Distances and Ordination

Community Distance

Communities are a vector of abundances:

$$\mathbf{x} = \{x_1, x_2, x_3, \ldots\}$$

E. coli:	● ● ●
P. fluorescens:	●
B. subtilis:	●
P. acnes:	
D. radiodurans:	
H. pylori:	● ● ● ● ● ● ●
L. crispatus:	

$$\mathbf{x} = \{3,1,1,0,0,7,0\}$$

Community Distance Properties

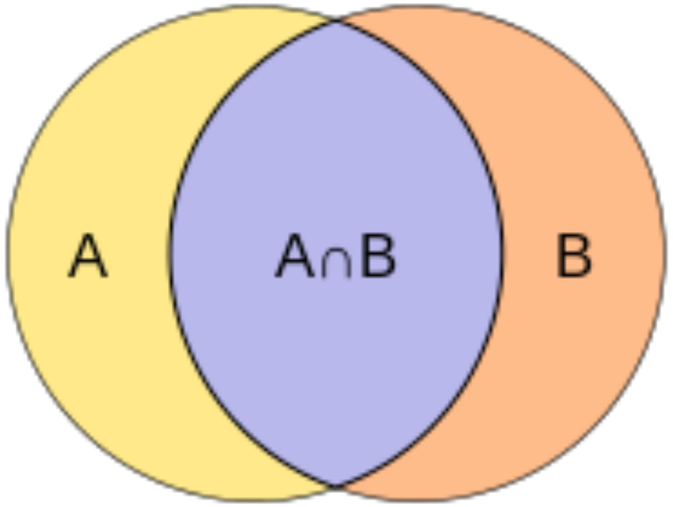
- Range from 0 to 1
- Distance to self is 0
- If no shared taxa, distance is 1
- Triangle inequality (metric)
- Joint absences do not affect distance (biology)
- Independent of absolute counts (metagenomics)

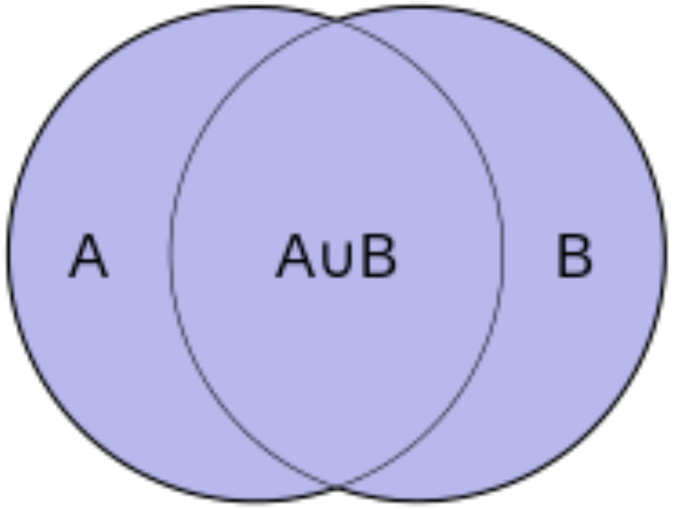
The Distance Spectrum

	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

Jaccard

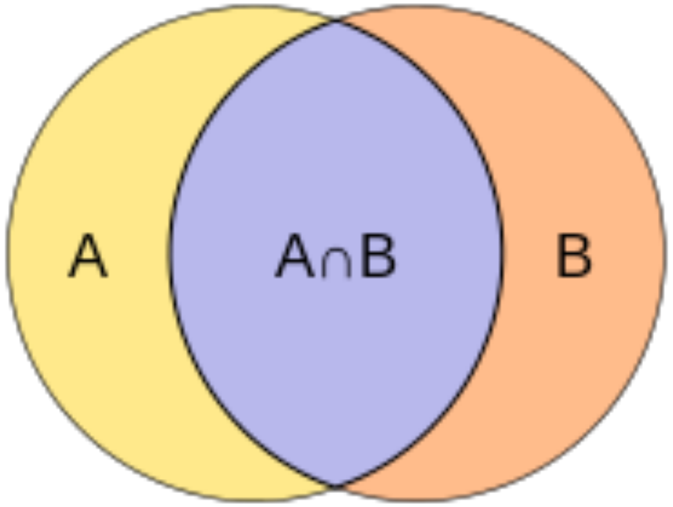
$$\text{Dist}(A, B) = 1 - (A \cap B)/(A \cup B)$$
$$= ((\mathbf{x}_A>0)\ \&\ (\mathbf{x}_B>0))/((\mathbf{x}_A>0)\ |\ (\mathbf{x}_B>0))$$

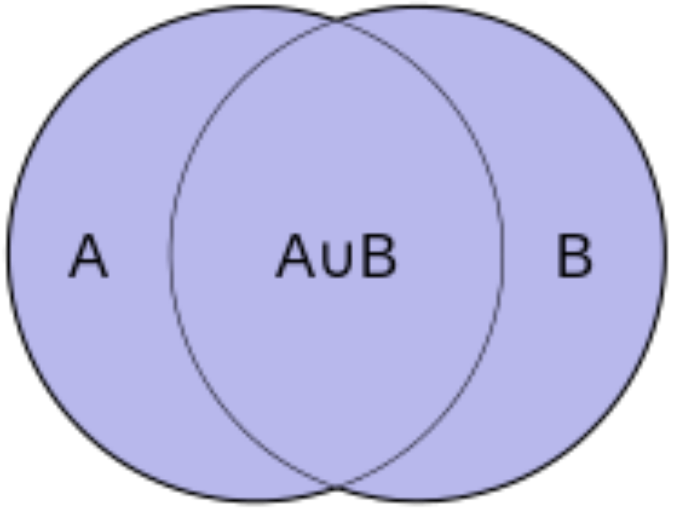




Jaccard

$$\text{Dist}(A, B) = 1 - (A \cap B)/(A \cup B)$$
$$= ((\mathbf{x}_{\mathbf{A}}>0)\ \&\ (\mathbf{x}_{\mathbf{B}}>0))/((\mathbf{x}_{\mathbf{A}}>0)\ |\ (\mathbf{x}_{\mathbf{B}}>0))$$

