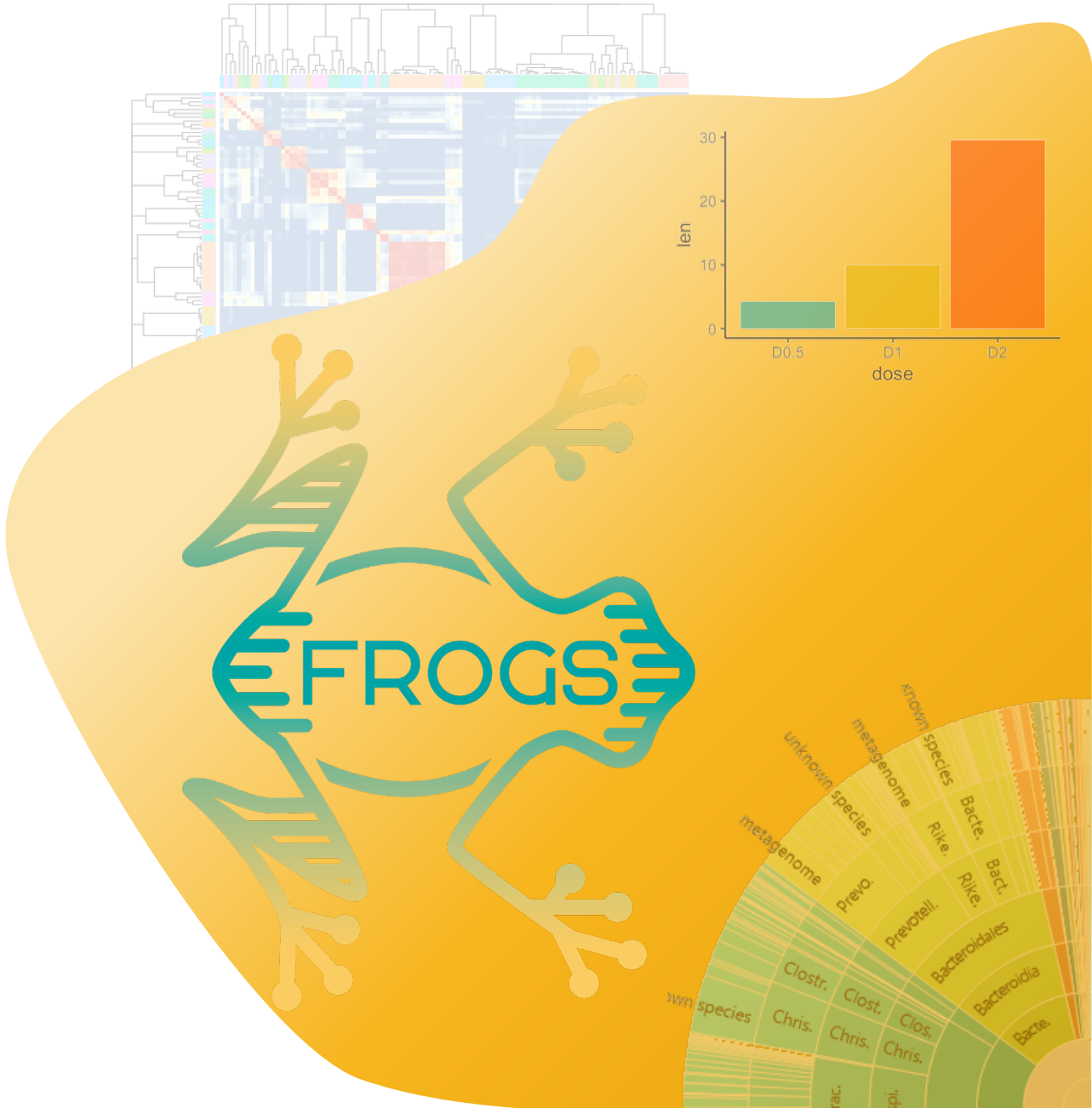
# Exploratory Data Analysis and Statistics

Lucas Auer, Gabryelle Agoutin, Maria Bernard, Géraldine Pascal, Maëlle Pomiès & Olivier Rué




What are the composition of the microbial communities?

How diverse and rich are the microbial communities?

At which taxonomic level do significant differences emerge?

Are samples in the same conditions similar?

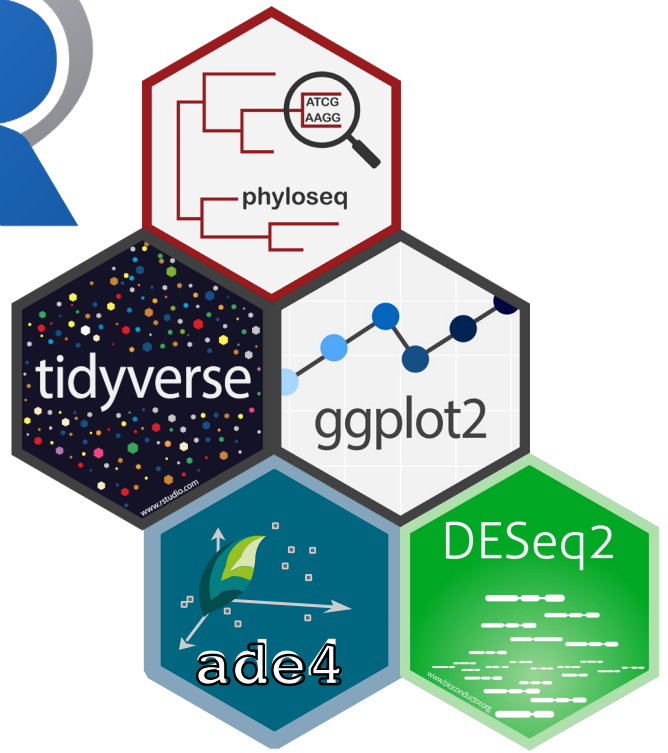
Does/how the samples organise according to conditions?

Which/how conditions/factors influence communities?

What are the content and structure of the metadata file?

Are there species with different abundances between conditions?

# Resources and packages




Exploratory Data Analysis is mainly based on phyloseq:

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

Paul J. McMurdie, Susan Holmes

Published: April 22, 2013 • https://doi.org/10.1371/journal.pone.0061217

It also relies on several additional R packages:

- *vegan* and *ade4* for community ecology analyses
- *ape* for phylogenetic tree handling
- *Ggplot2* for data visualisation
- *DESeq2* for differential abundance analysis
- A set of in-house scripts (accessible here : https://github.com/mahendra-mariadassou/phyloseq-extended/)

# Practice session

## Exercise: Available data


What data do we have at our disposal after running FROGS Core?

What are the content and structure of the metadata file?

# Practice session

## Exercise: Available data

What data do we have at our disposal after running FROGS Core?

- a FROGS .biom containing
  - an ASV count table
  - taxonomic descriptors when available
- (optional) a phylogenetic tree in Newick format
- a metadata file containing sample descriptions (in .tsv format)

# Practice session

## Exercise: Available data

What are the content and structure of the metadata file?

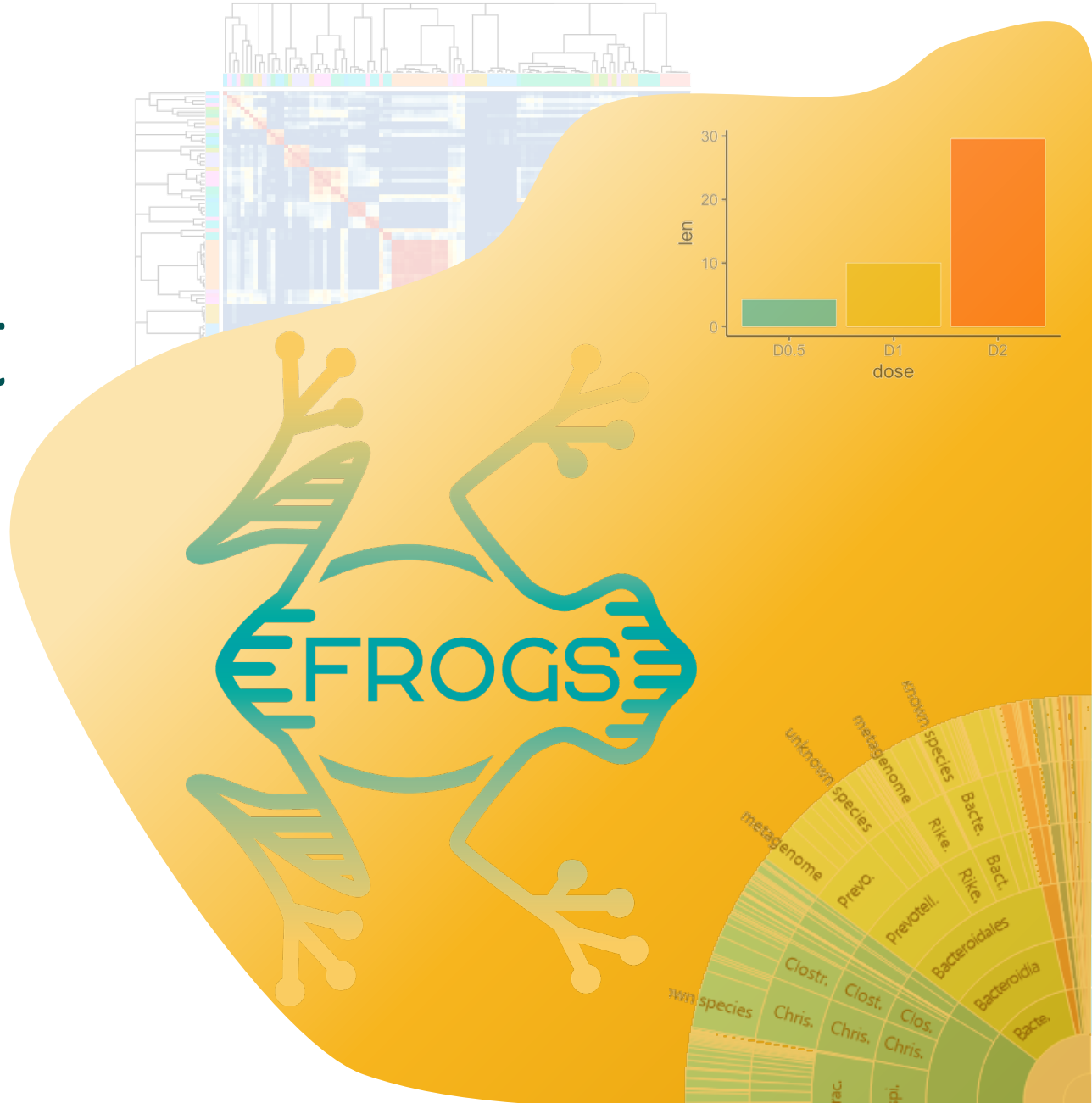
| Sample_ID | Tech_family | AOP | | Season | Season_code | | Localization | Dairy_species | French_Area |
|---|---|---|---|---|---|---|---|---|---|
| AOP1_PPC_S1 | PPC | AOP1 | 1 | Summer | S | 1 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_S2 | PPC | AOP1 | 1 | Summer | S | 2 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_S3 | PPC | AOP1 | 1 | Summer | S | 3 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_S4 | PPC | AOP1 | 1 | Summer | S | 4 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_S5 | PPC | AOP1 | 1 | Summer | S | 5 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_S6 | PPC | AOP1 | 1 | Summer | S | 6 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_W1 | PPC | AOP1 | 1 | Winter | W | 1 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_W2 | PPC | AOP1 | 1 | Winter | W | 2 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_W3 | PPC | AOP1 | 1 | Winter | W | 3 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_W4 | PPC | AOP1 | 1 | Winter | W | 4 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_W5 | PPC | AOP1 | 1 | Winter | W | 5 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP1_PPC_W6 | PPC | AOP1 | 1 | Winter | W | 6 | Rind | Cow | Bourgogne _Franche-Comte |
| AOP2_PPC_S1 | PPC | AOP2 | 2 | Summer | S | 1 | Rind | Cow | Auvergne_Rhone_Alpes |
| AOP2_PPC_S2 | PPC | AOP2 | 2 | Summer | S | 2 | Rind | Cow | Auvergne_Rhone_Alpes |
| AOP2_PPC_S3 | PPC | AOP2 | 2 | Summer | S | 3 | Rind | Cow | Auvergne_Rhone_Alpes |

A table with

- in the first column the sample IDs (identical to .fastq files names)
- Several other qualitative metadata (samples/experimental plan description):
  - Type of cheese (PPC, PPNC, PPS)
  - AOP
  - Season
  - …
- Quantitative external measures:
  - pH,
  - qPCR on bacteria or fungi
  - …

# FROGS Stat

# Phyloseq Import



# Practice session

## Import data in R Phyloseq format

With the R package phyloseq, all the different data (count table, taxonomy, metadata, and optionally phylogenetic tree) are regrouped in a single complex R object.
This object thus needs to be created with our FROGS Core outputs.

In FROGS tools, select FROGS Stat and choose Phyloseq: Import data

The required inputs are:
- A .biom file with count table and taxonomic annotations
- A metadata table (in tabular separated values, .tsv or .tabular)

Other information:
- (optional) a phylogenetic tree in Newick format (.nwk or .nhx)
- Names of the taxonomic ranks (in case of specific databanks)
- Rarefaction or not (subsampling to the smallest sequencing depth)

Choose subsample

# Practice session

## Exercise: Phyloseq Import outputs


What are Phyloseq Import outputs?

What are the differences between with and without normalisation?

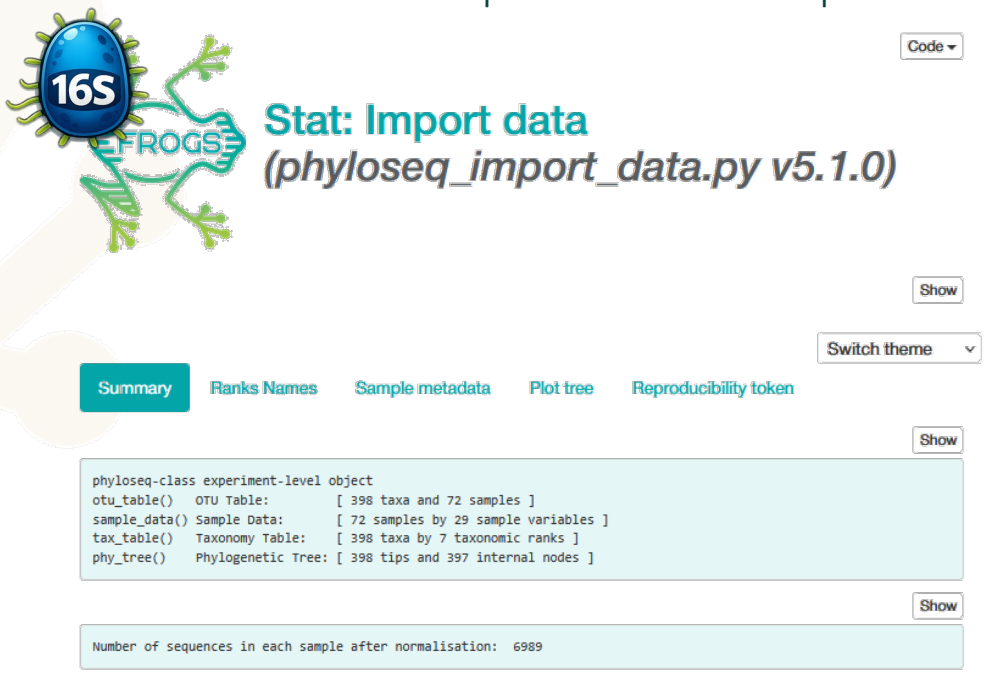
# Practice session

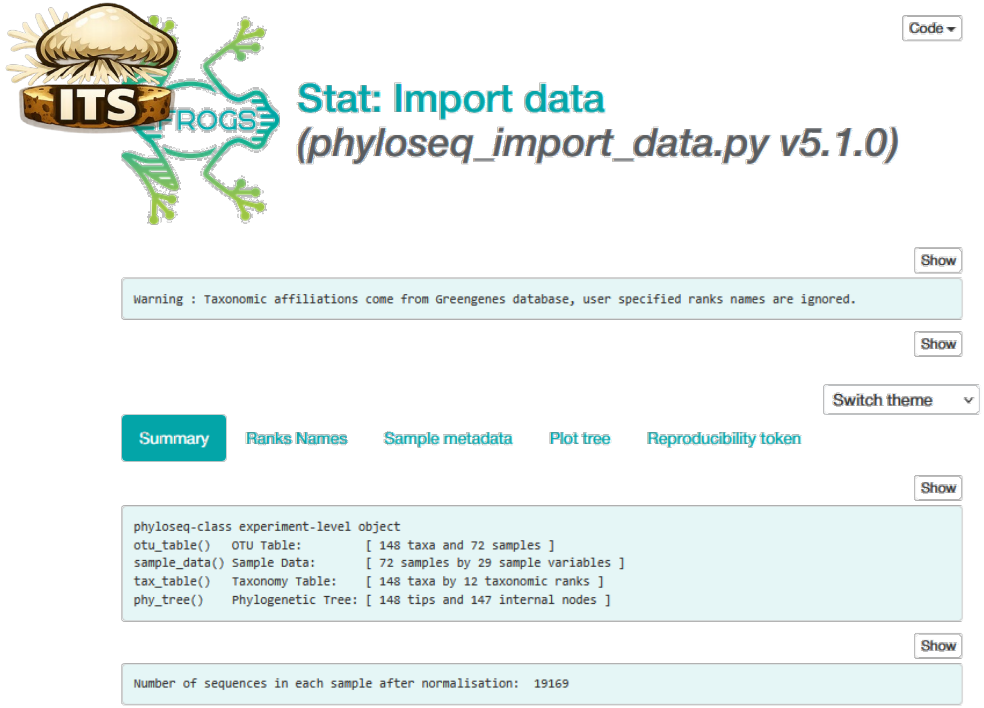
## Exercise: Phyloseq Import outputs

What are Phyloseq Import outputs?

“OTU” and otu_table() are Phyloseq names.

- an asv_data.Rdata file: an R object, here at the phyloseq format
- a .html report with a description of the phyloseq object



# Exercise: Phyloseq Import outputs

What are Phyloseq Import outputs?

Summary | **Ranks Names** | Sample metadata | Plot tree | Reproducibility token

Show

```
Rank names :  Kingdom, Phylum, Class, Order, Family, Genus, Species
```

Summary | Ranks Names | **Sample metadata** | Plot tree | Reproducibility token

Show

```
Sample variables:  Tech_family, AOP, AOP_code, Season, Season_code, Replicate, Localization, Dairy_species, French_Area, Ripening
_time, Wooden_shelf_use, Rind_treatment, pH, Cheese_aerobic_bacteria_counts, Cheese_fungi_counts, Cheese_lactic_acid_bacteria_cou
nts, study_accession, sample_accession, experiment_accession, run_accession, tax_id, scientific_name, fastq_ftp, original_Sample_
ID, PDO, original_AOP, Production, original_Replicate, SampleID
```

Show

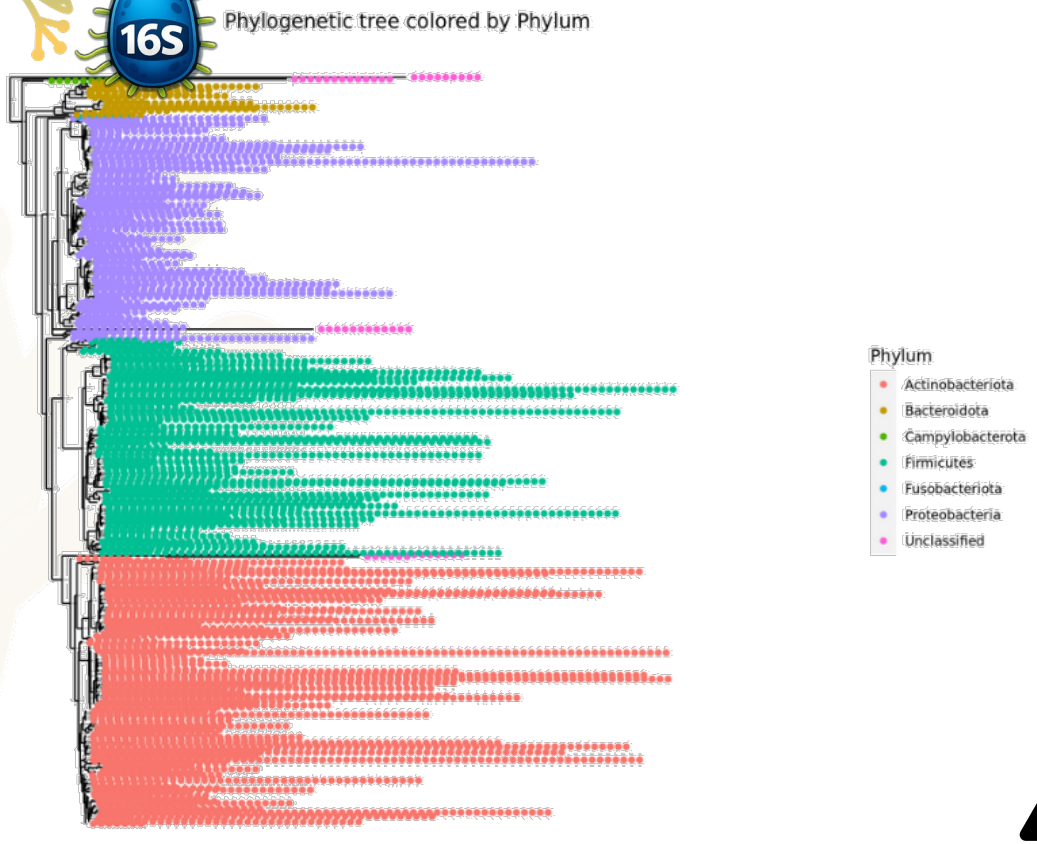
```
Tech_family :  PPC, PPNC, PPS

AOP :  AOP1, AOP2, AOP3, AOP4, AOP5, AOP6

AOP_code :  1, 2, 3, 4, 5, 6

Season :  Summer, Winter

Season_code :  S, W

Replicate :  1, 2, 3, 4, 5, 6

Localization :  Rind

Dairy_species :  Cow, Goat

French_Area :  Bourgogne _Franche-Comte, Auvergne_Rhone_Alpes
```

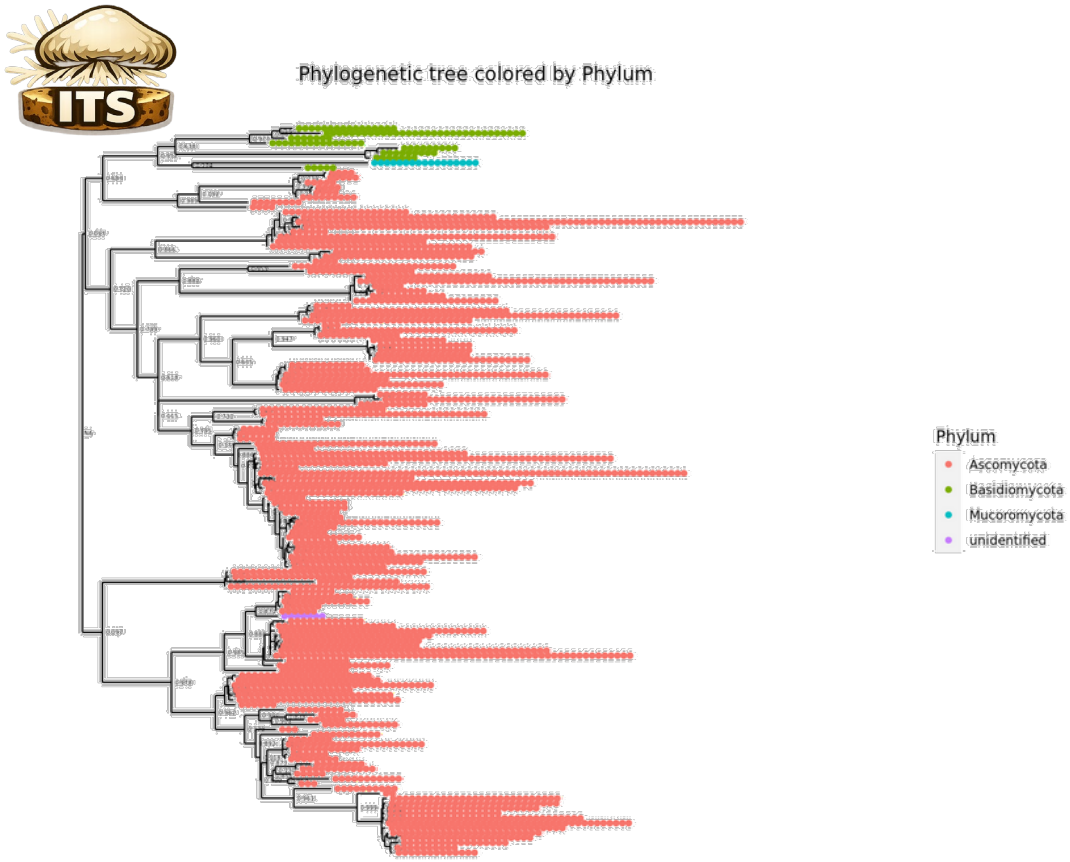
When organising a figure according to a variable, the order of apparition of its different modalities will be used.
It is thus important that this order make sense in the metadata file.

# Practice session

## Exercise: Phyloseq Import outputs

What are Phyloseq Import outputs?




For fungi, ITS (and particularly ITS length) variation does not always reflect true fungal phylogeny, and phylogenetic trees based on ITS may contain inconsistencies.

# Practice session

## Exercise: Phyloseq Import outputs

What are the differences between with and without normalization?

