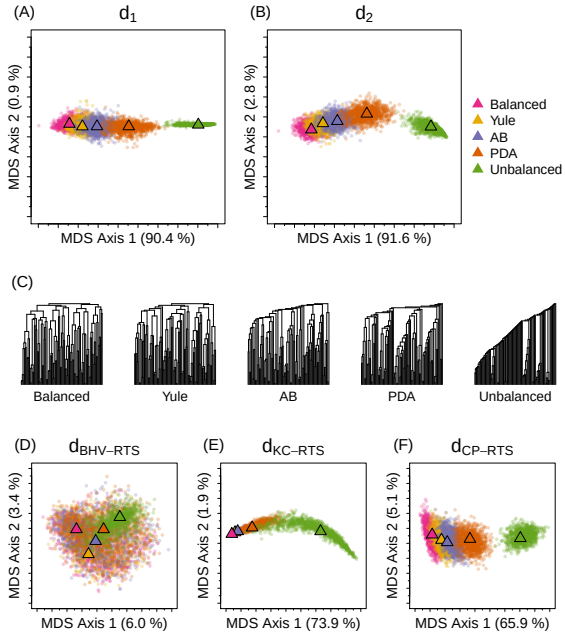


Figure 6: **MDS representation of distances between 5000 simulated isochronous ranked tree shapes of $n = 100$ leaves, aggregated from five different beta-splitting models.** 1000 isochronous ranked tree shapes were simulated from each of the following models: the balanced model ($\beta = 100$), the Yule model ($\beta = 0$), the Aldous branching (AB) model ($\beta = -1$), the proportional-to-distinguishable-arrangements (PDA) model ($\beta = -1.5$), and the unbalanced model ($\beta = -1.9$). (A) MDS of the d_1 metric. (B) MDS of the d_2 metric. (C) L_2-medoid trees from each distribution using the d_1 metric. MDS plots for (D) $d_{\text{BHV-RTS}}$, (E) $d_{\text{KC-RTS}}$, and (F) $d_{\text{CP-RTS}}$. In each MDS plot, the triangle represents the L_2-medoid tree of 1000 points for a specified model.

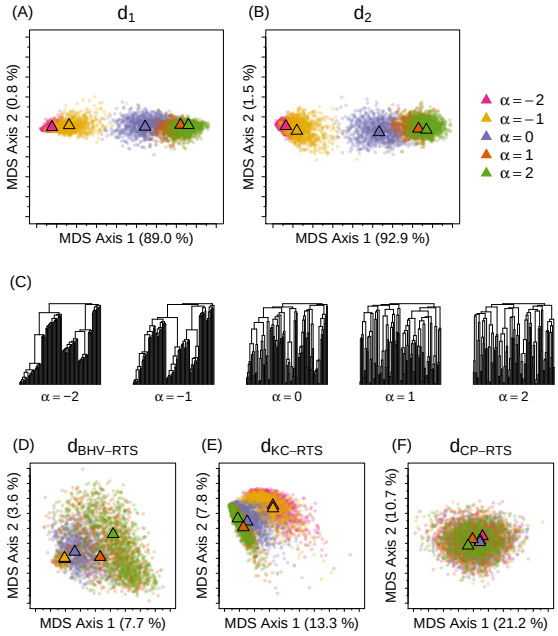


Figure 7: **MDS representation of distances between 5000 simulated isochronous ranked tree shapes of $n = 100$ leaves, aggregated from five different alpha-beta splitting models.** 1000 isochronous ranked tree shapes were simulated for each of α values in $\{-2, -1, 0, 1, 2\}$. Different α generates different distributions of internal node ranking while keeping the same tree shape distribution at $\beta = 0$. (A) MDS of the d_1 metric. (B) MDS of the d_2 metric. (C) L_2-medoid trees from each distribution using the d_1 metric. MDS plots for (D) $d_{\text{BHV-RTS}}$, (E) $d_{\text{KC-RTS}}$, and (F) $d_{\text{CP-RTS}}$. In each MDS plot, the triangle represents the L_2-medoid tree of 1000 points for a specified model.

The first two dimensions in MDS explain more than 90% of the total d_1 and d_2 distances, whereas the other metrics explain less than 35%. Although the MDS visualizations in Figures 7(A) and (B) show that most clusters of trees are well separated according to their sampling distributions with d_1 and d_2, a large overlap exists between groups of trees with $\alpha = 1$ and $\alpha = 2$. The observed similarity is evident from the L_2-medoid trees (Figure 7C). Figure S6 confirms these conclusions obtained by visually inspecting the MDS plots of Figure 7. We note that $d_{\text{KC-RTS}}$ has better performance than $d_{\text{BHV-RTS}}$ and $d_{\text{CP-RTS}}$ in discriminating simulated trees from the alpha-beta distributions with different α values.

The dispersion statistics in Table S1(B) indicate that trees from the alpha-beta splitting model with $\alpha = 0$ are the most dispersed group according to our d_1 and d_2 metrics. This, however, is not the case with the other distances. In particular, no variations in dispersion across different distributions are observed using $d_{\text{CP-RTS}}$. $d_{\text{BHV-RTS}}$ shows a correlated trend between α and the dispersion of the corresponding distribution: trees from a distribution with large α have higher dispersion than trees with small α.

Our results confirm that our metrics capture both shape and rankings more effectively than the other adapted metrics on ranked tree shapes in that simulation models differing in these features are more easily discriminated with our distances than with the others. In addition, our d_1 and d_2 distances show good embedding in two-dimensional MDS, with a high proportion of the total distance explained.

Separation between genealogies with different branch length distributions

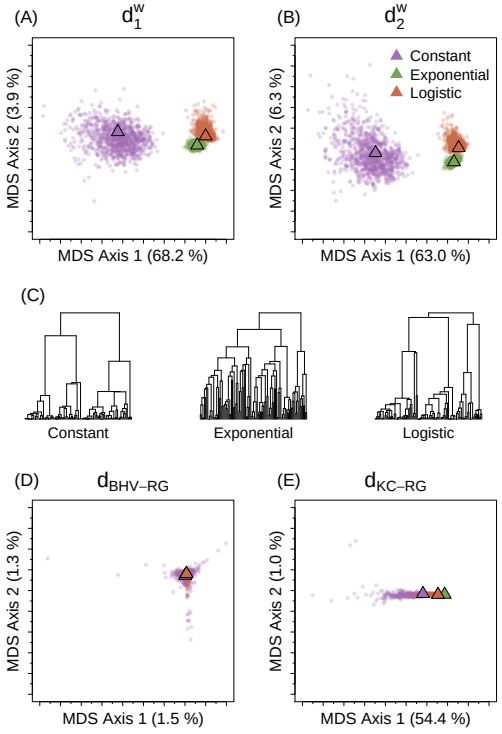


Figure 8: **MDS representation of distances between 3000 simulated isochronous ranked genealogies of $n = 100$ leaves under different demographic models.** The number of simulated genealogies per population model is 1000. (A) d_1^w metric, (B) d_2^w metric, (C) L_2-medoid genealogies from each distribution using the d_1^w metric, (D) $d_{\text{BHV-RG}}$, and (E) $d_{\text{KC-RG}}$.

When a population undergoes changes in population size $\lambda(t)$, the branch length distribution of the ranked genealogies depends on the population history $\lambda(t)$ as shown in Eq. 3. Under the neutral coalescent, the tree topology distribution in Eq. 1 is independent of the branch lengths. To investigate the performance of our weighted metrics d_1^w or d_2^w on genealogies in separating trees according to their branch length distribution, we generated 1000 random ranked genealogies with $n = 100$ leaves from the Tajima coalescent with each of the following demographic scenarios:

1. Constant effective population size: $\lambda(t) = N_0$;
2. Exponential growth: $\lambda(t) = N_0 e^{-0.01t}$;

3. Seasonal logistic trajectory:

$$\lambda(t) = \begin{cases} 0.1N_0 + \dfrac{0.9N_0}{1+\exp\left[6-2(t \bmod 12)\right]}, & (t \bmod 12) \leq 6 \\ 0.1N_0 + \dfrac{0.9N_0}{1+\exp\left[-18+2(t \bmod 12)\right]}, & (t \bmod 12) > 6 \end{cases}.$$

The coalescent trees under each specified population trajectory were simulated with $N_0 = 10000$. The functional form chosen and parameter values for the seasonal logistic trajectory mimic estimated trajectories of Human Influenza A virus in temperate regions (Vijaykrishna et al., 2015). We removed the leaf labels and retained branch lengths to produce isochronous ranked genealogies. We produced 3000×3000 pairwise distance matrices using the weighted metrics d_1^w and d_2^w.

Figures 8(A) and (B) show that our weighted metrics distinguish the three different demographic scenarios. The two-dimensional MDS accounts for 72.1% of the total distance, with the first MDS coordinate separating the uniform trajectory from the others, and the second coordinate distinguishing the exponential trajectory from the logistic trajectory.

A comparison of panels within Figure 8 shows that $d_{\text{BHV-RG}}$ is far less informative than any of our metrics. A cline differentiating the three distributions is somewhat present using $d_{\text{KC-RG}}$ (Figure 8(E)), far less so than in Figures 8(A) and (B). Our metrics' ability to distinguish different ranked genealogy distributions is also apparent in the confusion matrix in Figure S7, where our metrics display near perfect performance. Across all three distributions, 99.8% of the trees are closest to the L_2-medoid of their originating distribution with d_1^w and d_2^w; corresponding values are 33.4% with $d_{\text{BHV-RG}}$ and 74.0% with $d_{\text{KC-RG}}$.

Separation between heterochronous ranked tree shapes with different sampling events

To demonstrate that our metrics are sensitive to sampling schedules, we simulated trees with $n = 100$ leaves under heterochronous sampling with two sampling scenarios. For the first set of trees, we selected 90 samples at time 0 and the remaining 10 samples at distinct times with sampling times drawn uniformly at random from $(0, 10^4]$, i.e., $\mathbf{n} = (90, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$. In the second set of trees, 80 samples were drawn at time 0 and the remaining 20 were sampled in pairs at ten distinct random sampling times drawn uniformly from $(0, 10^4]$, i.e., $\mathbf{n} = (80, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2)$. We generated 1000 coalescent trees per sampling scheme assuming a constant effective population size trajectory of $N_0 = 10000$. We then removed leaf labels to produce the 2000×2000 distance matrices with all applicable distances.

The resulting MDS visualizations are displayed in Figures 9. Our metrics and $d_{\text{BHV-RTS}}$ show a clear separation between the two distributions along the first two MDS axes. The total distance explained in the two-dimensional space is higher using our metrics, 66.5% and 76.4% for d_1 and d_2, respectively, compared to 13.5% of $d_{\text{BHV-RTS}}$. $d_{\text{KC-RTS}}$ and $d_{\text{CP-RTS}}$ exhibit less discrimination than the other three distances. The confusion matrices in Figure S8 confirm that our metrics distinguish different sampling schemes better than the other three metrics compared.

