A full characterization of evolutionary tree topologies

C. Colijn[1*] and G. Plazzotta[1]

July 11, 2016

1 Department of Mathematics, Imperial College, London SW7 2AZ, UK.
* Corresponding author: c.colijn@imperial.ac.uk.

Abstract

The topologies of evolutionary trees are shaped by the nature of the evolutionary process, but comparisons of trees from different processes are hindered by the challenge of completely describing tree topology. We present a full characterization of the topologies of rooted branching trees in a form that lends itself to natural tree comparisons. The resulting metric distinguishes trees from random models known to produce different tree topologies. It separates trees derived from tropical vs USA influenza A sequences, indicating that the different epidemiology of tropical and seasonal flu leaves strong signatures in the tree topology. Our approach allows us to construct addition and multiplication on trees, and to create a convex metric on tree topologies which formally allows computation of average trees.

1 Introduction

The availability and declining cost of DNA sequencing mean that data on the diversity, variation and evolution of organisms is more widely available than ever before. Increasingly, thousands of organisms are being sequenced at the whole-genome scale [1, 2, 3]. This has had particular impact on the study of pathogens, whose evolution occurs rapidly enough to be be observed over relatively short periods. As the numbers of sequences gathered annually grow to the tens of thousands in many organisms, comparing this year's evolutionary and diversity patterns to previous years', and comparing one location to another, has become increasingly challenging. Despite the fact that evolution does not always occur in a tree-like way due to the horizontal movements of genes, phylogenetic trees remain a central tool with which we interpret these data.

The topologies of phylogenetic trees are of long-standing interest in both mathematics and evolution [4, 5, 6, 7, 8, 9, 10, 11]. A tree's topology refers to the tree's connectivity structure, without reference to the lengths of its branches. A key early observation was that trees reconstructed from evolutionary data are more asymmetric than simple models predict. This spurred an interest in ways to measure tree asymmetry [8, 12, 13, 14, 15], in the power of asymmetry measures to distinguish between random models [16, 8, 17], and in constructing generative models of evolution that produce imbalanced trees [13, 18, 10]. Tree topologies carry information about the underlying evolutionary processes, and distributions of tree topologies under simple null models can be used to test hypotheses about evolution [9, 10, 19, 7, 11]. Recent work also relates fitness, selection and a variety of ecological processes to tree topology [20, 21, 22, 23, 24, 18]. An additional motivation for studying the topologies of phylogenetic trees is that reconstructing branch lengths is challenging, particularly deep in a tree; there may be weak support for

a molecular clock, and coalescent inference procedures may produce trees with consistent topology but differing root heights.

Tree topology is well established as carrying important information about macroevolutionary processes, but also carries information about evolution in the short term. In the context of pathogens, diversity patterns represent a combination of neutral variation that has not yet become fixed, variation that is under selection, complex demographic processes (host behaviour and contact patterns), and an array of ecological interactions. The extent to which tree topologies are informative of these processes is not well understood, though there have been studies on the frequency of cherries and tree imbalance [25, 26, 27] and simulation studies aiming to explore the question [29, 28, 30, 31].

A key limitation in relating tree topologies to evolution and ecology has been the limited tools with which trees can be quantified and compared. Comparing tree topologies from different models of evolution or from different datasets requires comparing *unlabelled* trees, whereas most established tree comparison methods (eg the Robinson-Foulds [32] and Billera-Holmes-Vogtmann [33] metrics) compare trees with one particular set of organisms at the tips (ie one set of taxa, with labels). The tools at our disposal to describe and compare tree topologies from *different* sets of tips are limited, and have focused on scalar measures of overall asymmetry [5, 34, 17, 14, 12, 15, 35, 36] and on the frequencies of small subtree topologies such as cherries [37, 31, 25] and r-pronged nodes [38]. Recently, kernel [39] and spectral [40] approaches also have been used.

Here we give a simple and complete characterization of all possible topologies for a rooted tree. Our scheme gives rise to natural metrics (in the sense of true distance functions) on unlabelled tree topologies. It provides an efficient way to count the frequencies of sub-trees in large trees, and hence can be used to compare empirical distributions of sub-tree topologies. It is not limited to binary trees and can be formulated for any maximum size multifurcation, as well as for trees with internal nodes with only one descendant (sampled ancestors). The resulting topology-based tree metrics separate trees derived from different random tree models. We use the approach to compare trees from human influenza A (H3N2); it can distinguish between trees from influenza sampled in the tropics vs that sampled in the USA.

2 Results

Briefly, with details in Materials and Methods, our approach is to label any possible tree topology, traversing the tree from the tips to the root and assigning labels as we go. The simplest case is to assume a binary tree, in which all internal nodes have two descendants. We give a tip the label 1. For every internal node, we list its descendants' labels (k, j). Using lexicographic sorting, list all possible labels (k, j): $(1), (1, 1), (2, 1), (2, 2), (3, 1), (3, 2), (3, 3), ...$ We define the label of a tree topology whose root node has descendants (K, J) to be the index at which (K, J) appears in this list. Accordingly, a "cherry" (a node with two tip descendants) is labelled 2 because its descendants are $(1, 1)$, which is the second entry in the list. A node with a cherry descendant and a tip descendant (a $(2, 1)$, or a pitchfork) has label 3. The tree topology (k, j) (a tree whose root has a descendant with label k and one with label j) has label $\phi_2(k, j) = \frac{1}{2}k(k-1) + j + 1$. The scheme takes a different explicit form if there are multifurcations or internal nodes with a single descendant, but proceeds in the same way (see Supporting Information; the form for trees with no multifurcations but allowing for internal nodes with one descendant is $\hat{phi}_2(k, j) = \frac{1}{2}k(k+1) + j + 1$). We continue until the root of the tree has a label.

Figure 1 illustrates the labels at the nodes of two binary trees. The label of the root node uniquely defines the tree topology. Indeed, tree isomorphism algorithms use similar labelling [41, 42, 43, 44, 45]. If R_a and R_b are the root nodes of binary trees T_a and T_b, the tree topologies are the same if and only if $\phi_2(R_a) = \phi_2(R_b)$. The map between trees and labels is bijective: every positive integer corresponds

