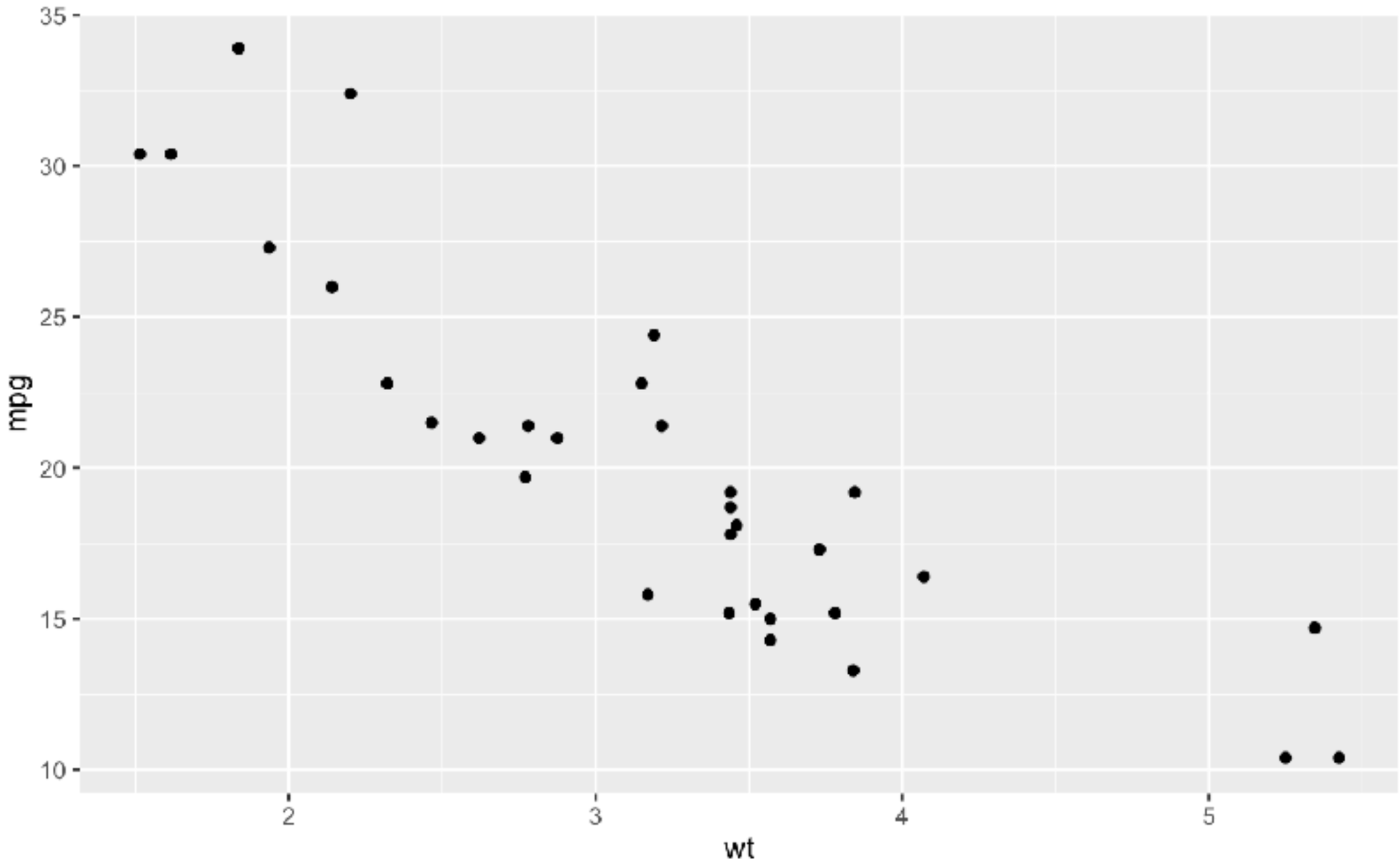
β-Diversity: Basics

Useful Reference

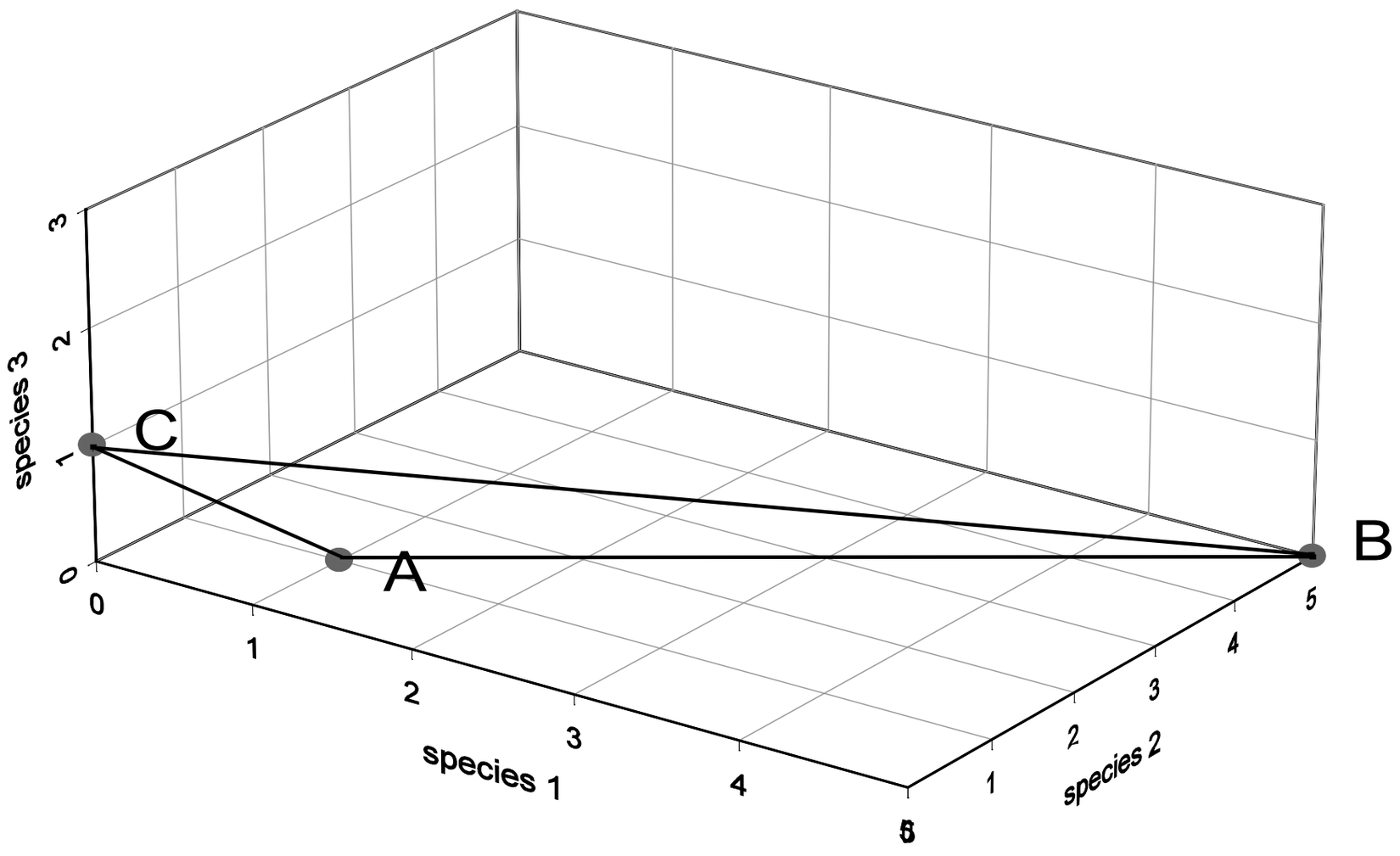
- GUide to STatistical Analysis in Microbial Ecology (GUSTA ME)!: https://mb3is.megx.net/gustame

Visualizing Samples



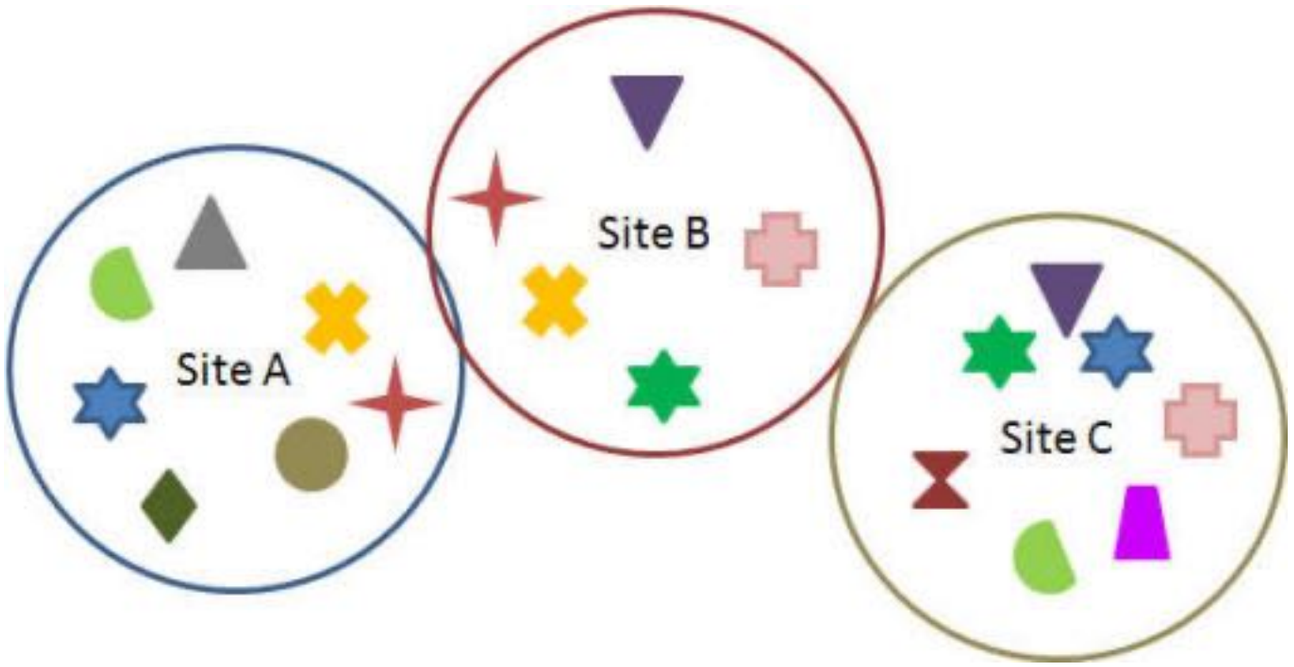
Beta Diversity:
How different are sites?

		Species		
		1	2	3
Samples	A	1	1	0
	B	5	5	0
	C	0	0	1



Tree diversity analysis, Kindt and Coe <http://www.worldagroforestry.org/downloads/Publications/PDFS/b13695.pdf>

More than 3 Taxa?

