Polynomial Phylogenetic Analysis of Tree Shapes

Pengyu Liu*, Matthew Gould, and Caroline Colijn,
Department of Mathematics, Simon Fraser University, Burnaby, V5A 1S6, Canada

**E-mail: pengyu_liu@sfu.ca*

Abstract

Phylogenetic trees are a central tool in evolutionary biology. They demonstrate evolutionary patterns among species, genes, and with modern sequencing technologies, patterns of ancestry among sets of individuals. Phylogenetic trees usually consist of tree shapes, branch lengths and partial labels. Comparing tree shapes is a main challenge in comparing phylogenetic trees as there are few tools to describe tree shapes in a quantitative, accurate, comprehensive and easy-to-interpret way. Current methods to compare tree shapes are often based on scalar indices reflecting tree imbalance, and on frequencies of small subtrees. Polynomials are important tools to describe discrete structures and have been used to study various objects including graphs and knots. There exist polynomials that describe rooted trees. In this paper, we present methods based on a polynomial that fully characterizes trees. These methods include tree metrics and machine learning tools. We use these methods to compare tree shapes randomly generated by simulations and tree shapes reconstructed from data. Moreover, we also show that the methods can be used to estimate parameters for tree shapes and select the best-fit model that generates the tree shapes.

Key words: Phylogenetics, Polynomials, Tree Shapes, Tree Metrics

A tree is a natural data structure that represents hierarchical relations between objects. In phylogenetics, a tree structure usually includes its tree shape, that is, the unlabeled underlying graph, as well as branch lengths reflecting either evolutionary distance or time. Estimating the branch lengths can be a challenge for tree reconstruction methods, with Bayesian and maximum likelihood methods yielding inconsistent results (Brown, 2010), high demands on memory and processor time (Binet, 2016), and/or lack of strong support for a molecular clock (in the case of timed trees). As a consequence, the inferred phylogenetic trees may have a consistent tree shape but differing root heights. In this paper, we mainly focus on tree shapes, which are of both evolutionary and mathematical interests.

The shapes of phylogenetic trees can carry information about macroevolutionary processes, as well as reflecting the data used and the choice of the evolutionary model (Kirkpatrick, 1993; Purvis, 2011; Aldous, 1996). The ecological fitness and the presence of selection can also affect the shapes of trees (Dayarian, 2014; Maia, 2004). In the study of infectious diseases, where the shapes of phylogenetic trees of pathogens reveal diversity patterns that represent a combination of unfixed neutral variation, variation under selection, demographic processes and ecological interactions, it is not clear how informative the tree shapes are of the underyling evolutionary and epidemiological processes. However, effort is being made to explore this question, with the main focus often on the frequency of cherries and tree imbalance (Grenfell, 2004; Lambert, 2013; Plazzotta, 2016; Volz, 2013).

One of the main topics of inquiry in phylogenetic tree shapes has been asymmetry, since a key observation was made that the shapes of phylogenetic trees reconstructed from data are more asymmetric than tree shapes simulated by simple models (Aldous, 1996). Various ways to measure the asymmetry were developed (Aldous, 1996; Colless, 1982; Fusco, 1995; Sackin, 1972; Stich, 2009) and it was shown that these asymmetric measures can distinguish random trees generated by different models (Agapow, 2002; Kirkpatrick, 1993; Matsen, 2006). At the same time, mathematical models that produce imbalanced

trees were developed (Aldous, 2001; Blum, 2006). As statistical tools, the distributions of tree shapes under simple models can be used to test evolutionary hypotheses (Blum, 2006; Mooers, 1997; Wu, 2016). In (Manceau, 2015), and mathematical models can be developed to match the macroevolutionary patterns observed in the phylogenetic trees reconstructed from data.

As the cost of DNA sequencing is decreasing, more genomic data are being collected and becoming available. More organisms are being sequenced progressively at the whole-genome scale (Bedford, 2015; Chewapreecha, 2014; Colijn, 2018) and the evolution of certain pathogens is being tracked in real time (Hadfield, 2018). As a consequence, both the number and the size of trees reconstructed from data are increasing. Accordingly, a major challenge in tree shape analysis is that there are few tools to describe and compare trees in a quantitative, accurate, comprehensive and easy-to-interpret way, especially for large trees. Scalar indices describing asymmetry or the frequency of subtrees have a limitation in that many different tree shapes may have the same index. A labelled tree is a tree shape whose vertices have unique labels. An alternative approach to comparing tree shapes is using metrics defined for labelled trees, for example, the well known Robinson-Foulds metric (Robinson, 1981), Billera-Holmes-Vogtmann metric (Billera, 2001) and Kendall-Colijn metric (Kendall, 2016), among others. These metrics depend on the labels of the vertices, that is, two labelled trees with the same tree shape but the labels re-arranged are not identical and the distances between them can be very large. Recently, metrics defined for rooted unlabelled trees or rooted tree shapes have also been introduced (Colijn, 2018), making use of integer labels assigned to tree shapes. However, these metrics have several limitations, including the challenge of interpreting the integer labels, the treatment of non-binary trees, and the metrics' performance in distinguishing trees from different processes or datasets.

Polynomials are important tools in the mathematics study of discrete structures, and can be used to describe discrete structures in interpretable ways. For example, the

Tutte polynomial (Tutte, 1954) is a renowned polynomial for graphs and the Jones polynomial (Jones, 1985) is one of the most important tools to study knots. In (Liu, 2019), a method to assign a unique polynomial to each tree shape is introduced. These polynomials provide a new way to describe tree shapes quantitatively and comprehensively. The coefficients of the polynomial of a tree can be considered as a generalization of the clade size distribution of the tree. In addition, the set of coefficients of a tree polynomial can be treated as a vector, and vectors are natural objects on which to define metrics. Here, we define and examine a metric and a binary similarity measure on rooted tree shapes. We show that the polynomial metric and the binary similarity measure can separate trees that are known to have different shapes. We also show that the coefficients of the polynomials of trees can be used to estimate parameters of model to match a fixed tree, and can identify the model that generates a tree most similar to a fixed tree.

Materials and Methods

Tree Polynomials, Distances and Binary Differences

A tree or a tree shape represents an unlabeled tree, that is a graph with no cycles, without information about branch lengths or tip labels unless otherwise stated. We define a polynomial $P(T)$ for each rooted unlabeled tree T in the following way. If T is the trivial tree with a single vertex, then $P(T) = x$. Otherwise T has k branches at its root and each branch leads to a subtree of T. So T has k rooted unlabeled subtrees T_1, T_2, ... , T_k whose roots are internal nodes of T and are adjacent to the root of T. We define the polynomial for T by $P(T) = y + \prod_{i=1}^{k} P(T_i)$. If all the subtrees are the trivial tree, then we have defined the polynomial. If T_i is not trivial, then we apply the same definition to compute $P(T_i)$. We apply the definition recursively until we reach all tips of T. It is proved in (Liu, 2019) that the polynomial distinguishes unlabeled rooted trees and can be generalized to distinguish unlabeled unrooted trees. A rooted tree can be reconstructed from its

polynomial by computing its Newick code, which can be obtained by recursively subtracting y and factoring the rest of the polynomial. Methods to factor large multivariate polynomials can be found in (Monagan, 2018).

The coefficients of a tree polynomial can be written as a matrix or vector of integers. Let T be a rooted tree with n tips. Its coefficient matrix $C(T)$ or $(c^{(\alpha,\beta)})$ is displayed as follows, where $c^{(\alpha,\beta)}$ is the coefficient in the term $c^{(\alpha,\beta)}x^\alpha y^\beta$.

$$C(T) = \begin{array}{c} \\ 1 \\ y \\ y^2 \\ \vdots \\ y^n \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & x & x^2 & \dots & x^n \end{array} \\ \begin{pmatrix} c^{(0,0)} & c^{(1,0)} & c^{(2,0)} & \dots & c^{(n,0)} \\ c^{(0,1)} & c^{(1,1)} & c^{(2,1)} & \dots & c^{(n,1)} \\ c^{(0,2)} & c^{(1,2)} & c^{(2,2)} & \dots & c^{(n,2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c^{(0,n)} & c^{(1,n)} & c^{(2,n)} & \dots & c^{(n,n)} \end{pmatrix} \end{array} = (c^{(\alpha,\beta)})$$

These coefficients are interpretable. The coefficient $c^{(\alpha,\beta)}$ in $C(T)$ indicates the number of ways to choose β clades that together contain $n-\alpha$ tips in total, in a tree with n tips. In particular, the second row in the matrix $C(T)$ represents the clade size distribution such that $c^{(n-k,1)}$ indicates the number of clades with k tips (Liu, 2019). We define the polynomial metric by comparing the corresponding coefficients in the polynomials $P(T_1)$ and $P(T_2)$, where T_1 and T_2 are two arbitrary trees. Let $C(T_1) = (c_1^{(\alpha,\beta)})$ and $C(T_2) = (c_2^{(\alpha,\beta)})$ be the coefficient matrices of T_1 and T_2; we define the metric as follows.

$$d(T_1, T_2) = \sum_{0 \leqslant i,j \leqslant n} \log\left(\left|c_1^{(i,j)} - c_2^{(i,j)}\right| + 1\right)$$

This is a metric since non-negativity, identity, symmetry are trivial to check and the subadditivity follows the triangular inequality of the absolute value and the monotonicity and the concavity of the logarithm function. If T_1 and T_2 are of different sizes, for example say T_1 has n tips and T_2 has m tips where $m < n$, then we align the coefficient matrices so that the corresponding clade sizes are compared, that is, we align $c_2^{(m,0)}$ with $c_1^{(n,0)}$ and make $C(T_2)$ an $n \times n$ matrix by filling the remaining entries with zeros.

Counting the number of terms that are present in $P(T_1)$ but are absent in $P(T_2)$, or

the number of terms that are present in $P(T_2)$ but are absent in $P(T_1)$, provides another way to compare polynomials $P(T_1)$ and $P(T_2)$ for arbitrary trees T_1 and T_2. These are not strictly metrics, because two trees could be different but have the same presence-absence pattern in their coefficients. However, binary similarities are commonly used in various sciences for classification and clustering (Choi, 2010).