16S

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 398 taxa and 72 samples ]
sample_data() Sample Data:       [ 72 samples by 29 sample variables ]
tax_table()   Taxonomy Table:    [ 398 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 398 tips and 397 internal nodes ]
```

Show

```
Number of sequences in each sample after normalisation:  6989
```

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 398 taxa and 72 samples ]
sample_data() Sample Data:       [ 72 samples by 29 sample variables ]
tax_table()   Taxonomy Table:    [ 398 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 398 tips and 397 internal nodes ]
```

ITS

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 148 taxa and 72 samples ]
sample_data() Sample Data:       [ 72 samples by 29 sample variables ]
tax_table()   Taxonomy Table:    [ 148 taxa by 12 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 148 tips and 147 internal nodes ]
```

Show

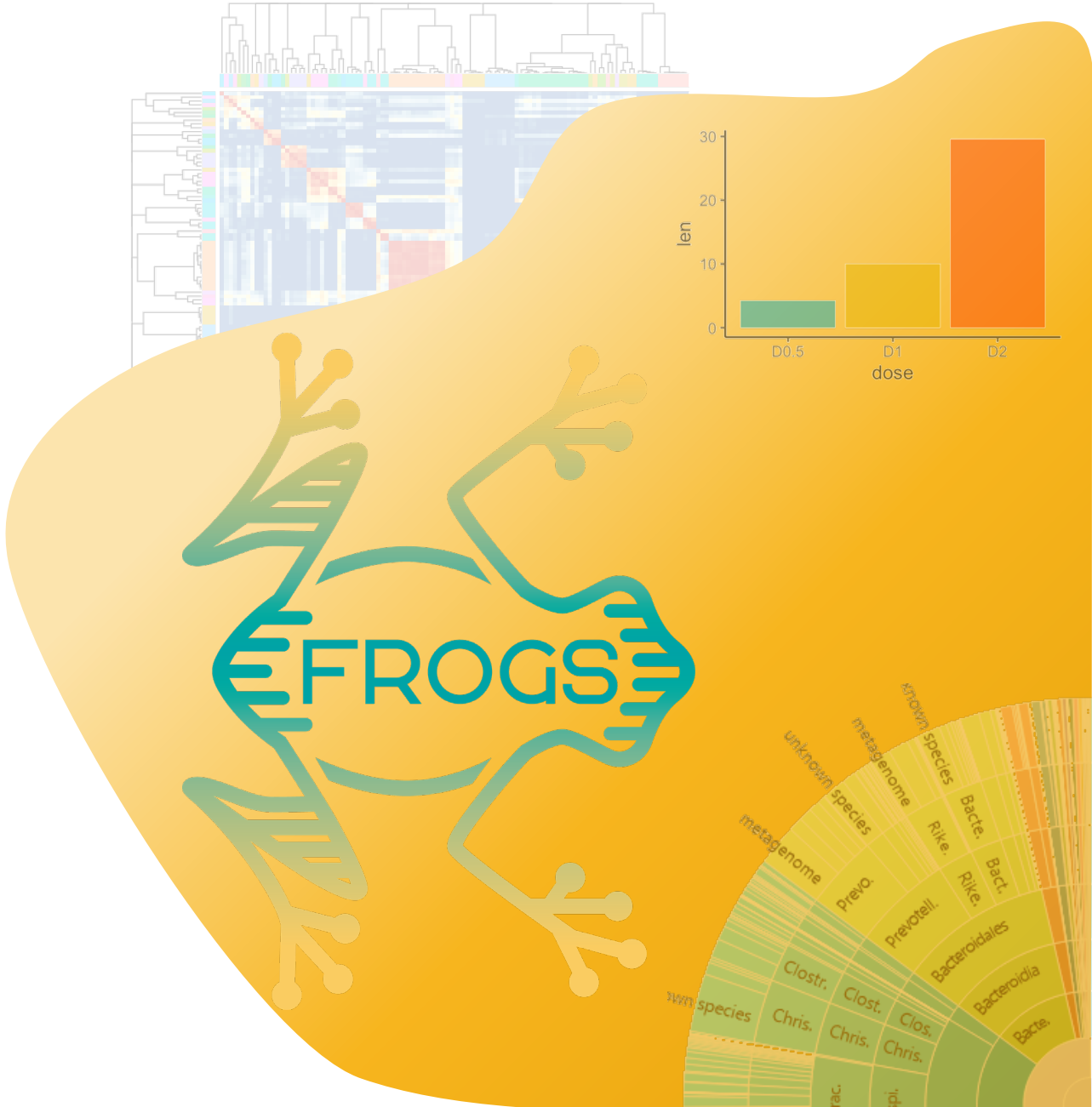
```
Number of sequences in each sample after normalisation:  19169
```

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 148 taxa and 72 samples ]
sample_data() Sample Data:       [ 72 samples by 29 sample variables ]
tax_table()   Taxonomy Table:    [ 148 taxa by 12 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 148 tips and 147 internal nodes ]
```

It is not the case here, but normalisation by subsampling can lead to a loss of taxa (rare taxa that randomly will not be kept), and some ASV may no longer satisfy the abundance filters thresholds (but won't be filtered again)

# FROGS Stat

# FROGS Stat & Easy16S



# FROGSFUNC & Easy16S

Historically FROGS provided its Exploratory Data Analysis tools through the Galaxy platform. Although they remain available and fully functional in the FROGS Stat tool, Galaxy is not ideal for interactive data exploration. To improve usability, colleagues developed a ShinyR application that brings all FROGS Stat tools (already coded in R) into a more responsive interface : Easy16S.

You still can configure the next FROGS Stat tools as we just did with Phyloseq Import

or **download the .Rdata file to continue outside Galaxy**


### Online web instance

https://shiny.migale.inrae.fr/app/easy16S

### Local installation

https://easy16s.migale.inrae.fr/

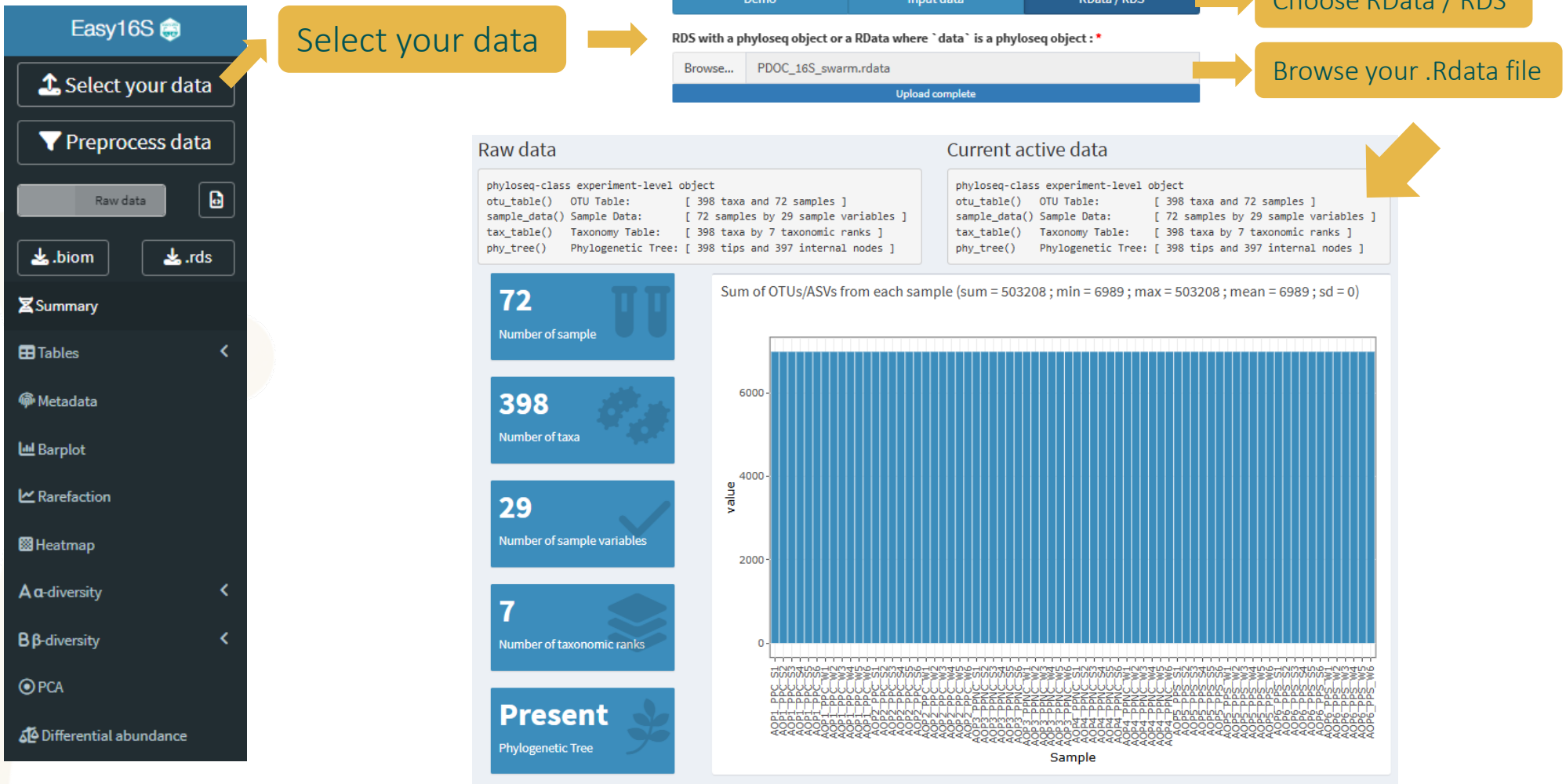
```
if (!requireNamespace("remotes", quietly = TRUE))
install.packages("remotes")
remotes::install_gitlab(repo="migale/easy16S@main", host="forge.inrae.fr")
easy16S::run_app()
```

Publication: Midoux et al. 2024 https://joss.theoj.org/papers/10.21105/joss.06704

# Practice session

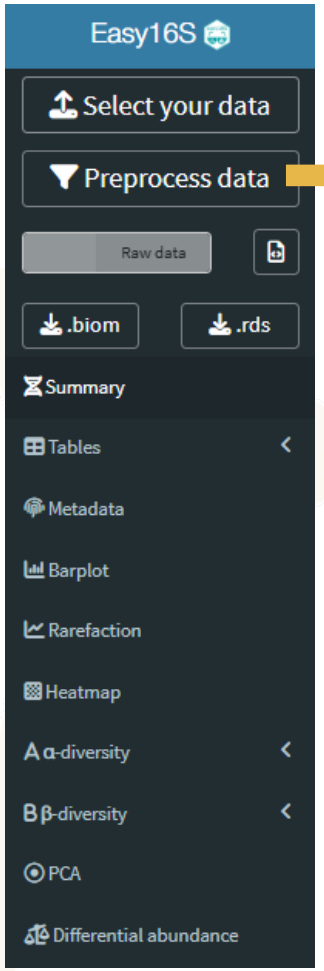
## Import .Rdata in Easy16S

Use the online instance (https://shiny.migale.inrae.fr/app/easy16S) or launch a local installation ( *easy16S::run_app()* )

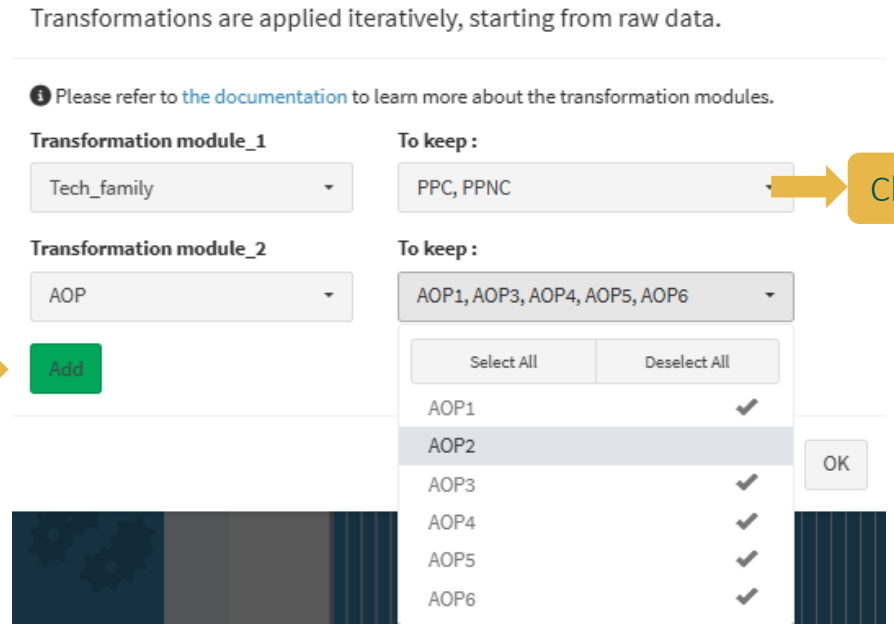


# Practice session

## Preprocess your data: select samples to keep for your analysis

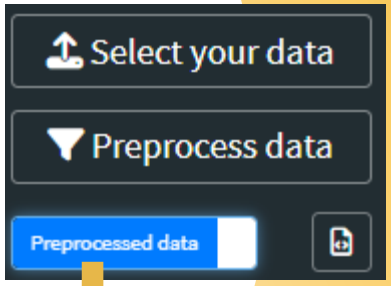



Preprocess




Choose samples to keep

With the Add button, choose the variable(s) to use to select your samples.
Here, only the PPC and PPNC modalities of the Tech_family will be selected, and the modality AOP2 of the AOP variable will be excluded.

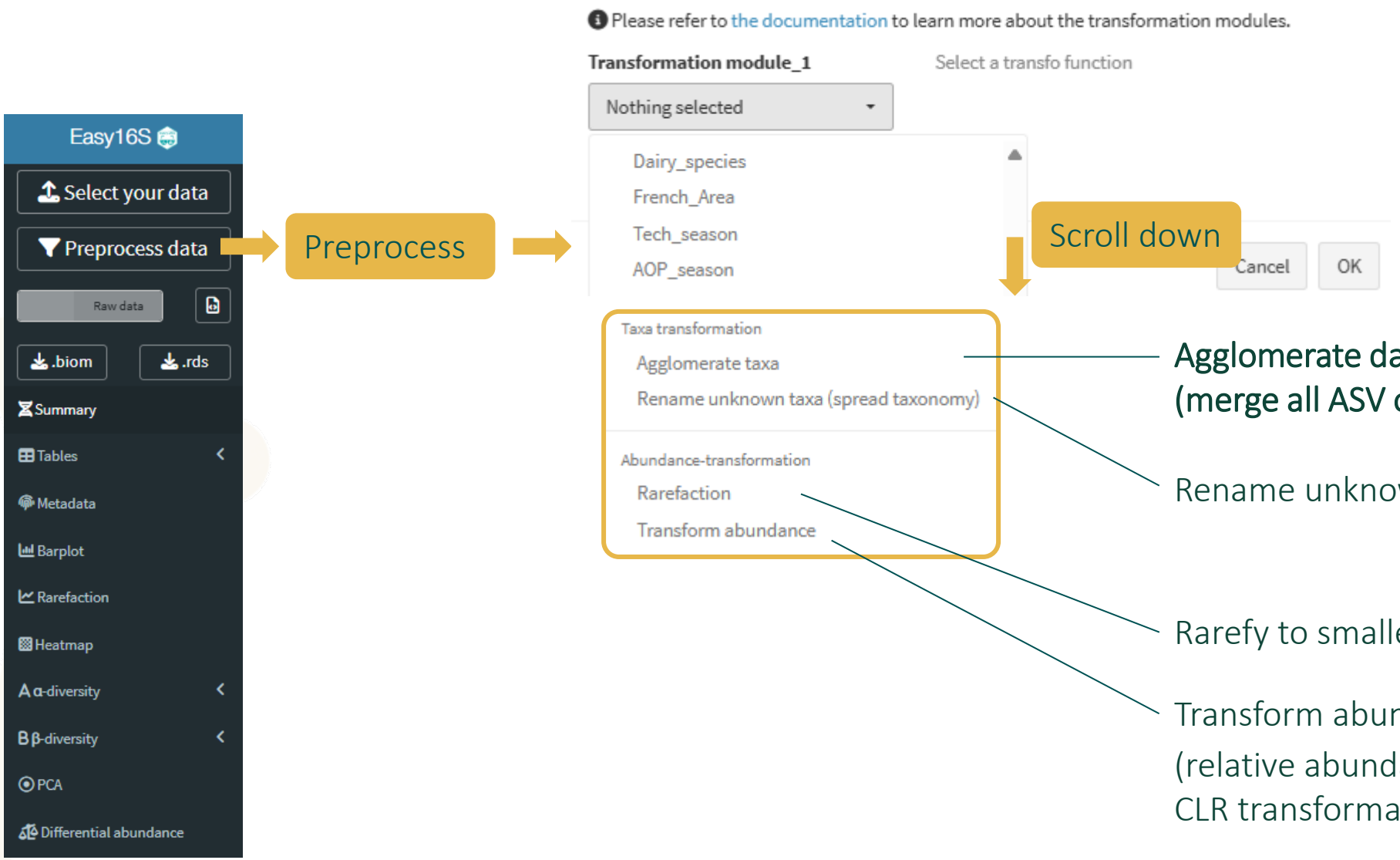



Switch between Preprocessed and raw data

It is ok to discard samples to perform separated analyses or figures,
but ALL THE SAMPLES used for swarm clustering or dada2 should be included in your future data deposit linked to your publication (for reproducibility issues).
It is thus recommended to do separated analyses for unrelated samples.

# Practice session

## Preprocess your data: transform your data



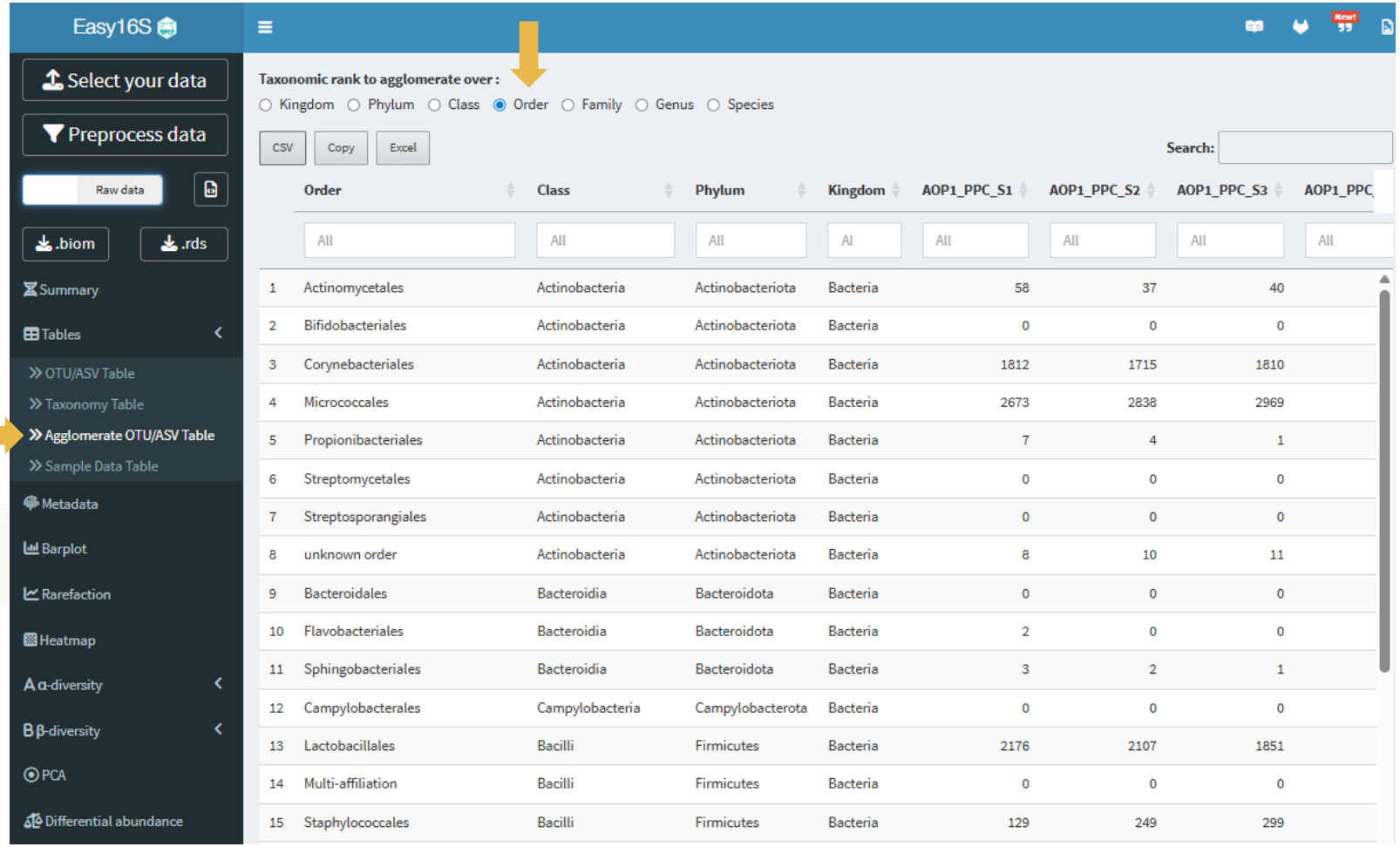

Agglomerate data at a taxonomic level (merge all ASV of the same species, genus, etc…)

Rename unknown taxa

Rarefy to smallest abundance

Transform abundances (relative abundances, $\sqrt{\ }$ transformation, CLR transformation)

# Abundance tables by taxonomic ranks

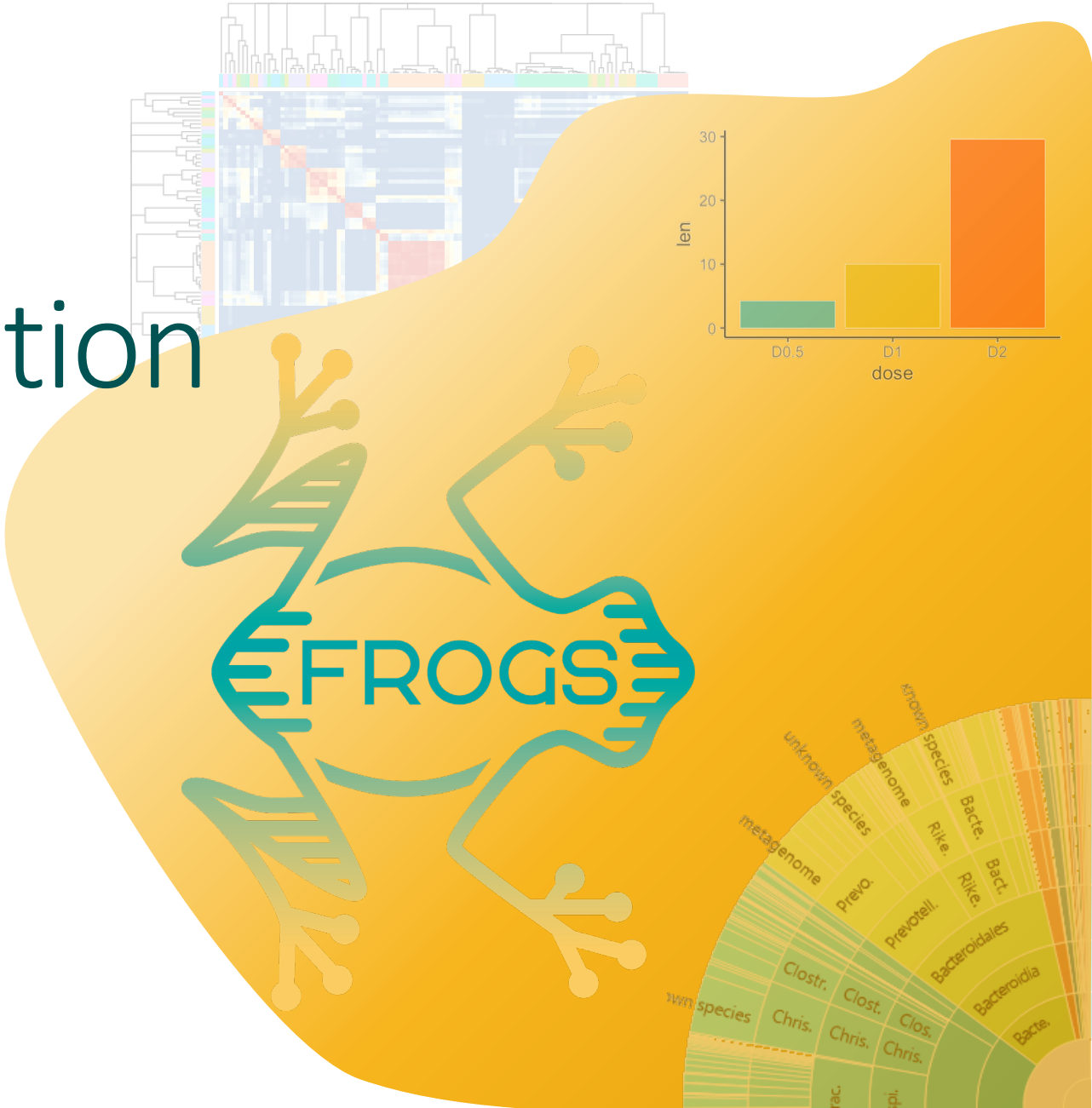



| | Order | Class | Phylum | Kingdom | AOP1_PPC_S1 | AOP1_PPC_S2 | AOP1_PPC_S3 | AOP1_PPC |
|---|---|---|---|---|---|---|---|---|
| | All | All | All | Al | All | All | All | All |
| 1 | Actinomycetales | Actinobacteria | Actinobacteriota | Bacteria | 58 | 37 | 40 | |
| 2 | Bifidobacteriales | Actinobacteria | Actinobacteriota | Bacteria | 0 | 0 | 0 | |
| 3 | Corynebacteriales | Actinobacteria | Actinobacteriota | Bacteria | 1812 | 1715 | 1810 | |
| 4 | Micrococcales | Actinobacteria | Actinobacteriota | Bacteria | 2673 | 2838 | 2969 | |
| 5 | Propionibacteriales | Actinobacteria | Actinobacteriota | Bacteria | 7 | 4 | 1 | |
| 6 | Streptomycetales | Actinobacteria | Actinobacteriota | Bacteria | 0 | 0 | 0 | |
| 7 | Streptosporangiales | Actinobacteria | Actinobacteriota | Bacteria | 0 | 0 | 0 | |
| 8 | unknown order | Actinobacteria | Actinobacteriota | Bacteria | 8 | 10 | 11 | |
| 9 | Bacteroidales | Bacteroidia | Bacteroidota | Bacteria | 0 | 0 | 0 | |
| 10 | Flavobacteriales | Bacteroidia | Bacteroidota | Bacteria | 2 | 0 | 0 | |
| 11 | Sphingobacteriales | Bacteroidia | Bacteroidota | Bacteria | 3 | 2 | 1 | |
| 12 | Campylobacterales | Campylobacteria | Campylobacterota | Bacteria | 0 | 0 | 0 | |
| 13 | Lactobacillales | Bacilli | Firmicutes | Bacteria | 2176 | 2107 | 1851 | |
| 14 | Multi-affiliation | Bacilli | Firmicutes | Bacteria | 0 | 0 | 0 | |
| 15 | Staphylococcales | Bacilli | Firmicutes | Bacteria | 129 | 249 | 299 | |

In Tables you can have access to the different tables (abundance, taxonomy and metadata) and to abundance tables agglomerated (summed) by taxonomic ranks.

