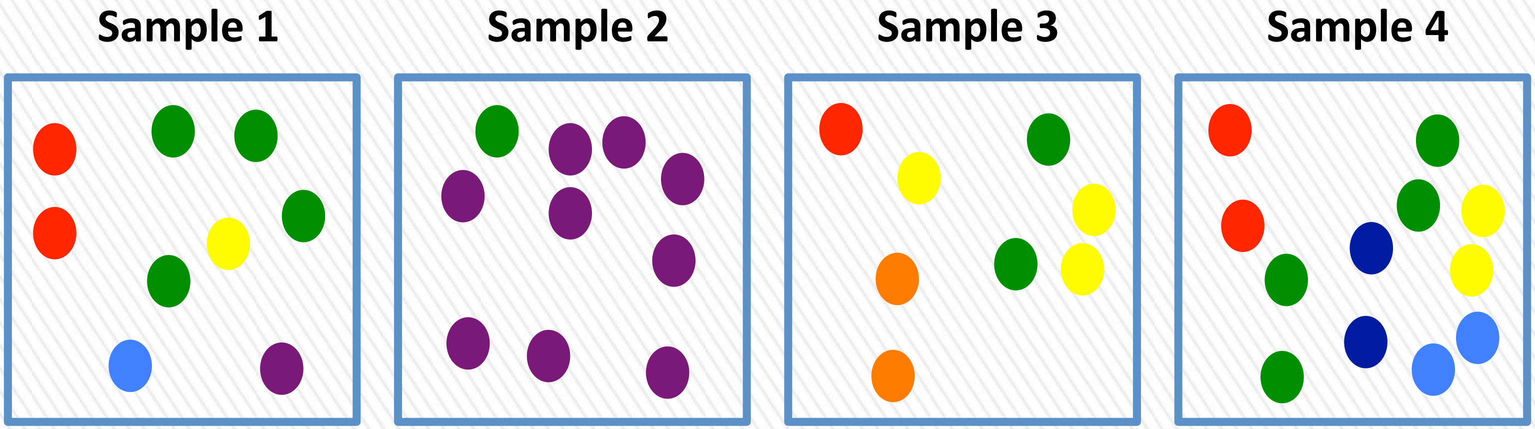
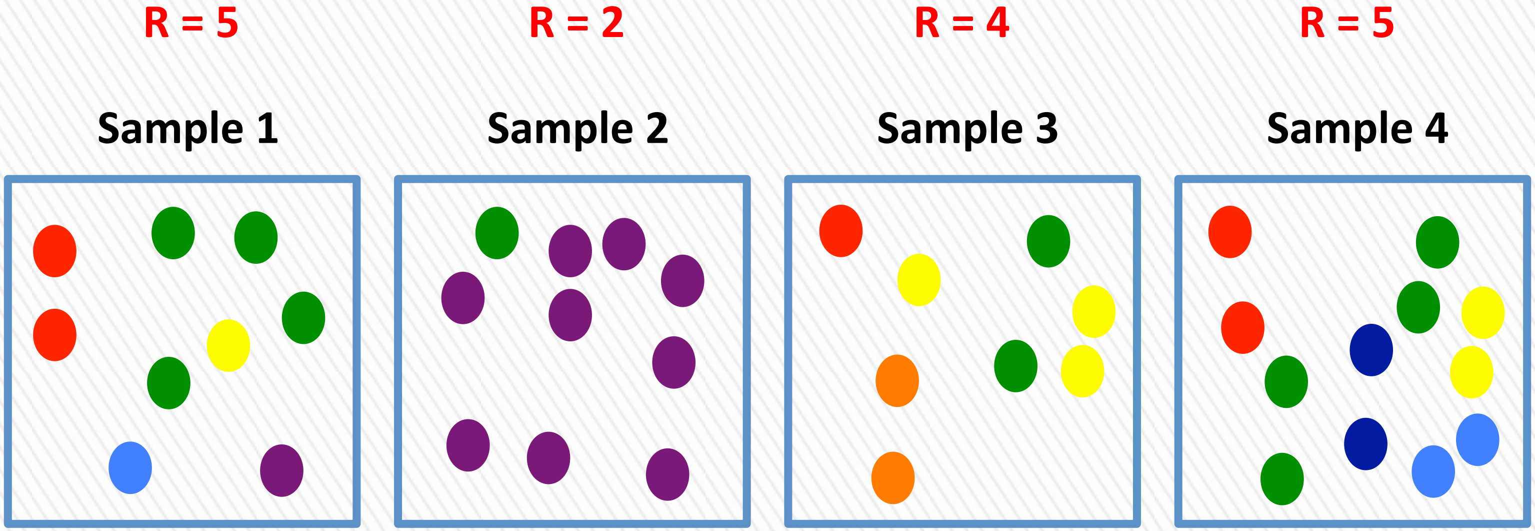
α-diversity metrics

Alpha Diversity: within sample diversity



Alpha Diversity: richness (*R*)



SPECIES RICHNESS (*S*) ESTIMATORS:

- **OTU richness** – count of different species/OTUs
- **Observed Species** – count of unique OTUs in each sample
- **Chao1 index** – estimate diversity from abundance data (importance of rare OTUs)
- **ACE index …**

Species richness indices

Let S_0 be the number of taxa observed at least once in a sample,

a_0 the unknown number of species present in the community but not observed.

1. OTU richness: R= S_0 (no correction for taxa not observed);

2. Chao-1 index: assumes that the number of observations for a taxa has a Poisson distribution and corrects for variance;

R= S_0 + a_0 *i.e.* $S_0 + a_1(a_1 - 1)/(2a_2 + 1)$

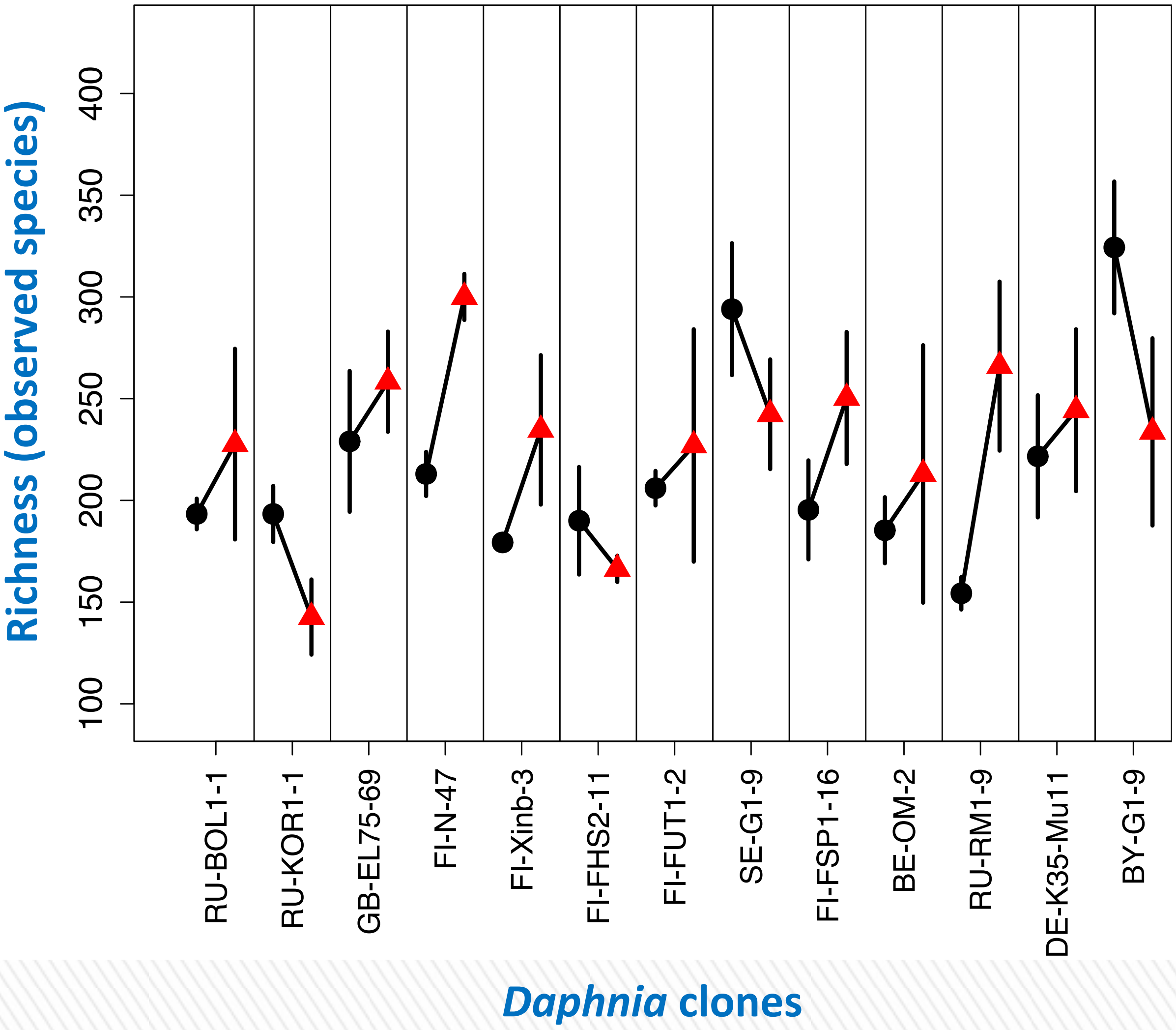
3. ACE (abundance-based coverage estimators): involves an arbitrary abundance threshold to label S_{abun} as the number of abundant taxa, S_{rare} as the number of rare taxa; The expression basically inflates the number of rare taxa and inflates again the number of taxa with abundance 1.

R= S_0 + a_0 *i.e.*

$$S_{\mathrm{abun}} + S_{\mathrm{rare}}/C_{\mathrm{ace}} + a_1/C_{\mathrm{ace}} * \gamma^2$$

$$\gamma^2 = \max S_{\mathrm{rare}}/C_{\mathrm{ace}} \sum_{i=1}^{10} (i(i-1)a_i/(N_{\mathrm{rare}}(N_{\mathrm{rare}} - 1)) - 1), 0$$

Species richness: example of use

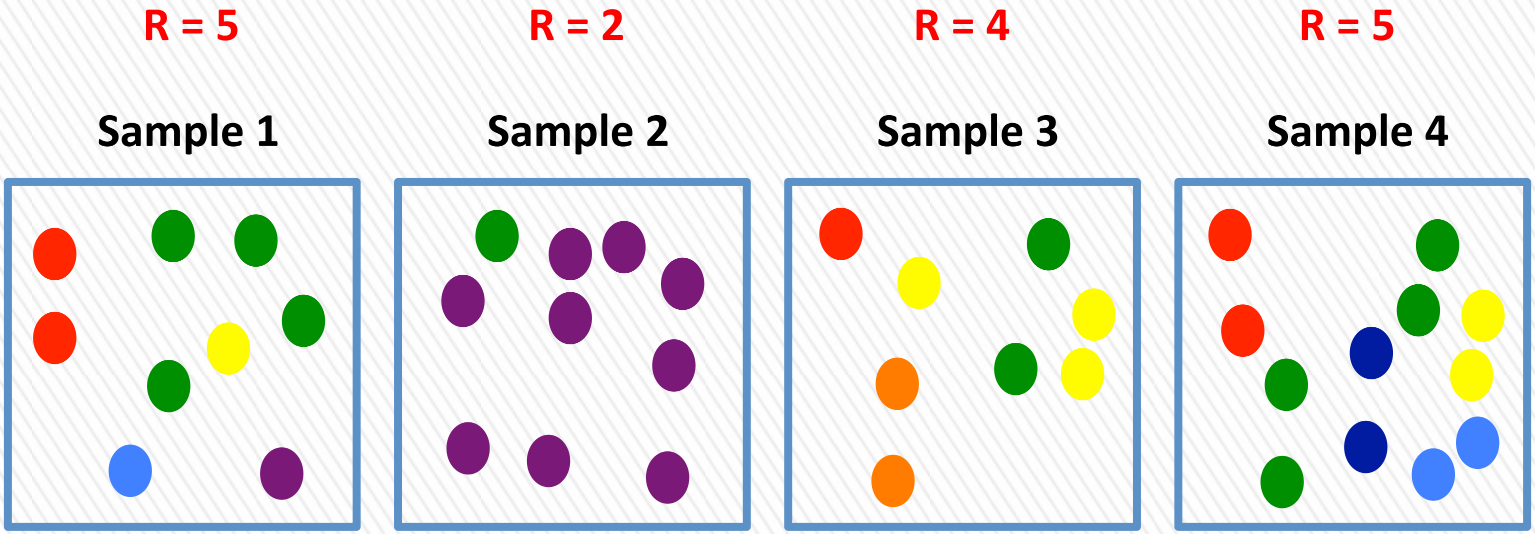


Influence of temperature on *Daphnia* microbiota:

● **20°C**

▲ **28°C**

Alpha Diversity: within sample diversity



R = 5 R = 2 R = 4 R = 5

Sample 1 Sample 2 Sample 3 Sample 4

SPECIES RICHNESS ESTIMATORS

La valeur calculée de la diversité augmente à la fois lorsque le nombre d'espèces augmente et lorsque l'équité (uniformité de répartition) augmente.

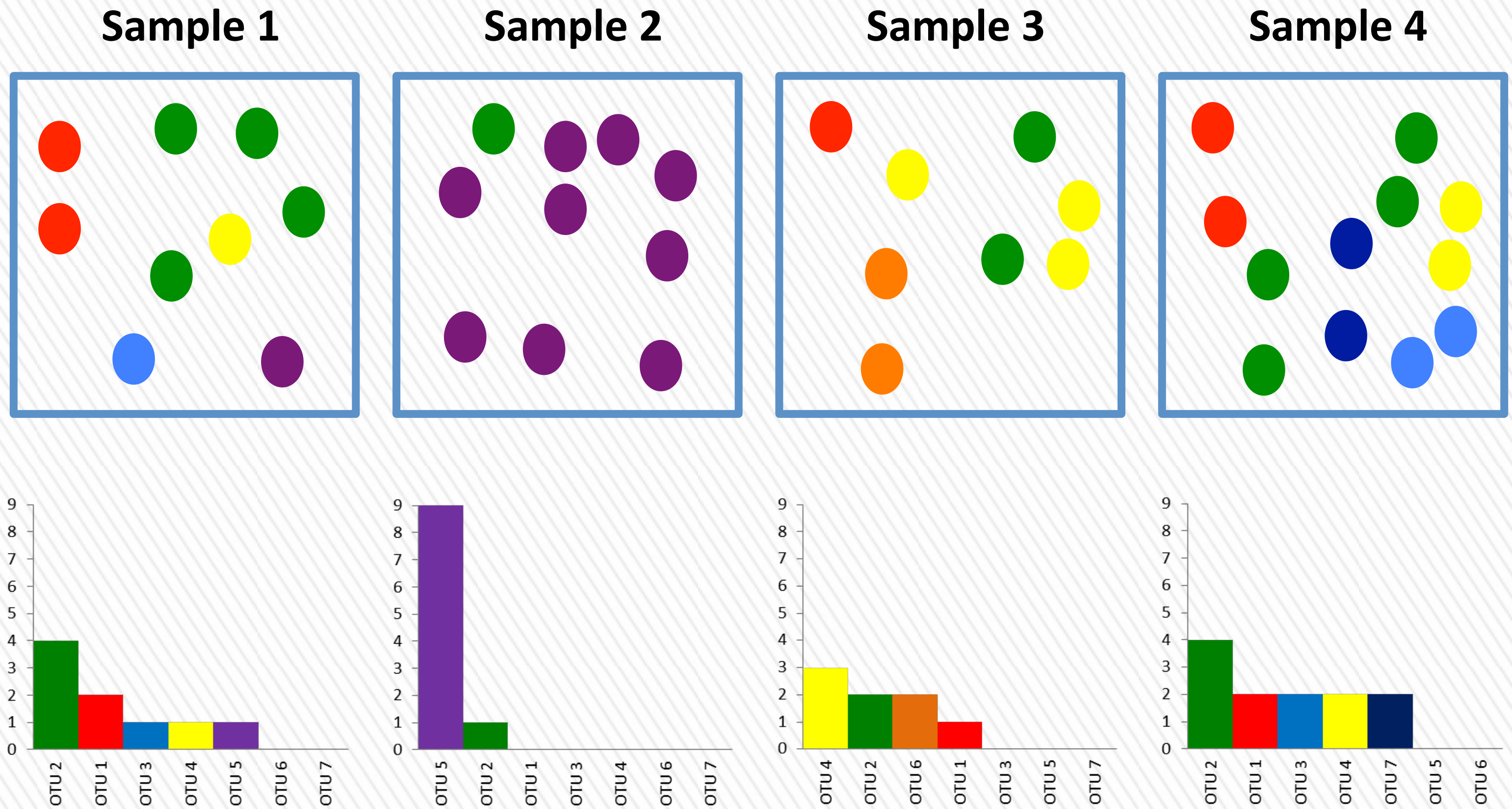
EVENESS c'est équité

RICHNESS and EVENNESS ESTIMATORS: the calculated value of diversity increases both when the number of species increases and when evenness increases.

- **Information statistics**: Shannon-Wiener, Shannon-Weaver, Shannon entropy
- **Dominance indices**: Inverse Simpson, Gini–Simpson, Berger–Parker index

Alpha Diversity: relative abundances

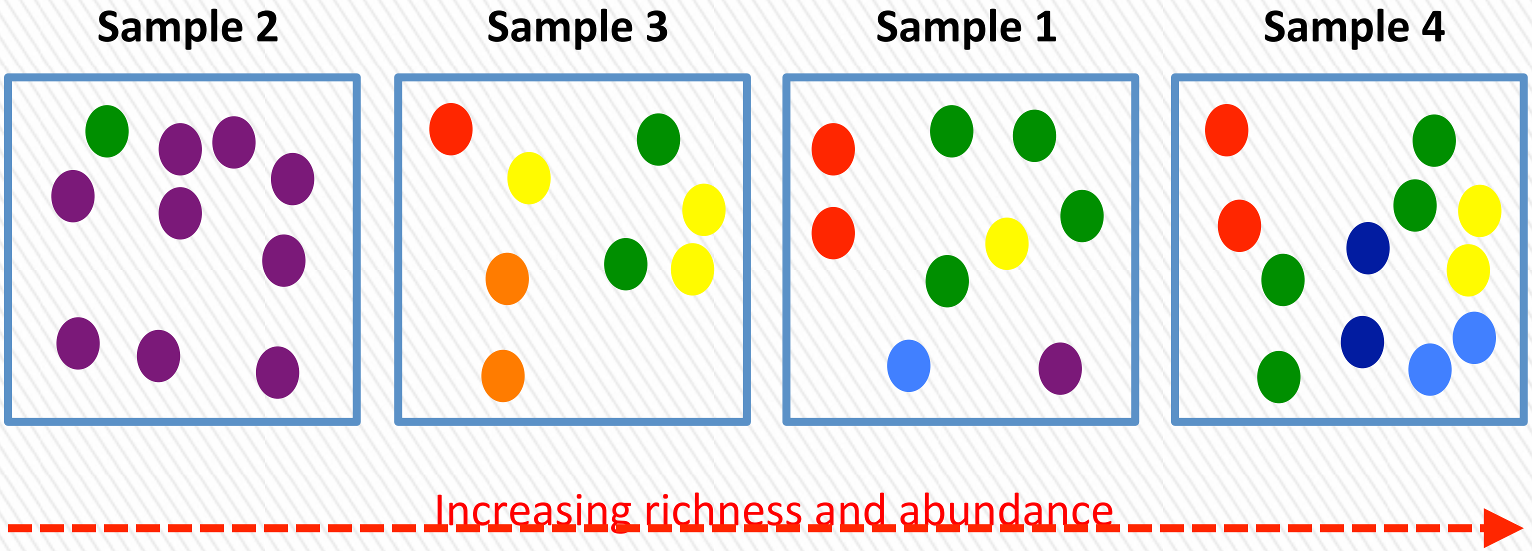
Refers to how common or rare a species is relative to other species in a community.

