Supplement for: A metric on phylogenetic tree shapes

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

February 9, 2017

1 Supplementary Results

The Euclidean nature of the d_2 metric allows techniques such as principal components analysis to be used to find low-dimensional representations of a set of trees. We have used discriminant analysis of principal components (DAPC) [**?**] to determine which components (sub-tree shapes) best distinguish the tropical vs USA influenza trees, and the simulated birth-death trees with different birth-to-death ratios (R_0). In both cases, a single axis separates the groups of trees almost entirely; Figure S1 illustrates this, showing the value of the first component.

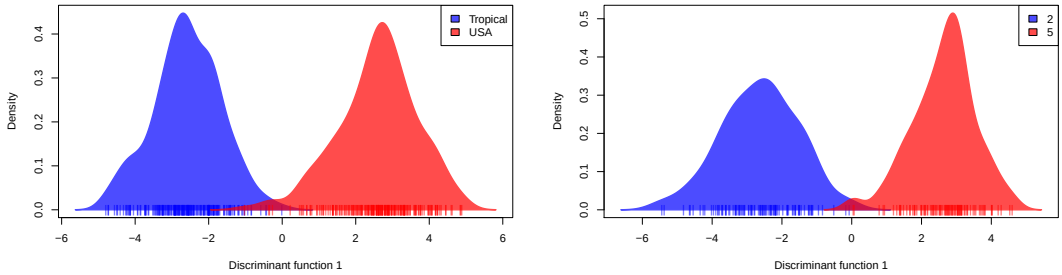

Figure S1: DAPC: One principal component separates the groups of trees in both cases. Left: tropical vs USA trees. Right: simulated birth-death trees.

Since there is only one principal component and it almost entirely separates the trees into two groups (this is the case both for the tropical vs USA and for the birth-death trees), it is straightforward to determine which sub-trees make up this principal component and therefore best separate the groups. The loadings of the vector entries v_i, corresponding to trees i (1 for a tip, 2 for a cherry and so on), reflect the importance of the $i'th$ tree in distinguishing the groups. Figure S2 shows the sub-trees that are more frequently present in the two groups of birth-death trees, in parallel with Figure 3 in the main text (for the tropical vs USA trees).

Imbalance, or asymmetry, is the most widely-discussed scalar measure of tree shape. Indeed, for rooted full binary trees, in the absence of branch length considerations, the most natural quantity to examine at each node is the difference between the numbers of descendants on the two sides (the key quantity in the Colless imbalance), and/or the path lengths from the tips to the root. Imbalance does a good job in separating the groups, but cannot in itself reveal which imbalanced sub-trees are more highly represented in which groups.

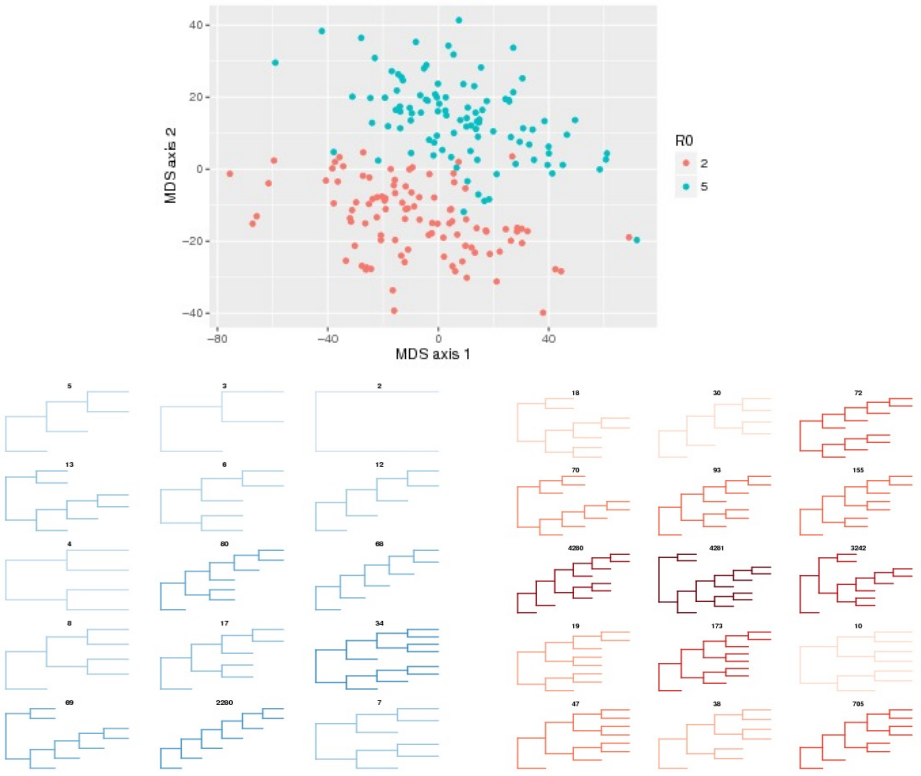

Figure S2: Multi-dimensional scaling plot comparing the trees from the simulated birth-death process. Lower: tree shapes that distinguish the two groups as determined by discriminant analysis of principal components. Colour represents groups in the lower panel, so the blue subtrees are more prevalent among the trees with $R_0 = 5$ and the red more prevalent in $R_0 = 2$. The trees' integer labels correspond to the labelling scheme. Depth of colour corresponds to Sackin imbalance, with darker shading corresponding to higher imbalance.