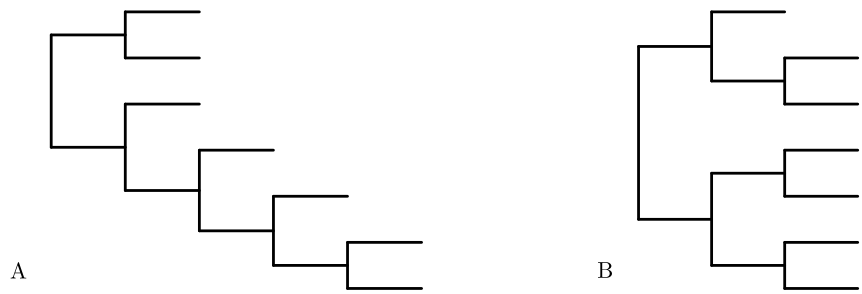


Figure 1. The polynomial for tree A is $P(A) = x^7 + 2x^5y + x^4y + x^3y^2 + x^3y + x^2y^2 + x^2y + xy^2 + y^2 + y$. The polynomial for tree B is $P(B) = x^7 + 3x^5y + x^4y + 3x^3y^2 + x^3y + 2x^2y^2 + xy^3 + xy^2 + y^3 + y^2 + y$. The polynomial distance between the trees is $d(A, B) = \sum_{0\leqslant i,j\leqslant 7} \log(|c_1^{(i,j)} - c_2^{(i,j)}| + 1) = 5\log(1 + 1) + \log(2 + 1) = 4.5643$. The term that are present in $P(A)$ but are absent in $P(B)$ is x^2y and the terms that are absent in $P(A)$ but are present in $P(B)$ are xy^3 and y^3. So the the polynomial binary differences are 1 and 2 respectively.

Simulations

The random tree shapes compared in this paper are generated using the R package *apTreeshape*. Four different models are used to generate the random trees. The Yule model or the Markov model is a non-parametric model such that each extant lineage has the same probability to branch into two new lineages. The proportional to distinguishable arrangements or the PDA model generates trees by uniformly randomly choosing a tree from the set of all tip-distinguished trees with a fixed number of tips. As there are more tip-distinguished trees that are based on imbalanced tree shapes, the PDA model is more likely to generate imbalanced trees. The Aldous model or Aldous' branching model is defined with a specific symmetric split distribution using harmonic numbers (Aldous, 1996; Blum, 2006). The biased speciation model is a parametric model with a probability

parameter p such when a lineage with branching rate r splits, one of the descendant lineages has a speciation rate of pr and the other has a speciation rate of $(1-p)r$. Because this is a simple one-parameter model, and because it generates imbalanced trees, we use this model to demonstrate parameter estimation.

To simplify the statements, a Yule (PDA, Aldous or biased) tree stands for a random tree generated by the Yule (PDA, Aldous or biased) model in the rest of the paper.

Parameter and Model Estimation

Parameter estimation We use two methods to estimate the parameter p of the biased speciation model. One of the methods is based on the polynomial metric and the k-nearest neighbor algorithm (with $k = 3$). To estimate the parameter p for a given tree T^*, we generate 1,000 biased trees of the same size as T^* with the parameter p uniformly randomly chosen from the interval $(0, 0.5]$. Then we compute the polynomial distances between T^* and the set of biased trees. Let T_1, T_2, T_3 be the 3 nearest trees to T^*, $d_1 = d(T^*, T_1)$, $d_2 = d(T^*, T_2)$, $d_3 = d(T^*, T_3)$ and let p_1, p_2, p_3 be the parameters that were used to generate T_1, T_2, T_3 respectively. We estimate p^*, the parameter for T^*, with the following formula:

$$p^* = \frac{(1/d_1)p_1 + (1/d_2)p_2 + (1/d_3)p_3}{1/d_1 + 1/d_2 + 1/d_3}$$

In the case where some $d_i = 0$, $p^* = (\sum_{d_i=0} p_i)/(\sum_{d_i=0} 1)$.

The Sackin index is a scalar that measures the imbalance of a tree. Let $s(T)$ be the Sackin index and $\lambda(T)$ be the number of cherries of an arbitrary tree T. The imbalance and the number of cherries are two of the main statistics that have been used previously to capture tree shapes and to fit evolutionary and epidemiological models (Frost, 2013). We associate a vector, the Sackin-cherry vector, $v(T) = (s(T), \lambda(T))$ with the tree T. To determine whether polynomial-based tree comparisons offer improved estimates compared to these scalar tree statistics, we compute the Euclidean distances between these vectors and use the same formula as above to estimate the parameter p (except that the distances

d_i are replaced with the Euclidean distances between Sackin-cherry vectors).

The second method is based on linear regression. To estimate the parameter p of a given tree T^*, we generate 100,000 biased trees of the same size as T^* with the parameter p uniformly randomly chosen from the interval $(0, 0.5]$. We consider the polynomial coefficients of a tree as variables and use the polynomial coefficients of biased trees to fit a linear model, where the parameter p of a tree is the scalar response. We apply the predictor function to the polynomial $P(T^*)$ to estimate the parameter p^* of T^*. This method is more computationally expensive than the first one as there exist $(n+1)^2$ variables for a tree with n tips.

Similarly, we also use the Sackin-cherry vectors to fit a linear model, where we model the predicted parameter p as a linear combination of the Sackin index and the number of cherries of a tree. We use the predictor function and the Sackin-cherry vector $v(T^*)$ to estimate p^*.

Model estimation We use classification to select the model that is the best fit for generating a given tree. To reduce the number of variables, we substitute the variable y in the polynomial with the complex number $1 + i$. It is proved in (Liu, 2019) that if we substitute any prime number for the variable y in the polynomial, the resulting polynomial also distinguishes rooted binary trees. We use the phrase "complex polynomial" for the polynomial with $i + 1$ in place of y. To estimate the model that is the best fit for generating a fixed tree T^*, We generate 200 random trees of the same size as T^* using the Yule model, the PDA model, the Aldous model and the biased speciation model with a chosen parameter p (for a total of 800 trees). We compute the complex polynomials of these trees and use their coefficients to train a naive Bayes classifier. We then use this classifier to estimate the model that generates T^*.

Data

HIV and influenza virus trees The HIV trees were described and analyzed previously (Chindelevitch, 2019). Briefly, HIV-1 sequence data from three studies were used. The Wolf et al. study (Wolf, 2017) provided data from a concentrated epidemic of HIV-1 subtype B, occurring primarily in men who have sex with men (MSM) in Seattle, USA. The Novitsky et al. study (Novitsky, 2013) describes data from a generalized epidemic of HIV-1 subtype C in Mochudi, Botswana, a village in which the HIV-1 prevalence in the adult population at the time was estimated to be approximately 20%. Hunt et al. (Hunt, 2013) describes data from a national survey of the generalized epidemic of HIV-1 subtype C in South Africa. These datasets reflect a diverse set of spatial scales and epidemiological contexts. Phylogenetic reconstruction was described in (Chindelevitch, 2019); briefly, trees were reconstructed using RAxML (Stamatakis, 2014), which is a maximum likelihood method, under a general time-reversible (GTR) model of nucleotide substitution. We use a GTRCAT model for rate variation among sites. Each tree was based on a random sample of 100 sequences. We use a subtype D sequence as an outgroup to root HIV-1 subtype B phylogenies.

Our influenza virus trees were previously described in (Colijn, 2018). We aligned HA protein sequences from NCBI, focusing on human influenza A (H3N2). Data were downloaded from NCBI on 22 Jan. 2016. We included full-length HA sequences with collection date. The USA dataset ($n = 2168$) includes sequences from the USA with collection dates between Mar. 2010 and Sep. 2015. The tropical dataset ($n = 1388$) includes sequences with a location listed as tropical, with collection dates within Jan. 2000 and Oct. 2015. Accession numbers are included in the Supporting Information of Colijn (2018). Fasta files were aligned with mafft, and for both the tropical and USA datasets, 500 taxa were selected uniformly at random 200 times. We inferred 200 corresponding phylogenetic trees with FastTree (Price, 2010). Where necessary we re-aligned the 500 selected sequences before performing tree inference. This process resulted in 200 “tropical”

influenza virus trees and 200 "USA" influenza virus trees, each with 500 tips, reconstructed from the HA region of human H3N2 samples. Note that this approach is distinct from the perhaps more familiar phylogenetic methods where bootstrapping or Bayesian reconstructions results in many trees on *one* set of tips. These are likely to share features and structures because they describe the ancestry of the same set of taxa. Here, each tree has a different set of tips (though there is some overlap).

WHO influenza virus clades We used several influenza virus clades, described in (Hayati, 2019). In that work we downloaded all human H3N2 full-length HA sequences with dates between 1980 and May 2018 and created a large, timed phylogeny of H3N2 using RAxML and Least Squares Dating (Stamatakis, 2014; To, 2016). This "full" tree has over 12,000 tips. We used the Nextflu (Neher, 2015) *augur* pipeline (`https://bedford.io/projects/nextflu/augur/`) to assign a WHO clade designation to the sequences. The WHO defines named clades using specific mutations in the HA1 and HA2 subunits of the HA protein. The full list of mutations is available at: `https://github.com/nextstrain/seasonal-flu/blob/master/config/clades_h3n2_ha.tsv`. We assign a sequence to a clade if it contains all the mutations defining that clade. We then extracted the subtrees of the "full" tree corresponding to specific WHO clades A1B/135N (60 tips), A1B/135K (63 tips), 3c3.B (117 tips) and A3 (227 tips). These are recent and appropriately-sized trees which we use here to demonstrate parameter estimation for simple models, and model selection among our four random tree models.

Implementation

We developed an R package named *treenomial*, which is available at CRAN. We also prepared a demonstration named *treeverse*, which displays a 3-dimensional projection of the polynomial metric space of all binary tree shapes up to 16 tips with interactive options available at `https://magpiegroup.shinyapps.io/treeverse/`.

Results

Polynomial Tree Comparison

Simulated random trees Random trees are generated using models that are known to produce different tree shapes, that is, the Yule model, the PDA model, the Aldous model and the biased speciation model with $p = 0.3$. See Methods. For each model, 100 random trees with 500 tips are generated. To determine how the polynomial tree comparisons distinguish the models' random trees, a visualization of the distances by classical multidimensional scaling or MDS is shown in Figure 2 A. The separated clusters suggest that the polynomial distance captures the distinctive tree shapes generated by these four models. These clusters are tighter, more distinct, and less noisy than the corresponding clusters for the same set of random trees generated by the four models using the earlier integer labelling metric in (Colijn, 2018). The horseshoe shape in the MDS plot suggests that there may be an underlying latent parameter for these random trees (Diaconis, 2008). The Sackin index and the number of cherries are the most studied scalars for tree shapes. To determine if they match the latent parameter, we compute the Sackin indices and the numbers of cherries for the random trees. Our results suggest that neither corresponds to the underlying latent parameter. See Supplementary Figure 2.