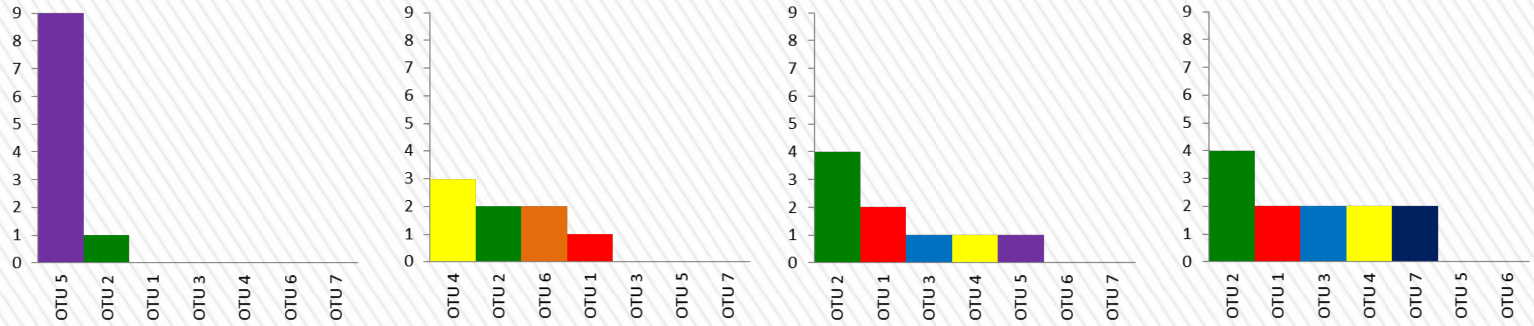


Relative species abundance distributions are graphed as rank-abundance diagrams.

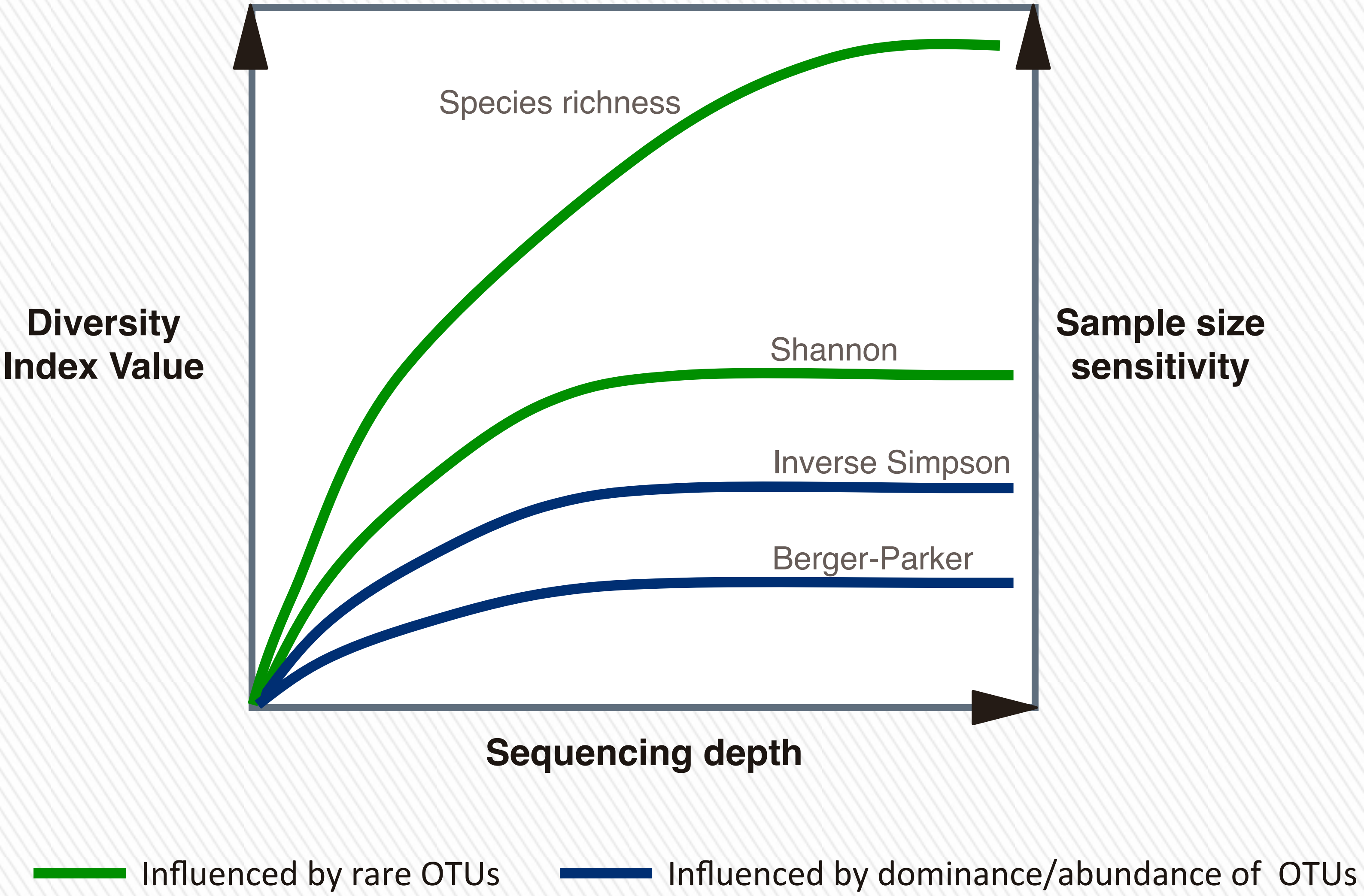
Alpha Diversity: species evenness

Species evenness refers to how equally abundant species in an environment are.

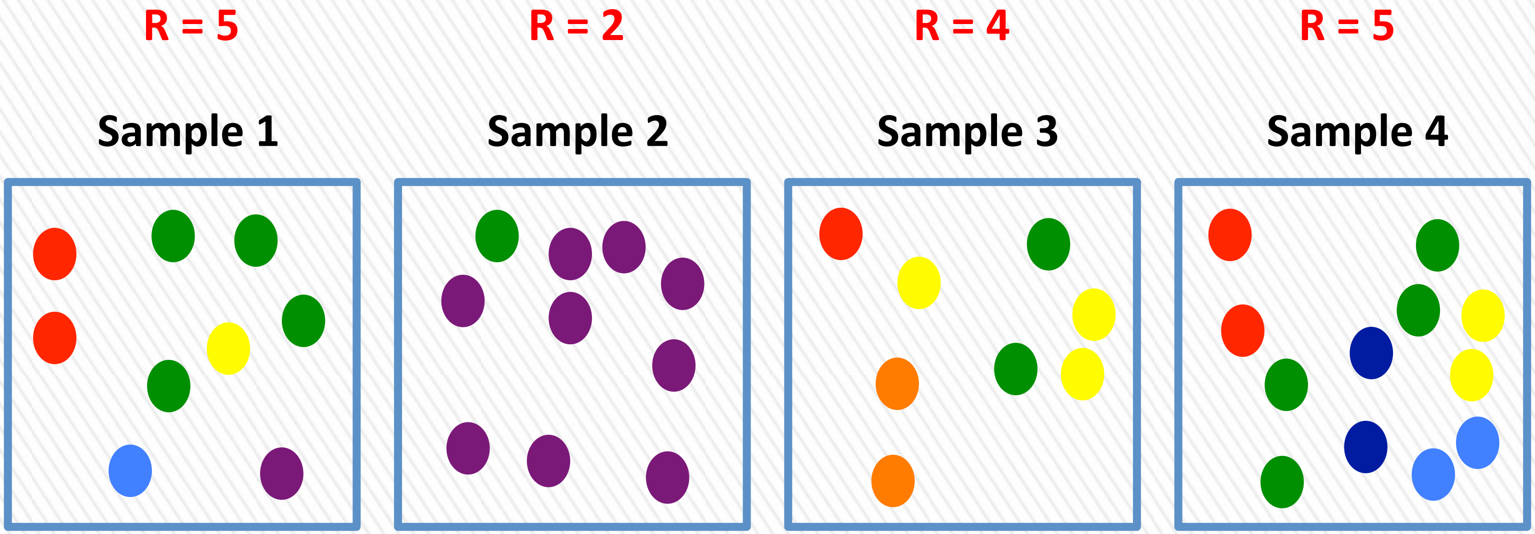




Alpha Diversity: indices’ performance



Alpha Diversity: phylogenetic diversity



R = 5 R = 2 R = 4 R = 5

Sample 1 Sample 2 Sample 3 Sample 4

SPECIES RICHNESS ESTIMATORS

RICHNESS and EVENNESS ESTIMATORS

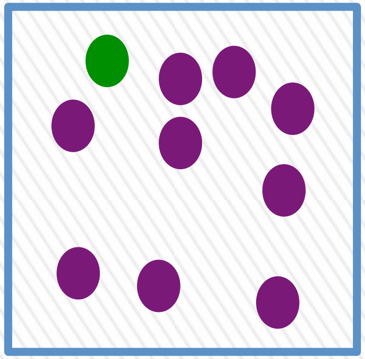
PHYLOGENETIC RICHNESS ESTIMATOR:

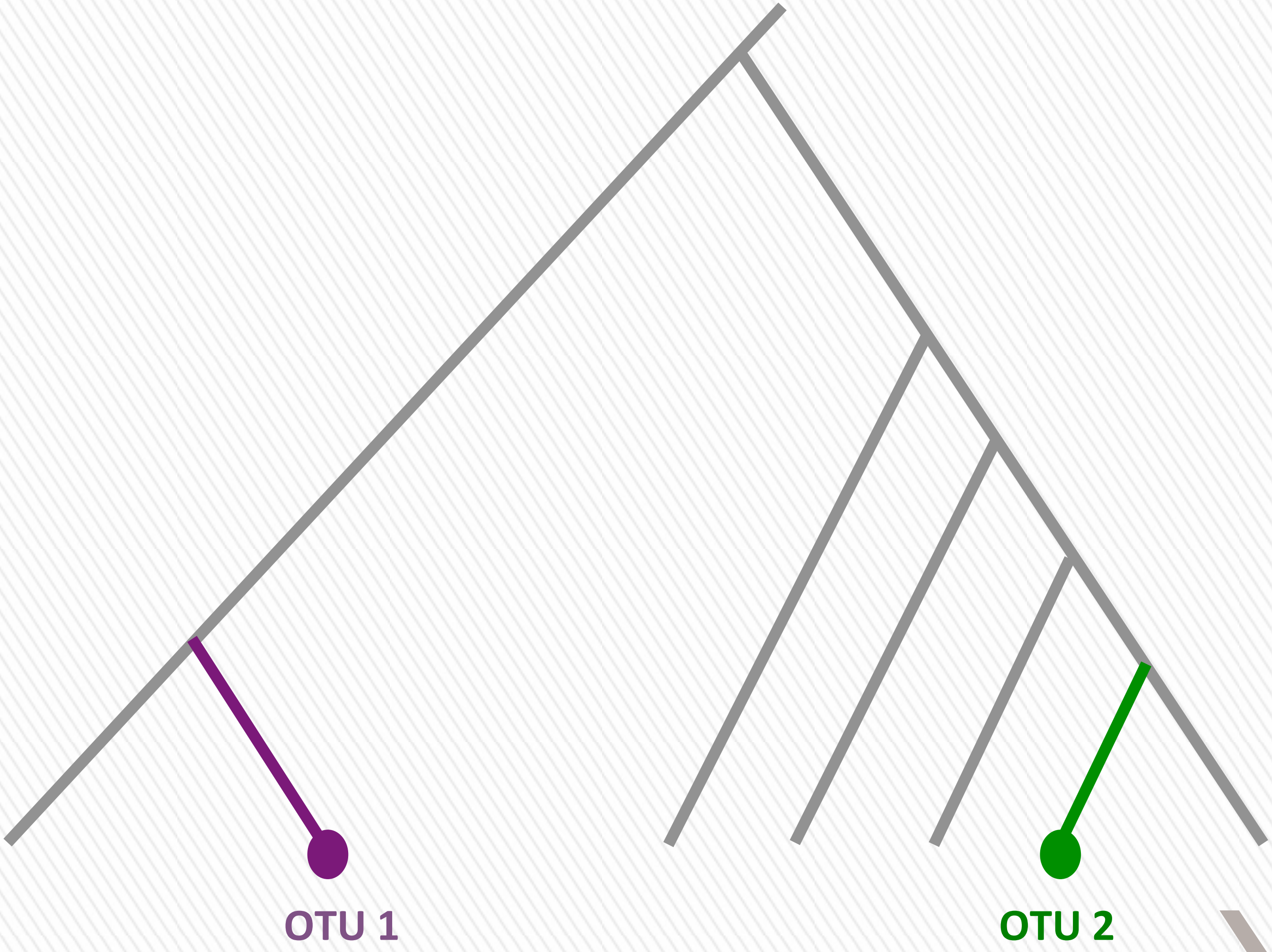
- **Phylogenetic diversity** (PD) – takes into consideration the phylogeny of microbes to estimate diversity across a tree