We illustrate how the shape metric can be extended to include branch lengths and other features of the trees, even if those themselves are not metrics and do not uniquely define a tree. Let V be a new vector, whose first k components are various summary features or other properties. Here, let $V(1)$ be the ratio of the mean terminal branch length in a tree to the mean internal branch length. This captures how "star-like" a tree is, where star-like trees have long terminal branches and short internal ones. Let the remaining components of V be the counts of the labels, as we have done throughout: number of 1s, 2s, 3s and so on. Then the metric $\hat{d}$ is

$$\hat{d}(T_1, T_2) = ||w \cdot V^a - w \cdot V^b|| = \sqrt{(w_0 V_1^a - w_2 V_1^b)^2 + \sum_{i=1}^{L} w_i (v_i^a - v_i^b)^2} \quad (1)$$

where superscripts refer to which tree a, b the entry is from, and subscripts 1 and i refer to the entry of the vector. $\hat{d}$ is Euclidean because it is the standard Euclidean distance between two vectors. The weights w should be chosen to reflect the desired weighting and the natural scaling of different variables; the counts in v are integers and the natural unit is 'number of occurrences'; to compare these to branch lengths in substitutions per site requires a scaling choice. In Figure S4 we use a weight of $w_0 = 1$ and $w_i = 0.00067$ for all i, to compensate for the fact that the mean branch length is much less than 1. Figure S4 illustrates that metric $\hat{d}$ retains the shape separation

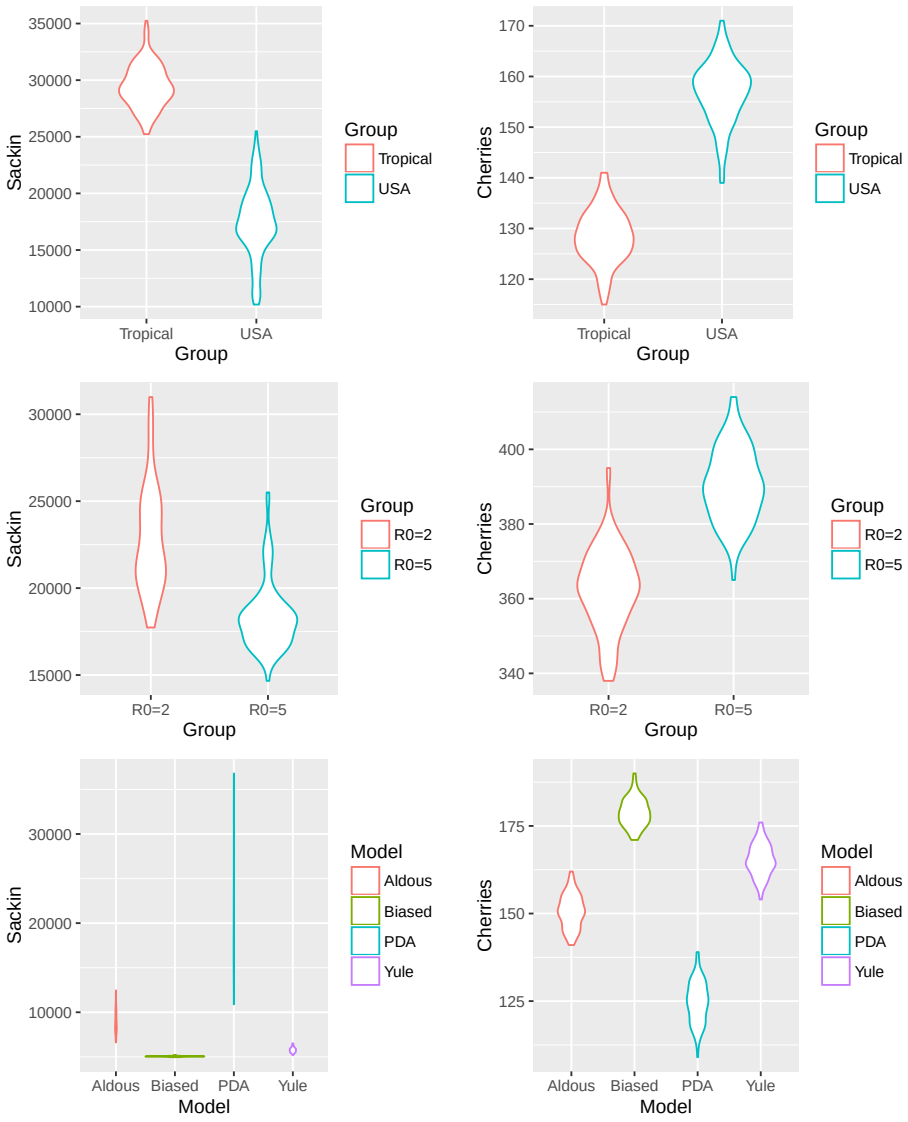

Figure S3: Two standard tree summary statistics, the Sackin imbalance and the number of cherry configurations, in the tropical and USA flu trees, the simulated birth-death trees with different values of the basic reproduction number R_0 and in the trees from different random models.

and additionally identifies similarity between outliers in the length statistic.