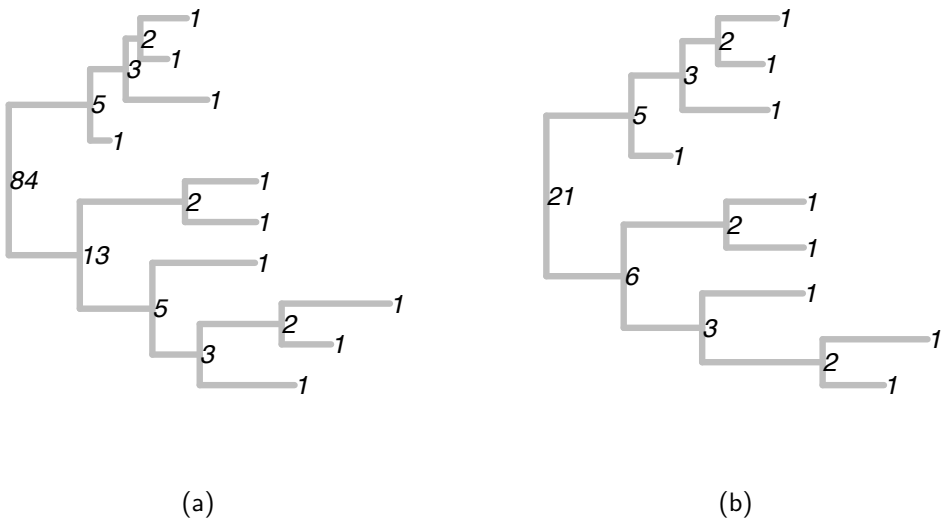

Figure 1: Illustration of the labels of the nodes of binary trees under the full bifurcating tree model $\phi_2(k, j) = \frac{1}{2}k(k-1) + j + 1$. Tips have the label 1. Labels of internal nodes are shown in black. The only difference between the trees in (a) and (b) is that in (b), the bottom-most tip from (a) has been removed. As a consequence, most of the labels are the same.

to a unique tree topology and vice versa.

Metrics are an appealing way to compare sets of objects; defining a metric defines a *space* for the set of objects – in principle allowing navigation through the space, study of the space's dimension and structure, and the certainty that two objects occupy the same location if and only if they are identical. The labelling scheme gives rise to several natural metrics on tree topologies, based on the intuition that tree topologies are similar when they share many subtrees with the same labels. In the context of relating tree topologies to underlying evolutionary processes, a useful metric will be one that both distinguishes trees from processes known to produce distinct topologies, and that fails to distinguish trees from processes known to produce the same distribution of tree topologies.

There are several ways to sample random trees, known to produce trees of different topologies. These include models capturing equal vs different speciation rates, continuous time birth-death processes with different rates and others (see Methods). We used the metric arising from our labelling scheme to compare these. Figure 2 shows a visualization of the tree-tree distances between trees from different random models. The metric groups trees from each process together and distinguishes between them well. Summary statistics such as tree imbalance also distinguish some of these groups well (particularly the PDA, Aldous, Yule and biased speciation model), but imbalance does not substantially differ between the continuous-time branching models.

We also compared trees inferred from sequences of the HA protein in influenza A H3N2 sequences. Influenza A is highly seasonal outside the tropics [46], with the majority of cases occurring in winter. In contrast, there is little seasonal variation in transmission in the tropics. In addition, over long periods of time, influenza evolves in response to pressure from the human immune system, undergoing evolution particularly in the surface HA protein. This drives the 'ladder-like' shape of long-term influenza phylogenies [47, 25, 48, 49], but would not typically be apparent in shorter-term datasets. With this motivation, we compared tropical samples to USA samples, and recent (2010-2015) global samples to early samples (pre-2010). Figure 3 shows that the tropical and USA flu trees are well separated by the metric. In

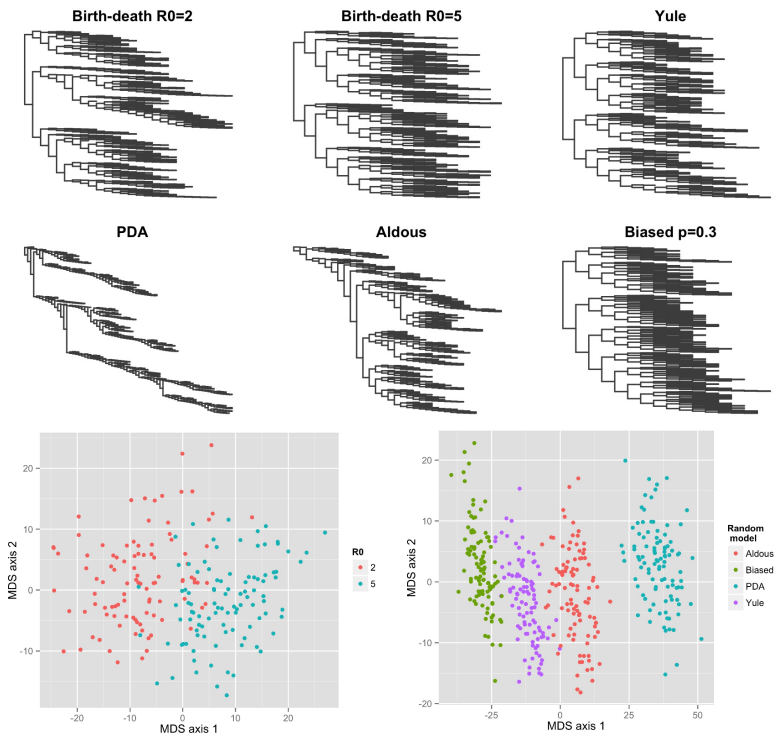

Figure 2: Top: Six sample trees, one from each of six different random processes. Bottom: Multi-dimensional scaling (MDS) plots showing that trees from each process are grouped together in the metric. Bottom left: trees with 700 tips under a birth-death model with different values of $R_0 = \lambda/\mu$. Bottom right: trees derived from the Yule, proportional to distinguishable arrangements (PDA), Aldous and biased models, each with 500 tips.

contrast, the five-year (post-2010) and pre-2010 global samples occupy different regions of the projected tree space, but have some overlap. In these groups of trees, the underlying processes are similar, but the time frames and sampling density differ.

Natural metrics associated with the labelling scheme are all based on the bijective map ϕ between the tree space $\mathbb{T}$ and the natural numbers $\mathbb{N}$. Composing ϕ with bijective maps between $\mathbb{N}$ and other countable sets like the integers ($\mathbb{Z}$), the positive rational numbers ($\mathbb{Q}^+$), or the rationals ($\mathbb{Q}$) opens up further possibilities because we can take advantage of the properties (addition, multiplication, distance, etc) of integer and rational numbers. If ψ is a bijective map between $\mathbb{N}$ and one of these sets, then the composition $\psi \circ \phi$ is also bijective, and we can use it to define addition and multiplication operations on trees:

$$\begin{aligned} T_1 +^T T_2 &= \psi^{-1}\big(\phi(T_1) + \phi(T_2)\big), \\ T_1 \cdot^T T_2 &= \psi^{-1}\big(\phi(T_1) \cdot \phi(T_2)\big), \end{aligned} \tag{1}$$

where $+$ and $\cdot$ are the usual addition and multiplication. Now the space of trees together with these definitions of addition and multiplication, $\big(\mathbb{T}, +^T, \cdot^T\big)$, inherits all the algebraic properties of the set it is mapped into. For instance, $\big(\mathbb{T}, +^T, \cdot^T\big)$ is a commutative ring if $\psi : \mathbb{N} \to \mathbb{Z}$. These constructions allow algebraic operations in the tree space $\mathbb{T}$. However the choice of the map ψ determines whether these operations are "meaningful" or "helpful" for applications of branching trees in biology or other fields. It