Alpha Diversity: phylogenetic diversity

R = 2

Sample 2

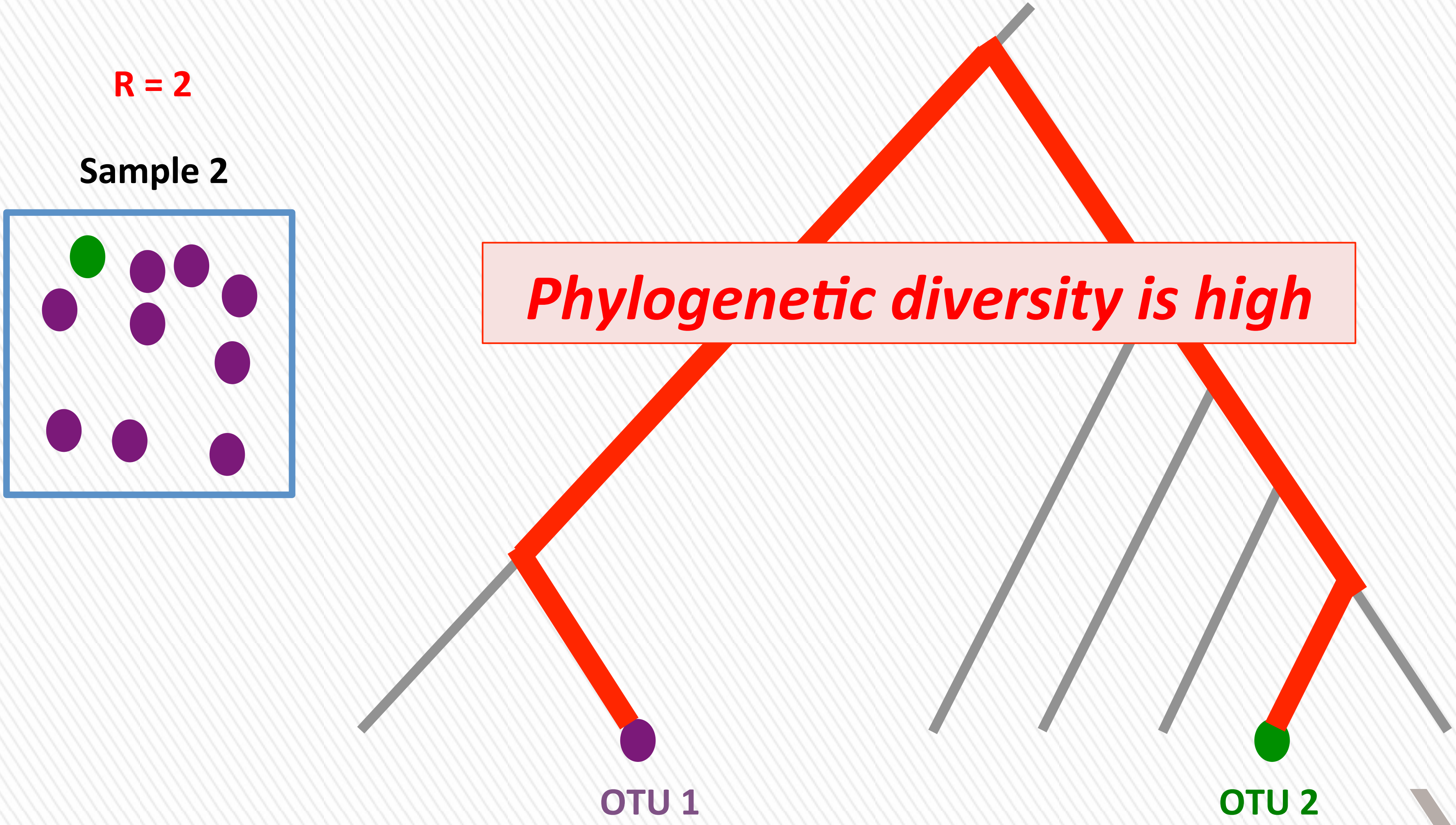




Alpha Diversity: phylogenetic diversity

R = 2

Sample 2

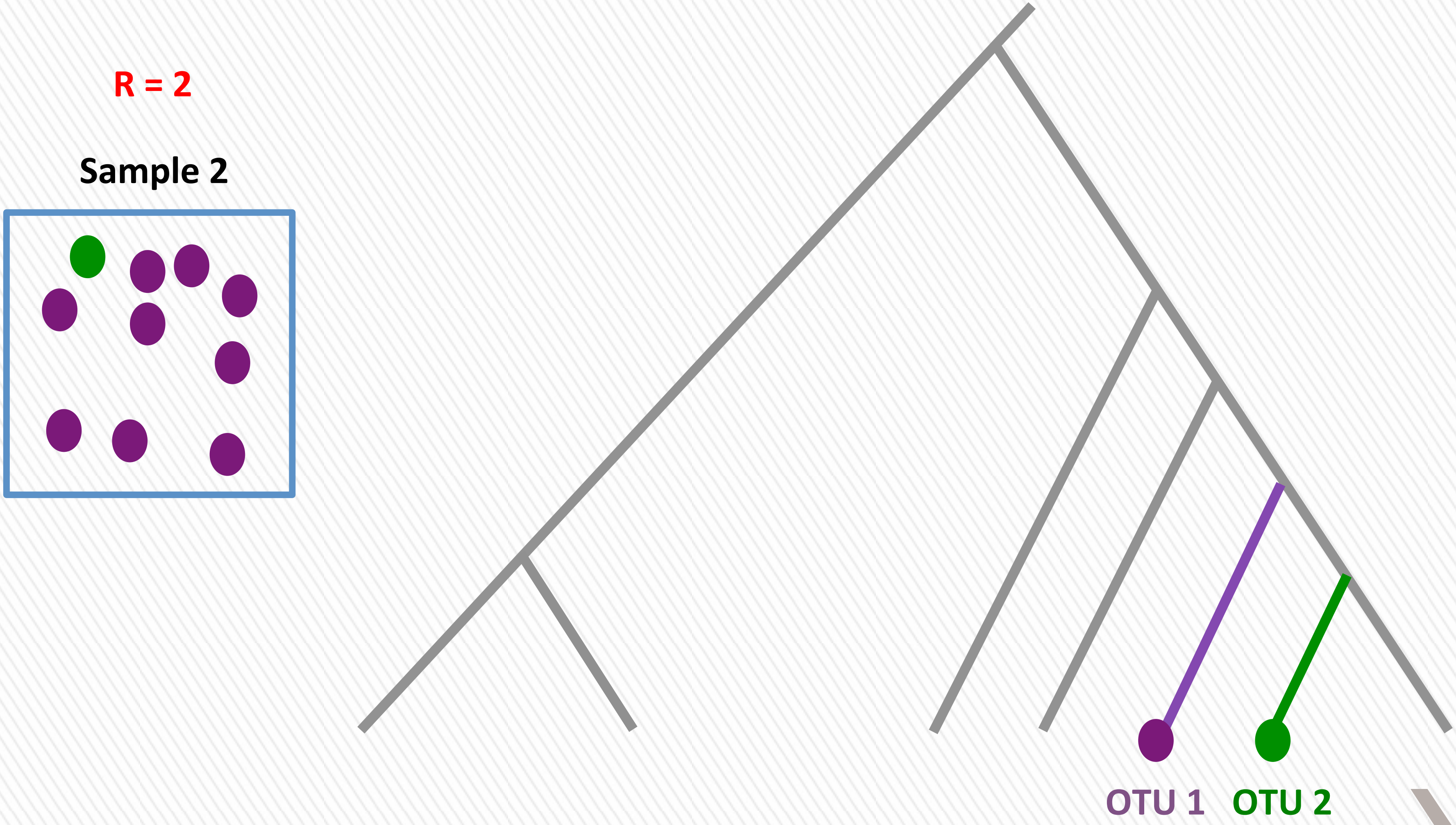


Phylogenetic diversity is high

Alpha Diversity: phylogenetic diversity

R = 2

Sample 2

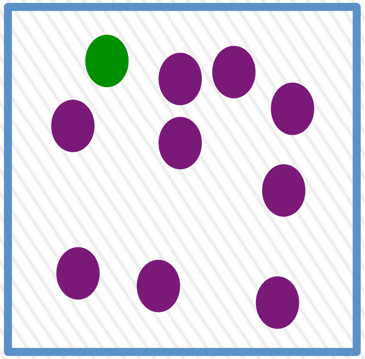


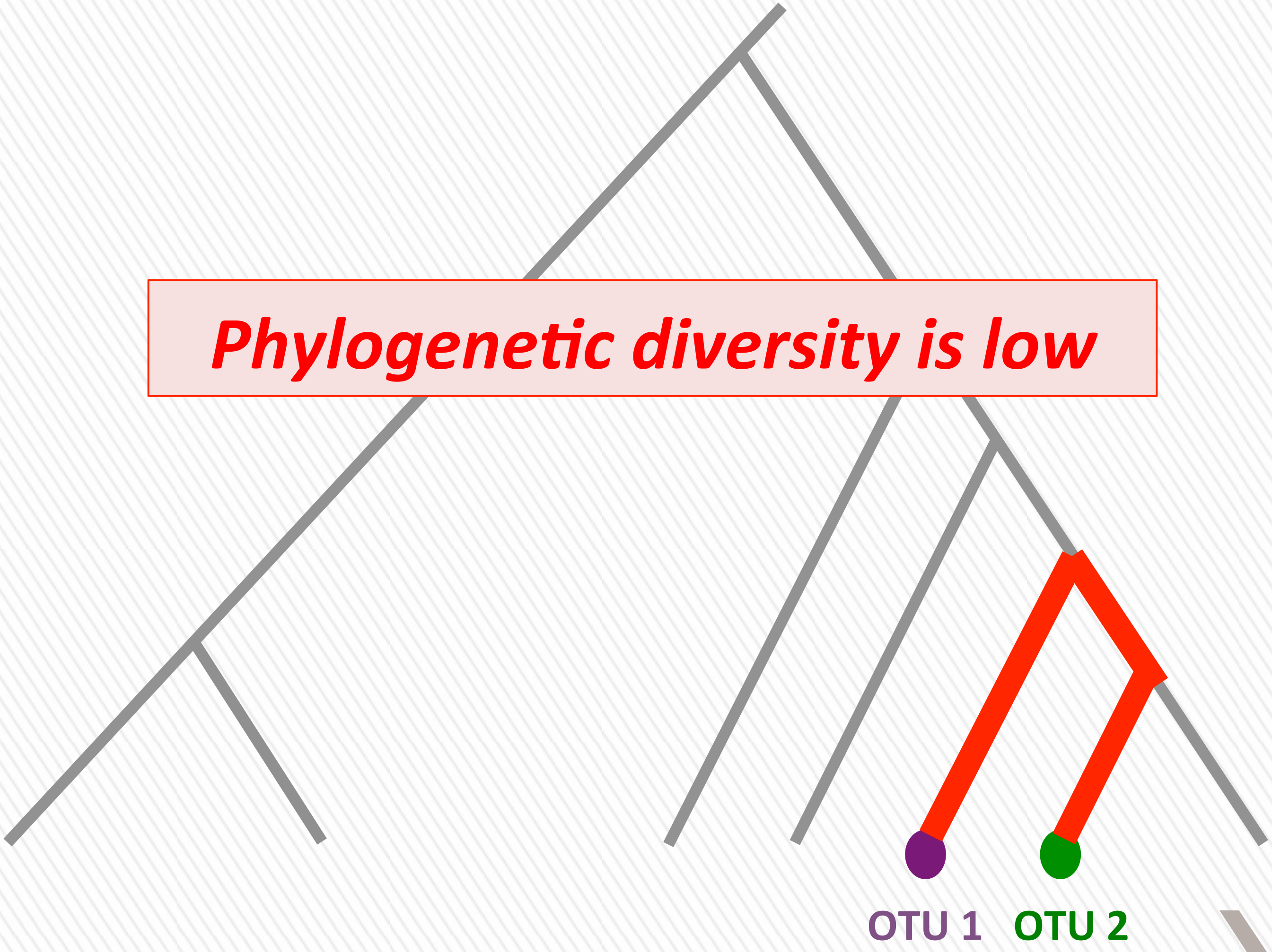
OTU 1 **OTU 2**

Alpha Diversity: phylogenetic diversity

R = 2

Sample 2





Phylogenetic diversity is low

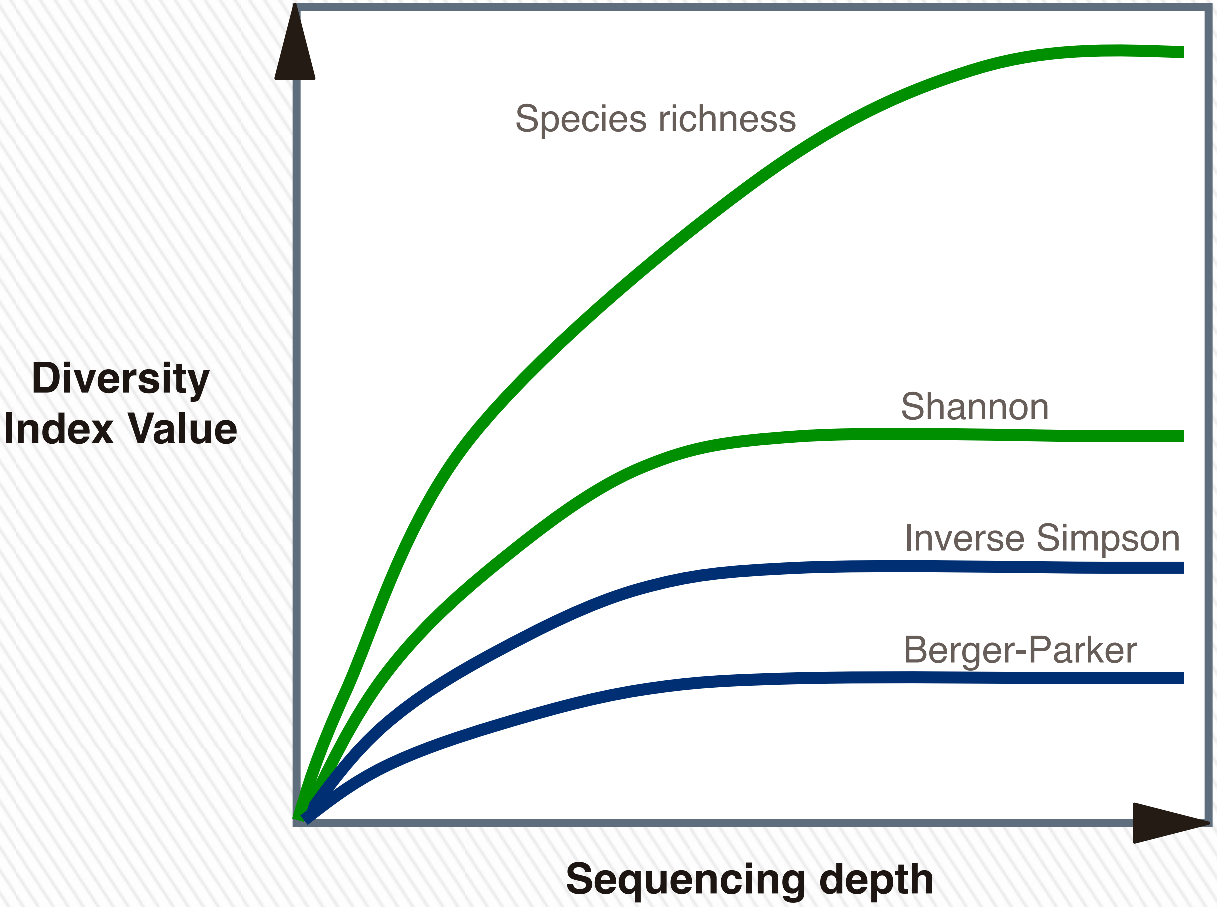
Your questions

- What you want to know determines how you analyze your data

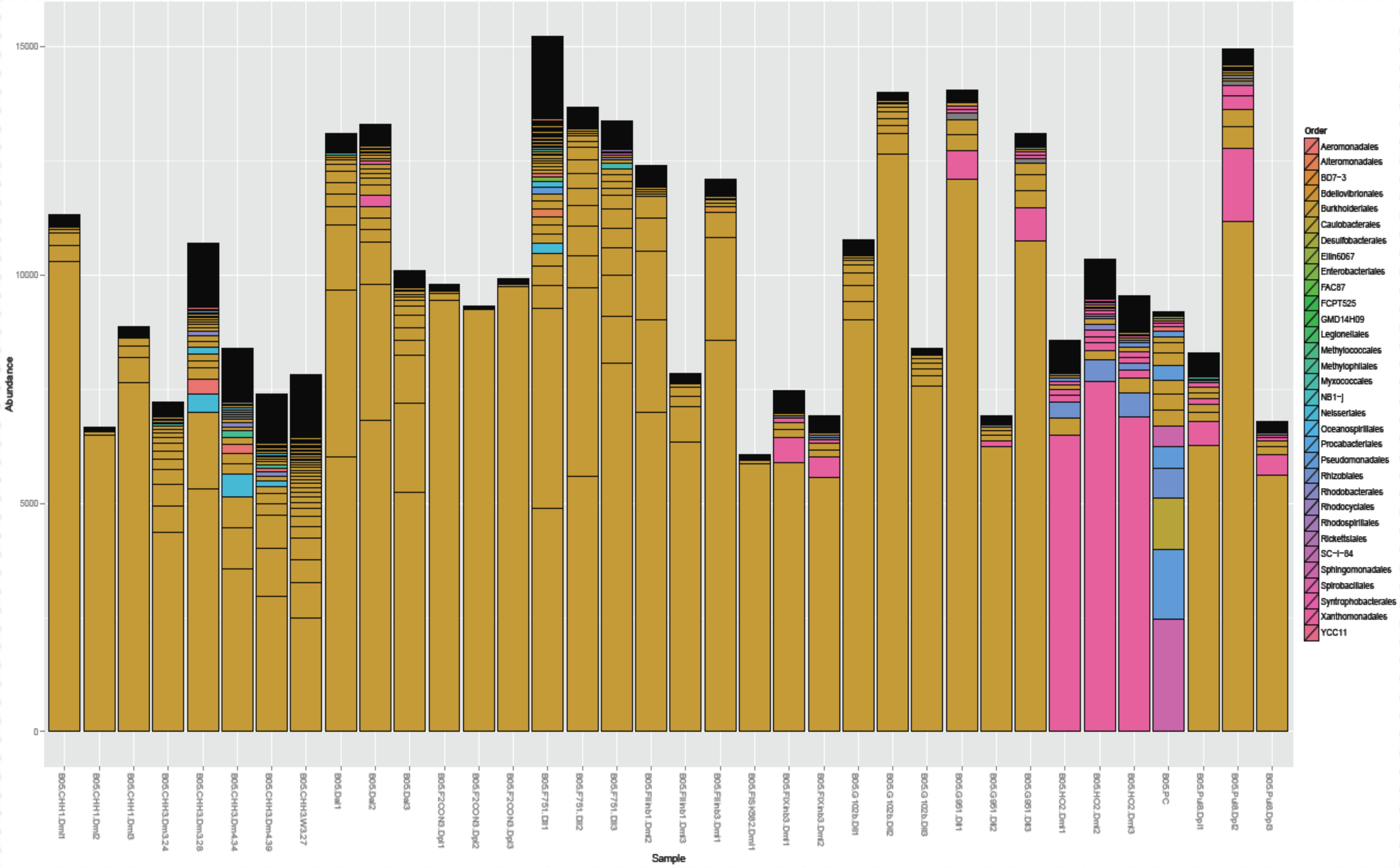
- How important is each aspect of diversity?
  - Richness?
  - Evenness?
  - Dominance?
  - Abundance?
  - Per-species (relative) abundance?
  - Taxon diversity?

Variation among samples

Diversity: influence of sequencing effort



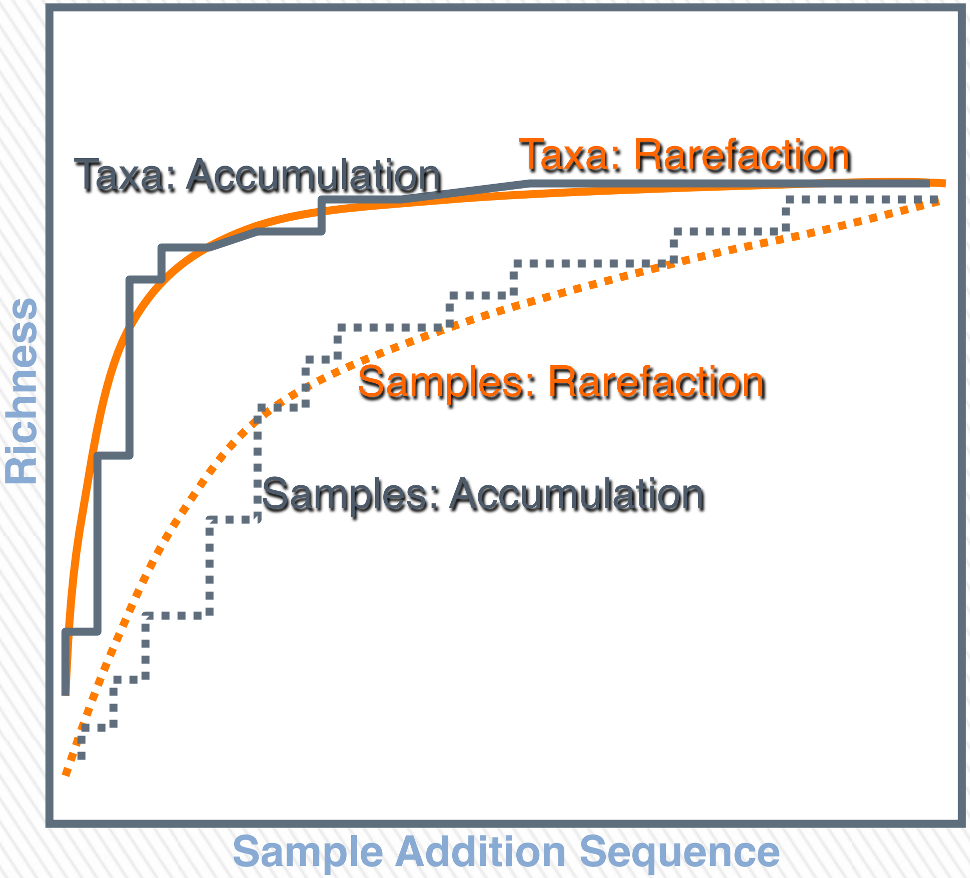
Variation in reads counts



Rarefaction

Rarefying was first recommended for microbiome counts in order to moderate differences in the presence of rare OTUs

(Lozupone *et al.* 2011 ISME J)



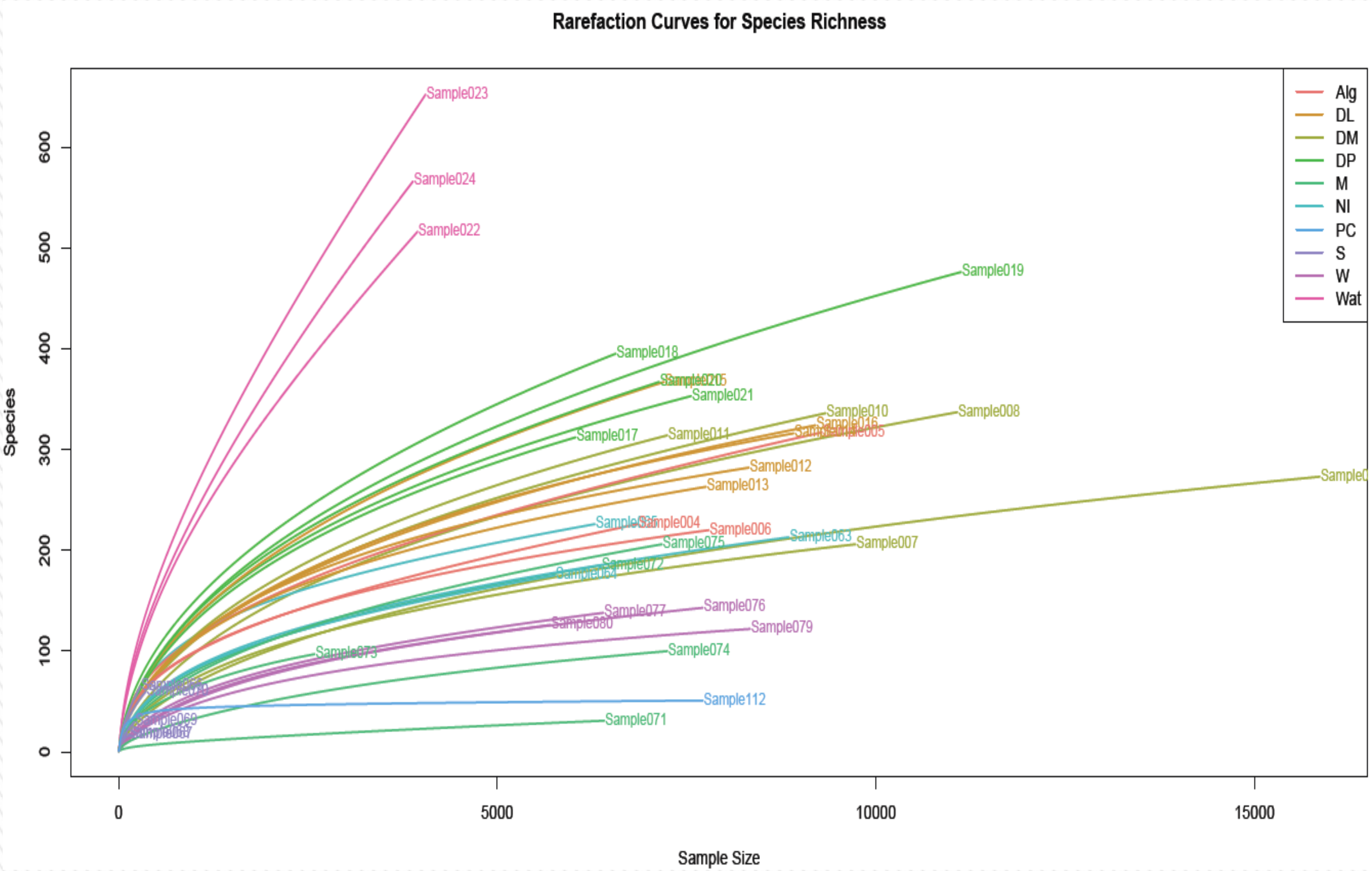
Goal:

- Standardize unequal sequencing effort.
- Enable similarity comparisons along a range of samples or a gradient.
- Enable comparison of different runs or replicates.