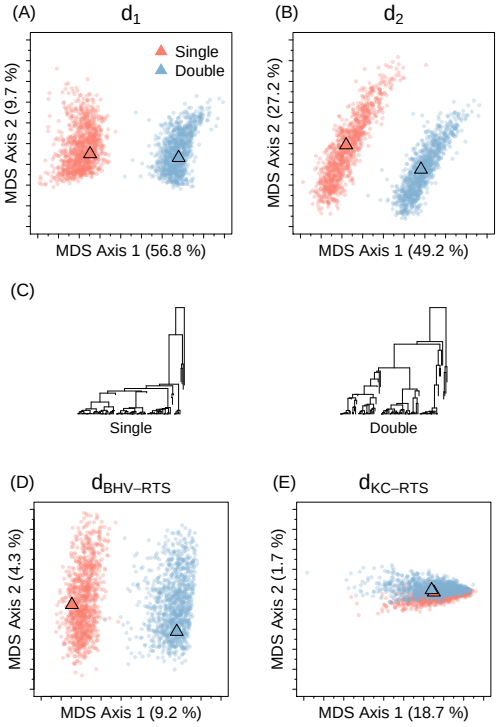


Figure 9: **MDS representation of distances between 2000 simulated heterochronous ranked tree shapes of $n = 100$ with different sampling events.** In the "Single" distribution, a single sample is drawn at each of the ten sampling events after the initial sampling event at time 0 with 90 samples: $\mathbf{n} = (90, 1, 1, 1, 1, 1, 1, 1, 1, 1)$. In the "Double" distribution, a pair of samples is drawn at each of the ten sampling events after the initial sampling event at time 0 with 80 samples: $\mathbf{n} = (80, 2, 2, 2, 2, 2, 2, 2, 2, 2)$. The sampling times of the ten sampling events after $t = 0$ are selected uniformly at random from $(0, 10^4]$ in both distributions. The resulting simulated trees with $n = 100$ taxa are from the neutral coalescent model with constant effective population size of 10^4. (A) d_1 metric, (B) d_2 metric, (C) L_2-medoid trees from each distribution using the d_1 metric, (D) $d_{\text{BHV-RTS}}$, (E) $d_{\text{KC-RTS}}$, and (F) $d_{\text{CP-RTS}}$.

Application to simulated influenza data

Before applying our metric to human influenza data, we examined our metric on simulated heterochronous genealogies that mimic hypothesized human influenza population size trajectories in temperate and tropical regions. In temperate climates, such as New York, influenza epidemics display seasonality, with peaks occurring during the winter months, whereas in regions with tropical or subtropical climates, influenza activity is persistent throughout the year (Tamerius et al., 2013).

For our hypothesized temperate region influenza dynamics, we assumed the seasonal logistic trajectory with $N_0 = 100$. We adopted preferential sampling (Karcher et al., 2016) to model probabilistic dependence of sampling times on effective population size by drawing sampling times from an inhomogeneous Poisson process with intensity proportional to the effective population size. The lower and upper sampling time bounds were set to 0 and 48, respectively. The resulting sampling event had $n = 104$ total samples. Parameter values were chosen following (Karcher et al., 2016) to mimic realistic Influenza A trajectories. Given the generated sampling event vectors $\mathbf{s}$ and $\mathbf{n}$, we generated 1000 random coalescent trees for the temperate region.

In tropical regions, influenza activity is more stable throughout the year and hence, we assumed a constant trajectory with $N_0 = 100$. We simulated 1000 random coalescent trees with sampling times of $n = 104$ samples randomly drawn from a uniform-$[0, 48]$ distribution.

We constructed 2000×2000 distance matrices with each of the metrics applicable for ranked genealogies: d_1^w, d_2^w, $d_{\text{BHV-RG}}$, and $d_{\text{KC-RG}}$. Figure 10 shows that the first two MDS components with our metrics d_1^w and d_2^w explain 34.7% and 39.2% of the total distance, respectively, with the first component clearly separating

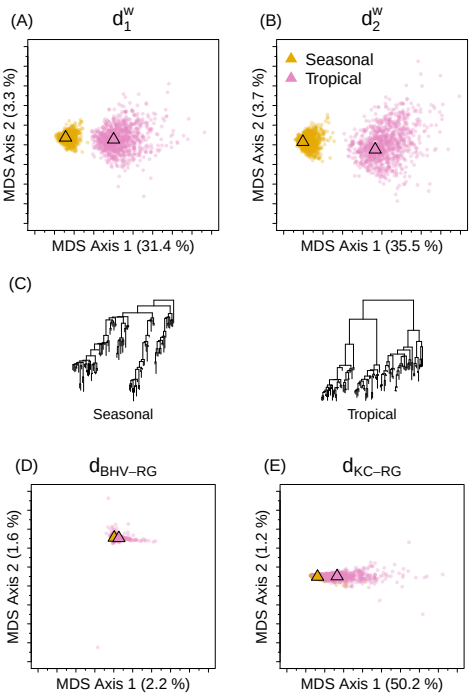


Figure 10: **MDS representation of distances between 2000 simulated heterochronous ranked genealogies of $n = 104$ leaves under different demographic models and sampling events.** The "Seasonal" distribution represents the population dynamics of influenza in the temperate region, which is modeled with logistic trajectory and preferential sampling reflecting the seasonality of the influenza dynamics. The "Tropical" distribution represents the population dynamics of influenza in the tropical region, which is modeled with a constant population trajectory and uniform sampling. (A) d_1^w metric, (B) d_2^w metric, (C) L_2-medoid trees from each distribution using the d_1^w metric, (D) $d_{\text{BHV-RG}}$, and (E) $d_{\text{KC-RG}}$.

between seasonal and tropical distributions. $d_{\text{BHV-RG}}$ do not distinguish the models to the same extent.

Although the $d_{\text{KC-RG}}$ metric shows greater variation in the first MDS axis than with our metrics, $d_{\text{KC-RG}}$ does not distinguish between the two distributions. The lack of separation explained by the first MDS axis in Figure 10(E) can be attributed to the large dispersion in the sampling-time distribution of the tropical-region simulated genealogies. This claim is supported by Figure S9, where ranked genealogies from the tropical distribution are approximately equally likely to be close to the L_2-medoid points of the seasonal and tropical distributions with $d_{\text{KC-RG}}$. By contrast, our metrics show near perfect diagonal entries. In particular, under the d_2^w metric, every tree is closer to the L_2-medoid tree of its own distribution than to the other L_2-medoid trees. Across two distributions, 99.7% of the trees are closest to the L_2-medoid of their originating distribution with d_1^w and 100% with d_2^w. Corresponding numbers are 51.0% with $d_{\text{BHV-RG}}$ and 74.1% with $d_{\text{KC-RG}}$.

Table S2(B) shows the dispersion statistics: as expected, all distances reflect higher dispersion in the tropical data than the dispersion observed in the seasonal data.

Analysis of human influenza A virus from different continents

We next apply our metrics to ranked tree shapes and ranked genealogies sampled from the posterior distributions of human influenza A/H3N2 genealogies from 3 geographic regions—New York, Chile, and Singapore—collected from January 2014 to January 2019 in the Global Initiative on Sharing All Influenza Data's (GISAID) EpiFlu database. The frequency of the collected samples by collection date and the estimated effective population size trajectories with the `BEAST` Bayesian Skygrid method are displayed in Figure S11 in Supplementary Material.

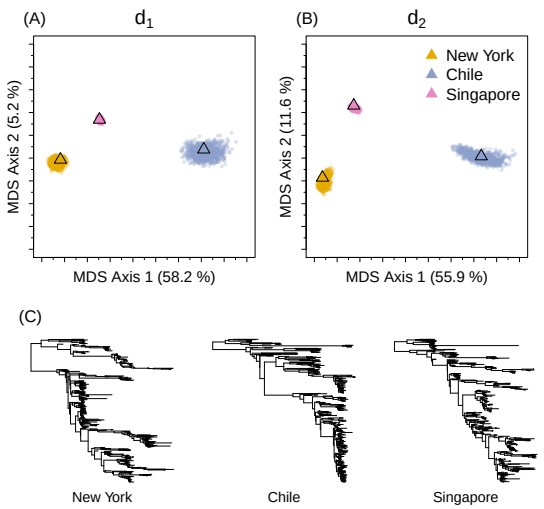


Figure 11: **MDS representation of distances between 3000 heterochronous ranked tree shapes sampled from three posterior distributions (1000 trees from each distribution). Observed data consist of 410 sequences of human influenza A virus from each geographic region: Singapore, New York and Chile.** (A) d_1 metric, (B) d_2 metric, and (C) L_2-medoid trees from each distribution using the d_1 metric.

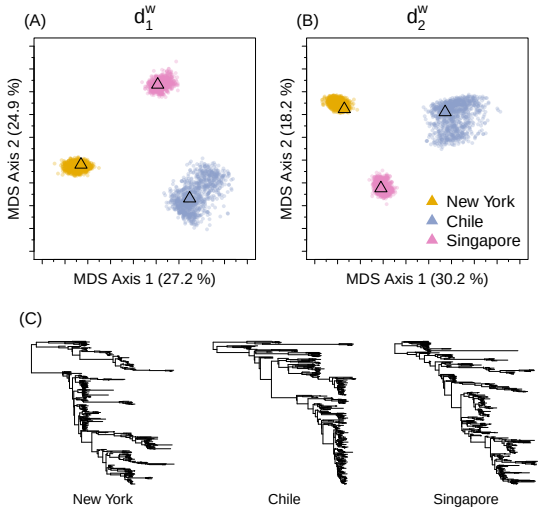


Figure 12: **MDS representation of distances between 3000 heterochronous ranked genealogies sampled from three posterior distributions (1000 trees from each distribution). Observed data consist of 410 sequences of human influenza A virus from each geographic region: Singapore, New York and Chile.** (A) d_1^w metric, (B) d_2^w metric, and (C) L_2-medoid trees from each distribution using the d_1^w metric.

For each of the three regions, we sampled 1000 trees from the posterior distribution with BEAST. We then computed our unweighted and weighted distance matrices of 3000×3000 pairwise distances. Figures 11(A) and (B) show the corresponding MDS plots of the ranked tree shapes using our d_1 and d_2 metrics. All three distributions are well separated in the two-dimensional MDS plots. The first MDS axis separates Singapore from the rest, and the second MDS axis splits Chile from the other two ranked tree shape distributions. Figures 11(A) and (B) show that the three tree distributions have varying degrees of dispersion. The trees from Singapore are more clustered together than the trees from the other two regions. Similarly, the trees from New York are more clustered together than the trees from Chile. The dispersion statistics in Table S2(C) quantify the observed differences in variation among different tree posterior distributions.

The same patterns of dispersion are observed with the weighted metrics d_1^w and d_2^w of ranked genealogies (Figure 12). The dispersion observed in the MDS plots reflects the high variance in branch lengths.

Discussion

Ranked tree shapes and ranked genealogies are used to model the dependencies between samples of molecular data in the fields of population genetics and phylogenetics. These tree structures have a particular value in studying heterochronous data and in providing a resolution that is fine enough to be informative but is computationally more feasible than finer resolutions. Although many other distances are defined on trees,

they are primarily defined on spaces of labeled tree topologies or tree-shape structures without ranking of internal nodes. The applications of such distances are limited when comparing detailed hierarchical tree structures from different sets of individuals with little or no overlap.

In this manuscript, we equipped the spaces of ranked tree shapes and ranked genealogies with a metric. By defining a bijection between ranked tree shapes and a triangular matrix of integers, we defined distance functions between ranked tree shapes as the L_1 and L_2 norms of the difference between two matrices. We extended our metrics to heterochronously sampled ranked tree shapes and genealogies. Our distances on ranked genealogies are the L_1 and L_2 norms of the difference between two matrices whose elements are the products of the corresponding entries of the ranked tree shape matrix representation and weights given by the genealogy branch lengths. To apply our distances to real data for comparing ranked genealogies or ranked tree shapes with different sampling events, we proposed an augmentation scheme, increasing the dimension of the matrix representations, and we defined our distances as L_1 and L_2 norms of the difference between extended matrices of the same size.

We compared our distance to projections of the BHV (Billera et al., 2001), Colijn-Plazotta (Colijn and Plazzotta, 2018), and Kendall-Colijn (Kendall and Colijn, 2016) distances on the spaces of ranked tree shapes and genealogies (when applicable). The comparison showed that our distances perform better than the other metrics in separating ranked tree shapes from distributions with different balance parameters (Figure 6), different family size compositions with respect to time (Figure 7), and different branch length distributions (Figure 8).