Data-derived trees We also compute the polynomial distances between trees inferred from data. The first data set consists of trees inferred from sequences of the HA gene in human influenza virus A H3N2. Influenza virus A is highly seasonal outside the tropics and most cases occur in the winter (Russell, 2008), whereas there is relatively little seasonal variation in the tropics. This demonstrative data set to provides trees for the same virus circulating with different epidemiological dynamics (seasonal forcing in temperate regions, vs lack of seasonality in the tropics). The second data set consists of three samples of trees inferred from HIV-1 sequences in different settings: subtype B among MSM in Seattle (Wolf, 2017), a generalised HIV subtype epidemic in at the village scale Botswana

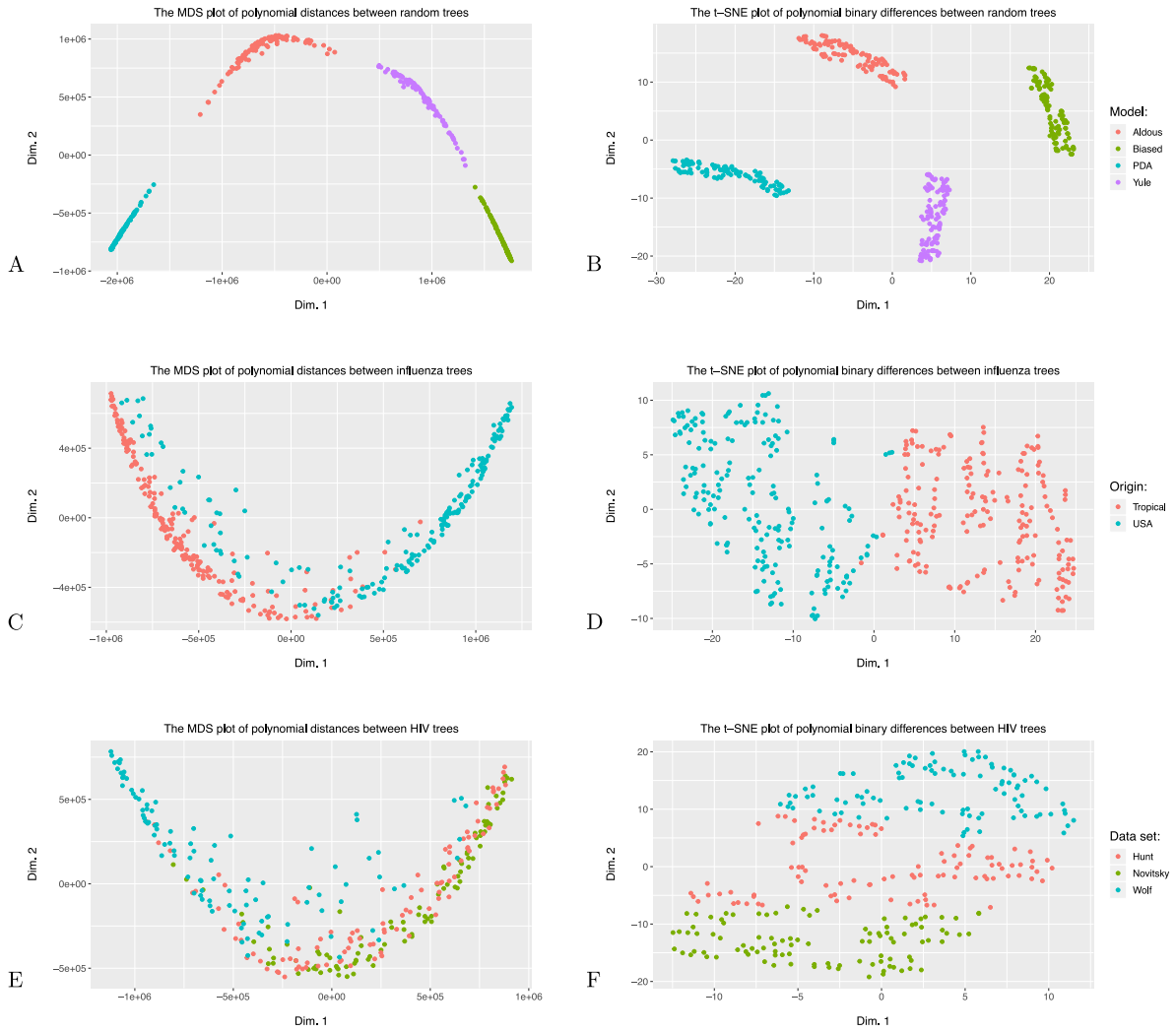


Figure 2. The MDS plots of the polynomial distances between (A) the random trees with 500 tips generated by the four models, (C) influenza trees and (E) HIV trees. The t-SNE plots of the polynomial binary differences between (B) random trees generated by the four models, (D) influenza trees and (F) HIV trees.

(Novitsky, 2013) and a national-level dataset from South Africa (Hunt, 2013). As with influenza virus, it is to be hoped that these different epidemiological patterns are revealed in the shapes of reconstructed phylogenetic trees (Chindelevitch, 2019; Colijn, 2018).

We visualize the polynomial distances between trees in these two sets by classical MDS in Figure 2 C,E. The MDS suggests that the polynomial metric may not cluster these trees into the desired groups. However, all the information of the trees' shapes is encoded in their polynomials. We compute the polynomial binary similarities between the these

trees. Binary similarities, based on presence and absence of components, are one of the commonly used indices in, for example, taxonomic, ecologic, biogeographic comparison and classification (Choi, 2010). They provide effective insights about clusters though they are not metrics in general. The polynomial binary similarity we choose is the number of terms that are present in the polynomial of one tree but are absent in the polynomial of the other. We visualize the polynomial binary similarities between these trees by the t-distributed stochastic neighbor embedding or t-SNE (van der Maaten, 2008). The results are displayed in Figure 2 B,D,F. The influenza trees and the HIV trees are very well separated into desired groups under the binary similarity measure with the t-SNE visualization; note that the group information (which trees were from tropical or USA flu, or which HIV trees were from which data source) is not supplied to the t-SNE. This indicates that classifying the epidemiological process behind a tree using the metric would likely be possible. For these particular challenges, however, typically a researcher would know whether their data were from the tropics or not, or what the broad epidemiological setting (village, national, Western population MSM) was at the time of collection. We therefore focus on more specific estimation questions (parameter estimation and model choice).

Parameter and Model Estimation

Parameter estimation The polynomial can be used to estimate the parameter p in the biased speciation model (see Methods).Figure 3 A displays the estimated values and the true values of the p parameters for the 1,000 simulated biased trees with 25 tips. We also conduct the same examination for the method on biased trees with 100 tips and 400 tips. The results are displayed in Figure 3 C,E. The method has better performance for trees of large sizes, and the results from the polynomial and the Sackin-cherry vector are similar.

The other method is linear regression; Figure 3 B displays the estimated and true values of the p parameters of the 1,000 trees. The results of the same examination of the

method for biased trees with 100 tips are displayed in Figure 3 D. We find that the polynomial performs better than the Sackin-cherry vector, and moreover, for biased trees with 100 tips, the linear regression method with polynomial coefficients produces better results than the polynomial metric. Linear regression also carries the advantage that we can identify the traits of trees that are most related to the biased speciation parameter p from the statistically significant coefficients in the model, since the coefficients of a polynomial are interpretable. In Supplementary Table 1, the most statistically significant terms in the polynomial are shown. For example, the coefficient of the term y^2 equals the number of ways to choose two clades that contains all the tips. For binary trees, this coefficient indicates if the tree has an adjacent tip to the root or not, which is an important factor of determining the imbalance of a tree.

We estimate the parameter p for the WHO influenza virus clades A1B/135N (60 tips), A1B/135K (63 tips), 3c3.B (117 tips) and A3 (227 tips). Due to the sizes of the clades, we use the method based on the metric (with both the polynomial and the Sackin-cherry vector) to estimate their parameters. For each clade, we estimate its parameter 1,000 times; the distributions of the estimates are displayed in Figure 3 F. For these clades, we find that the values estimated by the polynomials are smaller than those estimated values by the Sackin-cherry vectors. In general, we find that this difference between the two kinds of estimated values is common for trees generated by the PDA model and the Aldous model, though both polynomials and the Sackin-cherry vectors provide faithful estimates for trees generated by the biased speciation model (See Supplementary Figure 5). This suggests that the influenza clades are most similar to the trees generated by the PDA model or the Aldous model.