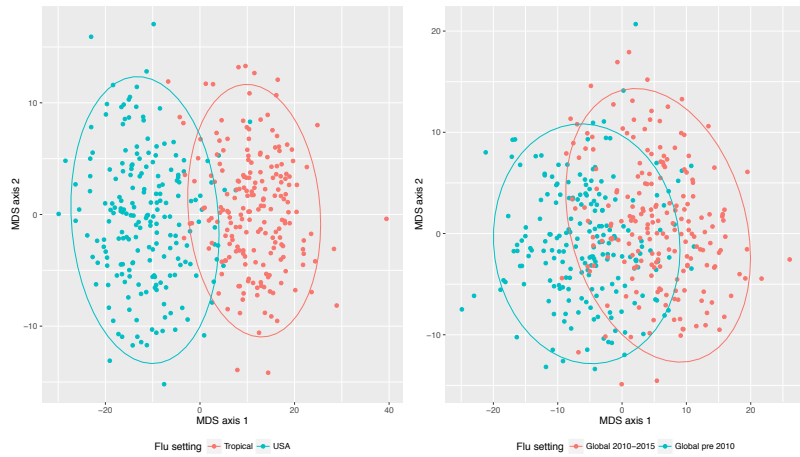

Figure 3: Comparisons between trees from H3N2 flu virus samples. Left: isolates from the tropics (red) and from the USA (blue) are separated. Right: isolates from 1979-2010 (blue) are distinct from those of 2010-2015 but there is some overlap.

turns out that the selection of a meaningful map is challenging.

For example, we can use the labelling scheme to map tree topologies to the (positive and negative) integers. We first extend ϕ with $\phi(0) = \emptyset$, i.e. the the empty tree no tips. Consider the following well-known map between $\mathbb{N}$ and $\mathbb{Z}$:

$$\psi_{\mathbb{Z}} : n \to \begin{cases} \frac{n}{2} & \text{if } n \text{ is even} \\ -\frac{n+1}{2} & \text{if } n \text{ is odd} \end{cases}.$$

$\psi_{\mathbb{Z}}$ is clearly bijective: each tree topology is mapped to a unique integer and each integer corresponds to a unique tree topology. A representation of ten trees is provided in Figure S1. To "add" or "multiply" trees, we can add or multiply their corresponding integers and then invert, as in Eq (1). This may seem intuitive for small trees; for example the sum of tree number 3 and tree number -1 gives tree number 2 which has one fewer tip than tree number 3. For larger trees, however, addition and multiplication operations are less intuitive and do not follow the numbers of tips.

Mapping tree topologies to other sets of numbers can help us to capture the space of tree topologies in new ways. A particularly nice property of a metric space is convexity - if given two trees T_1 and T_2, there exists a tree T_3 lying directly between them, i.e. $d(T_1, T_3) + d(T_3, T_2) = d(T_1, T_2)$. Convex metrics are appealing because in a convex metric on tree topologies we can find the average tree topology for a set of trees, define a centre of mass topology, and further develop statistics on the space of tree topologies.

We use the labelling scheme and a pairing of maps to construct a convex metric on tree topologies. To do this, we map tree topologies to the rational numbers, where the usual absolute value function is a convex metric (as there is always a rational number directly in between any two others). We use the prime decomposition, i.e. the unique product of prime factors of a number (e.g. $10 = 2 \cdot 5$). For a tree topology corresponding to integer n, we apply $\psi_{\mathbb{Z}}$ to the exponents of all the prime factors of $n + 1$, and multiply the result (see Methods). For example $\psi_{\mathbb{Q}^+}(19) = 2^{\psi_{\mathbb{Z}}(2)} 5^{\psi_{\mathbb{Z}}(1)} = 2^1 5^{-1} = 5/2$. We denote this map $\psi_{\mathbb{Q}}$; it takes each integer to a unique rational number, and vice versa (bijective). Applying $\psi_{\mathbb{Q}^+} \circ \phi$ to tree topologies maps them bijectively to the non-negative rational numbers. We can add or multiply trees' corresponding rational numbers to perform operations in the space of tree topologies. In particular, we can use the usual absolute value distance function to define a convex metric space of tree topologies $\left(\mathbb{T}, d^T\right)$:

$$d^T(T_1, T_2) = \left|\psi_{\mathbb{Q}^+}\left(\phi(T_1)\right) - \psi_{\mathbb{Q}^+}\left(\phi(T_1)\right)\right|.$$

In this space we can find the average tree of a group of trees, and a 'direct path' between two trees. Given n trees, the average tree is:

$$T_m = \phi^{-1} \circ \psi_{\mathbb{Q}^+}^{-1} \left(\frac{\sum_{i=1}^{n} \psi_{\mathbb{Q}^+} \circ \phi(T_i)}{n} \right).$$

In other words, the average of a set of trees is the tree corresponding to the average of the trees' rational numbers under the map we have defined. Figure 4 illustrates this operation.