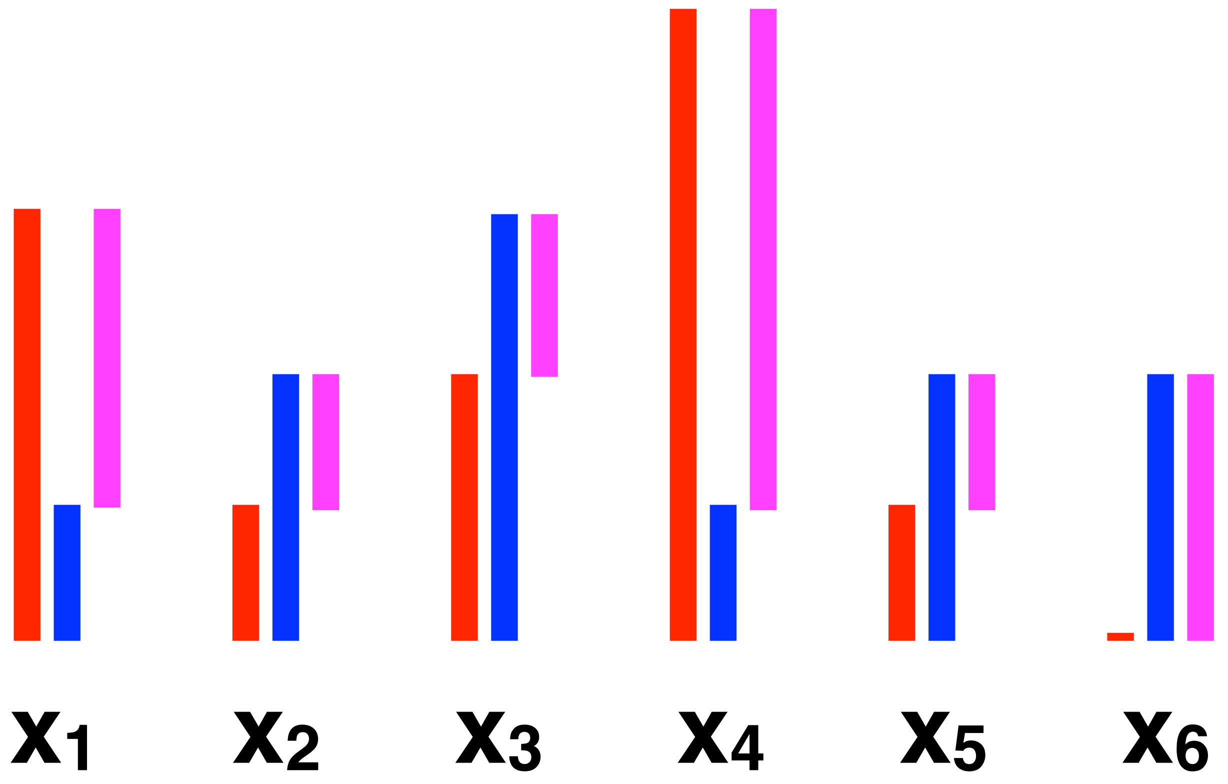




Intuition: Fraction of shared **types** unique to one of the communities

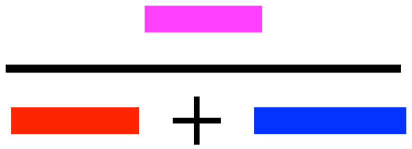
Bray-Curtis

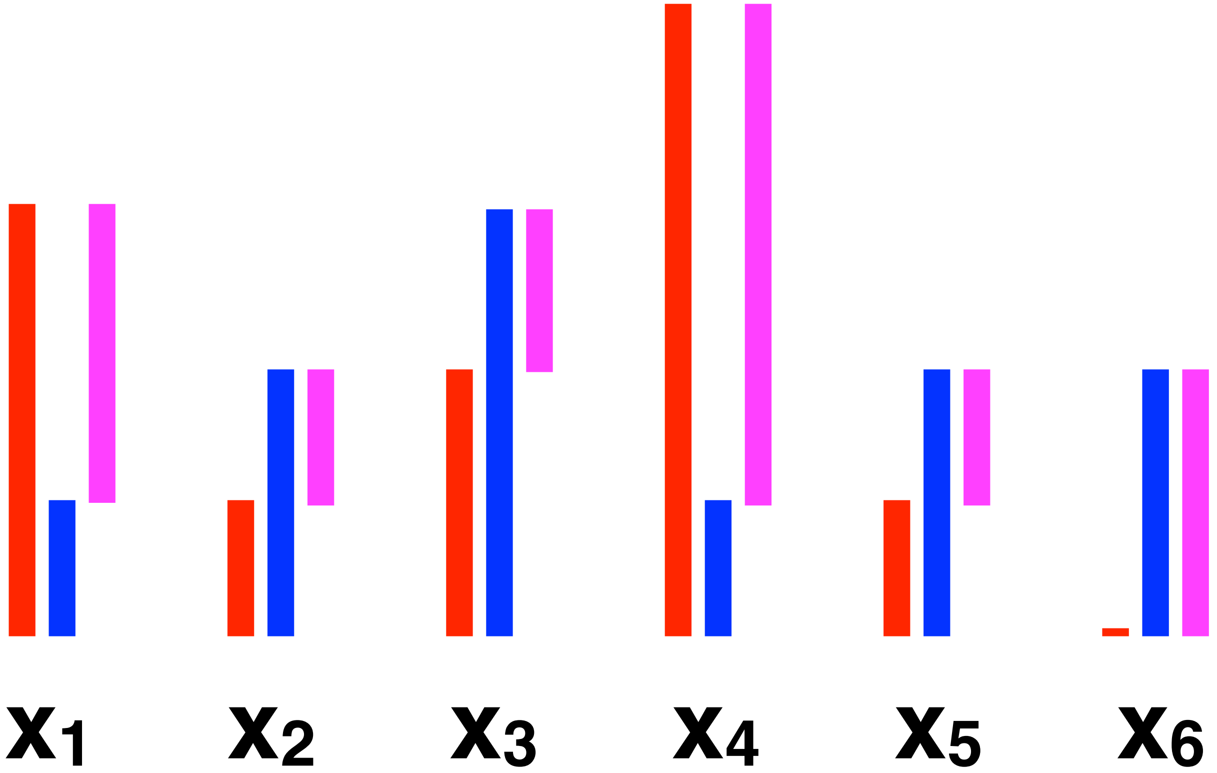
$$\mathrm{Dist}(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} = \frac{\text{🟪}}{\text{🟥} + \text{🟦}}$$



Bray-Curtis

$$\mathrm{Dist}(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} =$$

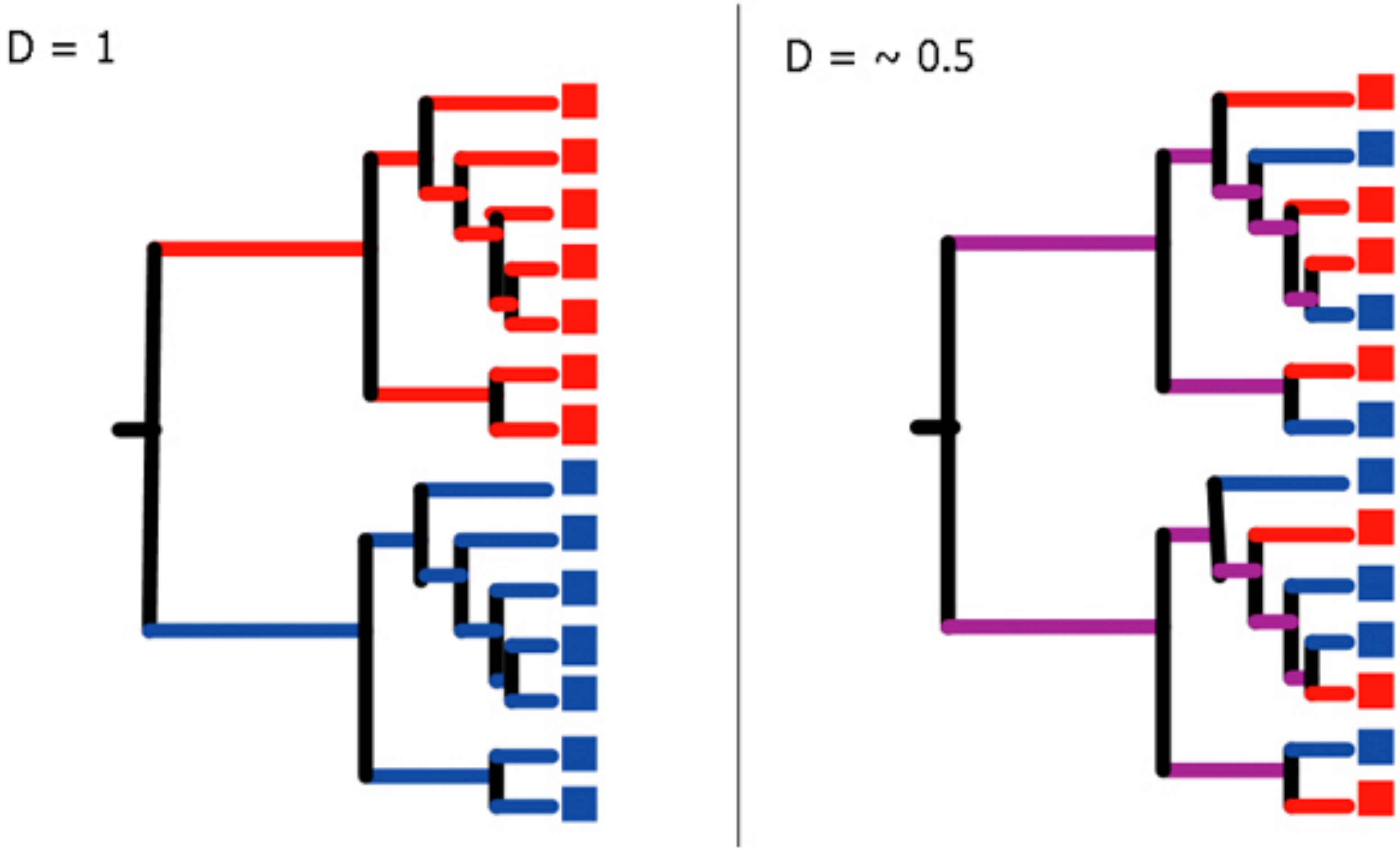




Intuition: *City block distance*. Sum of absolute differences over total abundance.

Unifrac

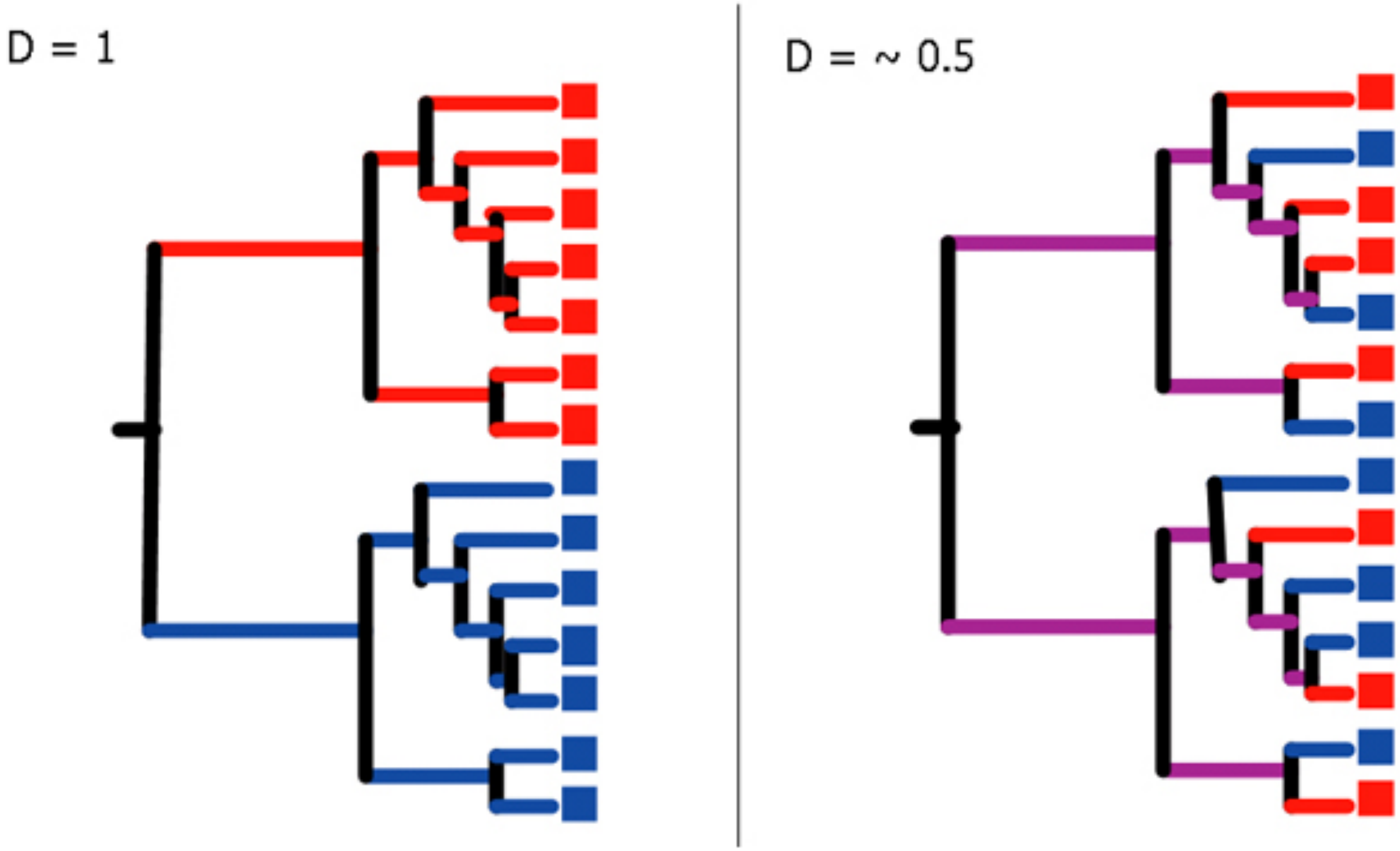
$$\text{Dist}(x, y) = \frac{\text{🟥} + \text{🟦}}{\text{🟥} + \text{🟦} + \text{🟪}}$$



