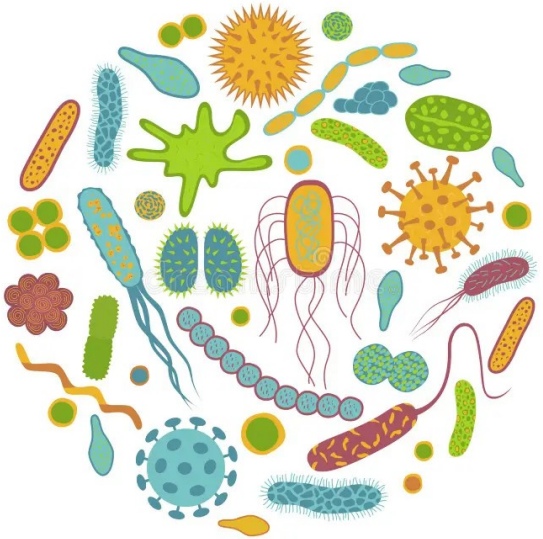
Descriptive statistics
α and β diversity

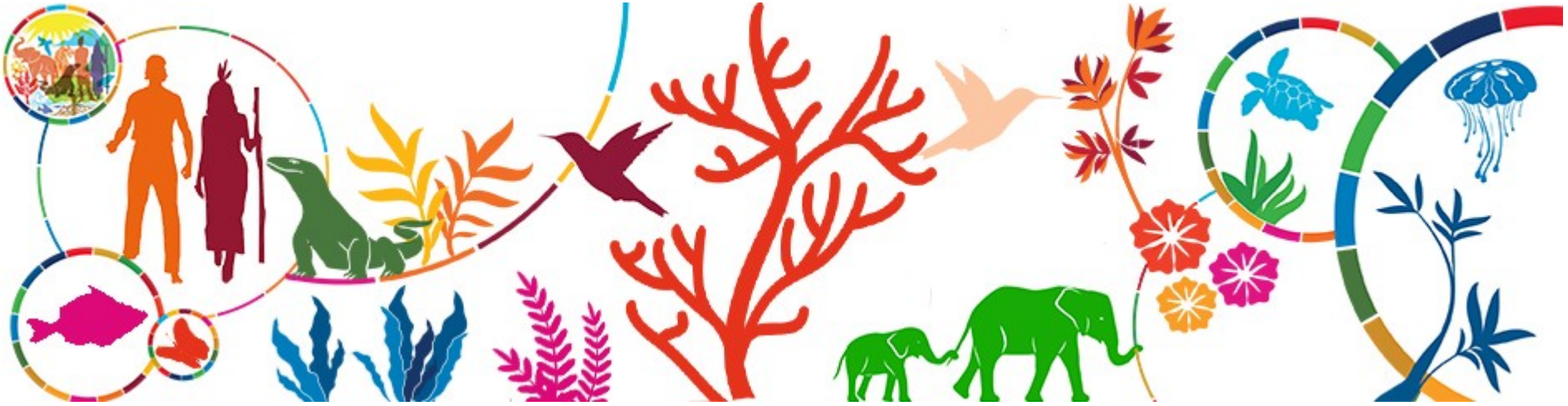


Vincent Manzanilla

intertryp | Host-vector-parasite-environment interactions in the Neglected Tropical Diseases due to Trypanosomatids

Institut de Recherche pour le Développement FRANCE
French National Research Institute for Sustainable Development

What is Biodiversity?



The **variety** and **variability** among living organisms on site, ecosystem and their **interactions** between them.

Diversity can be use to describe **variations** in several forms:

Genetic

Taxonomic

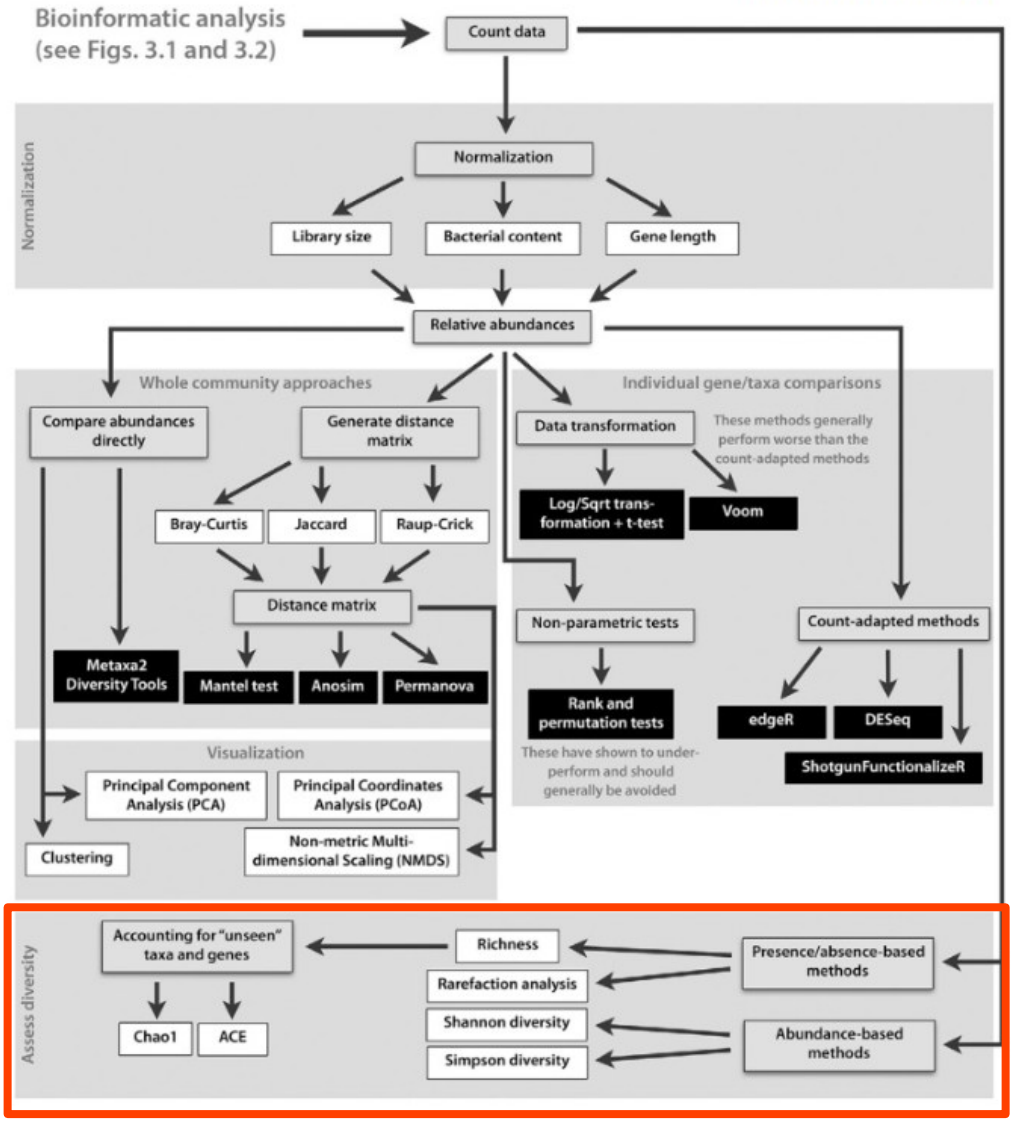
Funcional group (e.g. nitrogen-fixing)

Why measuring Biodiversity?

Biodiversity —proxy→ **Ecosystem health**

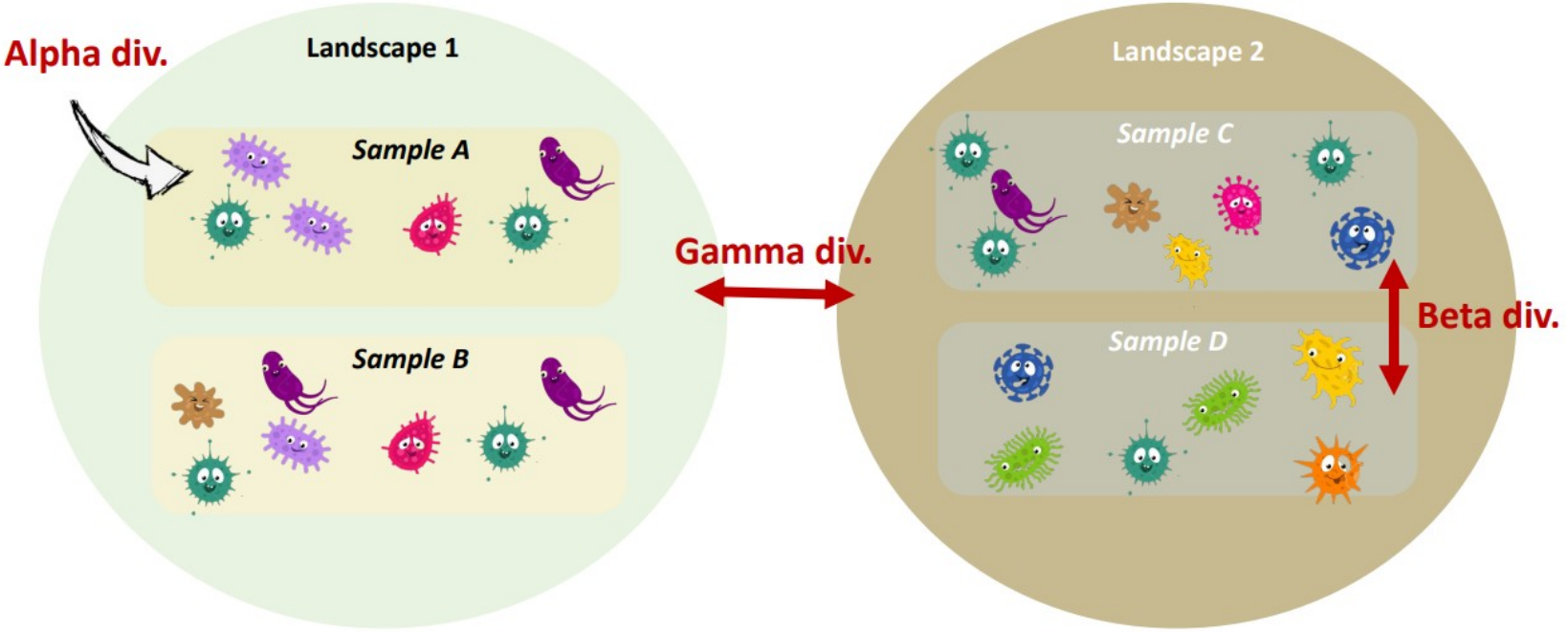
In general, diverse communities are believed to have increased stability, increased productivity, and resistance to invasion and other disturbances.

Bengtsson-Palme, 2018



Alpha vs Beta vs Gamma Diversity (Whi4aker, 1972)

Alpha, beta and gamma diversity are three types of biodiversity measures described over a special scale



α-Biodiversity ?

α-diversity is local diversity, measured within a closed system

→ **The diversity within an habitat of fixed size**

α-biodiversity has two components

- **Richness**
- **Evenness**

→ **Follow the evolution of populations over time, but also to compare studied stations**

Components of Diversity - Definition

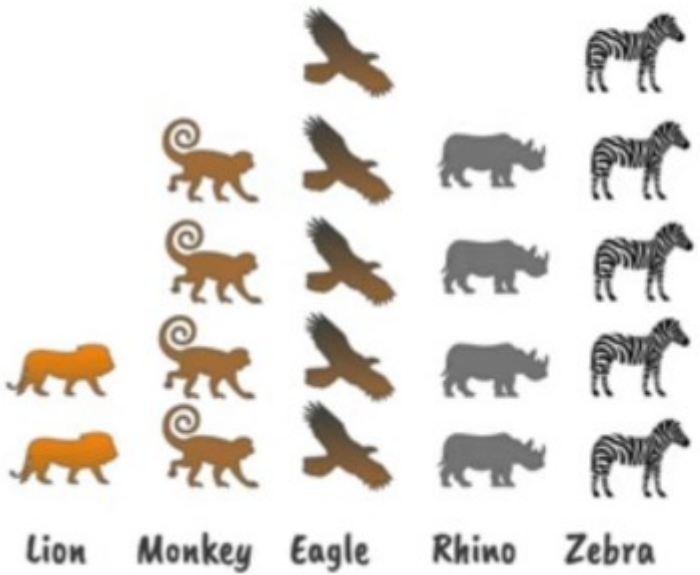
Specific Richness (S) = The measurement of the number of species present in a location/studied system

- The more species present, the richer sample is
- Gives equal weight to species which have many/few individuals

Evenness (equitability) = Relative population of each species

- Species represented by many individuals or by few ones do not give the same contribution Evenness index is independent of Richness!!