Procedure:

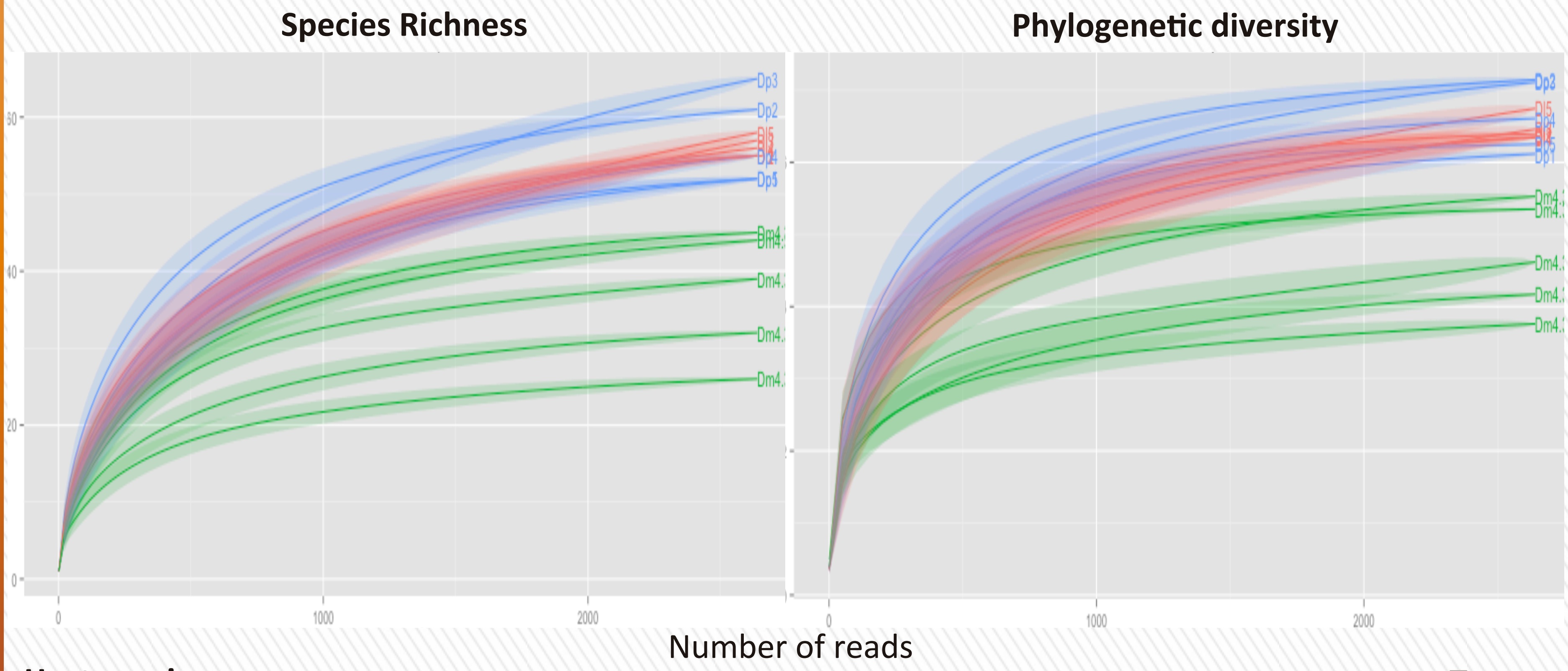
- Determine the minimum sequencing depth.
- Subsample without replacement sequences from the larger libraries so that all have the same smallest size.
- Note that the term is a bit misleading as this step should really be called "subsampling to a given depth".

Rarefaction curves



Rarefaction curves

Rarefaction curves represent the diversity as a function of sequencing depth.


Species Richness

Phylogenetic diversity

Number of reads

Host species:

- 🟦 *D. pulex*
- 🟥 *D. longispina*
- 🟩 *D. magna*

Rarefaction

This approach simultaneously addresses problems when

(1) DNA sequencing libraries are of widely different sizes (also loss of information in the larger sample)

(2) OTU count proportions vary more than expected under a Poisson model.

(3) power and accuracy is too low in the detection of differentially abundant OTUs

NB: A species/OTU is considered differentially abundant if its mean proportion is significantly different between two or more sample classes in the experimental design.

"In the case of differential abundance detection, it seems unlikely that the cost of rarefying is ever acceptable."

McMurdie 2014 *PLoS Comput Bio* 10-e1003531

Alternatives to rarefaction

All that is left after rarefaction is the expected number of species per sample, not a real value or real data.

Distinction between subsampling curves and normalization.

Randomly select evenly-sized samples from the larger sample

- Could be done iteratively to provide a normalized distribution of the expected number of species
- Akin to the Jackknife value

Kempton & Wedderburn (1978)

- Produce equal sized samples after fitting species abundances to gamma distribution
- Not commonly used

Procedure: use normalization tools or transform count data (*e.g.* use a log2(x + 1) transformation on count data to mitigate the impact of 0 and very high counts).

Alternatives to rarefaction

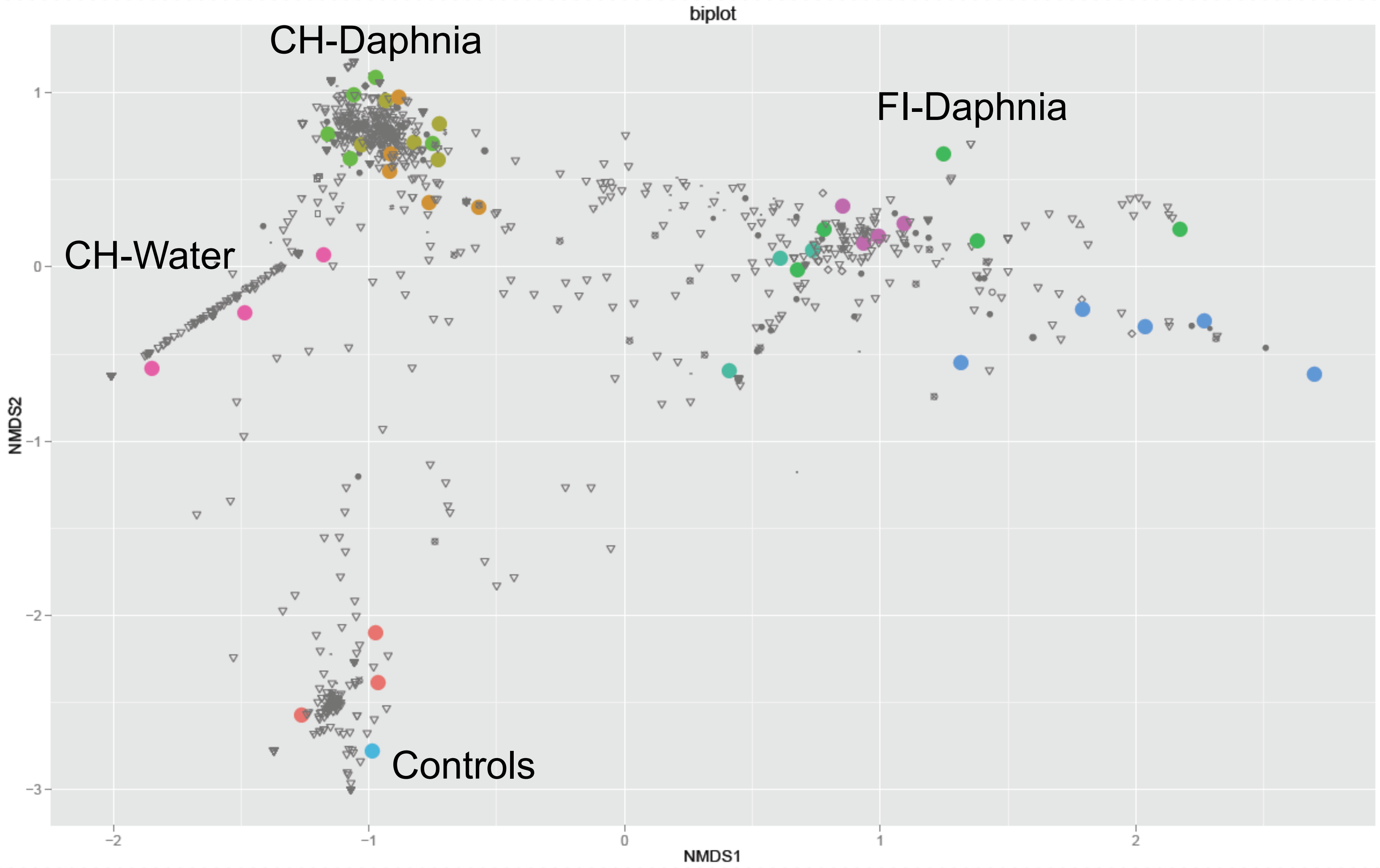
Normalization

- Equalize depths by simply scaling OTU counts to a common depth in all samples
- Transform counts into relative abundances for each sample. This lead the counts not to be integers any more. each number in the OTU table will represent the proportion of sequences from that samples belonging to that OTU.
- Note that the methods specifically tailored to integers will not apply but it will not change methods based on presence/absence or proportions, such as UniFrac, Bray-Curtis, *etc.* . .
- Other normalization: normalize data based on 16S copy number.
- DESeq (Anders and Huber 2010), DESeq2 (Love et al. Genome Biology 2014)
- MetagenomeSeq’s Cumulative Sum Scaling (CSS) (Paulson et al. Nature Methods 2013).

Filtration

- Remove taxa with 0 count (prune_taxa)
- Remove OTUs that appear less than n times or in less than n samples (genefilter_sample, filterfun_sample)