Model estimation Figure 2 A suggests that it is possible to estimate the model that is the best fit for generating a given tree using the polynomial. We use naive Bayes classifiers to estimate the model that is the best fit for a given tree. The results are displayed in Figure 4 A for 100 tips and Figure 4 B for 400 tips. There is very little

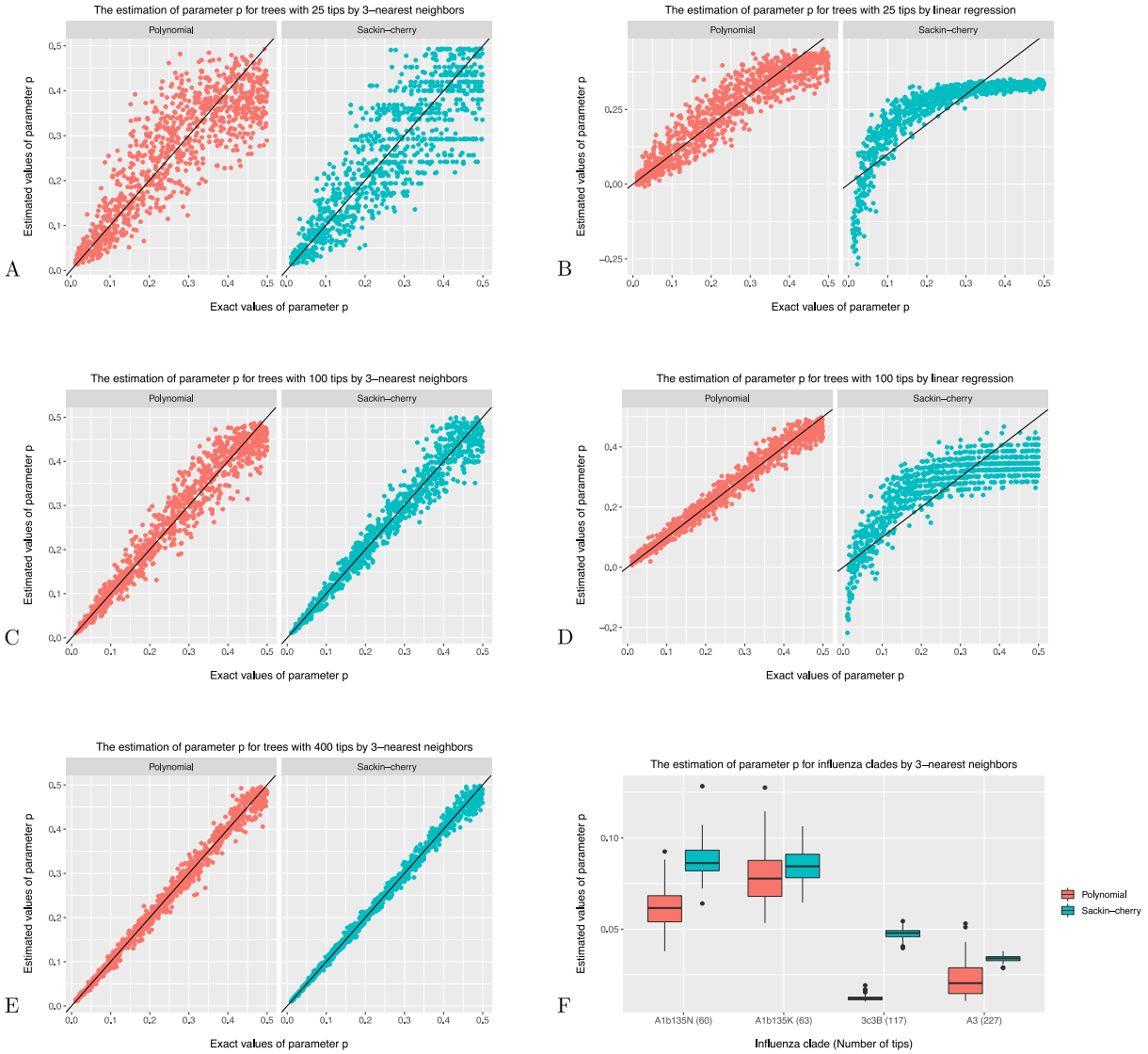


Figure 3. The plots of data points of true values and estimated values of parameter p by the 3-nearest neighbor method for trees generated by the biased speciation model with (A) 25 tips, (C) 100 tips and (E) 400 tips. The plots of data points of true values and estimated values of parameter p by the linear regression method for trees generated by the biased speciation model with (B) 25 tips and (D) 100 tips. (F) The distributions of the estimates of the parameter p for the chosen WHO influenza virus clades.

mis-classification between PDA and Yule trees, or between PDA and biased trees, but there is mis-classification among Aldous-Yule and biased-Yule. The performance is again better for larger trees, in the sense that there are fewer mis-classifications in Figure 4B than in A.

For each of the chosen WHO influenza virus clades, we train 1,000 naive Bayes classifiers, where the parameter p of the biased speciation model is set to be the clade's

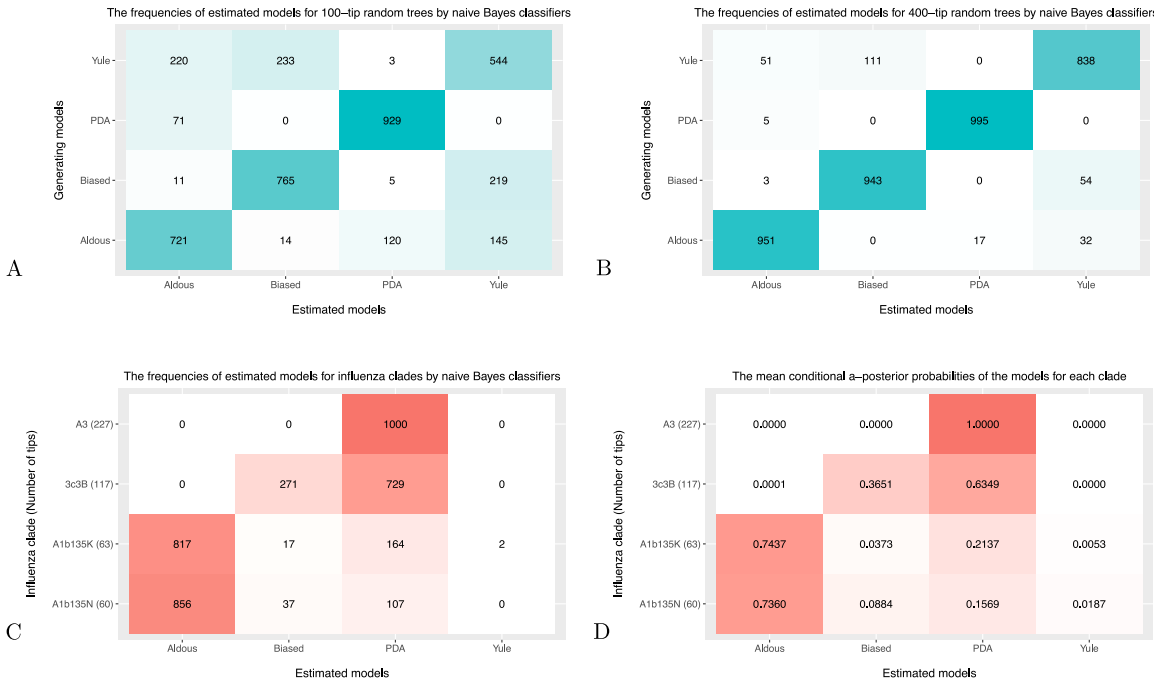


Figure 4. The frequencies of model estimation by naive Bayes classifiers for (A) random trees with 100 tips, (B) random trees with 400 tips and (C) the influenza clades. (D) The mean conditional *a posteriori* probabilities (over the 1,000 naive Bayes classifiers) of the model estimation for the influenza clades.

estimated value of p by the method based on the polynomial tree metric and the k-nearest neighbor algorithm, and we use the classifiers to estimate the model that is the best-fit for generating these clades. Figure 4 C,D display the results of, which show that the best-fit models for the clades are most likely to be the Aldous model or the PDA model among the four models that we use. This coincides with the results suggested by the differences between estimated values of parameter p by polynomials or Sackin-cherry vectors.

To further confirm the model estimates for the WHO influenza virus clades, we visualize the polynomial distances between the influenza virus clades and random trees generated by the PDA model, the Aldous model and the biased speciation model with the estimated values of parameter p for each of the clades. The results support naive Bayes classification results, in that the best-fit models for the clades are most likely to be the Aldous model or the PDA model (See Supplementary Figure 6, 7, 8 and 9). As an example, for the clade 3c3.B (117 tips), the 2-D MDS plot in Supplementary Figure 8 A

suggests that the clade is similar to the biased and PDA trees, and the 2-D t-SNE plot in Supplementary Figure 8 B shows that the clade is near the cluster of the PDA trees, which is also suggested by finding the nearest random tree to the clade. The nearest biased tree and the nearest tree (a PDA tree) to the clade are displayed in Supplementary Figure 8 C,E. It is observed that neither the nearest biased tree nor the nearest PDA tree resembles the clade. This is also observed in Figure 4 C,D: 271 out of the 1,000 estimates are the biased model and 729 estimates are the PDA model.

Discussion

We have developed polynomial-based tree comparison methods. Unlike other metrics and some comparisons on unlabelled trees, the polynomials are easy to compute, and the coefficients are interpretable. This opens up a wide range of dimension reduction and machine learning tools for application in phylogenetics. The methods discussed in this paper include a tree metric, a linear regression algorithm and classification with naive Bayes. These methods can distinguish trees from different models and different data sets, help estimate parameters, and aid in model selection. We have also applied the methods to estimate a parameter and select the best-fit model for the chosen WHO influenza virus clades. The results show that the tree shapes of the influenza clades are most similar to random trees generated by either the Aldous model or the PDA model among the models that we use. Moreover, the polynomials have the potential to be extended to record information about the branch lengths.

To compare trees with different sizes is another challenge in tree comparison. In this paper, we have compared trees with the same number of tips and we have proposed a way to compare trees with different sizes. We align the coefficient vectors so that the number of corresponding clade sizes are compared. In our demonstration *treeverse*, trees with different sizes are compared and the distances between the trees are visualized by an interactive 3-D MDS plot. There are other ways to align the coefficient vectors and

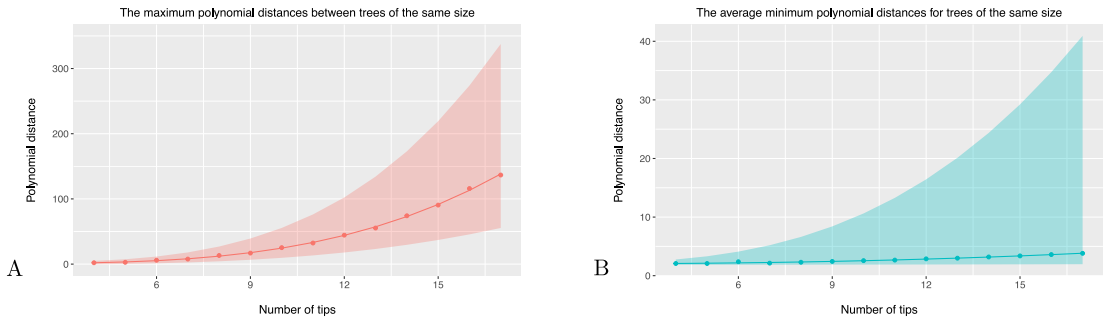
compare trees with different sizes, but for trees whose sizes are drastically different, the sizes of trees remain a dominating factor in comparing trees with the method.

