Analysis of multiple trees

General information

The 500 loaded trees all have the same tip labels.

The trees have different topologies. The maximum unweighted Robinson-Foulds distance between two trees is **14.00**. The maximum weighted Robinson-Foulds distance between two trees is **0.1073**. Figure 1 represents the trees in a 2-dimensional space, using the Robinson-Foulds metric. Figure 2 also represents the trees in a 2-dimensional space, using instead the weighted Robinson-Foulds metric.

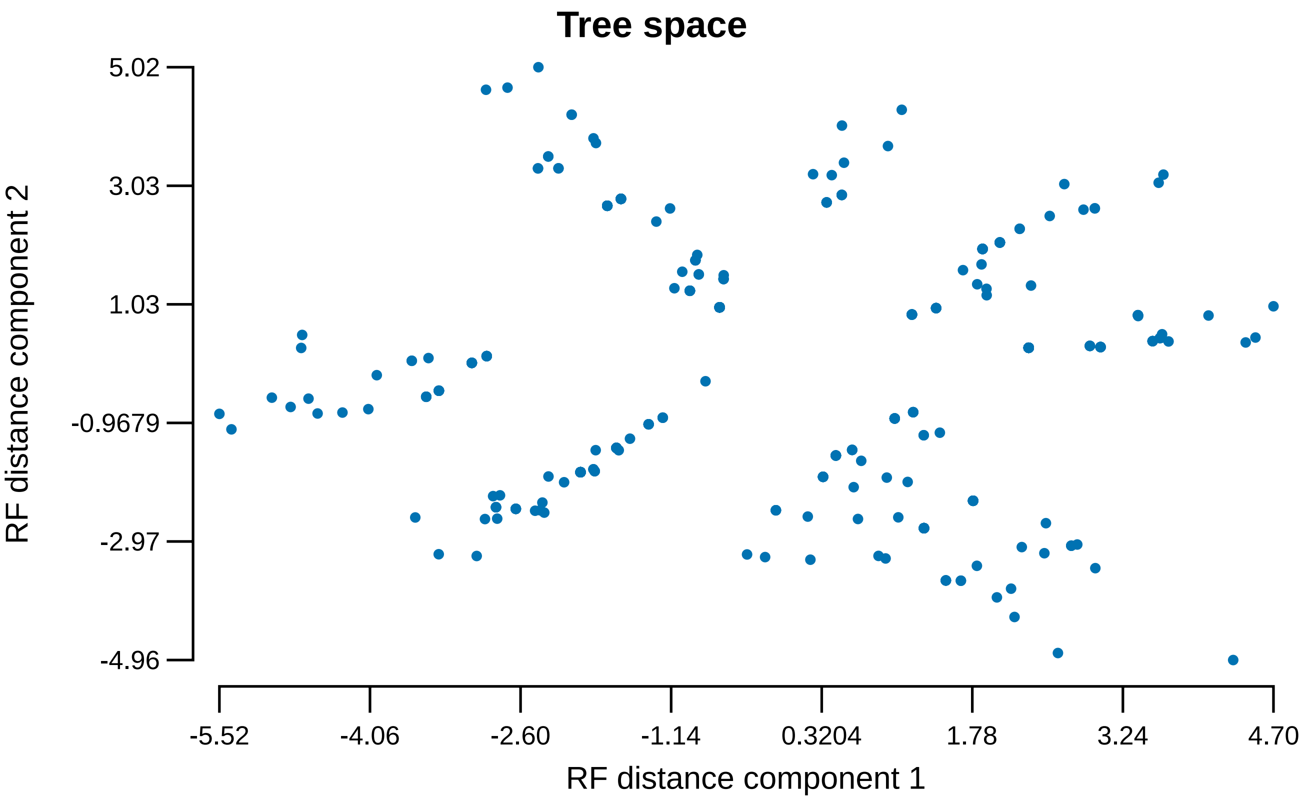

Figure 1. 2-dimensional representation of the trees. Each tree is represented by a point, whose position was determined using multidimensional scaling (MDS); the distance between two points is approximately proportional to the unweighted Robinson-Foulds distance between the corresponding trees. The Duda-Hart test was used to determine that the trees do not show significance evidence for clustering (p ≈ 0.1872, α = 0.001), based on the 2D metric obtained after the MDS analysis.

