Microbial ecology

Francisco Pascoal

Institutional organizers

ciimar

U.PORTO

FC FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Sponsors

SP ECO

fct Fundação
para a Ciência
e a Tecnologia

BIOPORTUGAL
Químico, Farmacêutica, Lda.

Objective

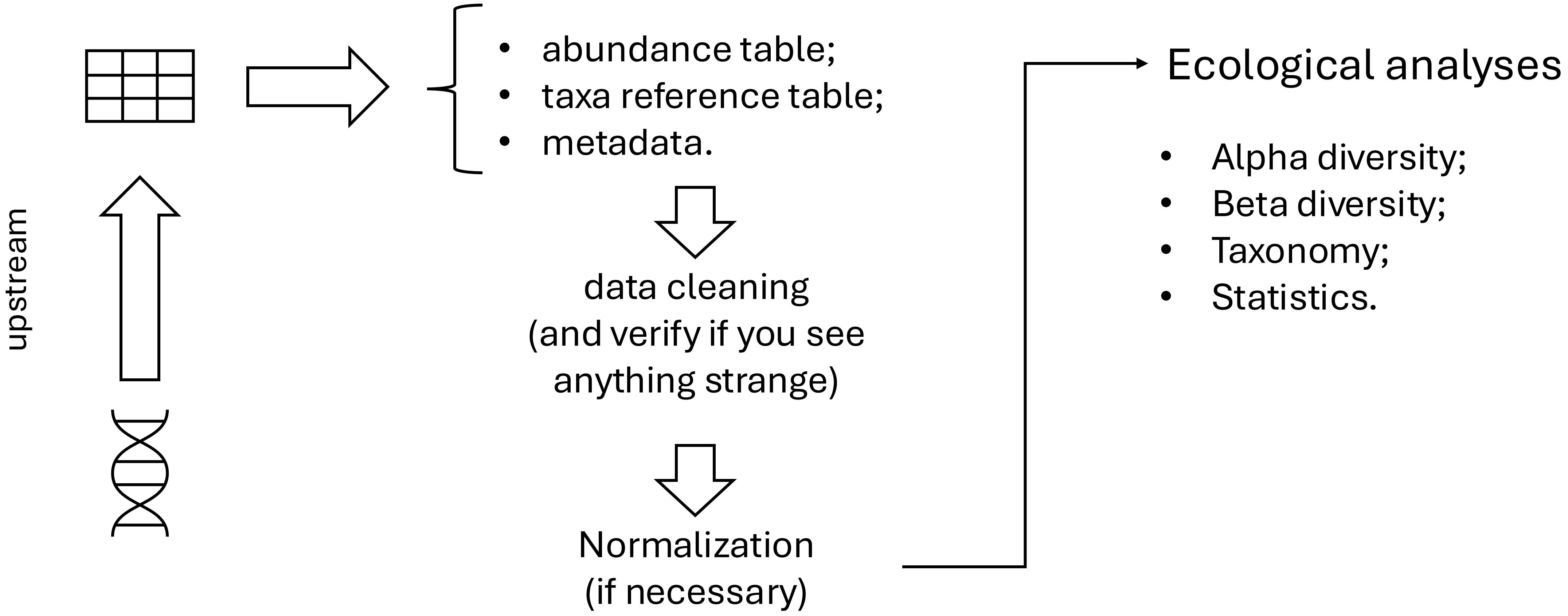
Ecology is the study of living organisms in their environment.

- Basis for the description of microbial communities in their environment;
- The specific aim of a given ecological analysis will vary according to the research question and experimental design.

Contents

- Overview of common steps in microbial ecology;
- Pre-processing and data cleaning;
- Normalization;
- Alpha diversity;
- Beta diversity;
- Taxonomic analysis;
- Statistical tests.

Overview of common steps in microbial ecology



upstream

- abundance table;
- taxa reference table;
- metadata.

data cleaning
(and verify if you see anything strange)

Normalization
(if necessary)

Ecological analyses

- Alpha diversity;
- Beta diversity;
- Taxonomy;
- Statistics.

Pre-processing and data cleaning

Common steps

- Verify the type of objects you have;
- Merge tables, if necessary;
- Match sample information in species abundance table with metadata;
- Clean metadata information;
- Clean taxonomy.
  - Remove NAs at domain/kingdom and phylum level;
  - Remove organelles;
  - Remove other unwanted groups.
- Other (singletons, etc).

Suggestion for R

dplyr

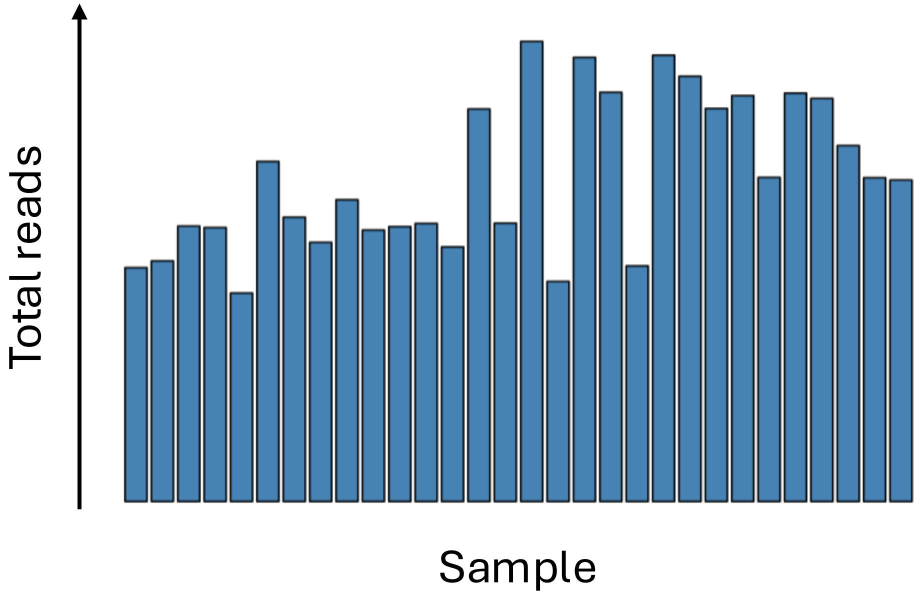
readr

stringr

You should always dedicate some time to inspect your data.