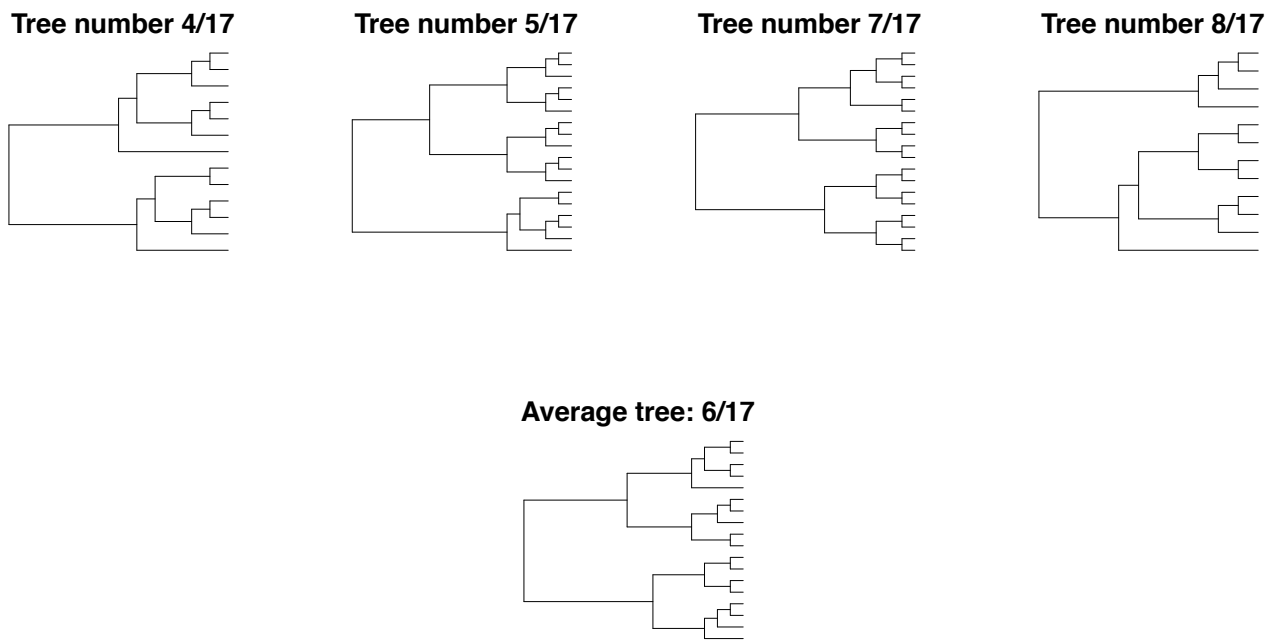

Figure 4: Trees associated to the rationals 4/17, 5/17, 7/17, 8/17, using the map in Example 2. Because the natural distance is convex in $\mathcal{Q}^+$, it is possible to find the "average" tree, which is the one mapped into 6/17. Moreover, trees mapped to 5/17, 6/17 and 7/17 are part of the direct path between the trees mapped to 4/17 and 8/17

There are infinitely many ways that we could map tree topologies to rational numbers. Any of them would give rise to a *convex* metric on the set of tree topologies. It would be most desirable if the resulting metric had some intuitive features - for example, if the trees lying directly between trees T_1 and T_2 (with n_1 and n_2 tips) had an intermediate number of tips between n_1 and n_2. The convex metric we have constructed does not have this particular intuitive property. This convex metric also relies on the prime factorisation of the tree labels, which is a challenge if large labels are encountered.

3 Discussion

The labelling scheme we present comprises a complete characterization of rooted tree topologies, not limited to fully bifurcating trees. Trees from processes known to produce different topologies are well separated in the metric that arises naturally from the scheme. This suggests applications in inferring evolutionary processes and to detecting tree shape bias [50, 24, 4]. The structure and simplicity of this comparison tool carry a number of advantages. Metrics have good resolution in comparing trees because the distance is only zero if tree topologies are the same. Empirical distributions of sub-tree topologies can easily be found and compared. And as we have shown, the approach can be extended to convex

metrics on tree topologies, allowing averaging as well as algebraic operations (addition, etc) in tree space. However, this approach does not seem likely to give rise to analytically tractable distributions of tree-tree distances, and in some cases, may not offer more useful resolution than a well-chosen collection of summary statistics.

Scalar measures of asymmetry are insufficient to characterize tree topologies. Here, imbalance measures do not distinguish between the continuous-time birth-death models with $R0 = 2, 5$ but are quite different between the random processes, whereas the metric distinguishes all cases. Matsen [51] developed a method to define a broad range of tree statistics. Genetic algorithms uncovered tree statistics that can distinguish between the reconstructed trees in TreeBase [52] and trees from Aldous' β-splitting model, whereas imbalance measures do not [10]. However, the search-and-optimize approach is vulnerable to over-fitting, as the space of tree statistics is large. It is also reasonable to believe that due to ongoing decreases in the cost of sequencing, studies will increasingly analyze large numbers of sequences and reconstructed trees will have many tips. Any single scalar measure will likely be insufficient to capture enough of the information in these large trees to perform inference, motivating the development of metric approaches.

Large trees present a problem for many approaches to inference, including phylodynamic methods that rely on computationally intensive inference methods. In contrast, our scheme is better able to distinguish between groups of large trees than small ones (fewer than 100 tips). The tip-to-root traversal means that it is very efficient to construct the label set on very large trees (and the same traversal could, with little additional computation time, compute other properties that are naturally computed from tip to root, such as clade sizes, some imbalance measures and many of Matsen's statistics [51]). However, due to the large number of tree topologies, the labels themselves become extremely large even for relatively small trees. Our implementation used MD5 hashing to solve this problem, but hashing removes the ability to reconstruct the tree from its label. Also, there are $2^{128} \approx 3 \cdot 10^{38}$ possible hashed strings, which while large is less than the number of possible tree topologies, even restricting to 500 tips. Alternative labelling schemes may partially alleviate this, for example by subtracting from the label the minimum label for n tips, and only comparing trees of size n or greater. A related approach was used by Furnas [53] in developing algorithms to sample trees.

The large size of the labels is also a challenge when they are mapped to $\mathbb{Z}$, $\mathbb{Q}^+$ or $\mathbb{Q}$ to define a tree algebra or a convex metric. Small changes in the label value can determine visible changes in the topologies. Because the bijective maps are sensitive to small perturbations, the implementation requires the full label, with no hashing compression. However, for trees with 500 tips, we encountered labels of about one million digits. Handling such large numbers with full accuracy required heavy and slow computation. The search for the average tree as found in Figure 4 was only possible for small trees, as the map requires the prime factorization of the label.

Perhaps as it should be, the dominant difference in our scheme between a tree with ten tips and one with one hundred tips is the size of the tree. In this work we have chosen to detect differences that are not simply a reflection of the size of the tree. If we relax this constraint, the largest contribution to the distances will result from comparing the number of instances of the label 1 (tip) in two trees; this is necessarily larger than any other label copy number, and furthermore, a tree with more tips can have more cherries, pitchforks and any other subtree than a tree with fewer tips. It is straightforward to modify the metric d_2 to be relatively insensitive to tree size (see Supporting Information).

Our scheme captures only the topology of the trees; there does not appear to be a natural way to incorporate branch lengths. One option is to add one or several terms to the distance function to incorporate more information (see Supporting Information). Linear combinations of our distances and other tree comparisons may turn out to be the most powerful approach to comparing unlabeled trees, allowing the user to choose the relative importance of scalar summaries, tree topology, spectra and so

on while retaining the discriminating power of a metric. Ultimately, discriminating and informative tools for comparing trees will be essential for inferring the driving processes shaping evolutionary data.

4 Methods

4.1 Definitions

A *tree topology* is a tree (a graph with no cycles), without the additional information of tip labels and branch lengths. We use the same terminology as Mooers and Heard [9]. We consider rooted trees, in which there is one node specified to be the root. Tips, or leaves, are those nodes with degree 1. A *rooted tree topology* is a tree topology with a vertex designated to be the root. We use "tree topology", as we assume rootedness throughout. Typically, edges are implicitly understood to be directed away from the root. A node's *descendants* are the node's neighbors along edges away from the root. A *multifurcation*, or a *polytomy*, is a node with more than two descendants, and its *size* is its number of descendants (> 2). Naturally, a rooted phylogeny defines a (rooted) tree topology if the tip labels and edge weights are discarded. Phylogenies typically do not contain internal nodes with fewer than two descendants (sampled ancestors), but we allow this possibility in the tree topologies.

