Distance metrics for ranked evolutionary trees

Jaehee Kim[a], **Noah A. Rosenberg**[a], **and Julia A. Palacios**[b,c,1]

[a]Department of Biology, Stanford University, Stanford, CA 94305; [b]Department of Statistics, Stanford University, Stanford, CA 94305; and [c]Department of Biomedical Data Science, Stanford School of Medicine, Stanford, CA 94305

Edited by David M. Hillis, The University of Texas at Austin, Austin, TX, and approved September 27, 2020 (received for review December 30, 2019)

Genealogical tree modeling is essential for estimating evolutionary parameters in population genetics and phylogenetics. Recent mathematical results concerning ranked genealogies without leaf labels unlock opportunities in the analysis of evolutionary trees. In particular, comparisons between ranked genealogies facilitate the study of evolutionary processes of different organisms sampled at multiple time periods. We propose metrics on ranked tree shapes and ranked genealogies for lineages isochronously and heterochronously sampled. Our proposed 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.

coalescent | distance metric | phylogenetics | ranked genealogy | ranked tree shape

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 among gene genealogies can provide information about differences in these forces. In this article, we propose a distance on genealogies that enables biologically meaningful comparisons between genealogies of nonoverlapping samples. Our proposed distance, in its more general form, is a distance on ranked genealogies.

Ranked genealogies are rooted binary unlabeled trees with branch lengths and ordered internal nodes (Fig. 1*B*). These genealogies are also known as Tajima's genealogies (1, 2), as they were examined by Tajima (3). Unlabeled ranked genealogies are coarser than labeled ranked genealogies but finer than unlabeled unranked genealogies (Fig. 1); in a labeled or unlabeled unranked genealogy, multiple bifurcations may occur simultaneously (4).

Recently, there has been increasing interest in modeling ranked genealogies for studying evolutionary dynamics (1, 5, 6). One method for inferring evolutionary parameters from molecular data is based on the Tajima coalescent for ranked genealogies (2): Modeling of ranked genealogies, as opposed to labeled ranked genealogies as in the Kingman coalescent, reduces the dimensionality of the inference problem. Thus, for example, in studying macroevolution, Maliet et al. (7) proposed a model on ranked tree shapes—ranked genealogies without branch lengths but with internal node orders retained—to investigate factors influencing nonrandom extinction and the loss of phylogenetic diversity.

Many metrics on labeled trees have been proposed, such as the Robinson–Foulds (RF) metric (8), the Billera–Holmes–Vogtmann (BHV) metric (9), and the Kendall–Colijn (KC) metric (10). These metrics have been used for summarizing posterior distributions and bootstrap distributions of trees on the same set of taxa (11, 12), for comparing estimated genealogies of the same organisms obtained with different procedures, and for quantifying uncertainty in inferring evolutionary histories (13). The Colijn–Plazzotta (CP) metric (14) on the coarser-resolution space of tree shapes—rooted binary unlabeled unranked trees without branch lengths—has recently been devised. To our knowledge, no other metrics on ranked tree shapes or ranked unlabeled genealogies have been proposed to date.

A tree metric on the space of ranked genealogies can facilitate 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 living at different geographical locations and different time periods. Moreover, a useful tree metric not only provides a quantitative measure for ranked genealogy comparison, but can also discriminate between key aspects of different evolutionary processes (15). We show that our metrics separate samples of genealogies originating from different sampling distributions of ranked tree shapes 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 Bayesian Evolutionary Analysis by Sampling Trees (BEAST) (16), our tree metrics enable the construction of credible sets and Markov chain Monte Carlo (MCMC) convergence assessment.

We first define a metric on ranked 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 L_1 norm and L_2 (Frobenius) norm, to define metrics on ranked tree shapes. The choice of these two distances produces computational benefits, as computations of the metrics are quadratic in the number of leaves. Our metrics on ranked tree shapes retain more information than metrics based on unlabeled unranked tree shapes alone. We expand our metric definition to ranked genealogies, including branch

Significance

Rooted binary trees inferred from molecular sequence data provide information about the evolutionary history of populations and species. We introduce metrics on ranked tree shapes and ranked genealogies, in which the shape and temporal branching order in a tree are considered, but not the taxon labels. Our metrics enable quantification of evolutionary differences, assessment of tree uncertainty, and construction of statistical summaries of a tree distribution. They are computationally efficient and particularly useful for comparing phylodynamics of infectious diseases involving heterochronous samples and for comparative analyses of organisms that live in different geographic regions.

Author contributions: J.K. and J.A.P. designed research; J.K., N.A.R., and J.A.P. performed research; J.K. and J.A.P. analyzed data; and J.K., N.A.R., and J.A.P. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: juliapr@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1922851117/-/DCSupplemental.

First published November 2, 2020.

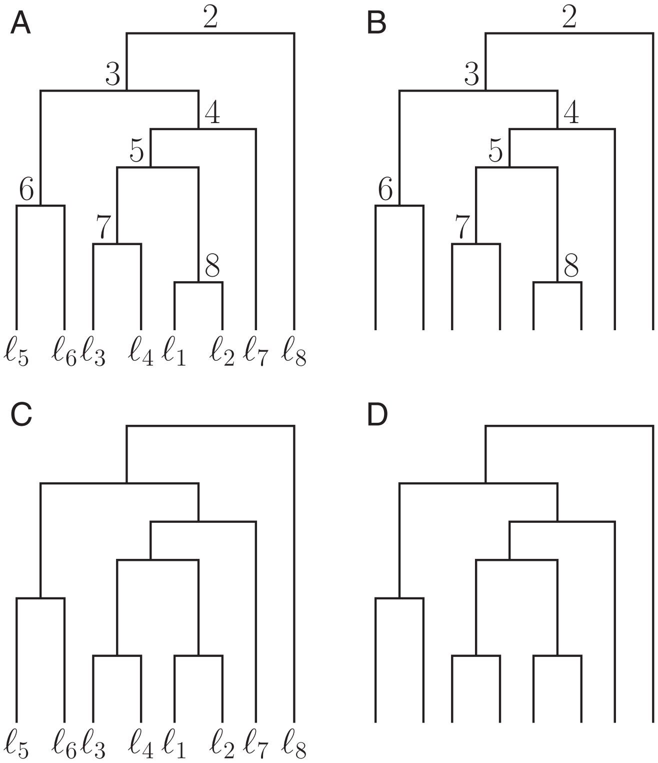
A
2
3
4
5
6
7
8
ℓ_5 ℓ_6 ℓ_3 ℓ_4 ℓ_1 ℓ_2 ℓ_7 ℓ_8
B
2
3
4
5
6
7
8
C
ℓ_5 ℓ_6 ℓ_3 ℓ_4 ℓ_1 ℓ_2 ℓ_7 ℓ_8
D

Fig. 1. Types of tree topology. (*A*) Labeled ranked tree shape. (*B*) Ranked tree shape (T^R). (*C*) Labeled unranked tree shape. (*D*) Tree shape. Genealogies corresponding to each topology include branch lengths.

lengths by weighting the matrix representation of ranked tree shapes by branch lengths.

We define metrics on isochronous ranked tree shapes and ranked genealogies, with all samples obtained at the same point in time. We next define metrics on heterochronous ranked tree shapes and ranked genealogies, in which samples are time stamped. While modeling isochronous genealogies is the standard practice for macroscopic organisms, such as animals and plants, modeling heterochronous genealogies is the standard approach for rapidly evolving organisms, such as viruses and bacteria (16–21). Heterochronous genealogies are also increasingly relevant in the study of ancient DNA samples (22–26).

We analyze properties of our proposed metrics and compare them to other tree-valued metrics—such as the BHV (9), KC (10), and CP (14) metrics—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. We use our distances to compare posterior distributions of genealogies of human influenza A/H3N2 virus from different geographic regions and to assess MCMC convergence. An implementation of our metrics is available for download at https://github.com/JuliaPalacios/phylodyn.

Definitions of Tree Topologies and Genealogies

All of 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 binary unlabeled tree with an increasing ordering of the $n-1$ interior nodes, starting at the root with label 2 (Fig. 1*B*). 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 focus mainly 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 (Fig. 1*A*) and a labeled ranked genealogy are the corresponding labeled counterparts of unlabeled ranked trees. A labeled unranked tree shape (Fig. 1*C*) and a labeled unranked genealogy are a rooted binary labeled tree shape and a rooted binary labeled genealogy without ranking of internal nodes, respectively. A tree shape (Fig. 1*D*) is a rooted binary unlabeled unranked tree. A genealogy is a tree shape equipped with branch lengths.