Higher biodiversity



Wide range of species = high richness

Relatively equal proportion = high evenness

Lower biodiversity

Few species = low richness

Unequal proportion = low evenness

Shannon index of Diversity

Species are considered as equidistant (= do not consider species relatedness)

$$H = -\sum_{i=1}^{s} p_i \ln(p_i)$$

where

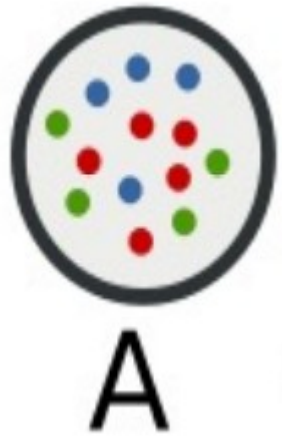
H = the Shannon index value

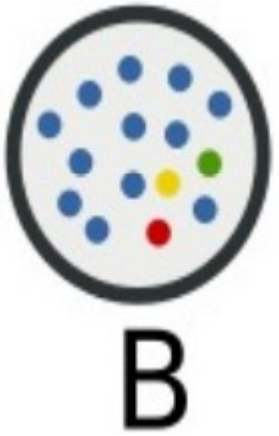
p_i = the proportion of individuals found in the ith species

ln = the natural logarithm

s = the number of species in the community

Calculate the Shannon diversity index for A and B

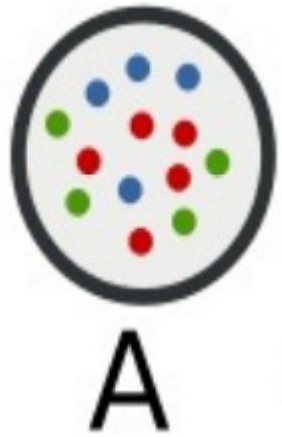


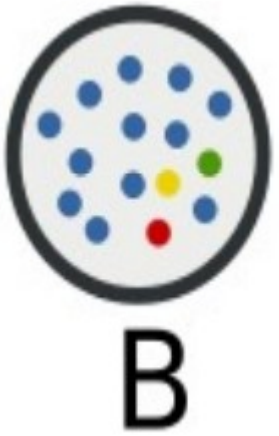


A: consists of 3 species, of which 4 green, 5 red & 4 blue

B consist of 4 species, of which 1 green, 1 red, 1 yellow & 11 blue

Calculate the Shannon diversity index for A and B





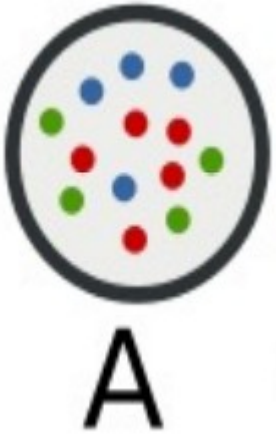
A: consists of 3 species, of which 4 green, 5 red & 4 blue

$$\frac{4}{13}log\frac{4}{13}+\frac{5}{13}log\frac{5}{13}+\frac{4}{13}log\frac{4}{13}$$

B consist of 4 species, of which 1 green, 1 red, 1 yellow & 11 blue

$$\frac{1}{14}log\frac{1}{14}+\frac{1}{14}log\frac{1}{14}+\frac{1}{14}log\frac{1}{14}+\frac{11}{14}log\frac{11}{14}$$

Calculate the Shannon diversity index for A and B



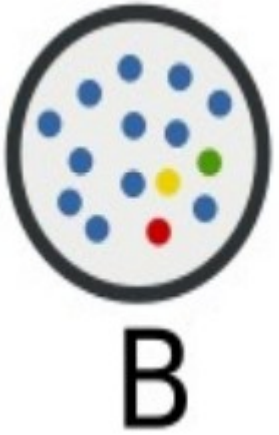
A

Shannon value:
Influenced by richness

H is generally between 1.5 – 4

H gets **higher** when :

- There are **more species**
- There is **greater evenness**



B

A: consists of 3 species, of which 4 green, 5 red & 4 blue

$$\frac{4}{13}log\frac{4}{13}+\frac{5}{13}log\frac{5}{13}+\frac{4}{13}log\frac{4}{13}=1.09$$

B consist of 4 species, of which 1 green, 1 red, 1 yellow & 11 blue

$$\frac{1}{14}log\frac{1}{14}+\frac{1}{14}log\frac{1}{14}+\frac{1}{14}log\frac{1}{14}+\frac{11}{14}log\frac{11}{14}=0.72$$

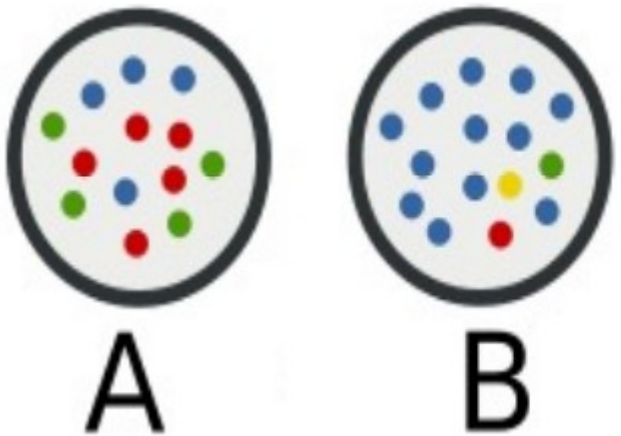
Simpson's Index of Diversity

Idea : Indicates the taxa dominance and gives the probability of two individuals that belong to the same taxa being randomly chosen

$$D = 1 - \sum_{i=1}^{N} (p_i)^2$$

$$1 - \left(\frac{3}{13}\right)^2 + \left(\frac{5}{13}\right)^2 + \left(\frac{4}{13}\right)^2 = 0.704$$

$$1 - \left(\frac{1}{14}\right)^2 + \left(\frac{1}{14}\right)^2 + \left(\frac{1}{14}\right)^2 + \left(\frac{11}{14}\right)^2 = 0.6$$

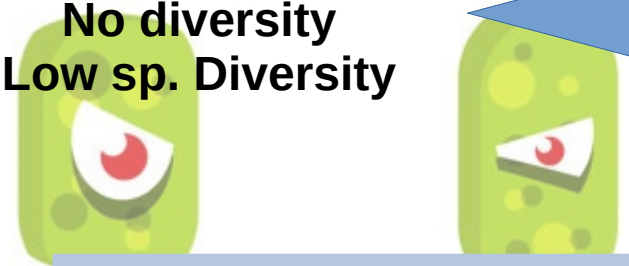


A value of 0.7 … 2 sequences randomly selected have 70% chance to belong to the same OTU

- Influenced by highly abundant Taxa
- Greater weight on evenness
- Range 0 to 1 (high)

The closer the value to 1, the more diverse the habitat is

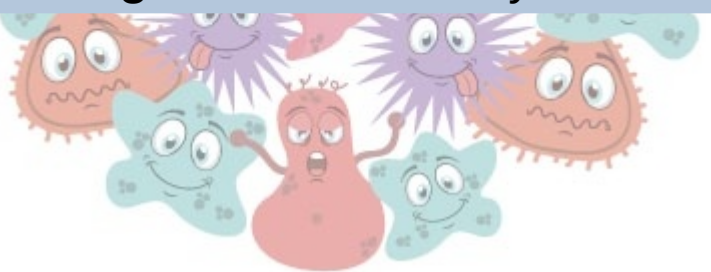
0
No diversity
Low sp. Diversity



→ **Few** successful species in the habitat
→ **Env stressful,** few niches, few organisms well adapted to env
→ Any **change** in env may have **serious effect on ecosystem**

1
Infinite diversity
Hig sp. Diversity

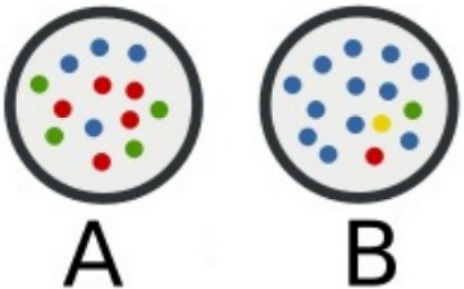
→ **Great number** of successful species, more **stable ecosystem**
→ Env is less likely to be Hostile
→ Complex food chain
→ Env change is **less likely to dammage** the whole ecosystem



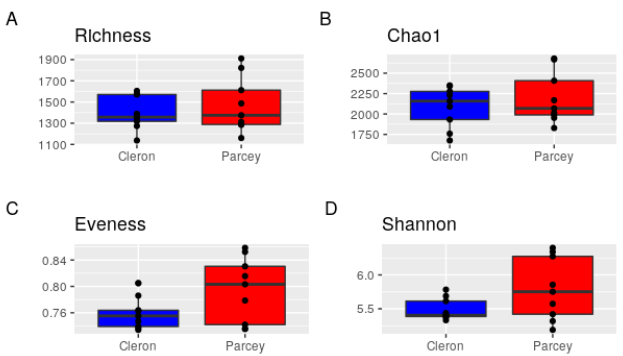
Diversity Estimators

- Chao1 & ACE are non-parametric estimators of taxa richness
- Chao1 estimates the total richness

$$S_{chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{n_2(n_2 + 1)}$$



$$S_{chao1} = 3 + \frac{0(0 - 1)}{3(3 + 1)}$$



S_{chao1} = the estimated richness

S_{obs} = the observed number of species

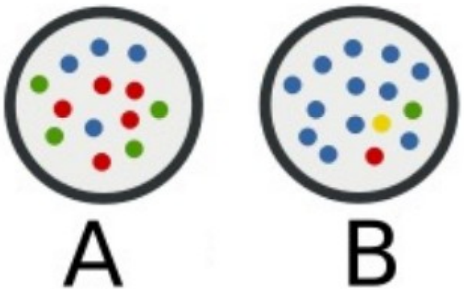
N_1 = 1 the number of OTUs with only one sequence (i.e. “singletons”)

N_2 > 1the number of OTUs with only two sequences (i.e. “doubletons”)

Diversity Estimators

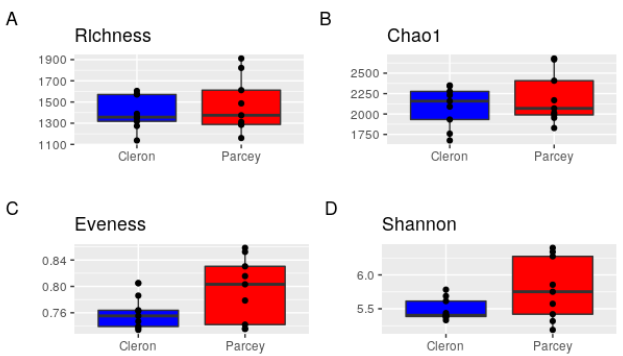
- Chao1 & ACE are non-parametric estimators of taxa richness
- Chao1 estimates the total richness

$$S_{chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{n_2(n_2 + 1)}$$



$$S_{chao1} = 3 + \frac{0(0 - 1)}{3(3 + 1)}$$

$$S_{chao1} = 4 + \frac{3(3 - 1)}{1(1 + 1)}$$



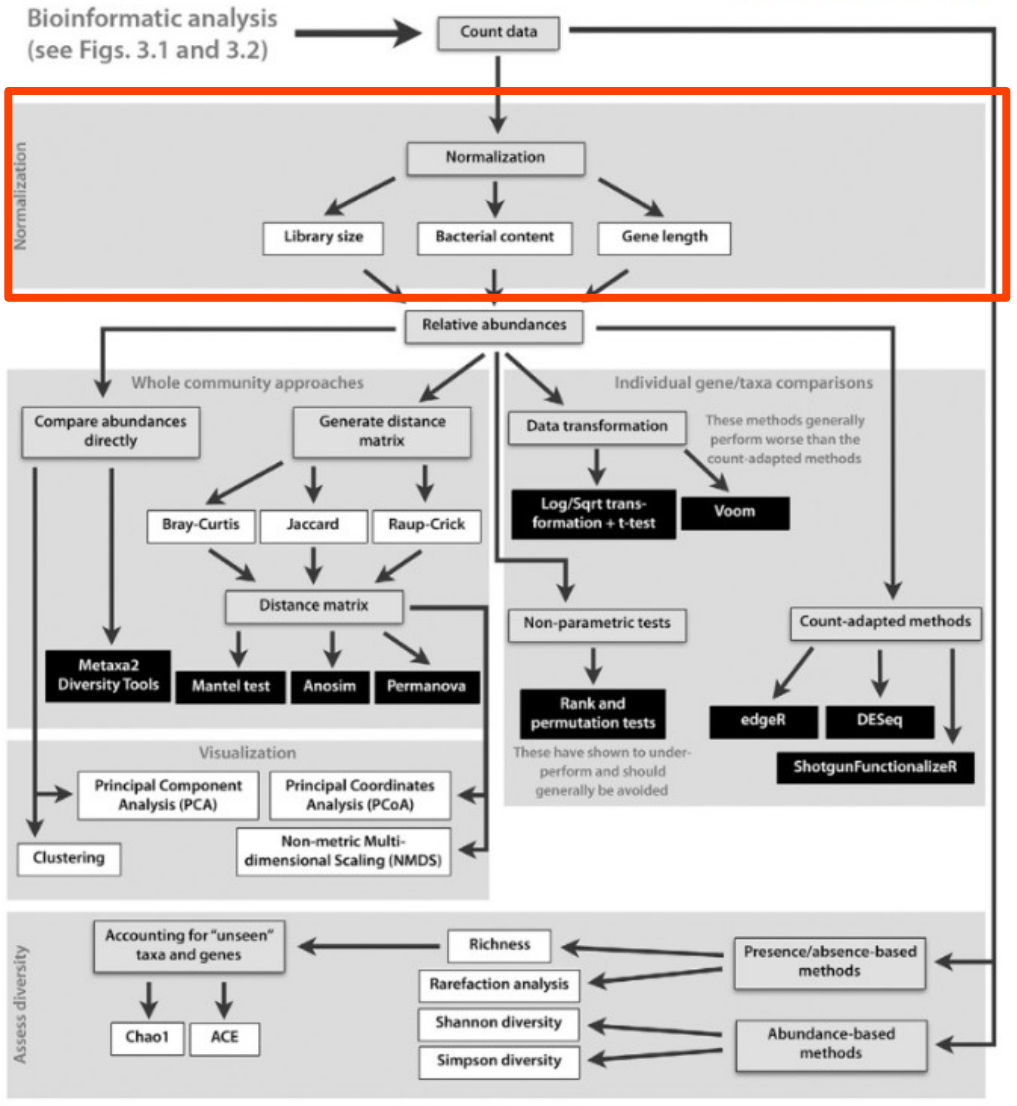
S_{chao1} = the estimated richness

S_{obs} = the observed number of species

N_1 = 1 the number of OTUs with only one sequence (i.e. “singletons”)

N_2 > 1the number of OTUs with only two sequences (i.e. “doubletons”)

Bengtsson-Palme, 2018

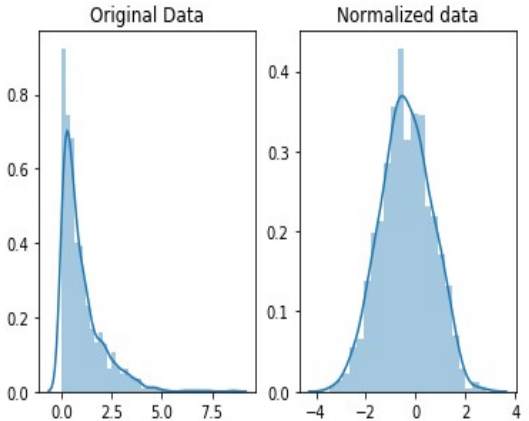


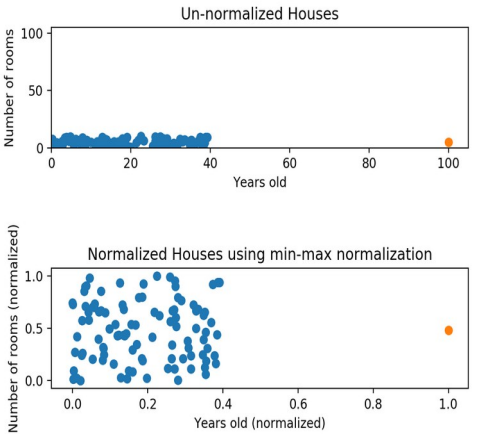
Normalization

Microbiome data often exhibit significant variance in sequencing depth among samples. Such variances, if unaccounted for, can skew the analysis and interpretation. Data normalization is crucial to account for this technical variation, ensuring that samples are comparable and results are robust.