4.2 Labelling scheme

We label each tree topology according to the labels of the two clades descending from the root. In the simplest case (full binary trees), we call this label function ϕ_2:

$$\phi_2(k, j) = \tfrac{1}{2}k(k-1) + j + 1. \tag{2}$$

The subscript 2 specifies that each node has a maximum of two descendants; the scheme can be extended to any fixed maximum number M of descendants, but then the explicit form of the label (ϕ_M) is different.

4.3 Metrics on the space of rooted unlabelled shapes

There are several natural metrics suggested by our characterisation of tree topologies. Given two binary trees T_a and T_b, we can write

$$d_0(T_a, T_b) = |L(R_a) - L(R_b)|. \tag{3}$$

Clear d_0 is symmetric and non-negative. The tree isomorphism algorithm and the above labelling clearly show that $d_0 = 0 \Leftrightarrow T_a = T_b$ and the absolute value obeys the triangle inequality. However, it is not a particularly useful metric, in the sense that a large change in root label can result from a relatively "small" change, in intuitive terms, in the tree topology (such as the addition of a tip).

While each tree is defined by the label of its root, it is also defined (redundantly) by the labels of all its nodes. If the tree has n tips, the list of its labels contains n 1s, typically multiple 2s (cherries) and so on. Let L_a denote the list of labels for a tree T_a: $L_a = \{1, 1, 1, ..., 2, 2, ..., \phi_2(R_a)\}$. The label lists are multisets because labels can occur multiple times. Define the distance d_1 between T_a and T_b to be the number of elements in the symmetric set difference between the label lists of two trees:

$$d_1(T_a, T_b) = |L_a \Delta L_b|. \tag{4}$$

Intuitively, this is the number of labels not included in the intersection of the trees' label lists. Formally, the symmetric set difference $A\Delta B = (A \cup B) \backslash (A \cap B)$ is the union of A and B without their intersection.

If A and B are multisets with A containing k copies of element x and B containing m copies of x, with $k > m$, we consider $A \cap B$ to contain m copies of x (these are common to both A and B). $A \Delta B$ has the remaining $k - m$ copies. Each tree's label list contains more 1s (tips) than any other label. Accordingly, this metric is most appropriate for trees of the same size, because if trees vary in size, the metric can be dominated by differences in the numbers of tips. For example, if $L_a = \{1, 1, 1, 1, 2, 2\}$ (four tips joined in two cherries) and $L_b = \{1, 1, 1, 2, 3\}$ (three tips, i.e. a pitchfork), then $L_a \Delta L_b = \{1, 2, 3\}$, because there is a 1 and a 2 in L_a in excess of those in L_b, and a 3 in L_b that is not matched in L_a.

Like d_0, d_1 is a metric: positivity and symmetry are clear from the definition. The cardinality of the symmetric difference is 0 if and only if the two sets are the same, in which case the root label is the same and the tree topologies are the same. That the symmetric difference obeys the triangle inequality is readily seen from the property $A \Delta C \subset (A \Delta B) \cup (B \Delta C)$.

Another natural metric that the labelling scheme induces is the L2 norm of the difference between two vectors counting the numbers of occurrences of each label. Let v_a be a vector whose k'th element $v_a(k)$ is the number of times label k occurs in the tree T_a. Define the metric

$$d_2(T_a, T_b) = ||v_a - v_b||. \tag{5}$$

Positivity, symmetry and the triangle inequality are evident, and again d_2 can only be 0 if T_a and T_b have the same number of copies of all labels (including the root label), which is true if and only if T_a and T_b have the same topology. This has a similar flavour to the statistic used to compare trees to Yule trees in [10], where the numbers of clades of a specific size were compared. We have used metric d_2 in the analyses presented in the Results.

4.4 Mapping tree topologies to the integers and rationals

Figure S1 illustrates tree topologies together with their labels under the map $\psi_{\mathbb{Z}}$. We use this map and a map to the rational numbers to define a convex metric on tree topologies.

Define the following map from $\mathbb{N}$ to $\mathbb{Q}$: $\psi_{\mathbb{Q}^+} : n \to \prod_{i=1}^{\infty} p_i^{\psi_{\mathbb{Z}}(a_i)}$ if $n > 0$, or 0 if $n = 0$. Here, p_i are all the prime numbers and $\prod_{i=1}^{\infty} p_i^{a_i}$ is the unique prime decomposition of $n + 1$. $\psi_{\mathbb{Z}}$ is as defined above, mapping the positive integers to all integers by $\psi_{\mathbb{Z}} = n/2$ if n is even and $-(n+1)/2$ if n is odd. For example $\psi_{\mathbb{Q}^+}(11) = 2^{\psi_{\mathbb{Z}}(2)} 3^{\psi_{\mathbb{Z}}(1)} = 2^1 3^{-1} = 2/3$. $\psi_{\mathbb{Q}^+}$ is injective, from the uniqueness of the prime factorization and the injectivity of $\psi_{\mathbb{Z}}$. Since any rational number is a ratio of integers and $\psi_{\mathbb{Z}}$ maps to all integers the map is also surjective, and hence bijective. Therefore $\psi_{\mathbb{Q}^+} \circ \phi$ maps tree topologies bijectively to the non-negative rational numbers. In turn, $\mathbb{T}$ inherits all of the properties and structure of $\mathbb{Q}^+$. A distance metric d^T on $\mathbb{T}$ can be defined from the usual distance $|\cdot|$ of $\mathbb{Q}$:

$$d^T(T_1, T_2) = \left| \psi_{\mathbb{Q}^+}\big(\phi(T_1)\big) - \psi_{\mathbb{Q}^+}\big(\phi(T_2)\big) \right|.$$

Because the absolute value is a convex metric in $\mathbb{Q}$, this is a convex metric on unlabelled tree topologies. It can be used to find averages of a set of trees.