Normalization (1)

Normalization might be necessary for fair comparison of samples.



- The diversity estimates might become biased, when we have samples with very different numbers of reads.
- How do we know if one sample really has more species than another, if they have a total number of reads that is different?
- Rarefaction is the most common solution.

Normalization (2) – rarefaction

What is rarefaction?

- Random sub-sampling of x reads;
- All samples end up having the same total number of reads;
- This method comes from general ecology;
- It automatically deals with singletons.
- Samples with less than x reads will be removed. So, you must balance having as much reads as possible versus losing as fewer samples as possible.

Suggestion for R

Vegan

Normalization (3) – pros and cons of rarefaction

Pros:

✓ All samples get the same number of total reads;

✓ Reads are selected randomly;

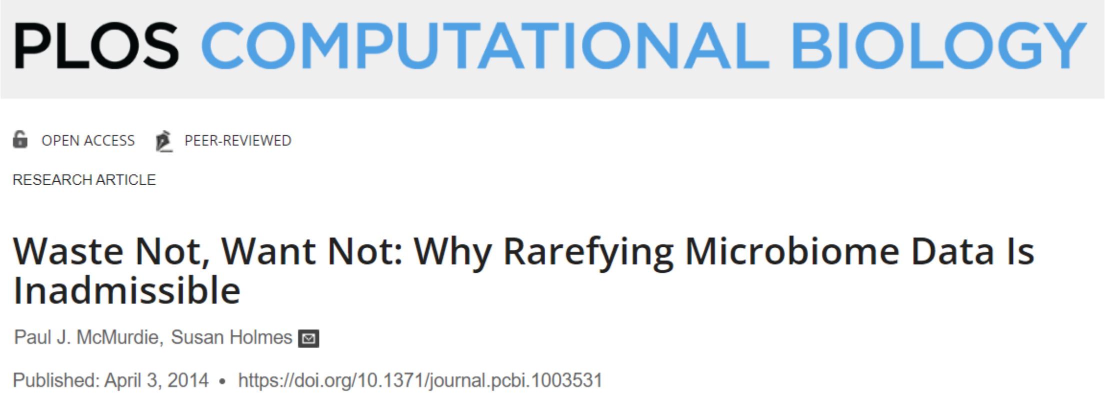
✓ Popular and easy to understand.

Cons:

X Loss of valid information (species);

X Unsuited for compositional data.

Normalization (4) – suggested reading



PLOS COMPUTATIONAL BIOLOGY

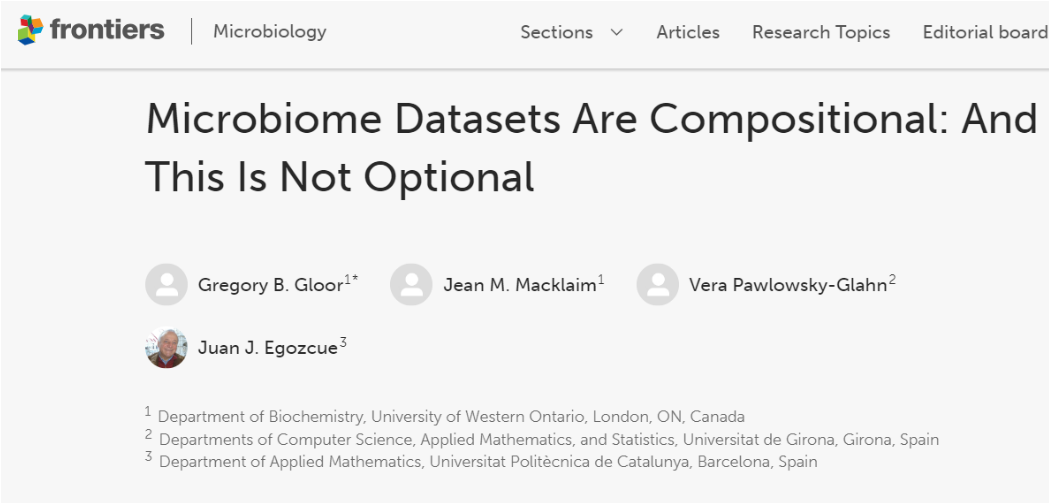
OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes

Published: April 3, 2014 • https://doi.org/10.1371/journal.pcbi.1003531

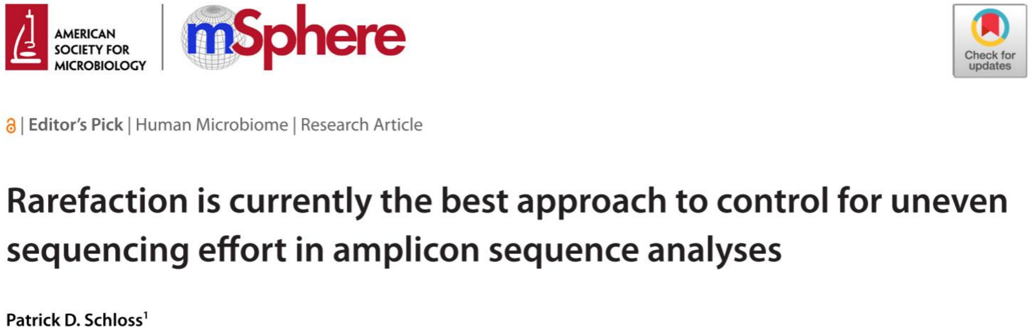


Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1*] Jean M. Macklaim[1] Vera Pawlowsky-Glahn[2]

Juan J. Egozcue[3]

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada
[2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain
[3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain



| Editor's Pick | Human Microbiome | Research Article

Rarefaction is currently the best approach to control for uneven sequencing effort in amplicon sequence analyses

Patrick D. Schloss[1]

Normalization (5) – alternatives to rarefaction

- Relative abundance;
  - note: you should replace absolute abundance with relative abundance independently of rarefying or not your data.
- Hellinger transformation – square root standardized to unit total;
- clr – central logo ratio.
  - will introduce negative abundance values.