Lozupone and Knight (2008)

Unifrac

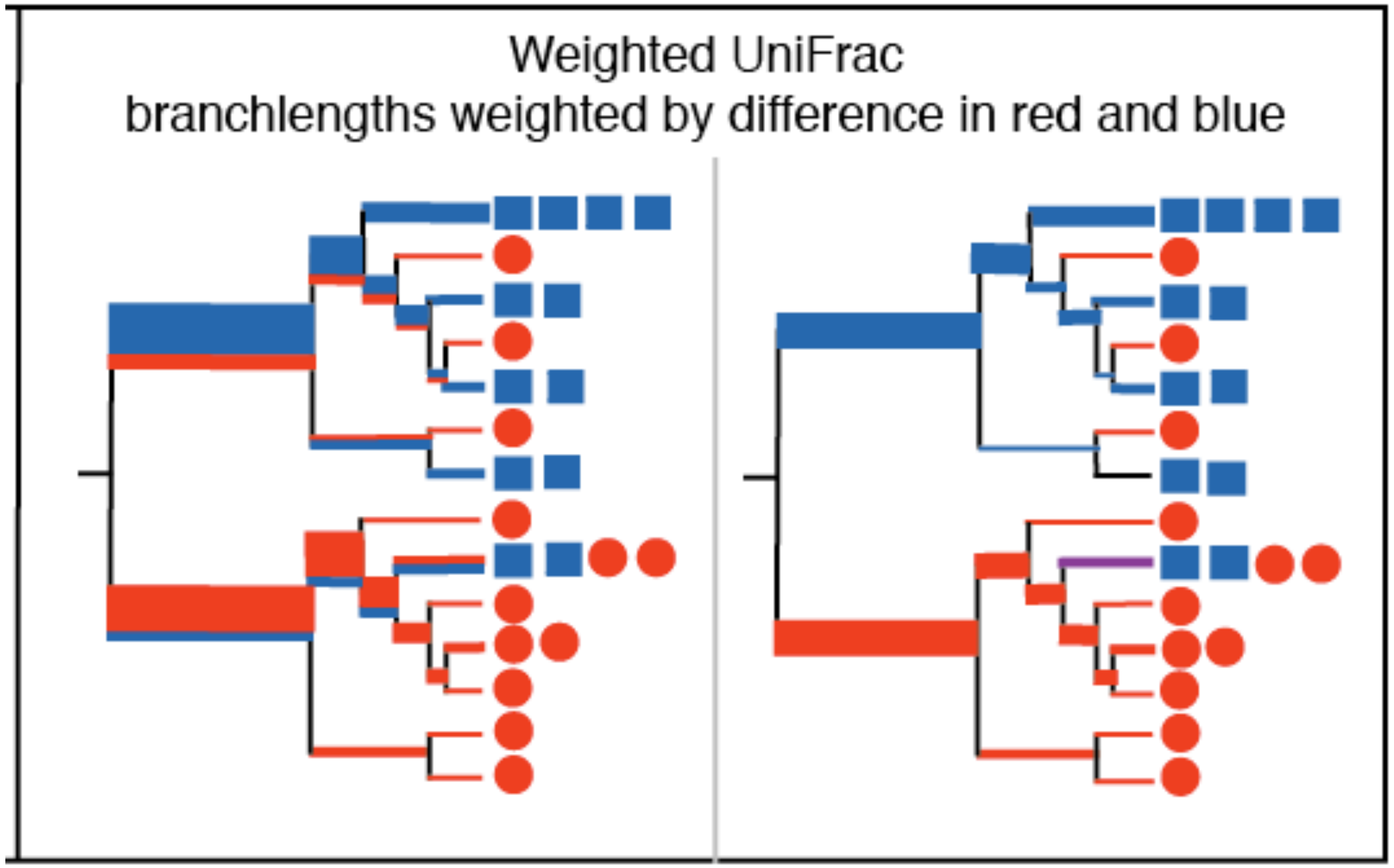
$$\text{Dist}(x, y) = \frac{\text{🟥} + \text{🟦}}{\text{🟥} + \text{🟦} + \text{🟪}}$$



Intuition: Fraction of shared **tree** unique to one of the communities

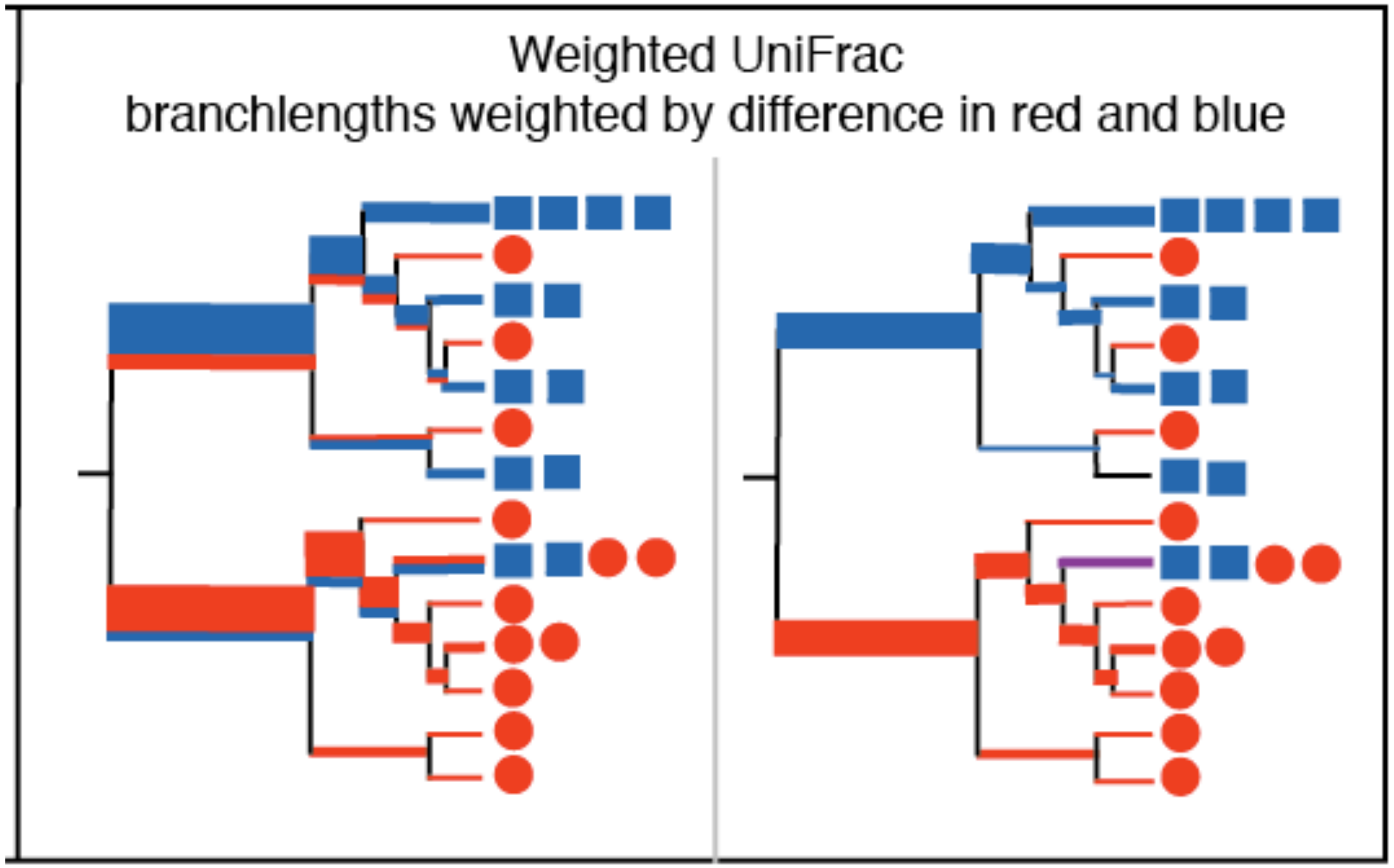
Lozupone and Knight (2008)

Weighted Unifrac



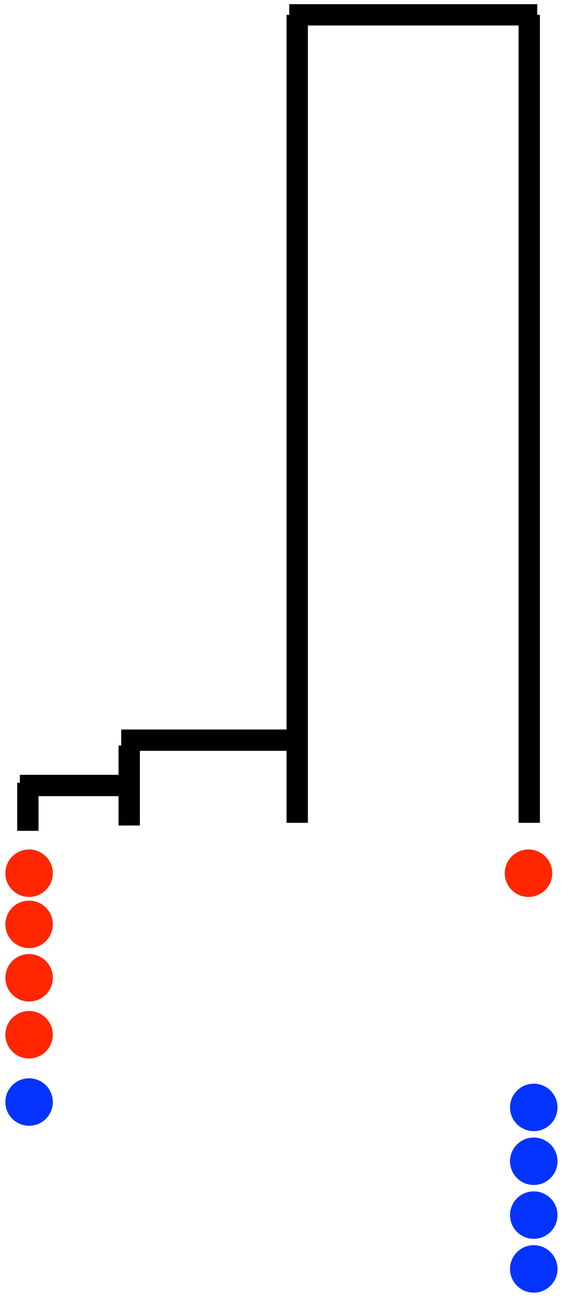
Lozupone et al. (2007)