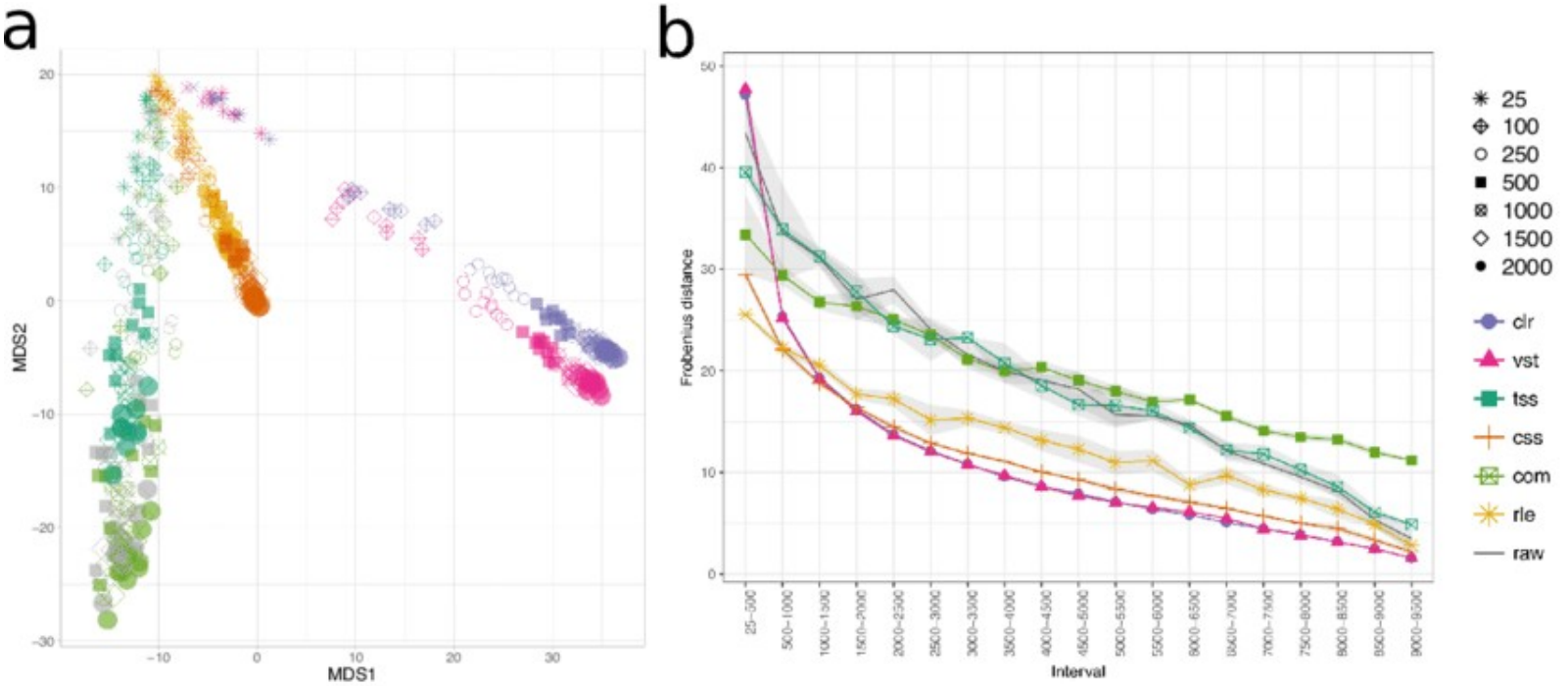
Why transformation?

- To reduce the variation range (e.g. give low weight to extreme values)
- Correcting sparsity and over-dispersion





Normalization, which one?



Transformations that remove compositional artifacts (CLR) and variance stabilizing transformation (VST) result in substantially different patterns of correlation. A) Multidimensional scaling representation distances between correlation structures of varying size estimated from different normalization methods. These estimates are also compared to untransformed or raw count values (dark grey points). B) Distance between sub-samples of different sizes. Lines represent mean and grey ribbon represent standard deviation from the mean. (color scheme as in A)

https://doi.org/
10.1101/406264

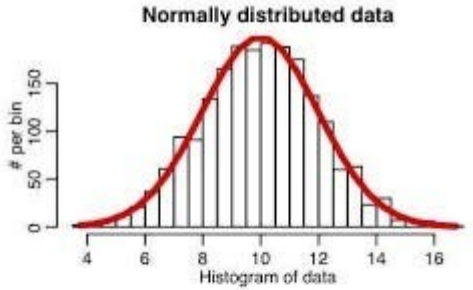
Compare your distribution with a normal distribution

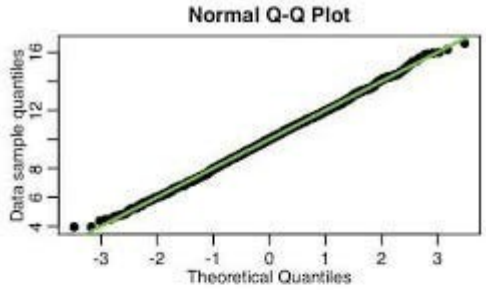
Check normality of data: Shapiro Test & QQ-plots.

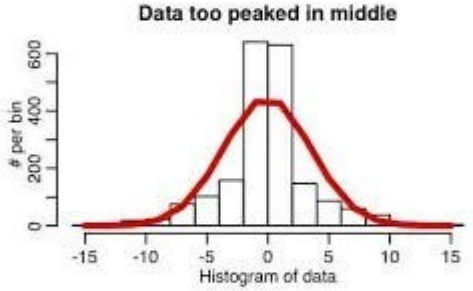
Shapiro:

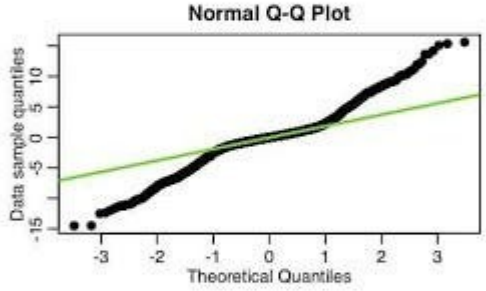
H0 is «data follow normal distribution»,

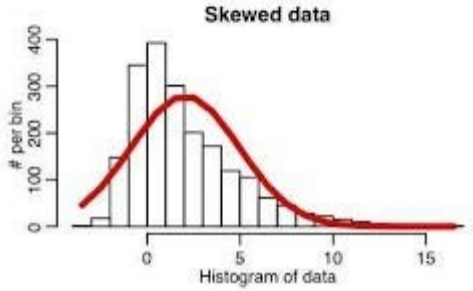
H1 is «data do not follow normal distribution».

