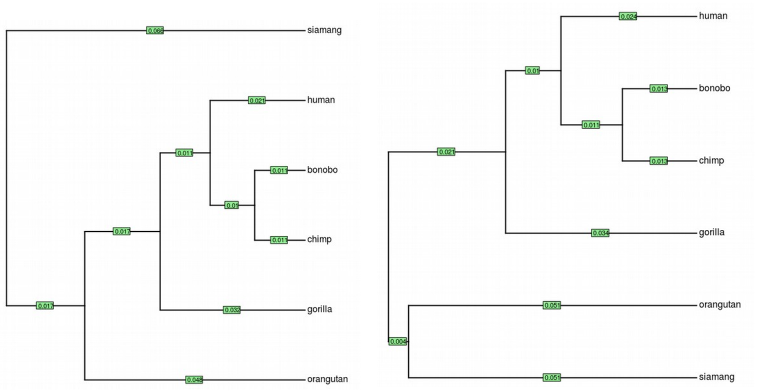
Theo Killian - r0649804
Management of Omics Data
IOU19a

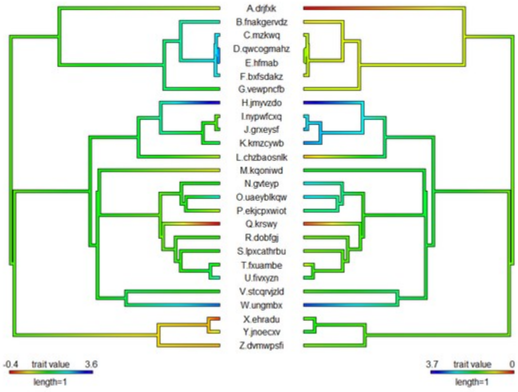
Data Visualization Report

In this report, we explore a datafile “ape.tree” in Nexus format containing 101 phyogenetic trees, visualized utilizing R and RStudio. First, we should briefly explain the properties of the phylogenetic trees in order to better describe the data. In the most abstract form, a tree is a special kind of graph which has no cycles. A phylogeny is a specialized type of tree that efficiently shows the evolutionary relationships among a group of organisms. The edge (or branch) of a phylogenetic tree represents the evolutionary transition from an ancestral taxon to a descendant taxon. Whereas edge length is a number associated with an edge and may represent time or be a measure of genetic distance. Phylogenetic trees may be visualized in several different ways, depending on the evolutionary relationships they describe. A rooted tree is one such example, and is drawn in a single direction with respect to time. There other types of tree visualizations possible, but they are mostly only different aesthetically, and will not be considered further. Binary trees will be used for the visualization of this data because of their simplicity. Binary trees represent an evolutionary history where all speciation events produce exactly two descendants from one common ancestor. All trees that represent this dataset are of this type.

In this datafile, each tree is a treated as an abstract ‘phylo’ object showing the evolutionary relationships among a group of organisms. The organisms found universally in all trees in this dataset are: bonobo, chimp, gorilla, human, orangutan and siamang. Each organism is represented as a taxon on a ‘leaf’ on the phylogenetic tree. For each tree three salient numerical features are present: edge, edge.length and number of nodes. If the total difference between metrics in a tree is small, as it is with edge.length and exact genetic distance of each leaf was desired, something like Figure 1 could be drawn. Whereas if a more general difference plot graphically comparing metrics is desired, something like Figure 2 may be drawn.

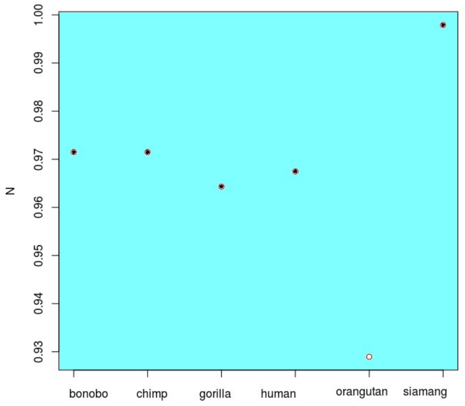
Figure 1. Tree plot of State_0 (left) and State_23000 (center) with edge.length as genetic distance. Figure 2. (right) Mapped phylogenetic trees colored by correlation heatmap. This heatmap can be used with edge distance or edge.length distance