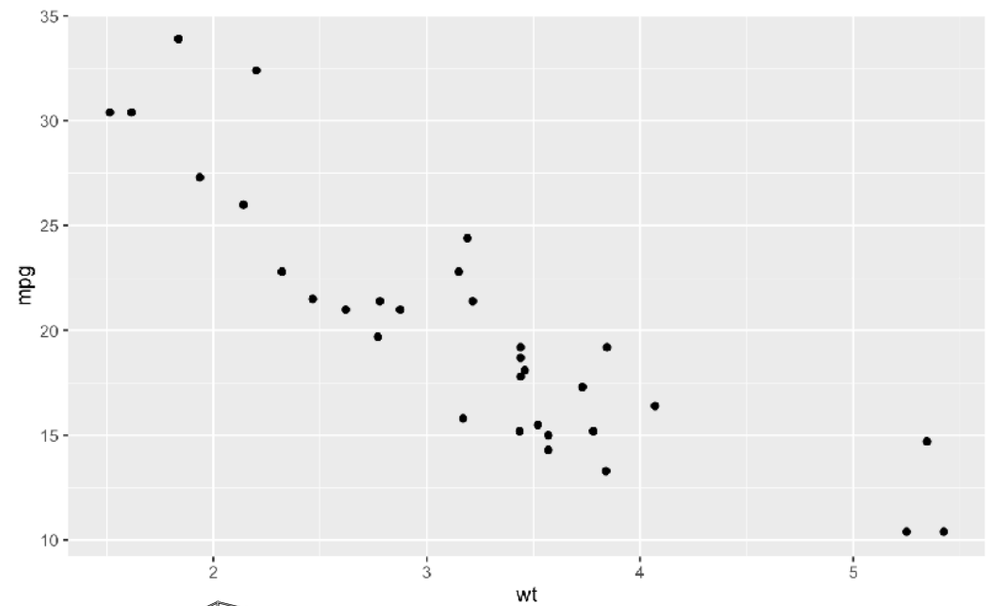


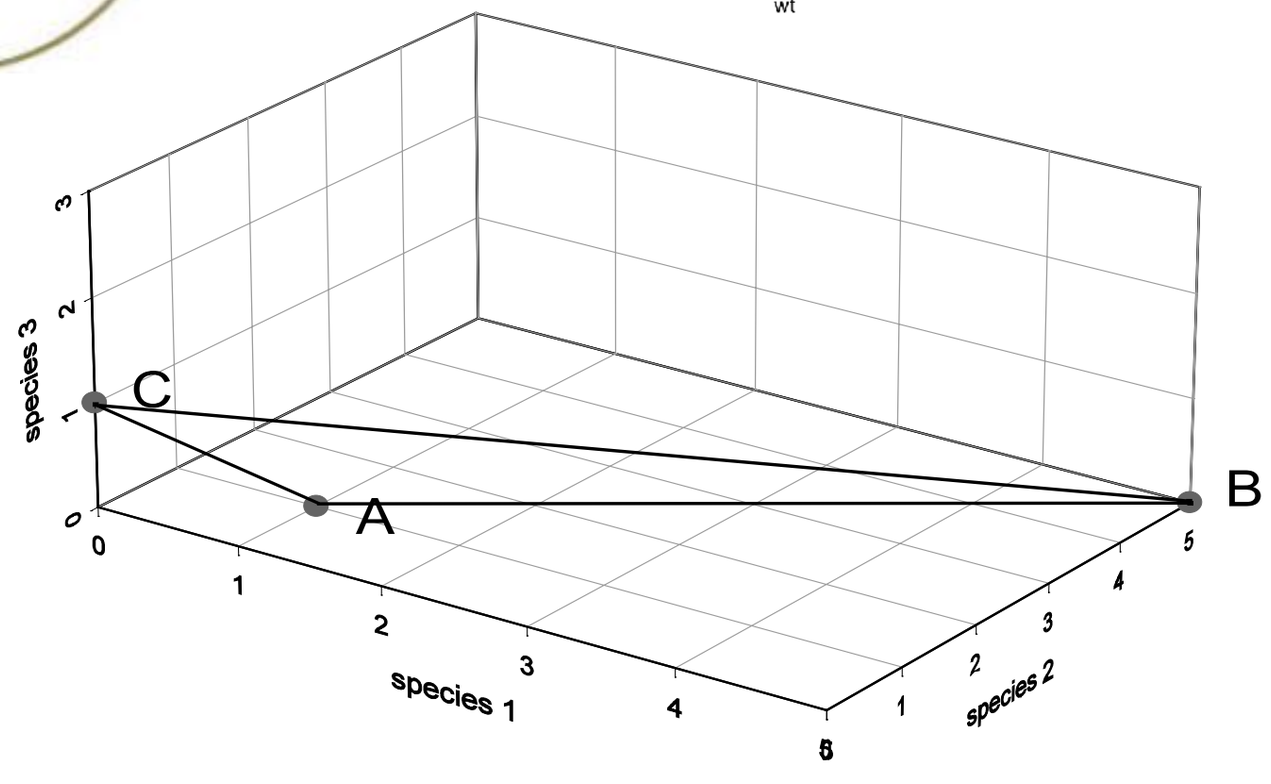
Site

A

B

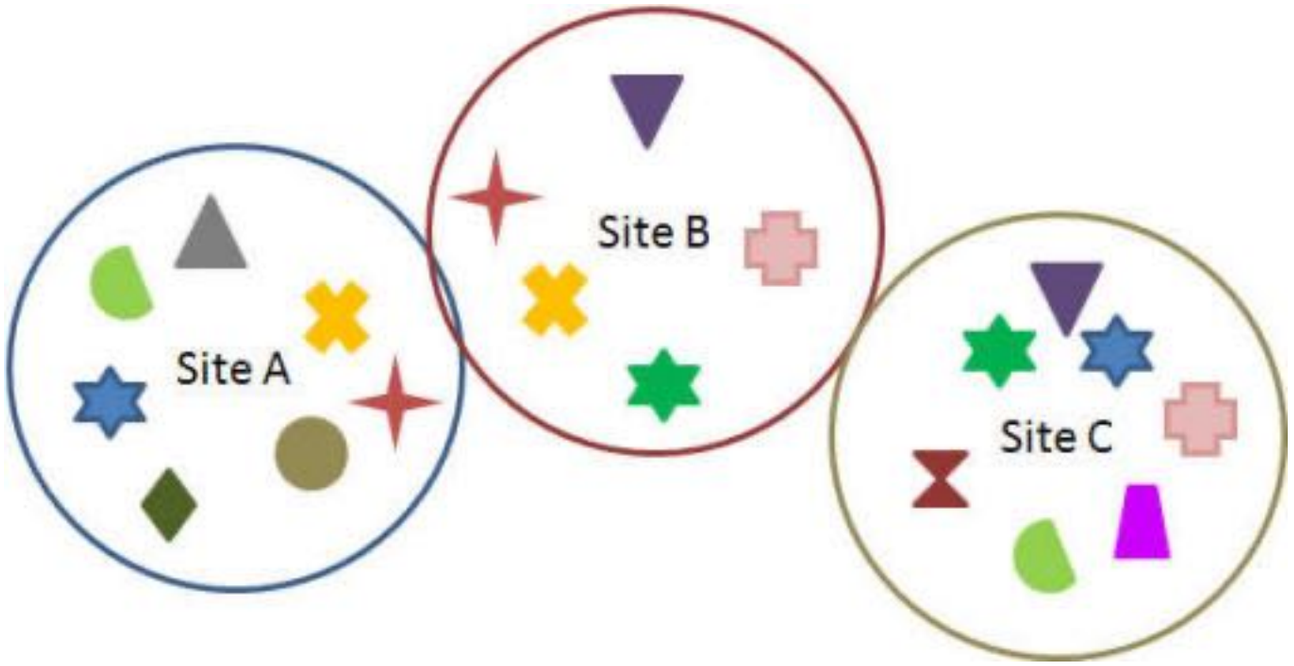
C





- http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209(Composition&Diversity)/9_2_Biodiversity.htm
- Tree diversity analysis, Kindt and Coe <http://www.worldagroforestry.org/downloads/Publications/PDFS/b13695.pdf>

How different are samples?