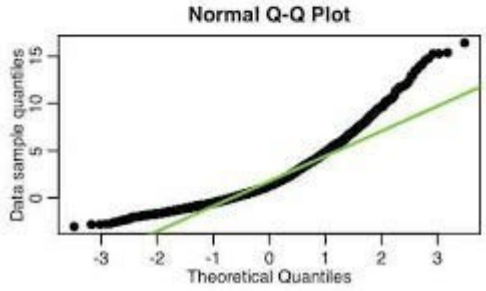


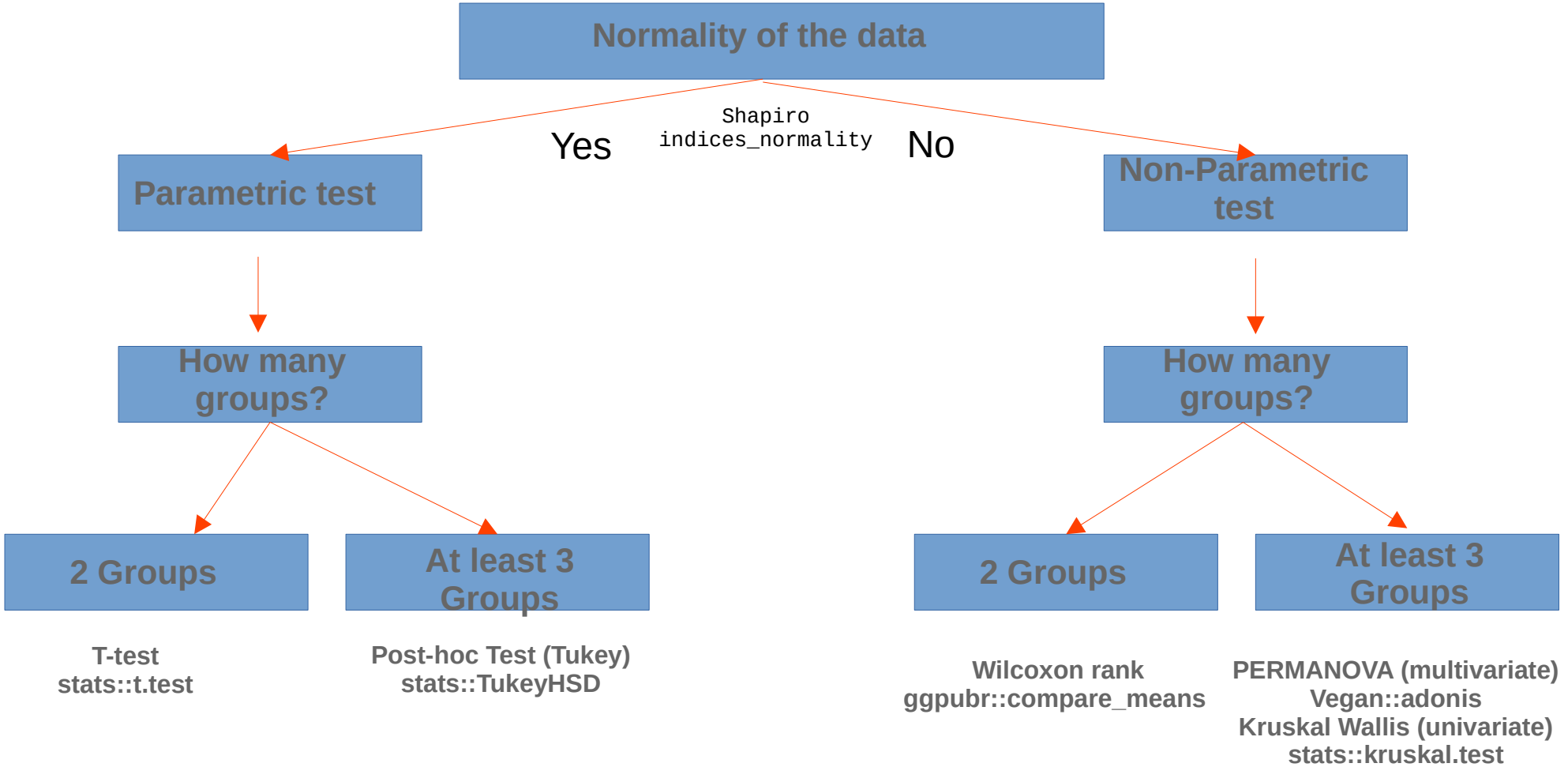




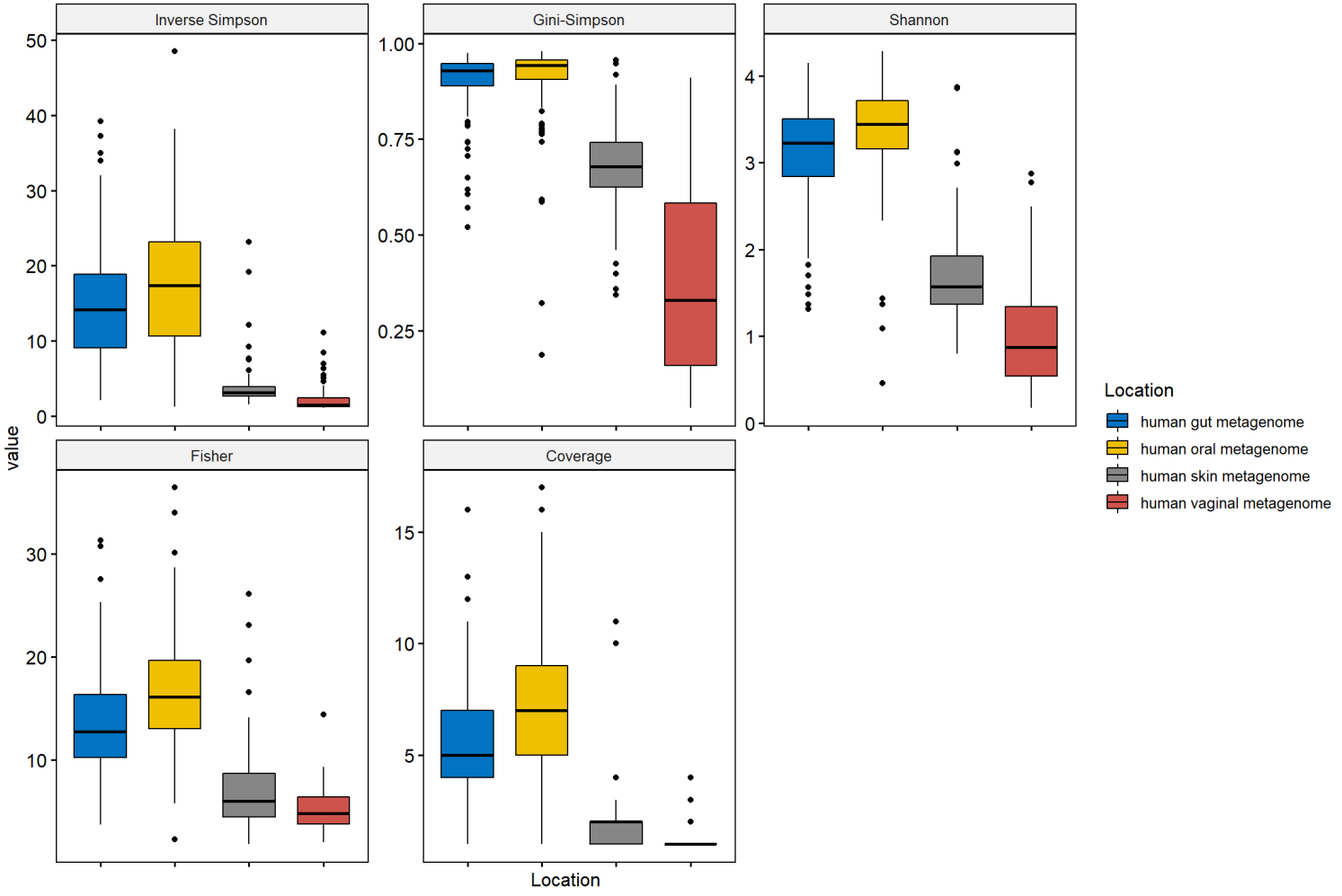


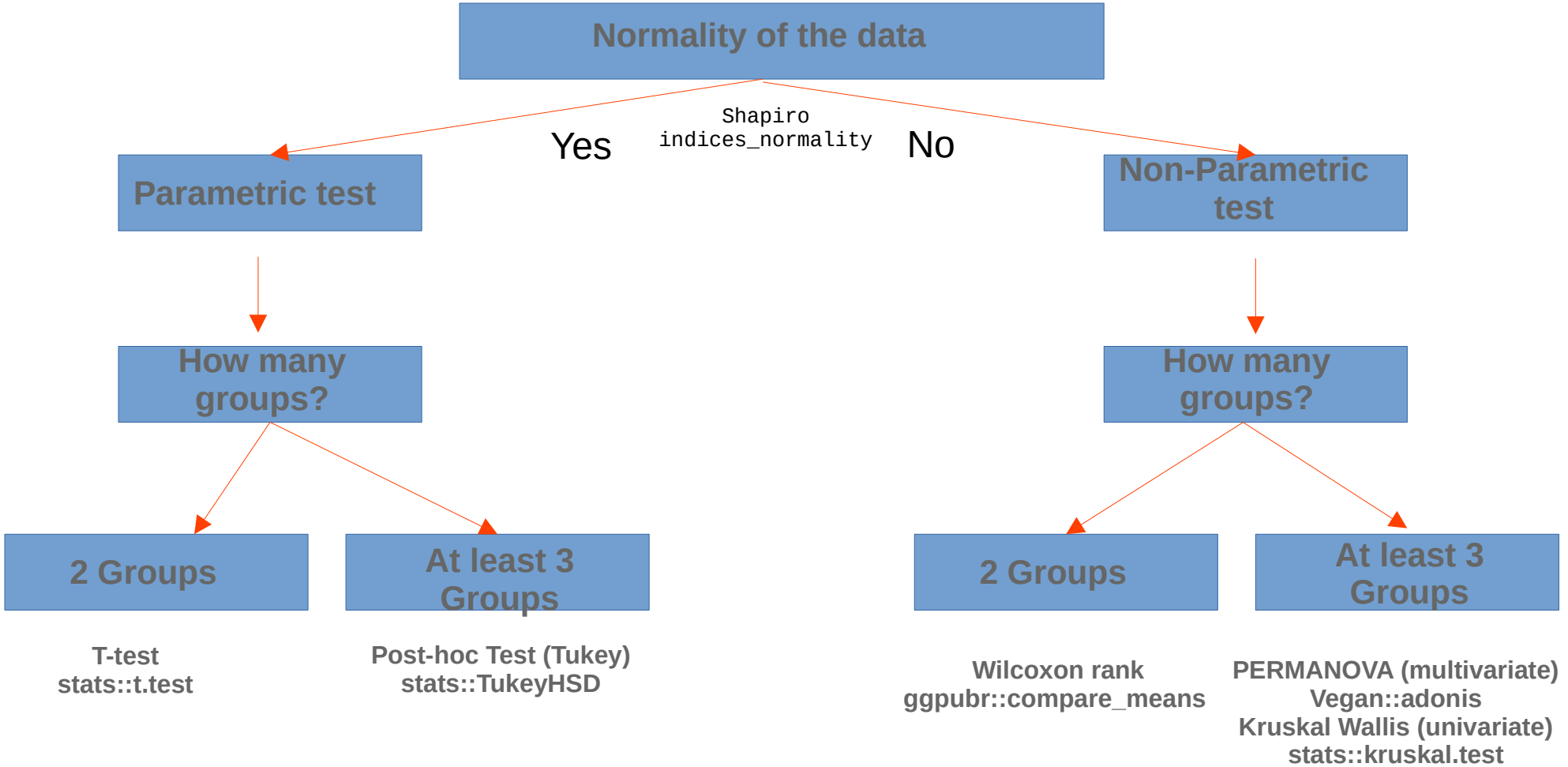





Normality of the data
Shapiro
indices_normality
Yes
No
Parametric test
Non-Parametric test
How many groups?
How many groups?
2 Groups
At least 3 Groups
2 Groups
At least 3 Groups
T-test
stats::t.test
Post-hoc Test (Tukey)
stats::TukeyHSD
Wilcoxon rank
ggpubr::compare_means
PERMANOVA (multivariate)
Vegan::adonis
Kruskal Wallis (univariate)
stats::kruskal.test

Are the groups significantly different?



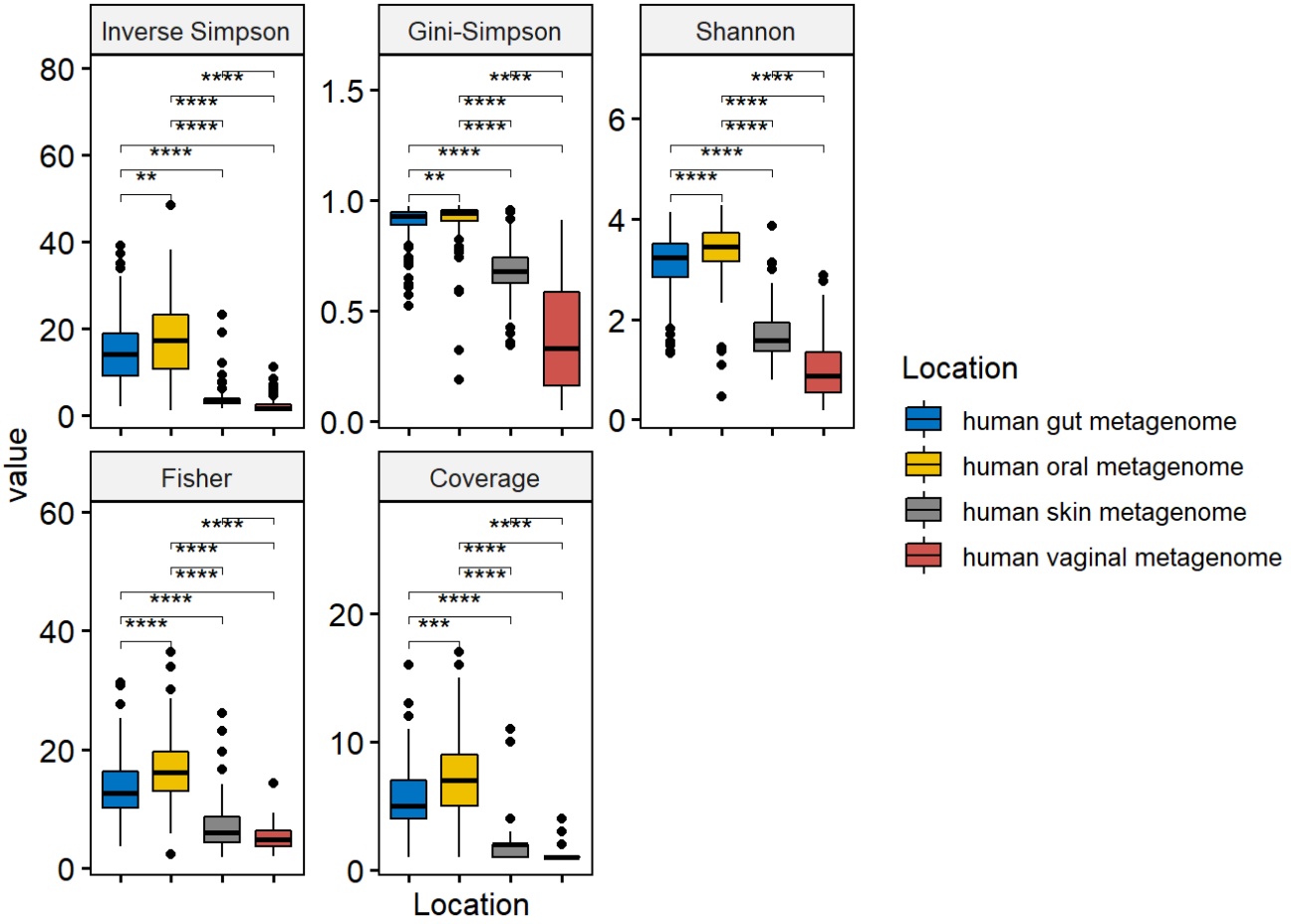

Normality of the data
Shapiro
indices_normality
Yes
No
Parametric test
Non-Parametric test
How many groups?
How many groups?
2 Groups
At least 3 Groups
2 Groups
At least 3 Groups
T-test
stats::t.test
Post-hoc Test (Tukey)
stats::TukeyHSD
Wilcoxon rank
ggpubr::compare_means
PERMANOVA (multivariate)
Vegan::adonis
Kruskal Wallis (univariate)
stats::kruskal.test

Are the groups significantly different?

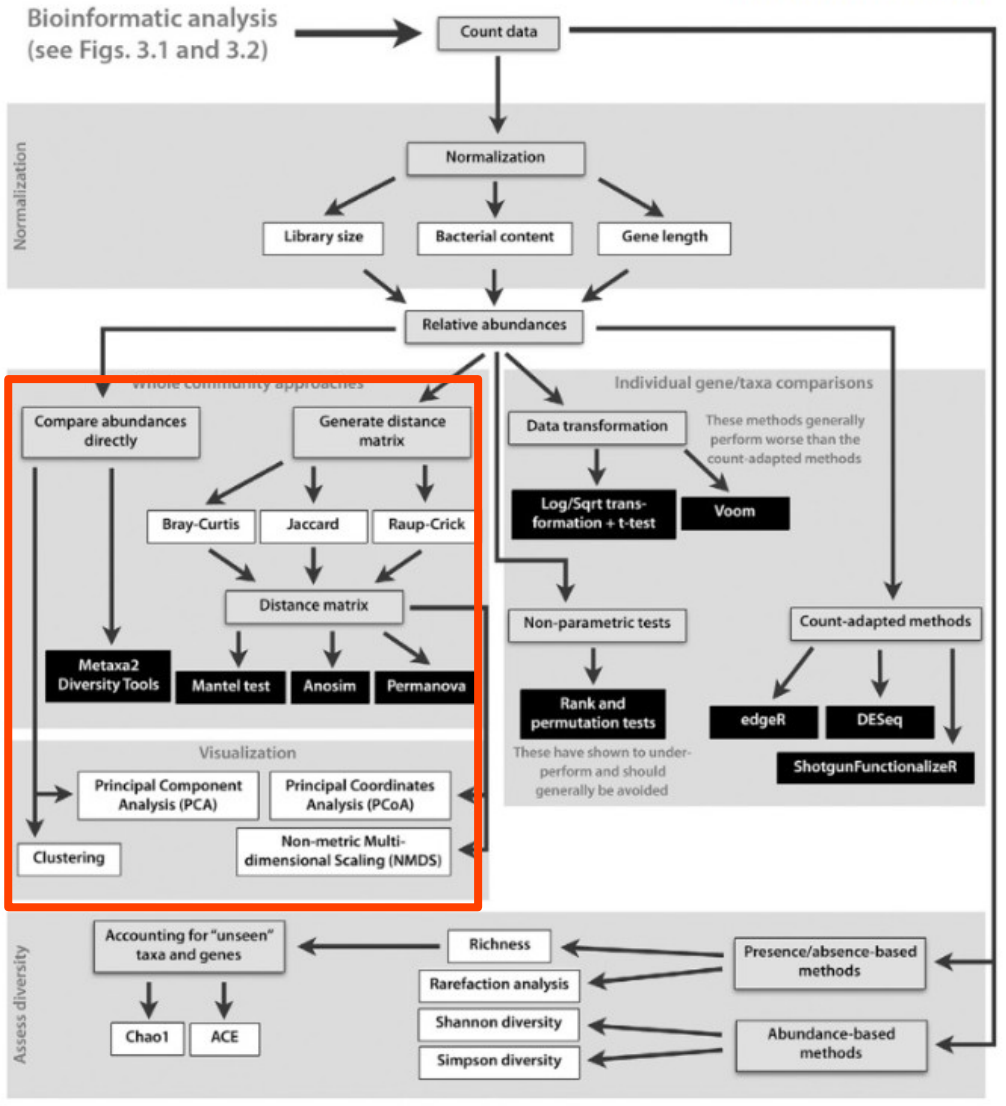
- Test with the appropriate test for more than 3 groups:

  Kruskal-Wallis multiple comparison

  (FSA::dunnTest)