Suggestion for R

None solve the problem of uneven sequencing power.
Some may be used together with rarefaction.

Alpha diversity (1)

Alpha diversity is the sample-level diversity

Common metrics:

- Species richness: the number of different species (ASVs/OTUs/etc);
- Shannon index: estimates entropy;
- Simpson: considers the relative abundance of species (downweighs rare species).

Suggestion for R

Vegan

phyloseq

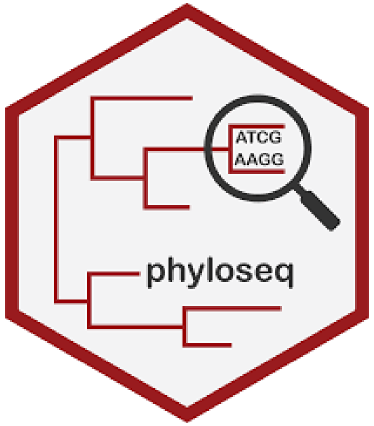
Alpha diversity (2) – with phylogenetic tree

Common metrics:

- Faith index: calculates the length of the tips of the phylogenetic tree;
- Weighted Faith: weights the length of the tips by the relative abundance.

Suggestion for R

geiger R package



Alpha diversity (3) – how to chose an index?

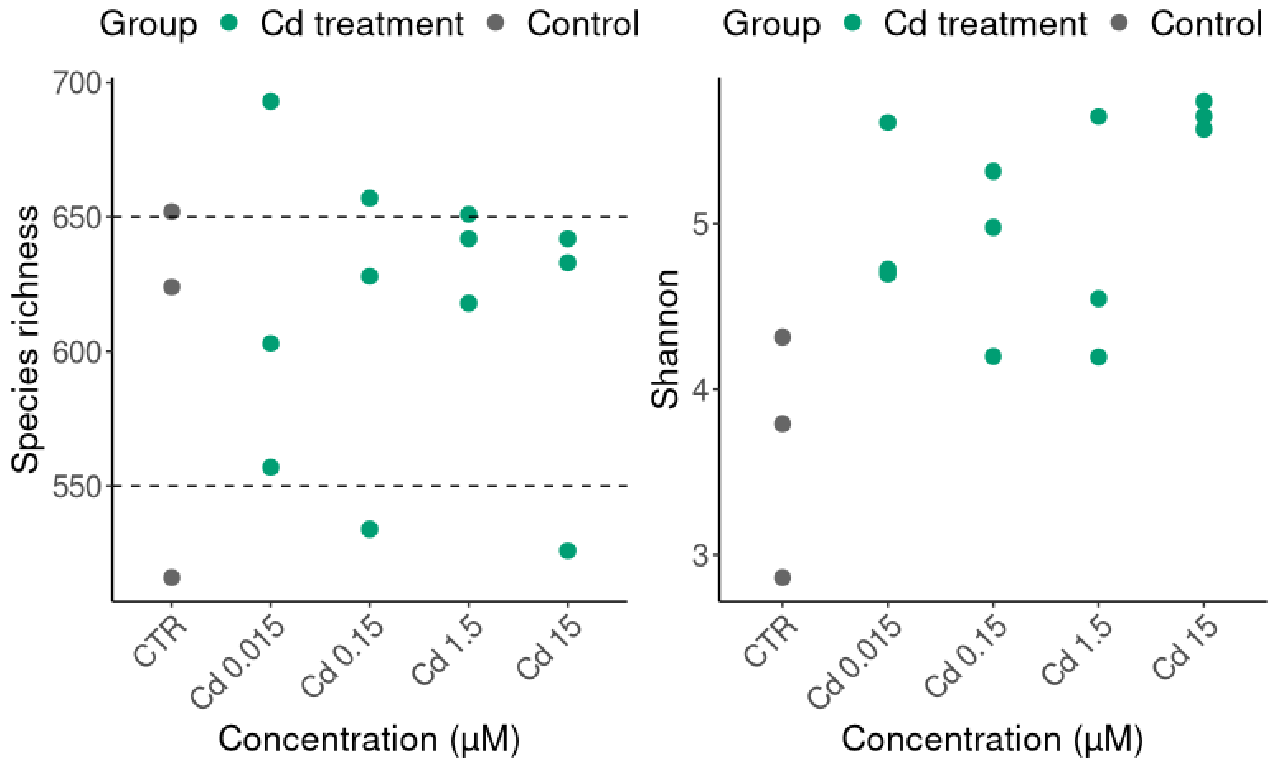
Some considerations:

- At least, calculate species richness;
- Add additional metrics to account for the difference in relative abundance;
- Avoid overwhelming your analysis with redundant metrics.
  - there are too many alpha diversity metrics, and they almost perfectly correlate with each other (see: Swenson, 2014).
- If phylogenetic trees are available, add some metric that accounts for phylogenetic diversity.

Alpha diversity (5) - examples

Alpha diversity as a function of Cadmium (Cd) concentration.

Sed037



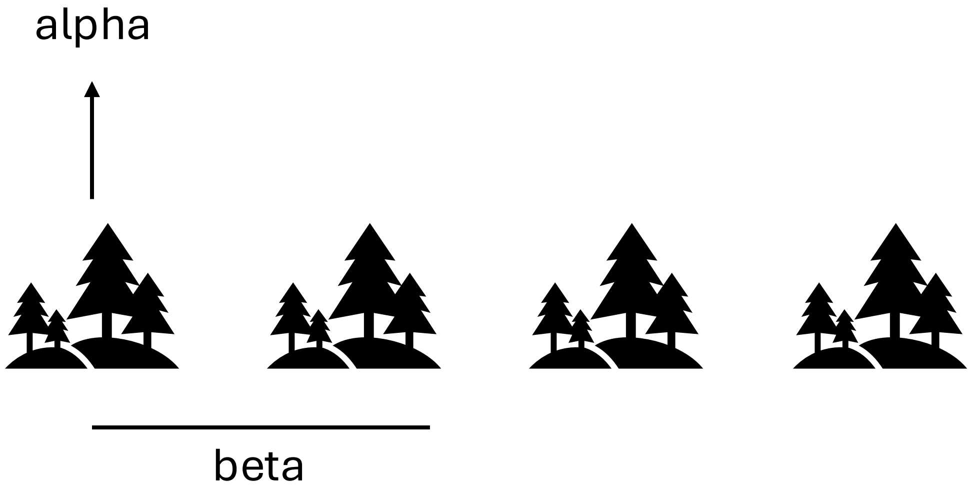
What were the differences between species richness and Shannon?