Our proposed distances do not equally penalize all types of tree differences. The analyses showed that our distances penalize coalescence-type changes—whether the coalescence is between two unlabeled leaves, a leaf and an internal (ranked) branch, or between two internal (ranked) branches— more than ranking changes. Events that change the tree shape incur a higher penalty. In addition, our metric penalizes changes near the root more than near the bottom of the trees. These attributes accord with desirable properties for making biologically meaningful metrics.

We used our distances to summarize the distribution of trees in 2-dimensional MDS representations and propose using the L_2-medoid tree, a sample version of the Fréchet mean, as the central point of a sample of trees. We proposed a sample version of the Fréchet variance to quantify uncertainty. In our simulations, the L_2-medoid trees summarized the center of the distributions, and in some cases, they looked similar to the maximal clade credibility trees. Despite this, however, the geometry of the spaces induced with our metrics is still unknown, and the mathematical properties of such summary statistics require further exploration.

Our metrics provide the basis for a decision-theoretic statistical inference that can be constructed by finding the best-ranked genealogy that minimizes the expected error or loss function, which, in turn, is a function of the tree distance. Further, our proposed metric provides a tool for evaluating convergence and the mixing of Markov chain Monte Carlo procedures on ranked genealogies.

Our proposed distances inherit the properties of L_1 and L_2 distances of symmetric positive definite (SPD) matrices. In fact, our matrix representation can be modified to a SPD matrix by replacing the upper matrix triangle with the transposed entries of the original lower triangular matrix. In this case, our new distance will be twice the original distance and will not change the observed properties. This view of SPD matrix

representation of ranked tree shapes makes it possible to explore new distances (Vemulapalli and Jacobs, 2015) on ranked tree shapes and ranked genealogies.

Materials and Methods

Metric spaces on heterochronous trees with different numbers of sampling events

We extend our distances to include cases in which the numbers of sampling events differ but the total number of samples remain the same.

Consider two heterochronous ranked tree shapes of n leaves, T_1^R and T_2^R, with different numbers of sampling events m_1 and m_2, respectively. In order to compute the distance between T_1^R and T_2^R with our metrics, we require the two ranked tree shapes to be represented as $\mathbf{F}$-matrices of the same dimension. We accomplish this by inserting artificial sampling events. The detailed steps are as follows. Note that the following formulation is done going backwards in time with time increasing from the present to the past.

For $i = 1, 2$, let $\mathbf{E}^{(i)} = (e^{(i)}_{m_i+n-1}, \ldots, e^{(i)}_1)$ be the vector of ordered sampling and coalescence events where $e^{(i)}_{m_i+n-1}$ denotes the most recent sample event ($e^{(i)}_{m_i+n-1} = s$) assumed to occur at time $u^{(i)}_{m_i+n-1} = 0$. $e^{(i)}_1 = c$ denotes the coalescent event at time $u^{(i)}_1$ corresponding to the most recent common ancestor of the samples in T_i^R. Each $e^{(i)}_j$ is either a sampling event ($e^{(i)}_j = s$) or a coalescent event ($e^{(i)}_j = c$). The event of type c occurs $n-1$ times and the event of type s occurs m_i times in $\mathbf{E}^{(i)}$. In the example illustrated in Figure S4, the event vectors for T_1^R (Figure S4(A)) and T_2^R (Figure S4(B)) are $\mathbf{E}^{(1)} = (s, c, s, c, s, c)$ and $\mathbf{E}^{(2)} = (s, s, s, c, s, c, c)$, respectively.

We first align all $n-1$ coalescent events between the two trees by adding empty spaces when needed as depicted in Figure S4(C). Once all the type-c events are aligned, we next align the sampling events between two successive coalescent events or between $t = 0$ and the first c event. When aligning the events of type s between the two trees, we pair the events of type s, starting from the most recent events. The event vector alignment is demonstrated in Figure S4(C). If one tree has more type-s events than the other in a given intercoalescent interval, we insert the excess artificial sampling events, denoted by a's, in the tree with fewer type-s events in that interval. We assign 0 new samples to type-a events. For example, in the tree of Figure S4(A), there is only one sampling event before the first coalescent event, whereas there are three sampling events in the tree of Figure S4(B). In Figure S4(D), we illustrate how the two artificial sampling events are added to the first tree in the first interval. The resulting augmented trees are shown in Figures S4(E) and (F) along with their corresponding $\mathbf{F}$-matrix representations in Figures S4(G) and (H). We note that by construction, $\mathbf{F}^{(1)} \neq \mathbf{F}^{(2)}$ in these cases. Hence $d_i(T_1^R, T_2^R) \neq 0$ for $i = 1, 2$.

Consider now two heterochronous ranked genealogies of n leaves, $\mathbf{g}_1^R$ and $\mathbf{g}_2^R$, with different number of sampling events m_1 and m_2 respectively. In order to compute the distance between $\mathbf{g}_1^R$ and $\mathbf{g}_2^R$ we first augment their $\mathbf{F}$-matrix representations as with ranked tree shapes. In addition, we augment their weight matrices $W^{(1)}$ and $W^{(2)}$ by assigning a time to each augmented artificial sampling event. If n_a artificial events are inserted between events $e^{(i)}_{j+1}$ and $e^{(i)}_j$, we subdivide the corresponding time interval $[u^{(i)}_{j+1}, u^{(i)}_j]$ into $n_a + 1$ intervals with equal length: the times assigned to the n_a augmented artificial events are $\{u^{(i)}_{j+1} + \Delta, u^{(i)}_{j+1} + 2\Delta, \ldots, u^{(i)}_{j+1} + n_a\Delta\}$, where $\Delta = \frac{u^{(i)}_j - u^{(i)}_{j+1}}{n_a+1}$.

Unique labeling scheme of ranked tree shapes

In order to adapt other distances defined on labeled trees to ranked tree shapes, we use the following labeling scheme. We start by labeling the leaves that descend directly from the internal node with the largest rank n. If the node has two direct descendent leaves with different edge lengths, we label the longer leaf ℓ_1 and the shorter leaf ℓ_2. If the two leaves have the same edge lengths, we label them ℓ_1 and ℓ_2 from left to right. If the node has only one direct descendent leaf, we label it ℓ_1. If there is no descending leaf, no labeling is done. We then move to the node with rank $n-1$ and continue labeling the leaves by traversing through the internal nodes in descending order of rank until all leaves are labeled. If the current node with rank k has two direct descendent leaves and if the last assigned leaf label is ℓ_j, we label the leaf with the longer edge ℓ_{j+1} and the leaf with shorter edge ℓ_{j+2}; if the leaves have the same edge lengths, we label the pair of leaves ℓ_{j+1} (left) and ℓ_{j+2} (right). If the node k has only one direct descendent leaf, we label it ℓ_{j+1}. If the node k has no direct descendent leaf, no label is assigned. Examples demonstrating our unique labeling scheme are in Figure 13.

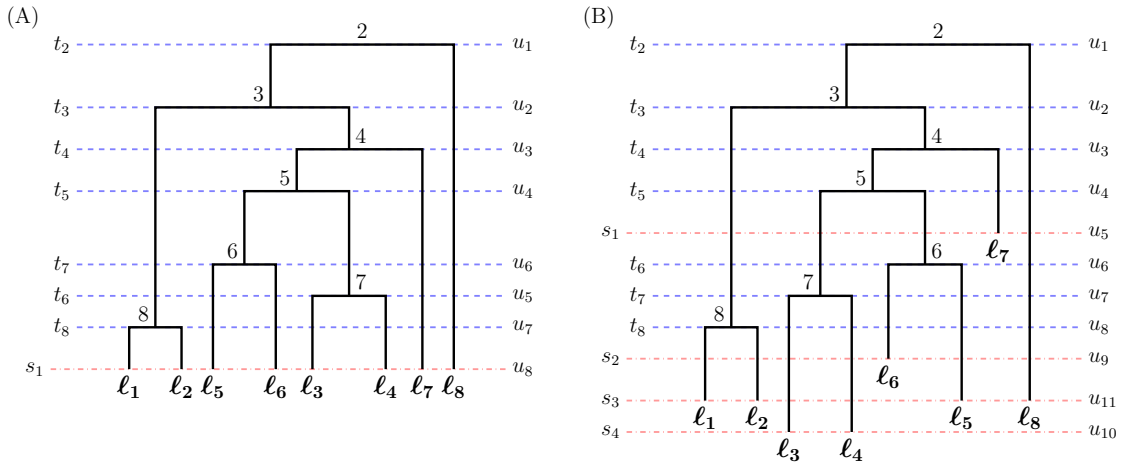


Figure 13: **Unique labeling of ranked tree shapes and ranked genealogies.** (A) Example of the unique labeling of a ranked genealogy with isochronous sampling. (B) Example of the unique labeling of a ranked genealogy with heterochronous sampling.

Analysis of human influenza A virus from different continents

We selected all available sequences collected from January 2014 to January 2019 with complete HA segments in the GISAID EpiFlu database: 438, 410, and 510 sequences from New York, Chile, and Singapore, respectively. In order to compare trees with the same numbers of leaves, we kept 410 randomly selected sequences. We aligned the sequences using the `FFT-NS-i` option of `MAFFT` v7.427 software (Katoh et al., 2002; Katoh and Standley, 2013).

We used `BEAST` v1.10.4 (Suchard et al., 2018) to sample from the posterior distribution of variable effective population size trajectories and ranked and labeled genealogies with the following settings: SRD06 codon-partition substitution model (Shapiro et al., 2005) to accommodate rate heterogeneity among sites, strict molecular clock, and a Skygrid model with a regular grid of 100 points for inferring the effective population

size trajectory (Gill et al., 2013). We ran the MCMC chain with 10^7 steps and thinned every 10^4.

Software

- The source code for our distance metrics is available at `https://github.com/JuliaPalacios/phylodyn`.
- To perform MDS, we used the `cmdscale` function in R package `stats`.
- For our adaptation to the BHV metric, we used the Geodesic Treepath Problem (GTP) software implemented in Java (Owen and Provan, 2011).
- For the KC metric, we used `multiDist` implemented in the R package `treespace` (Kendall and Colijn, 2016).
- For the CP distance, we used the `vecMultiDistUnlab` in the R package `treetop` (Colijn and Plazzotta, 2018) with default parameters.
- To generated random isochronous ranked tree shapes, we used the `simulate_tree` function in the R package `apTreeshape` (Maliet et al., 2018) with the following parameters. In the beta-distribution simulation, $\beta \in \{-1.9, -1.5, -1, 0, 100\}$, $\alpha = 1$, $\epsilon = 0.001$, and $\eta = 1$. In the alpha-beta distribution simulation, $\alpha \in \{-2, -1, 0, 1, 2\}$, $\beta = 0$, $\epsilon = 0.001$ and $\eta = 1$.
- To generate random isochronous coalescent trees with different branch length distributions, we used the `rcoal` function from the `ape` package in R.
- To generate random heterochronous coalescent trees with different sampling events or different branch length distributions, we used the `coalsim` function implemented in `phylodyn` package in R.
- The preferential sampling of the hypothesized temperate region influenza dynamics was performed using the `pref_sample` function in `phylodyn` (Karcher et al., 2017).

Acknowledgments

We would like to acknowledge Nina Miolane and Susan Holmes for useful discussion of distances and other embedding techniques. J.A.P. and N.A.R. acknowledge support from National Institutes of Health grant R01-GM-131404. J.A.P. acknowledges support from the Alfred P. Sloan Foundation.

References

Aldous D, 1996. Probability distributions on cladograms. In Aldous D, Pemantle R, editors, Random Discrete Structures, pages 1–18, New York, NY. Springer New York.