Bengtsson-Palme, 2018



β diversity

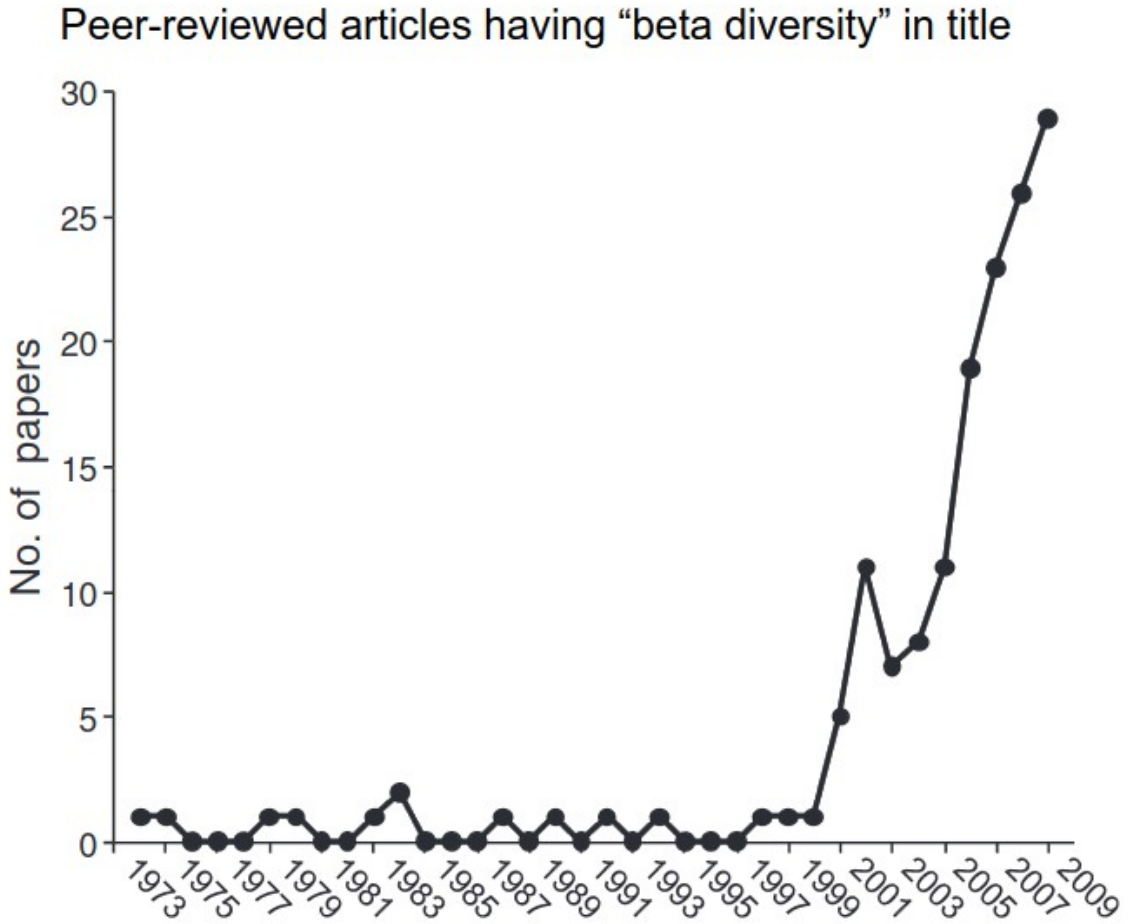
- Microbial ecologists typically use beta diversity as a broad umbrella term that can refer to any of several indices related to compositional differences (Differences in species content between samples)
- For some reason this is contentious, and there appears to be ongoing (and pointless?) argument over the possible definitions
- For our purposes, and microbiome research, when you hear “beta-diversity”, you can probably think:

Diversity of species composition

or

Analysis comparing whole microbiomes to one another Inter-sample comparison of the community composition

β diversity

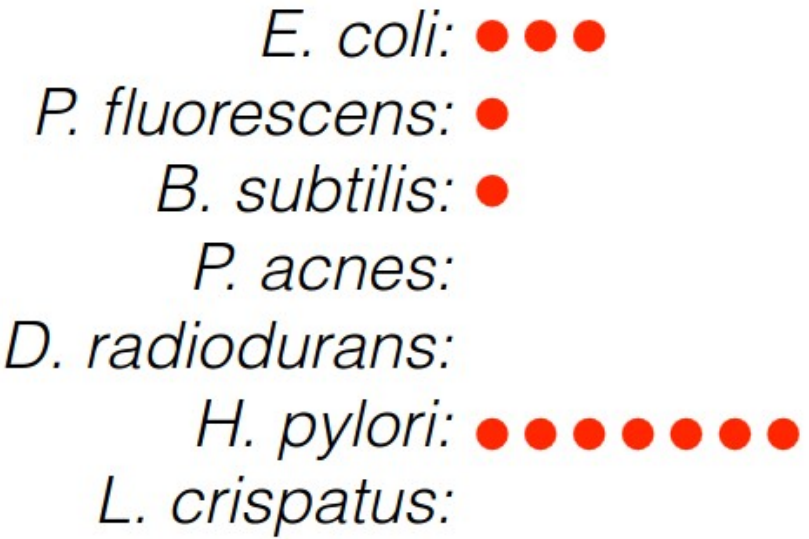


Anderson, M. J., et al. (2011). Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecology Letters, 14(1), 19–28.

β diversity

Communities are a vector of abundances:

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



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

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Community Distance Properties

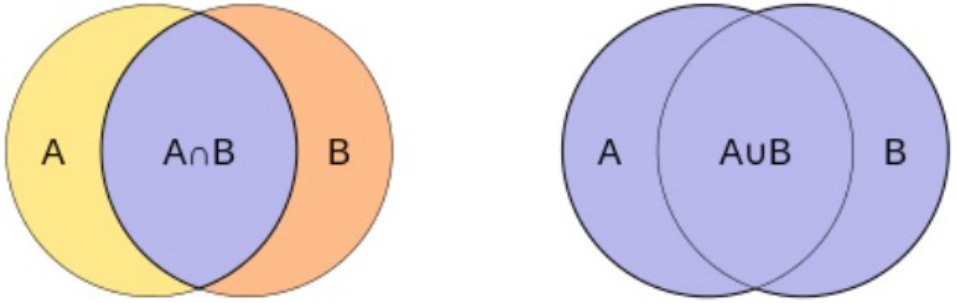
- Range from 0 to 1
- Distance to self is 0
- If no shared taxa, distance is 1

β diversity - Most common dissimilarities/distance used for species data

Dissimilarities Distances	Taxonomic	Phylogenetic
Compositional (Binary)	Sorensen Jaccard Ochiai	Unweighted Unifrac PhyloSor
Structural (Quantitative)	Bray-Curtis Chord Hellinger Aitchison	Weighted Unifrac Allen

Jaccard

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

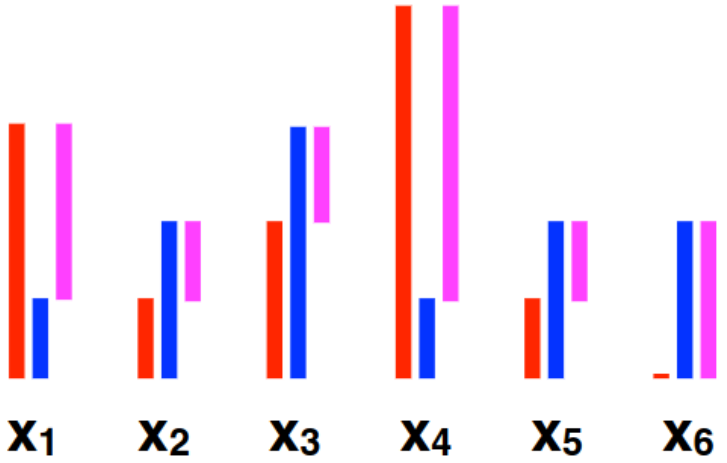


Intuition: Fraction of shared **types** unique to one of the communities

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Bray-Curtis

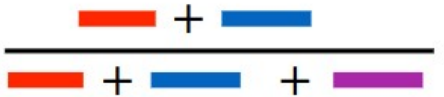
$$Dist(x, y) = \frac{\sum |x_i - y_i|}{\sum x_i + \sum y_i} = \frac{\quad}{\quad + \quad}$$

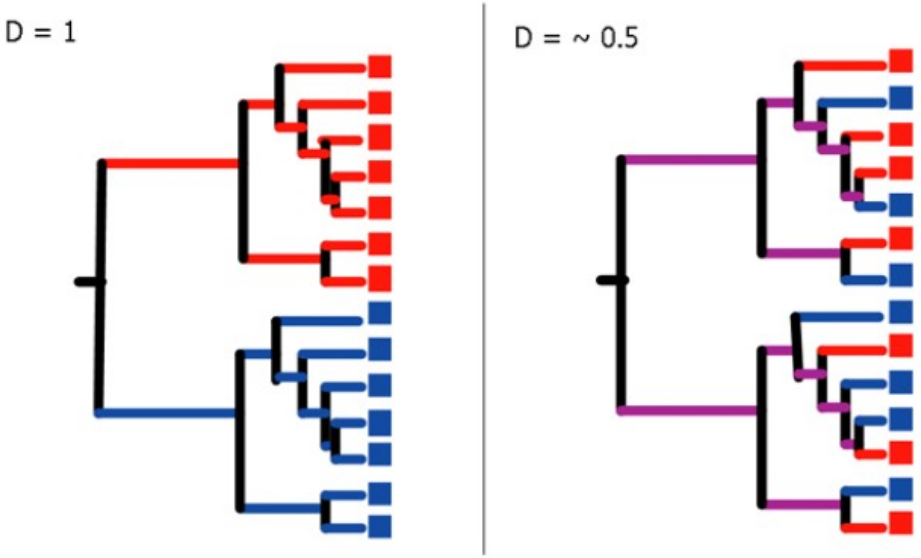


Intuition: *City block distance*. Sum of absolute differences over total abundance.

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Unifrac

Dist(x, y) = 

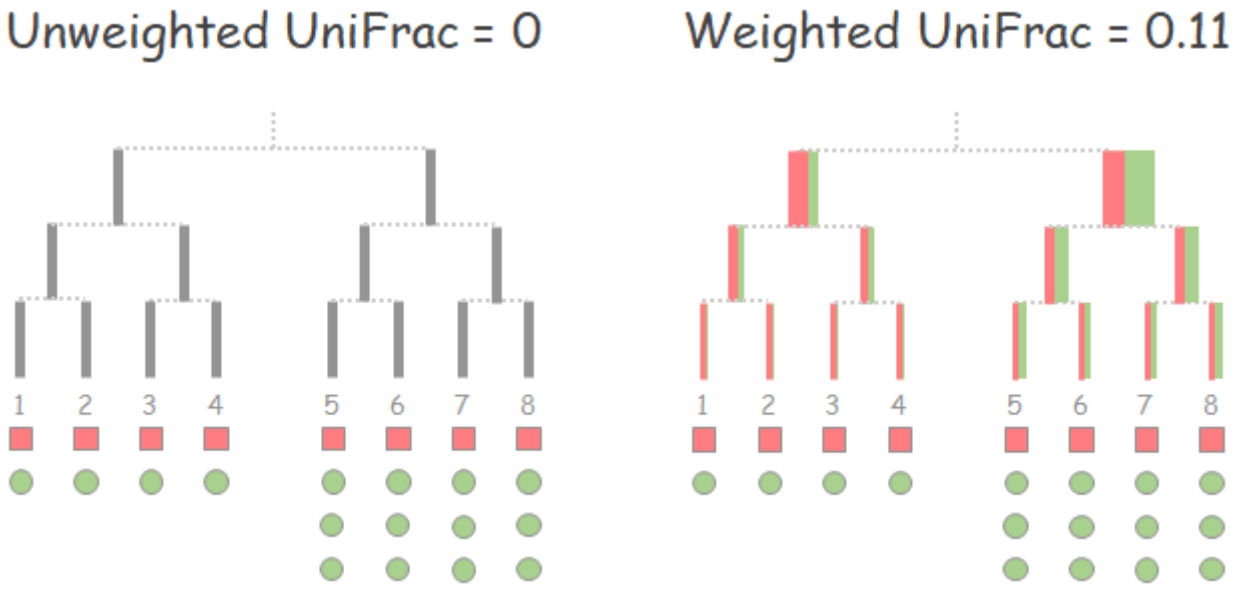


Intuition: Fraction of shared **tree** unique to one of the communities

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone and Knight (2008)

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.

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone et al. (2007)

Questions you should ask yourself before choosing a dissimilarity/distance metric

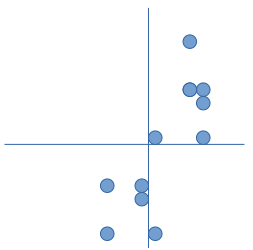
- Do I compare **variables** or **objects**?
- Do I use ASV/OTU/**species variables** or another type (e.g.; physico-chemical)? **Asymmetrical** Vs **symmetrical dissimilarity** or **distance index**
- What type of data do I have? **Binary** Vs **quantitative** Vs **multifactor**
  - Three broad categories of dissimilarity or distance index :
    - for **binary data** (presence/absence)
    - for **quantitative data**
    - for a **mix of numerical and categorical data** (multifactor)

Distance matrix

- Quantifying ecological **resemblances between samples**, including **similarities** (S) and **dissimilarities** (or distances), is the basic approach of handling **multivariate ecological** data

- Two samples, which contain the same species with the same abundances, have the highest similarity, the similarity decreases with the differences in species composition

- Ordination methods operate with **distances or dissimilarities** between samples



Ordination is a collective term for **multivariate techniques** which summarize a **multidimensional** dataset in such a way that when it is projected onto a **low dimensional space**, any intrinsic pattern the data may possess becomes apparent upon visual inspection (Pielou, 1984).

In **ecological terms**: Ordination summarizes community data (such as species abundance data: samples by species) by producing a **low-dimensional ordination space** in which **similar** species and samples are plotted close together, and **dissimilar** species and samples are placed far apart. Ideally and typically, dimensions of this low dimensional space will represent **important and interpretable environmental gradients**.

Similarity : How do deal with Double-zeros? Co-absence

- Species composition data are sparse matrix, which means that it contains lot of zeros, double zeros
- Double zero" is a situation when certain species are missing in both compared community samples for which similarity/distance will be next calculated!

	Sp A	Sp B	Sp C
Site 1	0	44	**0**
Site 2	11	50	**0**

Does not say anything about ecological similarity or difference between both samples.