Because polynomial coefficients can be treated as vectors, and vectors give rise to metrics, there are various alternative metrics that can be defined using tree polynomials, both those used here and others (Andrén, 2009; Chaudhary, 1991; Negami, 1996). Once trees are encoded as vectors, a range of regression, inference and dimension reduction tools are available that can be applied to trees. In addition, other tree shape statistics or other information about the trees can be easily appended to the vectors to integrate distinct sources of data. This provides a scheme to study phylogenetic trees comprehensively.

There remains considerable scope to improve the linear regression and classification tools used here, which we used to demonstrate that parameter estimation and model choice can be done. One challenge in this work is that there are too many polynomial coefficients; however, feature selection, hyperparameter optimization and dimension reduction tools could be used to reduce the number of features in a systematic way. Furthermore, we focused on a one-dimensional estimation task (estimating the bias parameter p). Realistic models of evolution are likely to contain multiple parameters (for example, time-dependent speciation and extinction rates; intra- and inter-group competition parameters, relative fitness), and the brute-force search we performed will not be possible in higher dimensions. However, given a smaller number of coefficients and other features that characterize a dataset, the tools of modern statistical inference are available for this task. The simpler estimation we have provided is a proof of principle for using polynomial coefficients in such estimation tasks.

Acknowledgements

This work was supported by the grant of the Federal Government of Canada's Canada 150 Research Chair program to Dr. Caroline Colijn. We would like to thank Art Poon, who provided the HIV trees.



Supplementary Figure 1. (A) The 95% confidence band of the fitted curve for maximum polynomial distances between trees with the same number of tips, (the data points fit the curve $y = 0.0112x^{3.324} + 0.9642$ with $R^2 = 0.9992$). (B) The 95% confidence band of the fitted curve for the average minimum distances between a tree and the nearest tree with the same number of tips, (the data points fit the curve $y = 0.0026x^{2.3046} + 2.0218$ with $R^2 = 0.9866$).

Supplementary Material

Interactive Figures

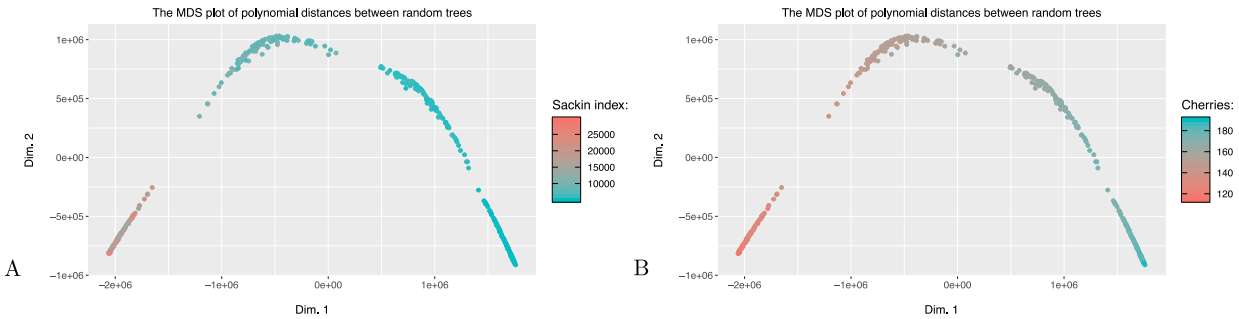
We made some interactive 3-D plots for the data sets discussed in the paper, which are available at: `https://magpiegroup.shinyapps.io/interactivefigures/`

Supplementary Results

Distance values in and of themselves are difficult to interpret without a sense of the overall scale or range. We compute how the distance scales with the number of tips n in the trees using trees with 4 to 17 tips. Supplementary Figure 1 A shows how the maximum polynomial distances between all rooted binary trees of a size n depends on n. For an arbitrary tree T with n tips, there exists a nearest tree T' in the polynomial metric space of trees with n tips. We define the polynomial distance between T and T' to be the minimum polynomial distance for T in the space. The average minimum polynomial distance for trees with n tips is the average of minimum polynomial distances for *all* trees with n tips. Supplementary Figure 1 B shows the average minimum polynomial distances between all rooted binary trees with the same number of tips.

The horseshoe shape in the MDS plot of the random trees generated by the four

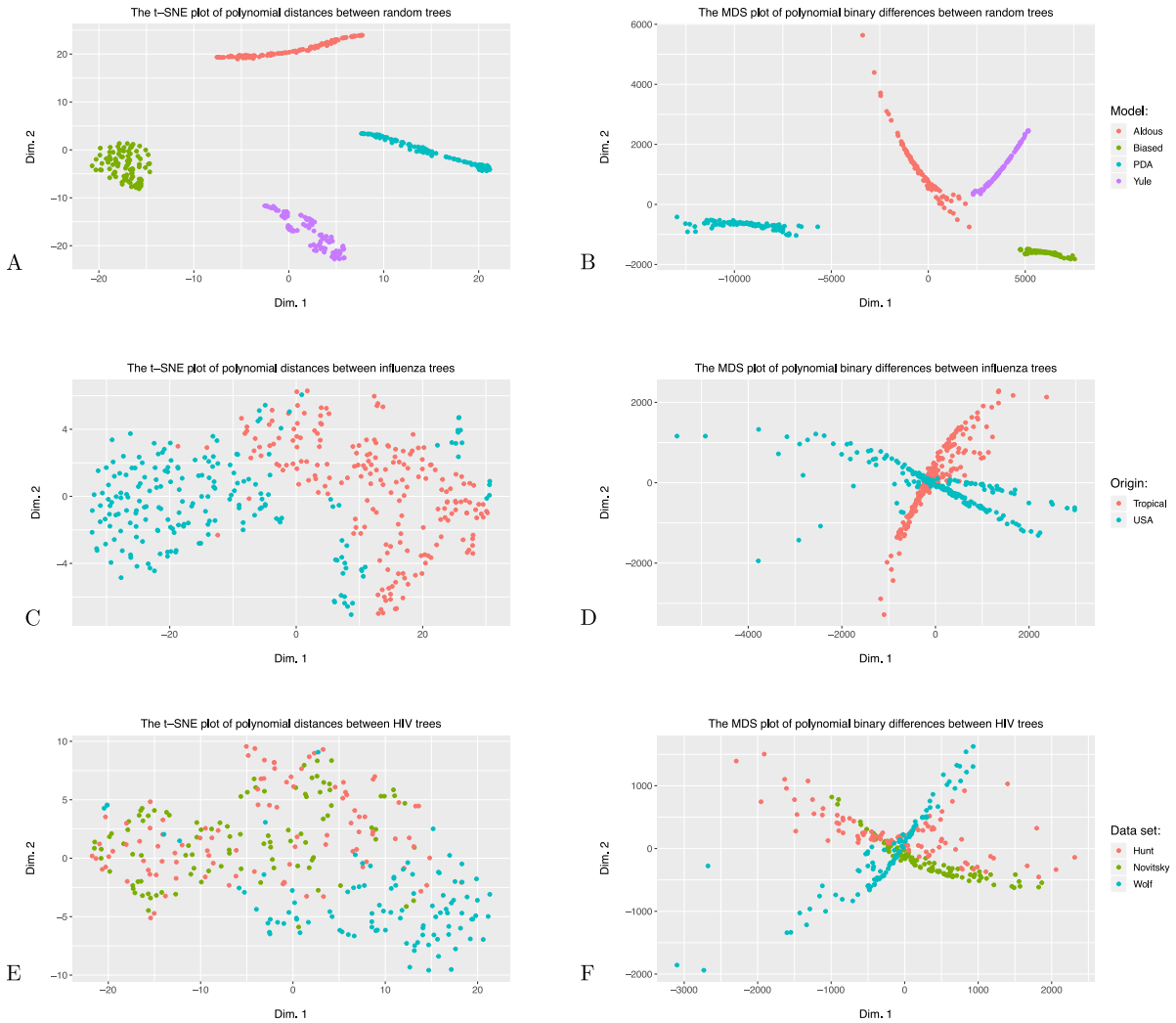
models suggests that there may be a latent parameter (Diaconis, 2008) ranging across the horseshoe. We compute the Sackin indices and the number of cherries of the random trees and display the results in Supplementary Figure 2. We find that neither the Sackin nor cherry frequency is likely to be this latent parameter, as neither range from large to small across the horseshoe; our results suggest that both are correlated with the (unknown) latent parameter.



Supplementary Figure 2. (A) The Sackin indices and (B) the number of cherries of random trees with 500 tips generated by the four models on the MDS plot of their polynomial distances as displayed in Figure 2 A.

We also display the t-distributed stochastic neighbor embedding plots of the polynomial distances and the classical multidimensional scaling plots of the polynomial binary differences between the random trees generated by the four models, the influenza trees and the HIV trees as complements to Figure 2 (Supplementary Figure 3). We compare the differences between the polynomial distances and the Euclidean distances in the visualizations by MDS and t-SNE. See the Shepard plots in Supplementary Figure 4 A,B, where we generate 25 trees with 100 tips using each of the four models. We also compare the polynomial metric with the metric d_2 defined by a labeling scheme defined in (Colijn, 2018) (Supplementary Figure 4 C).