Beta diversity (1)

Beta diversity is the between-samples diversity

Possible approaches:

- Calculate some metric (example: Bray-Curtis/Sorenson index);
- multivariate analysis and ordination methods – community structure analysis.



Suggestion for R

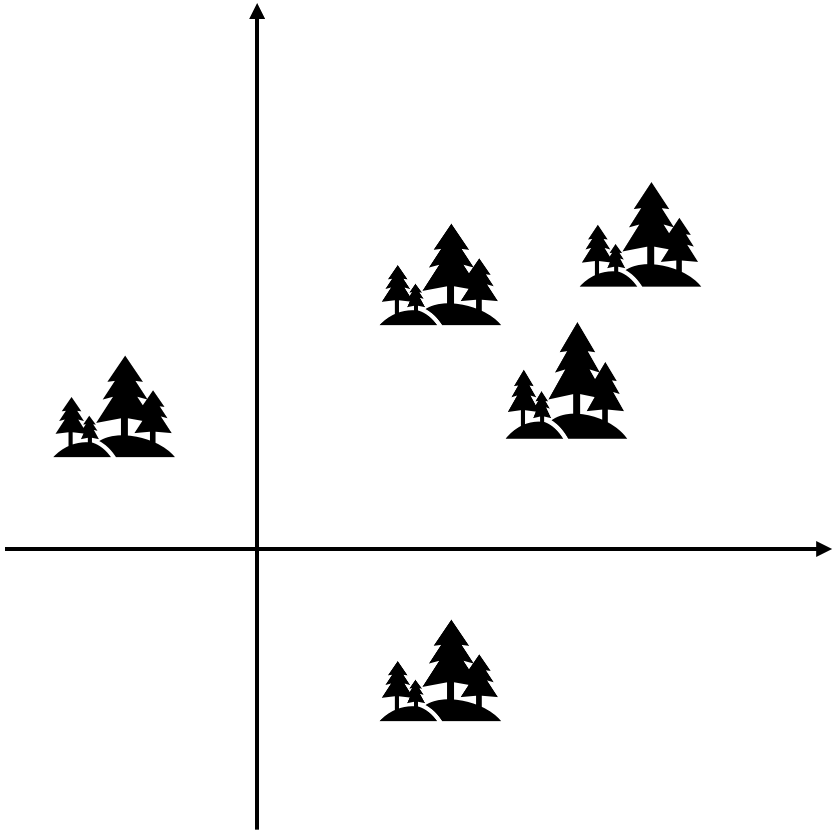
Vegan

phyloseq

Beta diversity (2) – multivariate analysis

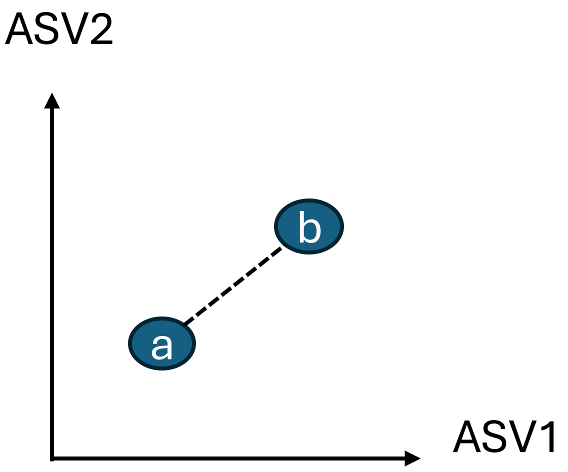
What is **multivariate analysis** in beta diversity context?

- It’s a group of statistical methods that we use to separate samples by community composition;
- These methods reduce many dimensions into two dimensions, for example, principal component analysis (PCA).



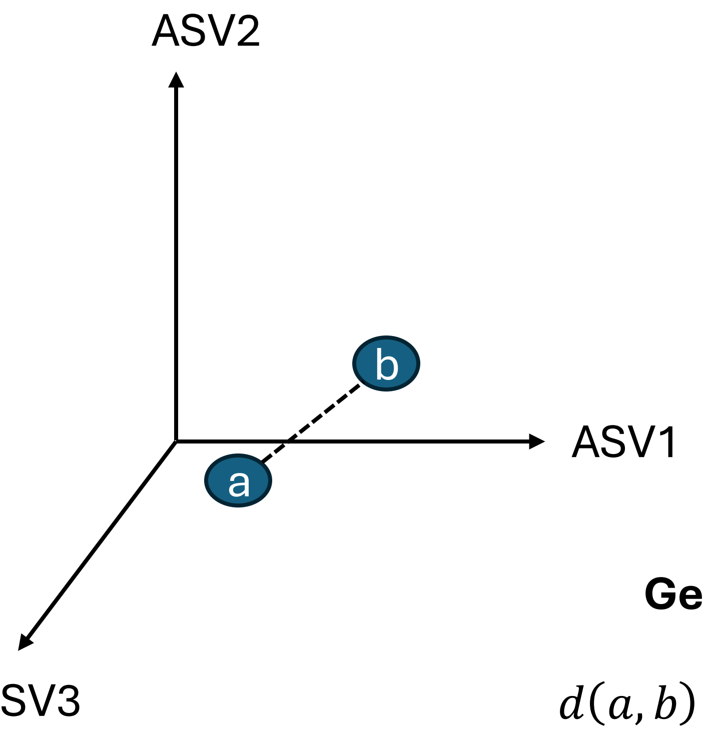
Beta diversity (2) – calculate dissimilarity