Aldous DJ, 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statistical Science, 16(1):23–34.

Allen BL, Steel M, 2001. Subtree transfer operations and their induced metrics on evolutionary trees. Annals of Combinatorics, 5(1):1–15.

Bacák M, 2014. Computing medians and means in Hadamard spaces. SIAM Journal on Optimization, 24(3):1542–1566.

Billera LJ, Holmes SP, Vogtmann K, 2001. Geometry of the space of phylogenetic trees. Advances in Applied Mathematics, 27(4):733–767.

Blum MGB, François O, 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. Systematic Biology, 55(4):685–691.

Bordewich M, Semple C, 2005. On the computational complexity of the rooted subtree prune and regraft distance. Annals of Combinatorics, 8(4):409–423.

Bourgain J, 1985. On Lipschitz embedding of finite metric spaces in Hilbert space. Israel Journal of Mathematics, 52(1-2):46–52.

Brown DG, Owen M, 2019. Mean and variance of phylogenetic trees. Systematic Biology, syz041.

Cappello L, Palacios JA, 2019. Sequential importance sampling for multi-resolution Kingman-Tajima coalescent counting. arXiv:1902.05527 [stat.AP], .

Chakerian J, Holmes S, 2012. Computational tools for evaluating phylogenetic and hierarchical clustering trees. Journal of Computational and Graphical Statistics, 21(3):581–599.

Colijn C, Plazzotta G, 2018. A metric on phylogenetic tree shapes. Systematic Biology, 67(1):113–126.

Felsenstein J, Rodrigo AG, 1999. Coalescent approaches to HIV population genetics. In Crandall KA, editor, The Evolution of HIV, pages 233–272. Johns Hopkins University Press, Baltimore, Maryland.

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

Ford DJ, 2005. Probabilities on cladograms: introduction to the alpha model.

Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA, 2013. Improving Bayesian population dynamics inference: A coalescent-based model for multiple loci. Molecular Biology and Evolution, 30(3):713–724.

Harding EF, 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Advances in Applied Probability, 3(1):44–77.

Hillis DM, Heath TA, John KS, 2005. Analysis and visualization of tree space. Systematic Biology, 54 3:471–82.

Holmes S, 2003. Statistics for phylogenetic trees. Theoretical Population Biology, 63(1):17–32.

Holmes S, Huber W, 2019. Modern Statistics for Modern Biology, chapter 9, pages 217–248. Cambridge University Press, Cambridge, United Kingdom.

Karcher MD, Palacios JA, Bedford T, Suchard MA, Minin VN, 2016. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. PLOS Computational Biology, 12(3):e1004789.

Karcher MD, Palacios JA, Lan S, Minin VN, 2017. phylodyn: an R package for phylodynamic simulation and inference. Molecular Ecology Resources, 17(1):96–100.

Katoh K, Misawa K, Kuma Ki, Miyata T, 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14):3059–3066.

Katoh K, Standley DM, 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution, 30(4):772–780.

Kendall M, Colijn C, 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. Molecular Biology and Evolution, 33(10):2735–2743.

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

Lambert A, Stadler T, 2013. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. Theoretical Population Biology, 90:113–128.

Lewitus E, Morlon H, 2015. Characterizing and comparing phylogenies from their Laplacian spectrum. Systematic Biology, 65(3):495–507.

Li M, Zhang L, 1999. Twist–rotation transformations of binary trees and arithmetic expressions. Journal of Algorithms, 32(2):155–166.

Maliet O, Gascuel F, Lambert A, 2018. Ranked tree shapes, nonrandom extinctions, and the loss of phylogenetic diversity. Systematic Biology, 67(6):1025–1040.

Mardia K, 1978. Some properties of clasical multi-dimesional scaling. Communications in Statistics - Theory and Methods, 7(13):1233–1241.

Mooers AO, Heard SB, 1997. Inferring evolutionary process from phylogenetic tree shape. The Quarterly Review of Biology, 72(1):31–54.

Owen M, Provan JS, 2011. A fast algorithm for computing geodesic distances in tree space. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(1):2–13.

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

Poon AFY, Walker LW, Murray H, McCloskey RM, Harrigan PR, Liang RH, 2013. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. PLOS ONE, 8(11):1–11.

Robinson D, Foulds L, 1981. Comparison of phylogenetic trees. Mathematical Biosciences, 53(1):131–147.

Robinson DF, Foulds LR, 1979. Comparison of weighted labelled trees. In Horadam AF, Wallis WD, editors, Combinatorial Mathematics VI, pages 119–126, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sainudiin R, Stadler T, Véber A, 2015. Finding the best resolution for the Kingman-Tajima coalescent: theory and applications. Journal of Mathematical Biology, 70(6):1207–1247.

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

Shapiro B, Rambaut A, Drummond AJ, 2005. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Molecular Biology and Evolution, 23(1):7–9.

Slatkin M, Hudson R, 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics, 129(2):555–562.

Steel M, 2016. Phylogeny: Discrete and Random Processes in Evolution. SIAM-Society for Industrial and Applied Mathematics, Philadelphia.

Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A, 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evolution, 4(1):vey016.

Tajima F, 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics, 105(2):437–460.

Tamerius JD, Shaman J, Alonso WJ, Bloom-Feshbach K, Uejio CK, Comrie A, Viboud C, 2013. Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. PLOS Pathogens, 9(3):e1003194.

Vemulapalli R, Jacobs DW, 2015. Riemannian metric learning for symmetric positive definite matrices. arXiv:1501.02393 [cs.CV], .

Vijaykrishna D, Holmes EC, Joseph U, Fourment M, Su YC, Halpin R, Lee RT, Deng YM, Gunalan V, Lin X, et al., 2015. The contrasting phylodynamics of human influenza B viruses. eLife, 4:e05055.

Wakeley J, 2009. Coalescent Theory: An Introduction. Roberts & Company Publishers, Greenwood Village, Colorado.

Willis A, Bell R, 2018. Confidence sets for phylogenetic trees. Journal of Computational and Graphical Statistics, 27(525):542–552.

Yule GU, 1925. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. Philosophical Transactions of the Royal Society of London. Series B, 213:21–87.

Supplementary Material

S1 Alpha-Beta splitting model by Maliet et al. (2018)

Algorithm 1 Simulation of a labeled ranked tree shape according to alpha-beta splitting model

Draw $u_1, \ldots, u_n \sim U(0,1)$
Set $i = 1$, $r_0 = 0$
while $i < n$ **do**
    Draw $R_i \sim \text{Beta}(1+\beta, 1+\beta)$
    Let $r_1, \ldots, r_i$ be the ordered permutation of $R_1, \ldots, R_i$ such that $r_1 < r_2 < \ldots < r_i$
    Let $y_j = \begin{cases} 1 & \text{if } \sum_{k=1}^{n} 1\left(u_k \in (r_{j-1}, r_j)\right) > 1 \\ 0 & \text{o.w} \end{cases}$, for $j = 1, \ldots, i$
    if $\sum_{j=1}^{i} y_j > 1$ **then**
        Pick y_k w.p. $\frac{(r_k - r_{k-1})^{\alpha y_k}}{\sum_{j=1}^{i}(r_j - r_{j-1})^{\alpha y_j}}$. The partition defined by the u_j's in (r_{k-1}, r_k) is chosen to bifurcate with ranking i.
        $i = i + 1$

S2 Proof of Theorem 1

Proof. Consider a ranked tree shape T^R with n leaves sampled at m different sampling times. We denote the total number of change points in T^R by $r = n+m-1$ and its ordered change point times by $(u_r, u_{r-1}, \ldots, u_1)$, $0 = u_r < u_{r-1} < \cdots < u_1$, with time increasing into the past. The internal nodes of T^R are labeled by the indices of their coalescent times, and all leaves of T^R are labeled by the indices of their sampling times. We note that for convenience, internal nodes are no longer labeled $2, \ldots, n$ from the root to leaves, but they are labeled by their time-event indices (see Figure S1). Each internal node has a unique label, but the leaf nodes with the same sampling time share the same label. We define $N = \{1, \ldots, r\}$ to be a set of all node labels, I to be a set of all internal node labels, and S to be a set of all leaf node labels. Note that I and S are disjoint and contain $n-1$ and m elements, respectively.

For $i \in I$, let $o_i = (x_{i,1}, x_{i,2})$ denote the ordered pair of labels of the two immediate descendants of internal node i, such that $i < x_{i,1} \leq x_{i,2}$. We denote the set of all pairs i and $o_i = (x_{i,1}, x_{i,2})$ choices in T^R by $X = \{(i, o_i) \mid i \in I\}$. Then X completely specifies T^R: T^R is a directed graph from the root to tips and X encodes its adjacency matrix and the order of the internal node indices $i \in I$ determines internal node rankings.

We define a function $\phi : X \to \{0,1,2\}^{r-1}$ as follows:

$$\phi_k(i, o_i) = \begin{cases} 0 & \text{if } 1 \leq k < i \\ 2 & \text{if } i \leq k < x_{i,1} \\ 1 & \text{if } x_{i,1} \leq k < x_{i,2} \\ 0 & \text{if } x_{i,2} \leq k < r. \end{cases}$$

The kth element of $\phi(i, o_i)$ is the number of immediate descendants of an internal node i present at the

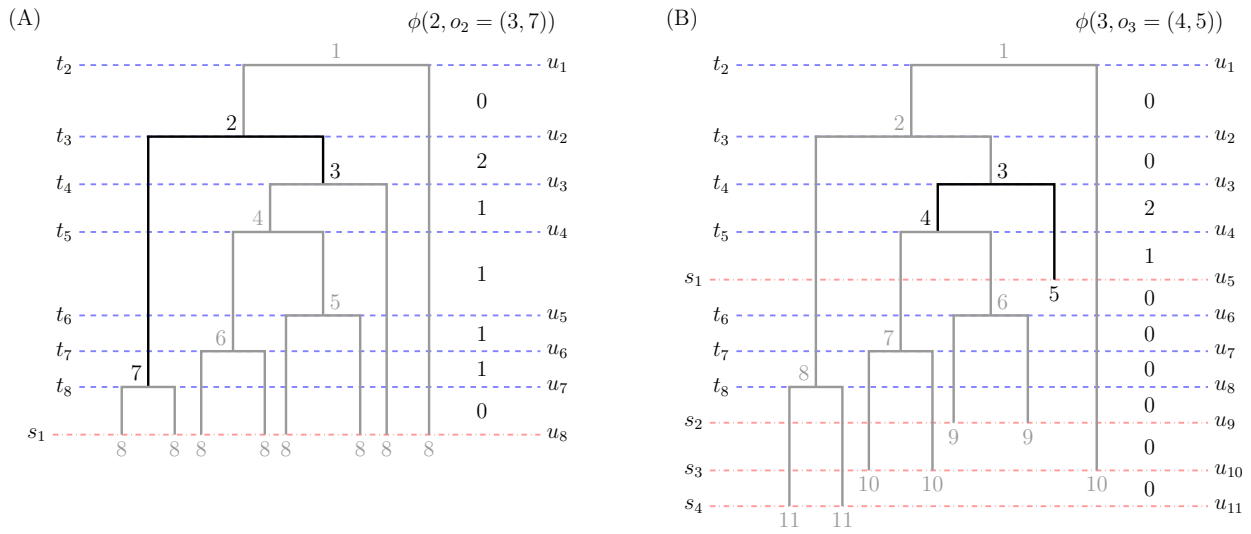


Figure S1: **Example of ϕ mapping.** (A) An example isochronous ranked tree shape. The set of internal node labels is $I = \{1,2,3,4,5,6,7\}$ and the set of leaf node labels is $S = \{8\}$. For convenience, internal nodes are labeled by their time-event indices throughout the proof. The internal node with label 2 at time u_2 has descendant nodes 3 and 7 at time u_3 and u_7, respectively ($o_2 = (3,7)$). The column vector $\phi(2, o_2 = (3,7)) = (0,2,1,1,1,1,0)$ indicates the number of direct descendants of node 2 at each change point time interval. (B) An example heterochronous ranked tree shape. The set of internal node labels is $I = \{1,2,3,4,6,7,8\}$ and the set of leaf node labels is $S = \{5,9,10,11\}$. The internal node with label 3 at time u_3 has descendants node 4 and node 5 at time u_4 and u_5, respectively ($o_3 = (4,5)$). The column vector $\phi(3, o_3 = (4,5)) = (0,0,2,1,0,0,0,0,0,0)$ indicates the number of direct descendants of node 3 at each change point time interval.

time interval (u_{k+1}, u_k). ϕ is an injective map. To prove this, let $(s, o_s), (t, o_t)$ be two elements in X and let $(s, o_s) \neq (t, o_t)$. Because internal nodes of T^R are ranked, $(s, o_s) \neq (t, o_t)$ implies $s \neq t$; without loss of generality, assume $s < t$. By the definition of the map ϕ, the sth element of $\phi(s, o_s)$ is $\phi_s(s, o_s) = 2$, while the sth element of $\phi(t, o_t)$ is $\phi_s(t, o_t) = 0$ because $s < t$ and $t < x_{t,1} \leq x_{t,2}$. Thus, $\phi(s, o_s) \neq \phi(t, o_t)$ and ϕ is injective.