Getting controls

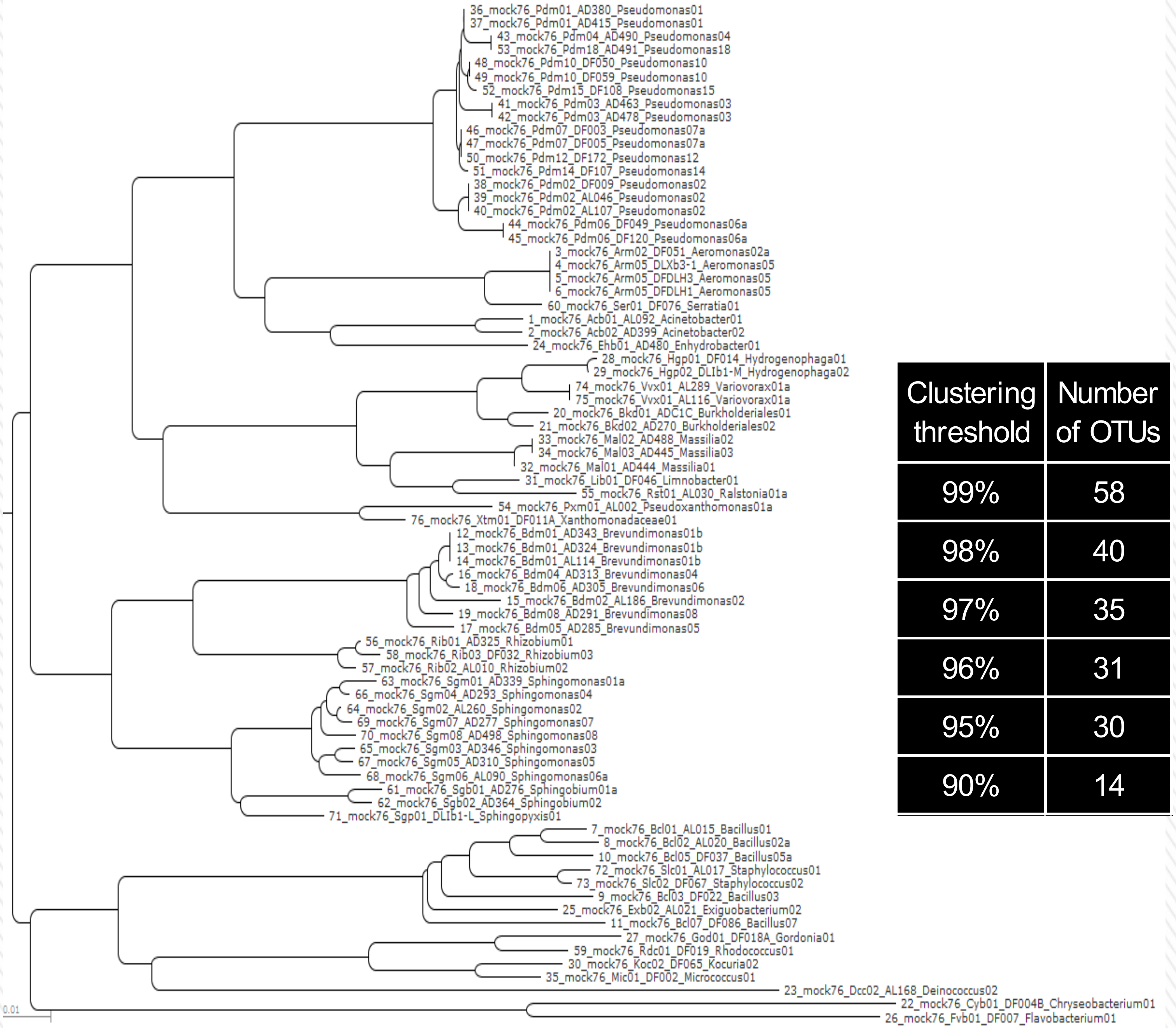


Daphnia, algae and ADaM substrates

Isolation

Stock

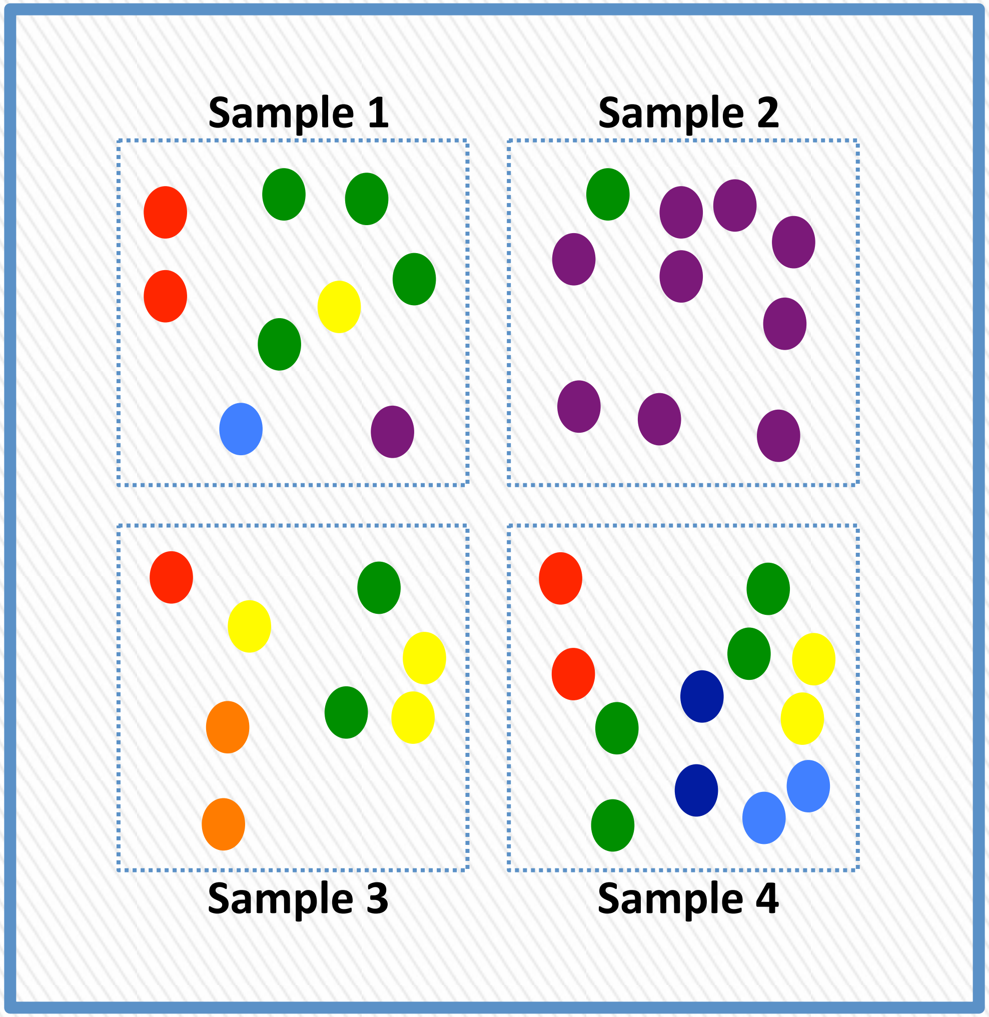
Mock 76 strains


36_mock76_Pdm01_AD380_Pseudomonas01
37_mock76_Pdm01_AD415_Pseudomonas01
43_mock76_Pdm04_AD490_Pseudomonas04
53_mock76_Pdm18_AD491_Pseudomonas18
48_mock76_Pdm10_DF050_Pseudomonas10
49_mock76_Pdm10_DF059_Pseudomonas10
52_mock76_Pdm15_DF108_Pseudomonas15
41_mock76_Pdm03_AD463_Pseudomonas03
42_mock76_Pdm03_AD478_Pseudomonas03
46_mock76_Pdm07_DF003_Pseudomonas07a
47_mock76_Pdm07_DF005_Pseudomonas07a
50_mock76_Pdm12_DF172_Pseudomonas12
51_mock76_Pdm14_DF107_Pseudomonas14
38_mock76_Pdm02_DF009_Pseudomonas02
39_mock76_Pdm02_AL046_Pseudomonas02
40_mock76_Pdm02_AL107_Pseudomonas02
44_mock76_Pdm06_DF049_Pseudomonas06a
45_mock76_Pdm06_DF120_Pseudomonas06a
3_mock76_Arm02_DF051_Aeromonas02a
4_mock76_Arm05_DLXb3-1_Aeromonas05
5_mock76_Arm05_DFDLH3_Aeromonas05
6_mock76_Arm05_DFDLH1_Aeromonas05
60_mock76_Ser01_DF076_Serratia01
1_mock76_Acb01_AL092_Acinetobacter01
2_mock76_Acb02_AD399_Acinetobacter02
24_mock76_Ehb01_AD480_Enhydrobacter01
28_mock76_Hgp01_DF014_Hydrogenophaga01
29_mock76_Hgp02_DLIb1-M_Hydrogenophaga02
74_mock76_Vvx01_AL289_Variovorax01a
75_mock76_Vvx01_AL116_Variovorax01a
20_mock76_Bkd01_ADC1C_Burkholderiales01
21_mock76_Bkd02_AD270_Burkholderiales02
33_mock76_Mal02_AD488_Massilia02
34_mock76_Mal03_AD445_Massilia03
32_mock76_Mal01_AD444_Massilia01
31_mock76_Lib01_DF046_Limnobacter01
55_mock76_Rst01_AL030_Ralstonia01a
54_mock76_Pxm01_AL002_Pseudoxanthomonas01a
76_mock76_Xtm01_DF011A_Xanthomonadaceae01
12_mock76_Bdm01_AD343_Brevundimonas01b
13_mock76_Bdm01_AD324_Brevundimonas01b
14_mock76_Bdm01_AL114_Brevundimonas01b
16_mock76_Bdm04_AD313_Brevundimonas04
18_mock76_Bdm06_AD305_Brevundimonas06
15_mock76_Bdm02_AL186_Brevundimonas02
19_mock76_Bdm08_AD291_Brevundimonas08
17_mock76_Bdm05_AD285_Brevundimonas05
56_mock76_Rib01_AD325_Rhizobium01
58_mock76_Rib03_DF032_Rhizobium03
57_mock76_Rib02_AL010_Rhizobium02
63_mock76_Sgm01_AD339_Sphingomonas01a
66_mock76_Sgm04_AD293_Sphingomonas04
64_mock76_Sgm02_AL260_Sphingomonas02
69_mock76_Sgm07_AD277_Sphingomonas07
70_mock76_Sgm08_AD498_Sphingomonas08
65_mock76_Sgm03_AD346_Sphingomonas03
67_mock76_Sgm05_AD310_Sphingomonas05
68_mock76_Sgm06_AL090_Sphingomonas06a
61_mock76_Sgb01_AD276_Sphingobium01a
62_mock76_Sgb02_AD364_Sphingobium02
71_mock76_Sgp01_DLIb1-L_Sphingopyxis01
7_mock76_Bcl01_AL015_Bacillus01
8_mock76_Bcl02_AL020_Bacillus02a
10_mock76_Bcl05_DF037_Bacillus05a
72_mock76_Slc01_AL017_Staphylococcus01
73_mock76_Slc02_DF067_Staphylococcus02
9_mock76_Bcl03_DF022_Bacillus03
25_mock76_Exb02_AL021_Exiguobacterium02
11_mock76_Bcl07_DF086_Bacillus07
27_mock76_God01_DF018A_Gordonia01
59_mock76_Rdc01_DF019_Rhodococcus01
30_mock76_Koc02_DF065_Kocuria02
35_mock76_Mic01_DF002_Micrococcus01
23_mock76_Dcc02_AL168_Deinococcus02
22_mock76_Cyb01_DF004B_Chryseobacterium01
26_mock76_Fvb01_DF007_Flavobacterium01
0.01

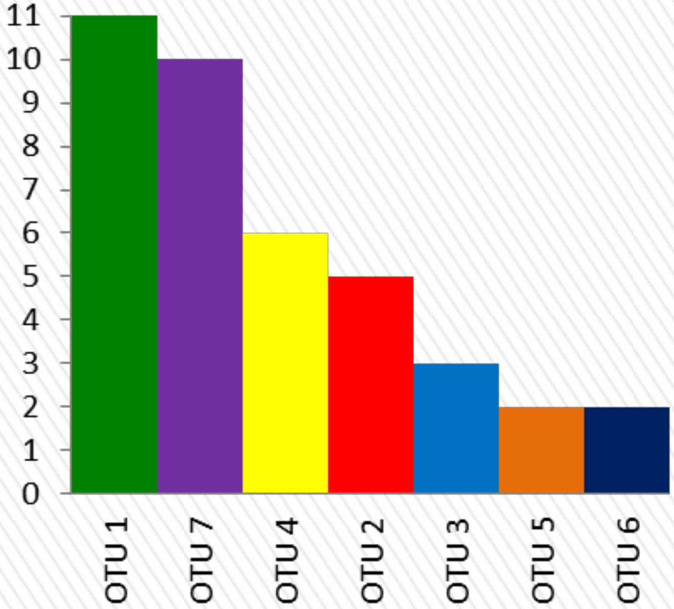
Clustering threshold	Number of OTUs
99%	58
98%	40
97%	35
96%	31
95%	30
90%	14

β-diversity metrics

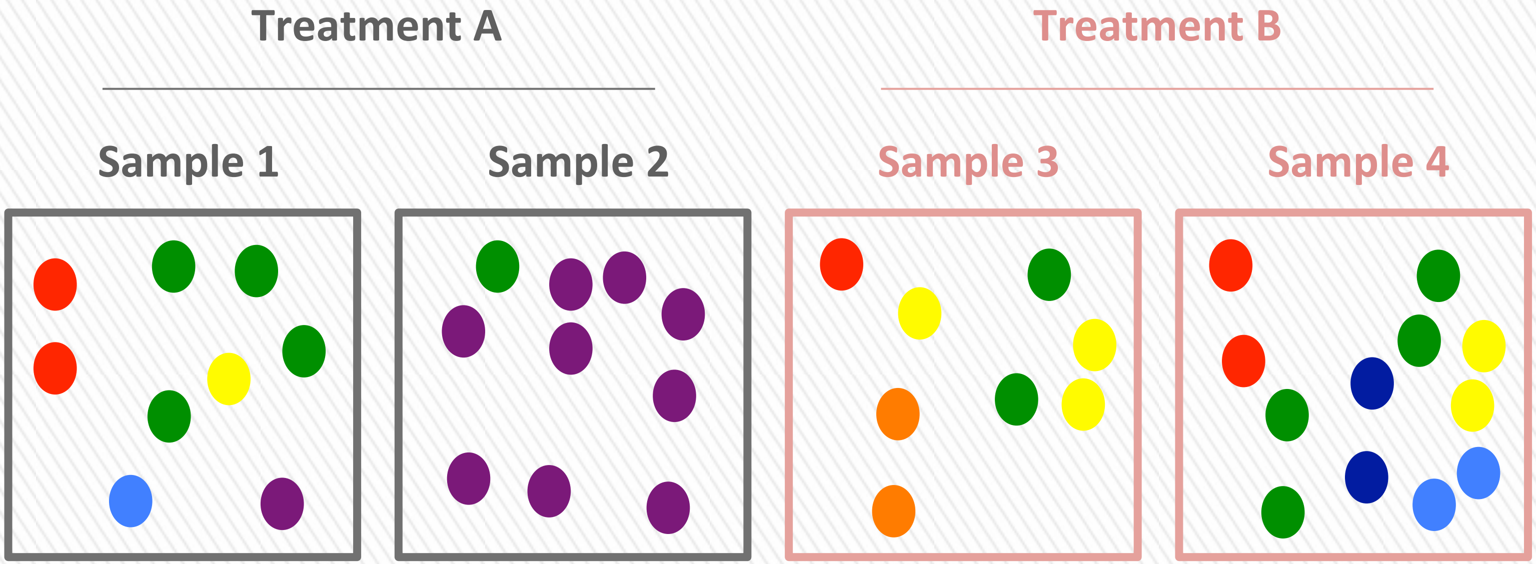
Gamma Diversity: total species diversity



R = 7



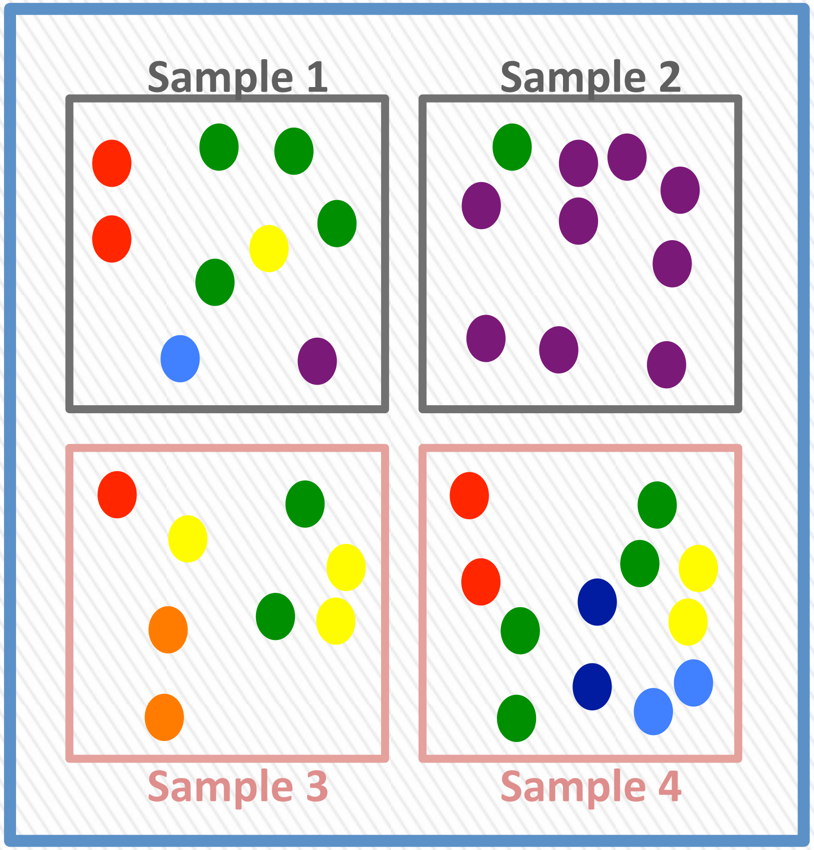
Beta Diversity: between sample diversity

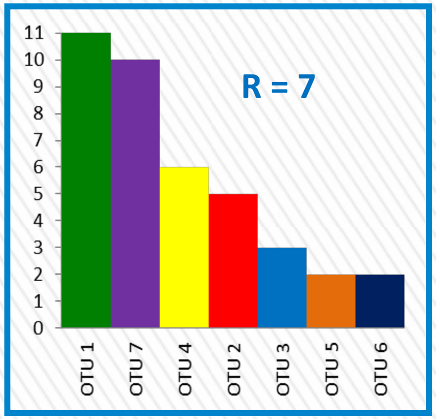


Main concepts of beta-diversity:
Question – What is the influence of treatments A & B?
What is the species diversity along transects & gradients?

Notion of similarity:
Species differences among samples are positively correlated to β-diversity values and inversely correlated to similarity.

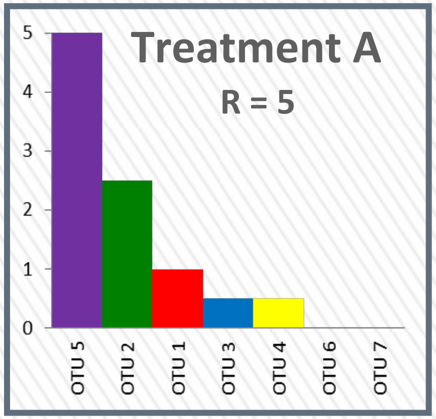
Beta Diversity *stricto sensu*

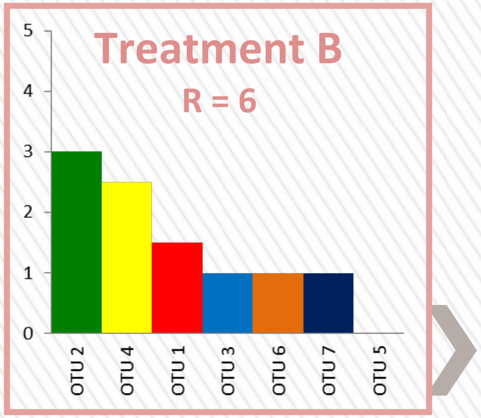




Mean diversity per treatment

$$\beta = \gamma / \alpha$$

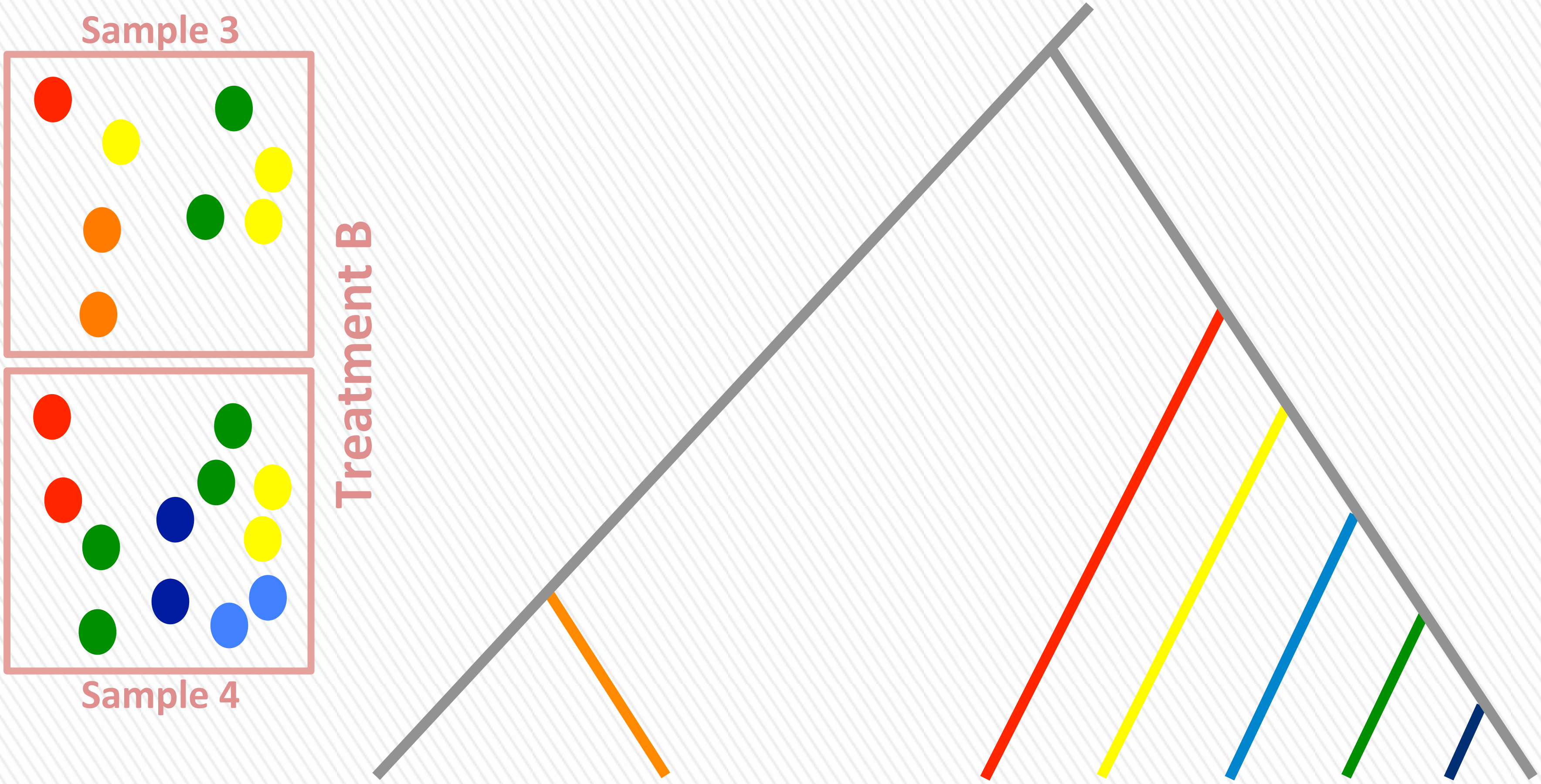




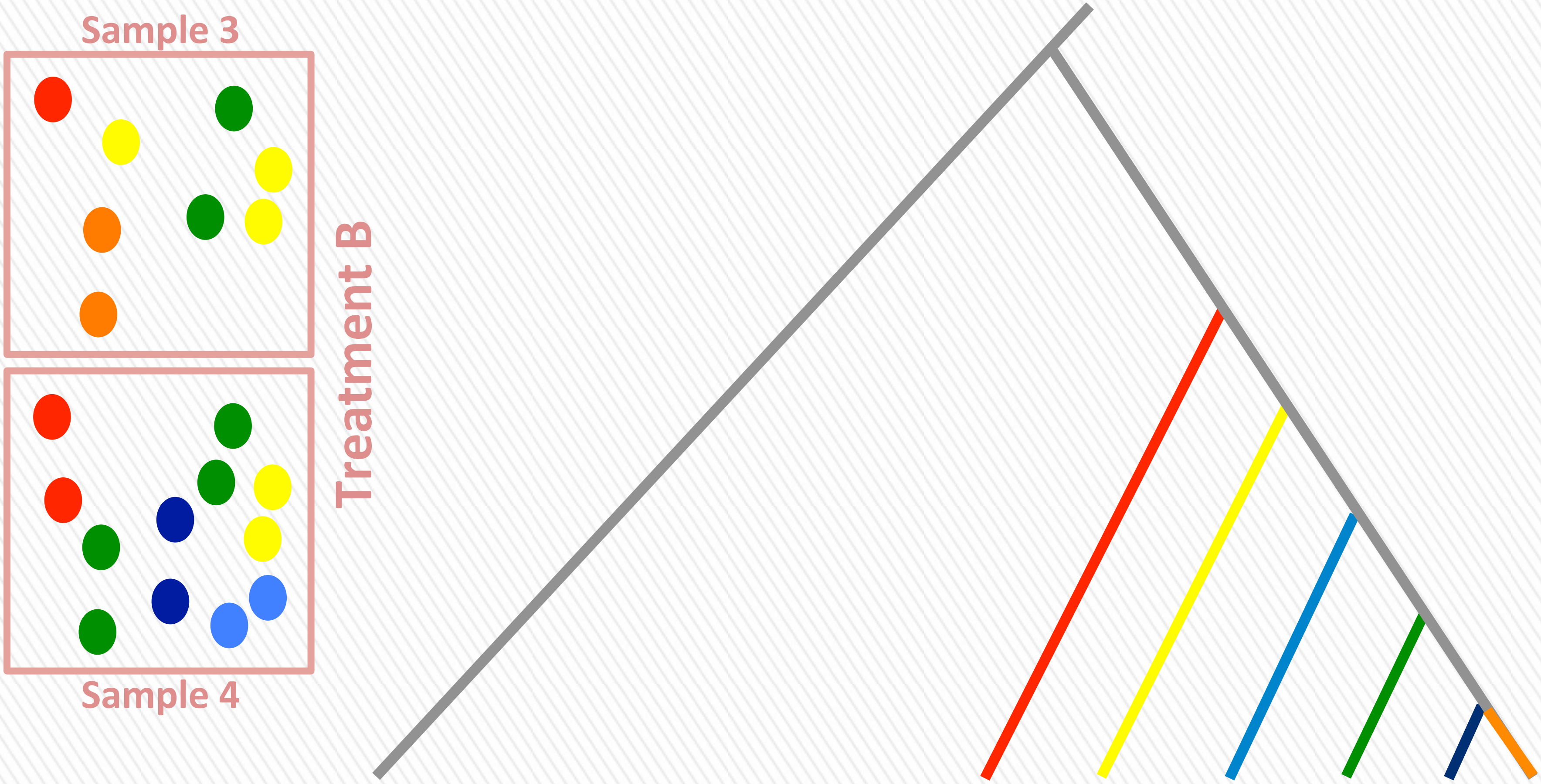
Beta Diversity: between sample diversity