Site

A

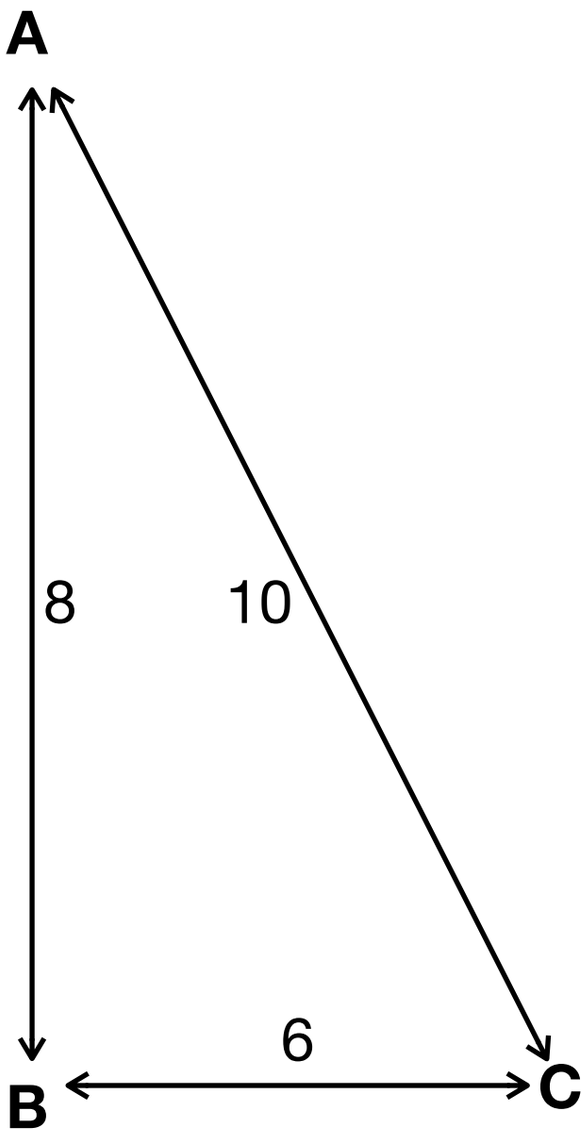
B

C

	Species Differences
A vs B	8
B vs C	6
A vs C	10

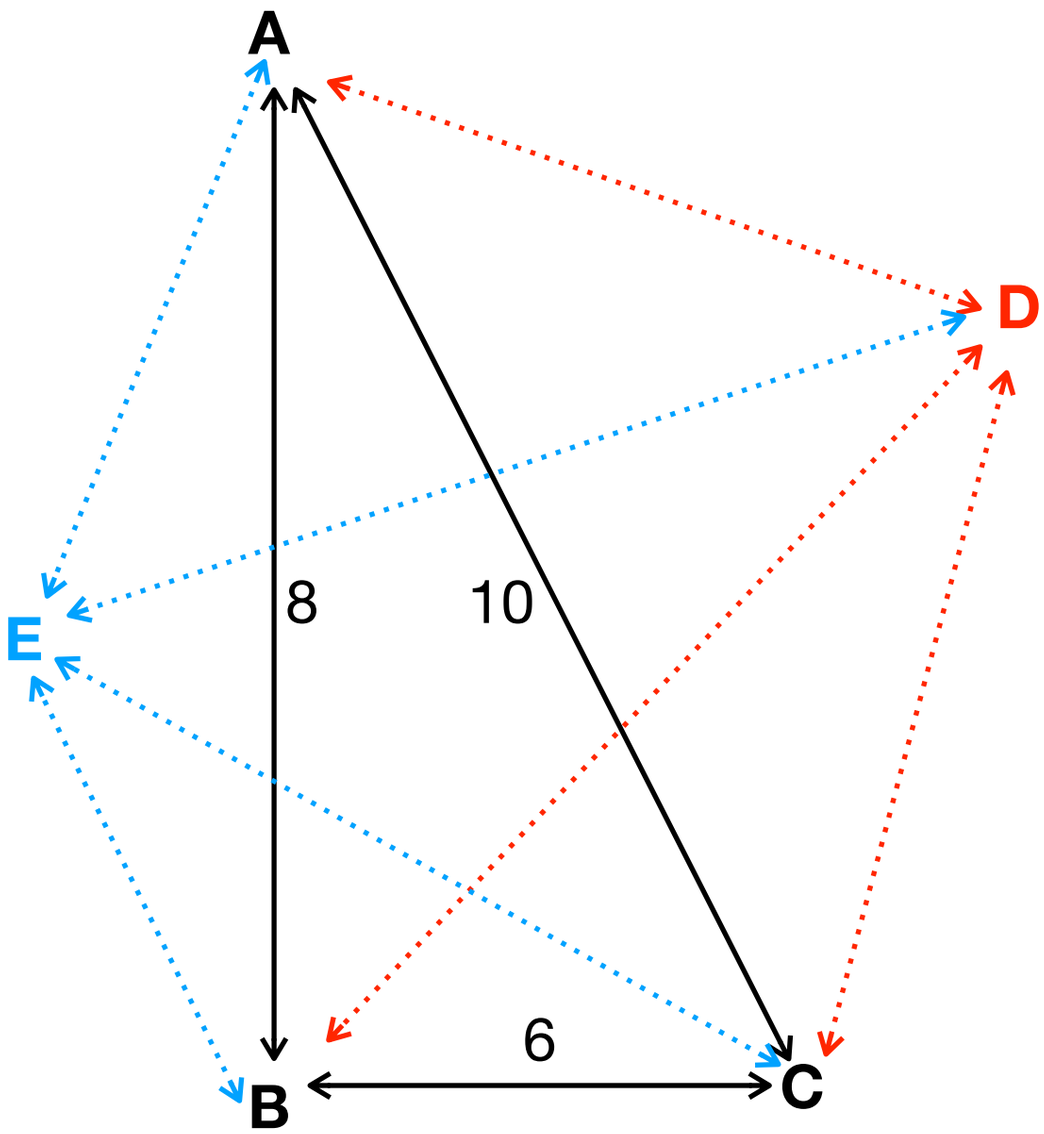
http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209(Composition&Diversity)/9_2_Biodiversity.htm

How different are samples?



	Species Differences
A vs B	8
B vs C	6
A vs C	10

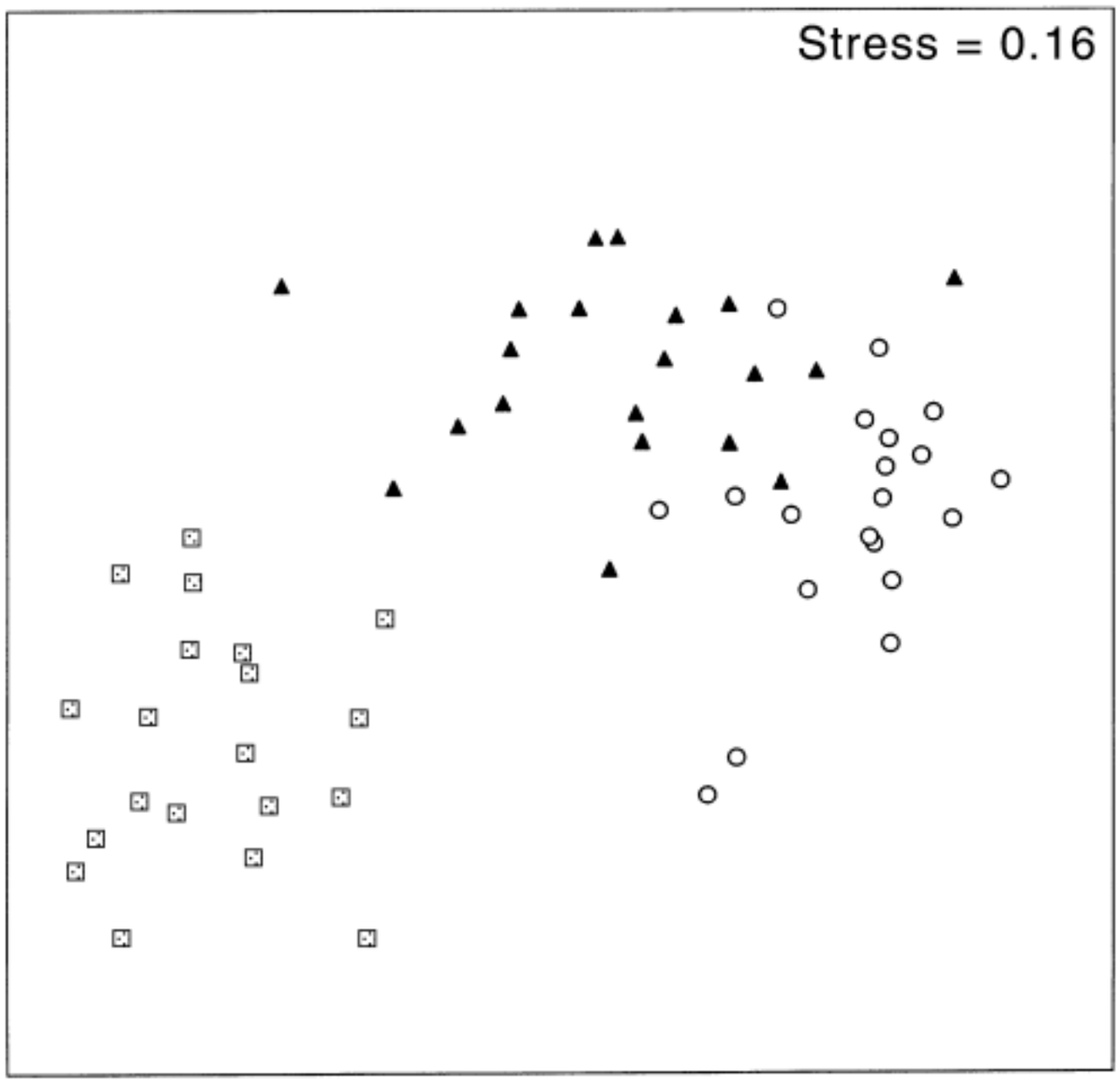
More Than 3 samples?



	Species Differences
A vs B	8
B vs C	6
A vs C	10

β-Diversity: Ordination

Dimensionality reduction



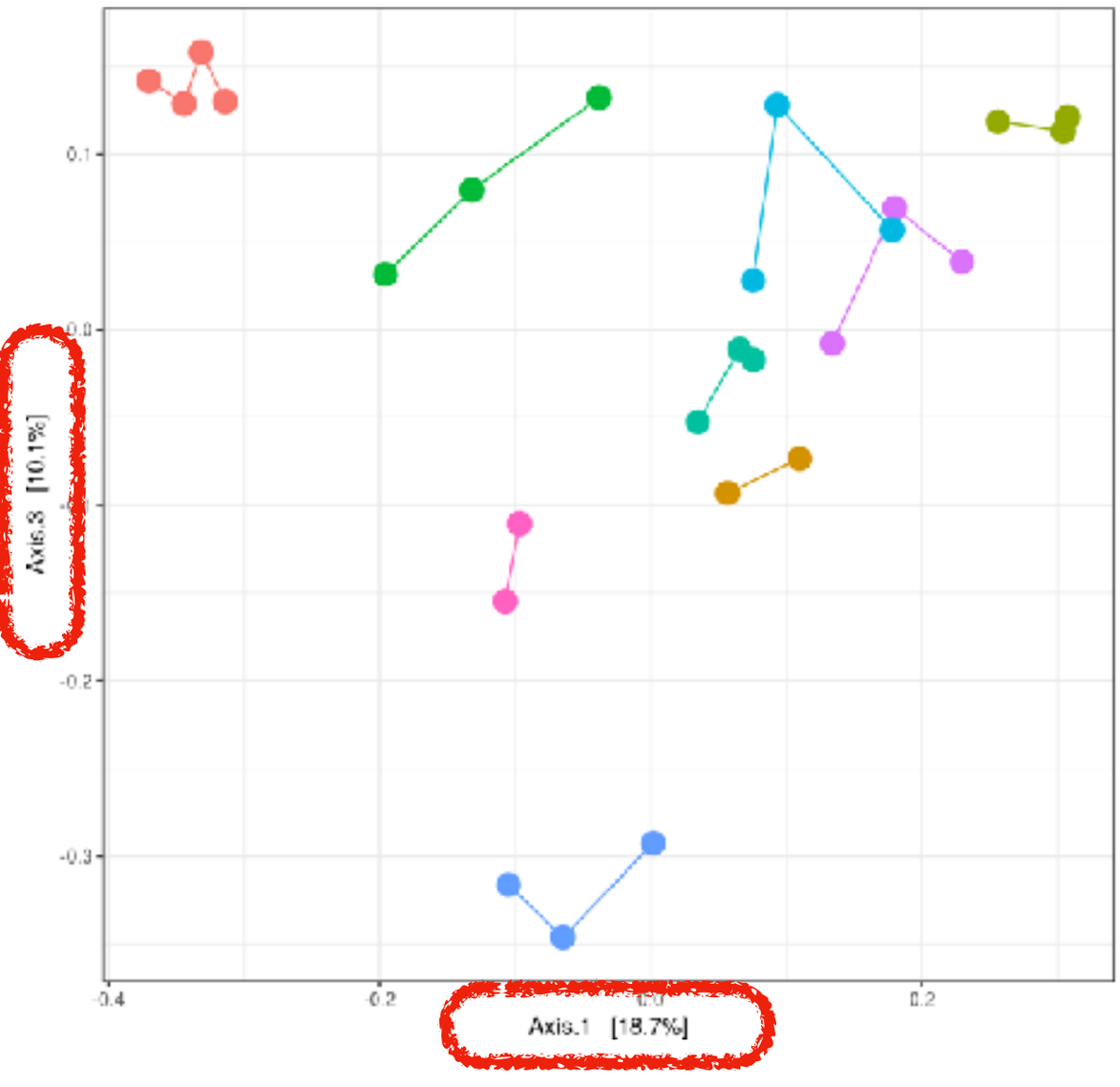
https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