Let $\eta : \{1, \ldots, r-1\} \times \{0,1,2\}^{r-1} \to \{0,1,2\}^{r-1}$ such that for $\mathbf{y} \in \{0,1,2\}^{r-1}$, the j-th element of η is

$$\eta(k, \mathbf{y})_j = \begin{cases} 0 & \text{if } 1 \leq j < k \\ y_j & \text{if } k \leq j < r. \end{cases}$$

That is, $\eta(k, \mathbf{y})$ sets all the first $k-1$ entry values of $\mathbf{y}$ to 0. Note that the first $i-1$ elements of $\phi(i, o_i)$ are 0 by definition and thus, $\eta(i, \phi(i, o_i)) = \phi(i, o_i)$.

Finally, for $T^R \in \mathcal{T}_n^R$, define $\psi : \mathcal{T}_n^R \to M_{r-1,r-1}(\mathbb{R})$, a function that maps a ranked tree shape with n leaves to a real valued square matrix of size $r-1$, with k-th column:

$$\psi(T^R)_{\cdot,k} = \sum_{\substack{i \in I, \\ i \leq k}} \eta(k, \phi(i, o_i)),$$

where $\psi(T^R)_{\cdot,k}$ indicates the kth column of $\psi(T^R)$ and I is the set of all internal node labels as defined at the beginning of this section. By the definition of η, the first $k-1$ values of the kth column $\psi(T^R)_{\cdot,k}$ are 0, i.e., $\psi(T^R)$ is a lower triangular matrix. Because ϕ records the number of immediate descendants of a single internal node present at each time interval, $\psi(T^R)_{\cdot,k}$ tracks the sum of all surviving immediate descendants of internal nodes with labels $i \leq k$ starting from time interval (u_{k+1}, u_k); thus, $\psi(T^R)_{s,k}$, with $k \leq s$, represents the number of lineages of T^R in (u_{k+1}, u_k) that are still present at the time interval (u_{s+1}, u_s).

We prove that ψ is an injective map. Let $T_1^R, T_2^R \in \mathcal{T}_n^R$ and $T_1^R \neq T_2^R$. Because $X = \{(i, o_i) \mid i \in I\}$ completely specifies T^R, $T_1^R \neq T_2^R$ implies that there exists an index $\ell \in \{1, \ldots, n-1\}$ such that $(i_\ell^{(1)}, o_{i_\ell}^{(1)}) \neq (i_\ell^{(2)}, o_{i_\ell}^{(2)})$. Here, i_ℓ indicates the ℓth element of I sorted in increasing order. If there is more than one such index, choose ℓ to be the smallest of them. Without loss of generality, let $i_\ell^{(1)} \leq i_\ell^{(2)}$. Then $\psi(T_1^R)_{\cdot,i_\ell^{(1)}} \neq \psi(T_2^R)_{\cdot,i_\ell^{(1)}}$ and thus $\psi(T_1^R) \neq \psi(T_2^R)$.

Hence, ψ maps each ranked tree shape T^R to a unique matrix, i.e., given an $\mathbf{F}$-matrix, if it encodes a ranked tree shape, it encodes exactly one ranked tree shape.

□

S3 Proof of Proposition 2

Proof. The non-negativity and symmetry are trivial. The triangle inequality follows from the Minkowski inequality of L_1 and L_2 norms. It remains to prove the identity property: $d_k^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = 0$ if and only if $\mathbf{g}_1^R = \mathbf{g}_2^R$ for $k = 1, 2$. It is clear that $d_k^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = 0$ if $\mathbf{g}_1^R = \mathbf{g}_2^R$ so we focus on $\mathbf{g}_1^R = \mathbf{g}_2^R$ if $d_k^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = 0$. The following proof is for d_1^w. The proof for d_2^w follows the same arguments.

We assume that the two genealogies have the same number of sampling events m and same number of leaves n, so that the $\mathbf{F}$-matrices of $\mathbf{g}_1^R$ and $\mathbf{g}_2^R$ have the same dimension $(n+m-2) \times (n+m-2)$ dimension. We define $r = n + m - 2$ for notational simplicity.

Because we allow only a single event at each change time point u_i, either coalescent or sampling, the first column of any $\mathbf{F}$-matrix is $(2, 1, \ldots, 1)$ or $(2, 1, \ldots, 1, 0, \ldots, 0)$. For the latter, we denote the row index of the last occurrence of 1 in the first column by k_1: $F_{k_1,1}^{(\ell)} = 1$ and $F_{k_1+1,1}^{(\ell)} = 0$ for some index $2 \leq k_1 \leq r$ and $\ell = 1, 2$.

If $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ have different first columns, then for some index $k_1 \geq 2$, $(F_{k_1,1}^{(1)}, F_{k_1,1}^{(2)}) = (0, 1)$ or $(F_{k_1,1}^{(1)}, F_{k_1,1}^{(2)}) = (1, 0)$. Because $\left| F_{i,j}^{(1)} W_{i,j}^{(1)} - F_{i,j}^{(2)} W_{i,j}^{(2)} \right| \geq 0$, $d_1^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = 0$ implies $F_{i,j}^{(1)} W_{i,j}^{(1)} = F_{i,j}^{(2)} W_{i,j}^{(2)}$ for all i, j. Therefore, $W_{k_1,1}^{(2)} = 0$ in the first case and $W_{k_1,1}^{(1)} = 0$ in the second case. However, this contradicts our assumption of positive time interval between two change points, and thus $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$ must have the same first column.

If both $\mathbf{F}$-matrices share the same first column, then $d_1^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = 0$ implies $W_{i,1}^{(1)} = W_{i,1}^{(2)}$ for all $i = 1, \ldots, r$. Recalling $W_{i,j} = u_j - u_{i+1}$, we have $u_1^{(1)} - u_{i+1}^{(1)} = u_1^{(2)} - u_{i+1}^{(2)}$. Because we assume $u_{r+1}^{(1)} = u_{r+1}^{(2)} = 0$, we can traverse through i in decreasing order starting from $i = r$ to get $u_i^{(1)} = u_i^{(2)}$ for all $i = 1, \ldots, r+1$, which gives $\mathbf{W}^{(1)} = \mathbf{W}^{(2)}$.

Along with $\mathbf{W}^{(1)} = \mathbf{W}^{(2)}$, $F_{i,j}^{(1)} W_{i,j}^{(1)} = F_{i,j}^{(2)} W_{i,j}^{(2)}$ implies $F_{i,j}^{(1)} = F_{i,j}^{(2)}$ for all i, j, i.e., $\mathbf{F}^{(1)} = \mathbf{F}^{(2)}$, and thus $\mathbf{g}_1^R = \mathbf{g}_2^R$. □

S4 Embeddings in MDS

We chose MDS in two dimensions to visualize matrices of pairwise distances. In general, our metrics are well explained in the MDS visualization; however, the other distances are usually poorly represented in this space. In this section, we propose a measure of distortion and correlation to assess how well the embedding preserves the pairwise distances for each metric. In our examples, the distortion measure shown in Table S3 suggests that our d_2 metric has the best MDS embedding in general of all distance functions considered with our d_1 metric a close second. Similarly, the correlation measure shown in Table S4 confirms that our d_1 and d_2 metrics, with near perfect correlations, have far better embedding in the 2-dimensional MDS space than the other distances considered.

S4.1 Distortion

To assess our distances and its MDS embedding in two dimensions, we compute the following distortion statistic (Bourgain, 1985) defined as follows:

$$\text{distortion} = \text{expansion} \times \text{contraction}.$$

where expansion and contraction are defined as follows. For a given sample of ranked tree shapes $\mathcal{T}_S = \{T_1^R, T_2^R, \ldots, T_s^R\}$ with n leaves in $\mathcal{T}_n^R$,

$$\text{expansion} = \max_{\substack{T_i^R, T_j^R \in \mathcal{T}_S; \\ i \neq j}} \frac{d_{\text{MDS}}(T_i^R, T_j^R)}{d(T_i^R, T_j^R)},$$

where d_{MDS} is the L_2-Euclidean distance in the reduced MDS space and d is any distance function on ranked tree shapes, and

$$\text{contraction} = \max_{\substack{T_i^R, T_j^R \in \mathcal{T}_S; \\ i \neq j}} \frac{d(T_i^R, T_j^R)}{d_{\text{MDS}}(T_i^R, T_j^R)}.$$

The distortion on the ranked genealogies is defined similarly. The comparison of distortions for simulated ranked tree shapes and ranked genealogies can be found in Table S3.

S4.2 Correlation

As a second measure for assessing our distances and its MDS embedding in two dimensions, we compute the Pearson correlation coefficient between the two vectors of pairwise distances between sampled ranked tree shapes, one from using any distance functions d on ranked tree shapes and the other from the L_2-Euclidean distance d_{MDS} in the reduced MDS space:

$$\text{correlation} = \frac{\sum_{i=2}^{s} \sum_{j=1}^{i} \left(d(T_i^R, T_j^R) - \mu_d\right)\left(d_{\text{MDS}}(T_i^R, T_j^R) - \mu_{d_{\text{MDS}}}\right)}{\sqrt{\sum_{i=2}^{s} \sum_{j=1}^{i} \left(d(T_i^R, T_j^R) - \mu_d\right)^2 \sum_{i=2}^{s} \sum_{j=1}^{i} \left(d_{\text{MDS}}(T_i^R, T_j^R) - \mu_{d_{\text{MDS}}}\right)^2}},$$

where s is the number of sampled ranked tree shapes. $\mu_{d_{\text{MDS}}}$ and μ_d are the mean of the pairwise distances using L_2-Euclidean distance in the MDS space and using any distance functions d on the sampled ranked tree shapes, respectively. The comparisons of correlations for simulated ranked tree shapes and ranked genealogies appear in Table S4.