Consider them as missing data!

Similarity : How do deal with Double-zeros? Co-absence

	Sp A	Sp B	Sp C
Site 1	0	44	**0**
Site 2	11	50	**0**

Recommendation is to use **dissimilarity indices** or **distance-based** method that **do not take into account the double zero as a resemblance**!!!

Symmetrical vs. Asymmetrical indices

- **Asymmetrical indices ignore** the double-zeros (e.g. bray-Curtis, Weighted Unifrac)
- **Symmetrical indices consider** the double-zeros as important (PCA!)! (e.g. Euclidian without transformation)

Bray-Curtis and Hellinger distances may be better choices than Euclidean or Chi-square distances.

Legendre & Legendre (2012).

Distance matrix

Variables

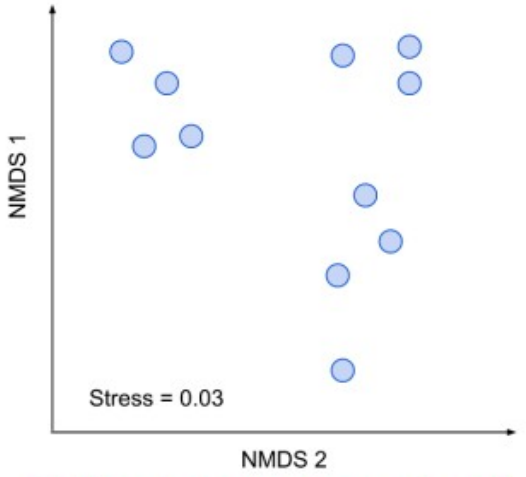
Samples	X1	X2	X3	X4
S1	14	2	14	14
S2	10	14	0	8
S3	0	5	0	2
S4	0	0	1	0

Abundance Matrix
Contingency table
OUT/ASV table

Samples

Samples	S1	S2	S3	S4
S1	0	...	...	...
S2	0.47	0	...	...
S3	0.84	0.64	0	...
S4	0.96	1	1	0

Dissimilarity/Distance matrix



Ordination plot in a reduced dimensional space

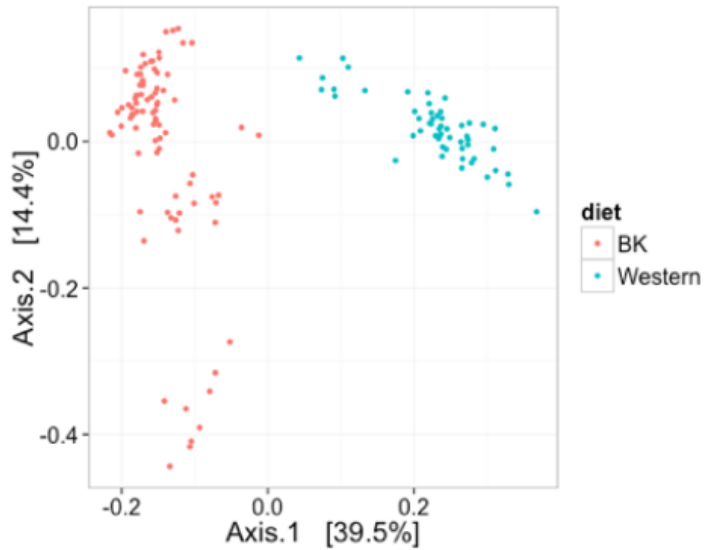
Ordination Methods

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

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Ordination Methods

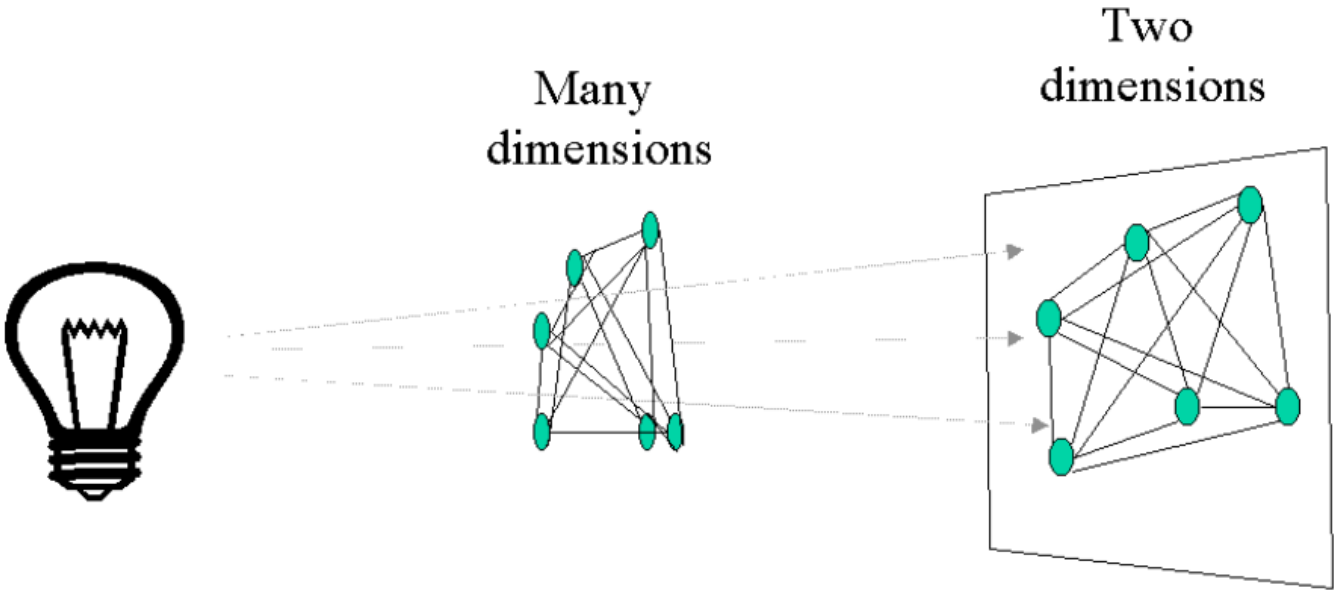
Intuition:

Each PC axis is projection that maximizes the area of the shadow
Equivalently - max(sum of square of distances between points)
Goal: “See” as much variation as possible

Slide graciously provided by Susan Holmes, not necessarily with permission O:-)

Multi-dimensional Scaling

Why MDS? It works with any distance!



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

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

MDS Details

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

In **PCA**, we start with n data points x_i, and then try to find a low dimensional projection of these points, in such a way that they minimize the reconstruction error (or **maximize the variance**).

The focus in **Multidimensional Scaling (MDS)** is somewhat different. Instead of being given the data X, our starting point is often a **matrix** of **distances** or **dissimilarities** between the data points, D.

Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

What we "learn" depends on the data.

- How many axes are probably useful?
- Are there clusters? How many?
- Are there gradients?
- Are the patterns consistent with covariates
- (e.g. sample observations)
- How might we test this?

Exploratory Data Analysis

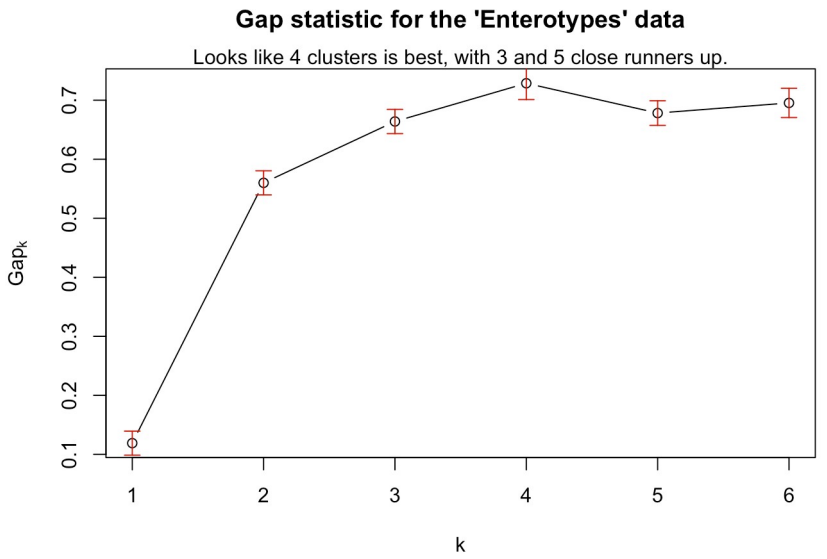
"Unsupervised Learning"

"Ordination Methods"

- Are there clusters? How many?

Technique:
Gap Statistic

Use the MDS ordination results in combinaison with the package: cluster



Exploratory Data Analysis

"Unsupervised Learning"
"Ordination Methods"

- Are the patterns consistent with covariates?

Technique:
Permutational Multivariate ANOVA
vegan::adonis( )
(note: this works with discrete and continuous variables)

Reproducible analysis of microbiome / metagenome data

- Why make the effort?
- What if I don't want someone else reproducing my analysis?
- What if I don't know how?
- Isn't it enough to provide a cursory description in the methods section with a light sprinkling of literature citations?
  - (I call this a "science poem" of your analysis)

phyloseq

OPEN ACCESS Freely available online

PLOS

phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data

Paul J. McMurdie, Susan Holmes*

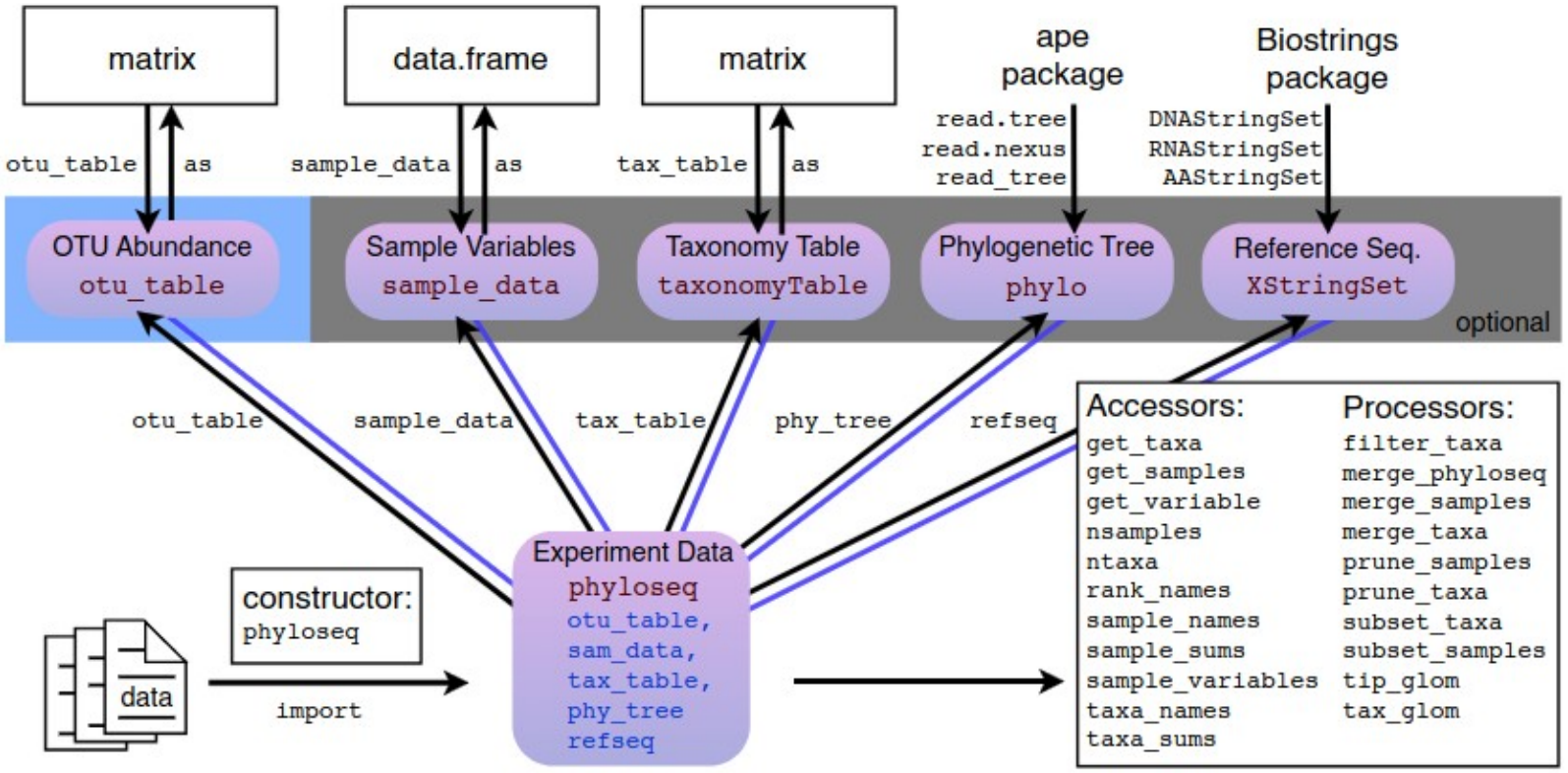
Department of Statistics, Stanford University, Stanford, California, United States of America