Tree Metrics for Comparative Analysis

Metrics on Labeled Trees. Many tree metrics have been proposed for phylogenetic studies (8–10, 27–30). We consider several of these. Billera et al. (9) introduced a metric on labeled unranked genealogies that is now commonly known as BHV space. Owen and Provan (31) and Chakerian and Holmes (11) provided polynomial algorithms and implementations for calculating the geodesic distance metric proposed by Billera et al. (9). Recently, Kendall and Colijn (10) proposed a metric on labeled unranked tree shapes and labeled unranked genealogies, representing each tree as a convex combination of two vectors—one encoding the number of edges from the root to the most recent common ancestor of every pair of tips in the tree and the other encoding the sum of the branch lengths of the corresponding paths. The most popular metric that can be computed in polynomial time is the symmetric difference of Robinson and Foulds (RF) (8) on labeled unranked tree shapes. This measure has an associated branch-length measure RFL on labeled genealogies (27).

Metrics on Unlabeled Trees. A metric on unlabeled trees is desirable because it enables comparison between trees of different samples. However, few such metrics are available. Colijn and Plazzotta (14) 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. (32) developed a kernel function that measures the similarity between two genealogies by accounting for differences in branch lengths and matching the number of descendants over all nodes for both trees. Lewitus and Morlon (33) proposed the Jensen–Shannon distance between the spectral density profiles of the corresponding modified graph Laplacians of the genealogies.

We are introducing metrics for unlabeled ranked tree shapes and unlabeled ranked genealogies. To define our metrics on ranked tree shapes, we need to introduce a unique encoding of a ranked tree shape as an integer-valued triangular matrix.

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 $n-1$ coalescence times. Here, time increases into the past (rootward), $u_{n-1} < \ldots < u_1$, 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) traversed forward in time (tipward).

In Fig. 2*B*, we show the corresponding **F** matrix to the ranked tree shape depicted in Fig. 2*A*. In the interval (u_2, u_1), there are two lineages, so $F_{1,1} = 2$. Traversing tipward, 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 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), $F_{2,2} = 3$. Traversing tipward, 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

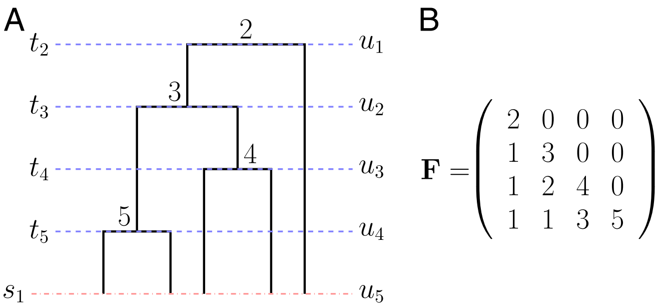
A
B
t_2
t_3
t_4
t_5
s_1
2
3
4
5
u_1
u_2
u_3
u_4
u_5
$\mathbf{F} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 2 & 4 & 0 \\ 1 & 1 & 3 & 5 \end{pmatrix}$

Fig. 2. 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 of the tree in *A*.

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, then 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 $0 = u_{n+m-1} < u_{n+m-2} < \ldots < u_1$, 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 *SI Appendix*, Fig. S5 *C* and *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, then it encodes exactly one ranked tree shape. The proof appears in *SI Appendix*, section 1. In the next section, we leverage the **F**-matrix representation of ranked tree shapes to define a distance between ranked tree shapes. From this point onward, we assume a matrix **F** represents a ranked tree shape.

Metrics 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 $r \times r$, where $r \geq n-1$,

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

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

The two distances are metrics as they inherit L_1 and L_2 elementwise matrix metrics with well-understood properties (34). The definitions are valid for both isochronous and heterochronous ranked tree shapes, as both types of trees have unique matrix encodings.

In the following definition, we include branch lengths to define distances between ranked genealogies. We 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 $0 = u_{n+m-1} < u_{n+m-2} < \ldots < u_1$, is the vector of $n-1$ coalescent times t_k and m sampling times s_ℓ, ordered with time increasing into the past (rootward). For a pair of ranked genealogies $\mathbf{g}_1^R$ and $\mathbf{g}_2^R$, and their corresponding **F**-matrix representations $\mathbf{F}^{(1)}$ and $\mathbf{F}^{(2)}$,

$$d_1^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = \sum_{i=1}^{r} \sum_{j=1}^{i} |F_{i,j}^{(1)} W_{i,j}^{(1)} - F_{i,j}^{(2)} W_{i,j}^{(2)}|, \qquad [3]$$

$$d_2^w(\mathbf{g}_1^R, \mathbf{g}_2^R) = \sqrt{\sum_{i=1}^{r} \sum_{j=1}^{i} \left(F_{i,j}^{(1)} W_{i,j}^{(1)} - F_{i,j}^{(2)} W_{i,j}^{(2)}\right)^2}, \qquad [4]$$

where $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are the weight matrices associated with $\mathbf{g}_1^R$ and $\mathbf{g}_2^R$, respectively. *SI Appendix*, Fig. S6 shows an example weight matrix **W**, associated with the example heterochronous ranked genealogy and its **F** matrix in *SI Appendix*, Fig. S5 *C* and *D*.

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

The proof appears in *SI Appendix*, section 2. Our distances on ranked tree shapes and ranked genealogies are distances between trees with the same number of leaves and, additionally in the heterochronous case, the same number of sampling events. An extension to cases in which the numbers of sampling events differ but the total numbers of leaves remain the same is described in *SI Appendix*, section 3.

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 and Ranked Genealogy Summary Statistics. We first 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

$$\overline{T}_i := \underset{T \in \{T_1^R, T_2^R, \ldots, T_s^R\}}{\operatorname{argmin}} \sum_{j=1}^{s} d_i^2(T, T_j^R) \qquad [5]$$

for $i = 1, 2$. We note that our definition of the L_2-medoid set corresponds to the ranked tree shape(s) in the sample minimizing the sum of squared distances as opposed to 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 do not require it to be a sampled tree, Eq. **5** corresponds to the Fréchet mean or barycenter under uniform sampling probabilities (35).

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

$$\sigma_i^2 := \frac{1}{s} \sum_{j=1}^{s} d_i^2(\overline{T}_i, T_j^R). \qquad [6]$$

When the L_2 medoid is not unique, the dispersion is defined with respect to any chosen L_2-medoid ranked tree shape from the set of L_2 medoids.

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

$$\bar{\mathbf{g}}_i := \underset{\mathbf{g} \in \{\mathbf{g}_1^R, \ldots, \mathbf{g}_s^R\}}{\operatorname{argmin}} \sum_{j=1}^{s} \left[d_i^w(\mathbf{g}, \mathbf{g}_j^R) \right]^2 \qquad \textbf{[7]}$$

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

Statistical Comparison of Ranked Tree Shape and Ranked Genealogy Sampling Distributions. To assess the utility of a metric in distinguishing between two different ranked tree shape (or ranked genealogy) sampling distributions, we propose a nonparametric test of equality of tree sampling distributions as follows. Given two sets of independent samples of ranked tree shapes (or ranked genealogies) from distributions P and Q, we are interested in testing the null hypothesis $H_0 : \mathrm{P} = \mathrm{Q}$.

Assume $X_1, \ldots, X_N$ are independent and identically distributed (i.i.d.) ranked tree shapes (or ranked genealogies) with n leaves according to a probability distribution P, and independently, $Y_1, \ldots, Y_N$ are i.i.d. ranked tree shapes (or ranked genealogies) with n leaves according to Q. For a given distance function d defined on the space of ranked tree shapes (or ranked genealogies), let $\bar{X}$ be the L_2 medoid of the $X_1, \ldots, X_N$ samples and let $\bar{Y}$ be the L_2 medoid of the $Y_1, \ldots, Y_N$ samples, as defined in Eq. **5**. The mean confusion is defined as

$$C^{x,y} = \frac{1}{2(N-1)} \sum_{j=1}^{N} \left[1_{d(X_j,\bar{Y}) \leq d(X_j,\bar{X})} + 1_{d(Y_j,\bar{X}) \leq d(Y_j,\bar{Y})} \right]. \qquad \textbf{[8]}$$

The closer the value of the mean confusion statistic is to 0, the more easily the two tree distributions can be distinguished from each other. The mean confusion as defined in Eq. **8** assumes the two L_2 medoids are unique, each observed only once, and ignores ties (trees equidistant to $\bar{X}$ and $\bar{Y}$). When the sample space of ranked tree shapes is large or when the sample space is the space of ranked genealogies, the expected value of the mean confusion statistic under the null hypothesis is close to 0.5. In *SI Appendix*, section 4, we define a more general mean confusion statistic that accounts for nonunique L_2 medoids and ties. We note that for our simulations and real data analyses, Eq. **8** is used since we observe a unique L_2 medoid in each ranked tree shape or ranked genealogy distribution, and no ties are present in pairwise comparisons.