Supplementary Figures

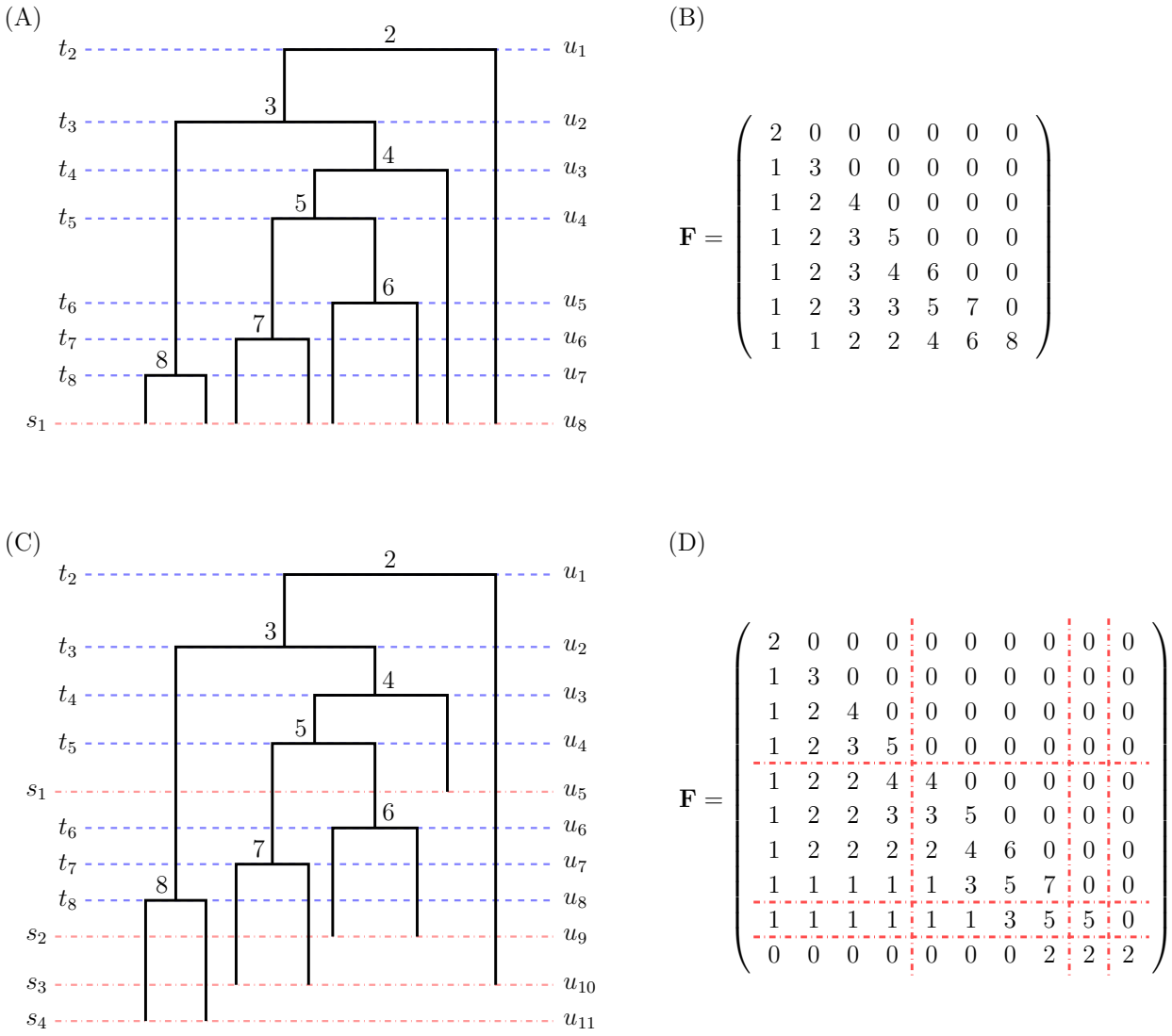


Figure S2: **Bijection of ranked tree shapes and F-matrices for isochronous and heterochronous trees**. (A) Example of a ranked genealogy with isochronous sampling. (B) The corresponding **F**-matrix that encodes the ranked tree shape information of the tree in (A). (C) Example of a ranked genealogy with heterochronous sampling. (D) The corresponding **F**-matrix of the heterochronous ranked tree shape in (C). Blue dotted lines indicate coalescent events and red dotted lines represent sampling events. In (C), coalescent times are denoted by $\{t_k\}_{k=2}^{8}$, sampling times by $\{s_k\}_{k=1}^{4}$, and the number of lineages changes at change points $\{u_k\}_{k=1}^{11}$.

$$\begin{pmatrix}
u_1-u_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_1-u_3 & u_2-u_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_1-u_4 & u_2-u_4 & u_3-u_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_1-u_5 & u_2-u_5 & u_3-u_5 & u_4-u_5 & 0 & 0 & 0 & 0 & 0 & 0 \\
u_1-u_6 & u_2-u_6 & u_3-u_6 & u_4-u_6 & u_5-u_6 & 0 & 0 & 0 & 0 & 0 \\
u_1-u_7 & u_2-u_7 & u_3-u_7 & u_4-u_7 & u_5-u_7 & u_6-u_7 & 0 & 0 & 0 & 0 \\
u_1-u_8 & u_2-u_8 & u_3-u_8 & u_4-u_8 & u_5-u_8 & u_6-u_8 & u_7-u_8 & 0 & 0 & 0 \\
u_1-u_9 & u_2-u_9 & u_3-u_9 & u_4-u_9 & u_5-u_9 & u_6-u_9 & u_7-u_9 & u_8-u_9 & 0 & 0 \\
u_1-u_{10} & u_2-u_{10} & u_3-u_{10} & u_4-u_{10} & u_5-u_{10} & u_6-u_{10} & u_7-u_{10} & u_8-u_{10} & u_9-u_{10} & 0 \\
u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_{10}
\end{pmatrix}$$

Figure S3: **Example of the weight matrix W.** The weight matrix associated with the example heterochronous ranked genealogy and its **F**-matrix in Figures S2(C) and (D). In the last row, u_{11} is suppressed because we set the initial sampling time to be $u_{11} = 0$.

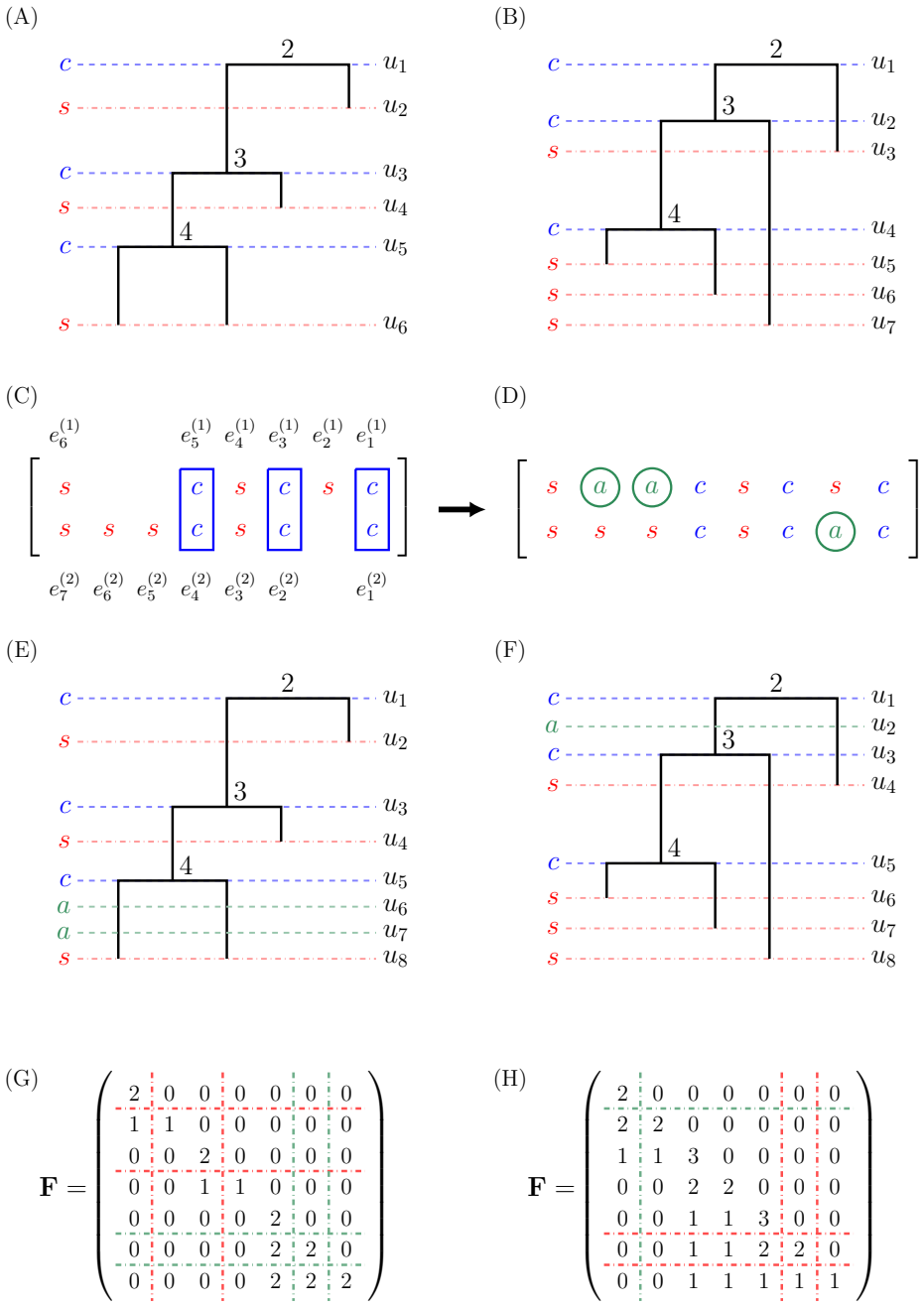


Figure S4: **Example of augmented F-matrix representation of ranked tree shapes.** In order to compute the d_1 or d_2 distances between ranked tree shapes with equally many samples but different numbers of sampling events, such as ranked tree shapes (A) and (B), we insert artificial sampling events with 0 samples in order to match their dimension. (A)-(B) Two ranked tree shapes of 4 samples with different sampling events. (C) Alignment of event vectors of two trees. The $n-1$ coalescent events are aligned first by matching ith coalescent event of a tree to the ith coalescent event of the other tree ($i = 1, \ldots, n-1$). The sampling events are then matched by increasing index order in the event vector. (D) Augmentation of artificial sampling events a between coalescent events or between the first sampling and the first coalescent event. (E)-(F) Augmented ranked tree shapes. (G)-(H) Augmented **F**-matrix representations.