When introducing using the naive Bayes classifiers to estimate the model that generates a given tree, we defined the complex polynomial, substituting the variable y by $1 + i$ in the polynomial. We can define a metric using the complex polynomial in the same fashion. More specifically, Let c_1^i and c_2^i be the coefficients of the term x^i in the complex

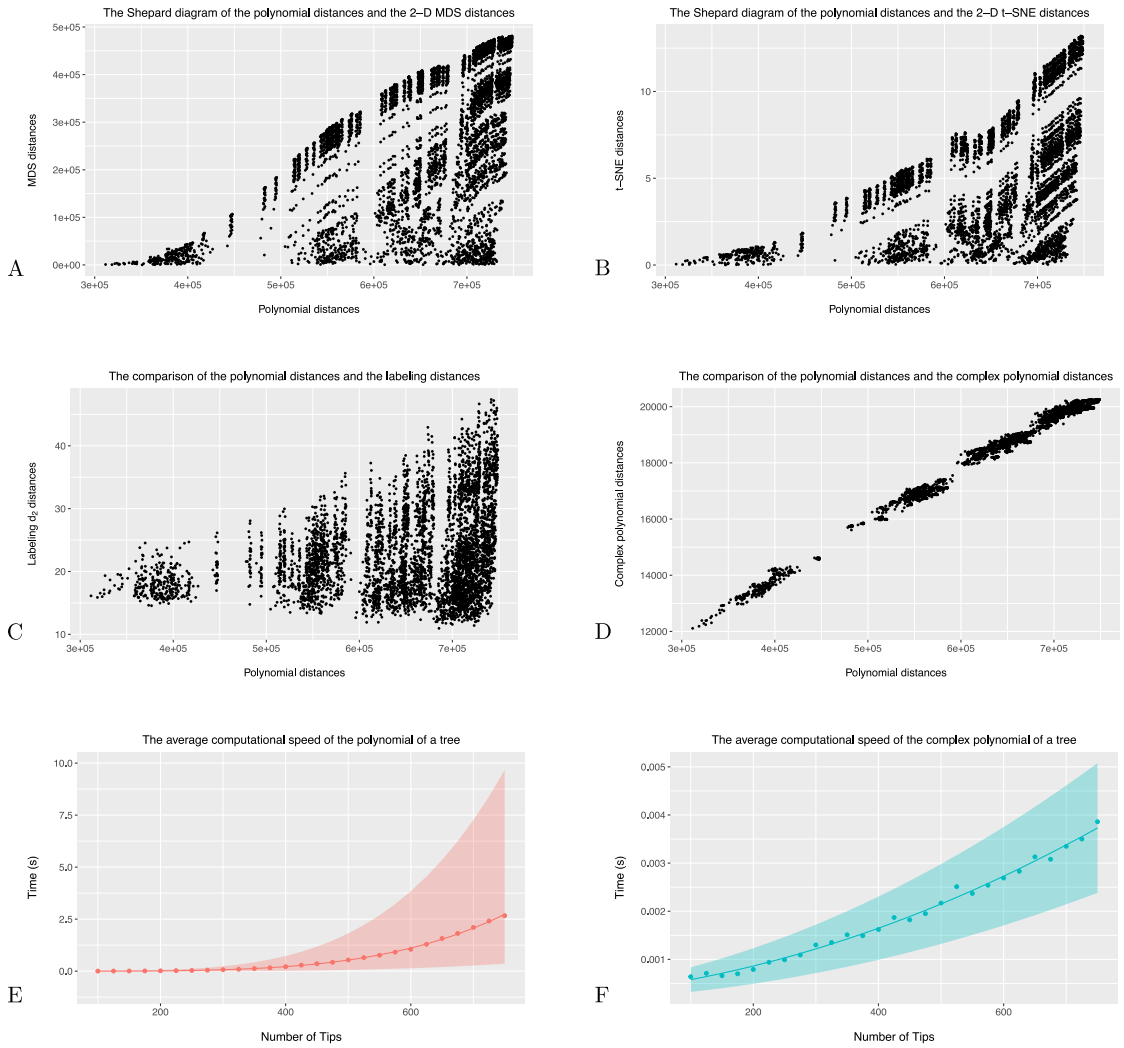


Supplementary Figure 3. The t-SNE plots of the polynomial distances between (A) random trees with 500 tips generated by the four models, (C) influenza trees and (E) HIV trees.The MDS plots of the polynomial binary differences between (A) random trees with 500 tips generated by the four models, (C) influenza trees and (E) HIV trees.

polynomials of two arbitrary tree T_1 and T_2 with n_1 and n_2 tips respectively. We define the complex metric as follows.

$$d_c(T_1, T_2) = \sum_{0 \leqslant i,j \leqslant n} \log\Big(\big|c_1^i - c_2^i\big| + 1\Big).$$

In the formula, $n = \max(n_1, n_2)$ and $|c_1^i - c_2^i|$ denotes the norm of the complex difference. The comparison of the polynomial metric and the complex polynomial metric is displayed in Supplementary Figure 4 D. The complex polynomial of a tree has fewer variables, which



Supplementary Figure 4. The Shepard plots of the polynomial distances and the Euclidean distances in (A) the 2-D MDS plot, (B) the 2-D t-SNE plot of random trees with 100 tips generated by the four models. The comparison of (C) the polynomial distances and the labeling distances d_2 (Colijn, 2018) and (D) the polynomial distances and the complex polynomial distances between random trees with 100 tips generated by the four models. The average computational speed (E) of the polynomial of a rooted binary tree with various number of tips (the data points fit the curve $6.658 \times 10^{-12} x^{4.039}$ with $R^2 = 0.9992$), and (F) of the complex polynomial of a rooted binary tree with various number of tips, (the data points fit the curve $y = 3.672 \times 10^{-9} x^2 + 1.729 \times 10^{-6} x + 3.687 \times 10^{-4}$ with $R^2 = 0.9932$).

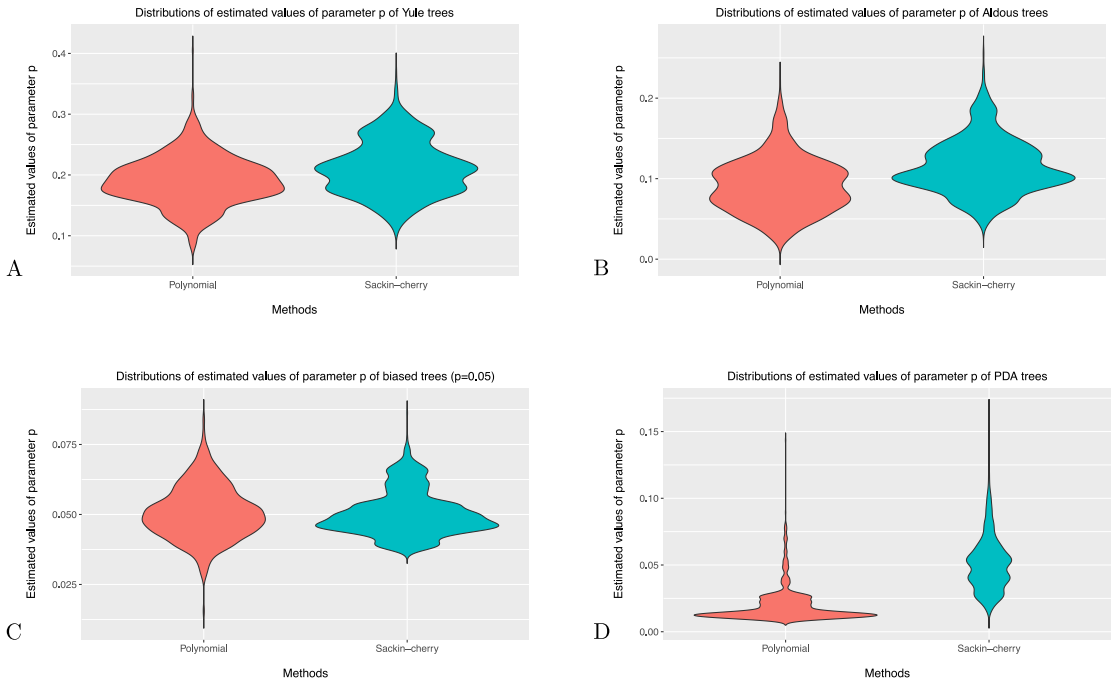
makes it good for regression or machine learning tools, and its computation is more efficient than the general polynomial. This is because computing the general polynomial requires repeated multiplication of polynomials with growing length and magnitude of coefficients. We use a direct approach where every combination of nonzero coefficients must be visited in the multiplication. Since the complex polynomial is a function of one variable,

it is convenient and more efficient to use a one-dimensional convolution for the polynomial multiplication. The computational speeds of the two polynomials are displayed in Supplementary Figure 4 E,F.

Residuals					
Min.	1Q	Median	3Q	Max.	
-0.216949	-0.034936	-0.000871	0.034743	0.258618	
Coefficients					
Interpretation	Term	Estimate	Std. Error	t value	$Pr(>\|t\|)$
2 clades, 25 tips	y^2	-0.0064270	0.0010688	-6.013	1.82e-09 ***
3 clades, 25 tips	y^3	-0.0153878	0.0017633	-8.727	< 2e-16 ***
2 clades, 24 tips	xy^2	-0.0117466	0.0008665	-13.556	< 2e-16 ***
3 clades, 24 tips	xy^3	-0.0115123	0.0017150	-6.713	1.92e-11 ***
4 clades, 24 tips	xy^4	-0.0300614	0.0043451	-6.918	4.59e-12 ***
2 clades, 23 tips	x^2y^2	-0.0057458	0.0008000	-7.182	6.91e-13 ***
3 clades, 23 tips	x^2y^3	-0.0083084	0.0018331	-4.533	5.84e-06 ***
3 clades, 22 tips	x^3y^3	0.0148055	0.0021410	6.915	4.70e-12 ***
2 clades, 21 tips	x^4y^2	0.0076429	0.0008286	9.224	< 2e-16 ***
3 clades, 21 tips	x^4y^3	0.0172132	0.0026160	6.580	4.73e-11 ***
2 clades, 19 tips	x^6y^2	-0.0066565	0.0010694	-6.224	4.85e-10 ***
2 clades, 18 tips	x^7y^2	-0.0086339	0.0012577	-6.865	6.69e-12 ***

Supplementary Table 1. The summary of fit report for estimating parameter p of random trees with 25 tips by linear regression with polynomial coefficients, where only the most statistically significant terms are displayed.

Estimating a parameter by linear regression with the polynomials carries the advantage that the most statistically significant terms in the polynomial can be identified. Since the terms in the polynomial are interpretable, the traits of trees that a related to the parameter can be spotted in the summary of fit report. Supplementary Table 1 displays a report of fitting the biased speciation parameter p with the polynomial coefficients of trees with 25 tips. The most statistically significant terms are listed in the table. Recall that the coefficient $c^{(\alpha,\beta)}$ in the polynomial $P(T)$ indicates the number of ways to choose β clades with $n-\alpha$ tips in total for a tree T with n tips. For example, the coefficient of the y^3 term represents the number of ways to choose 3 clades that contain all the tips of the tree and the coefficient of the x^4y^2 term represents the number of ways to choose 2 clades that contain 21 tips in total.