	Presence/Absence	**Abundance**
Without a phylogenetic tree	• Jaccard	• Bray-Curtis (PCoA) • Euclidean (PCA)
With a phylogenetic tree		

Beta Diversity: phylogenetic diversity



Beta Diversity: phylogenetic diversity

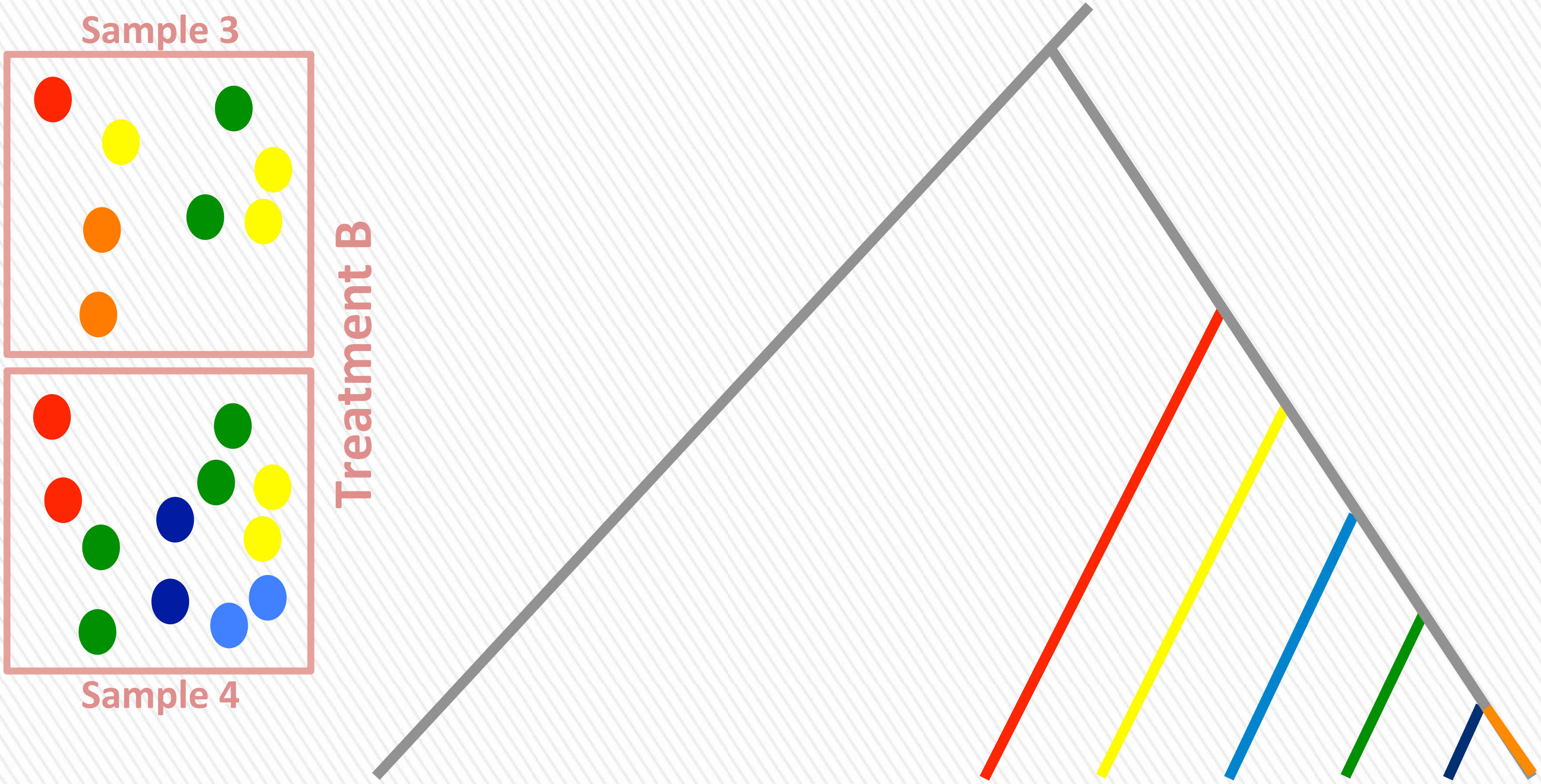


Sample 3

Treatment B

Sample 4

Beta Diversity: phylogenetic diversity



Distance between samples is reduced if phylogenetic approach is used

Beta Diversity: phylogeny-based assessment of differences in overall bacterial community composition.

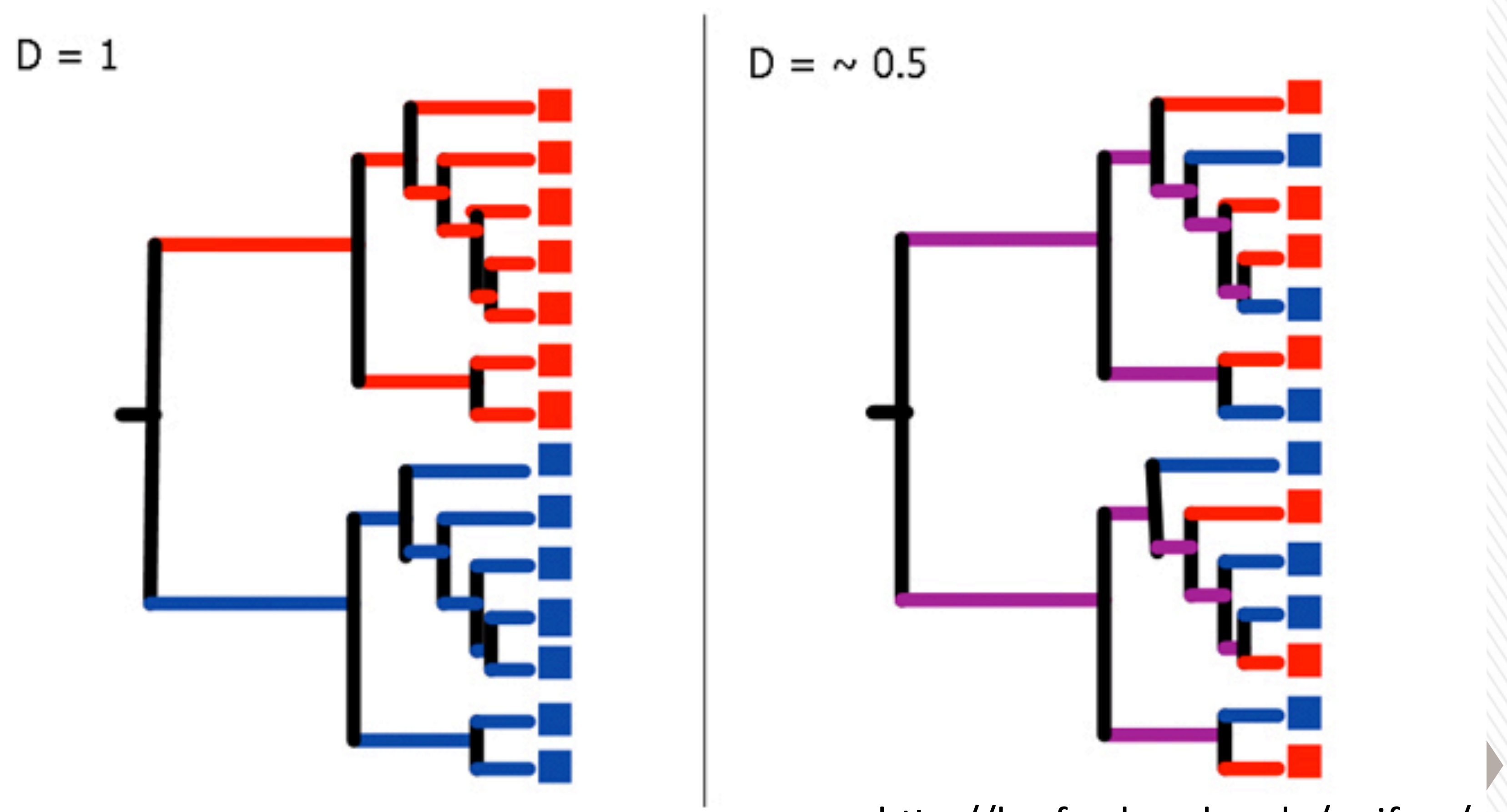
UniFrac distances are based on the fraction of branch length shared between two communities within a phylogenetic tree constructed from the 16S rRNA gene sequences from all communities being compared.

- With unweighted UniFrac, only the presence or absence of lineages are considered (community membership).

- With weighted UniFrac, branch lengths are weighted based on the relative abundances of lineages within communities (community structure).

Beta Diversity: phylogenetic diversity

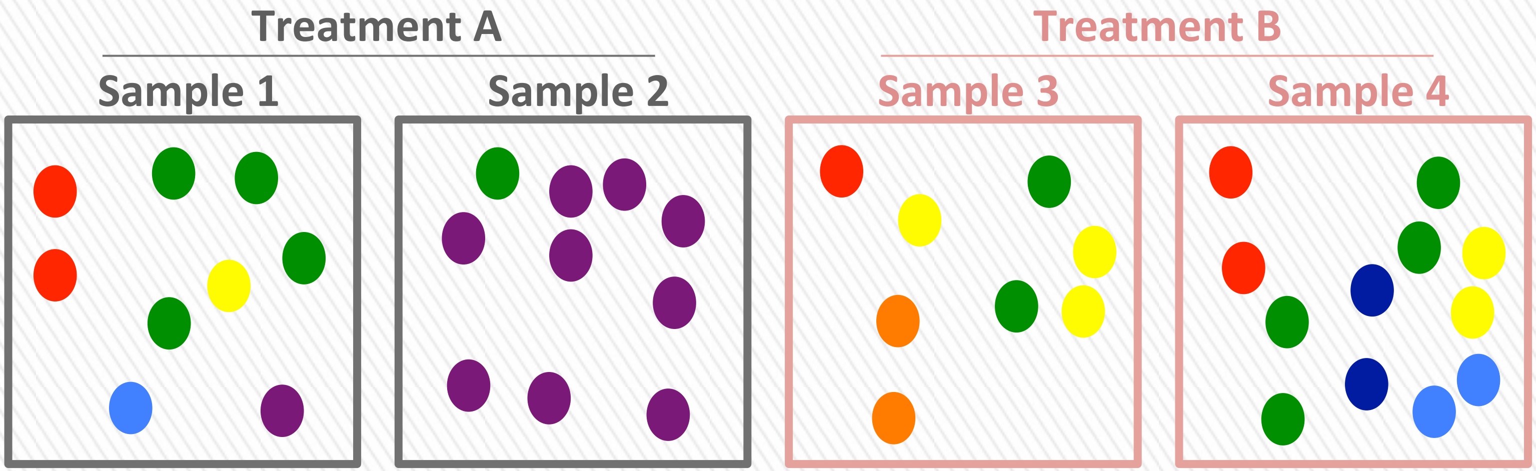
UniFrac:



http://bmf.colorado.edu/unifrac/

Beta Diversity: between sample diversity

	Presence/Absence	**Abundance**
Without a phylogenetic tree	• Jaccard	• Bray-Curtis (PCoA) • Euclidean (PCA)
With a phylogenetic tree	• Unweighted UniFrac • Comdist	• Weighted UniFrac • Comdist

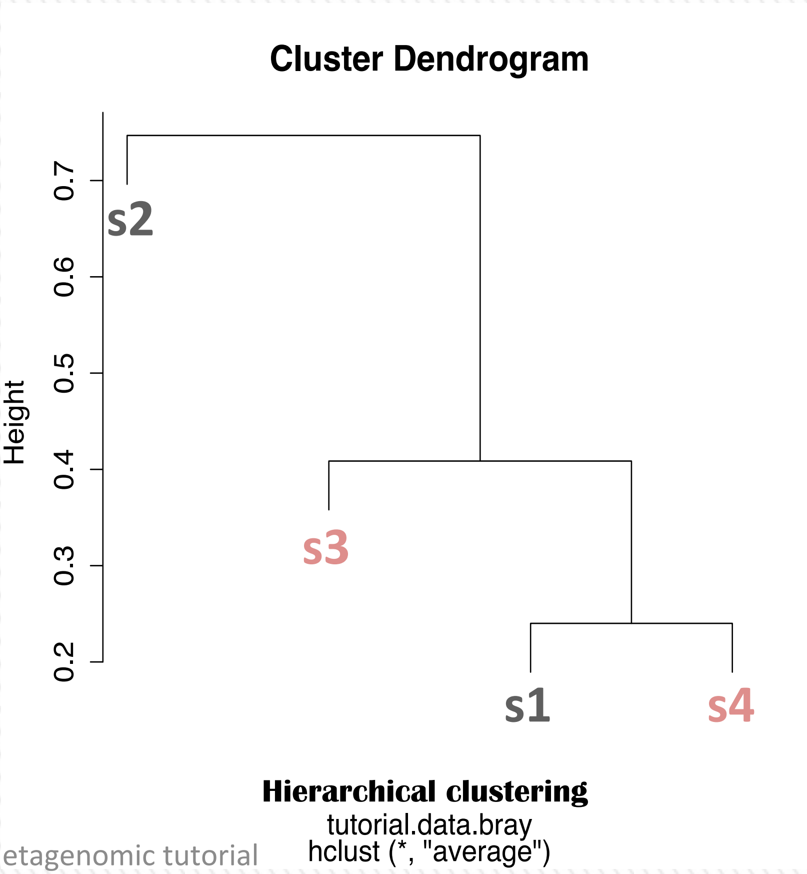


OTU_table

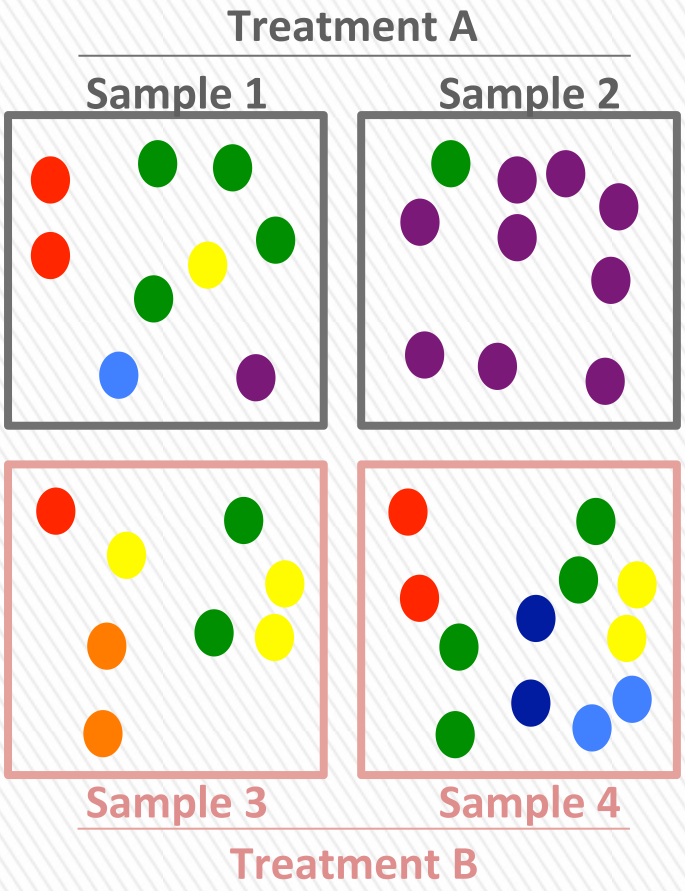
	Sample1	Sample2	Sample3	Sample4
OTU 1	2	0	1	2
OTU 2	4	1	2	4
OTU 3	1	0	0	2
OTU 4	1	0	3	2
OTU 5	1	9	0	0
OTU 6	0	0	2	0
OTU 7	0	0	0	2