Weighted Unifrac

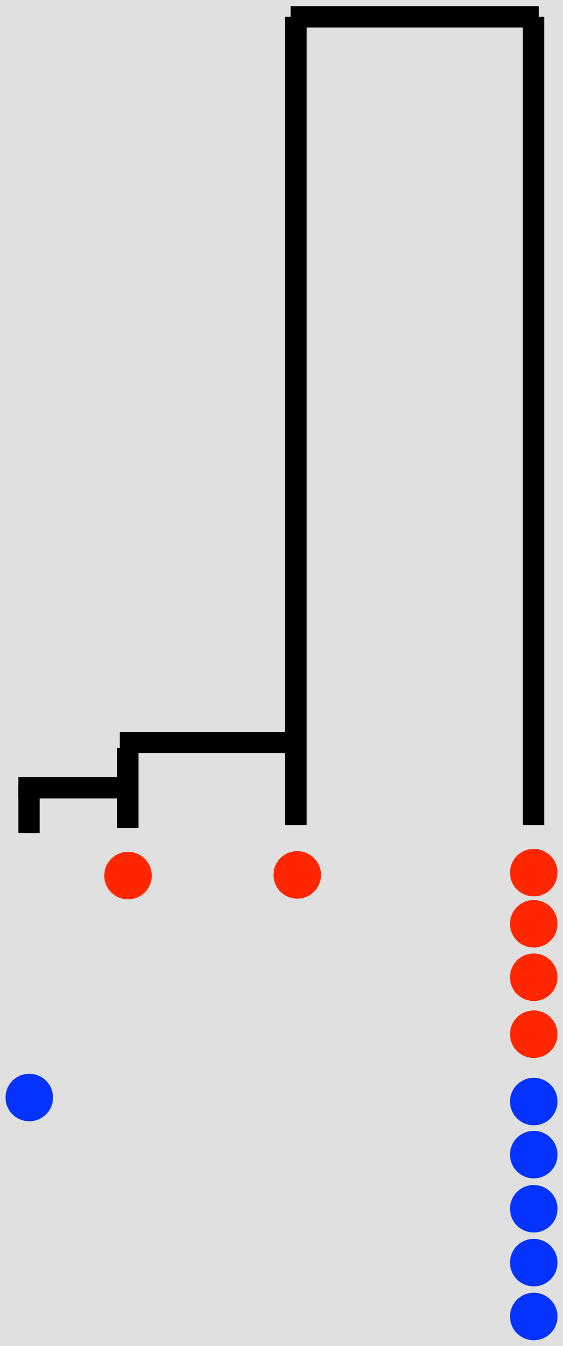


Intuition: The cost of turning one distribution into the other; where the cost is the amount of "dirt" moved times the distance by which it is moved.

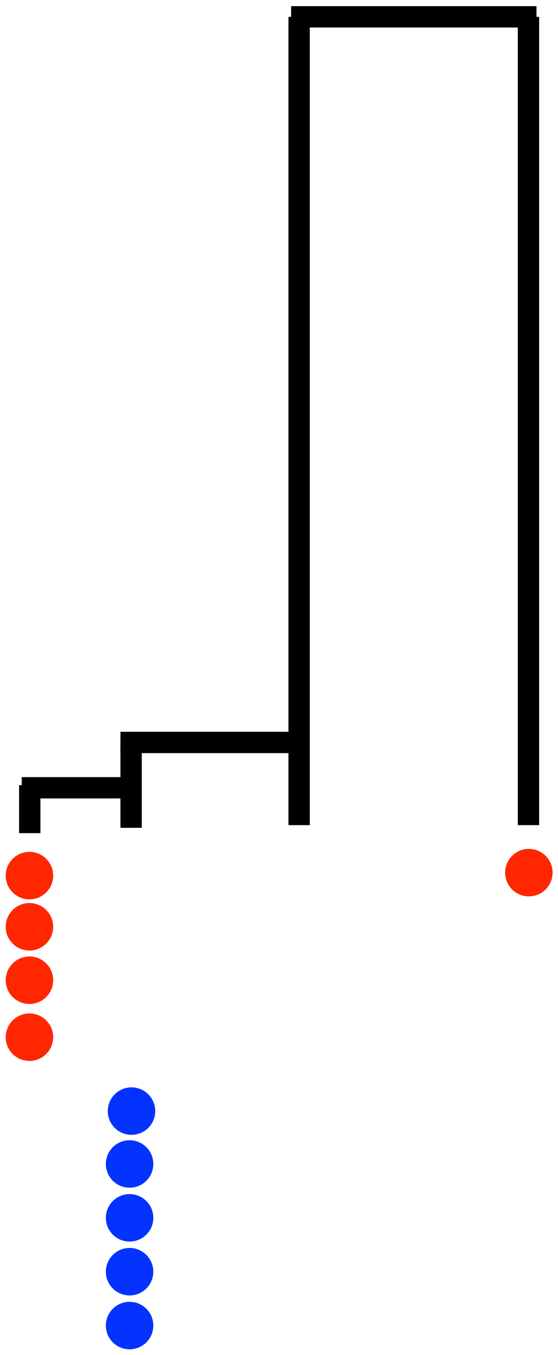
Lozupone et al. (2007)



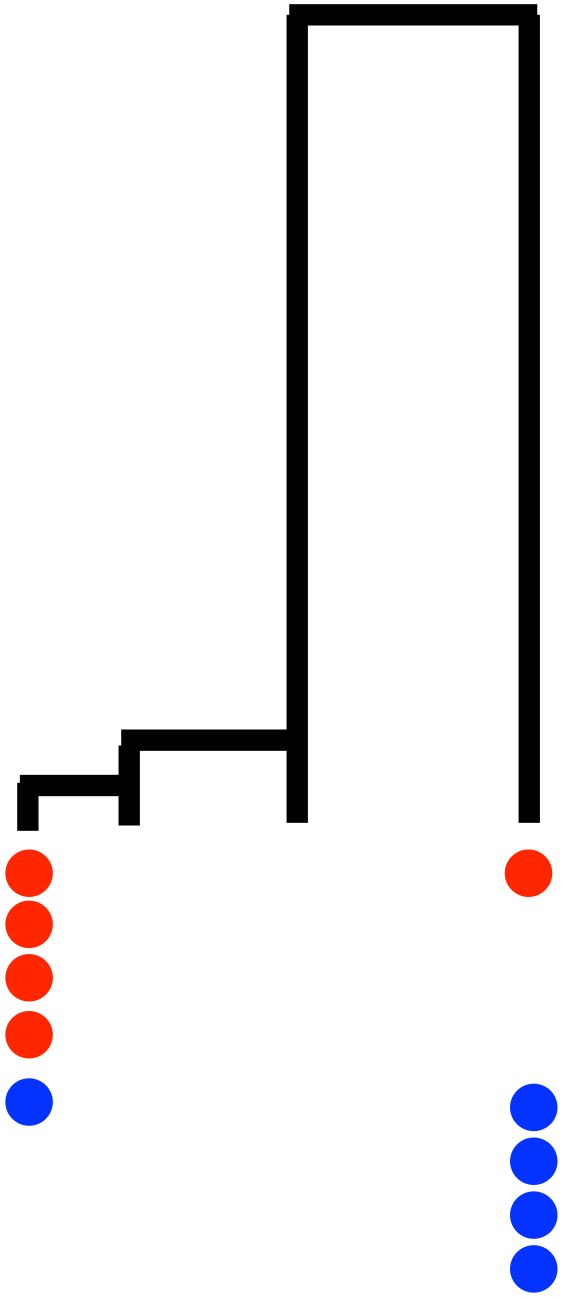
Jaccard:
Bray:
Unifrac:
W-Unifrac:



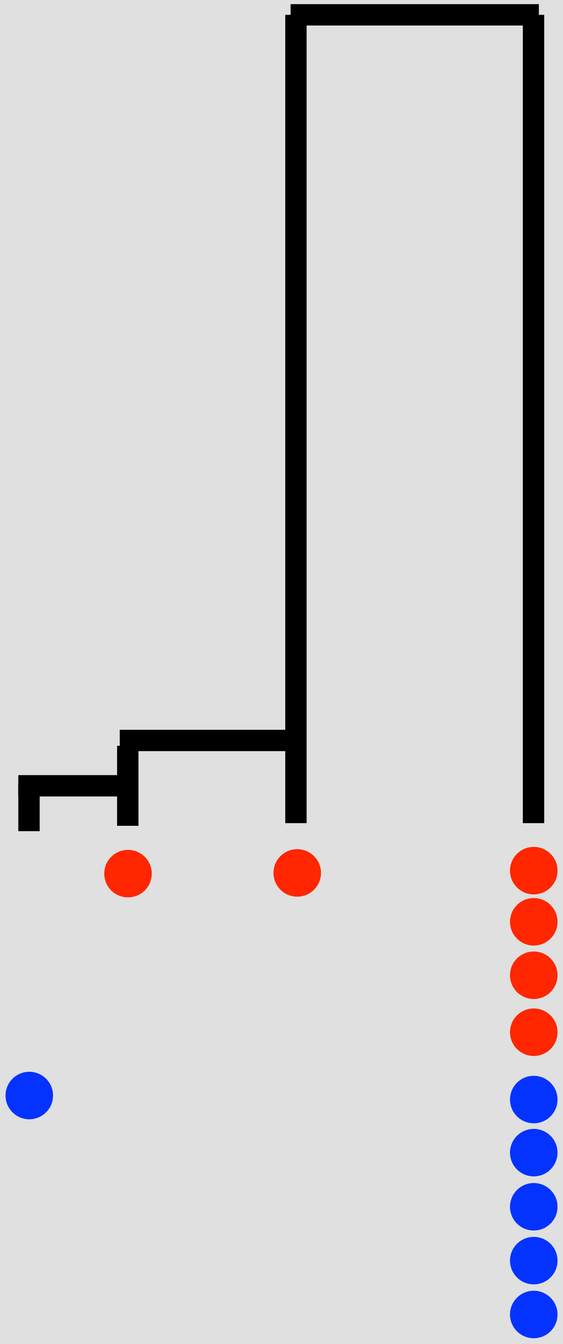
Jaccard:
Bray:
Unifrac:
W-Unifrac:



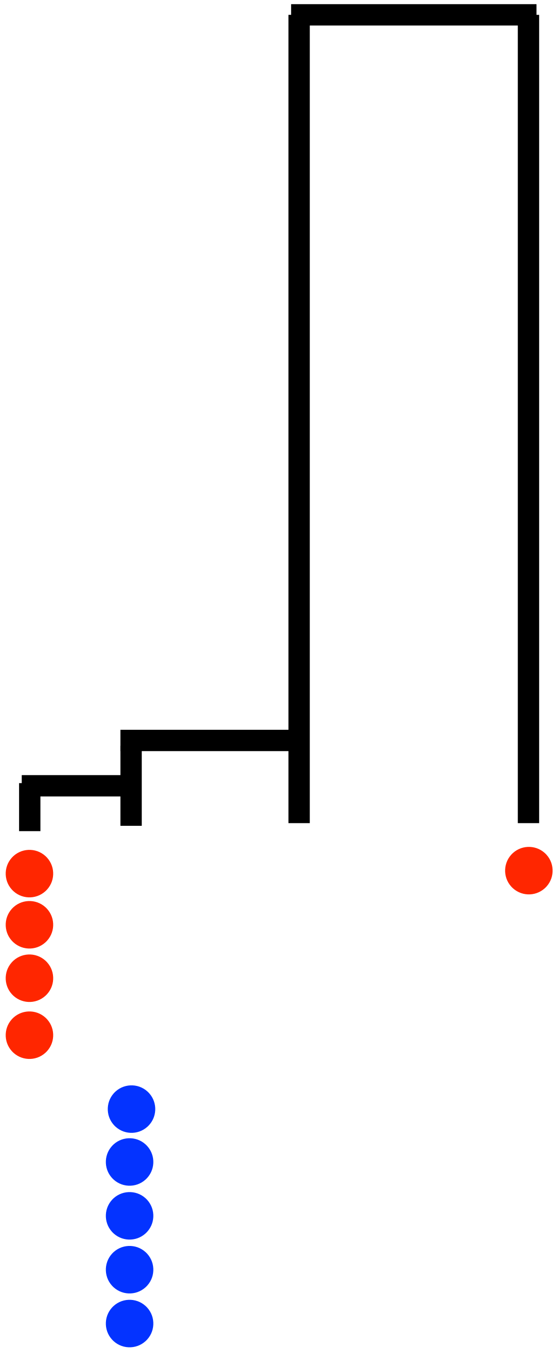
Jaccard:
Bray:
Unifrac:
W-Unifrac:



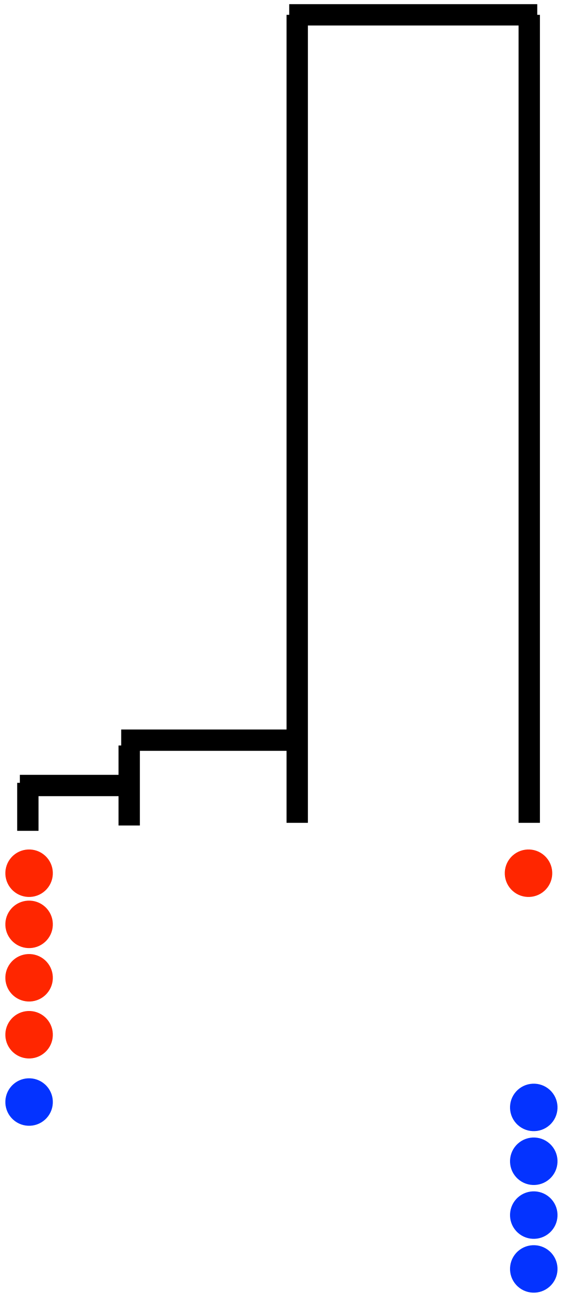
Jaccard: d=0
Bray:
Unifrac:
W-Unifrac:



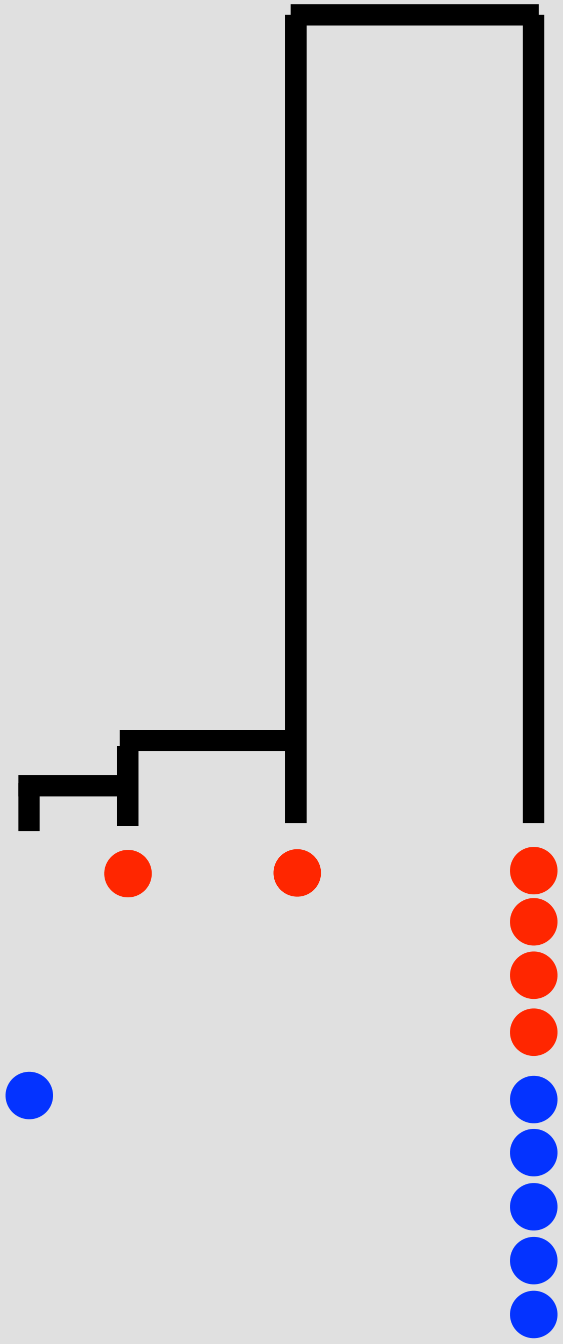
Jaccard: Distant
Bray:
Unifrac:
W-Unifrac:



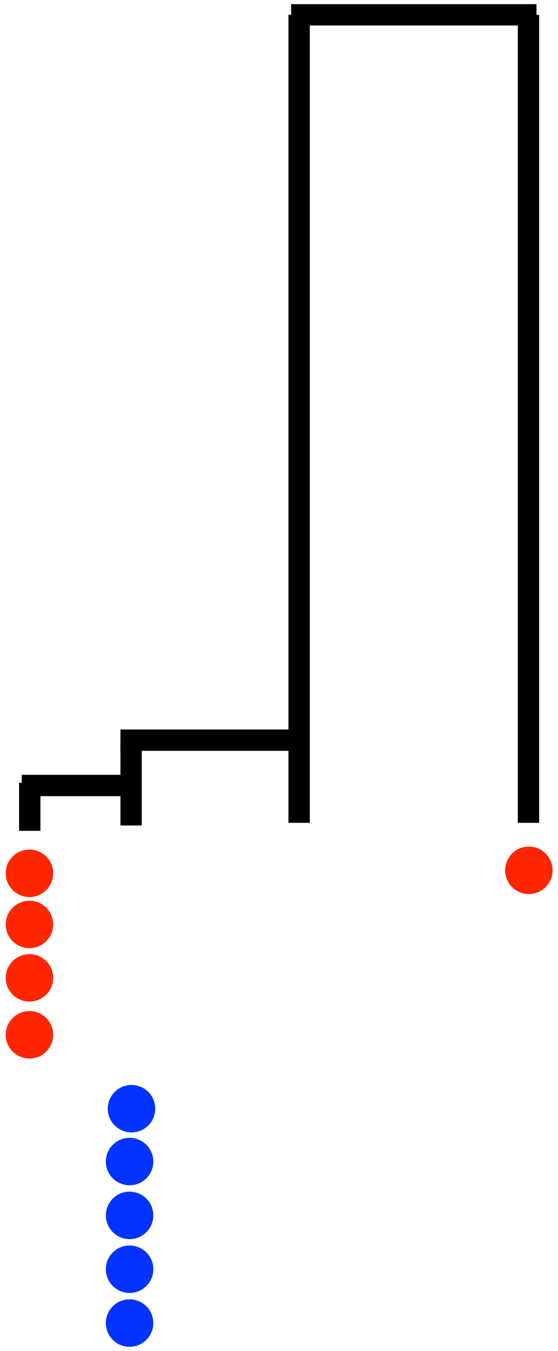
Jaccard: Distant
Bray:
Unifrac:
W-Unifrac:



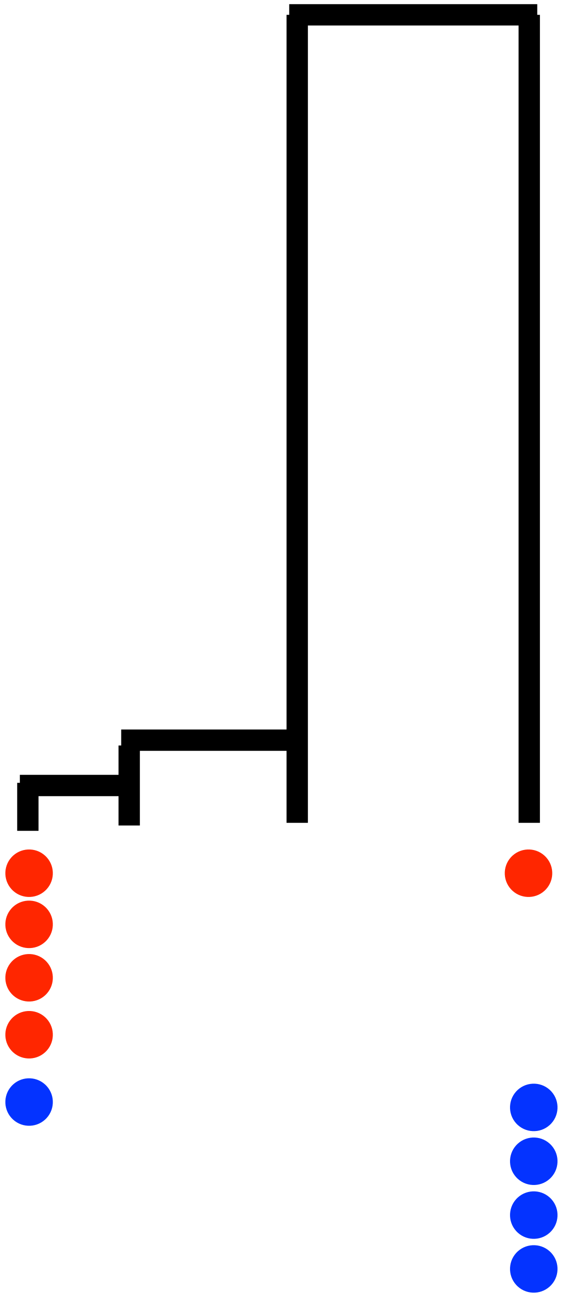
Jaccard: d=0
Bray: Distant
Unifrac:
W-Unifrac:



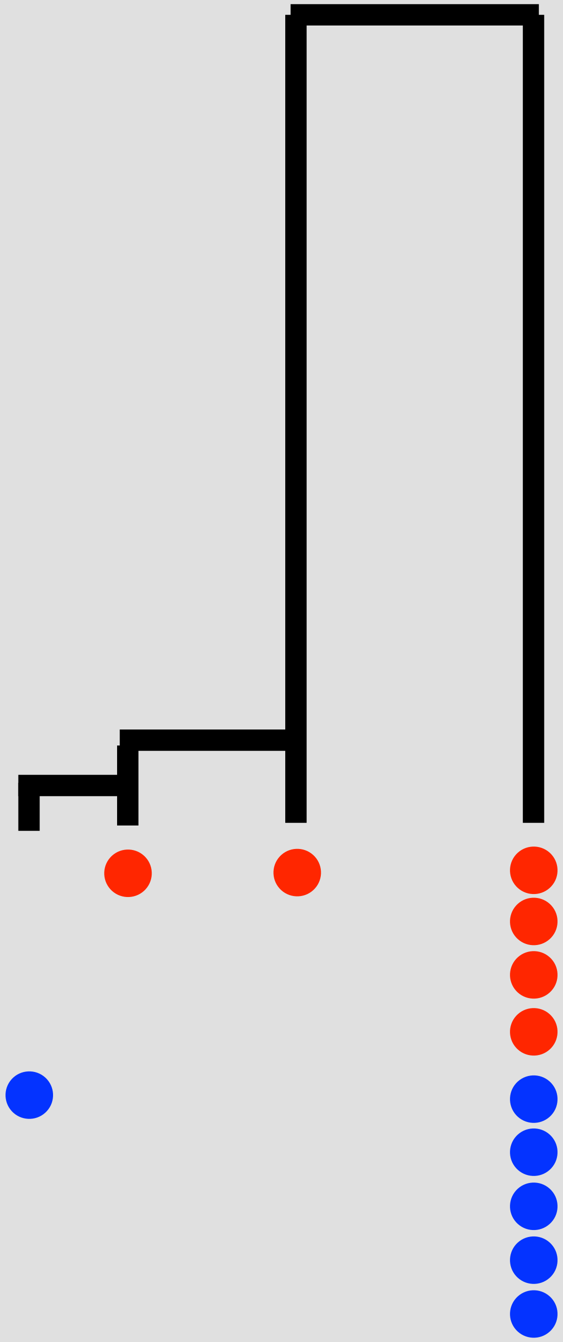
Jaccard: Distant
Bray: Similar
Unifrac:
W-Unifrac:



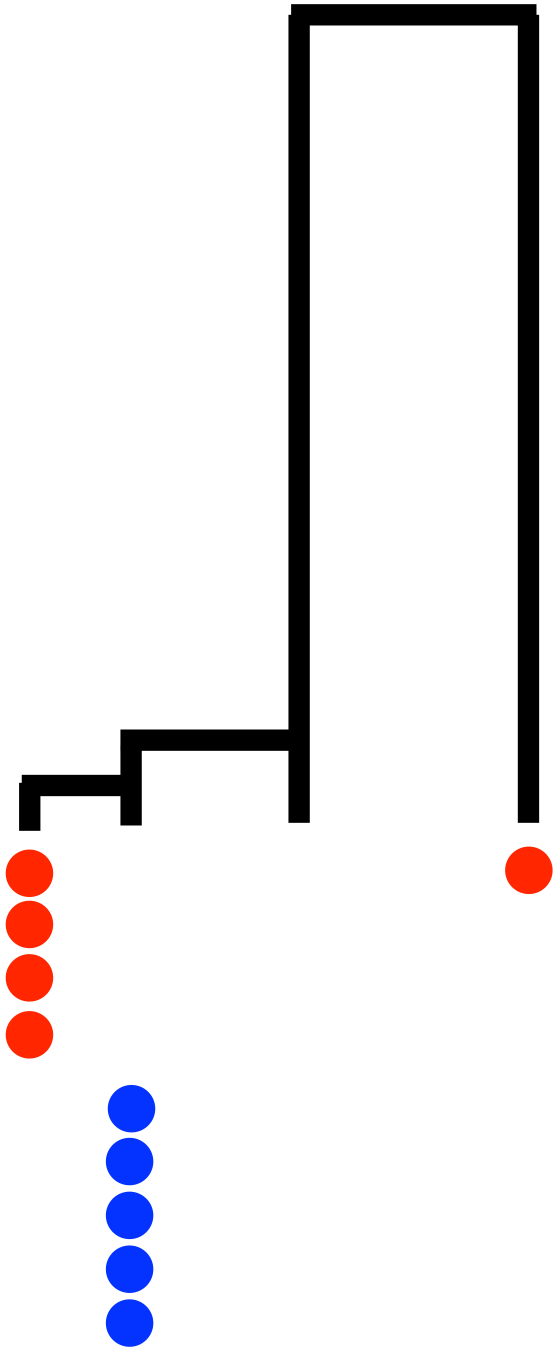
Jaccard: Distant
Bray: Distant
Unifrac:
W-Unifrac:



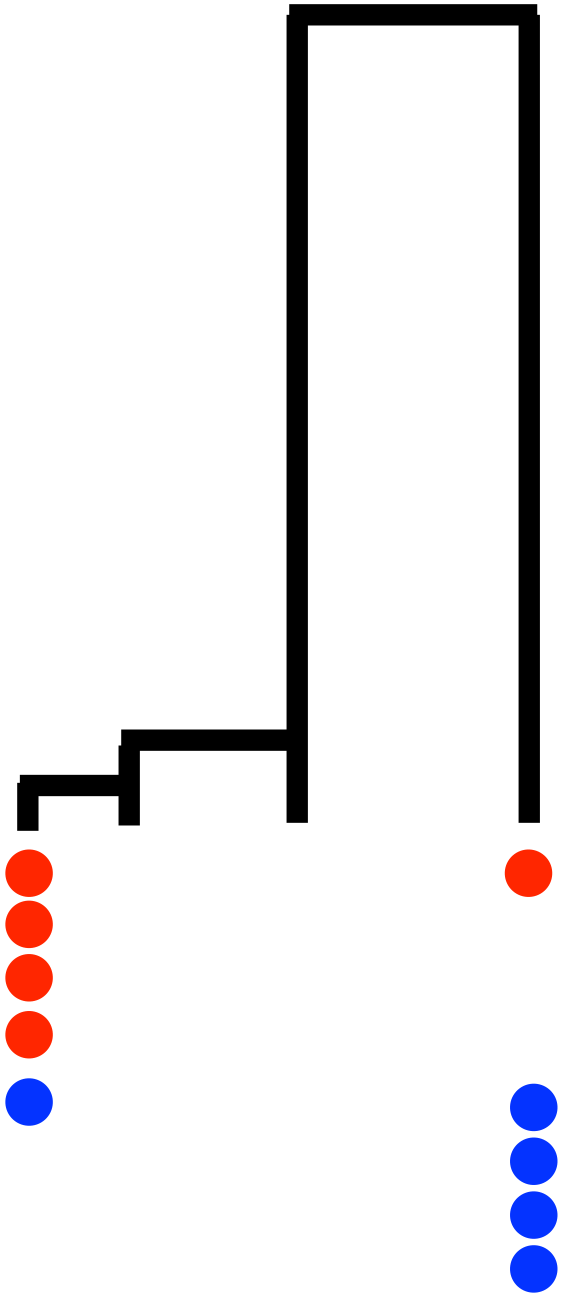
Jaccard: d=0
Bray: Distant
Unifrac: d=0
W-Unifrac:



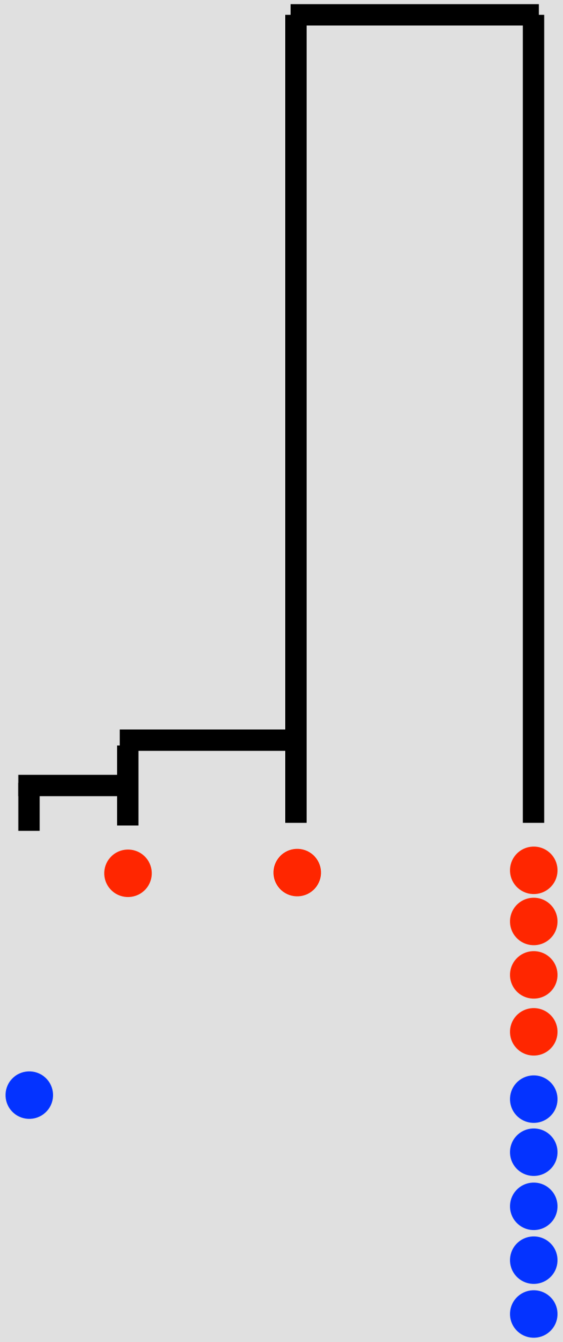
Jaccard: Distant
Bray: Similar
Unifrac: Similar
W-Unifrac:



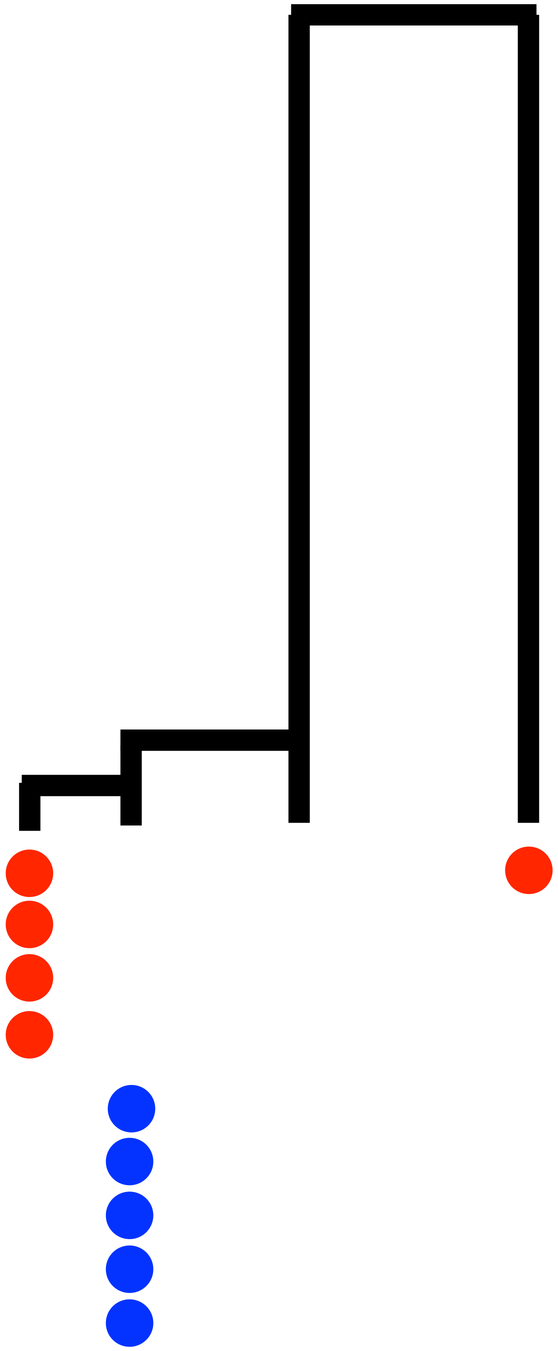
Jaccard: Distant
Bray: Distant
Unifrac: Distant
W-Unifrac:



Jaccard: d=0
Bray: Distant
Unifrac: d=0
W-Unifrac: Distant



Jaccard: Distant
Bray: Similar
Unifrac: Similar
W-Unifrac: Similar



Jaccard: Distant
Bray: Distant
Unifrac: Distant
W-Unifrac: Similar

The Distance Spectrum

	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

phyloseq distances

manhattan
euclidean
canberra
bray
kulczynski
jaccard
gower
altGower
morisita-horn
mountford
raup
binomial
chao
cao
jensen-shannon
unifrac
weighted-unifrac
...

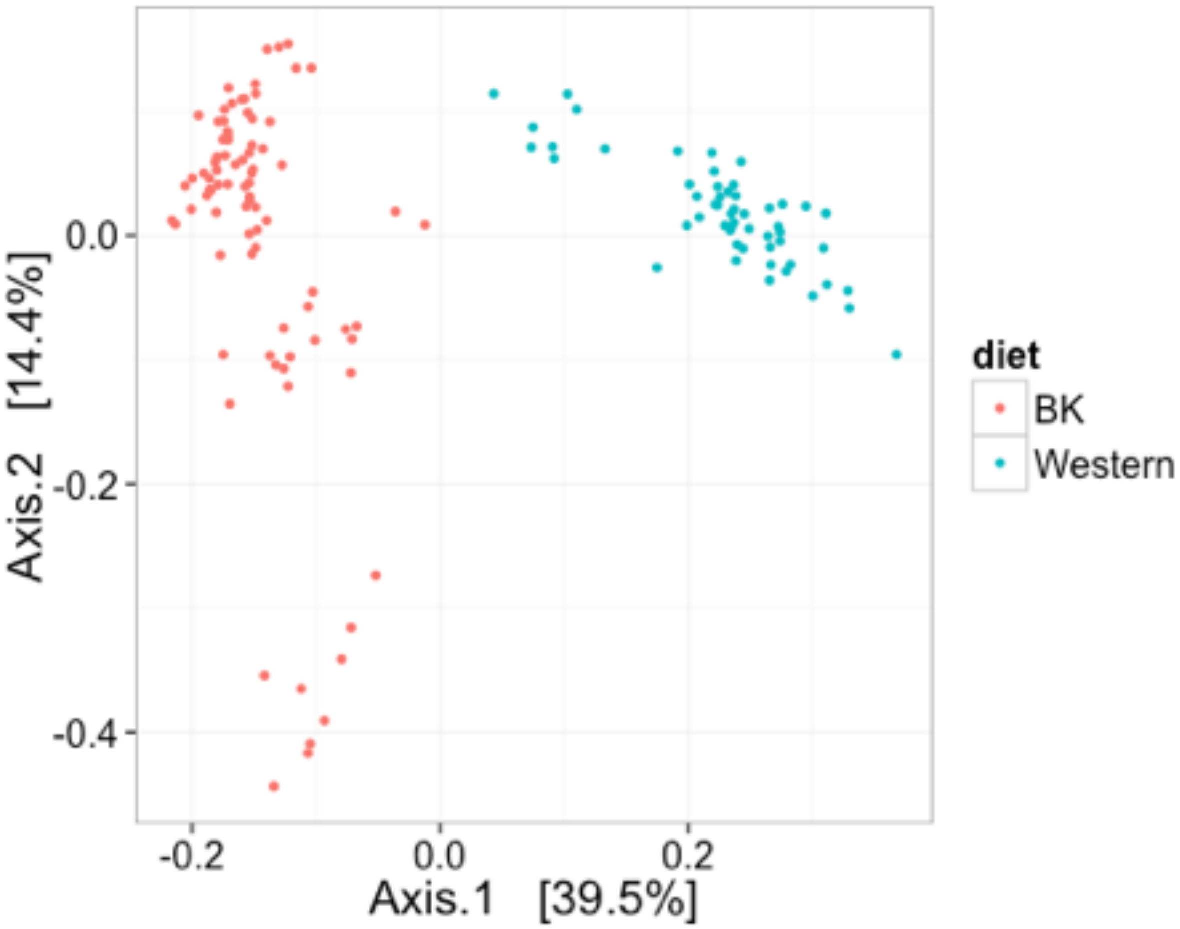
Ordination Methods

Project high-dimensional data onto lower dimensions

P taxa

N samples

0,1,5,1,0,1,2,1,0,0,9,…
7,2,0,0,0,0,0,0,1,0,0,…
0,0,0,0,0,0,8,0,0,0,1,…
0,0,0,1,0,1,2,0,0,0,5,…
0,1,0,2,0,0,0,1,0,0,4,…
0,0,0,1,9,1,2,5,2,0,1,…
0,0,0,0,0,1,2,1,8,0,0,…
0,0,0,0,9,4,0,0,0,0,1,…
.
.