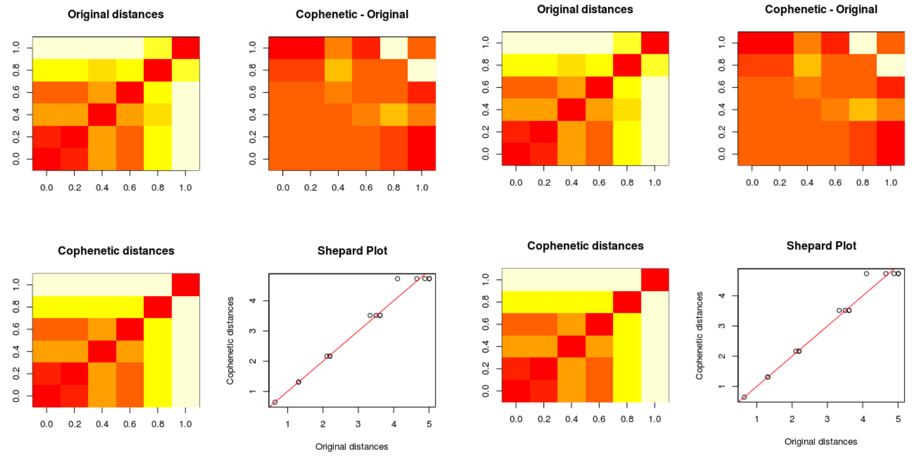




This way of visualizing these differences between each tree might be best served by merging the tree with a heatmap, where edge or edge length is correlated between species and visualized by a red-green or rainbow spectrum with red corresponding to the closest correlation and green/blue corresponding to the lowest correlation between values between particular species. (see Fig. 2) It is also possible to measure the edge and edge.length distance between species within the same tree (the cophenetic distance) and then compare these distances between trees. (see Fig. 3) From this, we can generate an original dissimilarity matrix, sorted on basis of cluster analysis groupings, the cophenetic distances, again sorted as above, the difference between the original dissimilarities and the cophenetic distances as well as generate a Shepard plot comparing the original and cophenetic distances (See Fig 4) ; the better the clustering at capturing the original distances the closer to the 1:1 line the points will lie. Having said all that, however, only the Shepard plot shows the "correlation between clustered data and (dis)similarity matrix," and that is not an image plot (levelplot). However, heatmaps are hard to read at a glance for highly or sparsely correlated data.

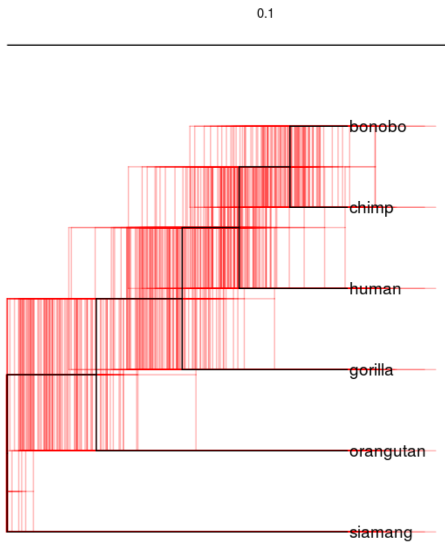
Figure 3. (left) Plot map of correlations between edge.length between species for State_0 and State_23000. Figure 4. Heatmap of cophenetic distance with Shepard plot for State_0 (center) and State_23000 (right).

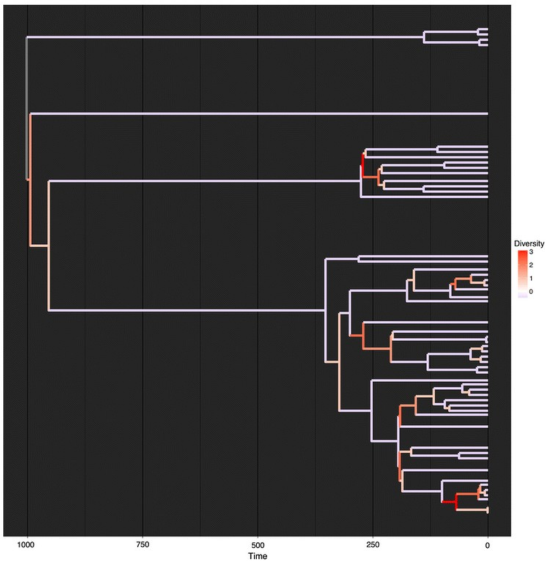




However, when it comes to comparing more than two trees at a time, or comparing them globally, the task of visualization becomes more complicated. While it is possible to plot the entire list of trees, this is cumbersome and not very informative. By averaging all 101 trees in the datafile together and super-imposing them, it is possible to illustrate the ideal mean tree. (see Fig. 4)