2 dimensions



$$d(a,b) = \sqrt{(a_{ASV1} - b_{ASV2})^2 + (a_{ASV2} - b_{ASV2})^2}$$

3 dimensions



$$d(a,b) = \sqrt{(a_{ASV1} - b_{ASV1})^2 + (a_{ASV2} - b_{ASV2})^2 + (a_{ASV3} - b_{ASV3})^2}$$

Step 1 – calculate distance between samples**:**

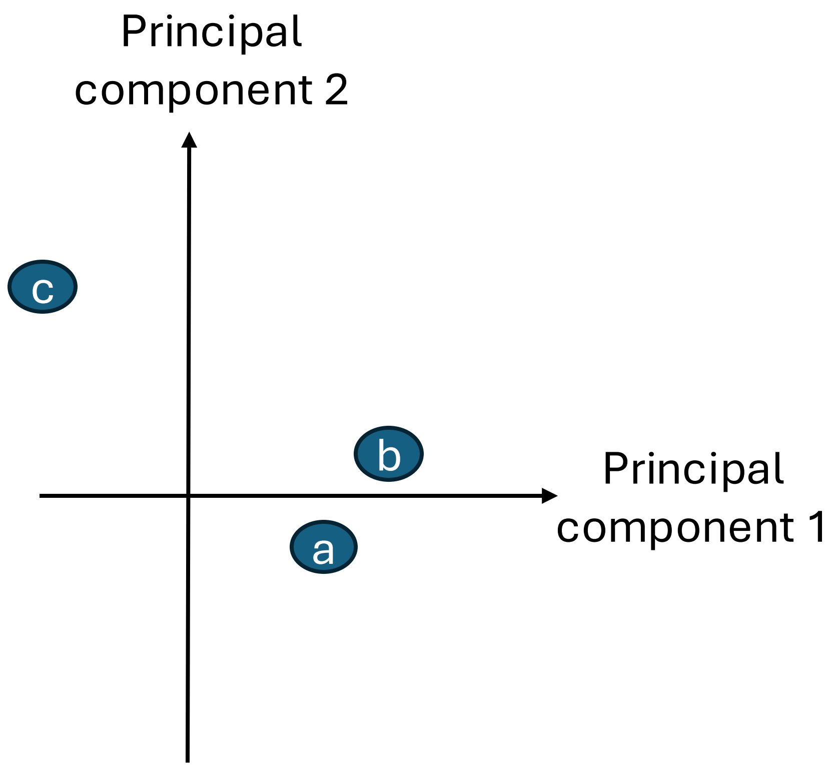
- Using **species abundance as orthogonal axis;**
- Two and three dimensions are easy to understand visually, but we use hundreds or thousands of dimensions;
- The process is repeated for all samples;
- Together, the axes represent the community composition;
- this step results in a **dissimilarity** matrix.

General formula (any number of dimensions)

$$d(a,b) = \sqrt{(a_{ASV1} - b_{ASV1})^2 + \cdots + (a_{ASVn} - b_{ASVn})^2}$$

Note: I'm using Euclidean distance as an example.

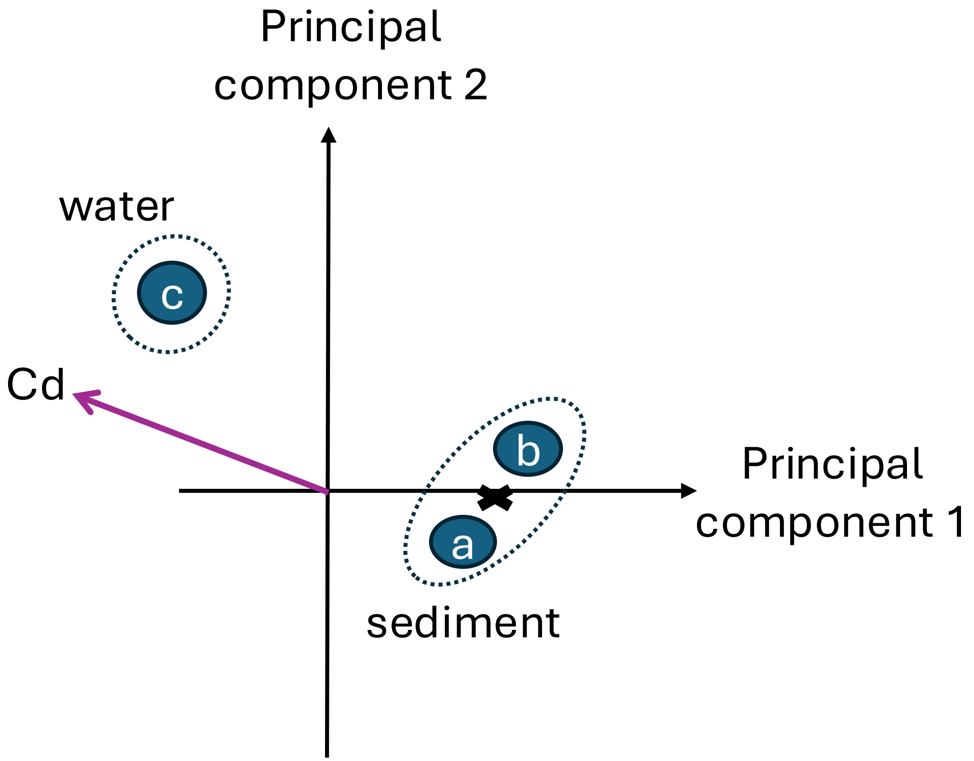
Beta diversity (3) – ordination plot



Step 2 - reduce **n** dimensions to **2** dimensions;

- the mathematical approach to reduce dimensionality is the **multivariate analysis method**;
- The plot of the multivariate analysis is the **ordination**;
- The two axis represent the community composition;
- The closer two samples are, the more similar they are, and vice-versa.