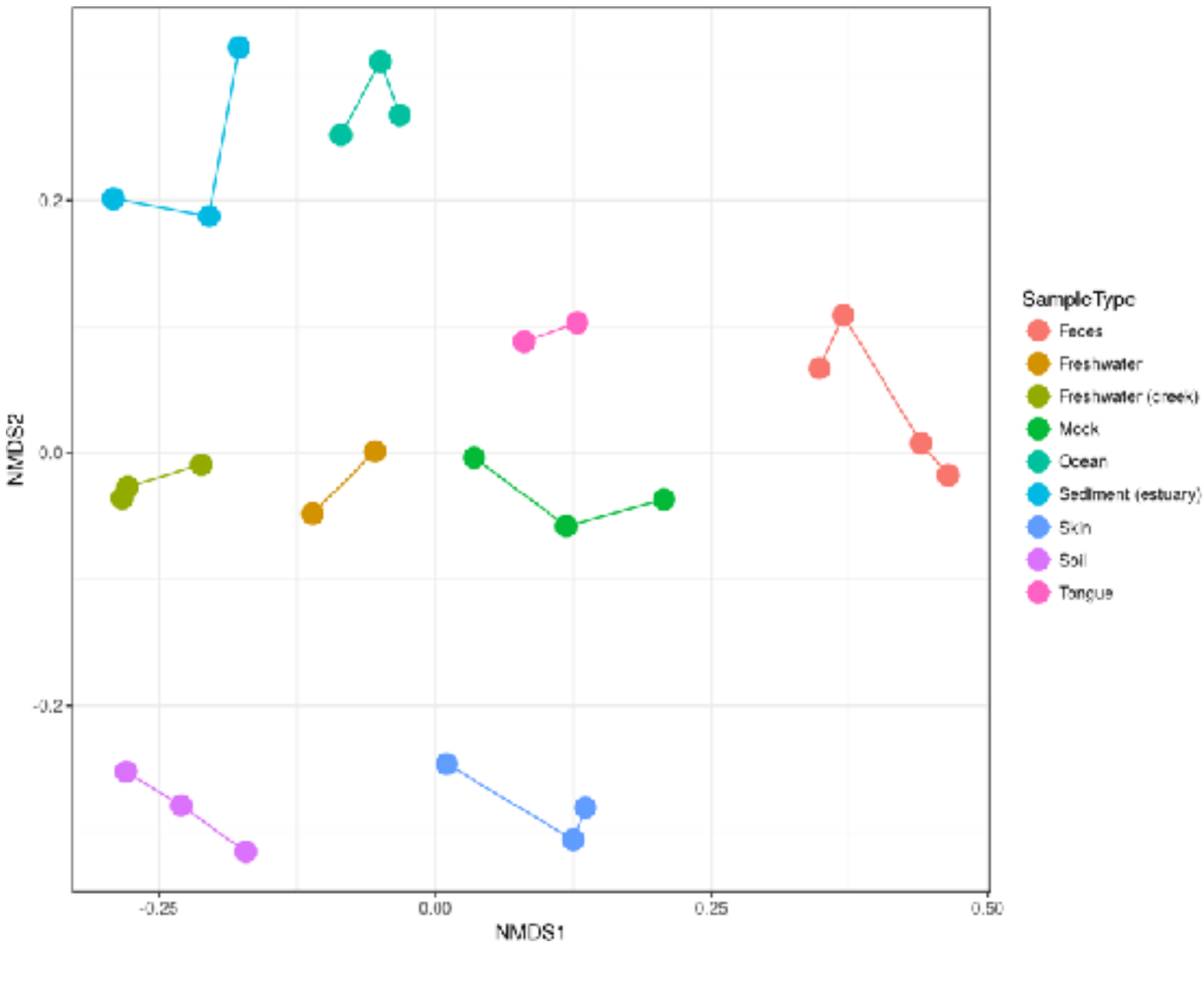
Dimensionality reduction

- **NMDS (Nonmetric Multidimensional Scaling)**: attempts to maintain pairwise dissimilarity

- **PCoA (Principal coordinates analysis):** Determines axes along which variance is maximized

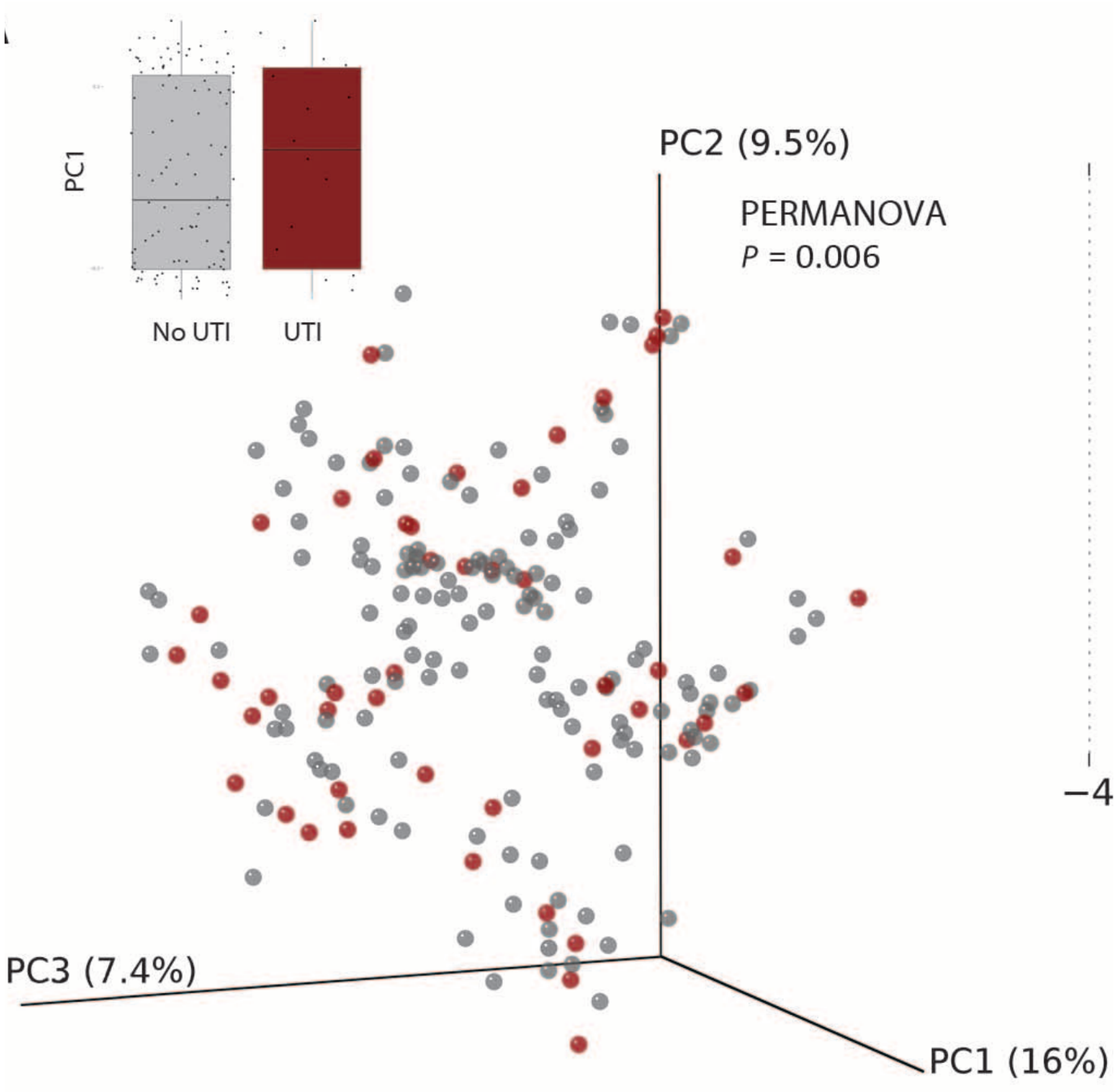
PCoA and NMDS





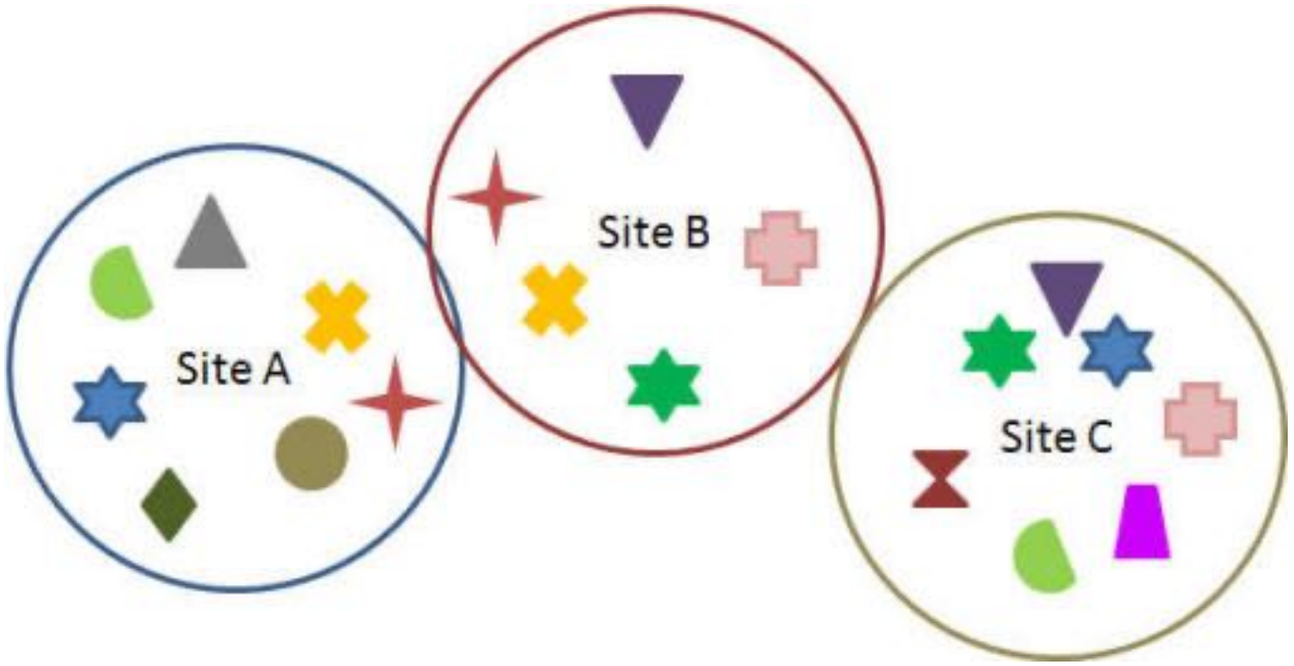
https://www.bioconductor.org/packages/devel/bioc/vignettes/phyloseq/inst/doc/phyloseq-analysis.html