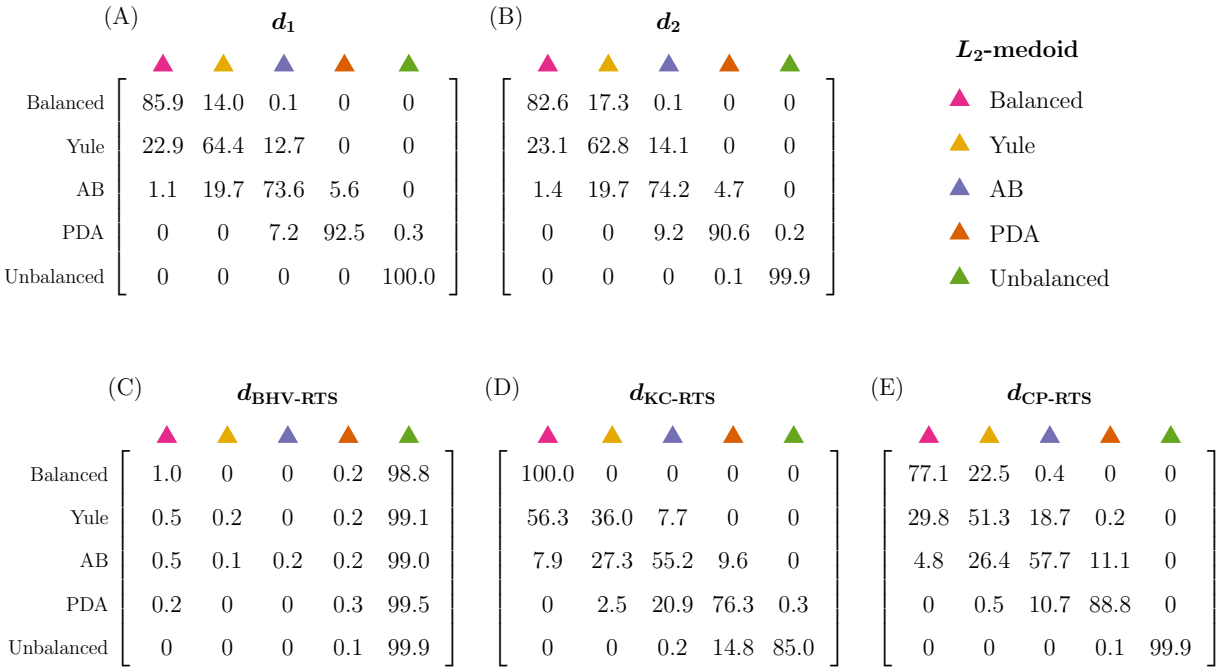


Figure S5: **Comparison of metrics: discrimination of isochronous ranked tree shapes under different beta-splitting models.** We compare the performance of different distances on ranked tree shapes according to how well they separate trees simulated from the beta-splitting distribution of ranked tree shapes with different balance parameters β. Rows indicate the sampling distribution and columns indicate the L_2-medoid of each distribution. Each matrix corresponds to a different distance metric. Entry (i, j) corresponds to the percentage of trees simulated from the i-th distribution that are closer to the j-th L_2-medoid than to the medoids of any other columns. The color scheme of the L_2-medoids follows Figure 6. The mean diagonal values are 83.28, 82.02, 20.32, 70.50, and 74.96 for matrices (A)-(E), respectively.

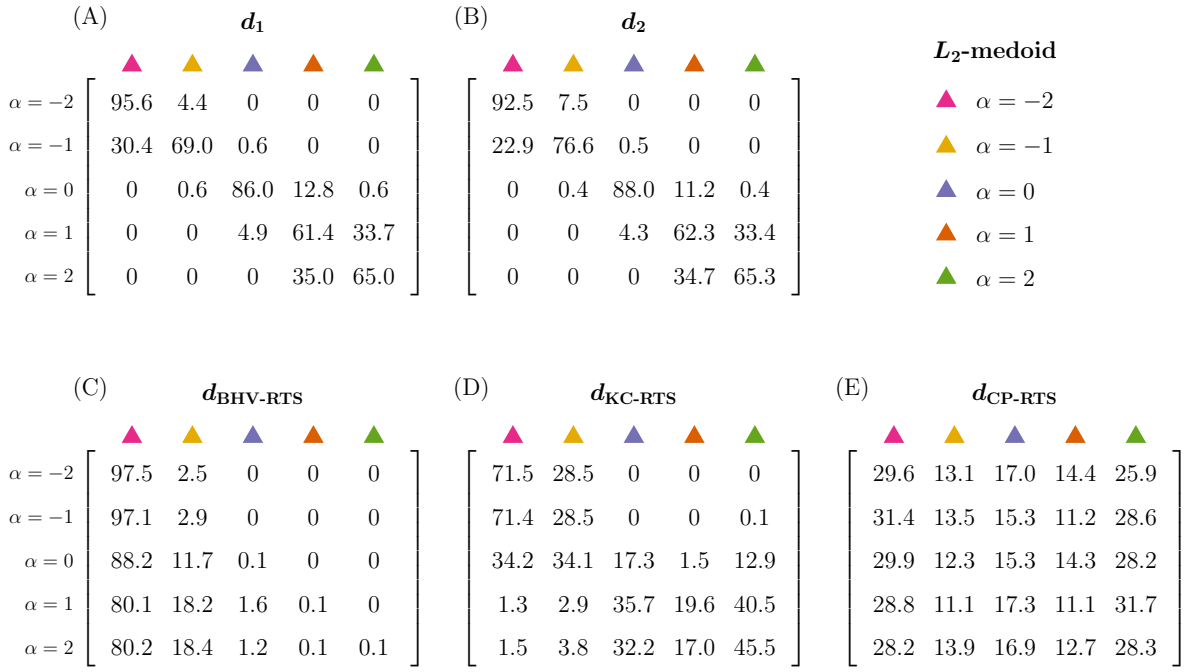


Figure S6: **Comparison of metrics: discrimination of isochronous ranked tree shapes under different alpha-beta splitting models.** We compare the performance of different distances on ranked tree shapes according to how well they separate trees simulated from the alpha-beta splitting distribution of ranked tree shapes with different parameter values α which regulates the internal node ranking of a given tree shape. The format of the matrices follows Figure S5. The simulation values and the color scheme of the L_2-medoids follow Figure 7. The mean diagonal values are 75.40, 76.94, 20.14, 36.48, and 19.56 for matrices (A)-(E), respectively.

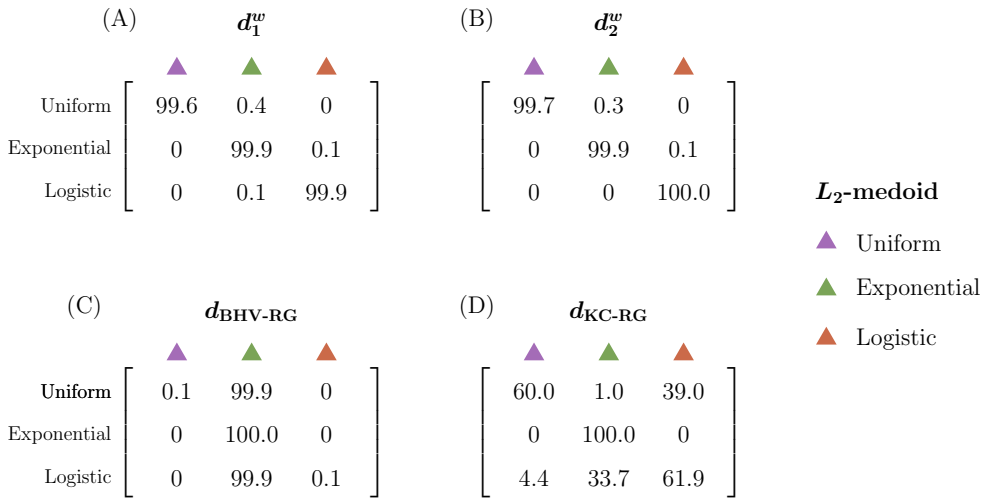


Figure S7: **Comparison of metrics: isochronous ranked genealogies under different demographic models.** We compare the performance of different distances on ranked genealogies according to how well they separate trees simulated from the $\lambda(t)$-coalescent with different population histories. The format of the matrices follows Figure S5. The simulation values and the color scheme of the L_2-medoids follow Figure 8. The mean diagonal values are 99.80, 99.87, 33.40, and 73.97 for matrices (A)-(D), respectively.

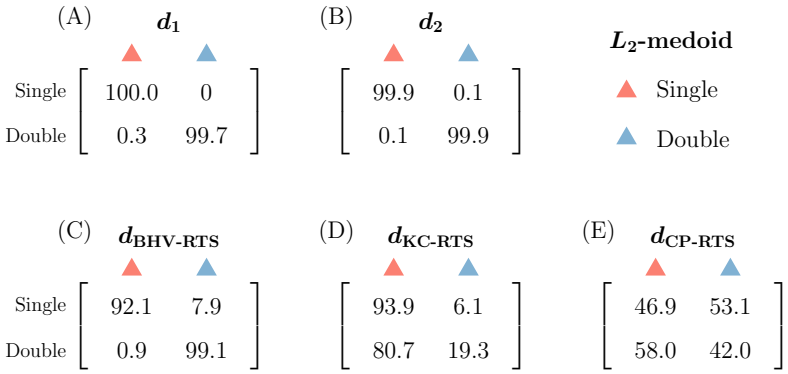


Figure S8: **Comparison of metrics: heterochronous ranked tree shapes with different sampling schemes.** We compare the performance of different distances on ranked genealogies according to how well they separate trees simulated from the heterochronous Tajima coalescent with different sampling sequences **s** and **n**. The format of the matrices follows Figure S5. The simulation values and the color scheme of the L_2-medoids follow Figure 9. The mean diagonal values are 99.85, 99.90, 95.60, 56.60, and 44.45 for matrices (A)-(E), respectively.

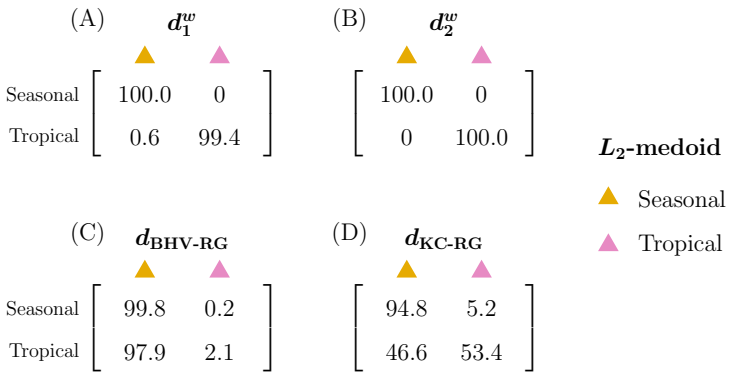


Figure S9: **Comparison of metrics: heterochronous ranked genealogies with different simulation models.** We compare the performance of different distances on ranked genealogies according to how well they separate trees simulated from the heterochronous $\lambda(t)$-coalescent with different population histories and sampling sequences **s** and **n**. The format of the matrices follows Figure S5. The simulation values and the color scheme of the L_2-medoids follow Figure 10. The mean diagonal values are 99.70, 100.00, 50.95, and 74.10 for matrices (A)-(D), respectively.