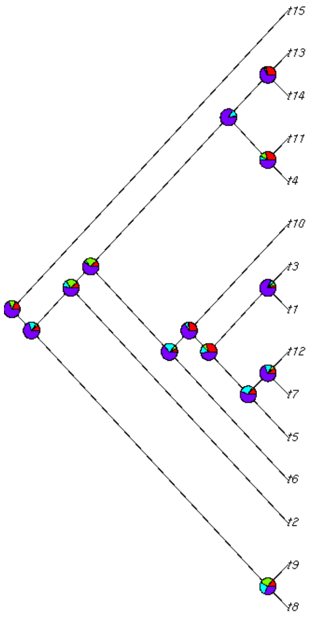
Figure 5. (left) plot of all 101 tree diagrams (red) and ideal mean tree (black). Figure 6. Intensity map of phylogenetic tree focusing on areas of highest diversity.

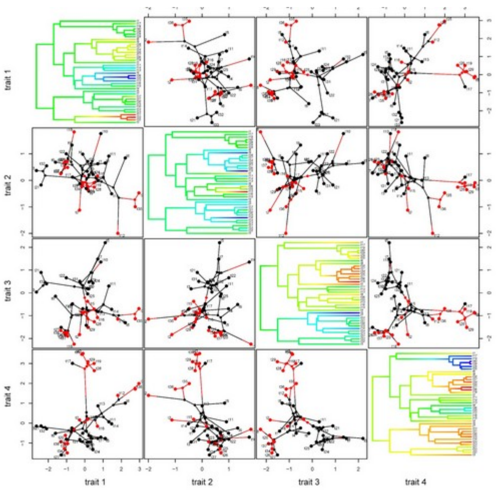


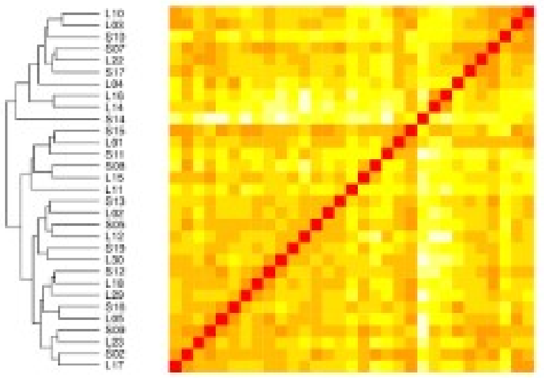


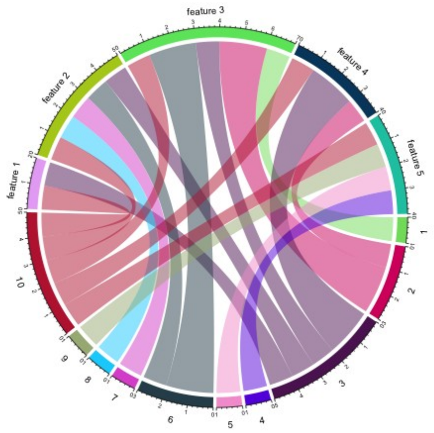
The ideal mean tree gives a global impression of the structure most of the trees in the datafile are similar to and demonstrates the most common relationships between tree branches and average genetic distances between species. Clustering (in red) of particular branches around the ideal tree implies degrees of similarity between branches. This has the advantage of combining similarities and dissimilarities of all metrics into one graphic, but it might be hard for someone to read without an explanation. It can be seen from Fig. 5 that globally, siamangs split off from the hominid lineage first, followed by orangutans, gorillas, then humans, while the ancestor of chimps and bonobos split off from humans to form their own clade. Edge length (e.g. time of divergence) of simangs did not seem to vary that much, whereas the genetic distance of all other lineages varied considerably. This is evidenced by the high R^2 correlations between different trees If the variance of a particular metric is desired to be visualized, a heatmap of the highest variability can be superimposed over the ideal mean tree. (see Fig. 6) Other possible visualizations which were unfortunately unable to able to be implemented in R were are shown below.

Figure 7. (upper left) plot of similarities between trees in an unrooted tree digram. Each pie chart at the node represents variability between the descending trees and branches. Figure 8. (upper right) Multivariate multipanel phylomorphospace plotting where (dis)similarity trees are plotted on the diagonal with metric differences plotted between the trees (in red and black) as geometric morphometric analysis (edge or edge length, for example). Figure 9. (lower left) Heatmap correlation of trees with the tree similarity mapped as a global tree. Figure 10. (lower right) Circular similarity comparison of all trees, where similarity of particular groups of trees by color









Admittedly, there are only so many ways to present data with only two different metrics (edge and edge length) while still preserving interpretability and meaningfullness. This dataset could definitely use more data points, which would add larger weight to patterns found in the data.