MapFile

#SampleID	Treatment
Sample1	A
Sample2	A
Sample3	B
Sample4	B



Beta Diversity: visualization / ordination



Treatment A

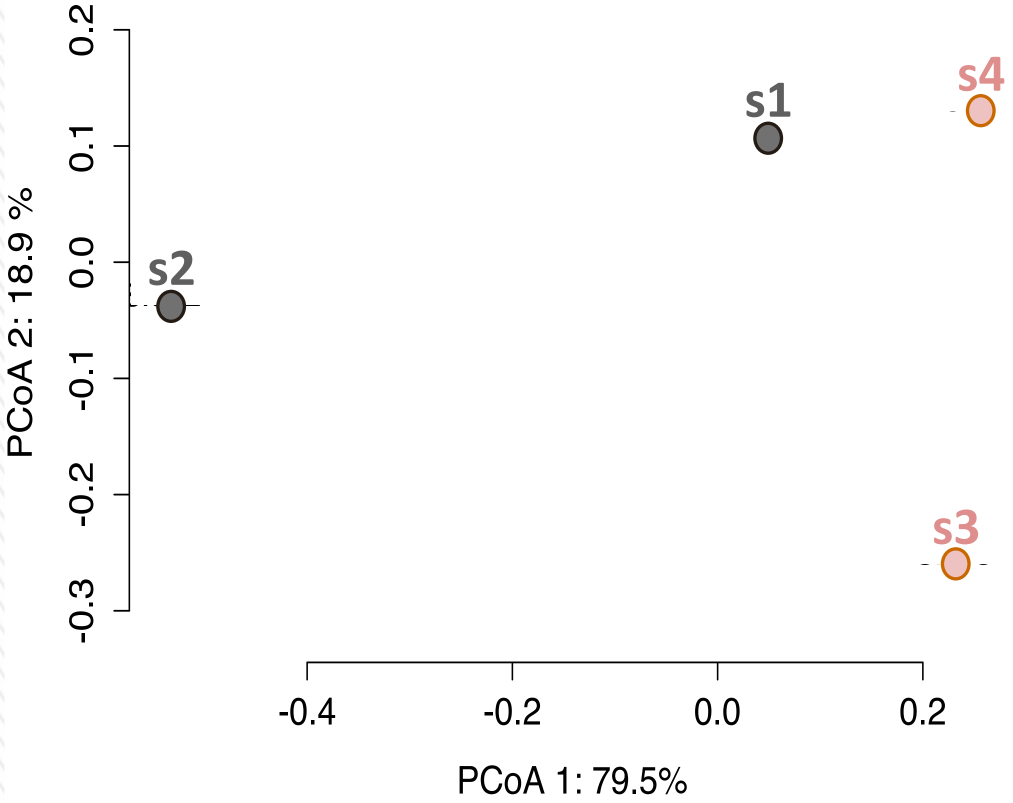
Sample 1

Sample 2

Sample 3

Sample 4

Treatment B



Your questions

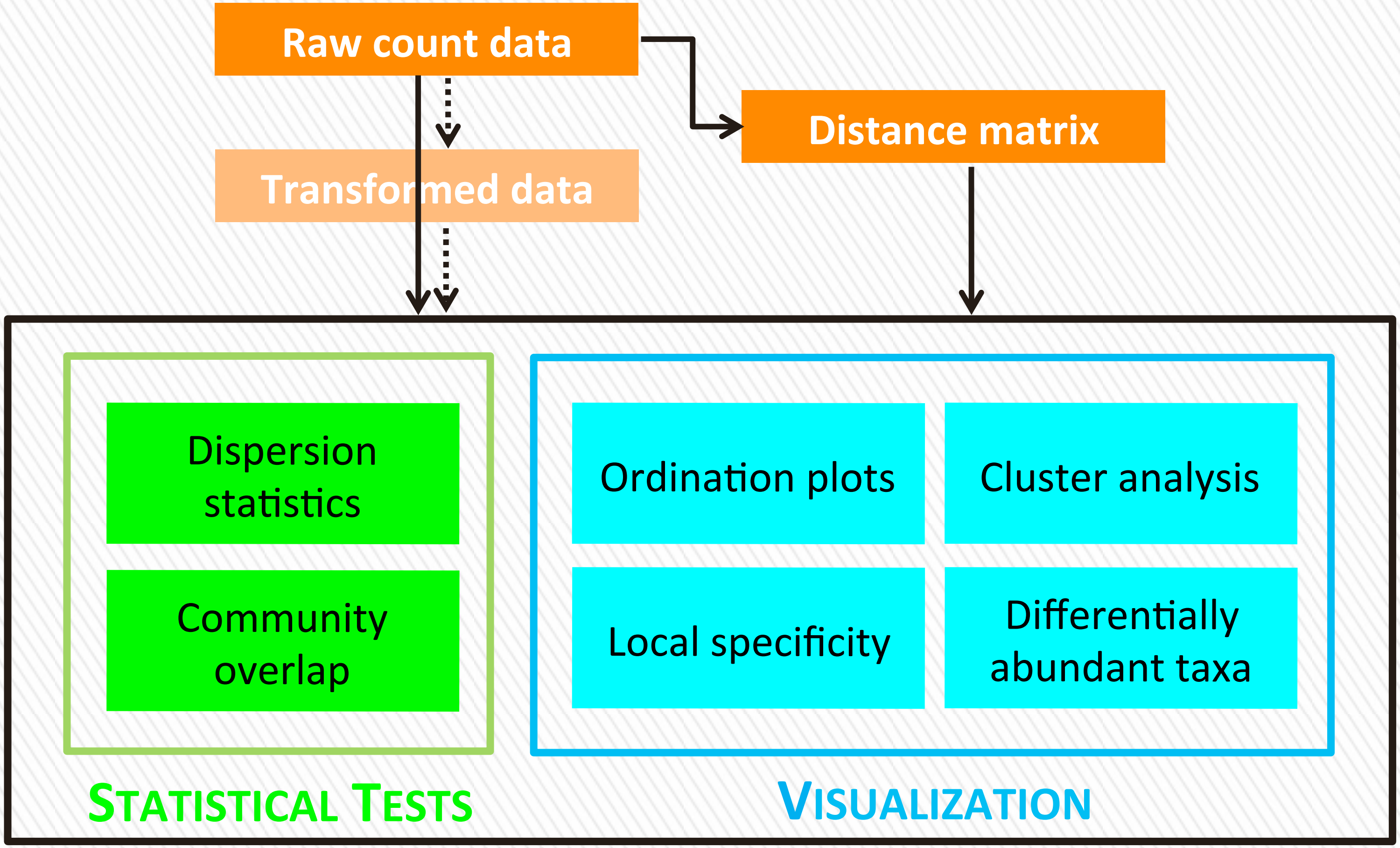
- What you want to know determines how you analyze your data

- How important is each aspect of diversity?
  - Richness?
  - Evenness?
  - Dominance?
  - Abundance?
  - Per-species (relative) abundance?
  - Taxon diversity?

Comparison of samples or group of samples:

How similar are communities?

Framework



Raw count data

Transformed data

Distance matrix

Dispersion statistics

Community overlap

Ordination plots

Cluster analysis

Local specificity

Differentially abundant taxa

STATISTICAL TESTS

VISUALIZATION

Beta Diversity: similarity coefficients

Based only on the number of species present in each sample
All species are counted & weighted equally

- Jaccard $C_J = j / (a + b - j)$
  a = richness in first site, b = richness in second site
  j = shared species

- Sorensen $C_S = 2j / (a + b)$
  makes an effort to weight shared species by their relative abundance

- Sorensen Quantitative $C_N = 2(jN) / (aN + bN)$
  jN = sum of the lower of the two abundances recorded for species found in each site

- Morisita-Horn C_{mH} is not influenced by sample size & richness but highly sensitive to the abundance of the most abundant OTUs

- Cluster Analyses use a similarity matrix of all samples
  - Group Average clustering
  - Centroid Clustering

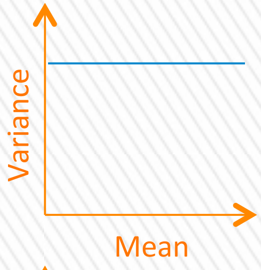
Beta Diversity: phylogenetic distance

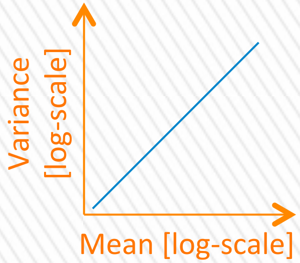
- Unique Fraction (UniFrac) metric
- Qualitative phylogenetic β-diversity = Unweighted UniFrac
- Distance = fraction of the total branch length that is unique to any particular environment
- Quantitative phylogenetic β-diversity = Weighted UniFrac

Lozupone and Knight (2005) *Appl. Environ. Microbiol.*
Lozupone *et al.* (2007) *Appl. Environ. Microbiol.*

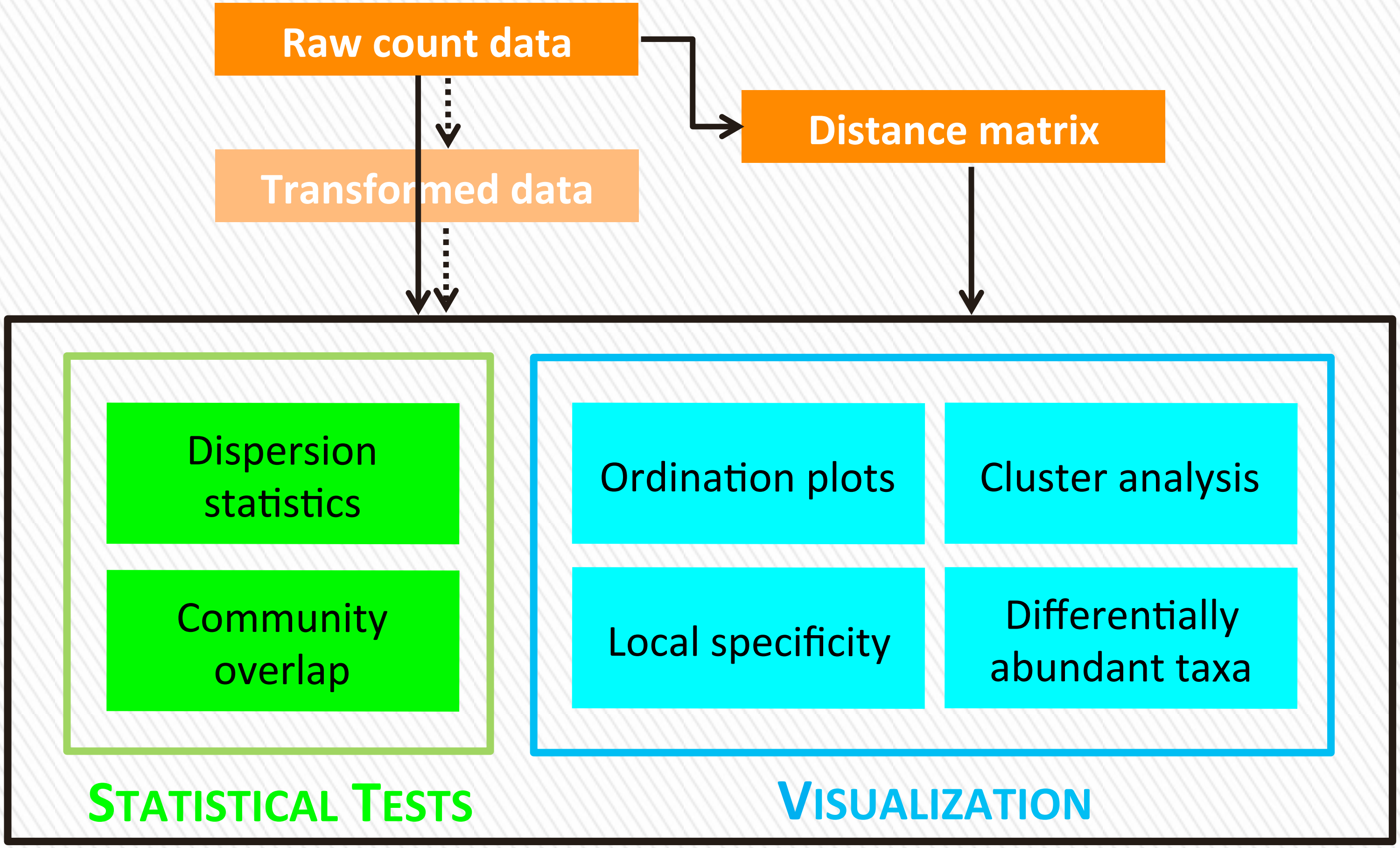
Distance metrics overview

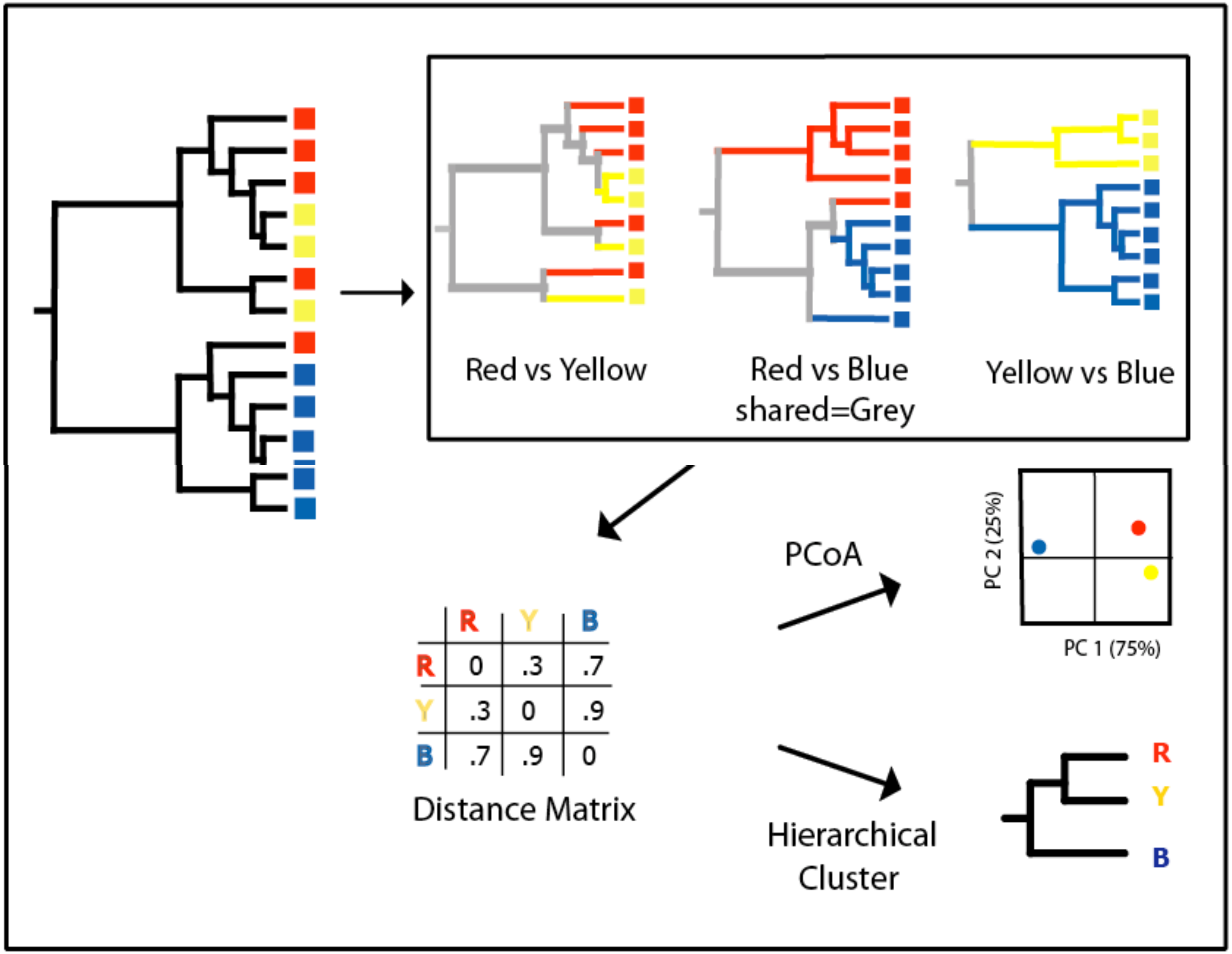
Distances	To be used when…
Euclidean Manhattan	variables are expected to have equal variance
Bray–Curtis Canberra	between-group differences in average absolute differenceare matched by proportionate changes in average abundance
Gower	variables are physical or chemical data
Jaccard Sorensen Ochiai	face presence/absence data
Whittaker Hellinger	analyze quantitative assemblage composition





Framework


Raw count data
Transformed data
Distance matrix
Dispersion statistics
Community overlap
Ordination plots
Cluster analysis
Local specificity
Differentially abundant taxa
STATISTICAL TESTS
VISUALIZATION



	R	Y	B
R	0	.3	.7
Y	.3	0	.9
B	.7	.9	0

Ordination techniques `plot_ordination / plot_samples`

1. **PCoA (Principal Coordinates Analyses)** also called MDS (Metric Dimensional Scaling) relies on a dissimilarity or distance matrix. A non-metric variant of PCoA is called NMDS (Non Metric Dimensional Scaling).

2. **DPCoA (Double Principal Coordinate Analysis)** is a two step PCoA. The procedure first computes a distance matrix for taxa using the patristic distance (length of the shortest path on a tree) over the taxa phylogeny. The position of the communities in coordinate space is then the average position (centroids) of their constituent taxa, weighted by relative abundances. The common space for taxa and communities allows for easier interpretation of the community. It also highlights leverage taxa that drive the differences between communities.

3. **PCA (Principal Components Analysis)** preserves the variance of samples. In particular, edge PCA, is an hybrid method where taxa abundance is combined to a phylogeny to create contrasts that are used as input variables. Unlike PCoA based on UniFrac distances and DPCoA, it does not use branch lengths. However, and unlike most other ordination methods, the principal components can be mapped onto the tree for easy visualization and interpretation.

Ordination techniques

Rely on presence/absence of taxa

PCoA/MDS with UniFrac

Puts more weight on shallow branches on the tree than either DPCoA or wUniFrac, sensitive to noise**, picks up shallower differences, linear runtime in the number of OTUs and samples

Rely on relative abundance of taxa

PCoA/MDS with wUniFrac

Less sensitive to outliers* / more sensitive to noise** than DPCoA, linear runtime in the number of OTUs and samples

DPCoA

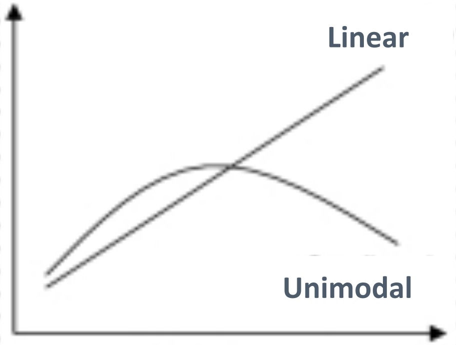
Most sensitive to outliers*, least sensitive to noise**, upweights deep differences, gives OTU locations, long runtime (depends on the number of OTUs but not of samples)