An α-level test can be constructed for testing the null hypothesis using N_{perm} random permutations. The permutation *P* value is computed as (36)

$$P = \frac{1 + \sum_{k=1}^{N_{\text{perm}}} 1_{C_k^{x,y} \leq C^{x,y}}}{1 + N_{\text{perm}}}, \qquad \textbf{[9]}$$

where $C_k^{x,y}$ is a confusion statistic from the kth replicate and $C^{x,y}$ is the observed value from data.

For comparing more than two sampling distributions, we compute the confusion matrix C, where entry $C_{i,j}$ corresponds to the percentage of trees in the ith sampling distribution that are closest to the L_2 medoid of the jth sampling distribution. Each row of the confusion matrix sums to 100%, and the diagonal elements represent the percentage of trees that are closer to the L_2 medoid of their originating distribution than L_2 medoids of the other sampling distributions. The average diagonal of the confusion matrix, which we term mean separation, provides a measure of overall separation between sampling distributions based on the distance used. It ranges from 0 to 100, with a larger value indicating better separation between sampling distributions.

Table 1. Summary of distances in the spaces of ranked tree shapes and ranked genealogies

Name	Tree space	Original tree space	Reference
d_1, d_2	Ranked tree shape	Ranked tree shape	This paper
d_1^w, d_2^w	Ranked genealogy	Ranked genealogy	This paper
$d_{\text{BHV-RTS}}$	Ranked tree shape	Labeled genealogy	Billera, Holmes, and Vogtmann (9)
$d_{\text{BHV-RG}}$	Ranked genealogy	Labeled genealogy	Billera, Holmes, and Vogtmann (9)
$d_{\text{KC-RTS}}$	Ranked tree shape	Labeled genealogy	Kendall and Colijn (10)
$d_{\text{KC-RG}}$	Ranked genealogy	Labeled genealogy	Kendall and Colijn (10)
$d_{\text{CP-RTS}}$	Ranked tree shape	Tree shape	Colijn and Plazzotta (14)

A brief description of existing distances that are originally defined on the space of tree shapes and the space of labeled genealogies, and their adaptations to the spaces of ranked tree shapes and ranked genealogies, can be found in *SI Appendix*, sections 5 and 6.

Adapting Other Tree Metrics to Ranked Tree Shapes and Ranked Genealogies. Although there are no other metrics designed specifically for ranked tree shapes and ranked genealogies, we can adapt other tree distances defined on finer or coarser tree spaces to the space of ranked tree shapes and ranked genealogies, and we can compare them to our metrics. We adapted metrics that are originally defined on the space of labeled genealogies—the BHV distance and the KC distance—to the space of ranked genealogies by artificially assigning uniquely defined leaf labels to ranked genealogies. The unique assignment of the leaf labels on a ranked tree shape consists of assigning labels in increasing index order starting with leaves subtended by an internal node closest to the tips and ending with leaves subtended closest to the root. For adapting the BHV distance and the KC distance to the space of ranked tree shapes, we assigned a unit length to each change point time interval and unique leaf labels to tips of a ranked tree shape. We also compared our distances on ranked tree shapes to the CP distance defined on the space of tree shapes by ignoring internal node labels. Details of these adaptations can be found in *SI Appendix*, sections 5 and 6.

We term the adapted BHV, KC, and CP distances on the space of ranked tree shapes $d_{\text{BHV-RTS}}$, $d_{\text{KC-RTS}}$, and $d_{\text{CP-RTS}}$, respectively. We denote the adapted BHV and KC distances on ranked genealogies by $d_{\text{BHV-RG}}$ and $d_{\text{KC-RG}}$, respectively. Table 1 shows a summary of the distances defined on the spaces of ranked tree shapes and ranked genealogies, as used in this paper.

Results

Having introduced the metrics on the spaces of ranked tree shapes and ranked genealogies, we first examine the behavior of our metrics in an illustrative example for $n = 5$. We then show their utility in simulated data under various evolutionary models. Finally, we apply them to analyze real human influenza A/H3N2 virus data.

Interpreting Proposed Distances between Ranked Tree Shapes of $n = 5$ Leaves. Before demonstrating the utility of our metrics in distinguishing different sampling distributions by simulations and a real data application, as a simple illustrative example, we first explore our distances on ranked tree shapes of $n = 5$

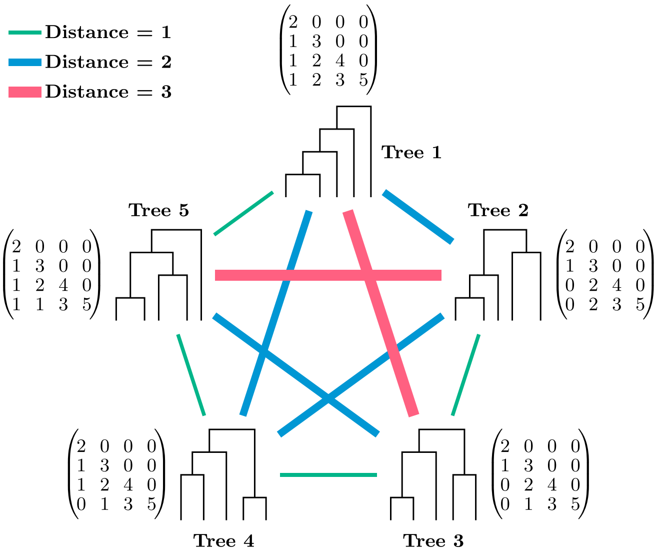

Fig. 3. d_1 distances between all pairs of ranked tree shapes of $n = 5$ leaves. Rank labels of internal nodes are removed for ease of visualization. There are three different distance values among the 10 pairs of distances.

leaves—the smallest number of taxa giving a nontrivial space of ranked tree shapes—and compare them with the adaptations of other metrics.

In Fig. 3, we show all five 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 trees in each of the pairs (T_2^R, T_3^R) and (T_3^R, T_4^R) differ by one rank switch, 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 rank switches. All pairwise distances are shown in *SI Appendix*, Fig. S7.

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 rank switch and a change of tree shapes to move from one tree to the other. The pairs of trees with the maximum $d_{\text{BHV-RTS}}$ 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 $d_{\text{BHV-RTS}}$ maximum, even though T_2^R and T_4^R differ by two rank change events but have the same tree shape.

Qualitative behaviors of distances $d_{\text{KC-RTS}}$ among the trees T_2^R, T_3^R, and T_4^R mirror d_1. $d_{\text{CP-RTS}}$, however, differs in that the trees T_2^R, T_3^R, and T_4^R are equidistant from T_5^R but not from the caterpillar tree T_1^R, whereas, using d_1 or d_2, all pairwise distances to the trees T_2^R, T_3^R, and T_4^R from T_1^R or T_5^R are distinct. We also note that because $d_{\text{CP-RTS}}$ is a distance originally defined on tree shapes, all pairwise $d_{\text{CP-RTS}}$ distances between the trees T_2^R, T_3^R, and T_4^R are 0, and all $d_{\text{CP-RTS}}$ distances to the trees T_2^R, T_3^R, and T_4^R from T_1^R or T_5^R have the same value of 2.

Separation between Ranked Tree Shapes with Different Distributions. Having demonstrated that our proposed metrics, d_1 and d_2, effectively differentiate both tree topologies and internal node rankings in the $n = 5$ case, we now evaluate by simulation how d_1 and d_2 can distinguish different ranked tree shape distributions from models with biological relevance.

We consider isochronous ranked tree shapes from a two-parameter branching model to which we refer as the alpha–beta-splitting model (7). These parameters capture two phenomena that have been widely used for phylogenetic studies. The parameter $\beta \in [-2, \infty)$, as in the beta-splitting model (37), determines the degree of balance, an important tree attribute that carries information on underlying evolutionary or epidemiological processes, including speciation and extinction (38–41), natural selection (42–45), and infectious disease transmission (46–48). The other parameter $\alpha \in (-\infty, \infty)$ regulates the relationship between clade family size and proximity to the root (clade size–age relation), which is fundamental to understanding the effects of ecological, evolutionary, geographic, and other factors on biodiversity (7, 49–52). More details about these models can be found in *SI Appendix*, sections 7 and 8.