Figures Josh Hates: 3D PCoA Plots



https://doi.org/10.1126/scitranslmed.3008599

Beta Diversity Metrics



Site

A

B

C

	Species Differences
A vs B	8
B vs C	6
A vs C	10

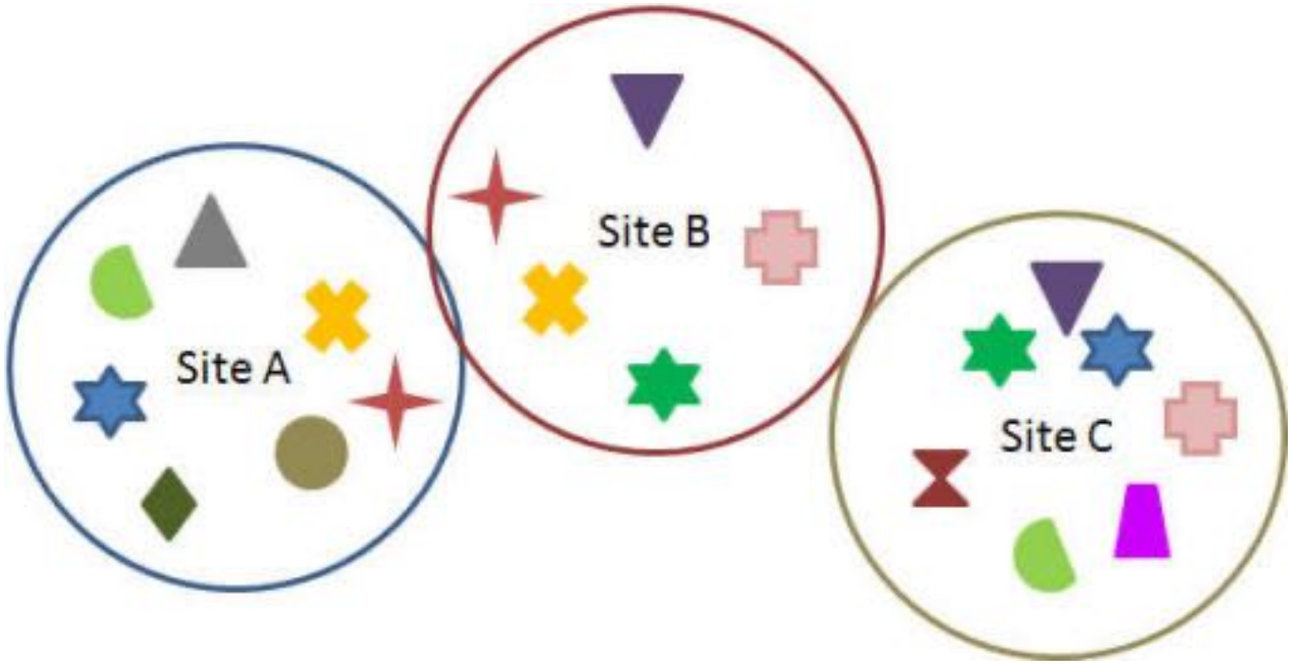
http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209(Composition&Diversity)/9_2_Biodiversity.htm

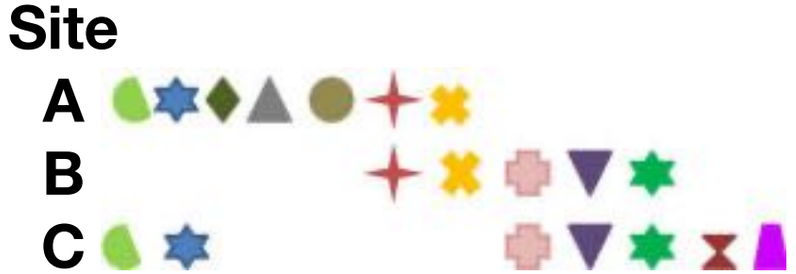
Beta Diversity Metrics

- Number of Taxa
- Bray-Curtis Dissimilarity
- UniFrac
- etc

Number of Taxa

- How different are samples?





	Species Differences
A vs B	8
B vs C	4
A vs C	10

http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209(Composition&Diversity)/9_2_Biodiversity.htm

Bray-Curtis Dissimilarity

$$D_{BC} = 1 - 2\frac{\sum_{i=1}^{S} min\left(a_i,\, b_i\right)}{\sum_{i=1}^{S} a_i + b_i}$$

	Species		
	1	2	3
A	1	1	0
B	5	5	0
C	0	0	1

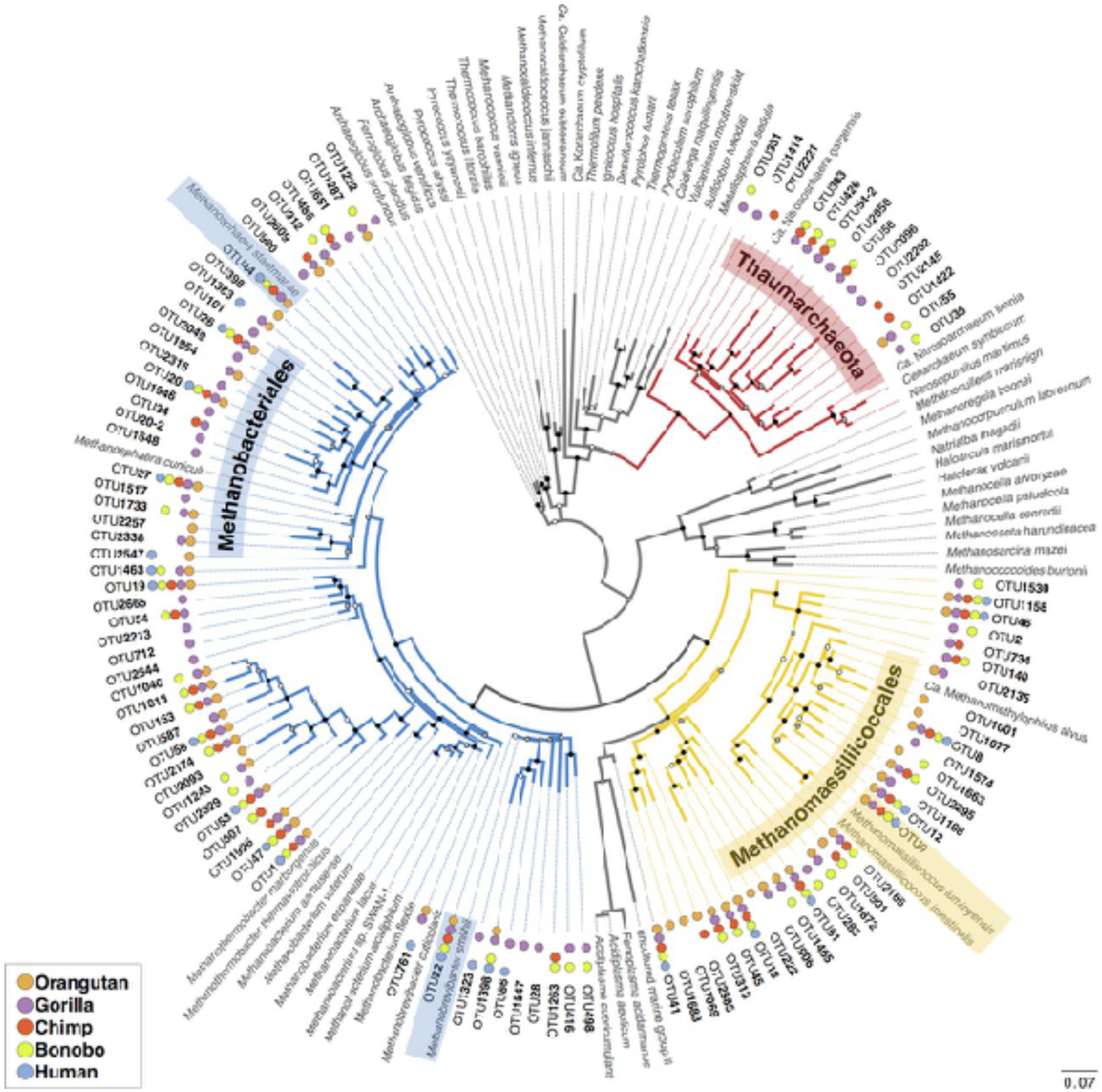
Where x_i is the abundance of species *i* in sample *X*

Bray-Curtis Dissimilarity

```
> y
  housing foodstuffs alcohol other services
1     640         328     147   169      196
2    1800         484     515  2291      912
3     640         328     147   169      196
4    6400        3280    1470  1690     1960
```

```
> vegdist(y, method="bray")
             1         2         3
2 0.6043839
3 0.0000000 0.6043839
4 0.8181818 0.4807230 0.8181818
```

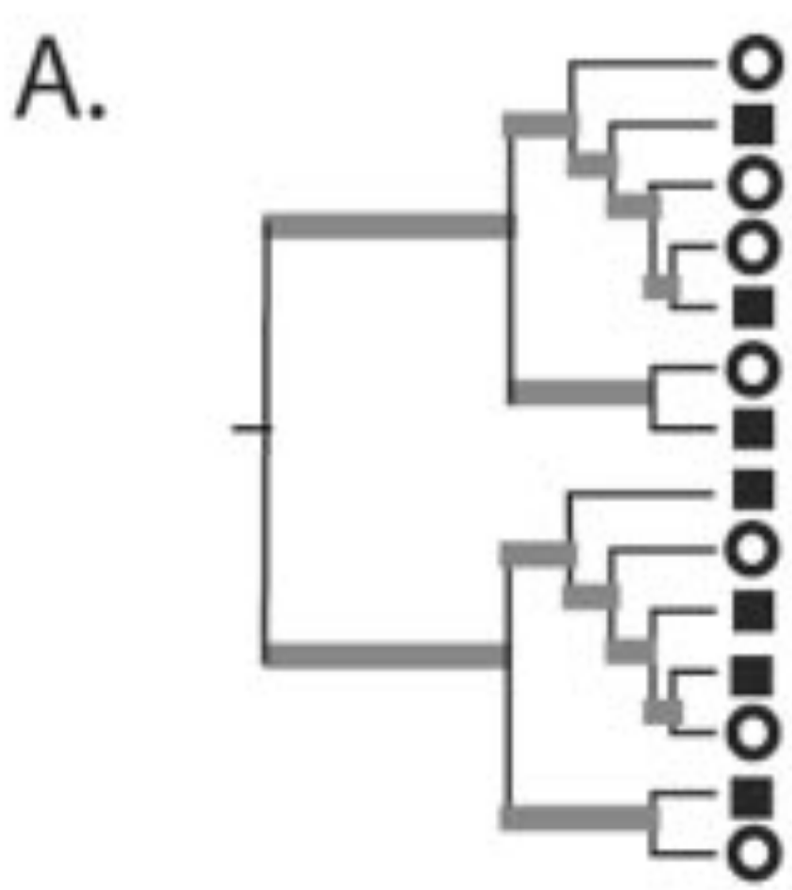
Phylogenetic Trees

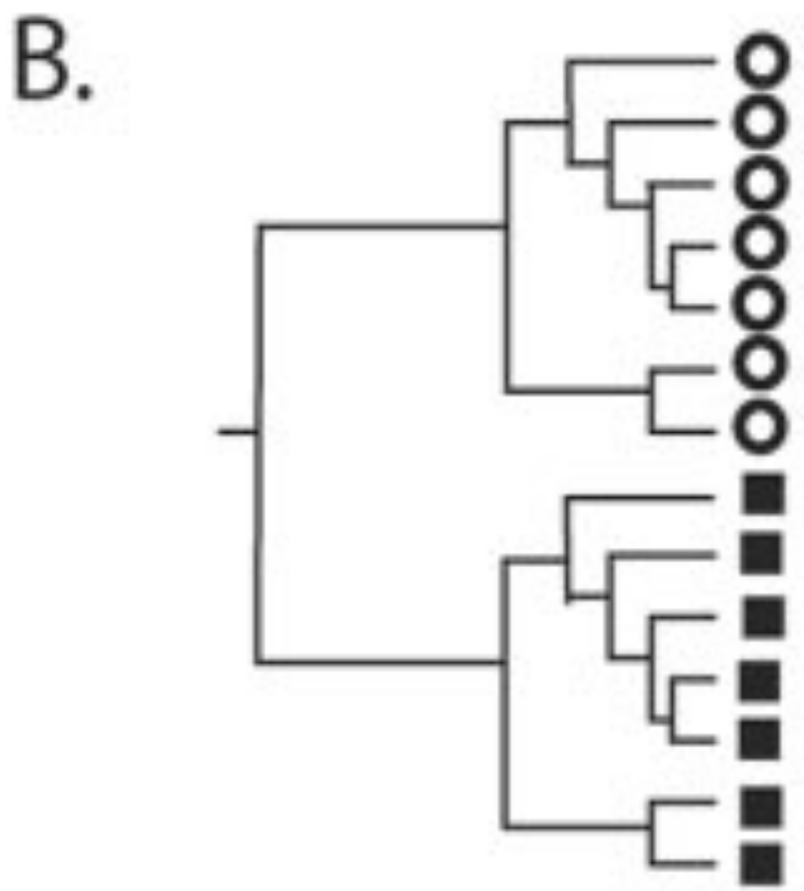


https://doi.org/10.1128/mSphere.00026-17

UniFrac

- The phylogenetic distance between sets of taxa in a phylogenetic tree as the fraction of the branch length of the tree that leads to descendants from either one environment or the other, but not both.