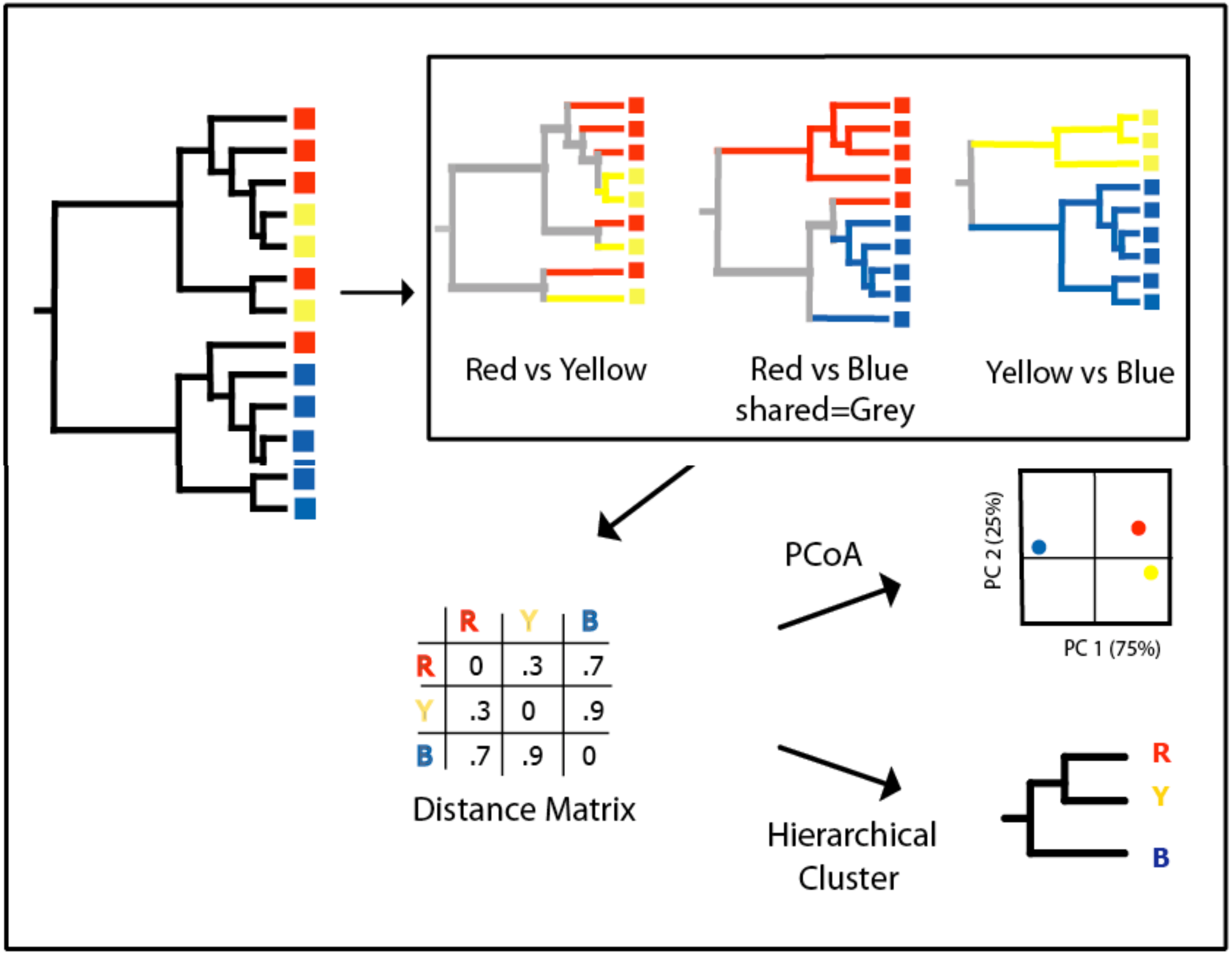
* Outliers = highly abundant OTUs

** Noise = noise in detecting low-abundance OTUs (around 0 count)

Ordination techniques

	Raw data (presence/absence or abundance data)		Distance-based (db)-data (distance matrix)	
	Response along a gradient			How you interprete results
	Linear	Unimodal		
Unconstrained by env. factors	tb-PCA	CA, DCA	PCoA, NMDS	Env. factors are used *post hoc*
Constrained by env. factors	tb-RDA	CCA	db-RDA	Test of significance and partitioning of variance explained by env. factors



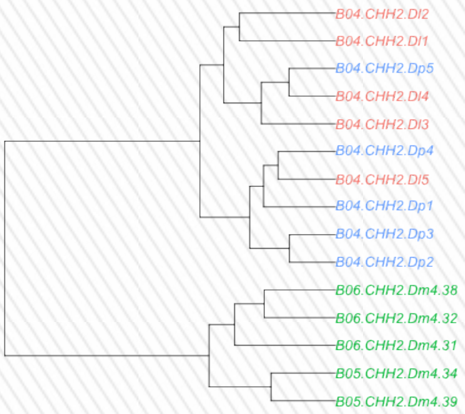


	R	Y	B
R	0	.3	.7
Y	.3	0	.9
B	.7	.9	0

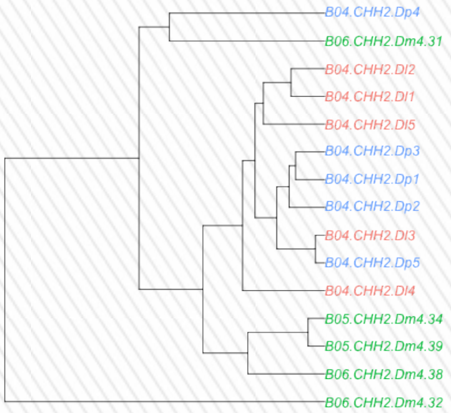
Distance Matrix

Hierarchical clustering

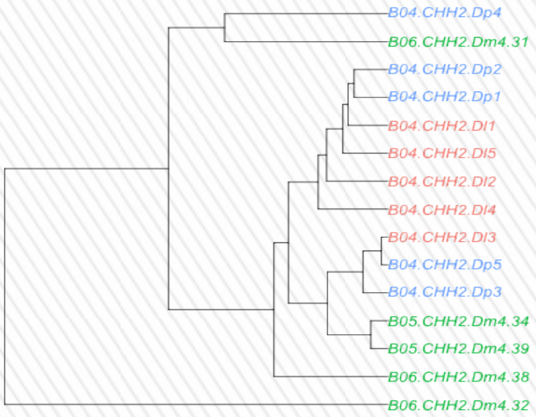
UniFrac distances



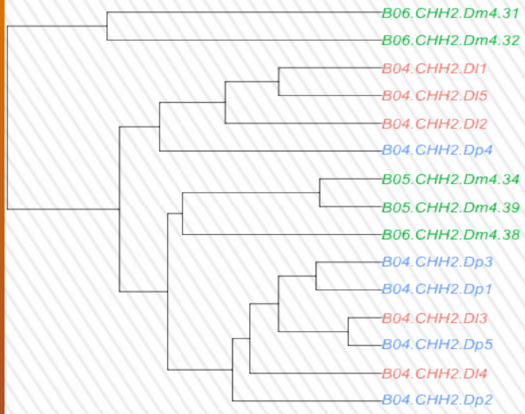
wUniFrac distances



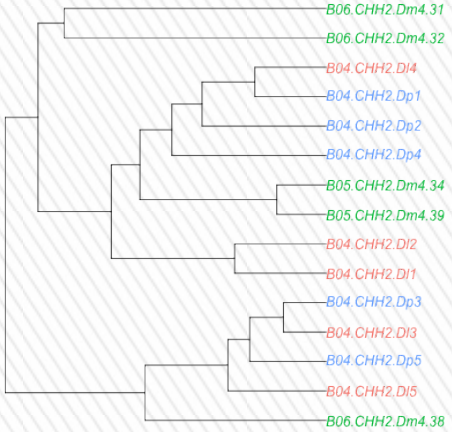
DPCoA dissimilarities



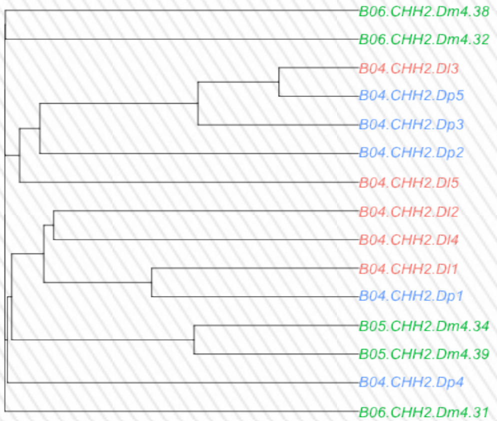
Bray-Curtis distances



Bray-Curtis distances (no filter)



'edge PCA' distances

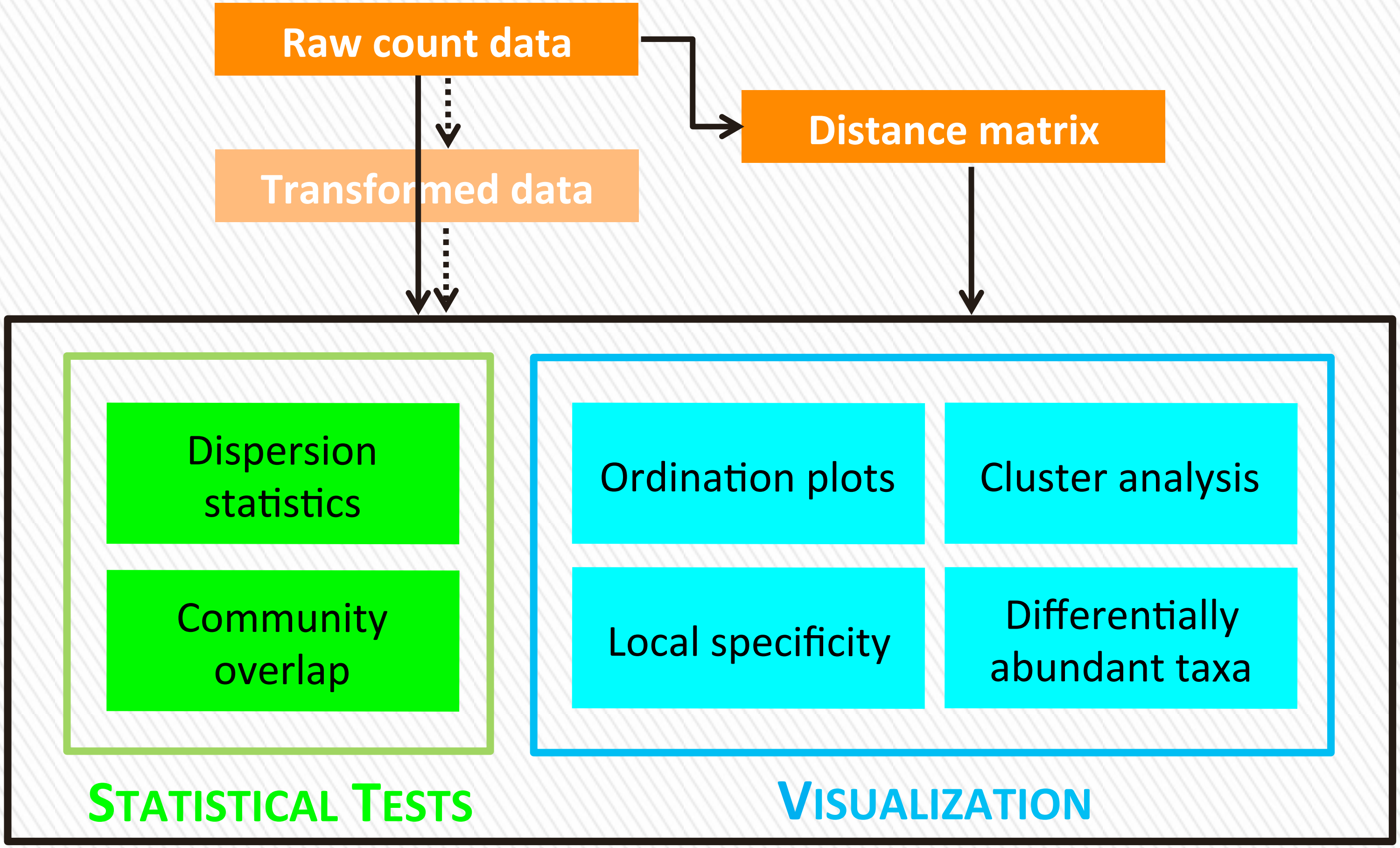


Distance-based analyses

1. Use a distance measure that is ecologically meaningful. A good analysis practice is to repeat the analysis with several good distance measures and investigate whether all these analyses lead to the same conclusion.

2. Investigate how well the distances in the ordination graph represent the total distances (e.g.

Framework


Raw count data

Transformed data

Distance matrix

Dispersion statistics

Community overlap

Ordination plots

Cluster analysis

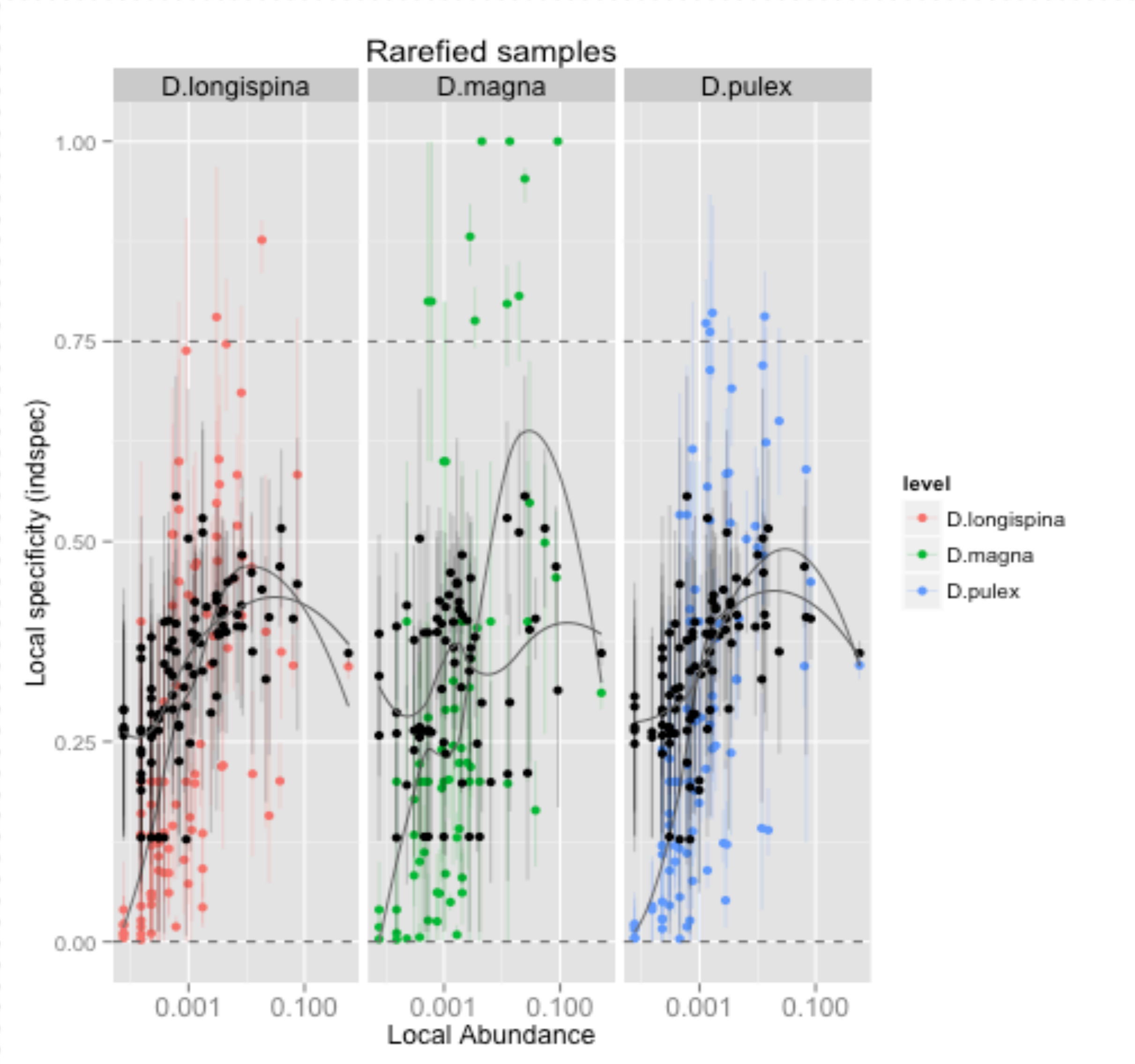
Local specificity

Differentially abundant taxa

STATISTICAL TESTS

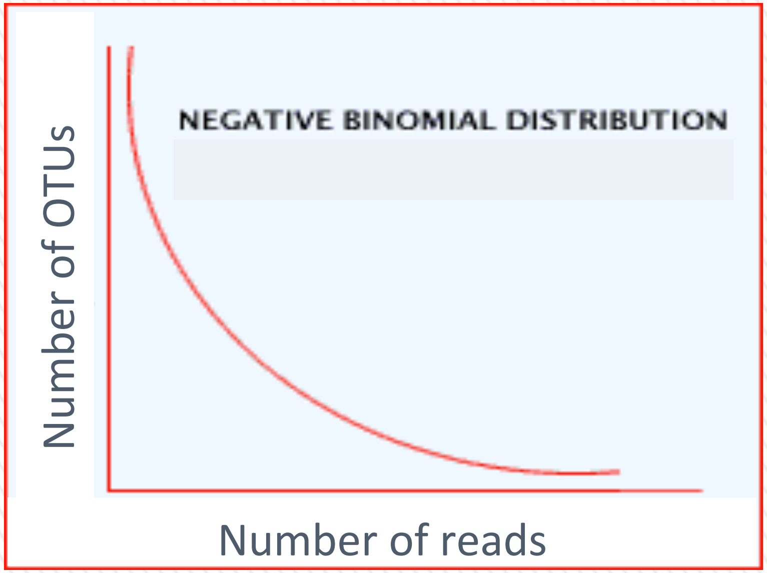
VISUALIZATION

Local Specicity: identify indicator species



Differentially Abundant OTUs

`edgeR` and `DESeq`
based on raw counts and use negative binomial distributions to model count data.
edgeR is usually more conservative than DESeq.



`A two-sided t-test with unequal variances`
use the mt wrapper in phyloseq and based either on rank test or moderated taxa-wise ANOVA.
Is usually more conservative than the negative binomial model implemented in edgeR and DESeq. Nevertheless, they are quite sufficient to detect large size effects and are easier to use (*e.g.* Fisher test).

Statistical tests for different membership

- UniFrac Significance: fraction of random trees that have more Unique branch length than the real tree.
- Phylogenetic (P) Test: based on the number of changes between states (samples) required to explain the distribution of sequences on the tree (Fitch parsimony). Sensitive to tree topology but not to branch lengths.
  See Martin AP (2002) *Appl. Environ. Microbiol.*
- LibShuff

Libshuff (Library shuffling)

The libshuff method is a generic test that describes whether two or more communities have the same structure using the Cramer-von Mises test statistic. The significance of the statistical test indicates the probability that the communities have the same structure by chance. Because each pairwise comparison requires two significance tests, a correction for multiple comparisons (*e.g.* Bonferroni's correction) must be applied.

The program calculates a homologous and a heterologous coverage curve for the libraries then calculates the distance between the two curves and use a Monte Carlo test procedure to compare them.

NB: Monte Carlo simulations: randomly permute the data (environment assignments) and determine how often the random data has a more extreme value than the real data.

Singleton DR *et al.* (2001) *Appl. Environ. Microbiol.*
Schloss PD *et al.* (2004) *Appl. Environ. Microbiol.*

https://toolshed.g2.bx.psu.edu/repos/jjohnson/mothur_toolsuite

Statistical tests for different membership

- UniFrac Significance: fraction of random trees that have more Unique branch length than the real tree.
- Phylogenetic (P) Test: based on the number of changes between states (samples) required to explain the distribution of sequences on the tree (Fitch parsimony). Sensitive to tree topology but not to branch lengths.
  See Martin AP (2002) *Appl. Environ. Microbiol.*
- LibShuff
- ADONIS: Analysis of variance using distance matrices (`vegan` package in R). formal testing of sample covariates is also done using a permutation MANOVA with the (squared) distances and covariates as response and linear predictors, respectively. See Anderson (2001) *Austral Ecology*
- ANOSIM
- Mantel test

Note that these multivariate analyses can be heavily influenced by heterogeneity of dispersion across groups in an unbalanced design.
See Anderson and Walsh (2013) *Ecological Monographs*

Dispersion

- Dispersion is defined as a change in mean–variance relationship

- Permutation test of homogeneity of group dispersion. It is an analogue to homogeneity of variances.

- Test if all groups share a common dispersion (*i.e.* if the variation between samples is similar to the variation between groups).

- Dispersion of sequences in the tree

See Webb CO (2000) *American Naturalist*