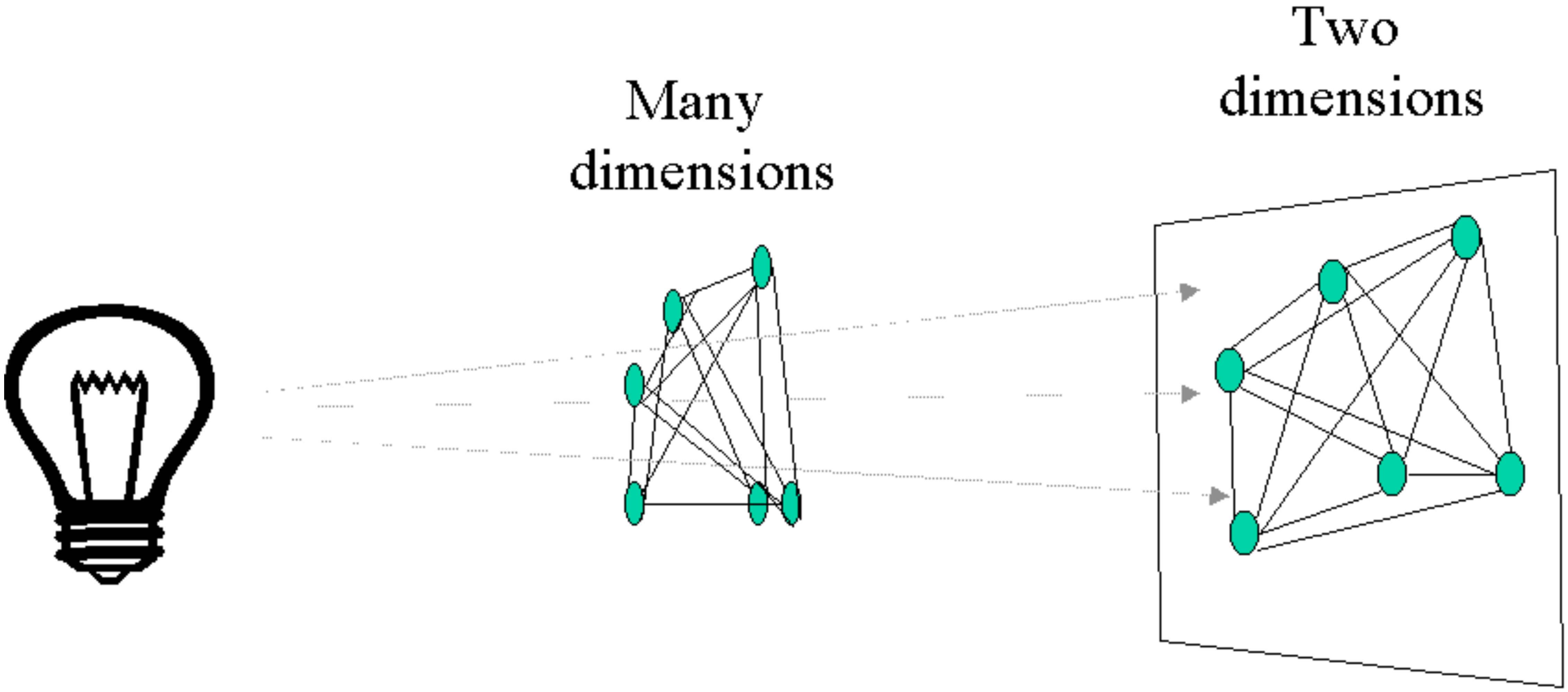


P-dimensions

2-dimensions

Multi-dimensional Scaling

Why MDS? It works with any distance!



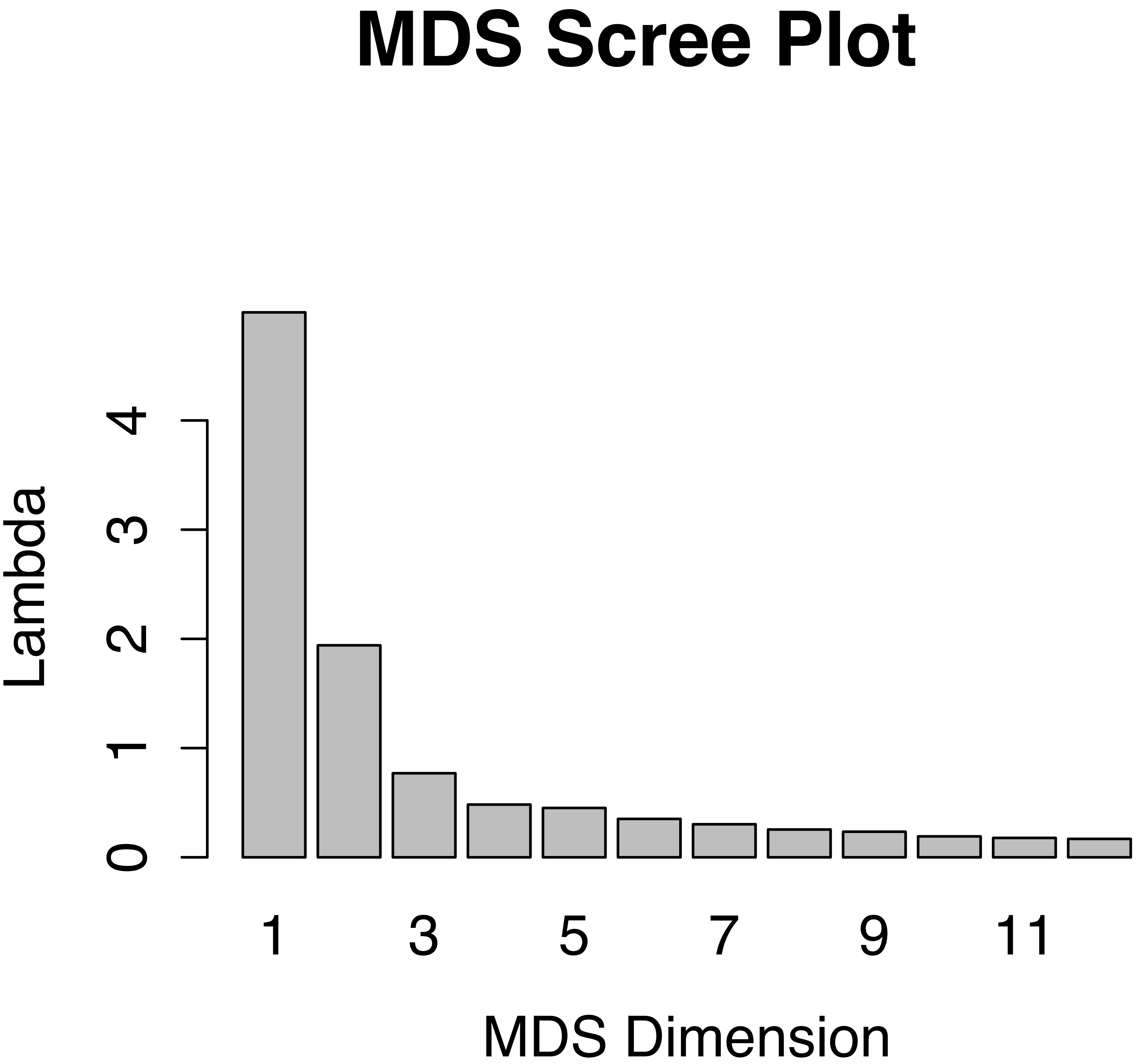
Input distance matrix can by Bray-Curtis, Unifrac, …

MDS Details

Given distances between each observation (sample), MDS finds the closest approximation of that in lower dimensional Euclidean space.

- Algorithm starts from **D** inter-point distances:
  - Center the rows and columns of the distance matrix: $\mathbf{S} = -1/2\ \mathbf{H}\ \mathbf{D}^{(2)}\ \mathbf{H}$
  - Compute SVD by diagonalizing S: $\mathbf{S} = \mathbf{U}\ \boldsymbol{\Lambda}\ \mathbf{U}^{T}$
  - Extract Euclidean representations: $\mathbf{X} = \mathbf{U}\ \boldsymbol{\Lambda}^{1/2}$
- The relative values of diagonal elements of Λ gives the proportion of variability explained by each of the axes.
- The valued of Λ should always be looked at in deciding how many dimensions to retain

NMDS is similar, but minimizes a different function (difference in distance ranks)


MDS Scree Plot
Lambda
0
1
2
3
4
1
3
5
7
9
11
MDS Dimension

Exploratory Analysis

- Looking for patterns (the “I-test”)
- Use multiple distances
- phyloseq makes this easy!