Key Packages:

vegan
ape
distory
phangorn
picante
metagenomeSeq
ggtree

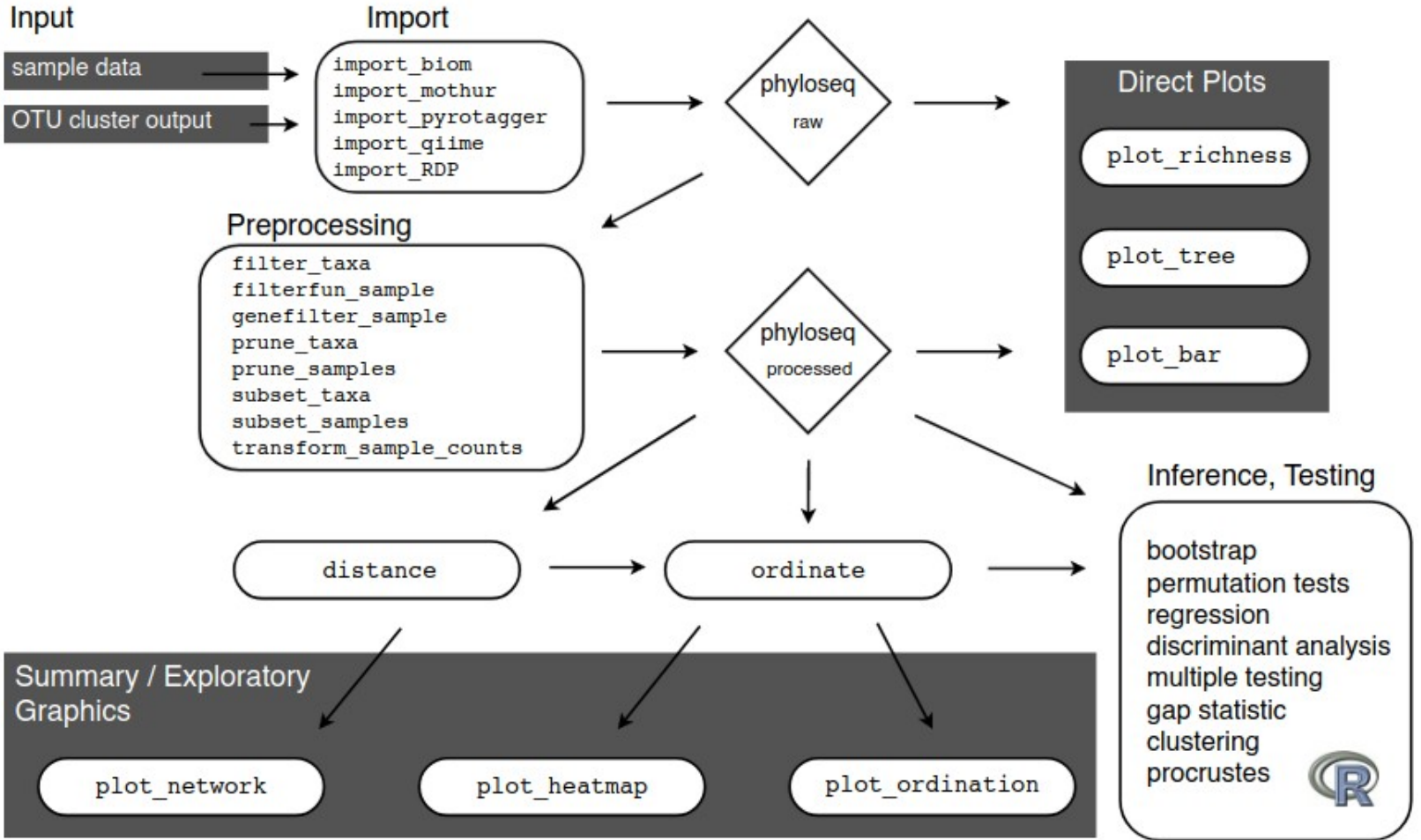
phyloseq

data structure & API



phyloseq

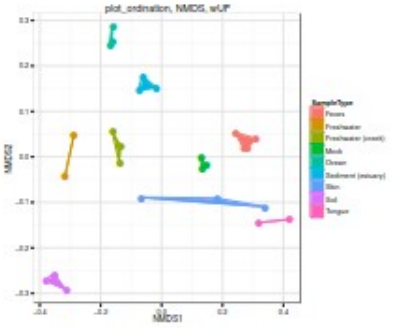
work flow

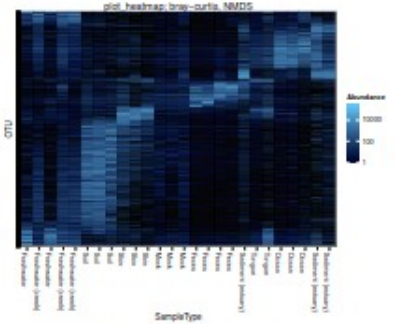


phyloseq

graphics

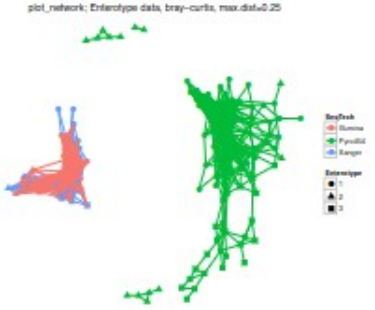
plot_ordination()

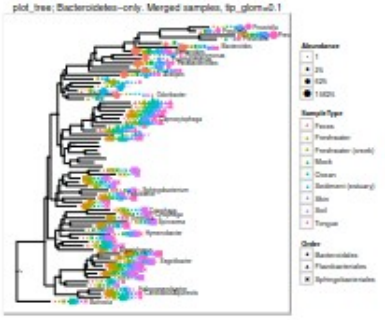




plot_heatmap()

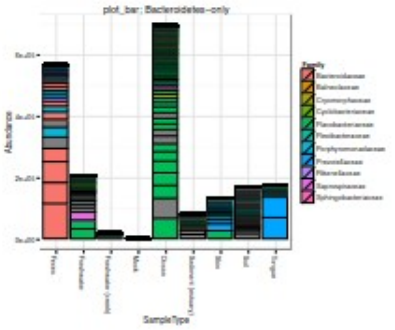
plot_network()

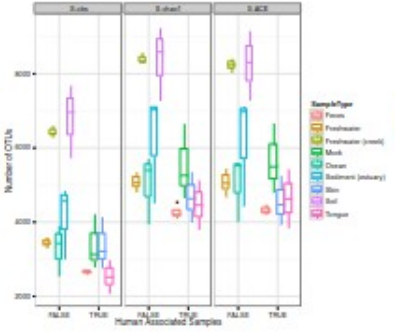




plot_tree()

plot_bar()



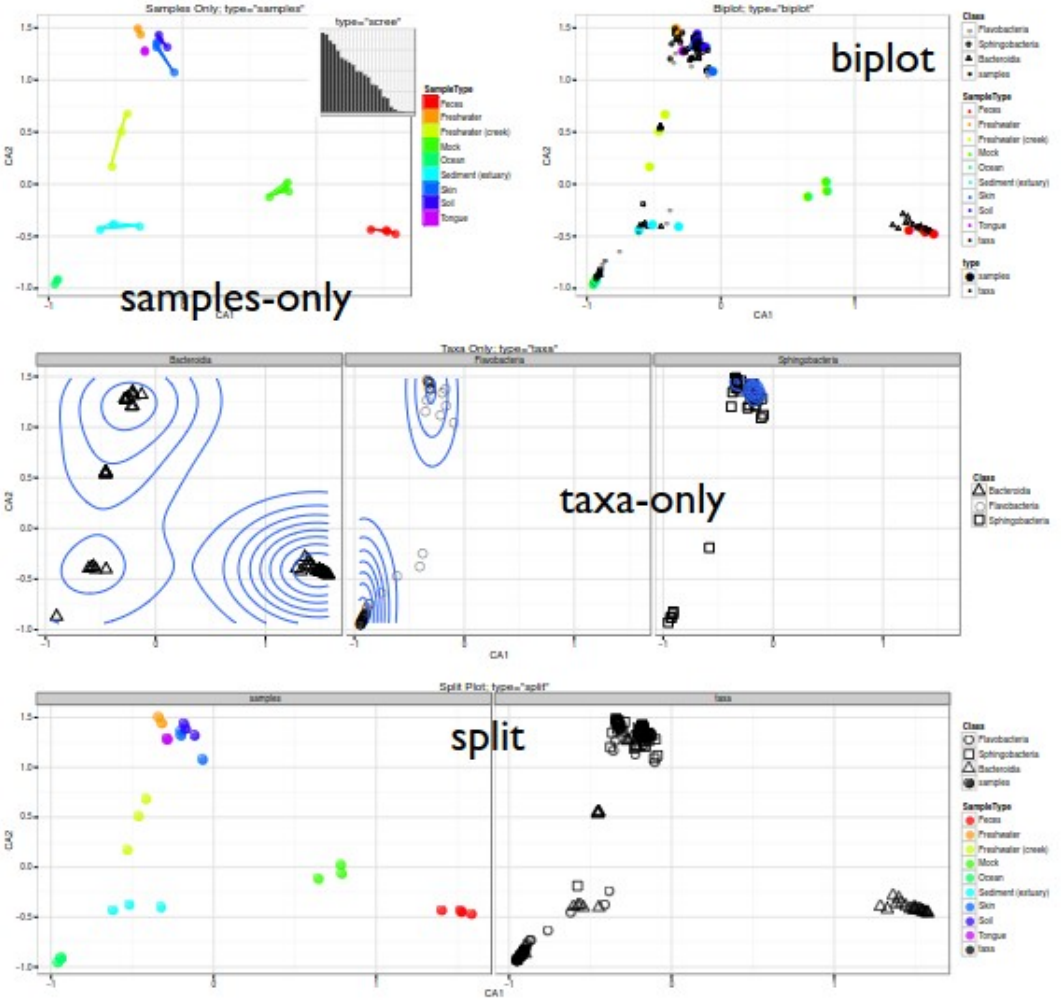


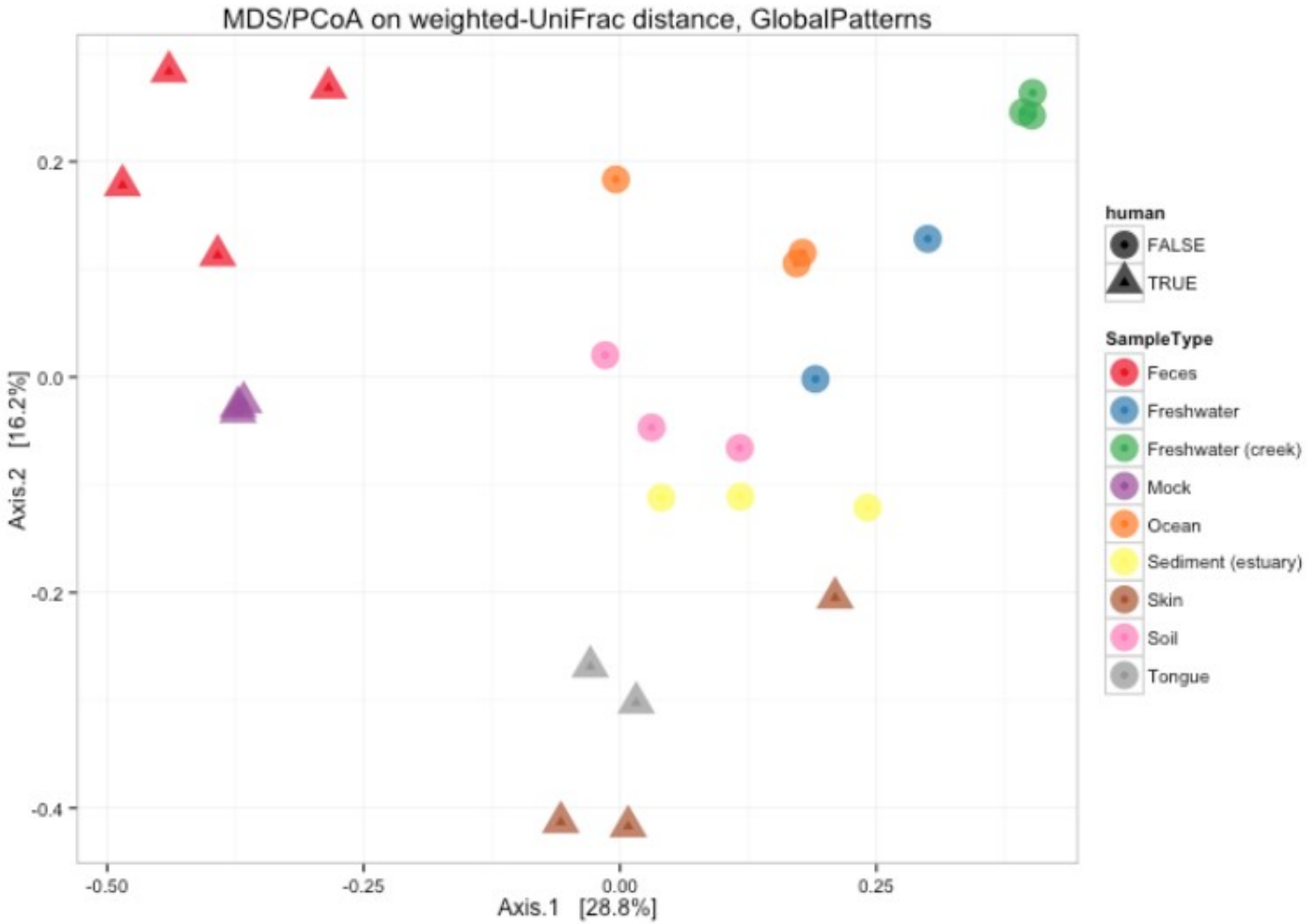
plot_richness()

phyloseq

plot_ordination()