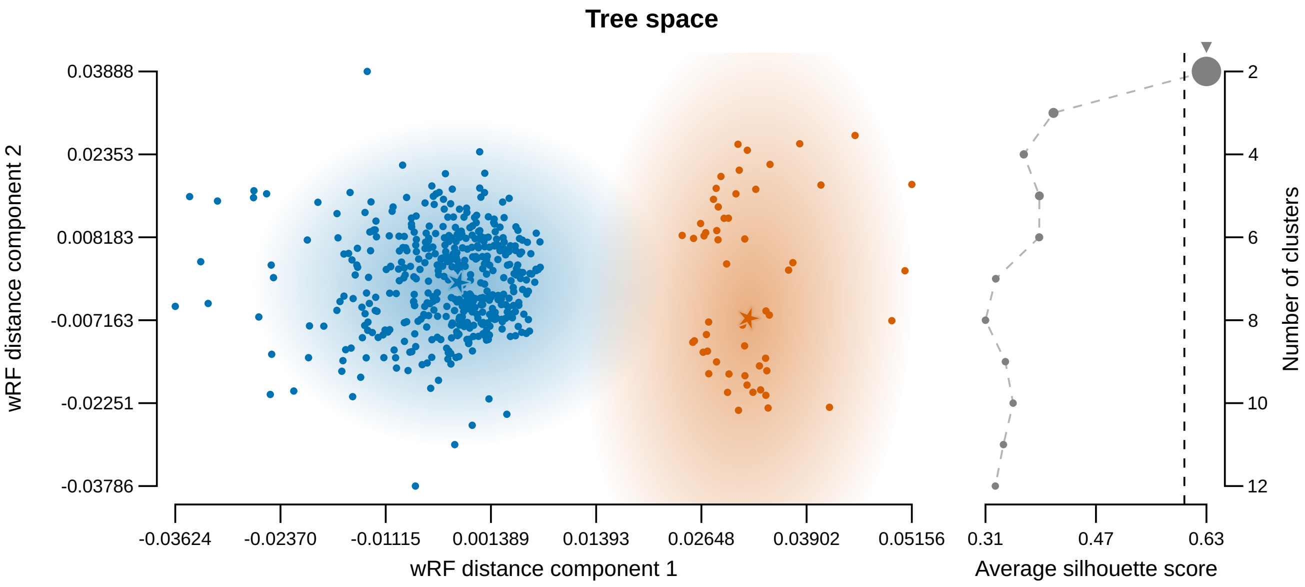

Figure 2. 2-dimensional representation of the trees. *Left plot*: each tree is represented by a point, whose position was determined using multidimensional scaling (MDS); the distance between two points is approximately proportional to the weighted Robinson-Foulds distance between the corresponding trees. Clustering was performed based on the 2D metric obtained after the MDS analysis. The presence of multiple clusters was determined using the Duda-Hart test ($p \approx 0.0003821$, $\alpha = 0.001$). 2 clusters of trees were identified using the K-medoids (PAM) method; the medoid for each cluster is displayed as a star. *Right plot*: the optimal number of clusters for the K-medoids clustering was determined using the average silhouette score of all the points when a certain number of clusters is used. The size of each point is proportional to the cluster size variance when the corresponding number of clusters is used. Clustering with up to 12 clusters was attempted, and for each number of clusters the average silhouette score and the variance of the cluster size were computed; the best number of clusters was selected as the one with the smallest cluster size variance, amongst those with a silhouette score greater than or equal to 95% of the highest observed silhouette score. This threshold is shown by the vertical dashed line. The number of clusters that was actually used in the plot is highlighted by an arrowhead.

Tree shape statistics

Not all the trees are rooted.

The average **Number of cherries** of the trees is **15.82** (89% highest-density interval: 15.00 — 16.00). **Figure 3** shows the distribution of the number of cherries among the trees.

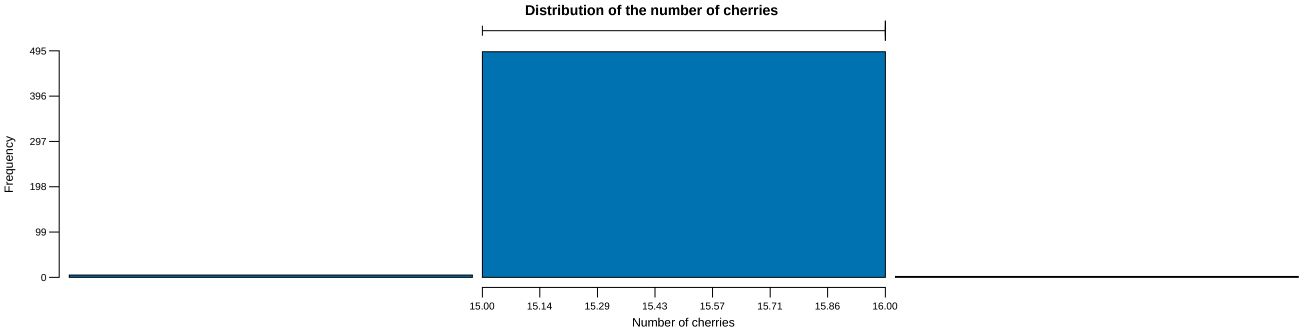

Figure 3. Distribution of the number of cherries. The histogram shows the distribution of the number of cherries among the trees. 5 values smaller than 15.00 are shown in the underflow bin; 2 values greater than 16.00 are shown in the overflow bin. The box and whisker plot at the top represents the median branch length, the interquartile range, and the 89% HDI.