We evaluate and compare our metrics, d_1 and d_2, with the adaptations of other metrics to the space of ranked tree shapes (Table 1)—$d_{\text{BHV-RTS}}$, $d_{\text{KC-RTS}}$, and $d_{\text{CP-RTS}}$—by the following methods. We summarize each sampling distribution by the summary statistics, the L_2 medoid (Eq. **5**) and the dispersion (Eq. **6**). We then assess the performance of the metrics in distinguishing different ranked tree shape distributions by several measures: the confusion matrix, mean separation, and mean confusion (Eq. **8**) and its permutation P value (Eq. **9**) for testing equality between pairs of sampling distributions (10^4 permutations). Additionally, for ease of visualization and interpretation, we use multidimensional scaling (MDS) (53) to embed our distance metrics into Euclidean spaces of two dimensions.

Distinguishing Different Tree Balance Distributions. To show how our proposed metrics can be used to differentiate ranked tree shapes sampled from distributions with different degrees of tree balance, we simulated ranked tree shapes under the alpha–beta-splitting model with different values of the tree balance parameter (β) while keeping the clade size–age relation parameter (α) fixed. We considered $\beta \in \{-1.9, -1.5, -1, 0, 100\}$, representing a sequence from unbalanced to balanced ranked tree shapes. We refer to the ranked tree shape distributions with their corresponding well-known models of speciation: the Yule model (54, 55) ($\beta = 0$), the proportional-to-distinguishable-arrangements (PDA) model (41) ($\beta = -1.5$), and the Aldous branching (AB) model (40) ($\beta = -1$). Additionally, we refer to $\beta = -1.9$ and $\beta = 100$ as "unbalanced" and "balanced", respectively. We simulated 1,000 ranked tree shapes with $n = 100$ leaves for each β value, generating 5,000 simulated ranked tree shapes. We then computed the pairwise distance matrices of size $5,000 \times 5,000$ with d_1, d_2, $d_{\text{BHV-RTS}}$, $d_{\text{KC-RTS}}$, and $d_{\text{CP-RTS}}$.

The confusion matrices displayed in *SI Appendix*, Table S2 show the greatest mean separation across the five distributions 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 $d_{\text{BHV-RTS}}$. Our d_1 and d_2 distances discriminate tree distributions with different tree balance parameters to a greater extent than $d_{\text{BHV-RTS}}$; $d_{\text{CP-RTS}}$ and $d_{\text{KC-RTS}}$ show more similar performance.

To test for equality in sampling distributions, we computed the pairwise mean confusions and associated P values. *SI Appendix*, Table S3 *A* and *B* shows mean confusions computed using our metrics, d_1 and d_2. All off-diagonal values are smaller than 0.21 and statistically significant ($P < 0.0001$). In contrast, comparisons with $d_{\text{BHV-RTS}}$ and $d_{\text{KC-RTS}}$ (*SI Appendix*, Table S3 *C* and *D*) produce higher pairwise mean confusion values across all pairs—with off-diagonal values larger than 0.45 with $d_{\text{BHV-RTS}}$ and values between 0.04 and 0.28 with $d_{\text{KC-RTS}}$—and thus, lower discrimination. Mean confusion values closer to 0 are indicative of good discrimination. Comparisons with $d_{\text{CP-RTS}}$ (*SI Appendix*, Table S3*E*) display similar pairwise mean confusion values and significance to our distances (all P values are significant at $P < 0.0001$).

SI Appendix, Table S1*A* shows the dispersion statistic for each tree distribution and for each distance function. The unbalanced sample shows the smallest dispersion value (all sampled

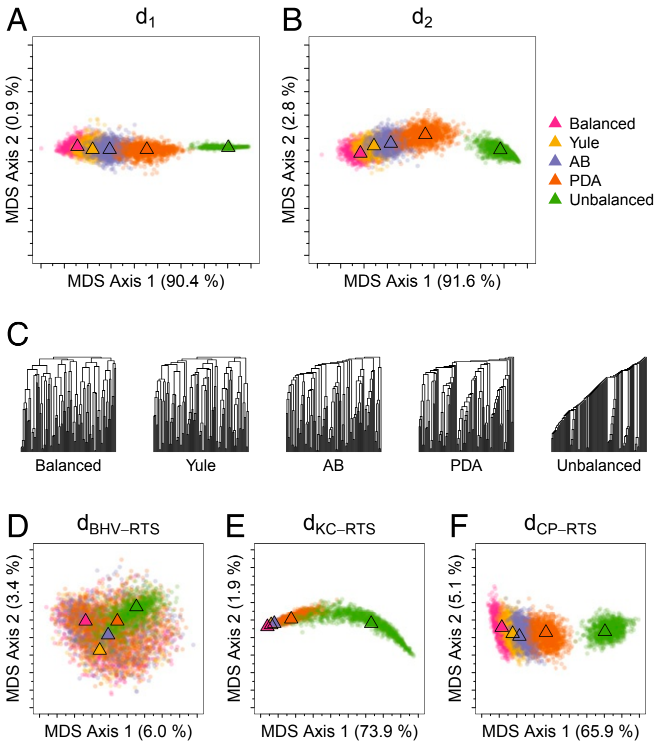

Fig. 4. MDS representation of distances between 5,000 simulated isochronous ranked tree shapes of $n = 100$ leaves, aggregated from five different beta-splitting models. A total of 1,000 isochronous ranked tree shapes were simulated from each of the following models: balanced model ($\beta = 100$), Yule model ($\beta = 0$), Aldous branching (AB) model ($\beta = -1$), proportional-to-distinguishable-arrangements (PDA) model ($\beta = -1.5$), and 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. (*D–F*) 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 1,000 points for a specified model.

trees are closer to each other) with our distances and $d_{\text{BHV-RTS}}$, whereas, with $d_{\text{KC-RTS}}$ and $d_{\text{CP-RTS}}$, it displays the largest dispersion value (more separation among trees within the same distribution).

Our findings are confirmed visually through the MDS plots displayed in Fig. 4. In the MDS plots, each dot corresponds to a tree, and each color corresponds to the sampling distribution for a specified β value. Fig. 4*C* shows the L_2-medoid trees using our d_1 distance. The total distances explained by the first two MDS dimensions are 91.4% using d_1 (Fig. 4*A*) and 94.4% using d_2 (Fig. 4*B*). The two-dimensional MDS mappings using the other distances explain less than 80%.

Distinguishing Different Internal Node Ranking Distributions. To show how our proposed metrics can be used to differentiate ranked tree shapes sampled from distributions with different internal node rankings, we generated 1,000 random trees from the alpha–beta-splitting model with $n = 100$ tips, fixed $\beta = 0$, and varying values of α in $\{-2, -1, 0, 1, 2\}$, producing 5,000 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 ($\alpha > 0$).

The confusion matrices displayed in *SI Appendix*, Table S4 show greater mean separation for d_1 and d_2 than for any of the other distances. 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.

The pairwise mean confusions computed using d_1 and d_2 (*SI Appendix*, Table S5 *A* and *B*) have values strictly smaller than the mean confusions computed with the other distances. All comparisons with our distances are statistically significant ($P < 0.0001$), indicating good distinguishability of distributions with different α values.

The dispersion statistics in *SI Appendix*, Table S1*B* indicate that, according to our d_1 and d_2 metrics, trees from the alpha–beta-splitting model with $\alpha = 0$ are the most dispersed group among the five α values considered. This, however, is not the case with the other distances. In particular, no variations in dispersion across different distributions are observed using $d_{\text{CP-RTS}}$, and a positive correlation between α and dispersion is seen with $d_{\text{BHV-RTS}}$.

Our two-dimensional MDS representations confirm the findings observed in the confusion matrix and mean confusions: d_1 and d_2 (Fig. 5 *A* and *B*), when compared to the other three metrics (Fig. 5 *D–F*), distinguish to a greater extent between trees with positive and negative α parameters. In addition, our d_1 and d_2 distances show tighter embeddings in two-dimensional MDS, with a high proportion of the total distance explained. Although the MDS visualizations in Fig. 5 *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 (Fig. 5*C*).