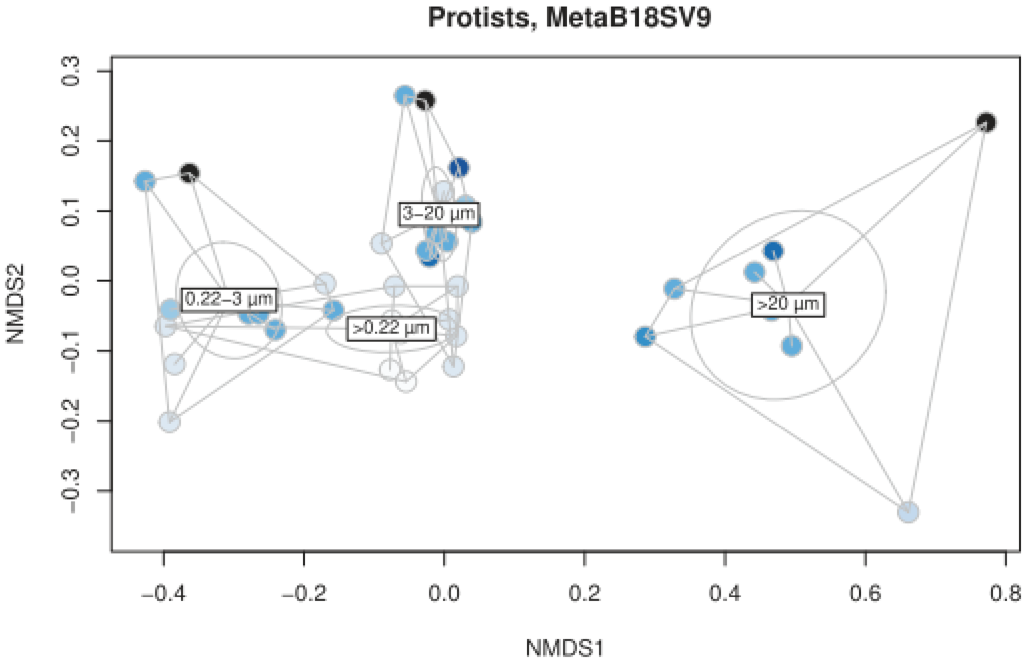
Beta diversity (4) – add environmental layers



Step 3 - Add environmental information

- we can add layers of environmental information on top of the ordination plot;
- **numerical** and **categorical** environmental variables are added differently;
- For numerical variables, we can use vectors or gradients;
- For categorical variables, we can group them explicitly (centroids, hulls, etc).

Beta diversity (5) – example



Pascoal F. et al., (2023). **Inter-comparison of marine microbiome sampling protocols.** *ISME Communications*.

Method used:

- Bray-Curtis distance (step 1);
- nMDS ordination plot (step 2);
- Centroids separate size fraction filtration (step 3 – categorical);
- Blue colors are proportional to filtered volume (step 3 – numerical);

What we can see:

- protist community composition was divided by size fractions;
- the effect of volume was contained within each size fraction, *i.e.,* filtration volume was less important than the size fraction.

Beta diversity (6) – dissimilarity options

Step 1 - Community dissimilarities:

- Euclidean distance (linear);
- Manhattan methods (non-linear);
  - Bray-Curtis/Sorenson;
  - Jaccard.
- Chi-squared;
- UniFrac
  - incorporates the phylogenetic distance.
  - Can be weighted or unweighted

These examples are generally good.

Suggestion for R

Vegan

phyloseq

Beta diversity (7) – dissimilarity options

Step 2 – Multivariate analysis:

- Eigenvector methods
  - Ordination by rotation and projection
    - PCA – Principal Component Analysis (linear)
      - Euclidean distance
    - CA – Correspondence Analysis
      - Chi-squared distance (weighted linear)
- Multidimensional Scaling (MDS)
  - takes any distance metric;
    - Principal Coordinates Analysis (PCoA) – classical MDS;
    - metric MDS – linear;
    - nMDS – non-linear.

Suggestion for R

Vegan

phyloseq

Beta diversity (8) – constrained vs unconstrained

Unconstrained methods:

- Calculate dissimilarities of community composition and make ordination plot;
- Add layers with environmental information afterwards;

Constrained methods:

- Incorporate the environmental information as constraints of the ordination.
- To do so, we need to specify a model:

community composition ~ variables

Examples:

- Redundancy Analysis (RDA) – constrained PCA;
- Constrained Correspondence Analysis (CCA) – constrained CA.

Note: if a model is constrained by all environmental variables, it becomes equivalent to the unconstrained version.

Suggestion for R

Vegan

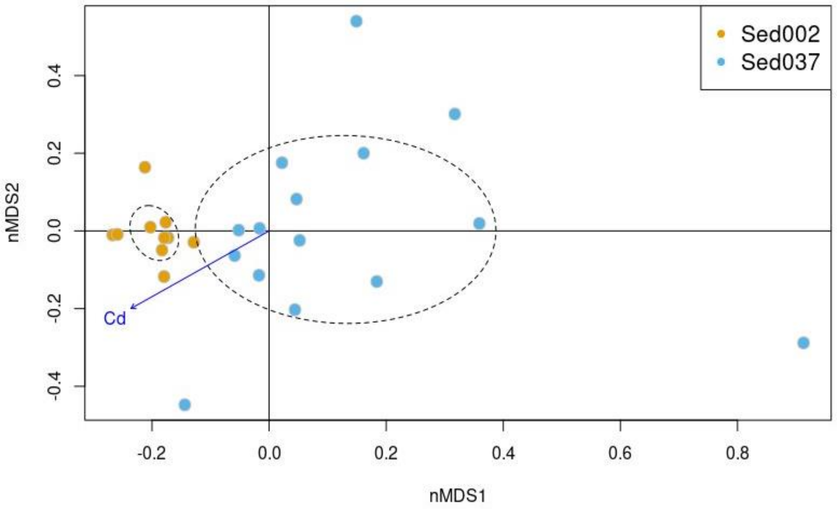
Beta diversity (9) – how to choose?