graphics

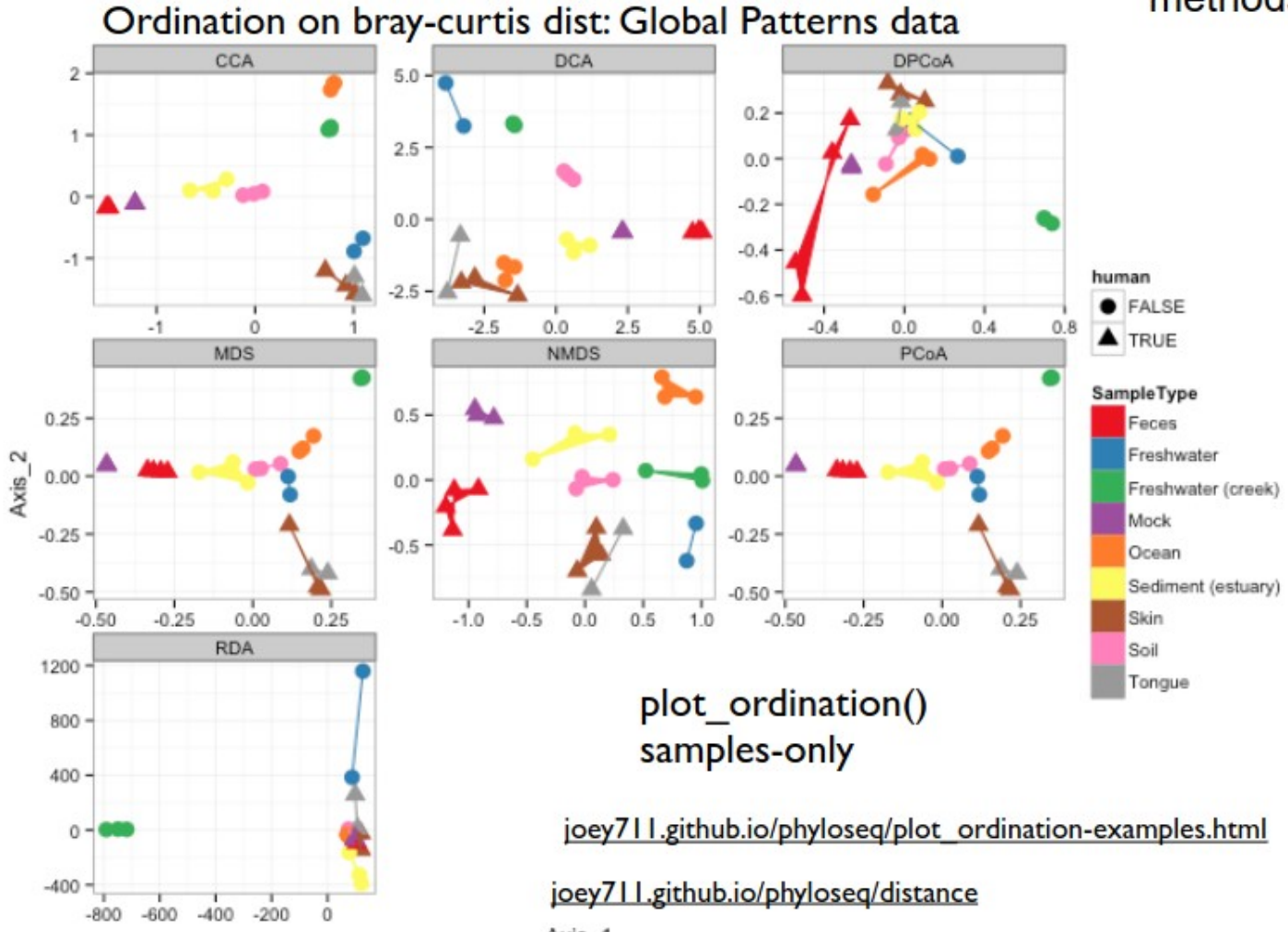





```
ordu = ordinate(GP1, "PCoA", "unifrac", weighted = TRUE)
plot_ordination(GP1, ordu, color = "SampleType", shape = "human")
```

phyloseq

supported ordination methods

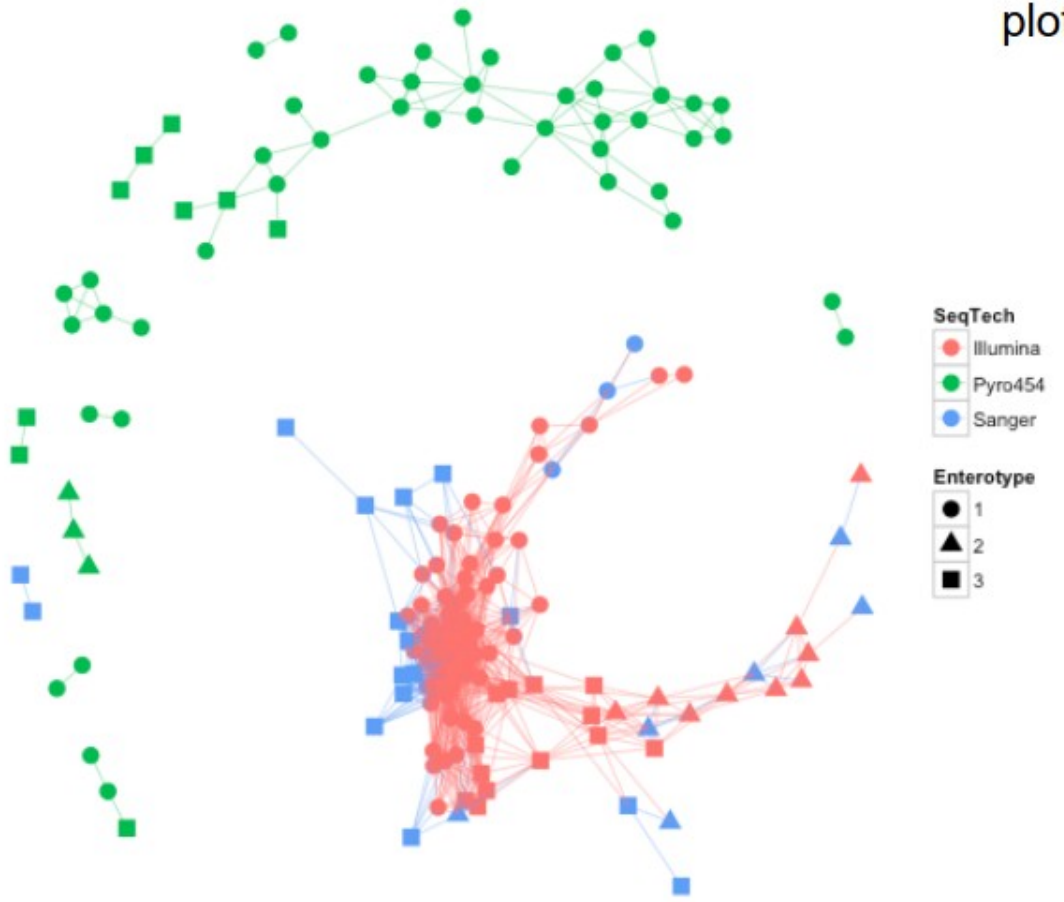


plot_ordination()
samples-only

joey711.github.io/phyloseq/plot_ordination-examples.html

joey711.github.io/phyloseq/distance

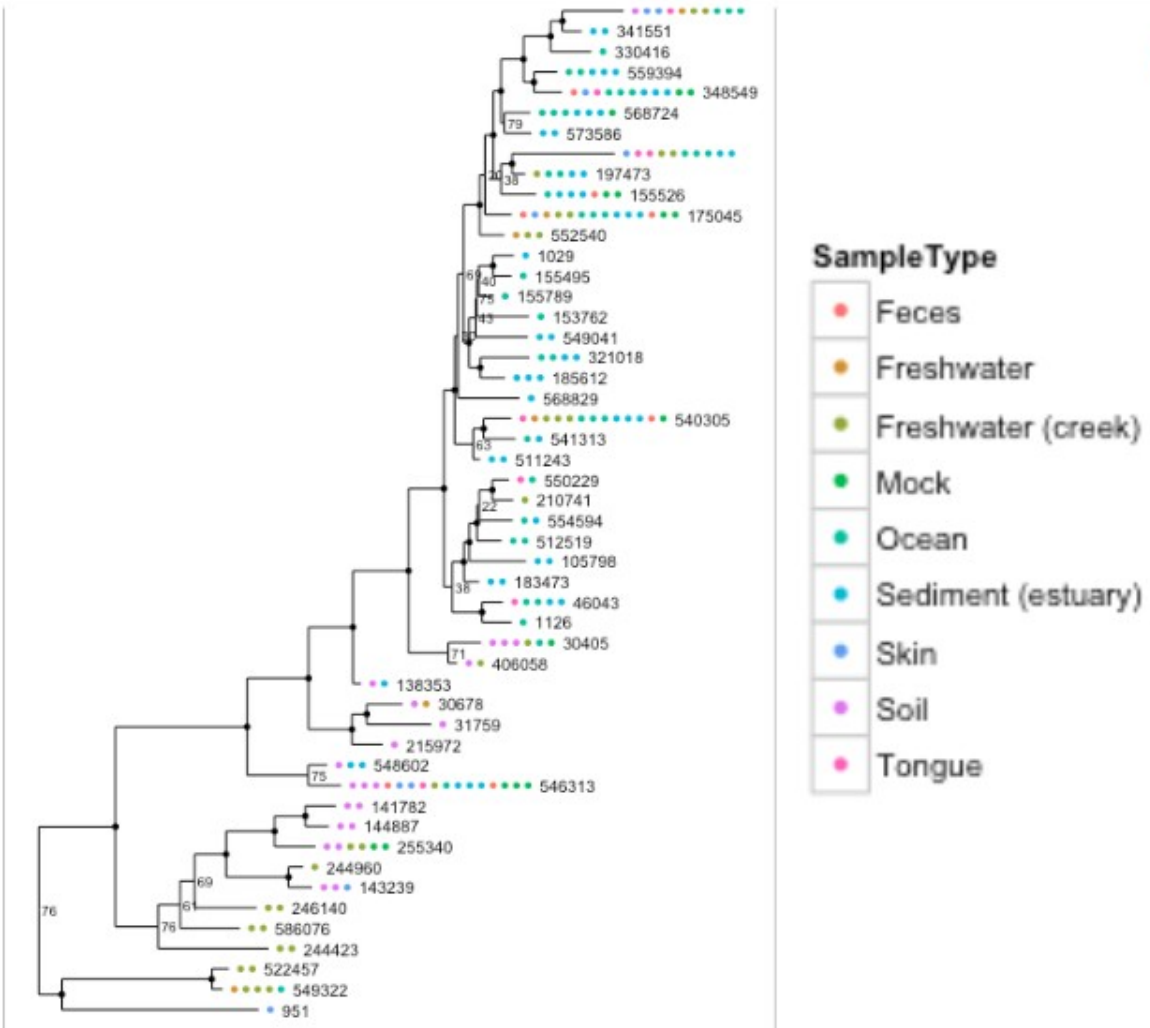
plot_network()




```
ig <- make_network(enterotype, dist.fun = "bray", max.dist = 0.3)
plot_network(ig, enterotype, color = "SeqTech", shape = "Enterotype",
    line_weight = 0.4, label = NULL)
```

joey711.github.io/phyloseq/plot_network-examples.html

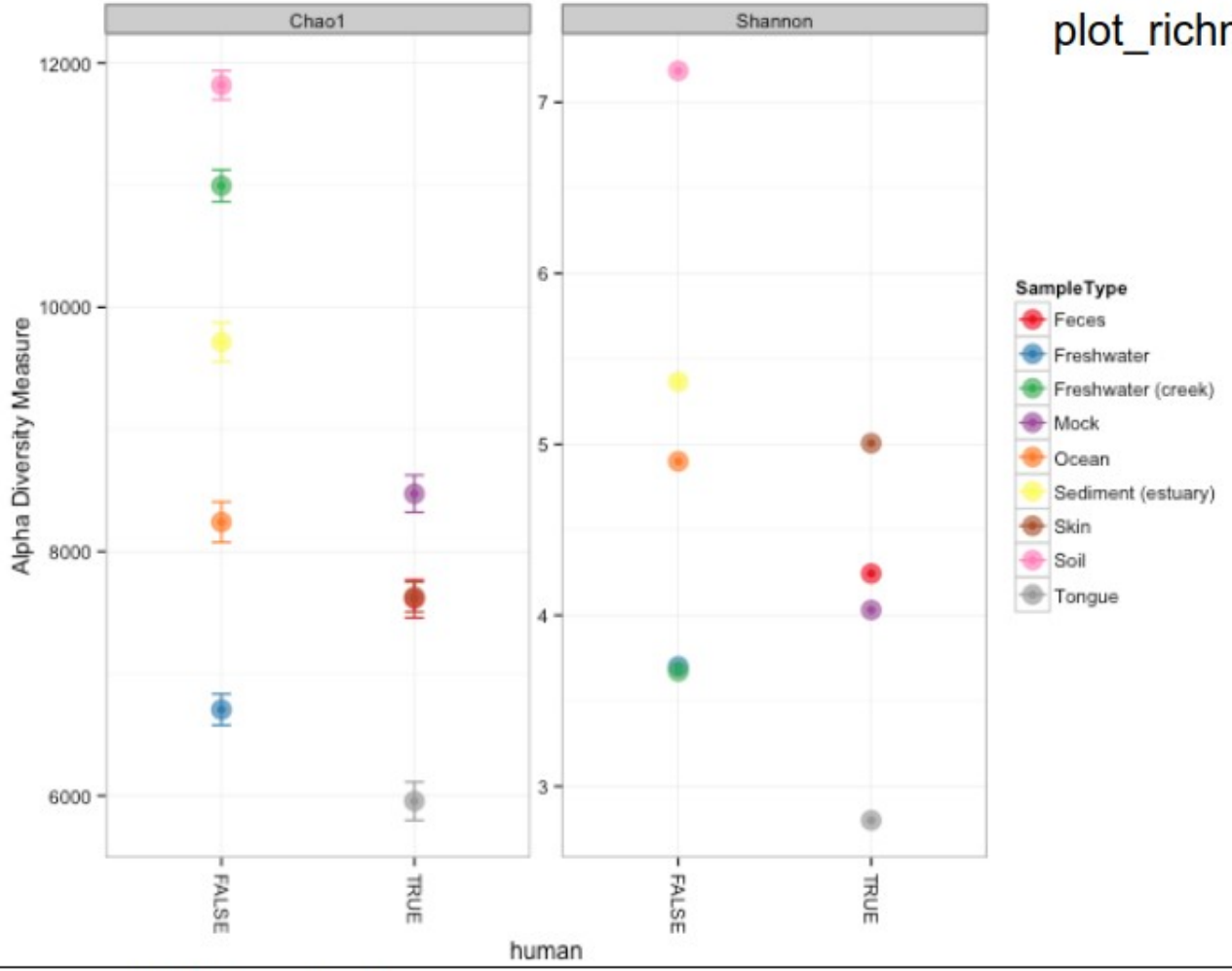
plot_tree()




```
plot_tree(physeq, nodelabf=nodeplotboot(80, 0, 3), color="SampleType",
        label.tips="taxa_names", ladderize="left")
```

joey711.github.io/phyloseq/plot_tree-examples.html

plot_richness()




```
GPst = merge_samples(GP, "SampleType")
p = plot_richness(GPst, x="human", color="SampleType", measures=c("Chao1", "Shannon"))
p + geom_point(size = 5, alpha = 0.7)
```

joey711.github.io/phyloseq/plot_richness-examples.html