# FROGS Stats

## Sample composition

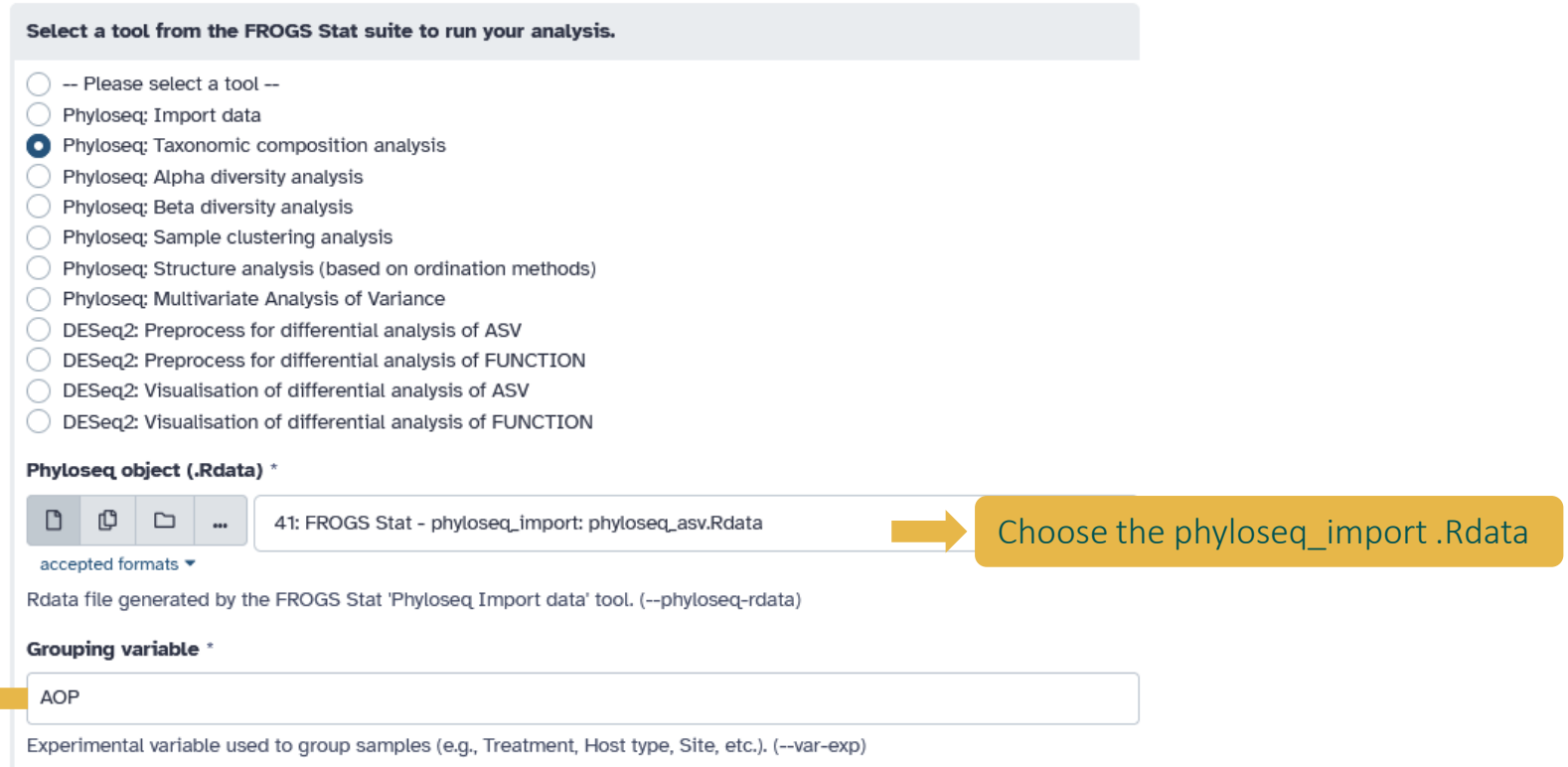


# Practice session

## Visualise abundances

In FROGS Stat select

Taxonomic composition analysis




Choose the phyloseq_import .Rdata

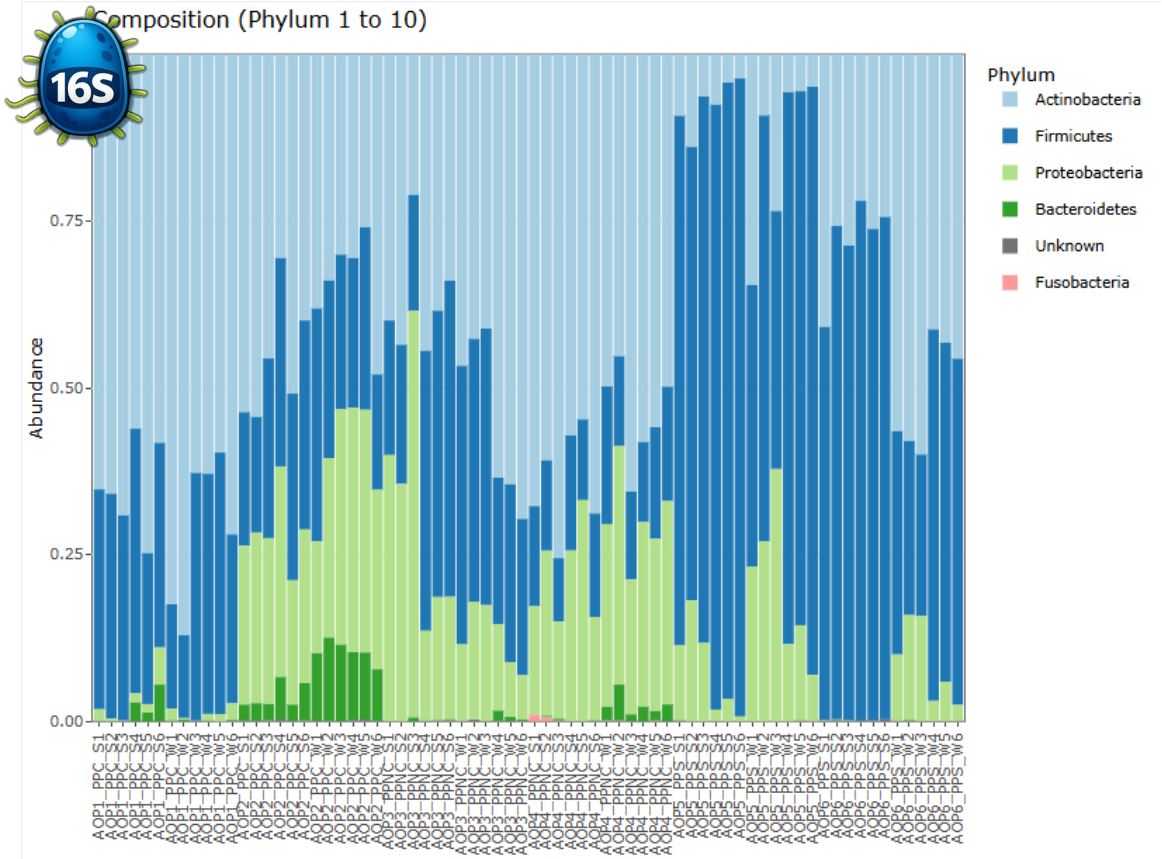
Choose a variable in the metadata (used to group samples)

The others parameters are used for filtering / focusing on specific taxonomies

# Practice session

## Visualise abundances

In Easy16S select **Barplot**




Settings :

**Taxonomic rank used for filtering :**

Root ○ Kingdom ○ Phylum ○ Class ○ Order
○ Family ○ Genus ○ Species ○ Rank7 ○ Rank6
○ Rank1 ○ Rank2 ○ Rank3 ○ Rank4 ○ Rank5

**Selected filter taxa :**

**Taxonomic rank used for coloring :**

○ Kingdom ● Phylum ○ Class ○ Order ○ Family
○ Genus ○ Species ○ Rank7 ○ Rank6 ○ Rank1
○ Rank2 ○ Rank3 ○ Rank4 ○ Rank5 ○ OTU

**Number of sub-taxa :**

1 10 30

1 4 7 10 13 16 19 22 25 28 30

**Subplot :**

...

Choose a rank and a taxa to filter (zoom on)

Choose a rank to color

Choose a number of taxa to color (the rest will be grouped in “Others”)

Choose a metadata variable to draw separate barplots

# Practice session

## Exercise: Sample composition visualisation

How to read the output?

What biological interpretation can be extracted?

# Practice session

## Exercise: Sample composition visualisation

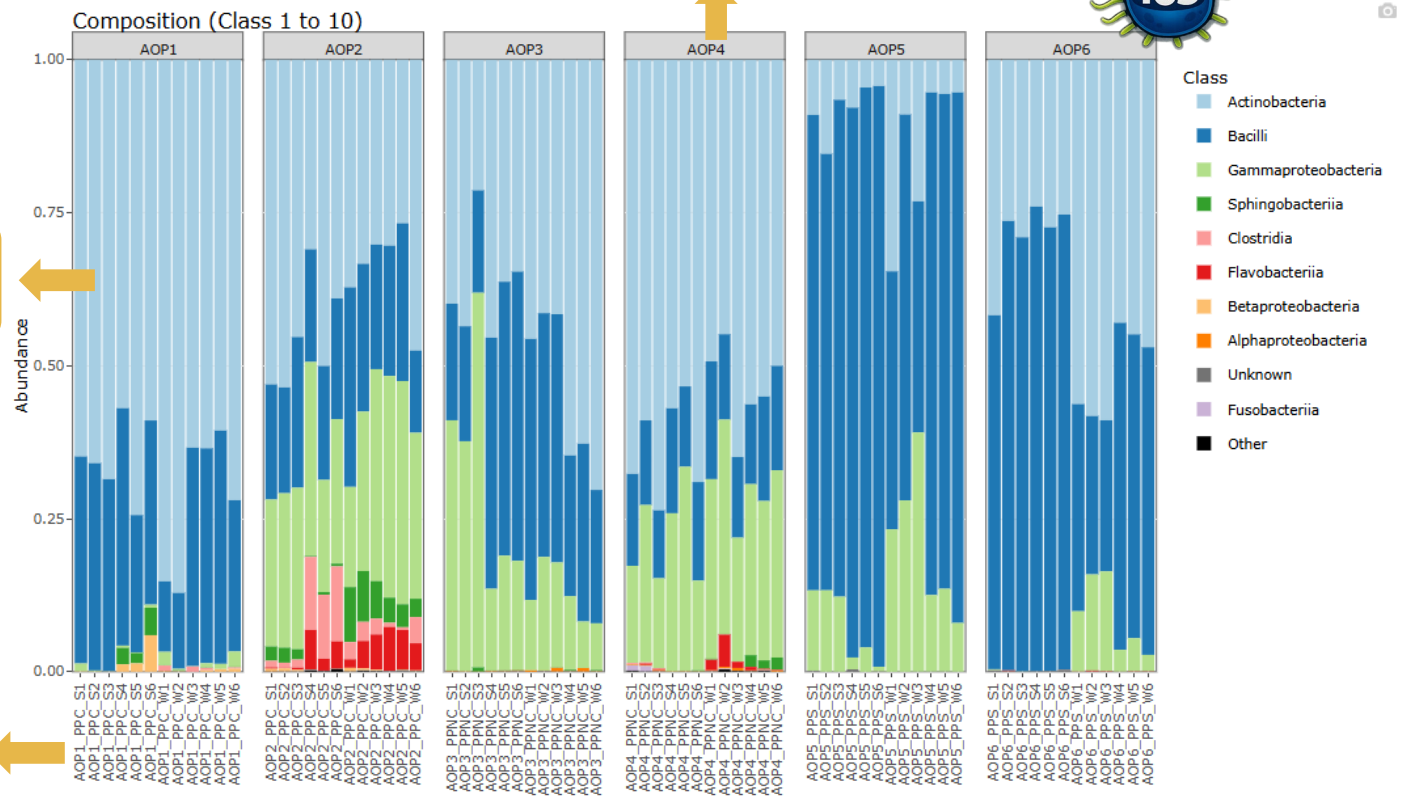
How to read the output?

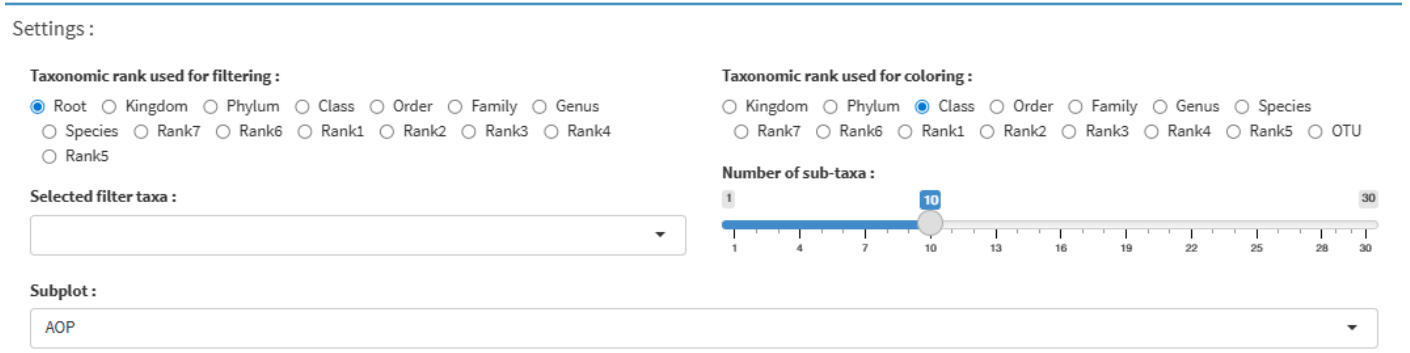
Relative abundances are stacked, colored by the chosen taxonomic rank

"Abundances" are amplicon abundances, not actual species abundances

Bars are grouped according to the variable chosen for subplot

Each bar is a sample

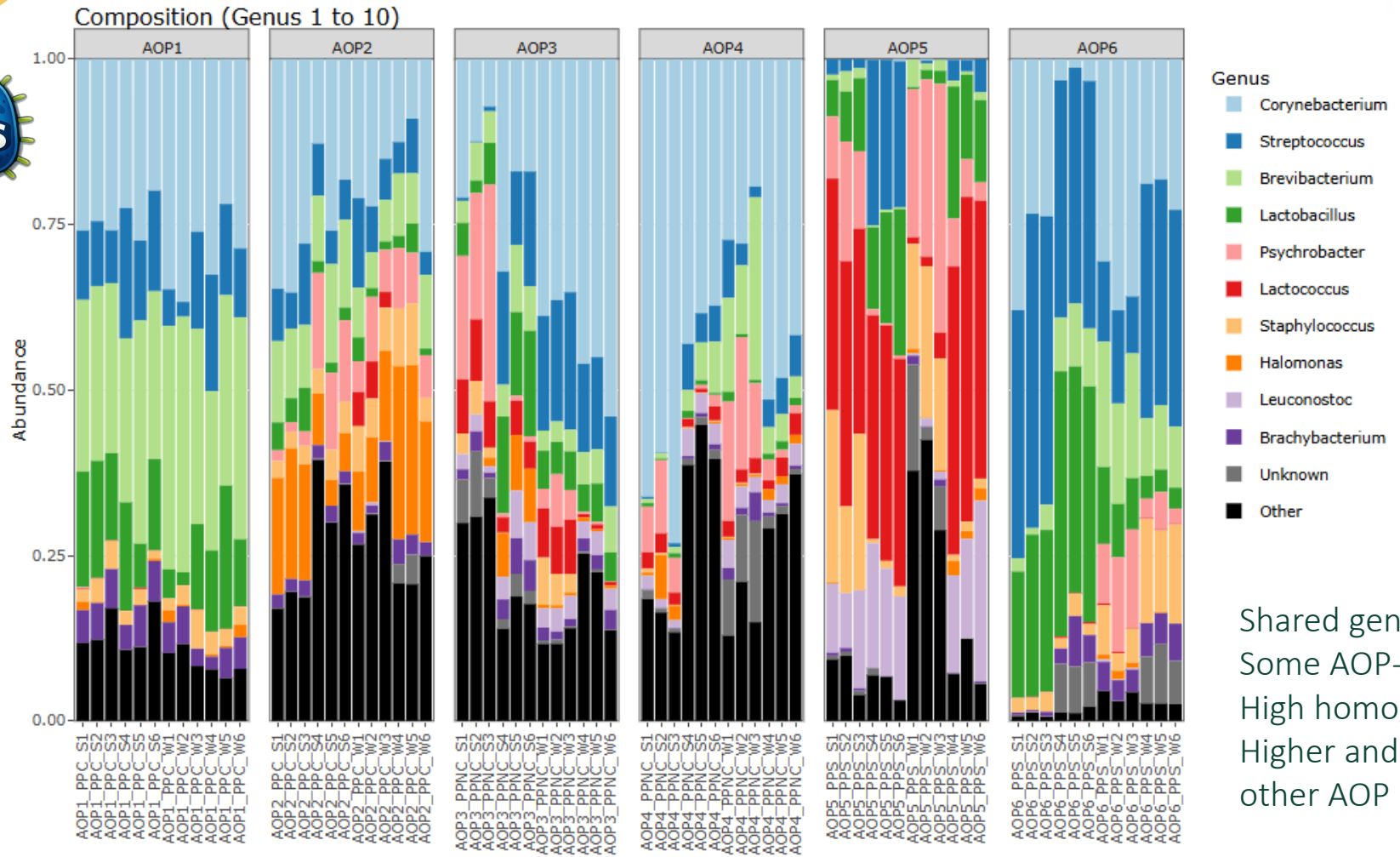


# Practice session

## Exercise: Sample composition visualisation

What biological interpretation can be extracted?

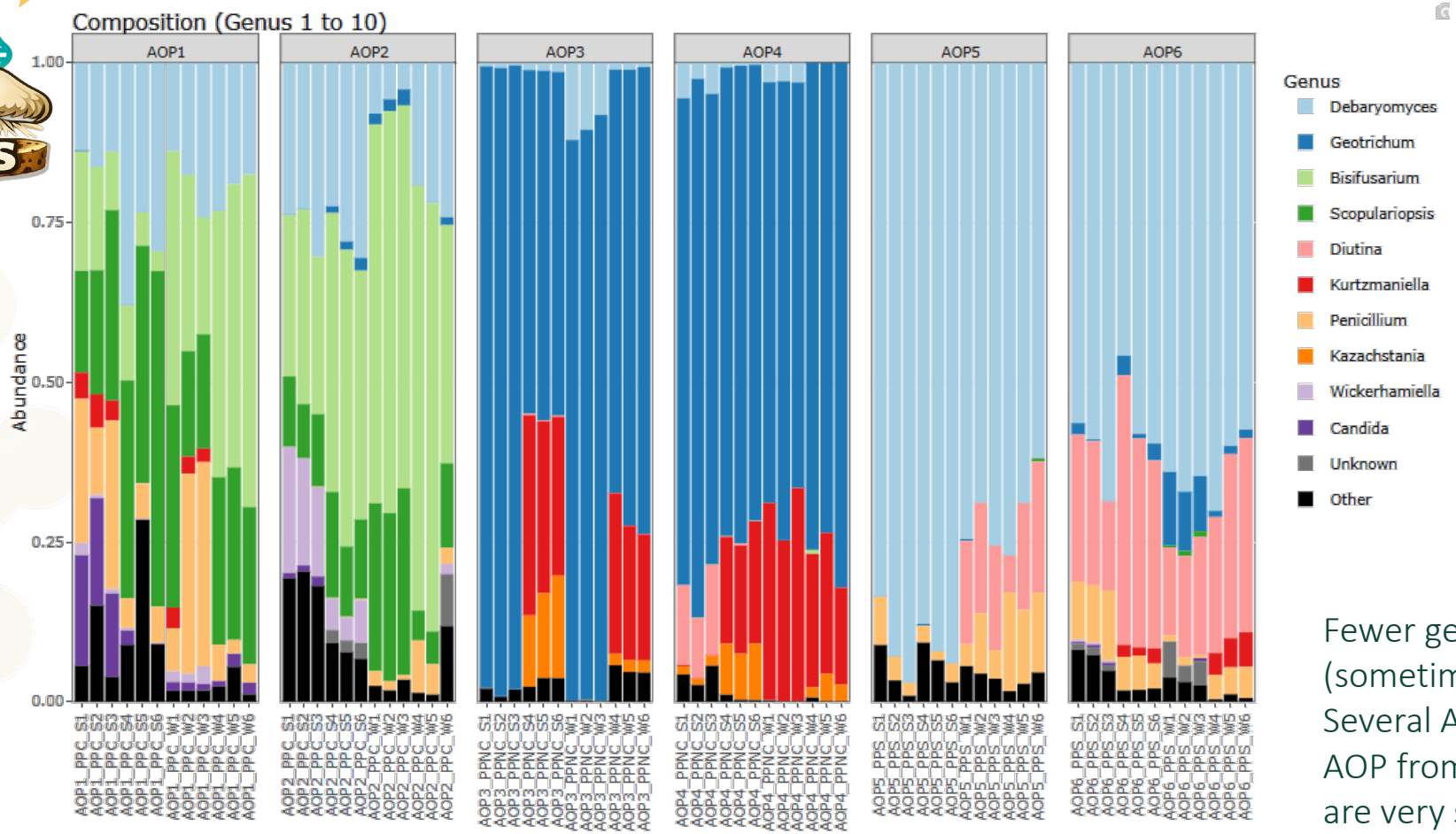



Shared genera (with variable abundances)
Some AOP-specific genera
High homogeneity in AOP1
Higher and variable heterogeneity in the other AOP

# Practice session

## Exercise: Sample composition visualisation

What biological interpretation can be extracted?




Fewer genera, with high abundances (sometime >90%)
Several AOP-specific genera
AOP from the same technological groups are very similar

# Practice session

## Exercise: Visualisation of details of composition


What is the composition inside the Phylum with the highest count? Inside the Genus?

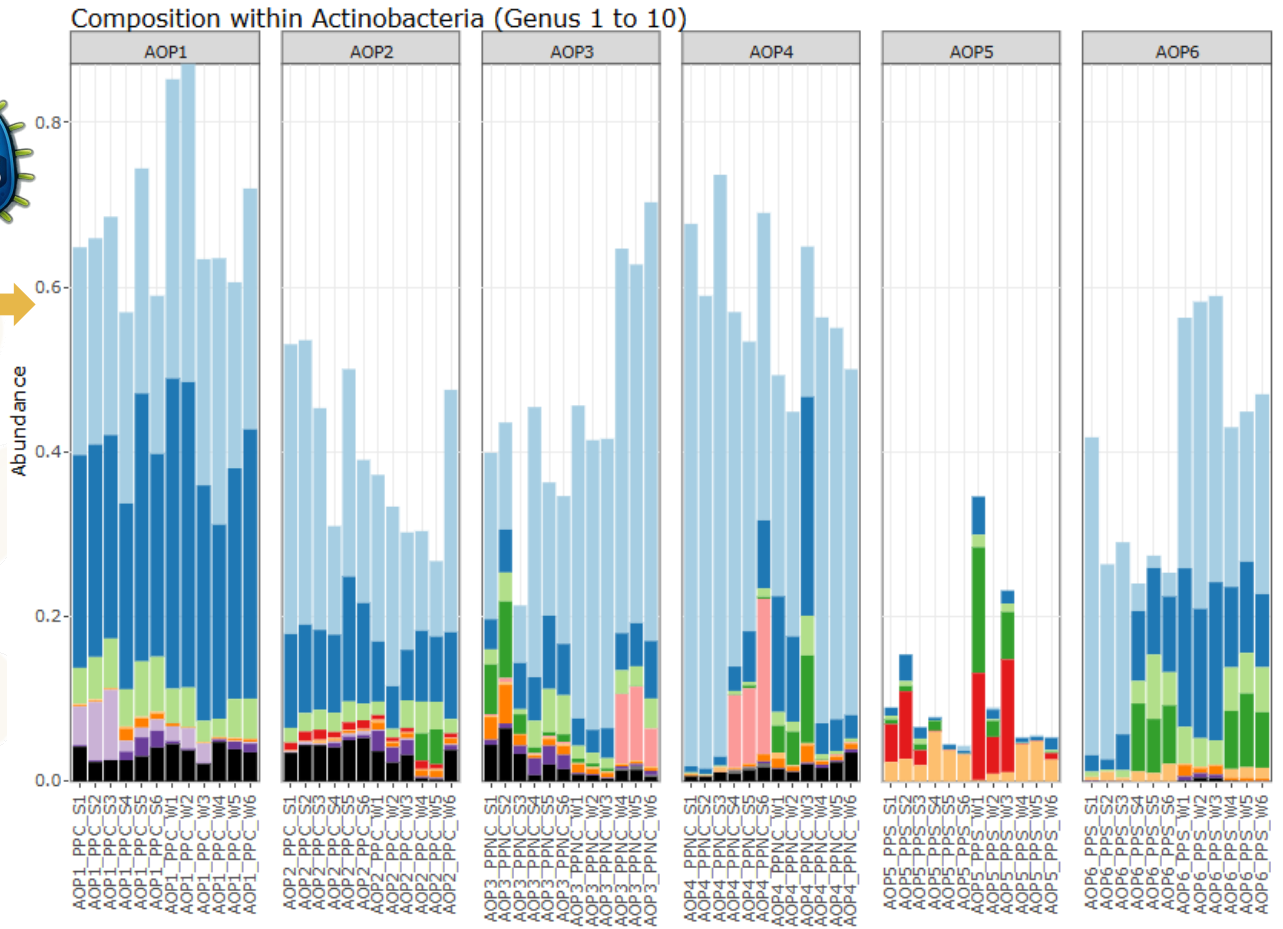
Are there seasonal effects?