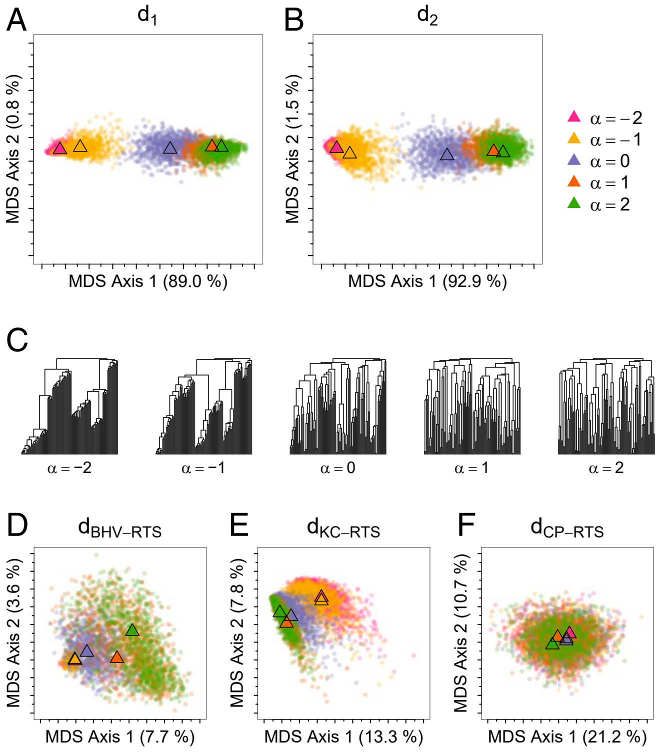

Fig. 5. MDS representation of distances between 5,000 simulated isochronous ranked tree shapes of $n = 100$ leaves, aggregated from five different alpha–beta-splitting models. A total of 1,000 isochronous ranked tree shapes were simulated for each α value in $\{-2, -1, 0, 1, 2\}$. Different α values generate 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. (*D–F*) 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 1,000 points for a specified model.

Our results confirm that our metrics capture both tree balance and clade size–age relation 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. To further demonstrate this point, we simulated 1,000 trees in each of the varying tree balance and internal node rankings: $(\alpha, \beta) = (-1, 0), (-1, -1.5), (1, 0), (1, -1.5)$. The groups with the same β value share the same tree balance distribution, and the groups with the same α value share the same internal node ranking distribution. In *SI Appendix*, Table S6 and Fig. S8 show that d_1 and d_2 effectively separate all four distributions, while $d_{\text{KC-RTS}}$ and $d_{\text{CP-RTS}}$, designed to measure differences in tree shapes only, separate tree distributions with the same tree balance but with different internal node rankings less effectively.

Separation between Ranked Genealogies with Different Distributions. Having investigated the utility of our metrics in separating different ranked tree shape distributions, we now demonstrate our metrics on the space of ranked genealogies. We consider two scenarios with evolutionary and epidemiological importance: isochronous ranked genealogy distributions resulting from different demographic histories and heterochronous ranked genealogies inferred from human influenza A/H3N2 virus data from two geographical regions. We evaluate and compare our metrics, d_1^w and d_2^w, with the adaptations of other metrics to the space of ranked genealogies (Table 1)—$d_{\text{BHV-RG}}$ and $d_{\text{KC-RG}}$—using the same measures we employed in the previous section. We note that there is no adaptation of the CP metric to the space of ranked genealogies, as the CP metric is defined on the space of tree shapes and, thus, does not incorporate tree branch lengths.

Distinguishing Different Demographic Histories. We assess how our proposed weighted metrics can differentiate genealogies with different branch length distributions. In coalescent models, the rate of coalescence (or branching) depends on the effective population size $\lambda(t)$, as shown in Eq. **10**. Under the neutral coalescent, the tree topology distribution (*SI Appendix*, Eq. **S2**) is independent of the branch lengths (see *Materials and Methods* for details). 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 1,000 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

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

The coalescent trees under each specified population trajectory were simulated with $N_0 = 10^4$. The functional form and parameter values chosen for the seasonal logistic trajectory mimic estimated trajectories of human influenza A virus in temperate regions (56). We removed the leaf labels and retained branch lengths to produce isochronous ranked genealogies. We produced pairwise distance matrices of $3{,}000 \times 3{,}000$ using the weighted metrics d_1^w and d_2^w.

The dispersion statistics in *SI Appendix*, Table S9A show that the sampled genealogies under the constant population size trajectory have the greatest dispersion, and the sampled genealogies under the exponential growth trajectory have the least dispersion.

Our metrics' ability to distinguish different demographic scenarios is evident in the confusion matrix in *SI Appendix*, Table S10, where they display good 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 $d_{\text{KC-RG}}$.

To test for equality in sampling distributions, we computed the mean confusion statistics and *P* values. In *SI Appendix*, Table S11 *A* and *B* displays all off-diagonal mean confusion statistics that are less than 0.002 and are statistically significant ($P < 0.0001$) with d_1^w and d_2^w, indicating a high level of discrimination between ranked genealogy distributions with different demographic histories. The mean confusions computed with $d_{\text{BHV-RG}}$ and $d_{\text{KC-RG}}$ have far higher values than the values with d_1^w and d_2^w (*SI Appendix*, Table S11 *C* and *D*).

The two-dimensional MDS plots, Fig. 6 *A* and *B*, display that our weighted metrics distinguish the three simulated ranked genealogy distributions, visually confirming the same finding using the confusion matrix and the mean confusion statistic. The first two dimensions of MDS account for 72.1% of the total distance, with the first MDS coordinate separating the constant trajectory from the others and the second coordinate

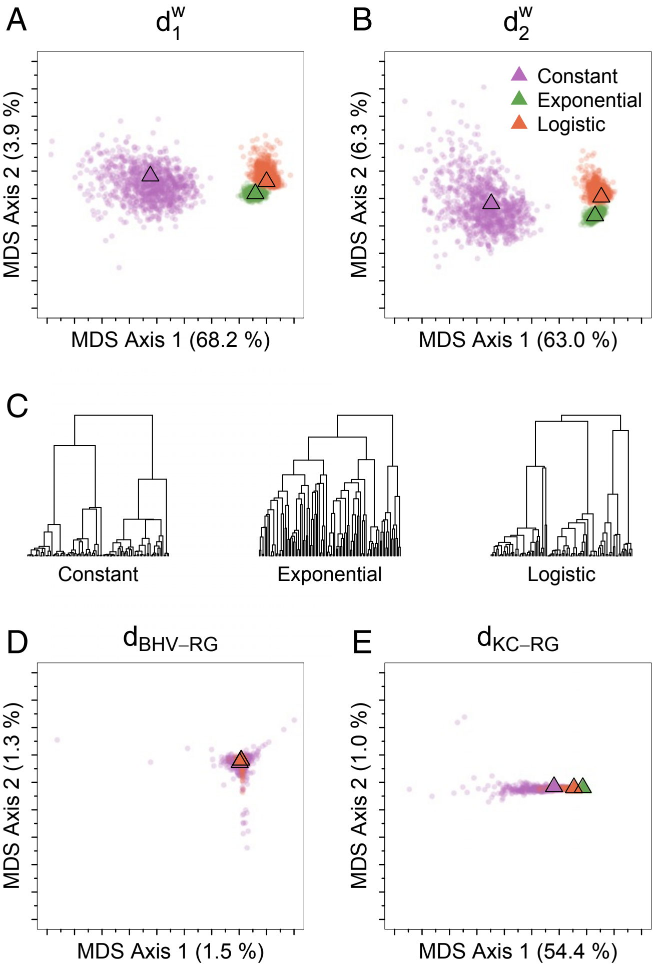

Fig. 6. MDS representation of distances between 3,000 simulated isochronous ranked genealogies of $n = 100$ leaves under different demographic models. The number of simulated genealogies per population model is 1,000. (*A*) d_1^w metric. (*B*) d_2^w metric. (*C*) L_2-medoid genealogies from each distribution using the d_1^w metric. (*D*) $d_{\text{BHV-RG}}$. (*E*) $d_{\text{KC-RG}}$.

distinguishing the exponential trajectory from the logistic trajectory. A comparison of panels within Fig. 6 shows that $d_{\text{BHV-RG}}$ is far less informative than either of our metrics.

Analysis of Human Influenza A/H3N2 Virus from Different Geographical Regions. We next apply our metrics to ranked genealogies sampled from the posterior distributions of human influenza A/H3N2 genealogies of hemagglutinin (HA) gene sequences from two geographic regions—New York and Southeast Asia. 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, such as Southeast Asia, influenza activity persists throughout the year (57). We reanalyzed a subset of the sequences used in a previous study of Bahl et al. (58) that showed patterns of seasonal dynamics of influenza in New York and stable dynamics in Southeast Asia. We use our metrics to assess the differences in genealogical posterior distributions between the two regions.

The frequency of the collected samples by collection date and the estimated effective population size trajectories (Fig. 7) show distinct patterns consistent with those in Bahl et al. (58): a seasonal effective population size trajectory in New York that peaks during the winter seasons and a flat effective population size trajectory in Southeast Asia.

For each geographical region, we sampled 1,000 ranked genealogies per region from posterior distributions obtained from BEAST. We refer to the resulting samples from New York and Southeast Asia as NY–NY and SEA–SEA, respectively.