Some tips:

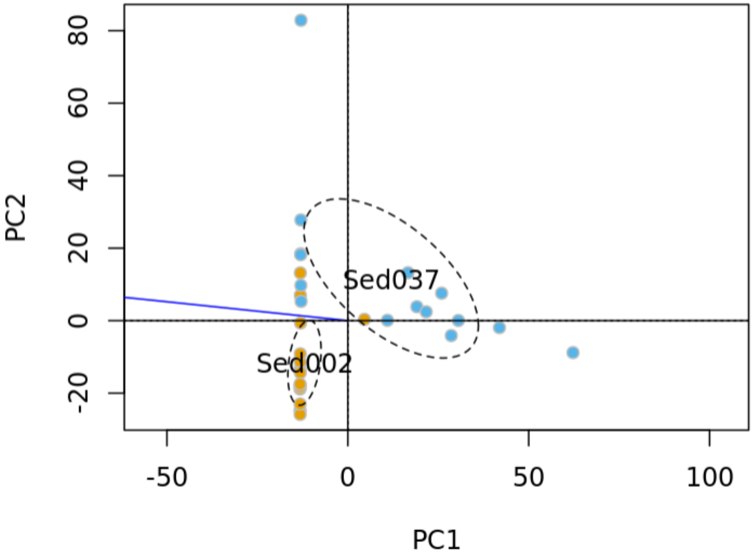
- Use the method that is a better description of your dataset;
- Experiment different options;
  - some methods favor abundant species, etc.
  - also experiment with normalization methods.
- There is no absolute best option;
- If you have domain knowledge of the interaction between a variable and your community, that knowledge can help you decide.

Beta diversity (9) – more examples

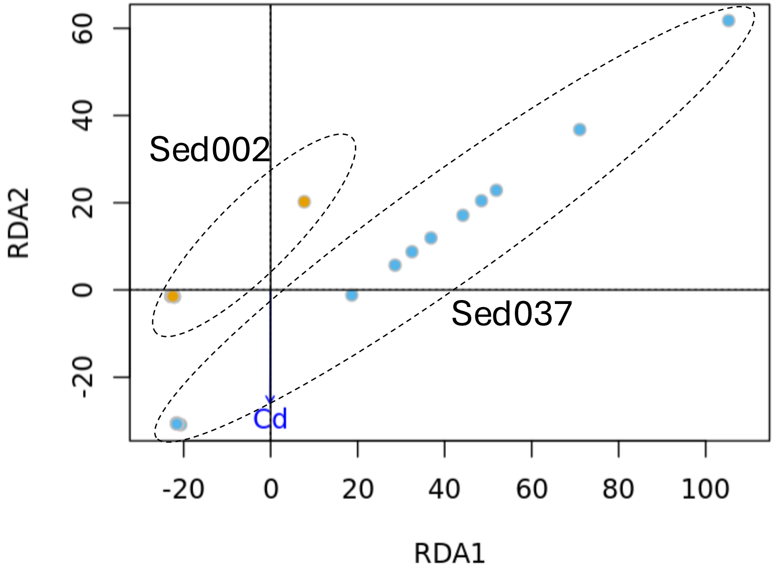
Bray-Curtis distance + nMDS



Euclidean distance + PCA (unconstrained)



Euclidean distance + RDA (constrained by Experiment)



Taxonomic analysis (1)

The taxonomy information will provide a more qualitative view of your microbial community.

Some general tips:

- Identify the most abundant phyla;
- Illustrate relative proportion of different groups within a taxonomic level;
    - this is tricky.
- Balance the depth of the analysis with the amount of information;
- Look at groups of interest for more detailed insights.

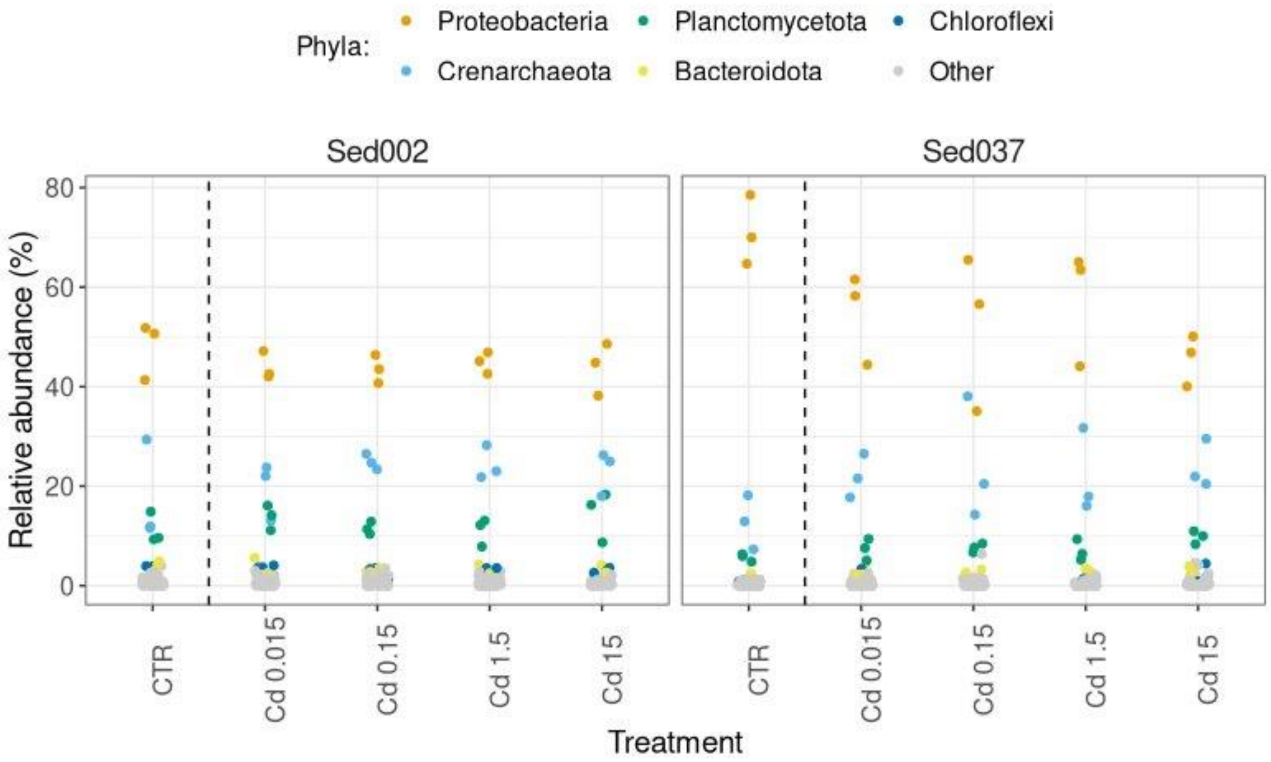
Taxonomic analysis (2) – examples

Balance:

- Overview of major groups;
- Vs detailed information.