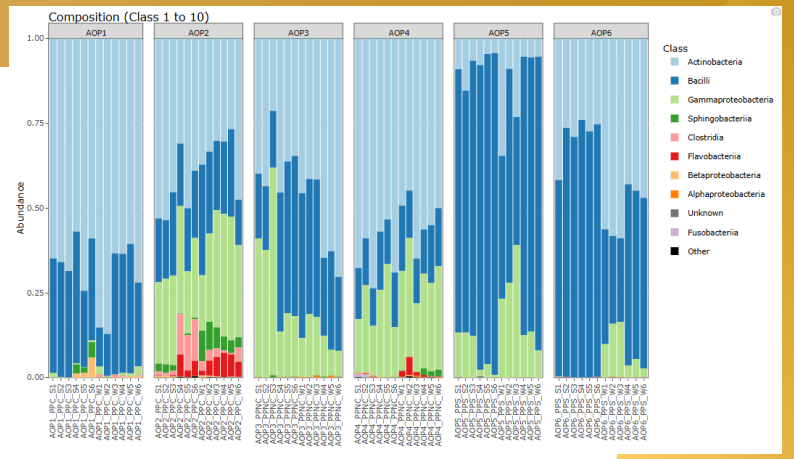
# Practice session

## Exercise: Visualisation of details of composition

What is the composition inside the Phylum with the highest count?

Scales adapt to max abundance



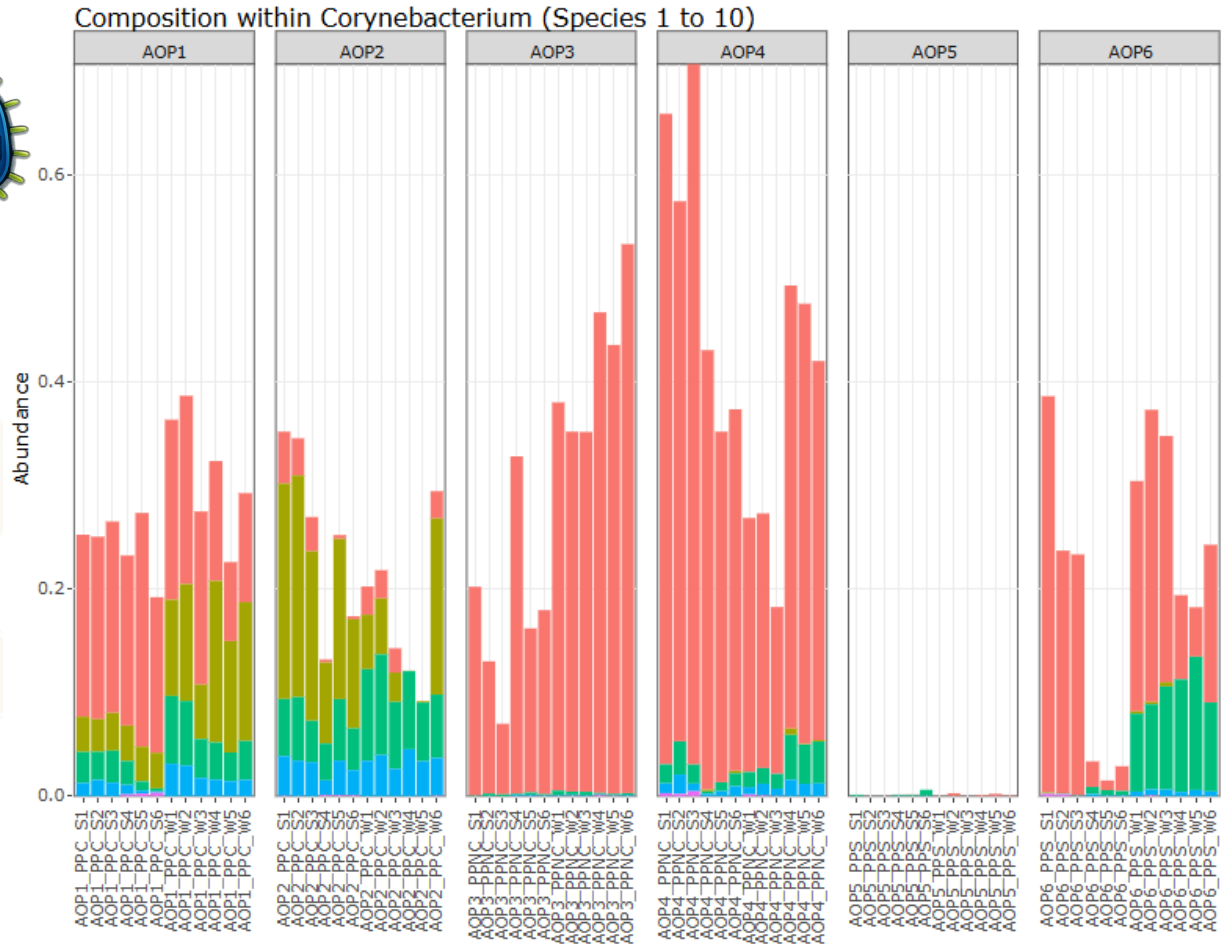



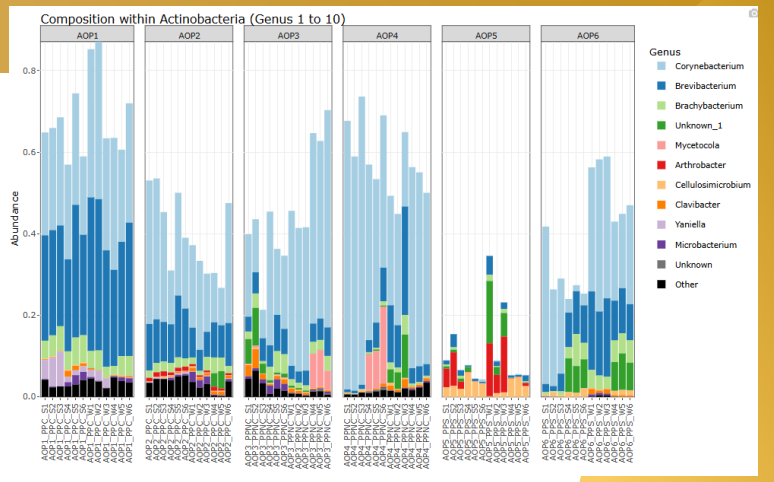

Comparable AOP in phylum composition can display different genus composition inside phyla. For example, AOP5 has lower Actinobacteria levels and a completely different composition

# Practice session

## Exercise: Visualisation of details of composition

What is the composition inside the Genus with the highest count?



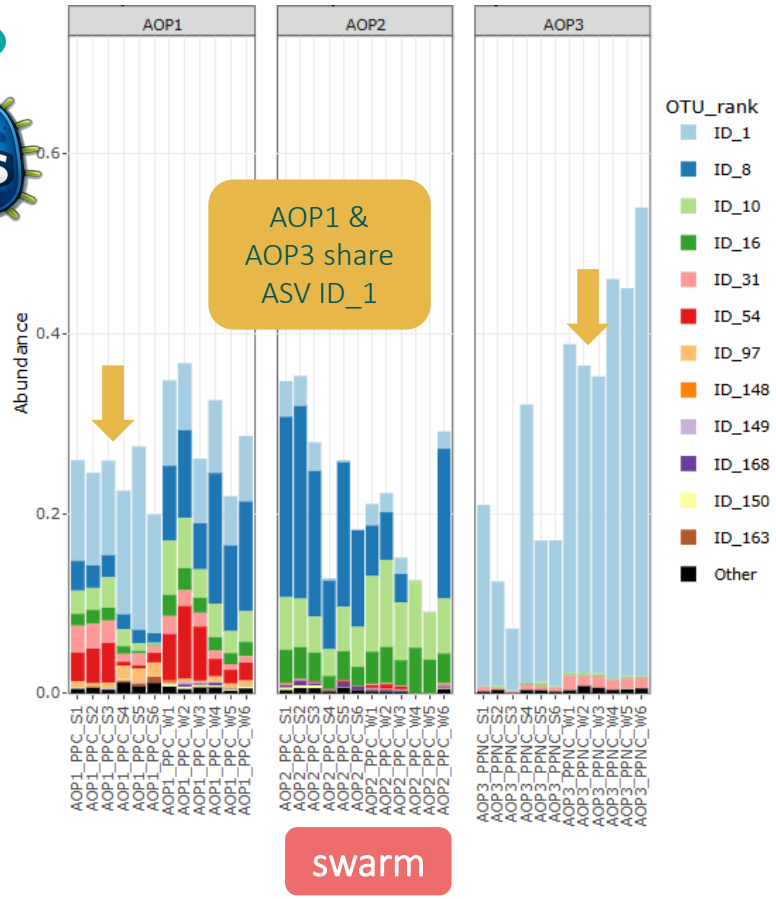

AOP3 and AOP4 had similarities in Corynebacterium levels, but different species compositions

swarm and dada2 produce very similar results at Genus and species level

# Practice session

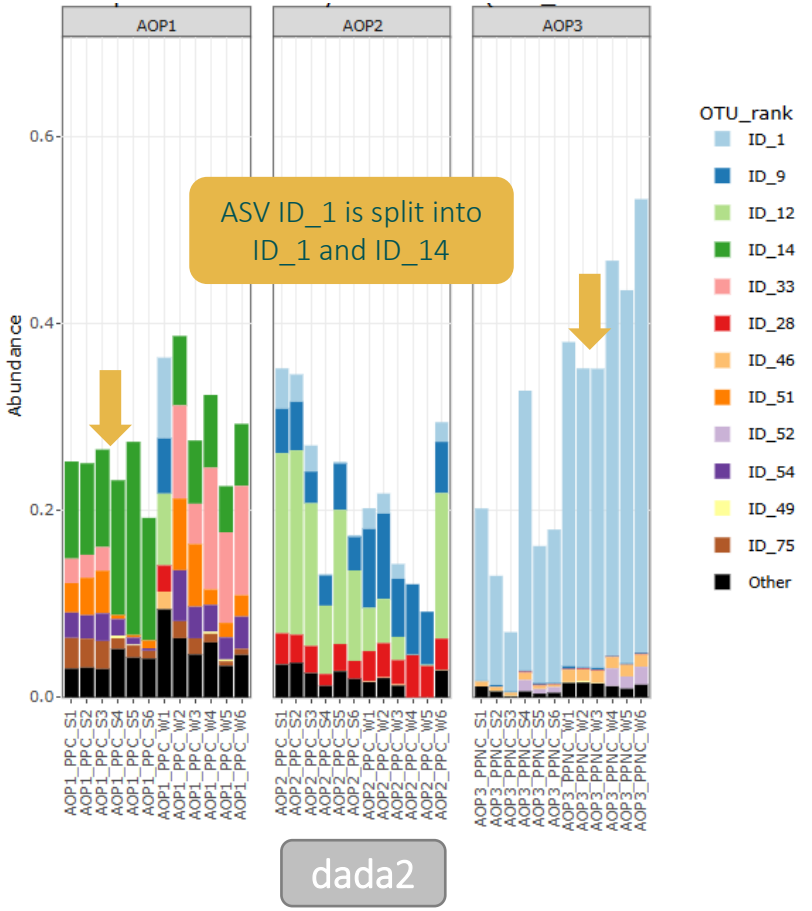
## Exercise: Visualisation of details of composition

What is the composition inside the Genus with the highest count?



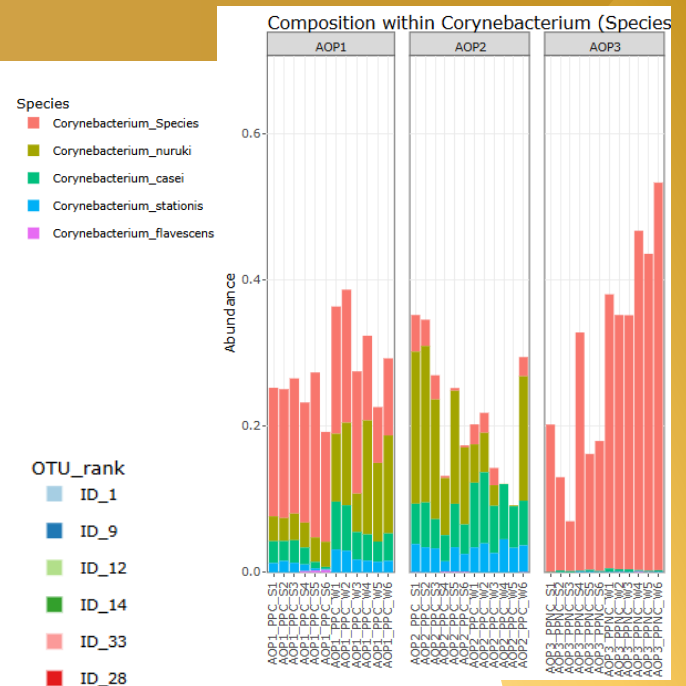

AOP1 & AOP3 share ASV ID_1

swarm




ASV ID_1 is split into ID_1 and ID_14

dada2

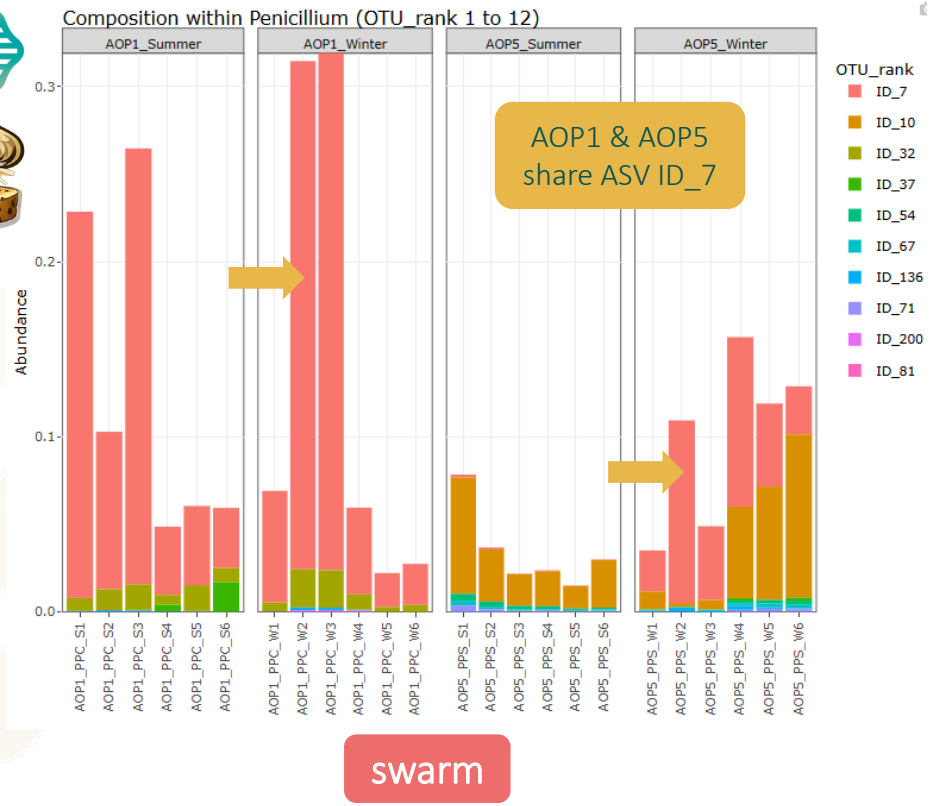



dada2 can produce more ASV than swarm inside a species (and these extra ASV could have biological consistency).

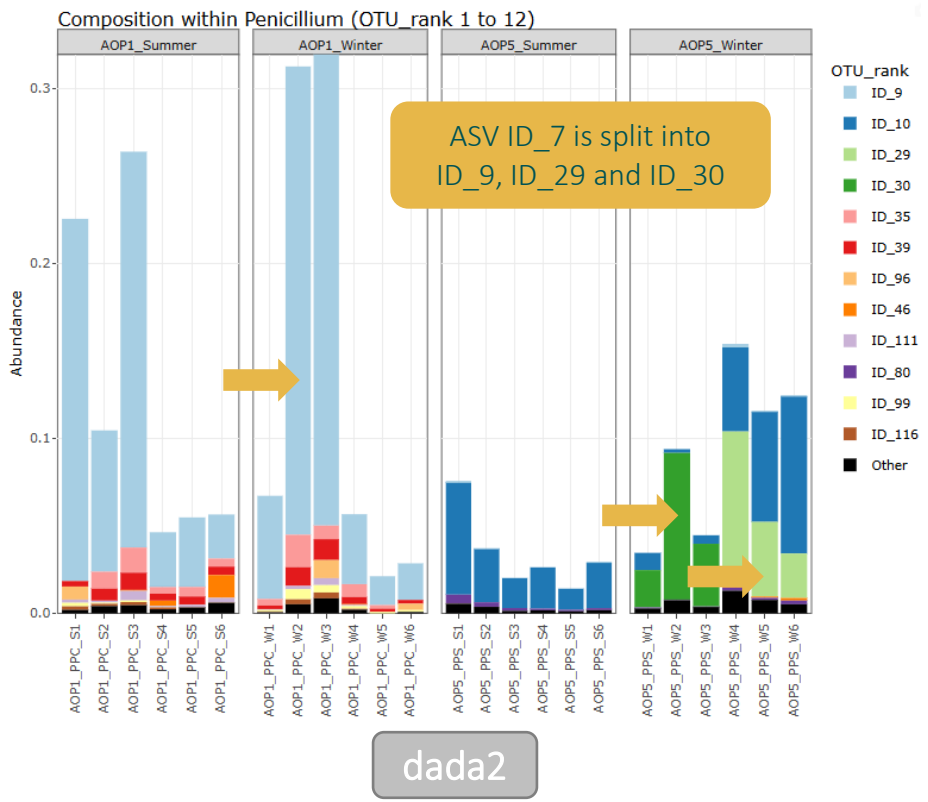
# Practice session

## Exercise: Visualisation of details of composition

What is the composition inside the Genus with the highest count?




swarm




dada2

dada2 can produce more ASV than swarm inside a species (and these extra ASV could have biological consistency).

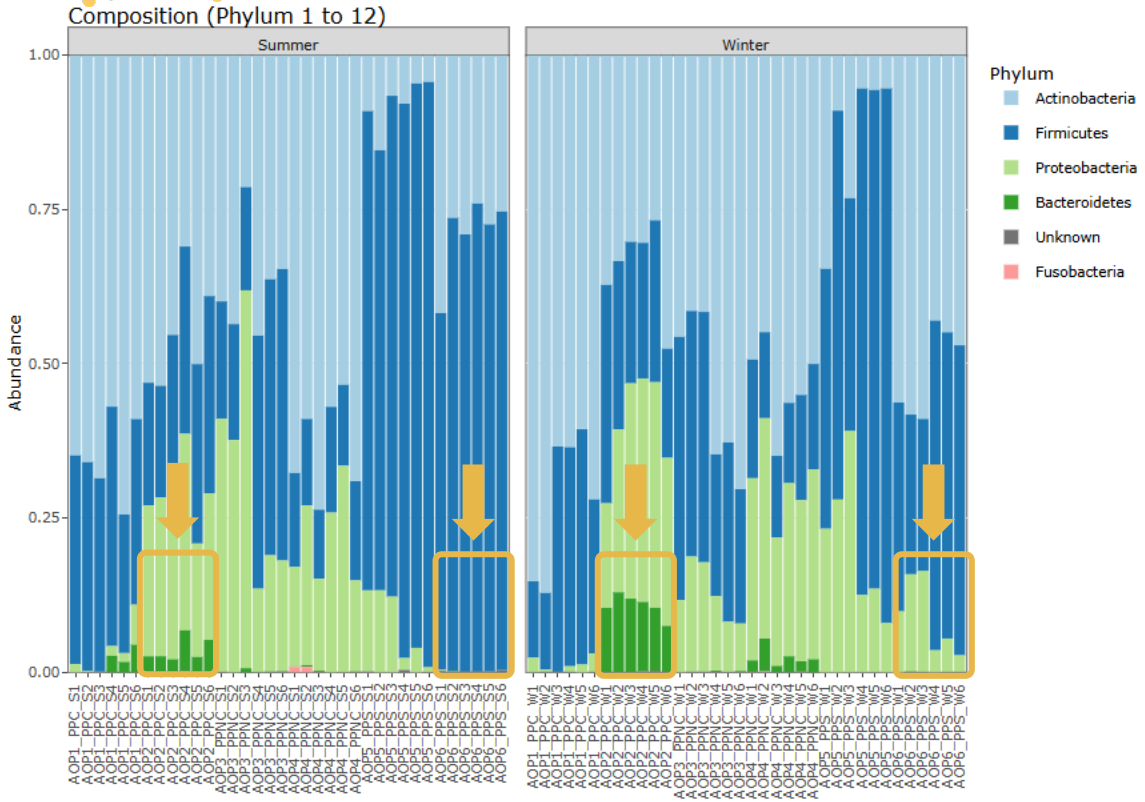
# Practice session

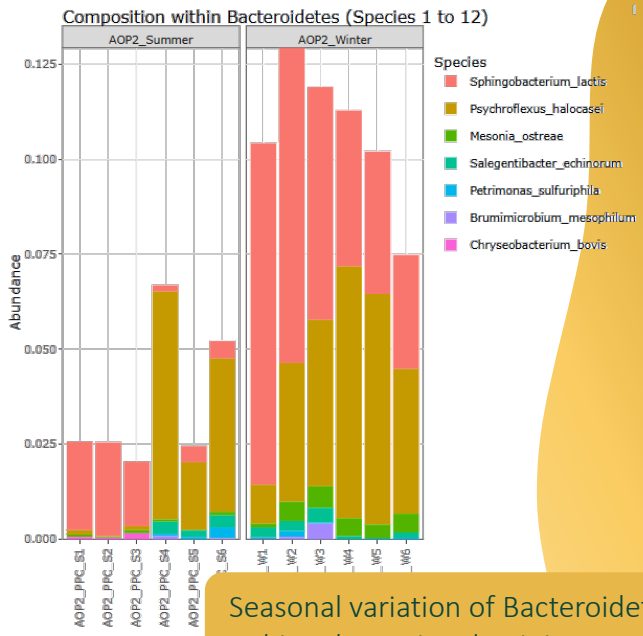
## Exercise: Visualisation of details of composition

Are there seasonal effects?