To investigate the effect of the different sampling schedules, we simulated two additional groups of 1,000 trees: one with fixed New York sampling schedule (Fig. 7*A*) and Southeast Asia estimated effective population size trajectory (Fig. 7*D*) and the other with fixed Southeast Asia sampling schedule (Fig. 7*B*) and New York estimated effective population size trajectory (Fig. 7*C*). We denote the former by NY–SEA and the latter by SEA–NY.

We then aggregated the 4,000 ranked genealogies—1,000 from each of NY–NY, SEA–SEA, NY–SEA, and SEA–NY—and computed pairwise distance matrices of $4,000 \times 4,000$ with d_1^w, d_2^w, $d_{\text{BHV-RG}}$, and $d_{\text{KC-RG}}$. The corresponding confusion matrices (*SI Appendix*, Table S12) show that d_1^w and d_2^w have good performance in discriminating all four distributions, both with mean separation of 99.9%, while $d_{\text{BHV-RG}}$ and $d_{\text{KC-RG}}$ display less discrimination, with mean separations of 28.9 and 65.6%, respectively.

As shown in *SI Appendix*, Table S13 *A* and *B*, the observed mean confusion statistics between all sampling distribution pairs with our distances are near zero and statistically significant ($P <$ 0.0001). While all observed off-diagonal mean confusion values with $d_{\text{KC-RG}}$ (*SI Appendix*, Table S13*D*) are larger than the values with d_1^w or d_2^w, all P values are significant, showing somewhat comparable performance with our distances. The test of equality in sampling distributions is not rejected at the significance level of 5% with $d_{\text{BHV-RG}}$ between the pair (NY–SEA, NY–NY) and between the pair (SEA–NY, NY–NY) (*SI Appendix*, Table S13*C*).

Fig. 8 *A* and *B* shows that all four distributions with different sampling and evolutionary histories are also well separated in the two-dimensional MDS plots using d_1^w and d_2^w. However, we note that the first two dimensions of MDS account for only 51.7% (d_1^w) and 50.5% (d_2^w) of the total distance.

SI Appendix, Table S9*B* displays the dispersion values computed with all distances on ranked genealogies. In particular, SEA–SEA ranked genealogies with relatively uniform sampling history and evolutionary trajectory have larger dispersion compared to the other distributions. The large dispersion of SEA–SEA ranked genealogies is consistent with the simulated results in Fig. 6 and *SI Appendix*, Table S9*A*, where the distribution of isochronous ranked genealogies with constant population size trajectory has larger dispersion than distributions with exponential or seasonal logistic trajectory.

Our proposed distances confirm that the HA segments of human influenza A/H3N2 experienced different evolutionary processes in New York and Southeast Asia between 2001 and 2005. While different sampling schedules of different datasets may obfuscate signals of underlying evolutionary dynamics in the application of our distances, we note that in this

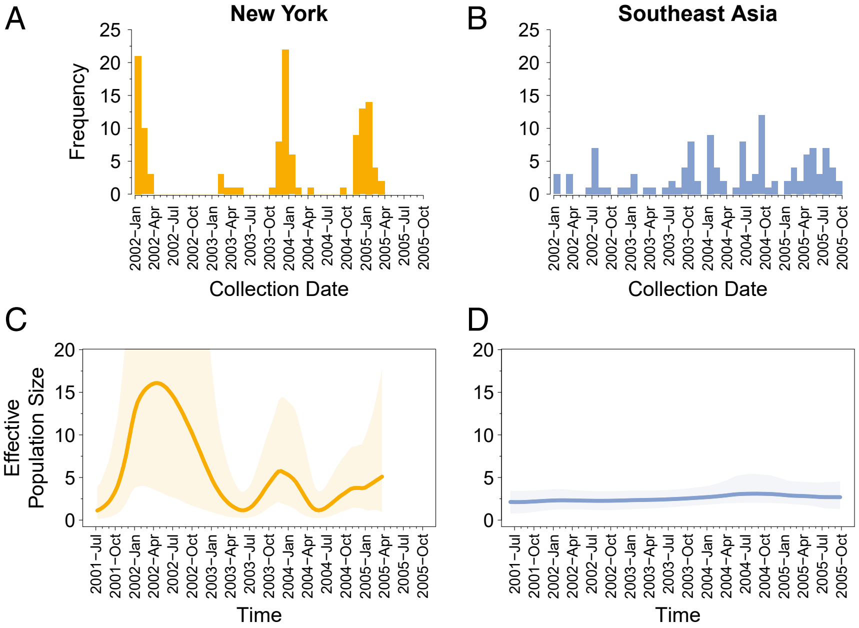

Fig. 7. Human influenza A/H3N2 virus collection dates and inferred effective population size trajectories. (*A*) Collection date histogram, New York. (*B*) Collection date histogram, Southeast Asia. (*C*) Inferred effective population size trajectory with BEAST, New York. (*D*) Inferred effective population size trajectory with BEAST, Southeast Asia.

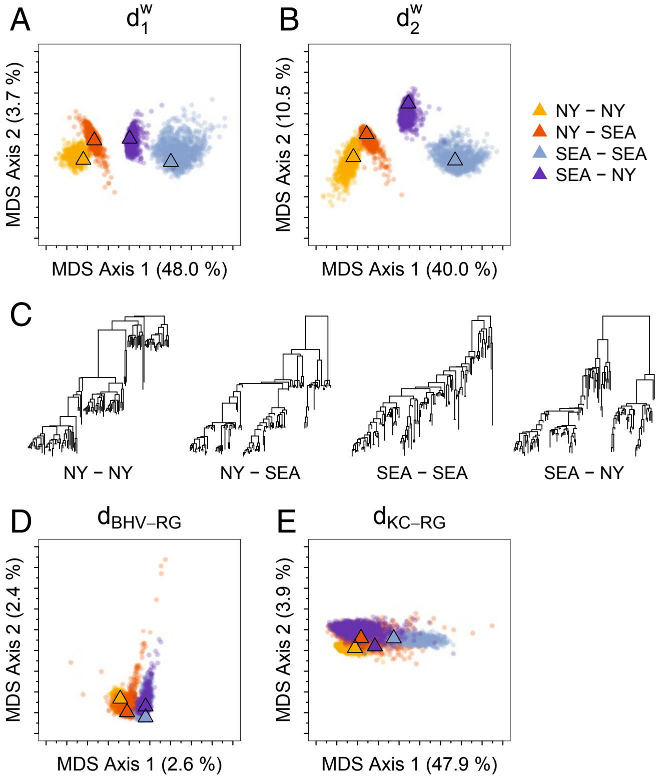

Fig. 8. MDS representation of distances between 4,000 heterochronous ranked genealogies sampled from two posterior distributions (1,000 trees from each distribution) and two sets of simulated trees (1,000 trees from each simulation). Observed data consist of 122 HA sequences of human influenza A/H3N2 virus from each geographic region, New York (NY–NY) and Southeast Asia (SEA–SEA). An additional two sets of trees were simulated using sampling times of New York and the effective population size trajectory of Southeast Asia (NY–SEA) and using sampling times of Southeast Asia and the effective population size trajectory of New York (SEA–NY). (*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*) $d_{\text{BHV-RG}}$ metric. (*E*) $d_{\text{KC-RG}}$ metric.

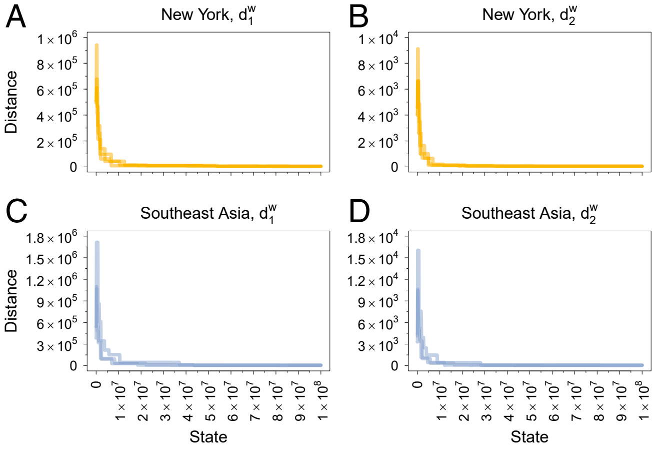

Fig. 9. Assessment of convergence of MCMC BEAST chains. Each curve corresponds to the distance between the running posterior L_2-medoid ranked genealogy and the global posterior L_2-medoid ranked genealogy after every 10^5 iterations for each chain. (*A*) New York, d_1^w. (*B*) New York, d_2^w. (*C*) Southeast Asia, d_1^w. (*D*) Southeast Asia, d_2^w.

particular analysis, sampling frequencies are highly correlated with effective population sizes and are thus informative about the underlying evolutionary processes. This situation of preferential sampling (59–61) is common in molecular phylodynamic studies of influenza viruses: More samples are sequenced when the effective population size is larger.