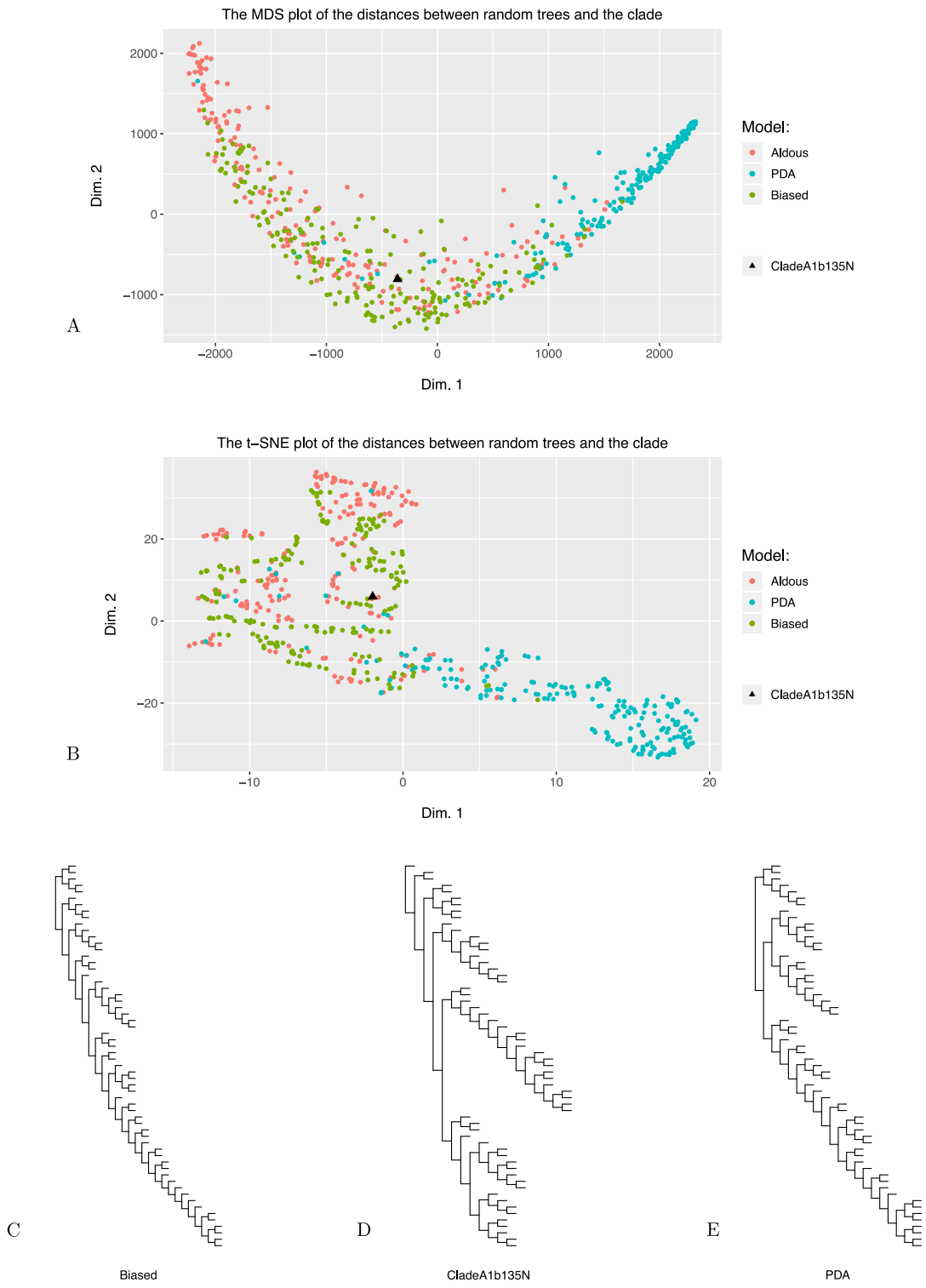


Supplementary Figure 5. Distributions of estimated values of parameter p for random trees by 3-nearest neighbors. 1,000 trees with 100 tips are generated for each model.

To investigate the observation in Figure 3 F that the polynomial-estimated values of the parameter for the influenza virus clades are smaller than the values estimated with Sackin-cherry vectors, we generate 1,000 trees with 100 tips using each of the Yule model, the PDA model, the Aldous model and the biased speciation model with $p = 0.05$. We estimate the parameter p of these trees by the 3-nearest neighbor method with polynomials and Sackin-cherry vectors respectively. The results are displayed in Supplementary Figure 5. We find that for trees generated by the PDA model and the Aldous model, the polynomial-estimated values of the parameter p are smaller than the values estimated with Sackin-cherry vectors. The results suggest that the influenza clades are more similar to the trees generated by the PDA model or the Aldous model than to the other models.

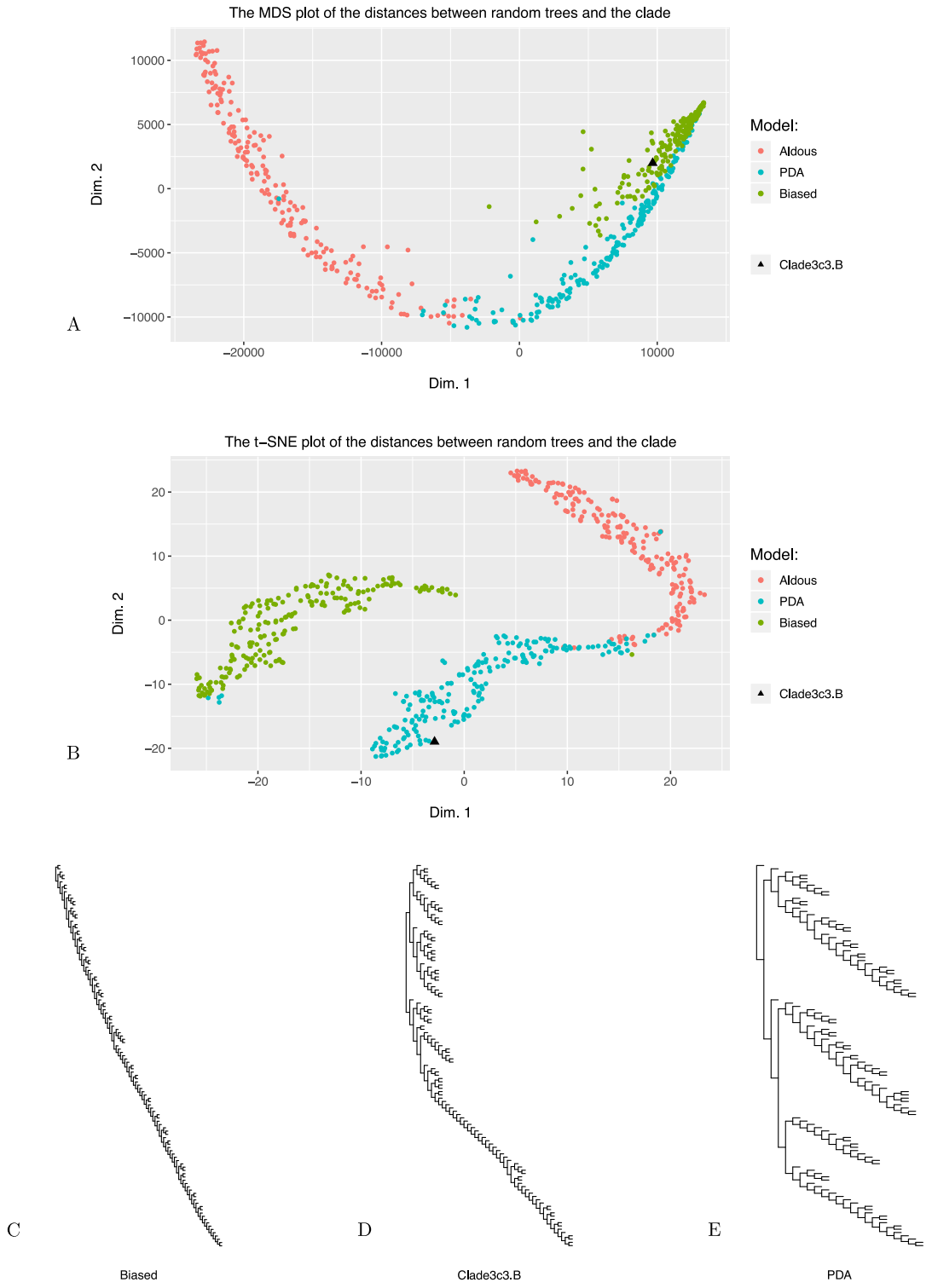
For each of the influenza clades, we generate 200 random trees using each of the Aldous model, the PDA model and the biased model with parameter p equal the polynomial estimated value for the clade. We compute the polynomial distances between

the random trees and the clade. We also plot the random tree and the biased tree that have the minimum polynomial distance to the clade. See Supplementary Figure 6, 7, 8 and 9. These results coincide with the results predicted by the naive Bayes classifiers.