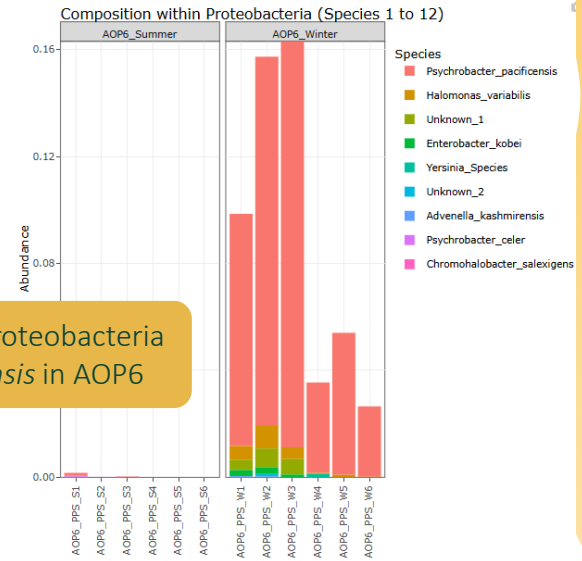
No strong seasonal variation visible at large scale, but some minor changes

⚠ Cognitive bias : easier to detect when at the barplot base




Seasonal variation of Bacteroidetes *Sphingobacterium lactis* in AOP2



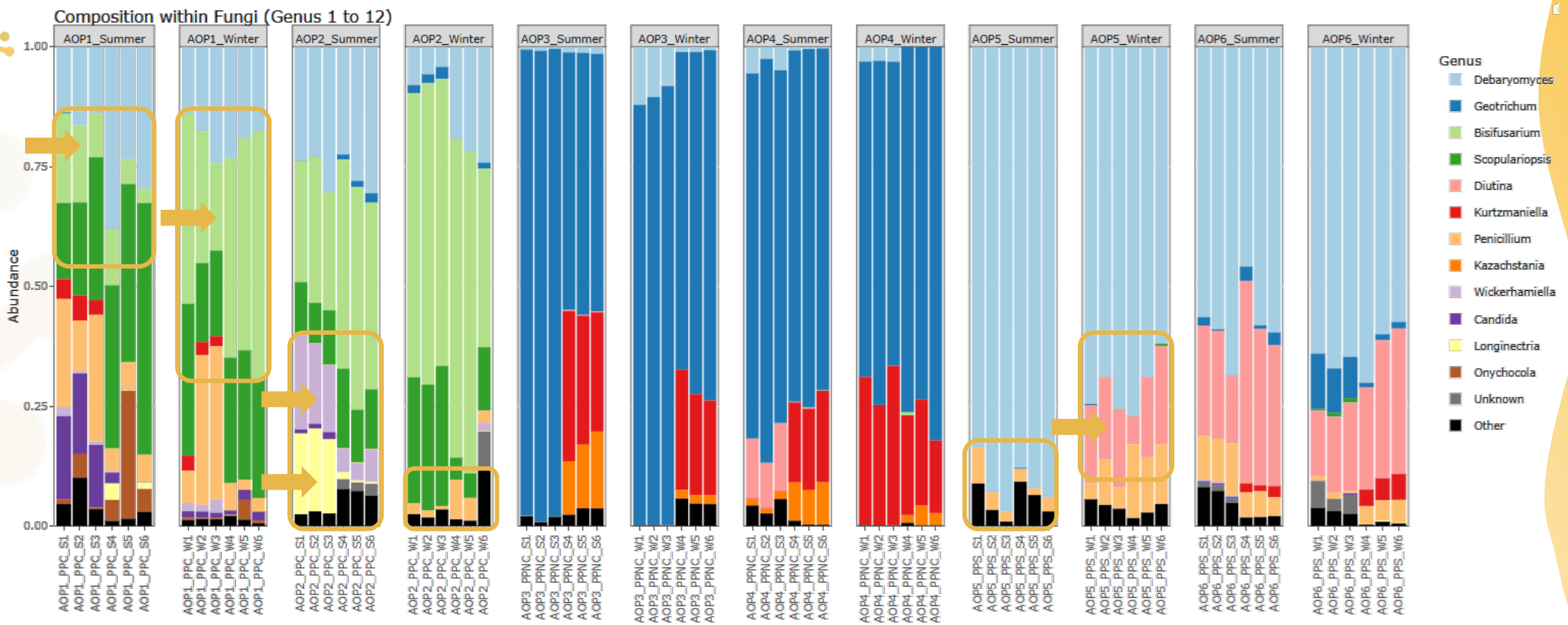

Seasonal variation of Proteobacteria *Psychorbacter pacificensis* in AOP6

# Practice session

## Exercise: Visualisation of details of composition

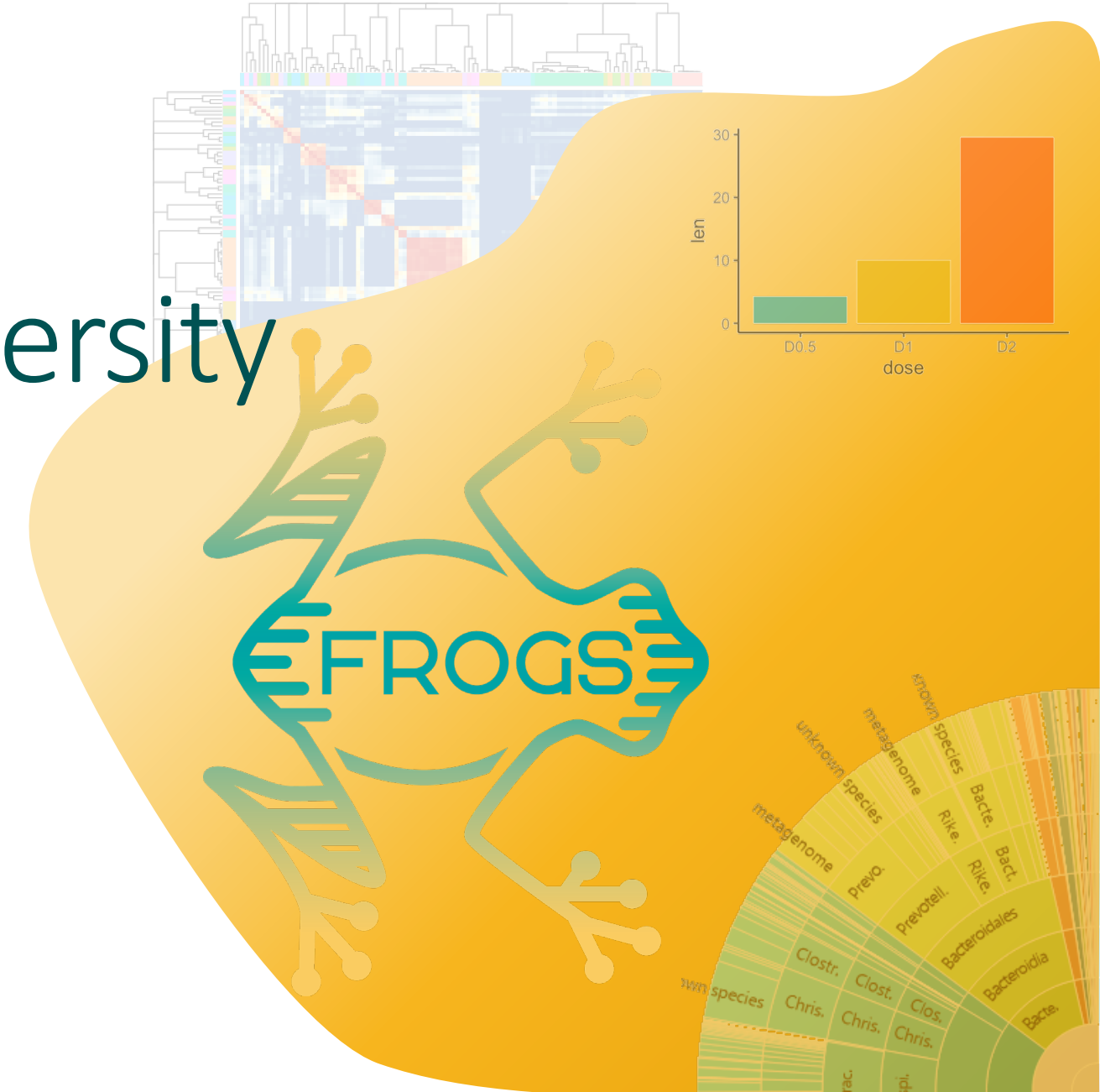
Are there seasonal effects?

No seasonal variation visible at Phylum, Class and Order levels, but observable starting at Family and Genus levels



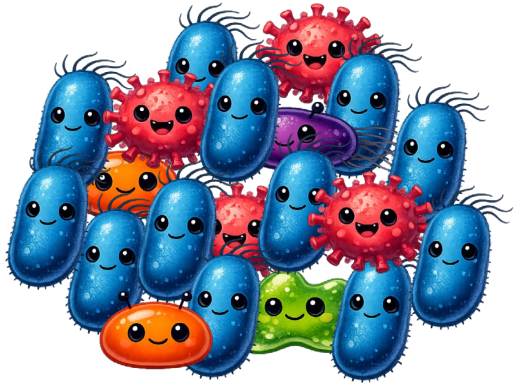
# FROGS Stats

# Exploring biodiversity



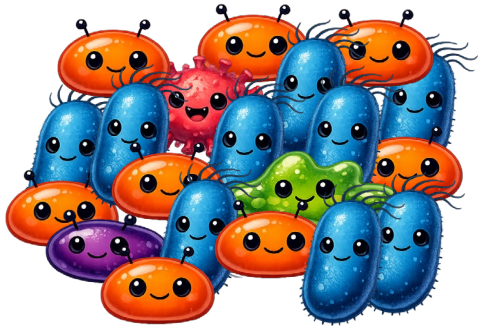
# What is microbial diversity?

Sample / Community 1



Sample / Community 2

Sample / Community 3



Which microbes are present? How are they distributed?

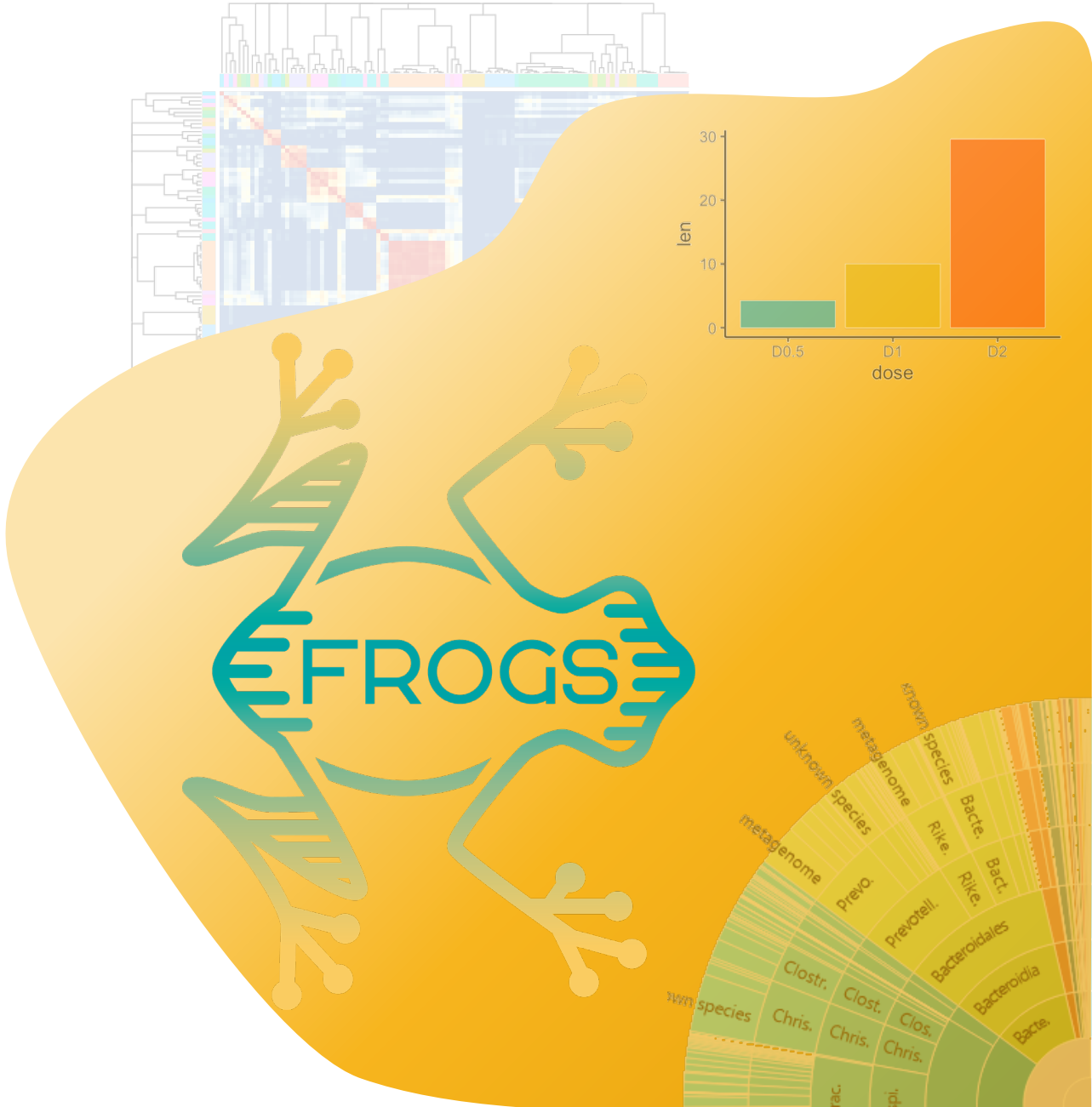
α-diversity captures and summarise what happens ***within a single sample.***

How communities differ from one sample to another?

β-diversity quantifies the similarity or difference ***between two samples.***

# FROGS Stats

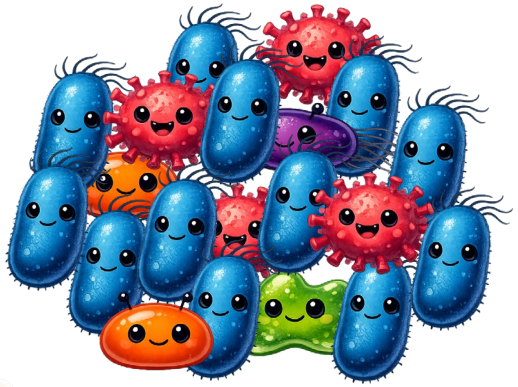
# α-diversity



# How to describe and measure biodiversity?

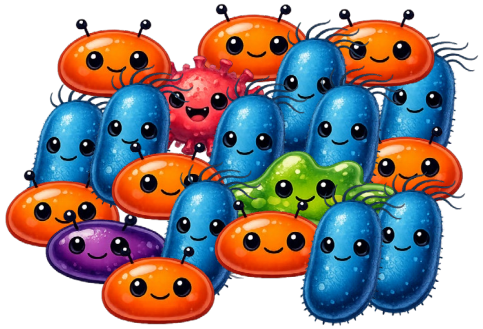
Which community is the more diverse?

Sample / Community 1



Sample / Community 2

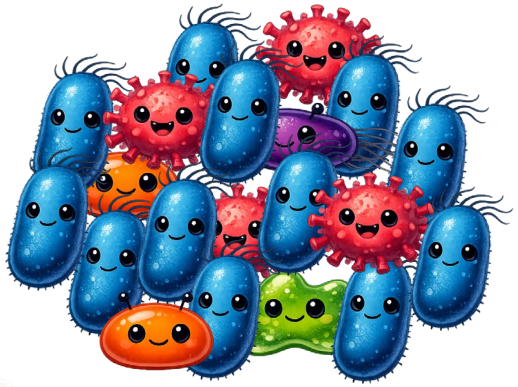
Sample / Community 3



# How to describe and measure biodiversity?

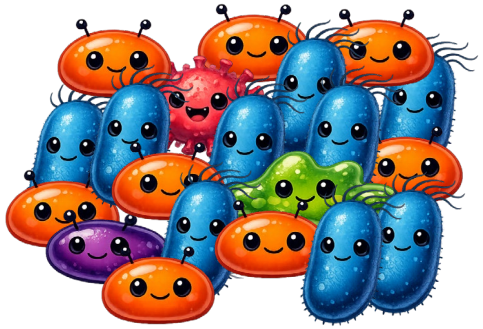
Which community is the more diverse?

Sample / Community 1



Sample / Community 2

Sample / Community 3



How many species in each sample?

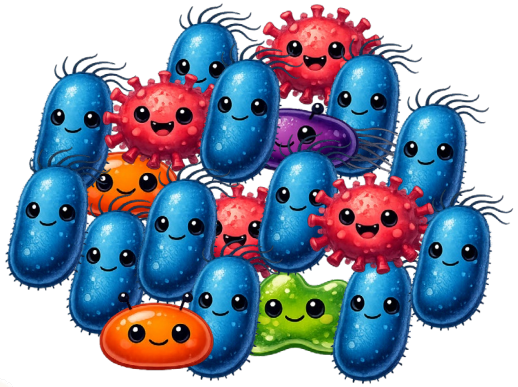
This number of observed species is called “**Richness**”

It can be measured at the ASV level or any other taxonomic level.

# How to describe and measure biodiversity?

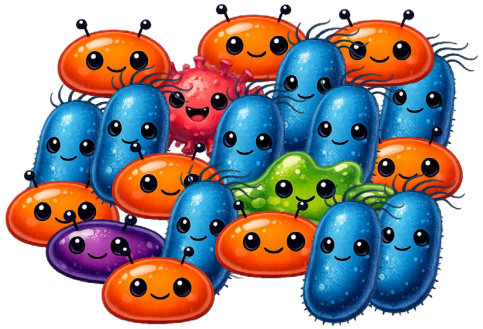
Which community is the more diverse?

Sample / Community 1



Sample / Community 2

Sample / Community 3



Same richness
Different distributions

Intuition that diversity of Sample 1 < Sample 3 < Sample 2

More balanced distributions involve more diverse communities

# How to describe and measure biodiversity?

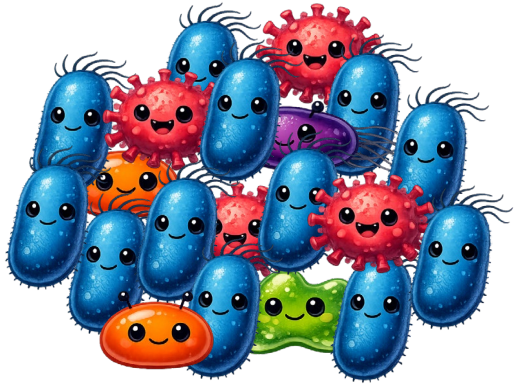
α-diversity indices are quantitative measures that describes the diversity **within a single sample**.
It summarises the number of present taxa and how they are distributed.

α-diversity indices can be calculated in many different ways, which can be grouped according to how much weight they give to three key components:

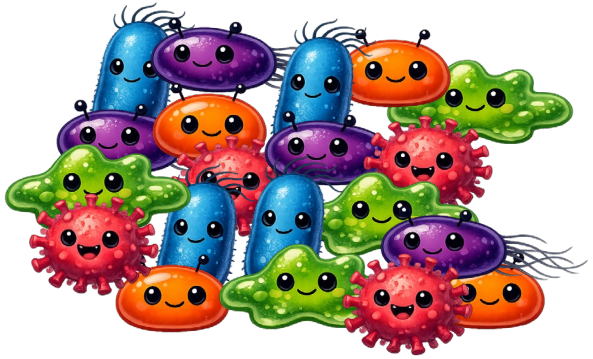
- Richness
- Abundance distribution : Shannon example
- Evenness : Simpson / invSimpson example

# Richness-based indices

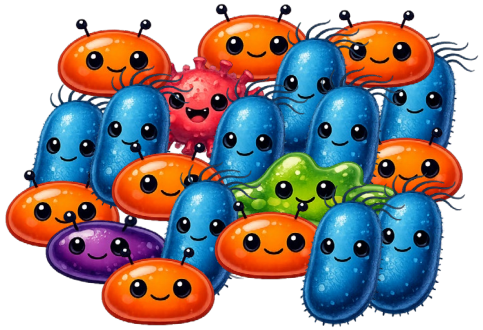
Sample / Community 1



Sample / Community 2

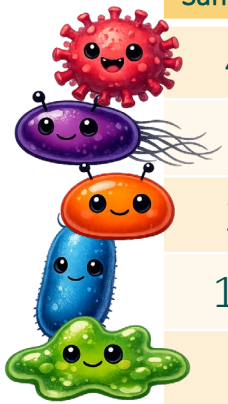


Sample / Community 3



Observed richness (*S*) is the simplest richness-based index: it just counts the number of observed species.

But could some species have been missed?
How many would we observed if we sampled more than 20 individuals?



| | Sample1 | Sample2 | Sample3 |
|---|---|---|---|
| | 4 | 4 | 1 |
| | 1 | 4 | 1 |
| | 2 | 4 | 8 |
| | 12 | 4 | 9 |
| | 1 | 4 | 1 |
| Richness | 5 | 5 | 5 |

# Richness-based indices : Chao1

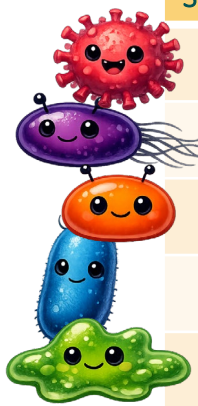
Chao1 was designed to estimate how many taxa are **missing** from your sample because rare taxa are easily overlooked in sequencing data.

The intuition is simple:
If you see many singletons or doubletons (taxa observed only once or twice), you probably missed species (you see them by chance but others stayed unseen)
If you see few or no singletons, your sampling was deep enough to capture most species.

$$\text{Chao1} = S + \frac{\text{singletons}^2}{2*\text{doubletons}}$$

In case there is no doubletons (to avoid division by 0)

$$\text{Chao1} = S + \frac{\text{singletons}(\text{singletons} - 1)}{2}$$



| | Sample1 | Sample2 | Sample3 |
|---|---|---|---|
| | 4 | 4 | 1 |
| | 1 | 4 | 1 |
| | 2 | 4 | 8 |
| | 12 | 4 | 9 |
| | 1 | 4 | 1 |
| Richness | 5 | 5 | 5 |
| Chao1 | 6 | 5 | 8 |

# Richness-based indices : Chao1

Weren’t all singletons removed??

| | AOP1_PPC_S1 | AOP1_PPC_S2 | AOP1_PPC_S3 | AOP1_PPC_S4 | AOP1_PPC_S5 | AOP1_PPC_S6 | AOP1_ |
|---|---|---|---|---|---|---|---|
| | All | All | All | All | All | All | All |
| ID_208 | 1 | 8 | 8 | 37 | 0 | 0 | |
| ID_138 | 0 | 7 | 6 | 7 | 0 | 0 | |
| ID_257 | 0 | 0 | 0 | 0 | 16 | 1 | |
| ID_118 | 2 | 18 | 7 | 2 | 0 | 0 | |
| ID_251 | 0 | 0 | 0 | 0 | 5 | 15 | |
| ID_171 | 0 | 11 | 7 | 20 | 7 | 5 | |
| ID_157 | 0 | 0 | 0 | 0 | 0 | 0 | |
| ID_62 | 11 | 28 | 12 | 58 | 15 | 7 | |
| ID_75 | 10 | 35 | 13 | 47 | 18 | 30 | |
| ID_104 | 8 | 13 | 8 | 15 | 5 | 3 | |
| ID_44 | 55 | 47 | 26 | 70 | 12 | 17 | |
| ID_3 | 3480 | 2871 | 1508 | 1952 | 896 | 479 | |
| ID_229 | 5 | 8 | 13 | 5 | 6 | 0 | |
| ID_143 | 15 | 11 | 39 | 7 | 5 | 2 | |
| ID_113 | 33 | 40 | 69 | 12 | 8 | 0 | |
| ID_169 | 2 | 12 | 0 | 17 | 9 | 15 | |

Singletons and rare ASVs were filtered at **the whole dataset scale**
“Sample-singletons” exist when the ASV filled the global abundance (0.005%) and occurrence filters.

Chao1 estimation is thus possible even on filtered data.

# Shannon index

To introduce the abundance, Shannon index (*H'*) expresses how unpredictable the identity of a randomly chosen individual is (physics notion of "entropy"). A community is considered more diverse when :

- It contains more taxa (higher richness)
- Their abundances are more evenly distributed

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

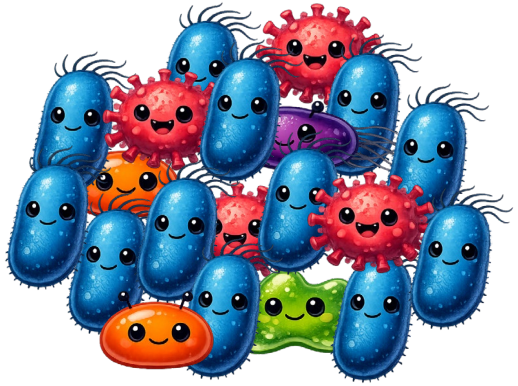
The index is computed by:

- Taking the relative abundance of each taxon ($p$)
- Evaluating how much each taxon contributes to the overall uncertainty ( $-p(\ln(p)$ )
- Summing these contributions across all taxa

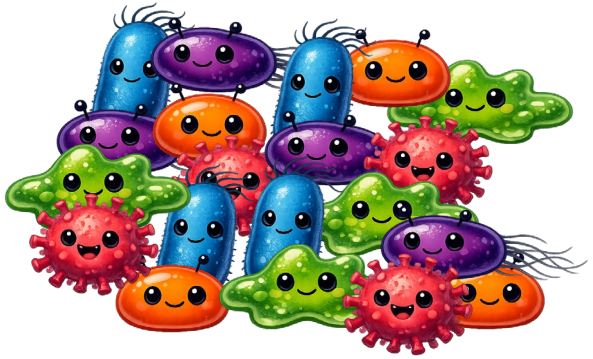
Any log function (*ln*, $log_2$, $log_{10}$…) can be used in the definition of Shannon index…
$log_2$ and *ln* and are the most common, but several tools / R packages use different log functions…
and thus provide different Shannon values, and this information is rarely (never?) indicated in publications.

# Shannon index

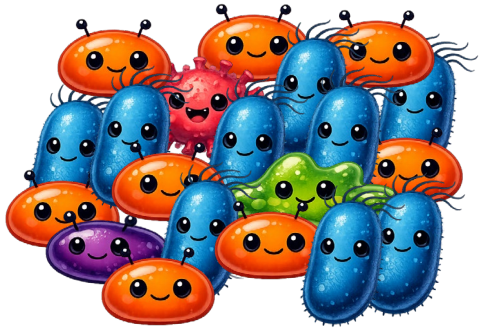
Sample / Community 1



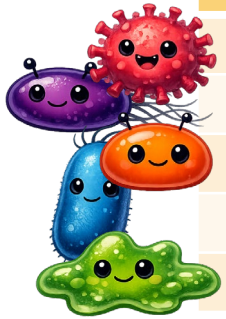
Sample / Community 2



Sample / Community 3



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



$p_i$(Relative abundances)

| Sample1 | Sample2 | Sample3 |
|---|---|---|
| 0.2 | 0.2 | 0.05 |
| 0.05 | 0.2 | 0.05 |
| 0.1 | 0.2 | 0.4 |
| 0.6 | 0.2 | 0.45 |
| 0.05 | 0.2 | 0.05 |

$-\ln(p_i)$

| Sample1 | Sample2 | Sample3 |
|---|---|---|
| 1.61 | 1.61 | 3.00 |
| 3.00 | 1.61 | 3.00 |
| 2.30 | 1.61 | 0.92 |
| 0.51 | 1.61 | 0.80 |
| 3.00 | 1.61 | 3.00 |

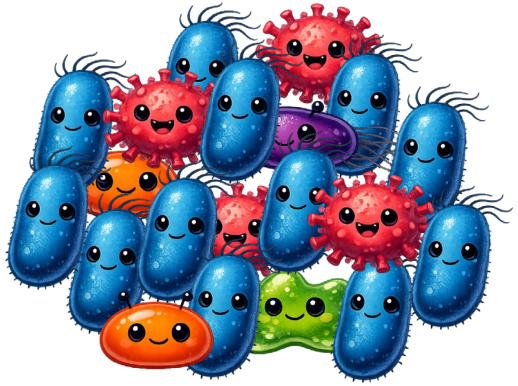
$-p_i \ln(p_i)$

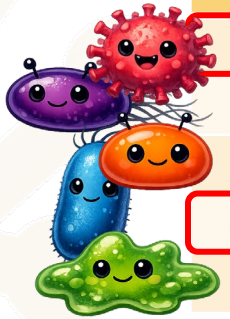
| Sample1 | Sample2 | Sample3 |
|---|---|---|
| 0.32 | 0.32 | 0.15 |
| 0.15 | 0.32 | 0.15 |
| 0.23 | 0.32 | 0.37 |
| 0.31 | 0.32 | 0.36 |
| 0.15 | 0.32 | 0.15 |

| Shannon | 1.16 | 1.61 | 1.18 |
|---|---|---|---|

# Shannon index

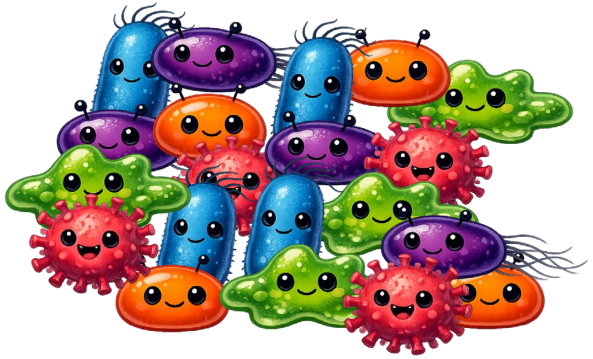
## Sample / Community 1





| %abund | Hcontrib |
|---|---|
| 20% | 0.32 |
| 5% | 0.15 |
| 10% | 0.23 |
| 60% | 0.31 |
| 5% | 0.15 |
| | 1.16 |

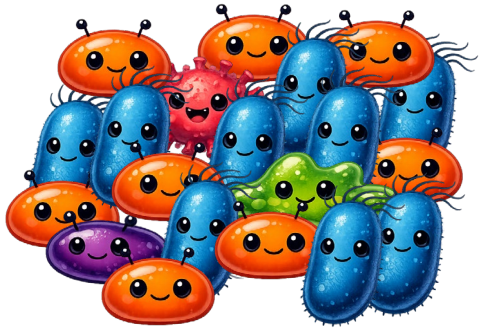
## Sample / Community 2



| %abund | Hcontrib |
|---|---|
| 20% | 0.32 |
| 20% | 0.32 |
| 20% | 0.32 |
| 20% | 0.32 |
| 20% | 0.32 |
| | 1.61 |

$= \ln(S)$

## Sample / Community 3

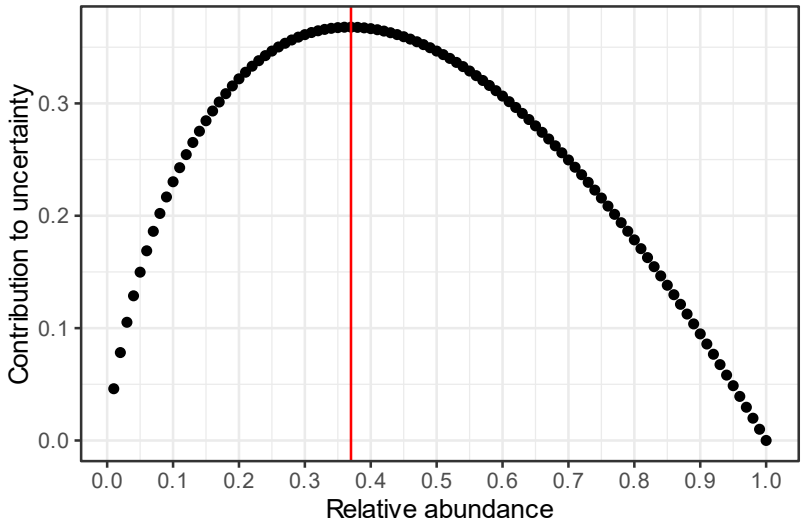


| %abund | Hcontrib |
|---|---|
| 5% | 0.15 |
| 5% | 0.15 |
| 45% | 0.37 |
| 40% | 0.36 |
| 5% | 0.15 |
| | 1.18 |

H’ has a maximum (linked to richness) when all abundances are equal

# Shannon index

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




$-p(\ln(p)$ has a maximum for $p$=0.3679

If a species is too rare, its contribution is small because $p$ is small.