4.5 Simulations

We compared trees from different random processes and models. One of the most natural random processes modelling phylogenetic trees is the continuous-time homogeneous birth-death branching process, in which each individual gives rise to a descendant at a constant rate in time, and also risks removal (death) at a constant rate. With birth rate λ and death rate μ, the ratio λ/μ specifies the mean number of offspring of each individual in this process, and affects the topologies and branching times of the

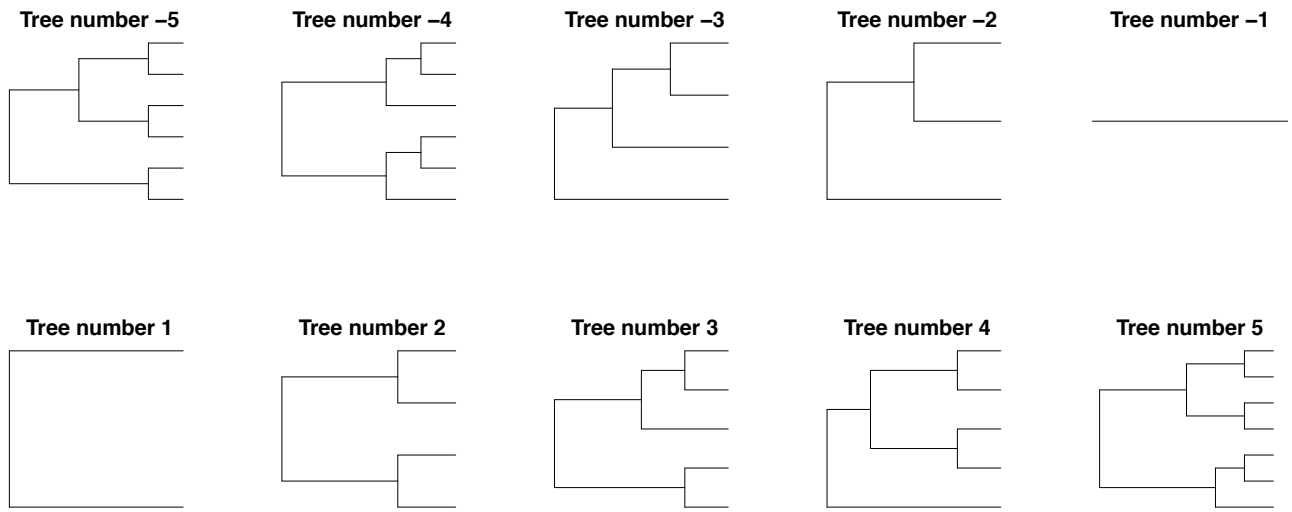

Figure S1: Some trees and their associated integers using the map $\psi_{\mathbb{Z}}$ of Example 1. The numbering goes from -5 to 5, with the exception of 0 which corresponds to the "empty tree".

resulting branching trees. In the epidemiological setting, the link to branching times has been used to infer the basic reproduction number R_0 from sequence data [54, 55]. We computed the distances between trees derived from constant-rate birth-death (BD) processes simulated in the package `TreeSim` in R [56]. One challenge is that the number of tips in the BD process after fixed time is highly variable and depends on λ/μ. We aimed to detect shape differences that were not dominated by differences in the number of tips. Accordingly, we conditioned the processes to have 1500 taxa and then pruned tips uniformly at random to leave 700 tips remaining.

There are several other random models for trees. The Yule model is a model of growing trees in which lineages divide but do not die; in terms of tree topology it is the same as the Kingman coalescent and the equal rates Markov models. In the 'proportional to distinguishable arrangements' (PDA) model, each unlabelled topology is sampled with probability proportional to the number of *labelled* trees on n tips with that unlabelled topology [57, 9]. The "biased" model is a growing tree model in which a lineage with speciation rate r has descendant lineages with speciation rates pr and $(1-p)r$. The Aldous' branching model that we use here is Aldous' β-splitting model with $\beta = -1$ [58]; in this model a β distribution determines the (in general asymmetric) splitting densities upon branching. The Yule, PDA, biased and Aldous $\beta = -1$ models are available in the package `apTreeshape` in R [59]. We used $p = 0.3$ for the biased model, and sampled trees with 500 tips.

4.6 Data

We aligned data of HA protein sequences from human influenza A (H3N2) in different settings reflecting different epidemiology. Data were downloaded from NCBI on 22 Jan 2016. In all cases we included only full-length HA sequences for which a collection date was available. The USA dataset ($n = 2168$) included USA sequences collected between March 2010 and Sept 2015. The tropical data ($n = 1388$) included sequences from the tropics collected between January 2000 and October 2015. The global set ($n = 8100$) included all (full-length HA, with date); dates ranged from July 1979 - Sept 2015. We also selected global sequences (from anywhere) within a five-year window (August 2010 - December 2015; $n = 2892$). Accession numbers are included in the Supporting Information. Fasta files were aligned with mafft. Within each dataset, we sampled 500 taxa uniformly at random (repeating 200 times) and inferred

a phylogenetic tree with the program FastTree. Where necessary we re-aligned the 500 sequences before tree inference. This resulted in 200 trees, each with 500 tips, from four datasets: global (pre-2010), USA, tropical and the five-year sample (global 2010-2015).

5 Supporting Information

5.1 Extension to multifurcations and sampled ancestors

A polytomy, or multifurcation, is an internal node with more than two descendants. In extending the scheme to handle polytomies we also extend it to allow for internal nodes with only one descendant.

We first explicitly work out the case where the maximum-size multifurcation is 4. Let 0 be the empty tree. Nodes may have 0, 1, 2, 3, or 4 descendants, and we write a general tree as (k, j, l, m), where k, j, l and m are the labels of the four trees descending from the root. Some of these may be empty (0) as not every node is a four-fold polytomy. As in the binary case, we use the convention that $k \geq j \geq l \geq m$, and sort the length-four strings lexicographically. Every possible tree T with a maximum-size multifurcation of four has a unique label $L_4(T)$ in this list. We seek to find an explicit expression for the label $L_4(T)$ – the order in the list – for the tree (k, j, l, m).

The number of possible labels in the scheme with four characters, starting with k and sorted lexicographically, is $\binom{k+3}{k}$. To see this, note that each (k, j, l, m) with $k \geq j \geq l \geq m$ can be thought of as a path on a lattice, starting on the left at height k and descending to height 0 after three horizontal steps. The path has a total length of $k + 3$ steps, and of these, three must be steps to the right and k must be downward. The number of such paths is the number of ways of placing three rightwards steps amongst $k + 3$ steps, ie. $\binom{k+3}{k}$. Extending this, we obtain the label L_4 of the tree $(k + 1, 0, 0, 0)$, noting that $L_4(k, k, k, k)$ is the sum of the numbers of labels beginning with 1, 2, ... k. $L_4(k + 1, 0, 0, 0) = 1 + L_4(k, k, k, k)$ (and we write 1 as $\binom{3}{3}$):

$$L_4(k + 1, 0, 0, 0) = \sum_{x=0}^{k} \binom{x + 3}{3}.$$

Rewriting the sum and making use of the identity $\sum_{y=0}^{k+c} \binom{y}{c} = \binom{k+c+1}{c+1}$, we have

$$L_4(k + 1, 0, 0, 0) = \sum_{x=0}^{k} \binom{x + 3}{3} = \sum_{y=3}^{k+3} \binom{y}{3} = \sum_{y=0}^{k+3} \binom{y}{3} = \binom{k + 4}{4}.$$

To obtain $L_4(k, j, l, m)$, we note that

$$L_4(k, j, l, m) = L_4(k, 0, 0, 0) + L_3(j, 0, 0) + L_2(l, m).$$

Following the same logic, this is

$$L_4(k, j, l, m) = \binom{k + 3}{4} + \binom{j + 2}{3} + \binom{l + 1}{2} + m.$$

As in the binary case, the labels will grow unfeasibly large, but in principle this is a bijective map between trees whose maximum-size polytomy is four and the non-negative integers.