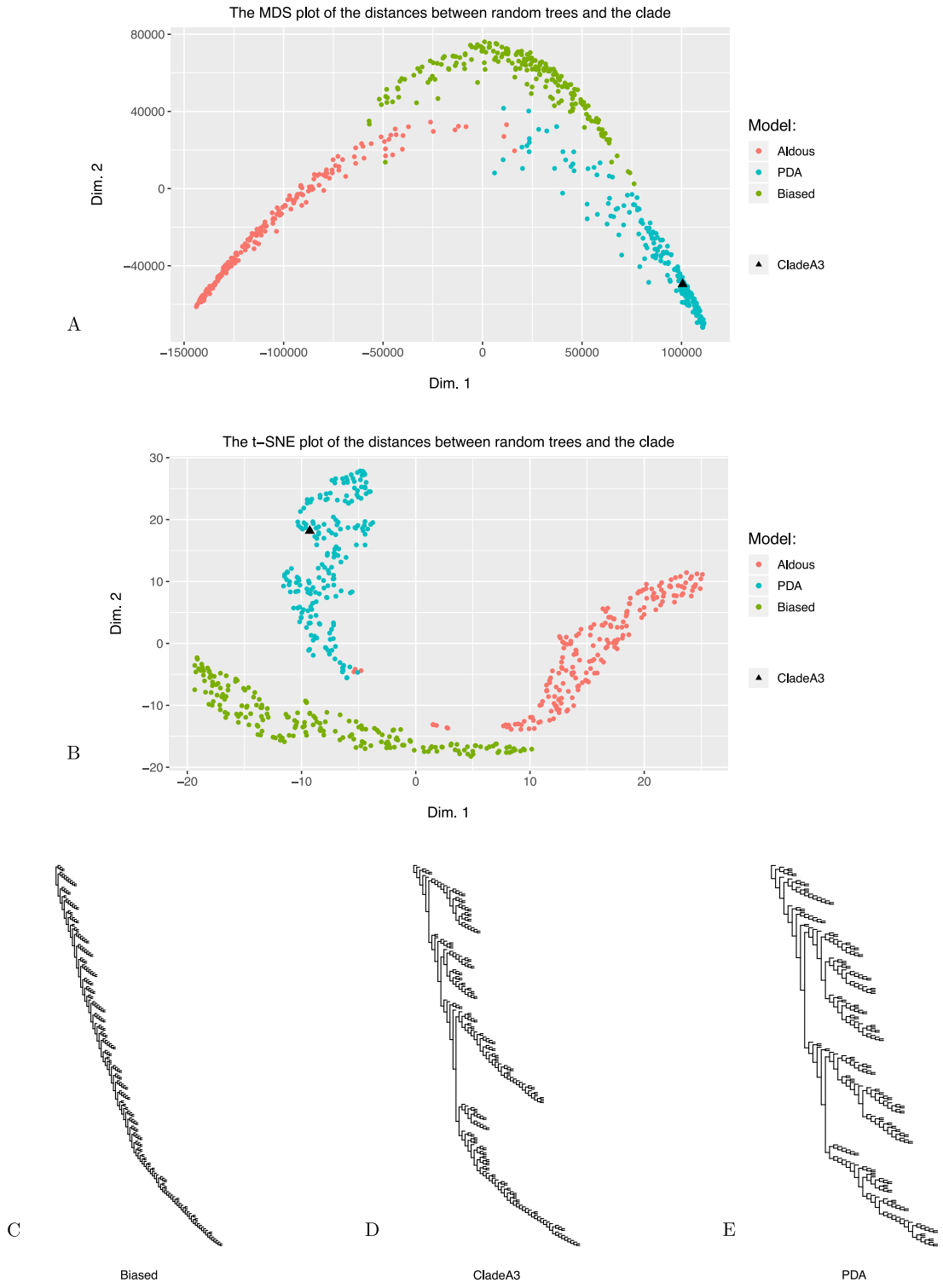


Supplementary Figure 6. The polynomial distances between the random trees and clade A1b135N visualized by (A) MDS and (B) t-SNE. The tree shapes of (C) the nearest biased tree in the sample to the clade (with polynomial distance 4331.28), (D) clade A1b135N, and (E) the nearest tree (an Aldous tree) in the sample to the clade (with polynomial distance 4210.54).

Supplementary Figure 7. The polynomial distances between the random trees and clade A1b135K visualized by (A) MDS and (B) t-SNE. The tree shapes of (C) the nearest biased tree in the sample to the clade (with polynomial distance 5302.63), (D) clade A1b135K, and (E) the nearest tree (an Aldous tree) in the sample to the clade (with polynomial distance 5240.62).



Supplementary Figure 8. The polynomial distances between the random trees and clade 3c3.B visualized by (A) MDS and (B) t-SNE. The tree shapes of (C) the nearest biased tree in the sample to the clade (with polynomial distance 27516.24), (D) clade 3c3.B, and (E) the nearest tree (a PDA tree) in the sample to the clade (with polynomial distance 27389.11).



Supplementary Figure 9. The polynomial distances between the random trees and clade A3 visualized by (A) MDS and (B) t-SNE. The tree shapes of (C) the nearest biased tree in the sample to the clade (with polynomial distance 293371.38), (D) clade A3, and (E) the nearest tree (a PDA tree) in the sample to the clade (with polynomial distance 258466.72).

References

P. Agapow and A. Purvis. 2002. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Systematic Biology*. 51(6):866–72.

C. Aggarwal, A. Hinneburg and D. Keim. 2001. On the surprising behavior of distance metrics in high dimensional spaces. *Proceedings of the International Conference on Database Theory*. 420–434.

D. Aldous. 1996. Probability distributions on cladograms. In: D. Aldous, R. Pemantle and editors, Random discrete structures. *Springer IMA Volumes in Mathematics and its Application*. 76:1–18.

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

D. Andrén and K. Markström. 2009. The bivariate Ising polynomial of a graph. *Discrete Appl. Math.* 157:2515–24.

T. Bedford et al.. 2015. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*. 523(7559):217–20.

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

M. Binet et al.. 2016. Fast and accurate branch lengths estimation for phylogenomic trees. *BMC Bioinformatics*. 17(23); doi: 10.1186/s12859-015-0821-8.

M. Blum and O. François. 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Systematic Biology*. 55(4):685–91.

J. Brown et al.. 2010. When Trees Grow Too Long: Investigating the Causes of Highly Inaccurate Bayesian Branch-Length Estimates. *Systematic Biology*. 59(2):145—161.

S. Chaudhary and G. Gordon. 1991. Tutte polynomials for trees. *J. Graph Theory*. 15:317–331.

C. Chewapreecha et al.. 2014. Dense genomic sampling identifies highways of pneumococcal recombination. *Nature Genetics* 46(3):305–309.

L. Chindelevitch et al.. 2019. Network science inspires novel tree shape statistics. *Preprint*. bioRxiv 608646; doi: https://doi.org/10.1101/608646.

S. Choi, S. Cha, and C. Tappert. 2010. A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*. 8(1):43–48.

C. Colijn and G. Plazzotta. 2018. A metric on phylogenetic tree shapes. *Systematic Biology*. 67:113–126.

D. Colless, 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*. 31(100).

A. Dayarian and B. Shraiman. 2014. How to infer relative fitness from a sample of genomic sequences. *Genetics*. 197(3):913–23.

P. Diaconis, S. Goel, and S. Holmes. 2008. Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*. 2(3):777–807.

S. Frost and E. Volz. 2013. Modelling tree shape and structure in viral phylodynamics. *Phil. Trans. R. Soc. B*. 368; doi: :http://doi.org/10.1098/rstb.2012.0208

G. Fusco and Q. Cronk. 1995. A new method for evaluating the shape of large phylogenies. *Journal of Theoretical Biology*. 175(2):235–243.

B. Grenfell et al.. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332.

J. Hadfield et al.. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 34(23):4121–4123.

M. Hayati, P. Biller and C. Colijn. 2019. Predicting the short-term success of human influenza A variants with machine learning. *Preprint*. bioRxiv 609248; doi: https://doi.org/10.1101/609248

G. Hunt et al.. 2013. Surveillance of transmitted HIV-1 drug resistance in 5 provinces in South Africa in 2011. *Communicable Diseases Surveillance Bulletin*. 11:122–124.

V. Jones. 1985. A polynomial invariant for knots via von Neumann algebras. *Bull. Amer. Math. Soc.* 12:103–111.

M. Kendall and C. Colijn. 2016. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular Biology and Evolution*. 33(10):2735–43.

M. Kendall, V. Eldholm and C. Colijn. 2018. Comparing phylogenetic trees according to tip label categories. *Preprint*. bioRxiv 251710; doi: https://doi.org/10.1101/251710.

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

A. Lambert and T. Stadler. 2013. Birth-death models and coalescent point processes: the shape and probability of reconstructed phylogenies. *Theoretical Population Biology*. 90:113–28.

P. Liu. 2019. A tree distinguishing polynomial. *Preprint*. arXiv: 1904.03332.

J. Losos et al.. 2013. Evolutionary biology for the 21st century. *PLoS Biology*. 11(1):e1001466.

L. Maia, A Colato and J.Fontanar. 2004. Effect of selection on the topology of genealogical trees. *Journal of Theoretical Biology*. 226(3):315–20.

M. Manceau, A. Lambert and H. Morlon. 2015. Phylogenies support out-of-equilibrium models of biodiversity. *Ecology Letters*. 18(4):347–56.

F. Matsen. 2006. A geometric approach to tree shape statistics. *Systematic Biology*. 55(4):652–61.

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

M. Monagan and B. Tuncer. 2018. Factoring multivariate polynomials with many factors and huge coefficients. *CASC*. 11077:319–34.

A. Mooers and S. Heard. 1997. Inferring evolutionary process from phylogenetic tree shape. *The Quarterly Review of Biology*. 31–54.

S. Negami and K. Ota. 1996. Polynomial invariants of graphs II. *Graphs Combin.* 12:189–198.

R. Neher and T. Bedford. 2015. nextflu: Real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*. 31(21):3546–48.

V. Novitsky et al.. 2013. Phylogenetic relatedness of circulating HIV-1C variants in Mochudi, Botswana. *PLoS One*. 8(12):e80589.

G. Plazzotta and C. Colijn. 2016. Asymptotic frequency of shapes in supercritical branching trees. *Journal of Applied Probability*. 53(4):1143–55.

M. Price, P. Dehal, and A. Arkin. 2010. Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*. 5(3):e9490, doi:10.1371/journal.pone.0009490.

A. Purvis et al.. 2011. The shape of mammalian phylogeny: patterns, processes and scales. *Philosophical Transactions of the Royal Society B*. 366(1577):2462–77.

D. Robinson and L. Foulds. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences*. 53(1-2):131–47.

C. Russell et al.. 2008. The global circulation of seasonal influenza a (H3N2) viruses. *Science*. 320:340–46 .

M. Sackin. 1972. “Good” and “bad” phenograms. *Systematic Zoology*. 21(2):225–26.

R. Safavian, and D. Landgrebe. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*. 21(3):660–74.

A. Stamatakis. 2014. RAxML version 8: a tool for phylogenetic analysis andpost-analysis of large phylogenies. *Bioinformatics*. 30(9):1312–13.

M. Stich and S. Manrubia. 2009. Topological properties of phylogenetic trees in evolutionary models. *The European Physical Journal B*. 70(4):583–92.

T. To et al.. 2016. Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology* 65(1):82–97.

W. Tutte. 1954. A contribution to the theory of chromatic polynomials. *Can. J. Math.* 6:80–91.

L. van der Maaten and G. Hinton. 2008. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research.* 9(11):2579–2605.

E. Volz, K. Koelle and T. Bedford. 2013. Viral phylodynamics. *PLoS Computional Biology.* 9(3):e1002947.

E. Wolf et al.. 2017. Phylogenetic evidence of HIV-1 transmission between adult and adolescent men who have sex with men. *AIDS Research and Human Retroviruses.* 33:318–22.

T. Wu and K. Choi. 2016. On joint subtree distributions under two evolutionary models. *Theoretical Population Biology*. 108:13–23.