If a species is too dominant, its contribution is small because $\ln(p)$ approaches 0.

The maximum contributions happens at an intermediate abundance : 37% abundance.

Shannon balances richness and evenness. It's maximum value is related to the richness ($\ln(S)$)
Rare or very dominant species have a weak effect on the index.
Same or close values can be due to different distributions **AND/OR** richness.

# Simpson / invSimpson

Simpson diversity ($D$) captures how dominated a community is by its most abundant taxa, asking "What is the probability that two randomly chosen individuals belong to the same taxon?"

$$D = \sum_{i=1}^{S} p_i^2$$

The index is computed by:

- Taking the relative abundance of each taxon ($p$)
- Calculate the probability of randomly drawing two times in each taxon ($p^2$)
- Summing these probabilities across all taxa

# Simpson / invSimpson

With this formula,

$$D = \sum_{i=1}^{S} p_i^2$$

higher D means lower diversity.

This counter-intuitive direction is why raw Simpson is rarely used and transformed versions are preferred.

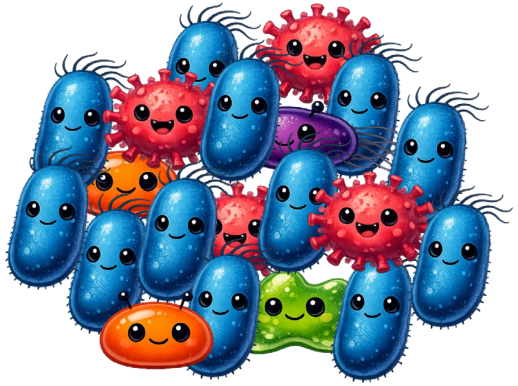
$$invSimpson = \frac{1}{D}$$

Interpretation:

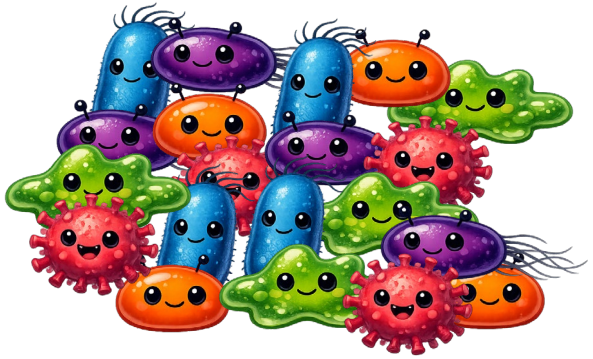
- If one species dominates completely : D = 1 and 1/D = 1 (minimal value)
- 1/D increases with diversity (as would Shannon, richness and others)

# Simpson / invSimpson index

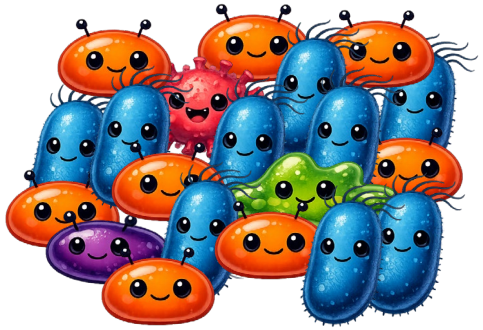
Sample / Community 1



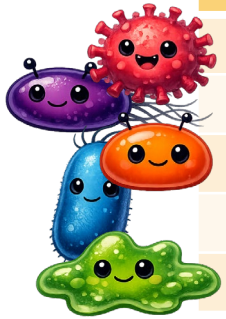
Sample / Community 2



Sample / Community 3



$$D = \sum_{i=1}^{S} p_i^2$$



$p_i$ (Relative abundances)

| Sample1 | Sample2 | Sample3 |
|---|---|---|
| 0.2 | 0.2 | 0.05 |
| 0.05 | 0.2 | 0.05 |
| 0.1 | 0.2 | 0.4 |
| 0.6 | 0.2 | 0.45 |
| 0.05 | 0.2 | 0.05 |

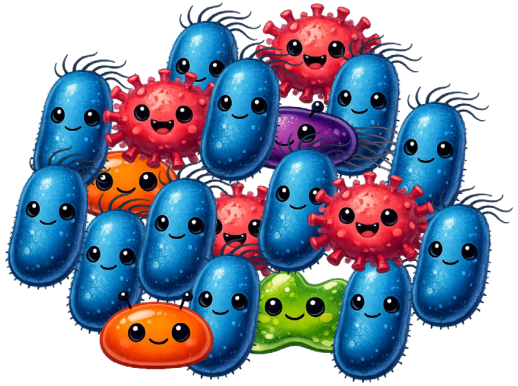
$p_i^2$

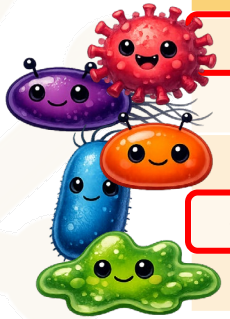
| Sample1 | Sample2 | Sample3 |
|---|---|---|
| 0.04 | 0.04 | 0.003 |
| 0.003 | 0.04 | 0.003 |
| 0.01 | 0.04 | 0.160 |
| 0.36 | 0.04 | 0.202 |
| 0.003 | 0.04 | 0.003 |

| | | | |
|---|---|---|---|
| Simpson | 0.415 | 0.02 | 0.37 |
| invSimpson | 2.41 | 5 | 2.70 |

# Simspon / invSimpson index

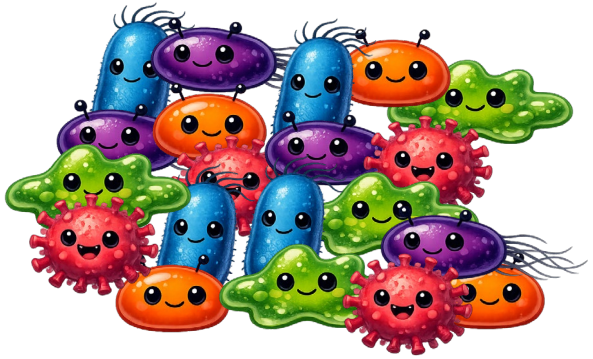
## Sample / Community 1





| %abund | Dcontrib |
|---|---|
| 20% | 0.04 |
| 5% | 0.003 |
| 10% | 0.01 |
| 60% | 0.36 |
| 5% | 0.003 |
| | 0.415 |
| | 2.41 |

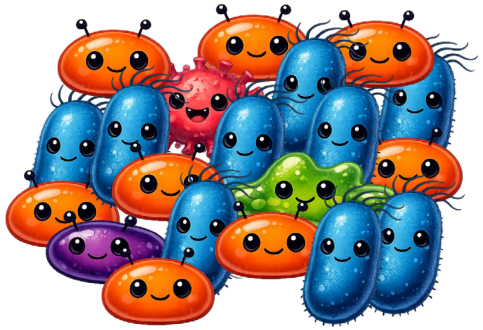
## Sample / Community 2



| %abund | Dcontrib |
|---|---|
| 20% | 0.04 |
| 20% | 0.04 |
| 20% | 0.04 |
| 20% | 0.04 |
| 20% | 0.04 |
| | 0.02 |
| | 5 |

$= S$

## Sample / Community 3

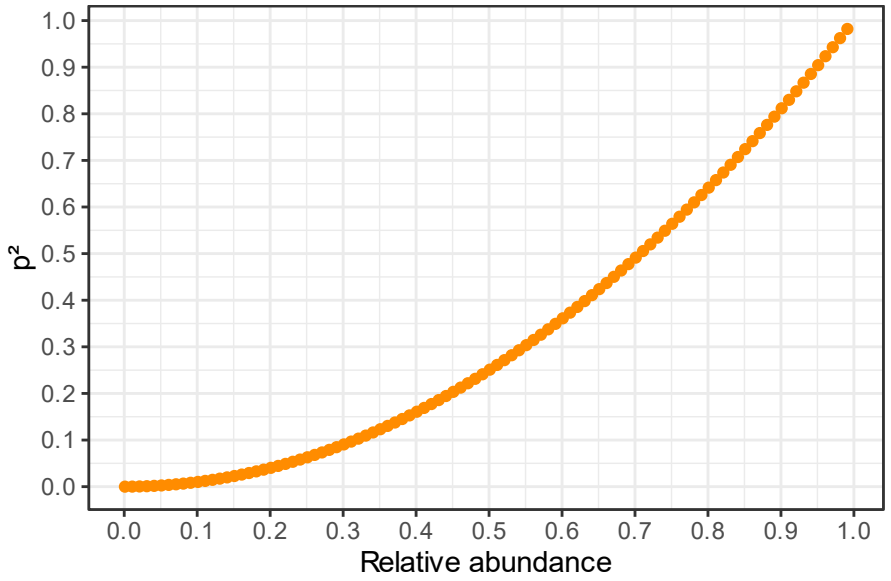


| %abund | Dcontrib |
|---|---|
| 5% | 0.003 |
| 5% | 0.003 |
| 45% | 0.160 |
| 40% | 0.202 |
| 5% | 0.003 |
| | 0.37 |
| | 2.70 |

1/D has a maximum (equal to richness) when all abundances are equal

# Simpson / invSimpson index

$$D = \sum_{i=1}^{S} p_i^2$$




$p^2$ is strictly increasing with relative abundance.

Squaring $p_i$ amplifies the contribution of dominant species and suppresses the influence of rare ones.

Simpson index captures dominance, and is driven almost entirely by the most abundant taxa.
invSimpson is more intuitive as it increases with diversity.

# α-diversity indices

## Sample / Community 1

| count | %abund |
|---|---|
| 4 | 20% |
| 1 | 5% |
| 2 | 10% |
| 12 | 60% |
| 1 | 5% |

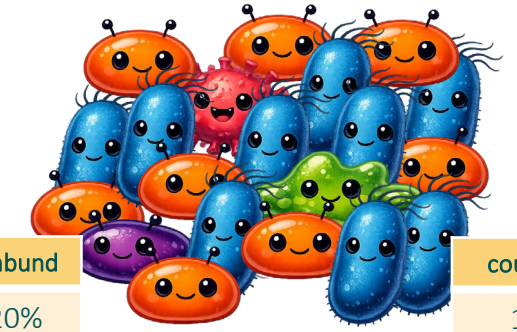
| | |
|---|---|
| Observed richness | 5 |
| Chao1 | 6 |
| Shannon | 1.16 |
| invSimpson | 2.41 |

## Sample / Community 2

| count | %abund |
|---|---|
| 4 | 20% |
| 4 | 20% |
| 4 | 20% |
| 4 | 20% |
| 4 | 20% |

| | |
|---|---|
| Observed richness | 5 |
| Chao1 | 5 |
| Shannon | 1.61 |
| invSimpson | 5 |

## Sample / Community 3



| count | %abund |
|---|---|
| 1 | 5% |
| 1 | 5% |
| 9 | 45% |
| 8 | 40% |
| | 5% |

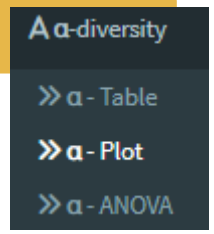
| | |
|---|---|
| Observed richness | 5 |
| Chao1 | 8 |
| Shannon | 1.18 |
| invSimpson | 2.70 |

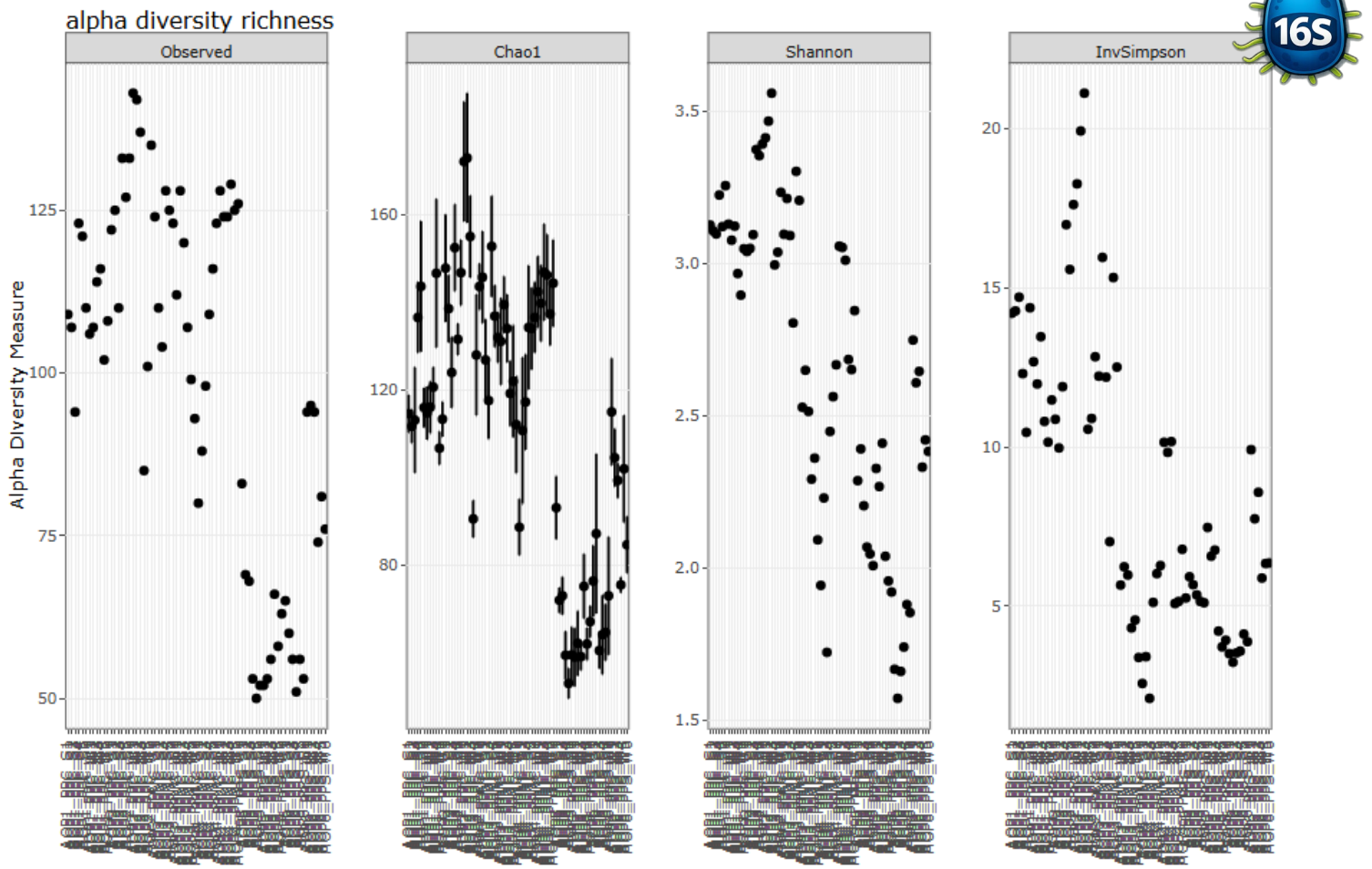
There are of course many others α-diversity indices, all equally valid, and their use often depend on disciplinary habits and the need to remain comparable with previous studies.

# Practice session

## Visualise α-diversity

In Easy16S select




Choose a variable to regroup points

Choose a variable to color points


Draw boxplots with the X variable

# Practice session

## Exercise: α-diversity analyses

How to read the output?

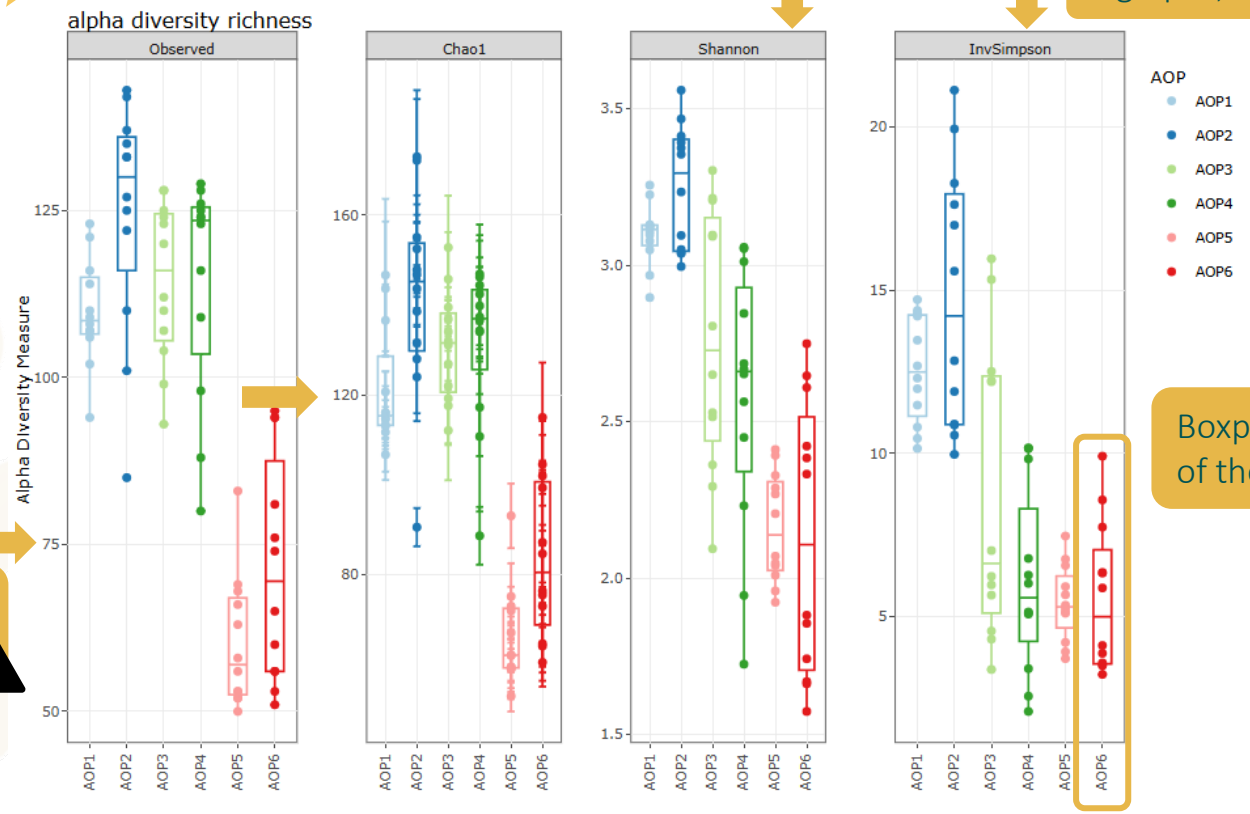
What biological interpretation can be extracted?

# Practice session

## Exercise: α-diversity analyses

How to read the output?

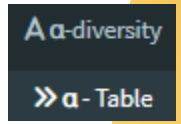
4 graphs, 1 per index



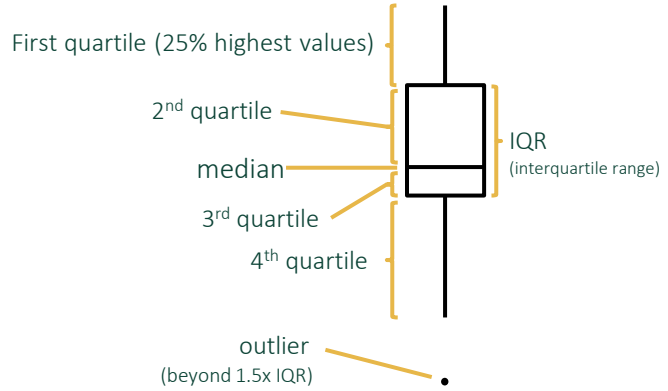

Scales adapt **per index**

Boxplots per modality of the selected variable


Calculated values are available through



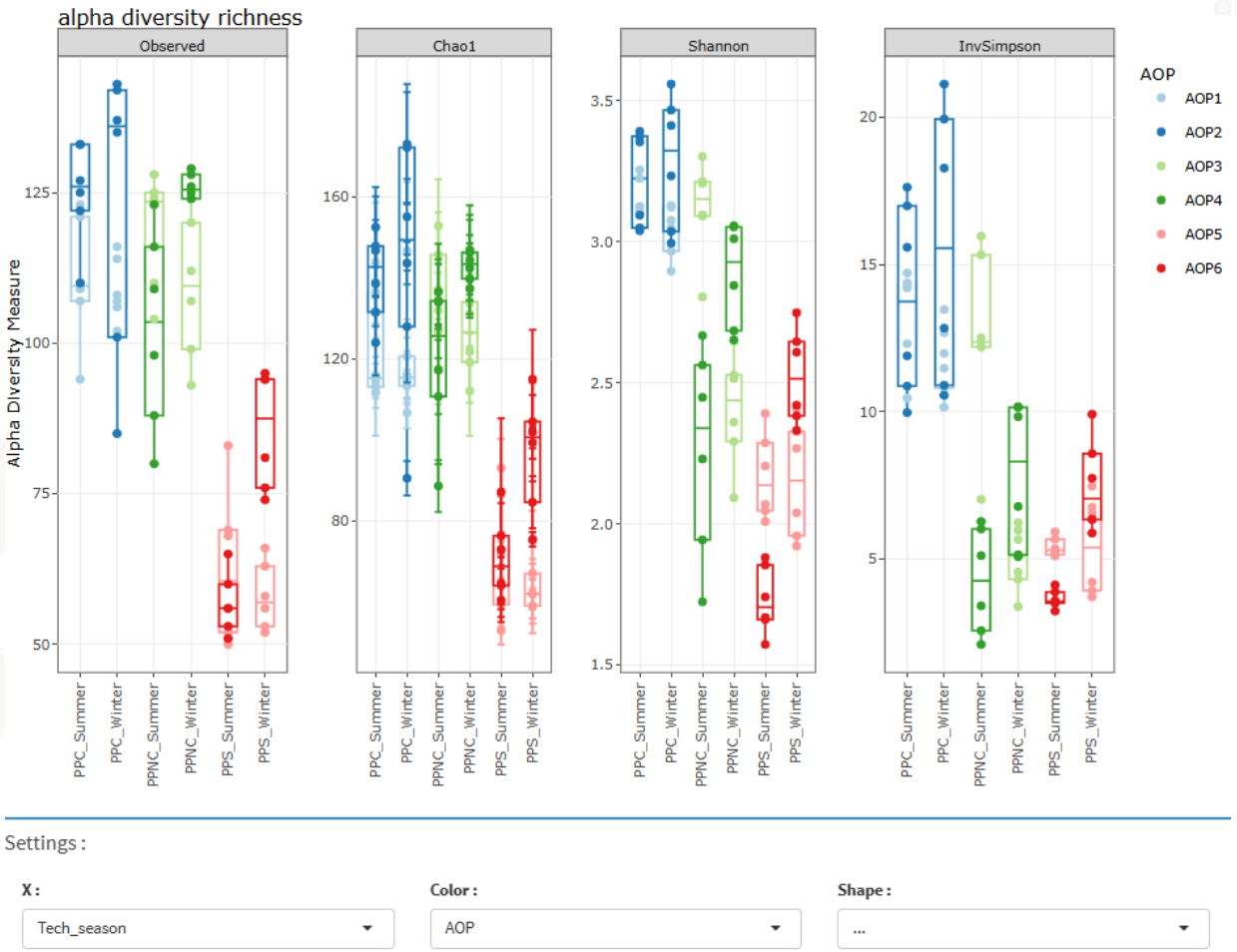

How to read a boxplot:

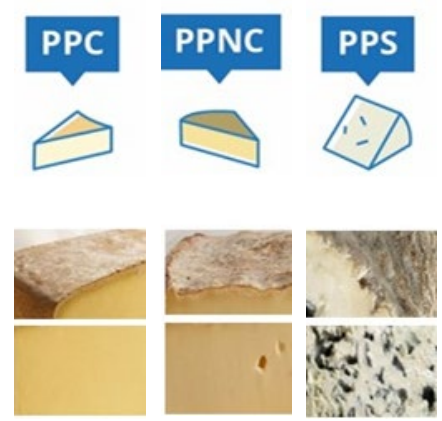


# Practice session

## Exercise: α-diversity analyses

What biological interpretation can be extracted?