For example:

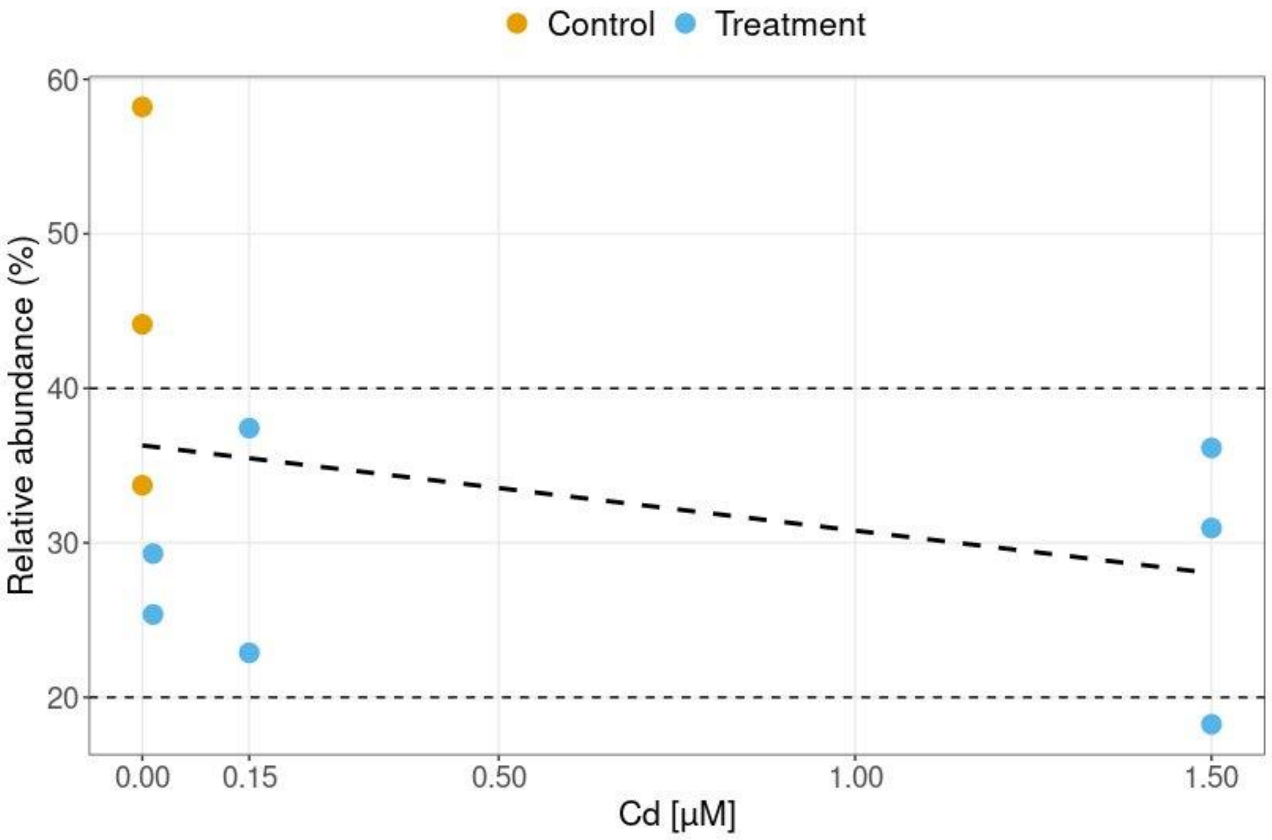
- Describe the most abundant phyla;
- Not very useful, but good starting point.



Taxonomic analysis (3) – examples

More specific analysis:

- *Stenotrophomonas* genus relative abundance as a function of Cd concentration.
- The abundance of *Stenotrophomonas* genus decreased after Cd treatment.
- And was responsible for major differences between sediments.



Statistical tests (1)

Suggestion for R

rstatix R package

Common tasks and tests in alpha diversity:

- Correlation between **two variables**;
  - correlation analysis.
- Correlation between **more than two variables**;
  - correlation matrix.

Statistical tests (2)

Common tasks and tests in alpha diversity (cont.):

- Comparing the means of **two groups** of samples;
    - Student’s t-test (parametric option);
    - Wilcoxon test (non-parametric option, has many other names).
- Comparing the means of **more than two groups** of samples;
    - one-way/two-way ANOVA test (parametric option) + post-hoc;
    - Kruskall-Wallis test (non-parametric option) + post-hoc;

Statistical tests (3)

Notes on parametric tests:

- Parametric tests are more powerful, so they should be preferred, when possible;
- Parametric tests require:
  - normal distribution of data (use Shapiro-Wilk test);
  - homogeneity of variance (use Levene test).
- The pre-requisites are more important if you have a **small sample size** (n < 30 samples).

Statistical tests (4)

Common tasks and tests in beta diversity:

- Verify if community composition is different between groups of samples
  - PERMANOVA test – permutation of MANOVA.

Suggestion for R

Vegan

Recommend reading

- Swenson, N. G. *Functional and Phylogenetic Ecology in R*. *Use R!* (Springer New York, 2014). doi:10.1007/978-1-4614-9542-0.
- Kassambara, A. *Practical Statistics in R II - Comparing Groups: Numerical Variables*. (Datanovia, 2019).
- Oksanen, J. Multivariate analysis of ecological communities in R: vegan tutorial. *Trends in Ecology & Evolution* **3**, 121 (2015)

Thank you