Convergence Diagnostic. In the previous section, we compared posterior samples of ranked genealogies derived from MCMC procedures that introduce correlation in the samples. Although we thinned our posterior samples, it is important to test the mixing of the Markov chains. We used our metrics to assess convergence of each of the three MCMC chains for the two geographical regions. We computed the distance between the running posterior L_2-medoid ranked genealogy and the global posterior L_2-medoid ranked genealogy after every 10^5 iterations for each chain. That is, in each chain, we sampled 1,000 ranked genealogies from the posterior distribution so that the kth sample $\mathbf{g}_k^R$ corresponds to iteration $k \times 10^5$ of the chain. For each $k = 1, \ldots, 1000$, we computed a running L_2-medoid $\bar{\mathbf{g}}_k$ of all samples up to and including the kth ranked genealogy using our metric. We then computed distances from the running L_2 medoids $\bar{\mathbf{g}}_k (k = 1, \ldots, 1{,}000)$ to the overall L_2 medoid of all of the 1,000 sampled ranked genealogies $\bar{\mathbf{g}}$ $(=\bar{\mathbf{g}}_{1,000})$.

Fig. 9 shows $d_i^w(\bar{\mathbf{g}}_k, \bar{\mathbf{g}})$ $(i = 1, 2)$ as a function of the chain iteration index for the three independent runs for each geographical region. The initial variability of the running-mean plot is relatively high, but, after an initial burn-in period (≈10%), the distance from the running mean to the overall mean stabilizes quickly with an increasing number of states indicating convergence.

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.

In this article, 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. Our distances on ranked genealogies are the L_1 and L_2 norms of the difference between two matrices weighted 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 used our distances to summarize the distribution of trees and proposed using the L_2-medoid tree set, a sample version of the Fréchet mean, as the central set of a sample of trees. We proposed a sample version of the Fréchet variance to quantify uncertainty. We demonstrated the utility of our metrics using simulated data from models with biological relevance and human influenza A/H3N2 virus between temperate and tropical regions.

The performance of the proposed metrics was evaluated through their ability to distinguish different tree distributions quantified by the confusion matrix, mean separation, and mean confusion statistic. We statistically assessed equality in distributions with a nonparametric test. This feature is particularly relevant where analytic expressions for underlying probability distributions are unknown. We note that our nonparametric test assumes samples are independent, and a future direction is to account for the nonindependence of samples. Another important future direction is to develop statistical tests based on our metrics for uncovering specific evolutionary events, for example, inferring mode and strength of natural selection and identifying specific genomic regions undergoing selective pressure, beyond detecting differences in evolutionary histories from tree distributions.

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 MCMC procedures on ranked genealogies. We demonstrate the applicability of our metrics as an MCMC convergence diagnostic with influenza data.

Our proposed distances inherit the properties of L_1 and L_2 norms, but other higher-order norms can also be applied to our matrix representation of ranked tree shapes. In general, the L_2 norm magnifies large differences more than the L_1 norm. Under the scenarios we investigated, no noticeable dissimilarities were observed between L_1- and L_2-based metrics in assessing different tree distributions, but L_2-based metrics overall provided better two-dimensional MDS embeddings by the distortion measure (*SI Appendix*, section 11 and Table S15). For general scenarios beyond those we examined, while we expect both d_1 and d_2 (d_1^w and d_2^w) would perform similarly well in distinguishing tree distributions, rigorous theoretical investigations of their properties will be needed.

We note that our matrix representation can be modified to a symmetric positive definite (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 an SPD matrix representation of ranked tree shapes will make it possible to explore additional distances (62) on ranked tree shapes and ranked genealogies.

Materials and Methods

Random Ranked Tree Shape Generation. To generate random isochronous ranked tree shapes, we used the simulate_tree function in the R package apTreeshape (7) with the following parameters. In distinguishing different tree balance distributions, we simulated 1,000 ranked tree shapes with $n = 100$ leaves for each β value in $\{-1.9, -1.5, -1, 0, 100\}$ with fixed $\alpha = 1$. In distinguishing different internal node ranking distributions, we simulated 1,000 ranked tree shapes with $n = 100$ leaves for each α value in $\{-2, -1, 0, 1, 2\}$ with fixed $\beta = 0$. All other parameters were kept at their default values. Heterochronous ranked tree shapes with different sampling events were generated using phylodyn::coalsim (63) in R.

Random Ranked Genealogy Generation. For all our simulations, we assume neutral evolution. Hence, the distribution of branching or coalescence waiting times depends on the number of extant branches and is independent of the particular topology. In the isochronous coalescent model, when there are k lineages, the coalescent time t_k backward in time from tips to root (Fig. 2*A*) has conditional density

$$f(t_k \mid t_{k+1}) = \frac{\binom{k}{2}}{\lambda(t_k)} e^{-\binom{k}{2} \int_{t_{k+1}}^{t_k} \frac{dt}{\lambda(t)}}, \text{ for } k = n, \ldots, 2, \qquad [10]$$

where $\lambda(t) > 0$ is the effective population size at time t and $t_{n+1} = 0$. In the heterochronous coalescent, sampling times are deterministic and the rate of coalescence depends on the number of extant lineages and the effective population size $\lambda(t)$. More details about coalescent models can be found in *SI Appendix*, sections 9 and 10.

We used phylodyn::coalsim (63) in R to generate random isochronous ranked genealogies with different branch length distributions and random heterochronous ranked genealogies with different sampling events and different branch length distributions.

Comparison with Other Metrics. For comparing our metrics to adaptations of other tree distances, we used the Geodesic Treepath Problem (GTP) software implemented in Java (31) for the BHV distance, multiDist implemented in the R package treespace (10) for the KC distance, and the vecMultiDistUnlab in the R package treetop (14) for the CP distance.

Human Influenza A/H3N2 Virus Analysis. We selected 122 complete HA sequences at random with overlapping collection dates between New York and Southeast Asia—from January 2002 to September 2005. The National Center for Biotechnology Information (NCBI) GenBank accession numbers for the 122 HA sequences from New York and Southeast Asia used in this article are provided in *SI Appendix*, Table S14. We aligned the sequences using the FFT-NS-i option of Multiple Alignment using Fast Fourier Transform (MAFFT) v7.427 software (64, 65).

We used BEAST v1.10.4 (16) to sample from the posterior distribution of model parameters with the following prior settings: Shapiro–Rambaut–Drummond codon-partition substitution model (66) to accommodate rate heterogeneity among sites, the uncorrelated log-normal relaxed molecular clock, and a time-aware Bayesian Skyride prior model (67) on evolutionary histories. For each geographical region, we ran three independent MCMC chains with 100 million steps, with 10% of burn-in, and thinned to obtain a total of 1,000 samples per region. We assessed convergences and effective sample sizes of the estimates using Tracer v1.7.1 (68) and with our distances.

Code and Data Availability. An implementation of our metrics is available at https://github.com/JuliaPalacios/phylodyn. The NCBI GenBank accession numbers for the human influenza A/H3N2 HA sequences are provided in *SI Appendix*, Table S14.

ACKNOWLEDGMENTS. We acknowledge Nina Miolane, Susan Holmes, and Emmanuel Candès for useful discussions. We thank reviewers for helpful comments and Julia Olivieri for her initial feasibility study of one of the proposed distances. 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.