Chao1 > richness so there are some rare species undetected at this sequencing depth

Lower richness for PPS

For PPC, differences between AOP for richness/Chao1 but not for Shannon/invSimpson

For PPNC, no differences between AOP for richness/Chao1 but differences for Shannon/Simpson

Seasonal variations for several AOP

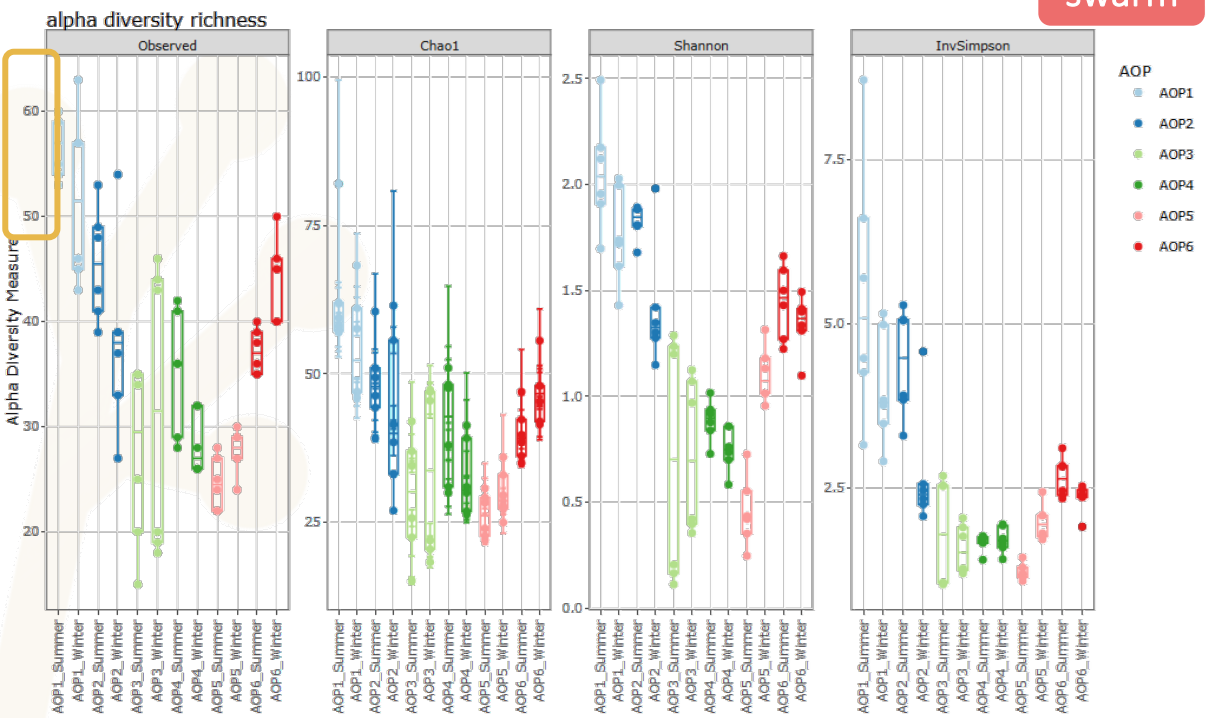
# Practice session

## Exercise: α-diversity analyses

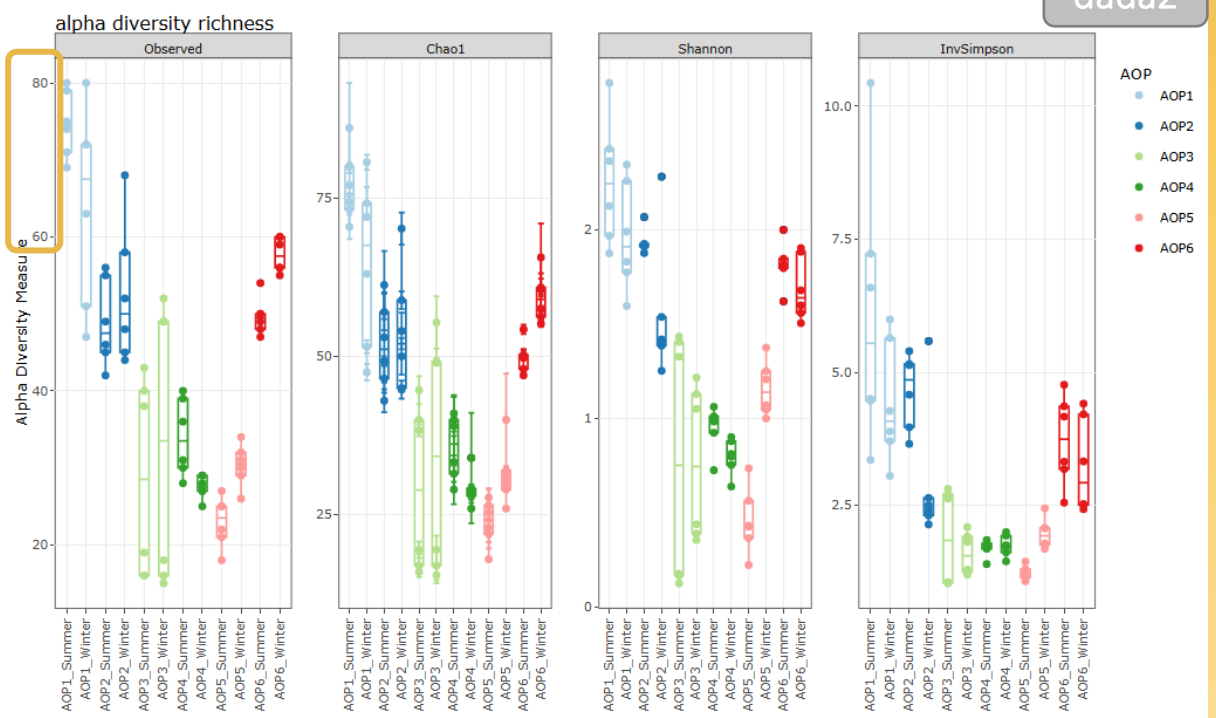
What biological interpretation can be extracted?

Different values with swarm and dada2 but same patterns / variations

swarm




dada2



# Testing α-diversity differences between groups

What does an ANOVA test?

ANOVA evaluates whether group means differ significantly
by **testing if there is a difference of means between at least two groups**.

ANOVA test the "null hypothesis" $H_0$ that all group means are equal.

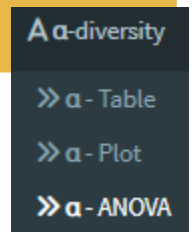
ANOVA has several assumptions that need to be filled:

- Observations must be independent (measures of samples are not linked)
- Normal ("bell-shaped") distribution of the values around the mean value
- Similar dispersion around the mean value between samples

# Practice session

## Test α-diversity differences

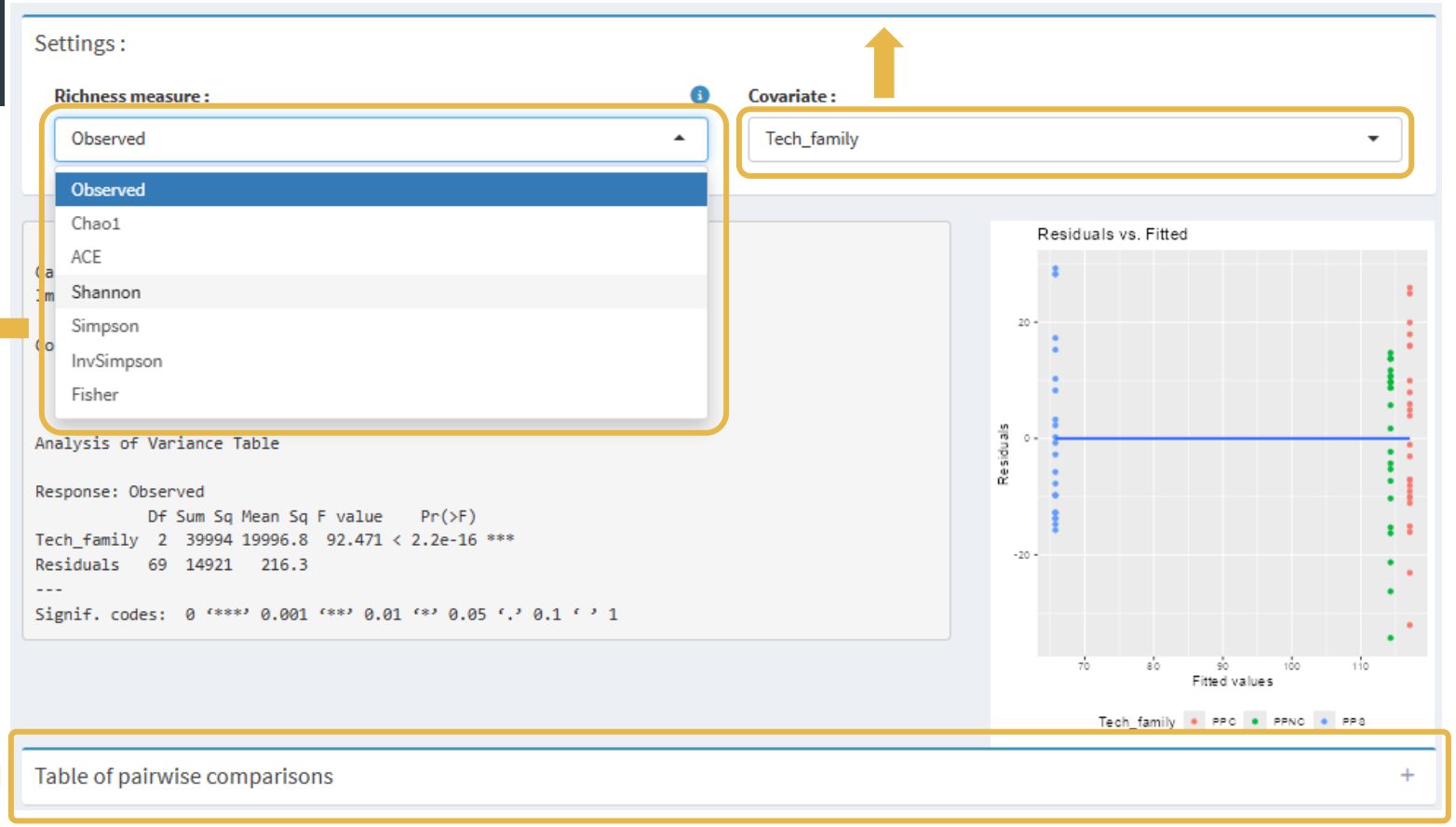
In Easy16S select




Choose the variable to test

Choose the index for the test

Visualise pairwise comparisons



# Practice session

## Exercise: test α-diversity differences


How to read the output?

Are there seasonal effects?

# Practice session

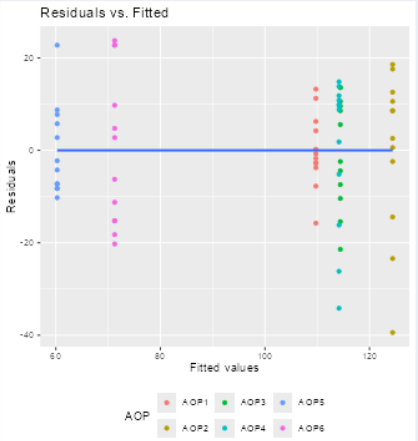
## Exercise: test α-diversity differences

How to read the output?

Richness measure : Observed

Covariate : AOP

```
Call:
lm(formula = Observed ~ AOP, data = anova_data)

Coefficients:
(Intercept)      AOPAOP2      AOPAOP3      AOPAOP4      AOPAOP5      AOPAOP6
    109.750       14.667        4.667        4.417      -49.500      -38.500

Analysis of Variance Table

Response: Observed
           Df Sum Sq Mean Sq F value    Pr(>F)
AOP         5  42011  8402.1  42.974 < 2.2e-16 ***
Residuals 66  12904   195.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```




Pr: Probability of observing this distribution by chance if AOP has no effect on observed richness

Usually we consider the tested variable to have a significant effect if **P<0.05**

P<0.05 means
at least one modality has a different mean richness

Table of pairwise comparisons

| contrast | estimate | conf.low | conf.high | adj.p.value |
|---|---|---|---|---|
| AOP5-AOP3 | | | | 0 |
| AOP6-AOP3 | | | | 0 |
| AOP5-AOP4 | -53.92 | -70.67 | -37.16 | 0 |
| AOP6-AOP4 | -42.92 | -59.67 | -26.16 | 0 |
| AOP2-AOP1 | 14.67 | -2.09 | 31.42 | 0.12 |
| AOP6-AOP5 | 11 | -5.75 | 27.75 | 0.4 |
| AOP4-AOP2 | -10.25 | -27 | 6.5 | 0.48 |
| AOP3-AOP2 | -10 | -26.75 | 6.75 | 0.5 |

Pairwise comparisons test if there are significant differences between modalities (*only if ANOVA is significant!*)

# Practice session

## Exercise: test α-diversity differences

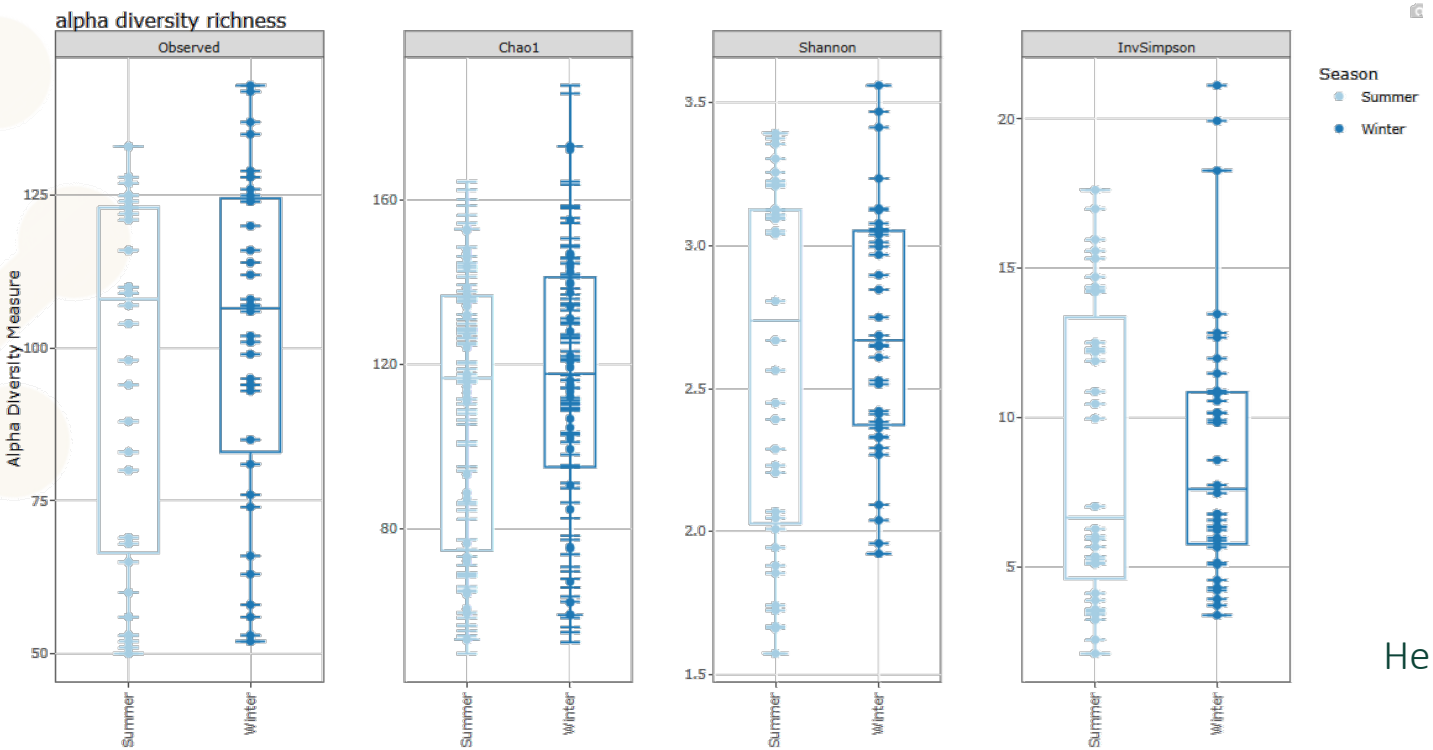
Are there seasonal effects?

**Richness measure :** Observed

**Covariate :** Season

```
Response: Observed
          Df Sum Sq Mean Sq F value Pr(>F)
Season     1    642  642.01  0.8281  0.366
Residuals 70  54273  775.33
```

No, and same for other metrics




Here we asked to compare the seasonal distributions across all AOP

# Practice session

## Exercise: test α-diversity differences

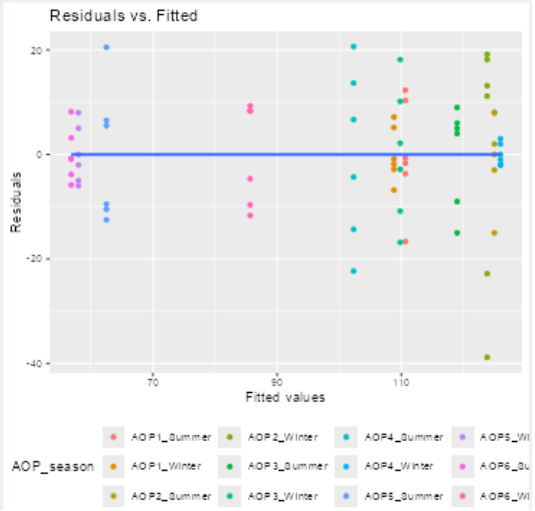
Are there seasonal effects?

If we want to consider AOP with ANOVA analysis, the variable should include it

**Richness measure :** Observed

**Covariate :** AOP_season

```
Call:
lm(formula = Observed ~ AOP_season, data = anova_data)

Coefficients:
          (Intercept)  AOP_seasonAOP1_Winter  AOP_seasonAOP2_Summer  AOP_seasonAOP2_Winter
              110.6667               -1.8333                14.3333                13.1667
AOP_seasonAOP3_Summer  AOP_seasonAOP3_Winter  AOP_seasonAOP4_Summer  AOP_seasonAOP4_Winter
                8.3333                -0.8333                -8.3333                15.3333
AOP_seasonAOP5_Summer  AOP_seasonAOP5_Winter  AOP_seasonAOP6_Summer  AOP_seasonAOP6_Winter
              -48.1667               -52.6667               -53.8333               -25.0000

Analysis of Variance Table

Response: Observed
            Df Sum Sq Mean Sq F value  Pr(>F)
AOP_season  11  46512  4228.4  30.192 < 2.2e-16 ***
Residuals   60   8403   140.0
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```




There are lot of non-relevant comparisons in the pairwise comparisons, but the intra-AOP are relevant

Table of pairwise comparisons

| contrast | estimate | conf.low | conf.high | adj.p.value |
|---|---|---|---|---|
| AOP6 | All | All | All | All |
| AOP6_Winter-AOP6_Summer | 28.83 | 5.6 | 52.06 | 0 |
| AOP6_Winter-AOP5_Winter | 27.67 | 4.44 | 50.9 | 0.01 |

# Practice session

## Exercise: test α-diversity differences

Are there seasonal effects?

Transformations are applied iteratively, starting from raw data.

Please refer to the documentation to learn more about the transformation modules.