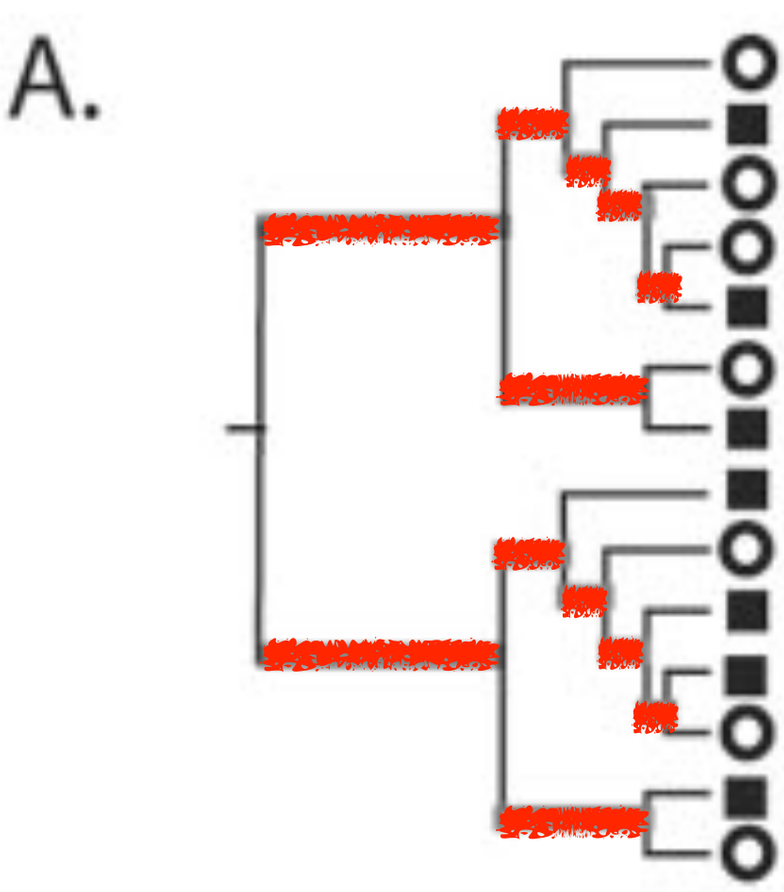
https://doi.org/10.1128/AEM.71.12.8228-8235.2005

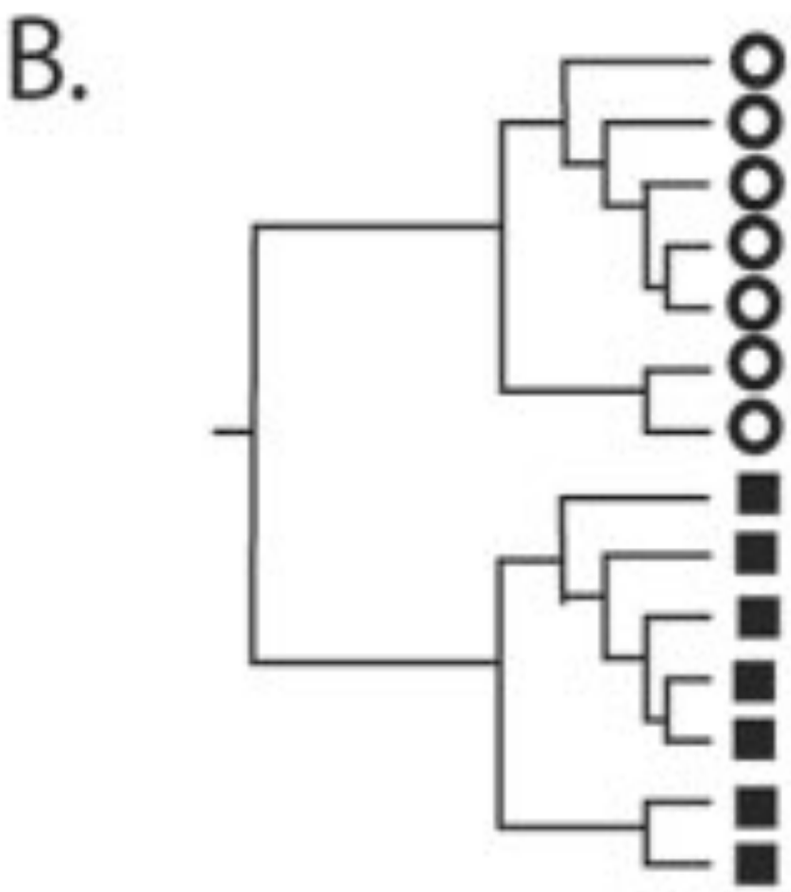
UniFrac

1. Label branches leading to taxa from both samples "shared"

2. Label branches leading to taxa which appears only in one sample "unshared".

3. Unifrac is the fraction of total branch length which is unshared. Alternatively, (the sum of "unshared" branch lengths)/(the sum of all tree branch lengths





https://doi.org/10.1128/AEM.71.12.8228-8235.2005

UniFrac

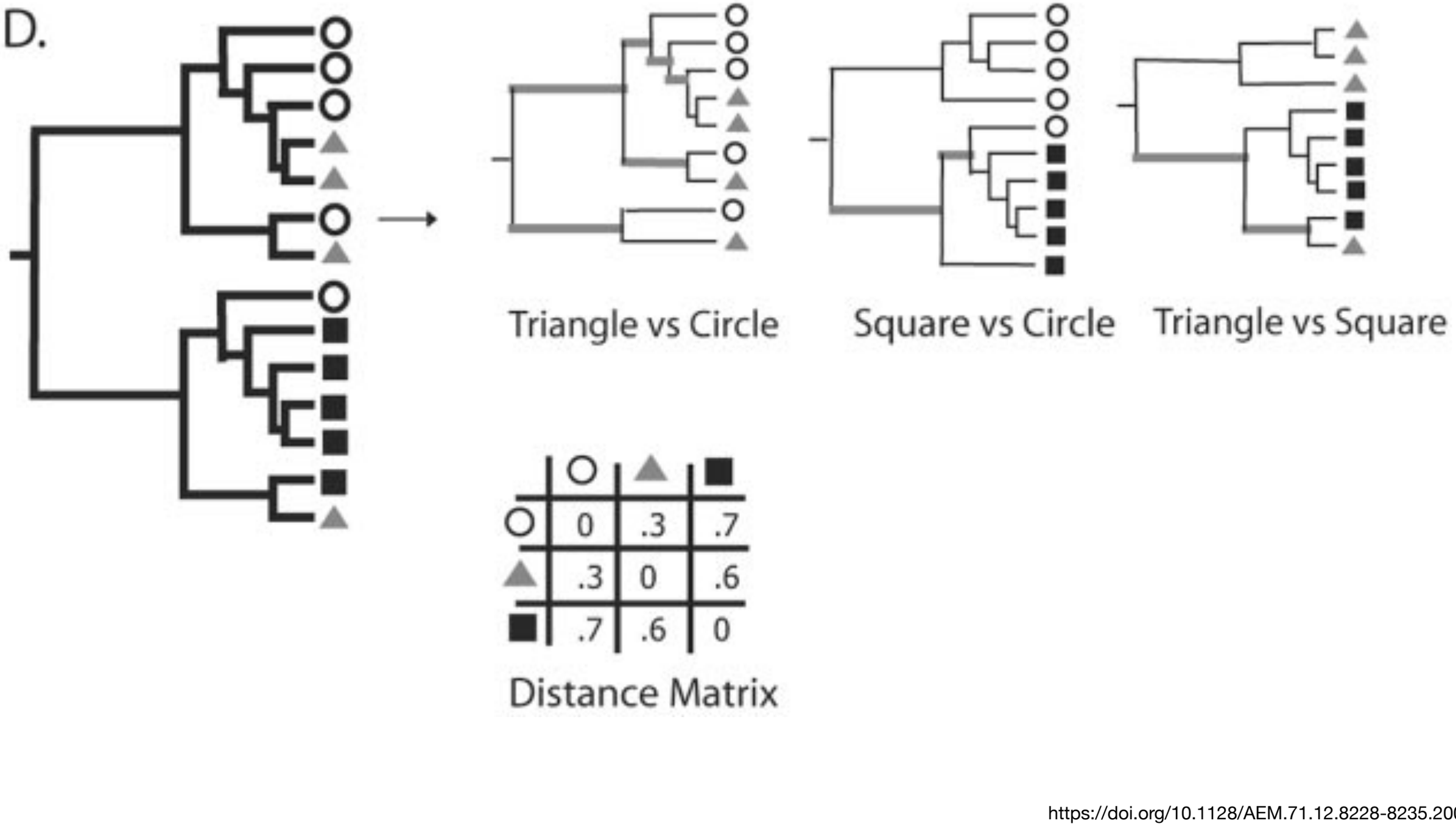
1. Label branches leading to taxa from both samples "shared"

2. Label branches leading to taxa which appears only in one sample "unshared".

3. Unifrac is the fraction of total branch length which is unshared. Alternatively, (the sum of "unshared" branch lengths)/ (the sum of all tree branch lengths)