Figure S5 illustrates tree shapes together with their labels under the map $\psi_{\mathbb{Z}}$.

Figure S4: Multi-dimensional scaling plot derived from distances $\hat{d}$, containing a weighted combination of the shape distance and a length-based comparison of the ratio between the mean terminal branch length and the mean internal branch length in each tree.

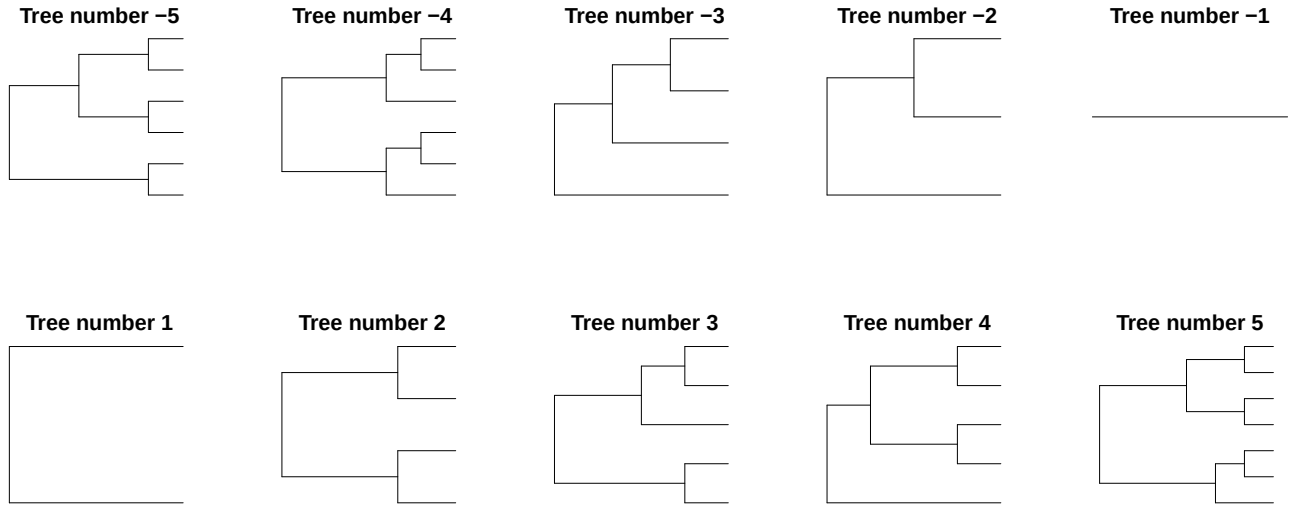

Figure S5: Some trees and their associated integers using the map $\psi_{\mathbb{Z}}$ of Example 1. The numbering goes from -5 to 5, with the exception of 0 which corresponds to the "empty tree".

2 Extension to multifurcations and sampled ancestors

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

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 children, 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 $\phi_4(T)$ in this list. We seek to find an explicit expression for the label $\phi_4(T)$ – the order in the list – for the tree (k, j, l, m). We begin by fixing k and finding how many such labels there are, going from $(k, 0, 0, 0)$ up to (k, k, k, k). Summing these over $m < k$ will give the explicit expression for the label.

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 ϕ_4 of the tree $(k + 1, 0, 0, 0)$, noting that $\phi_4(k, k, k, k)$ is the sum of the numbers of labels beginning with 1, 2, ... k. $\phi_4(k + 1, 0, 0, 0) = 1 + \phi_4(k, k, k, k)$ (and we write 1 as $\binom{3}{3}$):

$$\phi_4(k + 1, 0, 0, 0) = \sum_{x=0}^{k} \binom{x + 3}{3}.$$

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

$$\phi_4(k + 1, 0, 0, 0) = \sum_{x=0}^{k} \binom{x + 3}{3} = \sum_{y=3}^{k+3} \binom{y}{3} = \sum_{y=0}^{k+3} \binom{y}{3} = \binom{k + 4}{4}.$$

To obtain $\phi_4(k, j, l, m)$, we note that

$$\phi_4(k, j, l, m) = \phi_4(k, 0, 0, 0) + \phi_3(j, 0, 0) + \phi_2(l, m)$$

(where ϕ_2 is not precisely the same form as in the main text because here we allow for nodes with only one child). Following the same logic, this is

$$\phi_4(k, j, l, m) = \binom{k + 3}{4} + \binom{j + 2}{3} + \binom{l + 1}{2} + m.$$

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

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

References

[1] Thibaut Jombart, Sébastien Devillard, and François Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.*, 11:94, 15 October 2010.