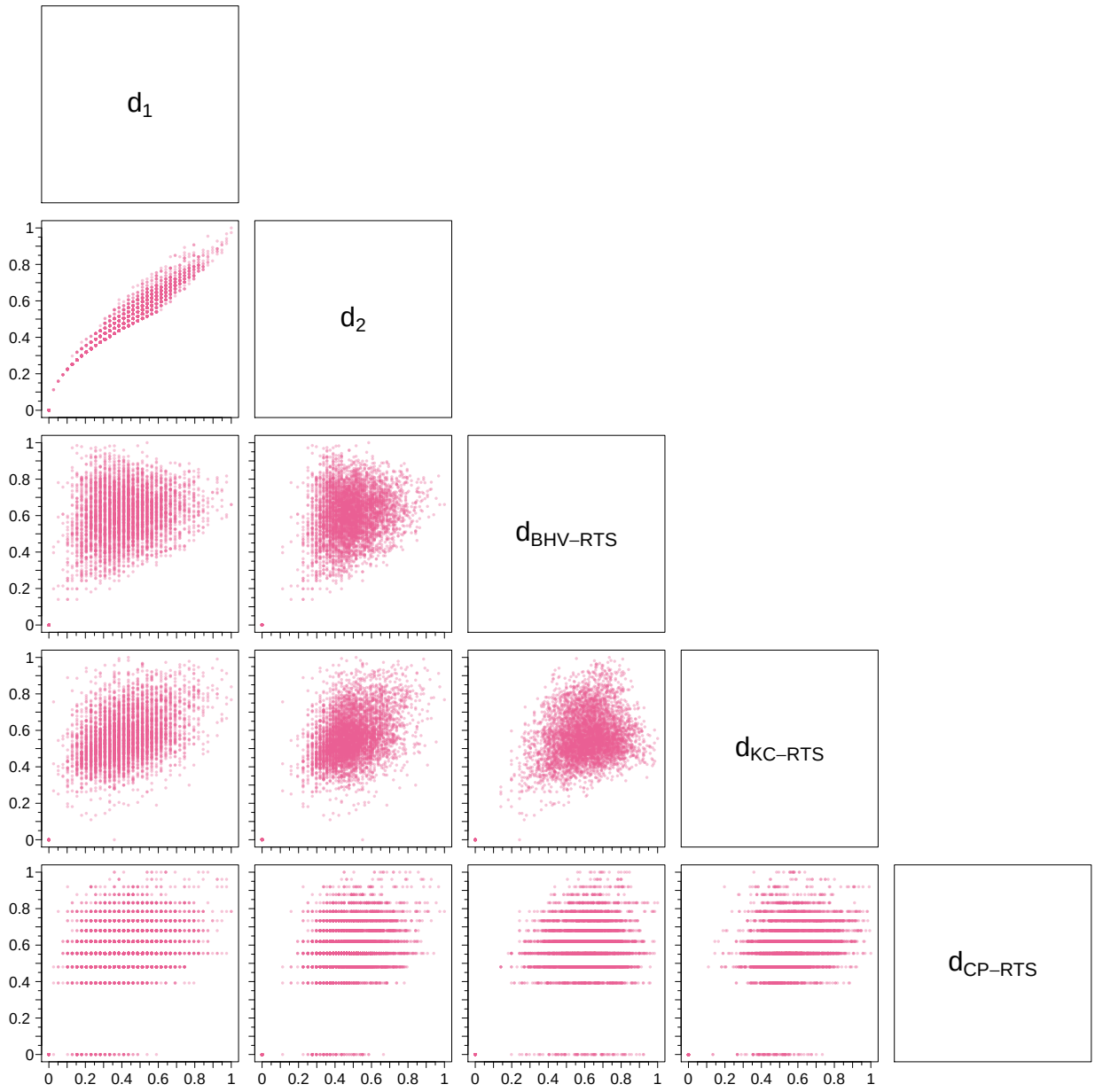


Figure S10: **Pairwise comparisons of metrics using 100 simulated ranked tree shapes with $n = 10$ leaves.** Each plot contains $\binom{100}{2}$ points, where each point represents two pairwise distances between simulated ranked tree shapes. $d_{\text{CP-RTS}}$ displays the lowest resolution among all the metrics plotted. Multiple pairs of different ranked tree shapes share the same $d_{\text{CP-RTS}}$ value while many of those pairs have distinct values using the other metrics.

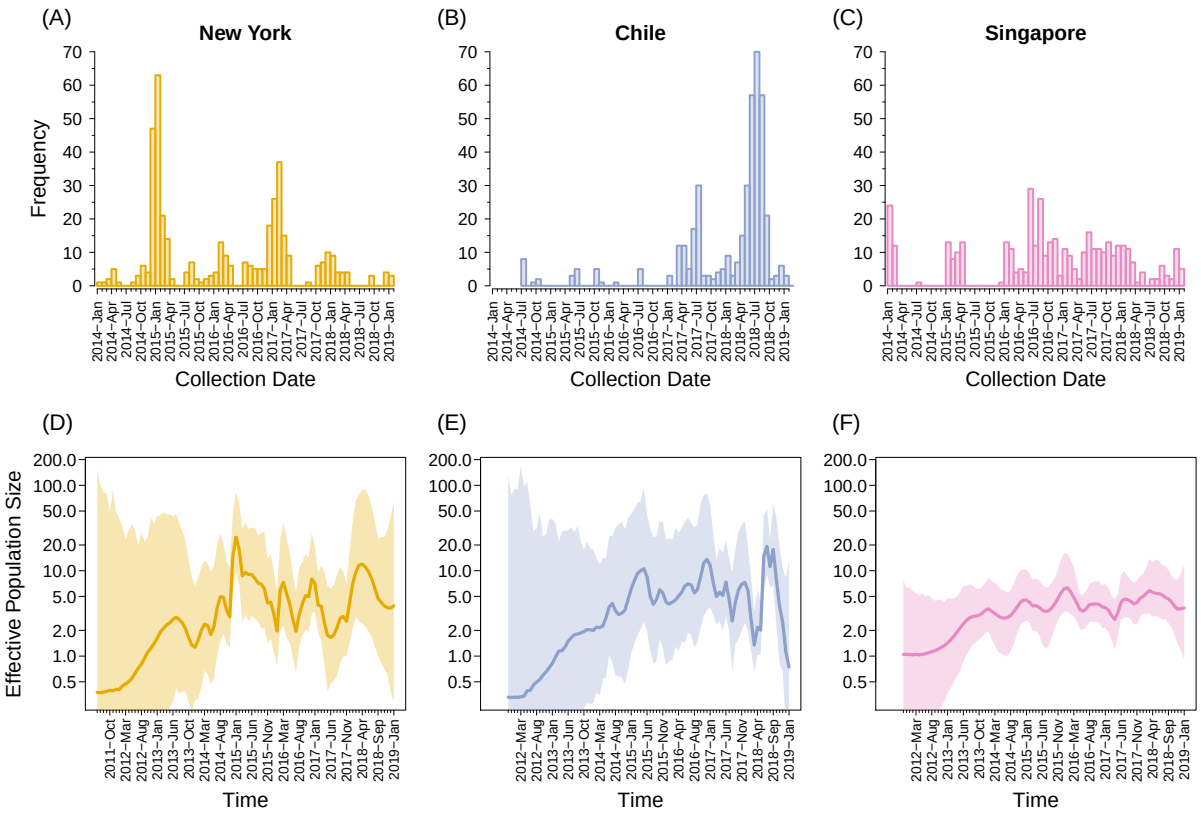


Figure S11: **Human influenza A virus collection dates and inferred effective population size trajectories.** (A) Collection date histogram, New York; (B) Collection date histogram, Chile; (C) Collection date histogram, Singapore; (D) Inferred population size trajectory with BEAST, New York; (E) Chile; and (F) Singapore. Data description and color scheme follow Figure 12.

Supplementary Tables

Table S1: **Summary of dispersion comparisons for ranked tree shapes.** Comparison of dispersion of ranked tree shape distribution (Section S4.1) using distance functions between ranked tree shapes. (A) Isochronous ranked tree shapes simulated from the beta-splitting model with varying β parameters (Figure 6). (B) Isochronous ranked tree shapes simulated from the alpha-beta splitting model with varying α parameters and fixed $\beta = 0$ (Figure 7). (C) Heterochronous ranked tree shapes simulated from different sampling schemes (Figure 9).

(A) Isochronous ranked tree shapes, beta-splitting model

	d_1	d_2	$d_{\text{BHV-RTS}}$	$d_{\text{KC-RTS}}$	$d_{\text{CP-RTS}}$
Balanced	5541.21	117.14	377.84	128.20	8.51
Yule	5889.31	119.40	363.33	182.52	8.97
AB	6579.90	127.37	345.82	271.78	9.58
PDA	6966.50	133.24	304.49	490.98	10.48
Unbalanced	4391.36	90.88	154.44	944.90	12.50

(B) Isochronous ranked tree shapes, alpha-beta splitting model

	d_1	d_2	$d_{\text{BHV-RTS}}$	$d_{\text{KC-RTS}}$	$d_{\text{CP-RTS}}$
$\alpha = -2$	4701.27	90.81	85.33	110.76	8.95
$\alpha = -1$	7202.89	135.70	110.64	126.06	9.00
$\alpha = 0$	7719.71	150.57	262.18	172.73	8.96
$\alpha = 1$	6084.73	122.77	363.17	182.49	9.06
$\alpha = 2$	5668.18	115.80	376.98	179.18	9.08

(C) Heterochronous ranked tree shapes

	d_1	d_2	$d_{\text{BHV-RTS}}$	$d_{\text{KC-RTS}}$	$d_{\text{CP-RTS}}$
Single	10359.43	277.88	350.26	230.17	9.47
Double	10423.86	238.90	324.47	257.65	9.62

Table S2: **Summary of dispersion comparisons for ranked genealogies.** Comparison of dispersion of ranked genealogies using distance functions between ranked genealogies. (A) Isochronous ranked genealogies simulated from different population trajectories under neutral coalescent model (Figure 8). (B) Heterochronous ranked genealogies simulated from different population trajectories under neutral coalescent model Figure 10). (C) Heterochronous ranked genealogies of human influenza A virus data from different continental groups (Figure 12).

(A) Isochronous ranked genealogies

	d_1	d_2	$d_{\text{BHV-RTS}}$	$d_{\text{KC-RTS}}$
Constant	6420771.00	142726.04	33007.42	391150.52
Exponential	1331068.00	24847.09	1397.17	4743.48
Logistic	1744504.00	39291.83	8751.45	109258.59

(B) Heterochronous ranked genealogies

	d_1	d_2	$d_{\text{BHV-RTS}}$	$d_{\text{KC-RTS}}$
Seasonal	98912.28	1074.80	105.49	1335.42
Tropical	307851.92	2703.92	344.64	4363.61

(C) Heterochronous ranked genealogies, human influenza A virus

	d_1	d_2	d_1^w	d_2^w
New York	668449.00	3982.78	139378.00	501.75
Chile	947326.20	4896.46	181856.90	568.68
Singapore	373360.50	1775.27	172247.40	544.05

Table S3: **Summary of distortion comparisons.** Comparison of distortion (Section S4.1) using distance functions between ranked trees shapes and ranked genealogies.

(A) Comparison of distortions on ranked tree shapes. The simulated data used for computation are the same considered for Table S1

	d_1	d_2	$d_{\text{BHV-RTS}}$	$d_{\text{KC-RTS}}$	$d_{\text{CP-RTS}}$
Isochronous ranked tree shapes (beta splitting)	5582.02	2369.98	8586.00	14331.64	27721.77
Isochronous ranked tree shapes (alpha-beta splitting)	3979.43	3028.30	10941.54	23989.64	5446.20
Heterochronous ranked tree shapes	2343.93	2969.33	4553.04	12034.03	1247.20

(B) Comparison of distortions on ranked genealogies. The simulated data used for computation are the same considered for Table S2.

	d_1^w	d_2^w	$d_{\text{BHV-RG}}$	$d_{\text{KC-RG}}$
Isochronous ranked genealogies	6751.44	1836.89	18569.28	13716.09
Heterochronous ranked genealogies	2283.54	2503.3	24424.35	3214.8

Table S4: **Summary of correlations.** Comparison of correlation (Section S4.2) between original distances and Euclidean distances in 2-dimensional MDS comparisons.

(A) Comparison of correlations of original distances and Euclidean distances in two-dimensional MDS plots on ranked tree shapes. The simulated data used for computation are the same considered for Table S1.

	d_1	d_2	$d_{\text{BHV-RTS}}$	$d_{\text{KC-RTS}}$	$d_{\text{CP-RTS}}$
Isochronous ranked tree shapes (beta splitting)	0.998	0.998	0.401	0.982	0.988
Isochronous ranked tree shapes (alpha-beta splitting)	0.998	0.999	0.648	0.518	0.909
Heterochronous ranked tree shapes	0.984	0.986	0.634	0.734	0.898

(B) Comparison of correlations of original distances and Euclidean distances in 2-dimensional MDS plots on ranked genealogies. The simulated data used for computation are the same considered for Table S2.

	d_1^w	d_2^w	$d_{\text{BHV-RG}}$	$d_{\text{KC-RG}}$
Isochronous ranked genealogies	0.981	0.954	0.527	0.946
Heterochronous ranked genealogies	0.926	0.888	0.377	0.924