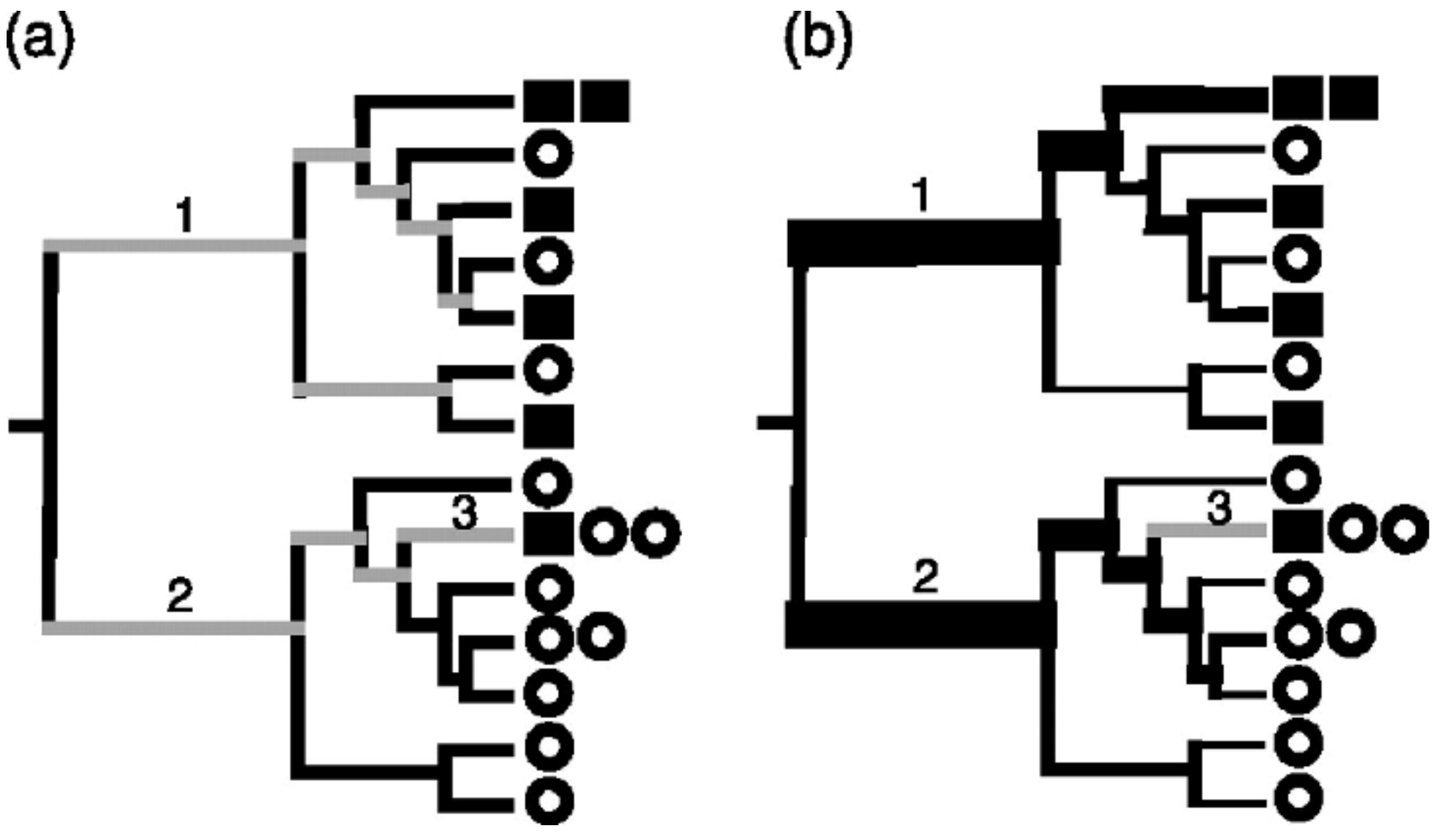
https://doi.org/10.1128/AEM.71.12.8228-8235.2005

UniFrac



https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Weighted UniFrac



https://dx.doi.org/10.1128%2FAEM.01996-06

UniFrac:
Unweighted vs Weighted

- Unweighted UniFrac
  - Qualitative
  - Sensitive to differences in overall community structure
    - Strongly influenced by differences in rare (low abundance) taxa
- Weighted UniFrac
  - Quantitative
  - Sensitive to differences in high abundance taxa

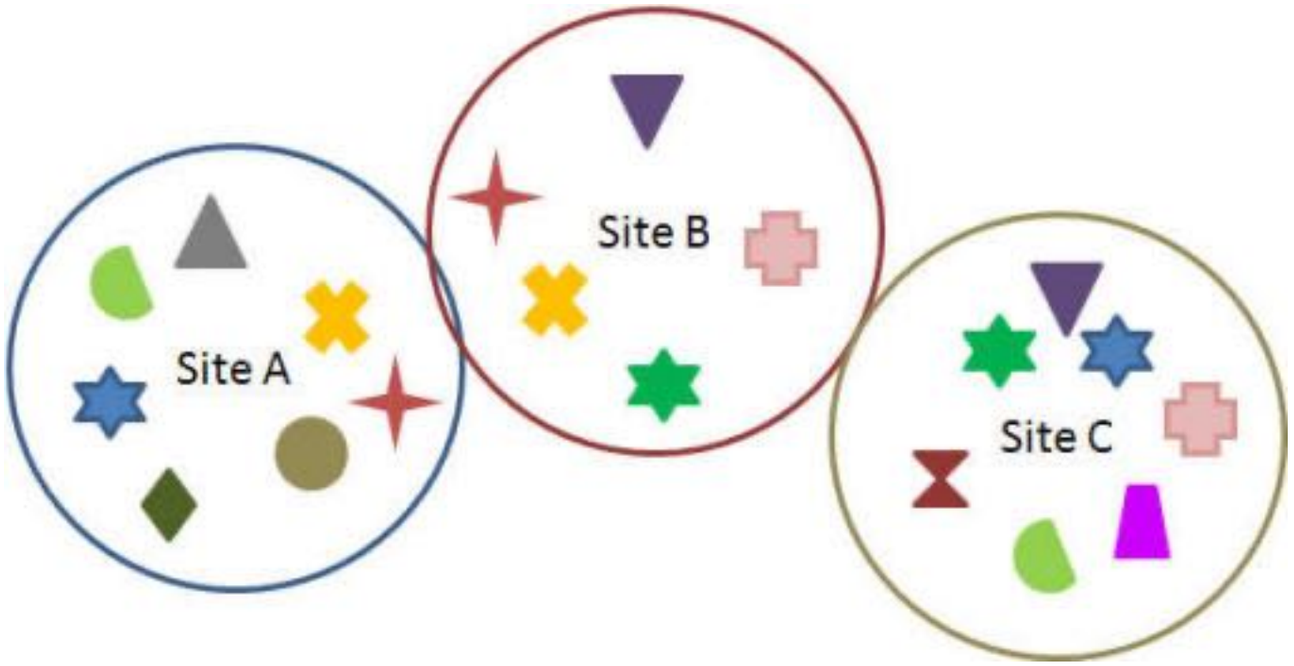
Generalized UniFrac

- Sensitive to differences in moderately abundant taxa

- Maintains ability to detect differences in high and low abundance taxa

https://dx.doi.org/10.1093%2Fbioinformatics%2Fbts342

Beta Diversity Metrics



Site

A

B

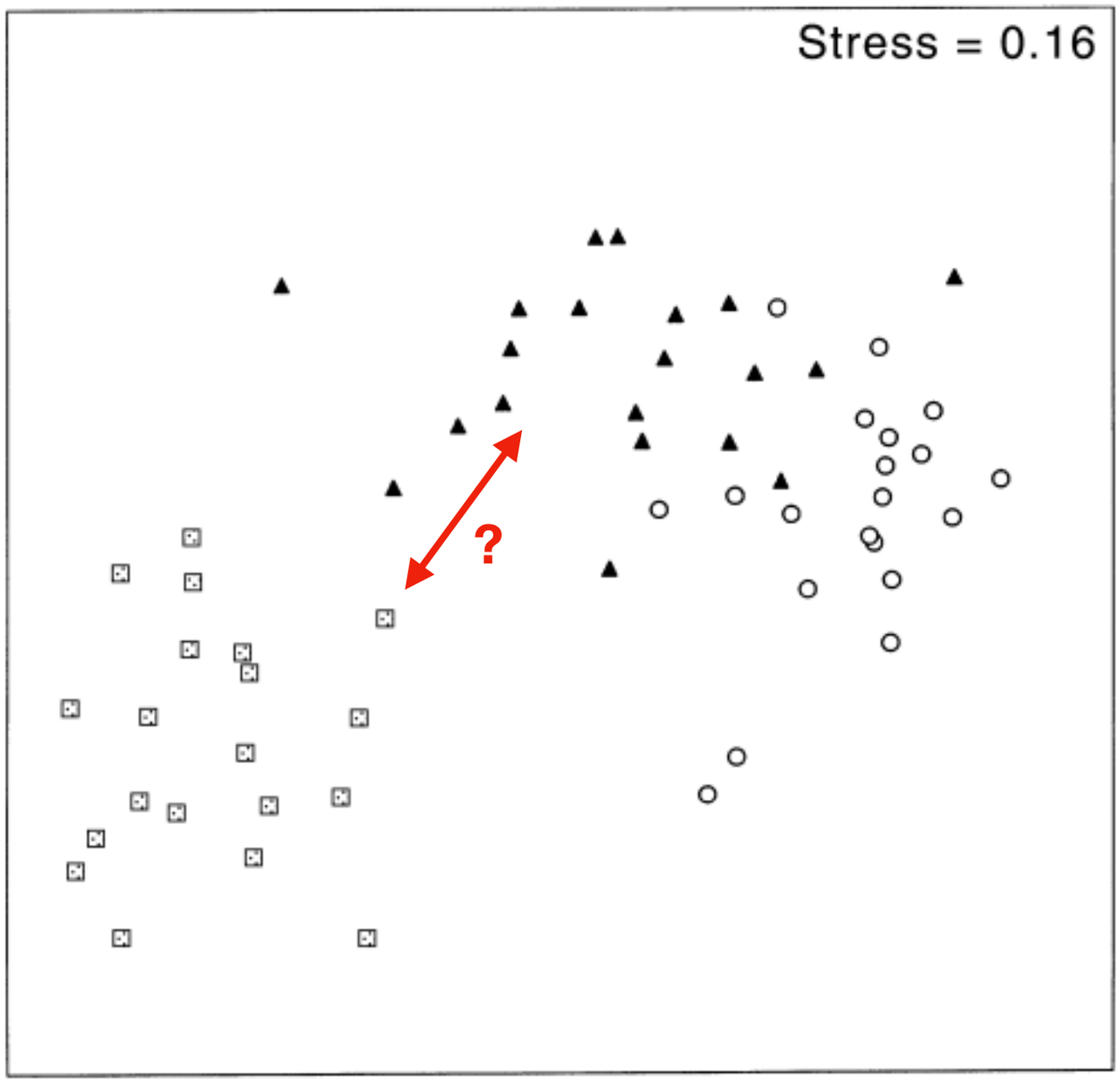
C

	Species Differences
A vs B	8
B vs C	6
A vs C	10

http://www.webpages.uidaho.edu/veg_measure/Modules/Lessons/Module%209(Composition&Diversity)/9_2_Biodiversity.htm