Naturally, there is nothing special about size-four polytomies. If the maximum size is c, the scheme is

$$L_c(x_c, x_{c-1}, x_{c-2}, .., .x_1) = \sum_{i=1}^{c} \binom{x_i + i - 1}{i}.$$

5.2 Extensions of the metric

As noted in the main text, the metrics d_1 and d_2 will be dominated by differences in the sizes of trees. It may be desirable to construct unlabelled metrics that are useful in comparing trees of different sizes with respect to their proportional frequencies of sub-trees. This is straightforward. We based the metric d_2 on vectors whose i^{th} components were the number of sub-trees of label i; we can divide these vectors by the number of tips in the tree: $\hat{v}_a = \frac{1}{n_a} v_a$ and define a new metric

$$\hat{d}_2(T_a, T_b) = ||\hat{v}_a - \hat{v}_b|| + \epsilon |n_a - n_b|$$

where $\epsilon > 0$. With small ϵ, $\hat{d}$ will be small when the proportional frequencies of sub-tree are very similar, but will only be 0 if the trees have identical vectors and the same number of tips.

Furthermore, if there are particular labels i that are of interest - for example those with relatively few tips, for a "tip-centric" tree comparison, weights w can be chosen and applied to the vectors to emphasize some entries more than others :

$$d_w(T_a, T_b) = ||w \cdot v_a - w \cdot v_b||.$$

The same weighting can of course be applied to $\hat{v}$ in $\hat{d}_2$.

The labelling schemes induce natural metrics on tree topologies, which we have applied to random tree-generating processes known to give rise to different shapes, and to data from human influenza A. The metric's use of a bijective mapping to $\mathbb{N}^+$ means that it extends to a convex metric in $\mathbb{Q}^+$. However, the nature of the scheme means that it does not capture the lengths of branches. These are biologically relevant in many examples, because they reflect the (inferred) amount of time or genetic distance between evolutionary events, although particularly for branches deep in the tree structure they may be difficult to infer accurately.

To date, we are unaware of a metric (in the sense of a true distance function) on unlabelled trees that captures branch lengths, but there are several non-metric approaches to comparing unlabelled trees. In particular, Poon's kernel method [39] compares subset trees that are shared by two input trees, after first "ladderizing" the trees (arranging internal nodes in a left-right order with branching events preferentially to one side). Using a kernel function, this approach can quantify similarity between trees. One challenge is that where branch length is included, differences in overall scaling or units of the branch lengths can overwhelm structural differences. Lengths can be re-scaled (for example such that the height of both trees becomes 1), but rescaling methods may be sensitive to outliers or to the height of the highest tip in the tree. Lengths could also be set to 1 to compare topologies only. Recently, Lewitus and Morlon (LM) [40] used the spectrum of a matrix of all the node-node distances in the tree to characterise trees; this is naturally invariant to any node and tip labels. They used the Kullback-Leibler divergence between smoothed spectra as a measure of distance. If the spectrum uniquely defined a tree this would be a metric, as it is non-negative and obeys the triangle inequality. As it uses all node-node distances, this approach, requiring the spectrum of a non-sparse $2n-1 \times 2n-1$ matrix for a tree of n tips, will become infeasible for large trees. Finally, it is always possible to compare summary features of trees, including the number of lineages through time, diversity measures, density of tip-tip distances, imbalance measures and other features of the topology.

These approaches can be combined with our metric to create novel metrics on unlabelled trees; as our metric satisfies $d(T_1, T_2) = 0 \iff T_1 = T_2$, any distance function of the form

$$\hat{d}(T_1, T_2) = w_1 d_i(T_1, T_2) + w_2 C(T_1, T_2)$$

where $C(T_1, T_2)$ is the LM tree difference, a kernel-based tree difference (not similarity), a distance between vectors of summary features, or a weighted sum of these, and w_i are positive, will be a metric.

In this way we can extend the metric to incorporate branch lengths and to emphasize features of interest (ie those believed to be informative of an underlying process of interest), while retaining the advantages of a true distance metric.

5.3 Implementation

We have used R throughout and are developing an R package. Code is available on github at `https://github.com/c`
The implementation assumes full binary trees and includes metrics d_1 and d_2 with the option of weighting.

References

[1] Chewapreecha C et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat Genet*, 46(3):305–309, March 2014.

[2] Bedford T et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*, 523(7559):217–220, 9 July 2015.

[3] Ag1000G: Anopheles gambiae 1000 genomes — www.malariagen.net. https://www.malariagen.net/projects/vector/ag1000g. Accessed: 2016-3-23.

[4] Stam E. Does imbalance in phylogenies reflect only bias? *Evolution*, 56(6):1292–1295, June 2002.

[5] Slowinski JB. Probabilities of n-trees under two models: A demonstration that asymmetrical interior nodes are not improbable. *Syst Zool*, 39(1):89–94, 1990.

[6] Guyer C and Slowinski JB. Adaptive radiation and the topology of large phylogenies. *Evolution*, 47(1):253–263, 1993.

[7] Rodríguez J Harvey PH Grenyer R Purvis A, Fritz SA. The shape of mammalian phylogeny: patterns, processes and scales. *Philos T Roy Soc B*, 366(1577):2462–2477, 12 September 2011.

[8] Kirkpatrick M and Slatkin M. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171–1181, 1 August 1993.

[9] Mooers AO and Heard SB. Inferring evolutionary process from phylogenetic tree shape. *Q Rev Biol*, pages 31–54, 1997.

[10] Blum MGB and François O. Which random processes describe the tree of life? a large-scale study of phylogenetic tree imbalance. *Syst Biol*, 55(4):685–691, August 2006.

[11] Wu T and Choi KP. On joint subtree distributions under two evolutionary models. *Theor Popul Biol*, 108:13–23, 29 November 2015.

[12] Fusco G and Cronk QCB. A new method for evaluating the shape of large phylogenies. *J Theor Biol*, 175(2):235–243, 21 July 1995.

[13] Aldous DJ. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. *Stat Sci*, 16(1):23–34, 1 February 2001.