1. R. Sainudiin, T. Stadler, A. Véber, Finding the best resolution for the Kingman-Tajima coalescent: Theory and applications. *J. Math. Biol.* **70**, 1207–1247 (2015).
2. J. A. Palacios *et al.*, Bayesian estimation of population size changes by sampling Tajima's trees. *Genetics* **213**, 967–986 (2019).
3. F. Tajima, Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
4. J. Schweinsberg, Coalescents with simultaneous multiple collisions. *Electron. J. Probab.* **5**, 1–50 (2000).
5. D. Ford, F. A. Matsen, T. Stadler, A method for investigating relative timing information on phylogenetic trees. *Syst. Biol.* **58**, 167–183 (2009).
6. A. Lambert, T. Stadler, Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theor. Popul. Biol.* **90**, 113–128 (2013).
7. O. Maliet, F. Gascuel, A. Lambert, Ranked tree shapes, nonrandom extinctions, and the loss of phylogenetic diversity. *Syst. Biol.* **67**, 1025–1040 (2018).
8. D. Robinson, L. Foulds, Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).
9. L. J. Billera, S. P. Holmes, K. Vogtmann, Geometry of the space of phylogenetic trees. *Adv. Appl. Math.* **27**, 733–767 (2001).
10. M. Kendall, C. Colijn, Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol. Biol. Evol.* **33**, 2735–2743 (2016).
11. J. Chakerian, S. Holmes, Computational tools for evaluating phylogenetic and hierarchical clustering trees. *J. Comput. Graph Stat.* **21**, 581–599 (2012).
12. D. G. Brown, M. Owen, Mean and variance of phylogenetic trees. *Syst. Biol.* **69**, 139–154 (2020).
13. A. Willis, R. Bell, Confidence sets for phylogenetic trees. *J. Comput. Graph Stat.* **27**, 542–552 (2018).
14. C. Colijn, G. Plazzotta, A metric on phylogenetic tree shapes. *Syst. Biol.* **67**, 113–126 (2018).
15. S. Holmes, Statistics for phylogenetic trees. *Theor. Popul. Biol.* **63**, 17–32 (2003).
16. M. A. Suchard *et al.*, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).
17. E. M. Volz, K. Koelle, T. Bedford, Viral phylodynamics. *PLoS Comput. Biol.* **9**, 1–12 (2013).
18. E. C. Holmes, G. Dudas, A. Rambaut, K. G. Andersen, The evolution of Ebola virus: Insights from the 2013-2016 epidemic. *Nature* **538**, 193–200 (2016).
19. M. Worobey *et al.*, 1970s and 'Patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature* **539**, 98–101 (2016).
20. X. Li *et al.*, Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J. Med. Virol.* **92**, 602–611 (2020).
21. M. Giovanetti, D. Benvenuto, S. Angeletti, M. Ciccozzi, The first two cases of 2019-nCoV in Italy: Where they come from?. *J. Med. Virol.* **92**, 518–521 (2020).

22. B. Shapiro *et al.*, Rise and fall of the Beringian steppe bison. *Science* **306**, 1561–1565 (2004).
23. D. Chang, B. Shapiro, Using ancient DNA and coalescent-based methods to infer extinction. *Biol. Lett.* **12**, 20150822 (2016).
24. B. Llamas *et al.*, Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Sci. Adv.* **2**, e1501385 (2016).
25. B. Mühlemann *et al.*, Ancient hepatitis B viruses from the bronze age to the medieval period. *Nature* **557**, 418–423 (2018).
26. S. Peyrégne *et al.*, Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. *Sci. Adv.* **5**, eaaw5873 (2019).
27. D. F. Robinson, L. R. Foulds, "Comparison of weighted labelled trees" in *Combinatorial Mathematics VI*, A. F. Horadam, W. D. Wallis, Eds. (Springer, Berlin/Heidelberg, Germany, 1979), pp. 119–126.
28. M. Li, L. Zhang, Twist–rotation transformations of binary trees and arithmetic expressions. *J. Algorithm* **32**, 155–166 (1999).
29. B. L. Allen, M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. Combinator.* **5**, 1–15 (2001).
30. M. Bordewich, C. Semple, On the computational complexity of the rooted subtree prune and regraft distance. *Ann. Combinator.* **8**, 409–423 (2005).
31. M. Owen, J. S. Provan, A fast algorithm for computing geodesic distances in tree space. *IEEE ACM Trans. Comput. Biol. Bioinf.* **8**, 2–13 (2011).
32. A. F. Y. Poon *et al.*, Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLoS One* **8**, 1–11 (2013).
33. E. Lewitus, H. Morlon, Characterizing and comparing phylogenies from their Laplacian spectrum. *Syst. Biol.* **65**, 495–507 (2015).
34. G. H. Golub, C. F. Van Loan, *Matrix Computations* (Johns Hopkins Studies in the Mathematical Sciences (Book 3), Johns Hopkins University Press, Baltimore, MD, ed. 4, 2013).
35. M. Bacák, Computing medians and means in Hadamard spaces. *SIAM J. Optim.* **24**, 1542–1566 (2014).
36. E. L. Lehmann, J. P. Romano, *Testing Statistical Hypotheses* (Springer, New York, NY, ed. 3, 2005).
37. D. Aldous, "Probability distributions on cladograms" in *Random Discrete Structures*, D. Aldous, R. Pemantle, Eds. (Springer, New York, NY, 1996), pp. 1–18.
38. M. Kirkpatrick, M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* **47**, 1171–1181 (1993).
39. A. O. Mooers, S. B. Heard, Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.* **72**, 31–54 (1997).
40. D. J. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.* **16**, 23–34 (2001).
41. M. G. B. Blum, O. François, Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Syst. Biol.* **55**, 685–691 (2006).
42. N. Galtier, F. Depaulis, N. H. Barton, Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155**, 981–987 (2000).
43. M. Przeworski, The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179–1189 (2002).
44. H. Li, T. Wiehe, Coalescent tree imbalance and a simple test for selective sweeps based on microsatellite variation. *PLoS Comput. Biol.* **9**, e1003060 (2013).
45. L. Ferretti, A. Ledda, T. Wiehe, G. Achaz, S. E. Ramos-Onsins, Decomposing the site frequency spectrum: The impact of tree topology on neutrality tests. *Genetics* **207**, 229–240 (2017).
46. G. E. Leventhal *et al.*, Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput. Biol.* **8**, 1–10 (2012).
47. S. D. W. Frost, E. M. Volz, Modelling tree shape and structure in viral phylodynamics. *Philos. Trans. Biol. Sci.* **368**, 20120208 (2013).
48. C. Colijn, J. Gardy, Phylogenetic tree shapes resolve disease transmission patterns. *Evol. Med. Public Health* **2014**, 96–108 (2014).
49. D. L. Rabosky, I. J. Lovette, Explosive evolutionary radiations: Decreasing speciation or increasing extinction through time?. *Evolution* **62**, 1866–1875 (2008).
50. A. L. Pigot, A. B. Phillimore, I. P. F. Owens, C. D. L. Orme, The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. *Syst. Biol.* **59**, 660–673 (2010).
51. F. Gascuel, R. Ferrière, R. Aguilée, A. Lambert, How ecology and landscape dynamics shape phylogenetic trees. *Syst. Biol.* **64**, 590–607 (2015).
52. O. Hagen, K. Hartmann, M. Steel, T. Stadler, Age-dependent speciation can explain the shape of empirical phylogenies. *Syst. Biol.* **64**, 432–440 (2015).
53. K. Mardia, Some properties of classical multi-dimensional scaling. *Commun. Stat. Theor. Methods* **7**, 1233–1241 (1978).
54. G. U. Yule, A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philos. Trans. R. Soc. Lond. Ser. B* **213**, 21–87(1925),.
55. E. F. Harding, The probabilities of rooted tree-shapes generated by random bifurcation. *Adv. Appl. Probab.* **3**, 44–77 (1971).
56. D. Vijaykrishna *et al.*, The contrasting phylodynamics of human influenza B viruses. *eLife* **4**, e05055 (2015).
57. J. D. Tamerius *et al.*, Environmental predictors of seasonal influenza epidemics across temperate and tropical climates. *PLoS Pathog.* **9**, e1003194 (2013).
58. J. Bahl *et al.*, Temporally structured metapopulation dynamics and persistence of influenza A H3N2 virus in humans. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 19359–19364 (2011).
59. M. D. Karcher, J. A. Palacios, T. Bedford, M. A. Suchard, V. N. Minin, Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Comput. Biol.* **12**, e1004789 (2016).
60. M. D. Karcher, M. A. Suchard, G. Dudas, V. N. Minin, Estimating effective population size changes from preferentially sampled genetic sequences. arXiv:1903.11797 (28 March 2019).
61. K. V. Parag, L. du Plessis, O. G. Pybus, Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. *Mol. Biol. Evol.* **37**, 2414–2429 (2020).
62. R. Vemulapalli, D. W. Jacobs, Riemannian metric learning for symmetric positive definite matrices. arXiv:1501.02393 (10 January 2015).
63. M. D. Karcher, J. A. Palacios, S. Lan, V. N. Minin, phylodyn: An R package for phylodynamic simulation and inference. *Mol. Ecol. Resour.* **17**, 96–100 (2017).
64. K. Katoh, K. Misawa, Ki. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
65. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
66. B. Shapiro, A. Rambaut, A. J. Drummond, Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol. Biol. Evol.* **23**, 7–9 (2005).
67. V. N. Minin, E. W. Bloomquist, M. A. Suchard, Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471 (2008).
68. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).