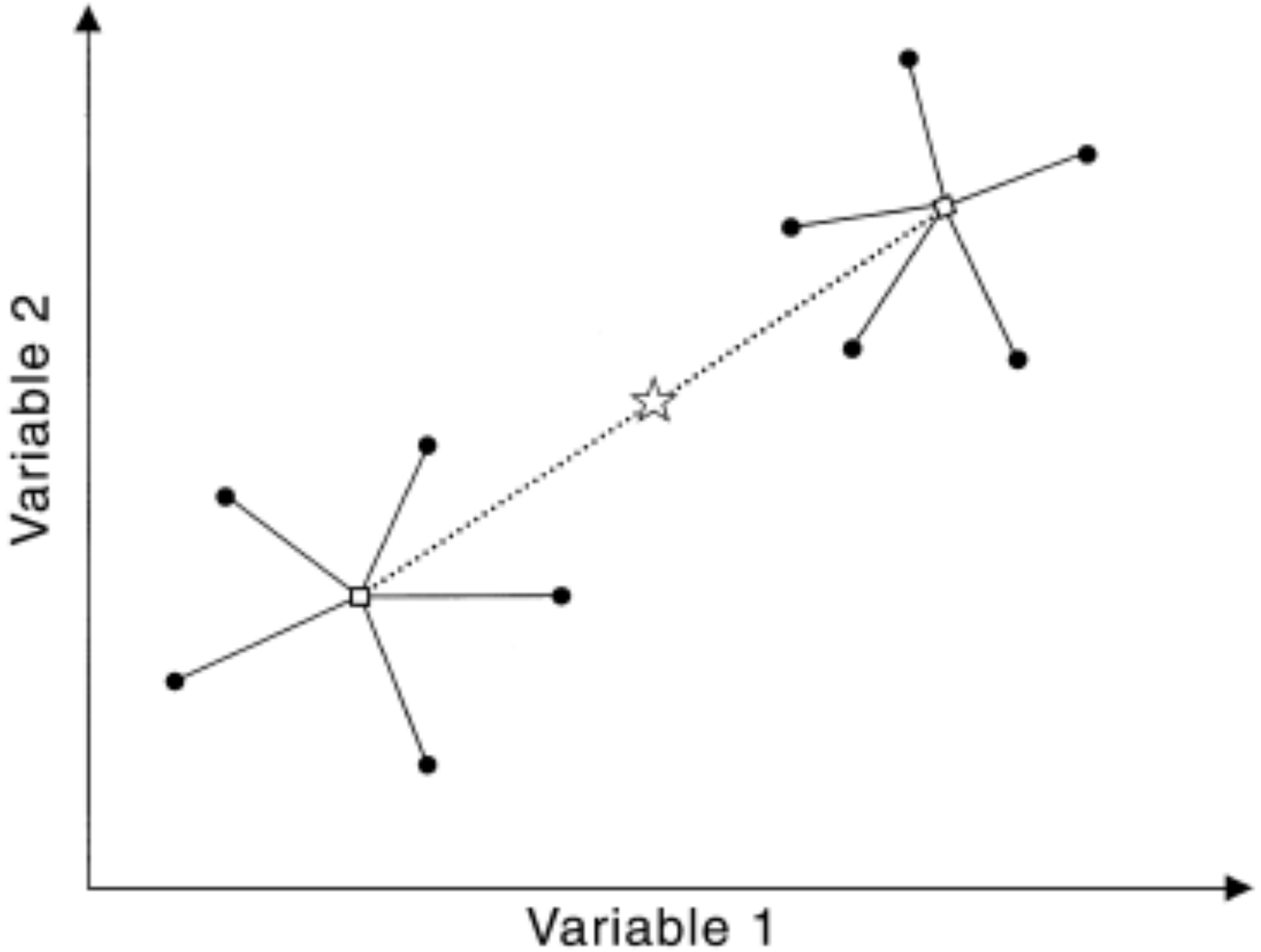
β-Diversity: Comparisons

Non-Parametric Multivariate Tests



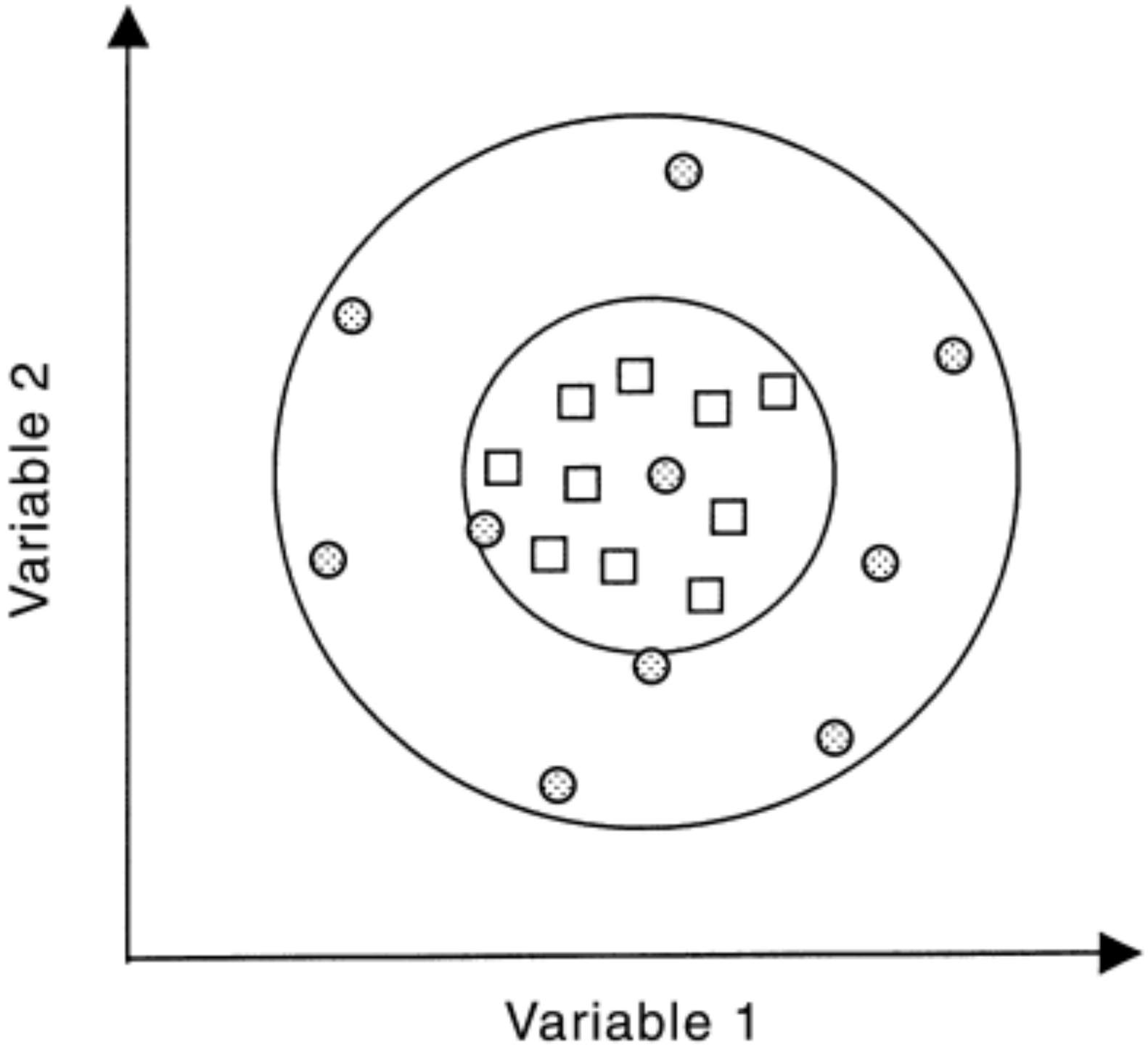
https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

PERMANOVA



https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

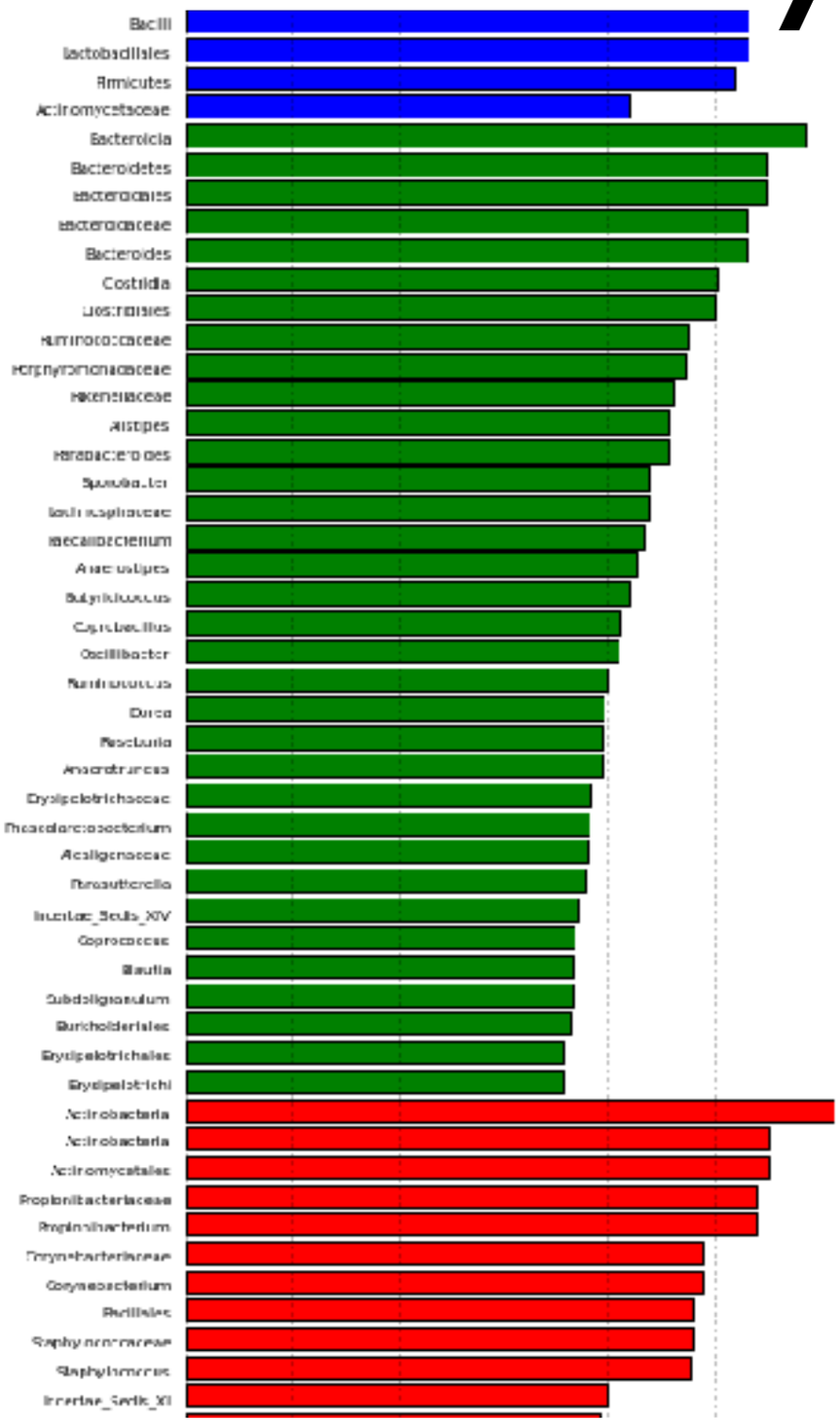
PERMANOVA Caveat



https://doi.org/10.1111/j.1442-9993.2001.01070.pp.x

Biomarker Discovery

- Question: Which OTUs* have different abundance between Site A and Site B



* or higher level taxonomic groups,

LEfSe Tutorial

Biomarker Discovery

- LEfSe
- MetaBoot
- Metastats
- LIBSVM
- mRMR
- Regularized Low Rank-Sparse Decomposition (RegLRSD)

P >> N

	OTU 1	OTU 2	...	Meta 1	Meta 2	...
Sample 1						
Sample 2						
...						
Sample N						

N

P

Machine Learning

- Random Forests
- Support Vector Machines
- Kernel Ridge Regression
- Regularized Logistic Regression
- Bayesian Logistic Regression
- K-Nearest Neighbors
- Probabilistic Neural Networks
- etc