[14] Tria F Pompei S, Loreto V. Phylogenetic properties of RNA viruses. *PLOS One*, 7(9):e44849, 20 September 2012.

[15] Stich M and Manrubia SC. Topological properties of phylogenetic trees in evolutionary models. *Eur Phys J B*, 70(4):583–592, 21 July 2009.

[16] Agapow PM and Purvis A. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Syst Biol*, 51(6):866–872, December 2002.

[17] Matsen FA. A geometric approach to tree shape statistics. *Syst Biol*, 55(4):652–661, August 2006.

[18] Morlon H Manceau M, Lambert A. Phylogenies support out-of-equilibrium models of biodiversity. *Ecol Lett*, 18(4):347–356, 2015.

[19] Janson S Blum M, François O. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *Ann Appl Probab*, 2006.

[20] Fontanari JF Maia LP, Colato A. Effect of selection on the topology of genealogical trees. *J Theor Biol*, 226(3):315–320, 7 February 2004.

[21] Dayarian A and Shraiman BI. How to infer relative fitness from a sample of genomic sequences. *Genetics*, 197(3):913–923, July 2014.

[22] Wiuf C Hein J, Schierup M. *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA, 2004.

[23] Wakeley J and Wakeley J. *Coalescent theory: an introduction*. 2009.

[24] Gascuel O. Evidence for a relationship between algorithmic scheme and shape of inferred trees. In *Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pages 157–168. Springer Berlin Heidelberg, 2000.

[25] Bedford T Volz EM, Koelle K. Viral phylodynamics. *PLOS Comp Biol*, 9(3):e1002947, 21 March 2013.

[26] Lambert A and Stadler T. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theor Popul Biol*, 90(0):113–128, December 2013.

[27] Boyd M Colijn C Plazzotta G, Kwan C. Effects of memory on the shapes of simple outbreak trees. *Sci Rep*, 6:21159, 18 February 2016.

[28] Robinson K et al. How the dynamics and structure of sexual contact networks shape pathogen phylogenies *PLOS Comp. Biol.*, 9(6): 2013.

[29] Leventhal GE, Kouyos R, Stadler T, Von Wyl V. Inferring epidemic contact structure from phylogenetic trees. PLoS Comp. Biol., 8(3): e1002413, 2012. e1003105, 2012.

[30] Colijn C and Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. *Evol Med Public Health*, 2014(1):96–108, 9 June 2014.

[31] Plazzotta G and Colijn C. Asymptotic frequency of shapes in supercritical branching trees. 9 July 2015.

[32] Robinson DF and Foulds LR. Comparison of phylogenetic trees. *Math Biosci*, 53(1–2):131–147, February 1981.

[33] Vogtmann K Billera LJ, Holmes SP. Geometry of the space of phylogenetic trees. *Adv Appl Math*, 27(4):733–767, November 2001.

[34] Guyer C and Slowinski JB. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution*, 45(2):340–350, 1991.

[35] Colless DH. Relative symmetry of cladograms and phenograms: an experimental study. *Syst Biol*, 1995.

[36] Sackin MJ. "good" and "bad" phenograms. *Syst Zool*, 21(2):225–226, 1972.

[37] McKenzie A and Steel M. Distributions of cherries for two models of trees. *Math Biosci*, 164(1):81–92, March 2000.

[38] Rosenberg N. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-Generated genealogical trees. *Ann Comb*, 10(1):129–146, June 2006.

[39] Poon A et al. Mapping the shapes of phylogenetic trees from human and zoonotic RNA viruses. *PLOS One*, 8(11):e78122, 1 November 2013.

[40] Lewitus E and Morlon H. Characterizing and comparing phylogenies from their laplacian spectrum. *Syst Biol*, 12 December 2015.

[41] Lueker GS and Booth KS. A linear time algorithm for deciding interval graph isomorphism. *J ACM*, 26(2):183–195, April 1979.

[42] Hopcroft JE and Tarjan RE. Isomorphism of planar graphs. In *Complexity of computer computations*, pages 131–152. Springer, 1972.

[43] Issel W. Aho, AV/Hopcroft, JE/Ullman, JD, the design and analysis of computer algorithms. London-Amsterdam-Don Mills-Sydney. Addison-Wesley publ. comp. 1974 x, 470 s., 24,–. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 59(2):141–141, 1979.

[44] Colbourn CJ and Booth KS. Linear time automorphism algorithms for trees, interval graphs, and planar graphs. *SIAM J Comput*, 10(1):203–225, 1981.

[45] Sayward C. The tree theory and isomorphism. *Analysis*, 41(1):6–11, 1 January 1981.

[46] Russell CA et al. The global circulation of seasonal influenza a (H3N2) viruses. *Science*, 320(5874):340–346, 18 April 2008.

[47] Kamradt M Kepler TB Koelle K, Khatri P. A two-tiered model for simulating the ecological and evolutionary dynamics of rapidly evolving viruses, with an application to influenza. *J R Soc Interface*, 7(50):1257–1274, 6 September 2010.

[48] Westgeest KB et al. Genetic evolution of the neuraminidase of influenza a (H3N2) viruses from 1968 to 2009 and its correspondence to haemagglutinin evolution. *J Gen Virol*, 93(Pt 9):1996–2007, September 2012.

[49] Luksza M and Lässig M. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, 6 March 2014.

[50] Huelsenbeck JP and Kirkpatrick M. Do phylogenetic methods produce trees with biased shapes? *Evolution*, 50(4):1418–1424, 1996.

[51] Matsen FA. Optimization over a class of tree shape statistics. *IEEE/ACM Trans Comput Biol Bioinform*, 4(3):506–512, July 2007.

[52] Piel WH Eriksson T Sanderson MJ, Donoghue MJ. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am J Bot*, 81(6):183, 1994.

[53] Furnas GW. The generation of random, binary unordered trees. *J Classif*, 1(1):187–233, 1984.

[54] Stadler T et al. Estimating the basic reproductive number from viral sequence data. *Mol Biol Evol*, 29(1):347–357, January 2012.

[55] Rasmussen DA du Plessis L Stadler T, Kühnert D. Insights into the early epidemic spread of ebola in sierra leone provided by viral sequence data. *PLOS Curr*, 6, January 2014.

[56] Stadler T. TreeSim in R-Simulating trees under the birth-death model. *R package*, 1, 2010.

[57] Rosen DE. Vicariant patterns and historical explanation in biogeography. *Syst Biol*, 27(2):159–188, 1 June 1978.

[58] Aldous D. Probability distributions on cladograms. In *Random Discrete Structures*, The IMA Volumes in Mathematics and its Applications, pages 1–18. Springer New York, 1996.

[59] Blum M François O Bortolussi N, Durand E. aptreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*, 22(3):363–364, 1 February 2006.