**Transformation module_1** AOP

**To keep :** AOP3

Add

Cancel OK

Another option is to focus on a specific AOP

**Richness measure :** Simpson

**Covariate :** Season

```
Call:
lm(formula = Simpson ~ Season, data = anova_data)

Coefficients:
 (Intercept)  SeasonWinter
      0.9144       -0.1228

Analysis of Variance Table

Response: Simpson
          Df   Sum Sq  Mean Sq F value   Pr(>F)
Season     1 0.045274 0.045274  25.906 0.000471 ***
Residuals 10 0.017477 0.001748
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

But **it is not recommended except on a very limited number of conditions**,
as repeating separated tests without multiple-test correction is not statistically rigorous.

# Practice session

## Exercise: draw rarefaction curves


How to read the output?

How to draw rarefaction curves at the species or Genus level?

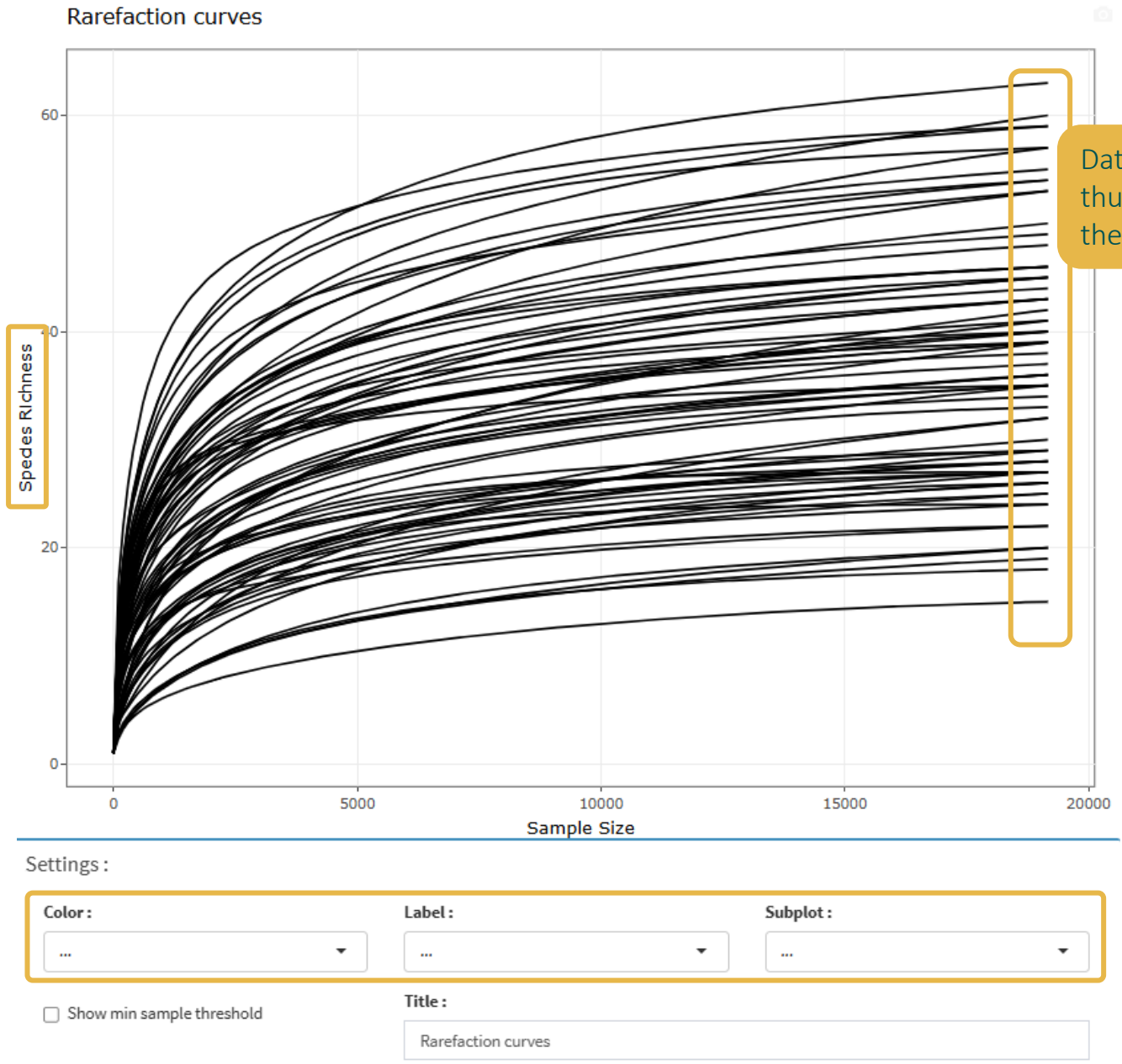
# Practice session

## Draw rarefaction curves

In Easy16S select Rarefaction

Detected richness

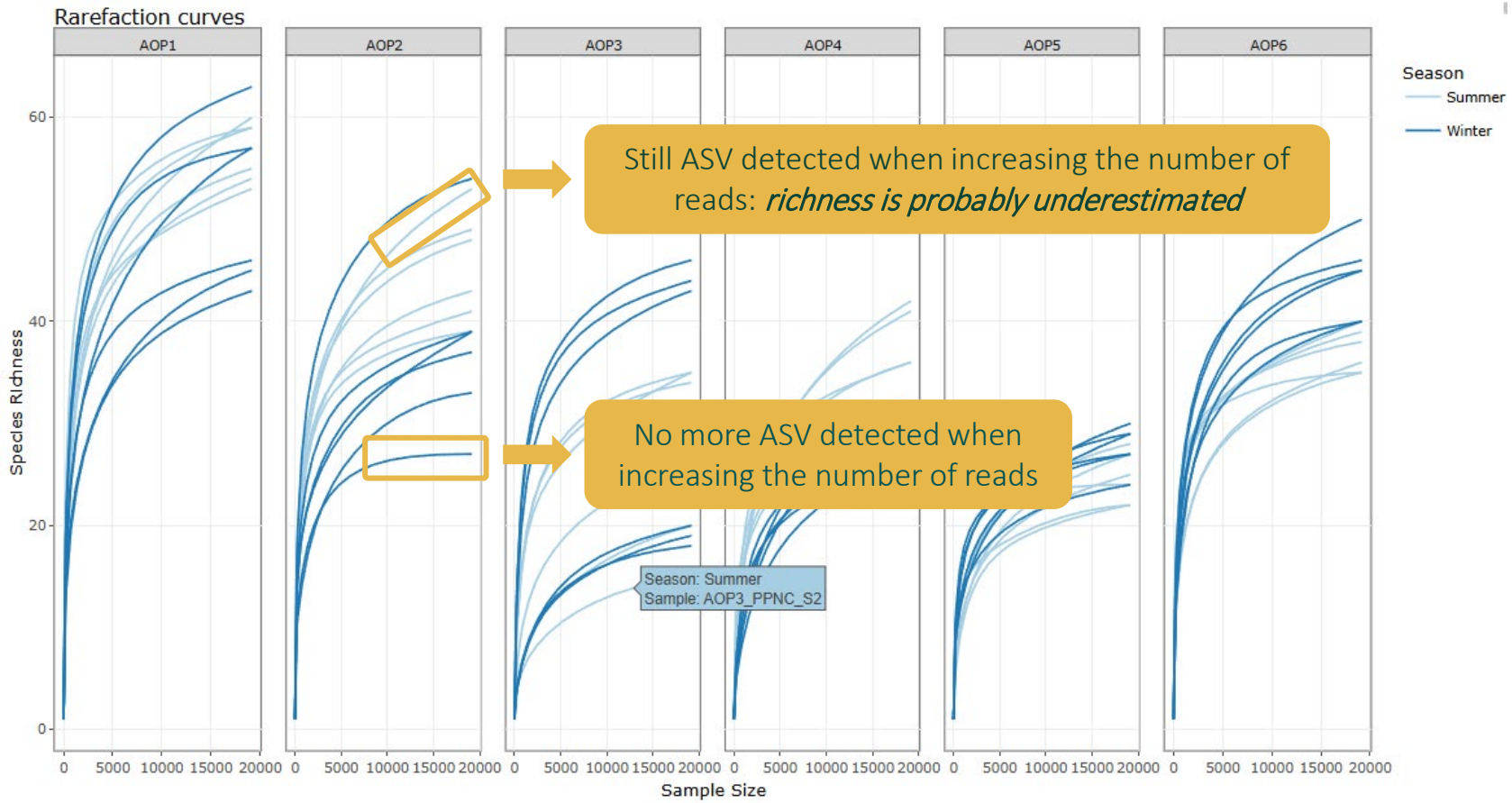
Draw separate plots and choose colors




Data were subsampled thus all samples have the same depth

## Draw rarefaction curves

How to read the output?




Rarefaction curves can easily mislead the eyes: depending on the axis scales, zoom level and plot dimensions, a *curve may or may not look flat and saturated* even though the data are identical

# Practice session

## Draw rarefaction curves

How to draw rarefaction curves at the species or Genus level?

Use “agglomerate taxa” in **Preprocess data**

Transformations are applied iteratively, starting from raw data.

Please refer to the documentation to learn more about the transformation modules.

**Transformation module_1**

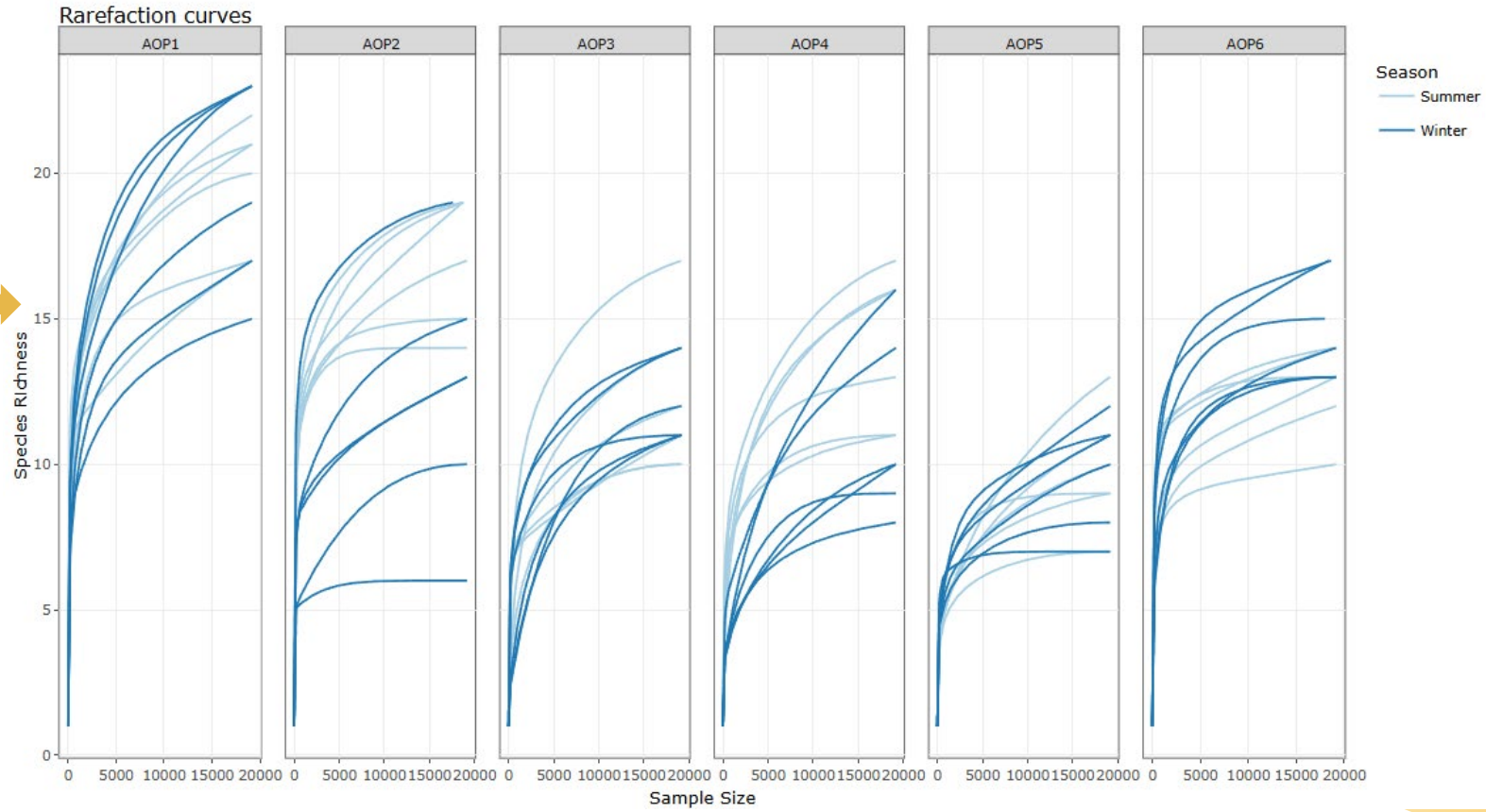
Agglomerate taxa

**Rank to agglomerate over :**

Kingdom Phylum Class
Order Family Genus
Species Rank7 Rank6
Rank1 Rank2 Rank3
Rank4 Rank5

Add

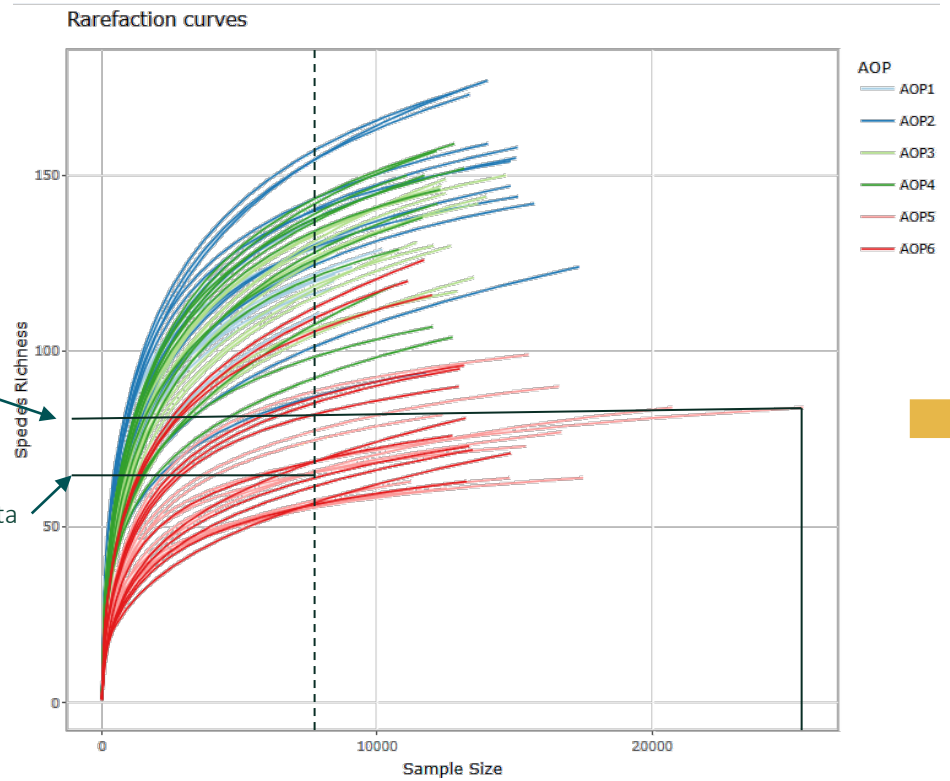
Cancel OK



# α-diversity and sequencing depth

When not at saturation, sequencing depth as a effect on the detected ASV, thus on richness…
and thus on all metrics depending on richness

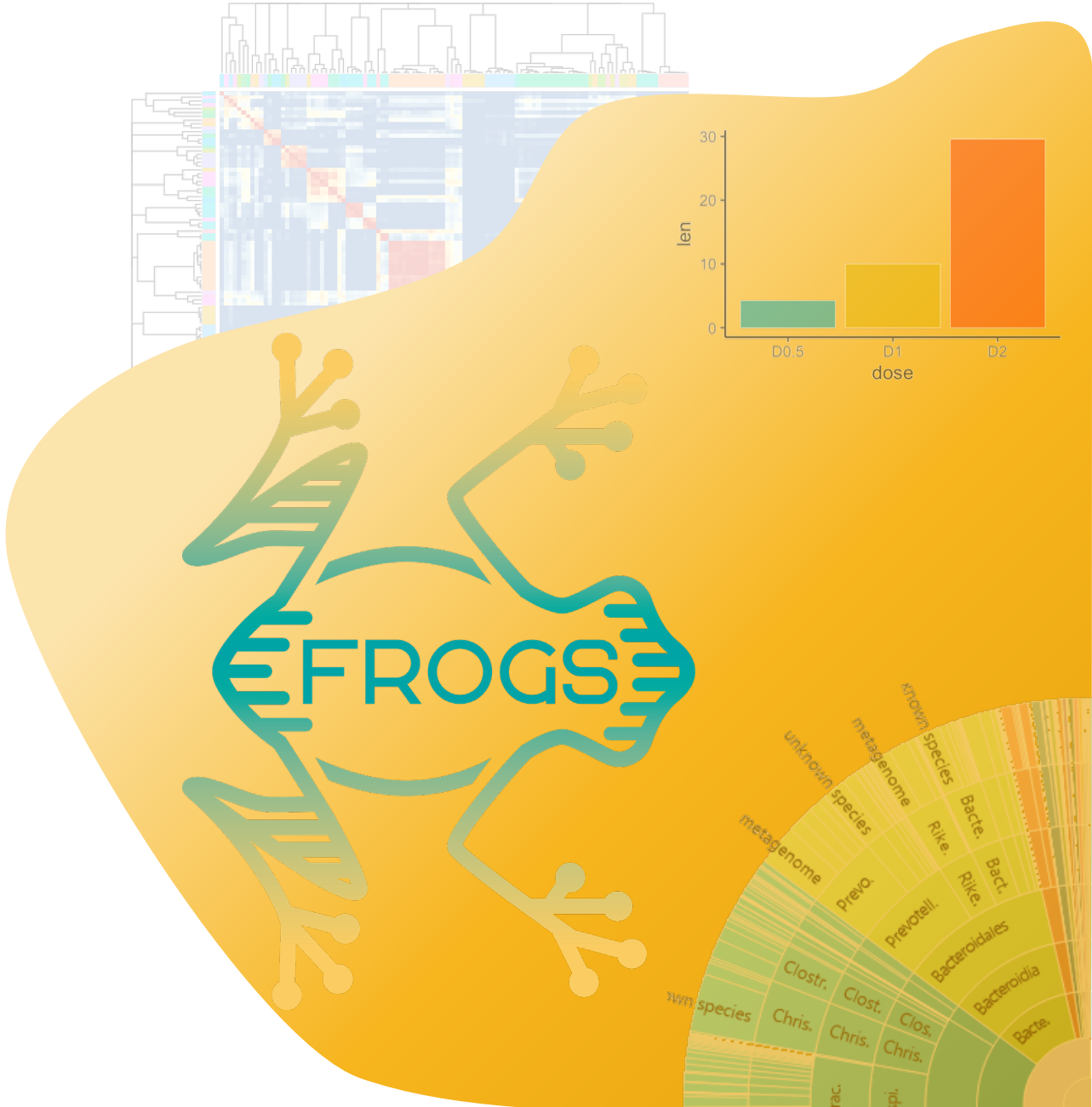
To mitigate this effect, subsampling samples to a common sequencing depth is the “least bad” option




All α-diversity and β-diversity analyses should be realised on subsampled / rarefied data

# FROGS Stats

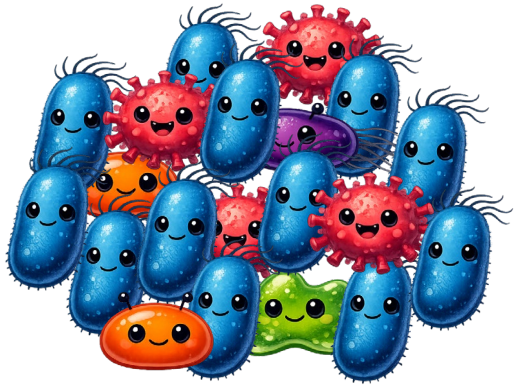
# β-diversity



# How to compare communities?

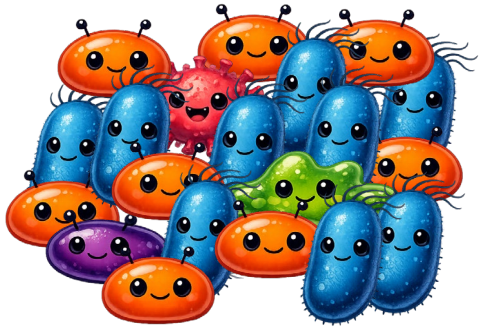
Which community are the more similar?

Sample / Community 1



Sample / Community 2

Sample / Community 3

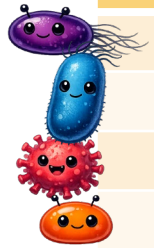


α-diversity tells what happens within a sample.

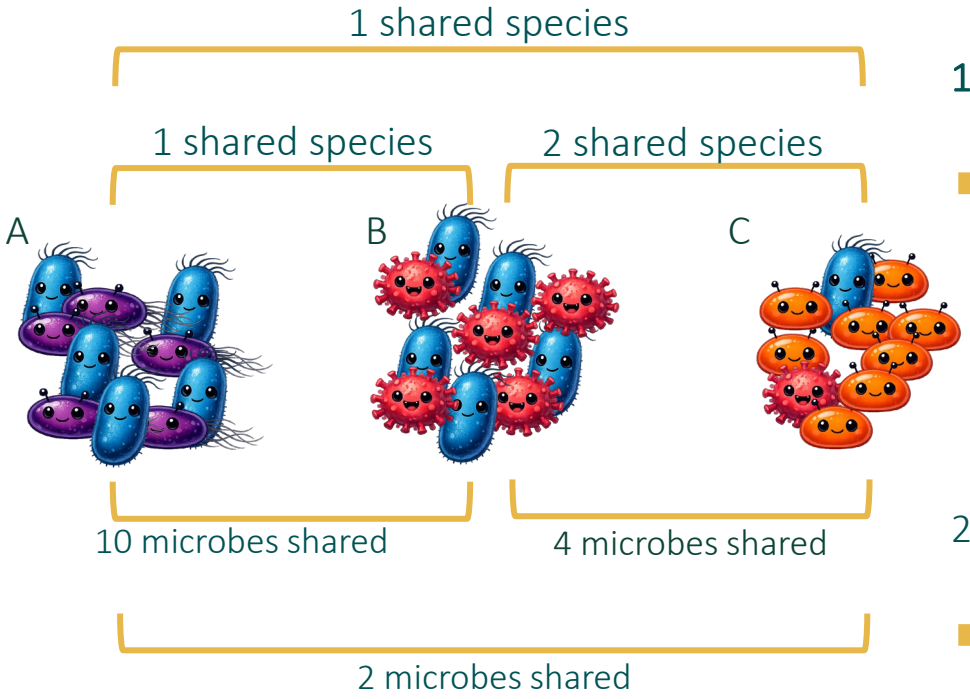
But microbial ecology also has to measure how (and how much) communities change ***across*** samples (or environments or conditions).

β-diversity measures how similar or different microbial communities are across samples.

# How β-diversity is measured?



| A | B | C |
|---|---|---|
| 5 | 0 | 0 |
| 5 | 5 | 1 |
| 0 | 5 | 1 |
| 0 | 0 | 8 |

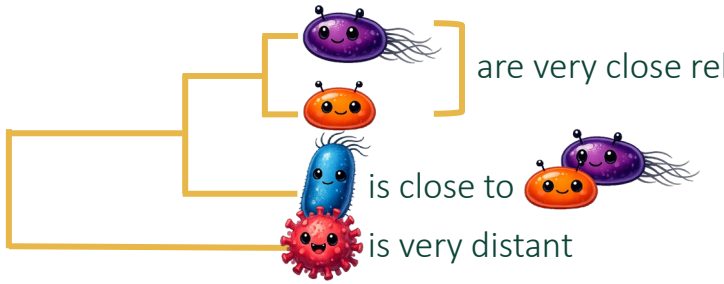



**1-Qualitatively** (no abundance consideration):

➡ B and C are the closest

2-**Quantitatively** (considering abundance):

➡ A and B are the closest




3-**Phylogenetically** (considering phylogeny):

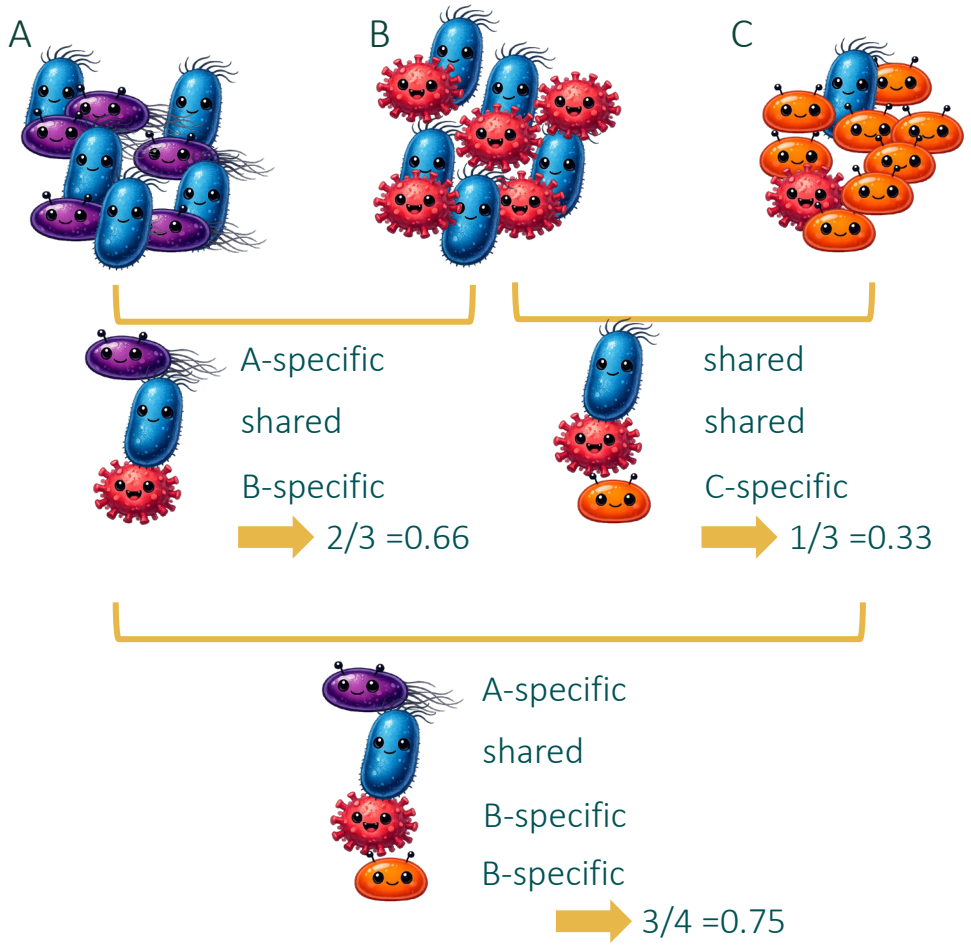
➡ qualitatively B and C are the closest

➡ quantitatively A and C are the closest

# Qualitative Jaccard index

| A | B | C |
|---|---|---|
| 5 | 0 | 0 |
| 5 | 5 | 1 |
| 0 | 5 | 1 |
| 0 | 0 | 8 |

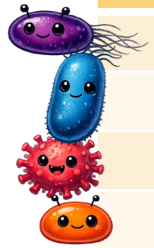
The qualitative Jaccard index measures
the **proportion** of taxa **specific** to either a sample or the other




$$Jaccard_A^B = \frac{taxa_A^{specific} + taxa_B^{specific}}{taxa_{total}}$$

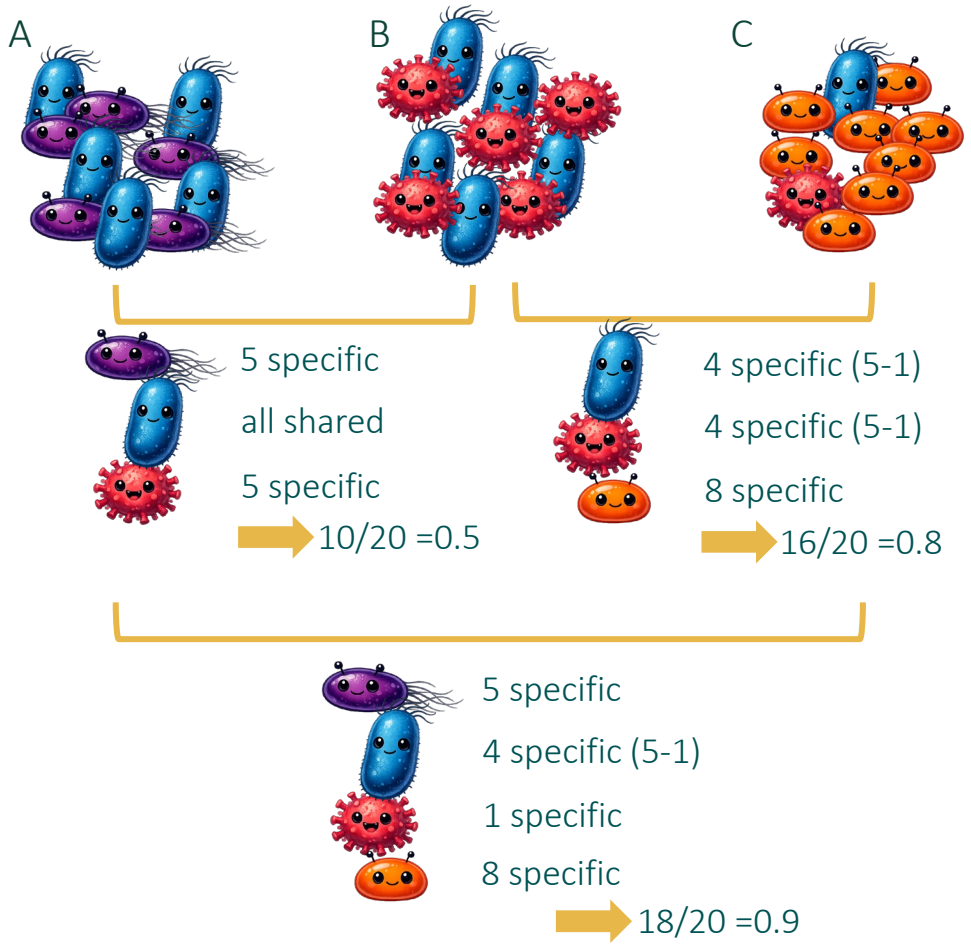
min(Jaccard) = 0 (all taxa are shared)
max(Jaccard) = 1 (all taxa are specific)

# Bray-Curtis index



| A | B | C |
|---|---|---|
| 5 | 0 | 0 |
| 5 | 5 | 1 |
| 0 | 5 | 1 |
| 0 | 0 | 8 |

The Bray-Curtis index measures
the **proportion** of ***abundances*** **specific** to either a sample or the other

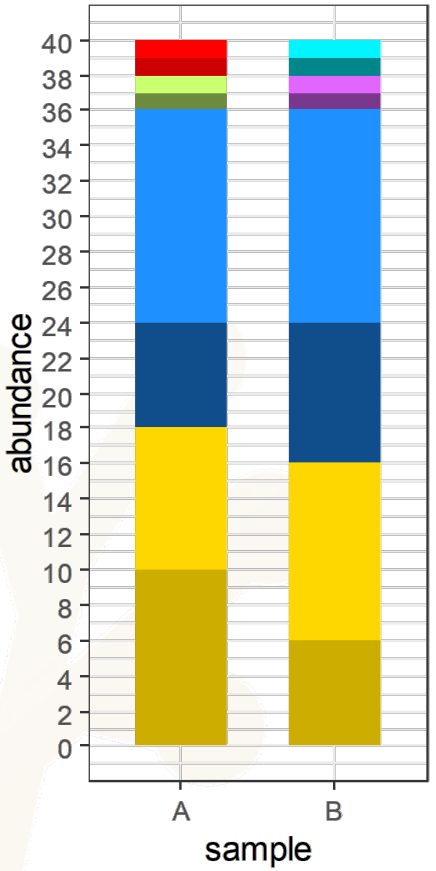



$$Bray_A^B = \frac{abundance_A^{specific} + abundance_B^{specific}}{abundance_{total}}$$

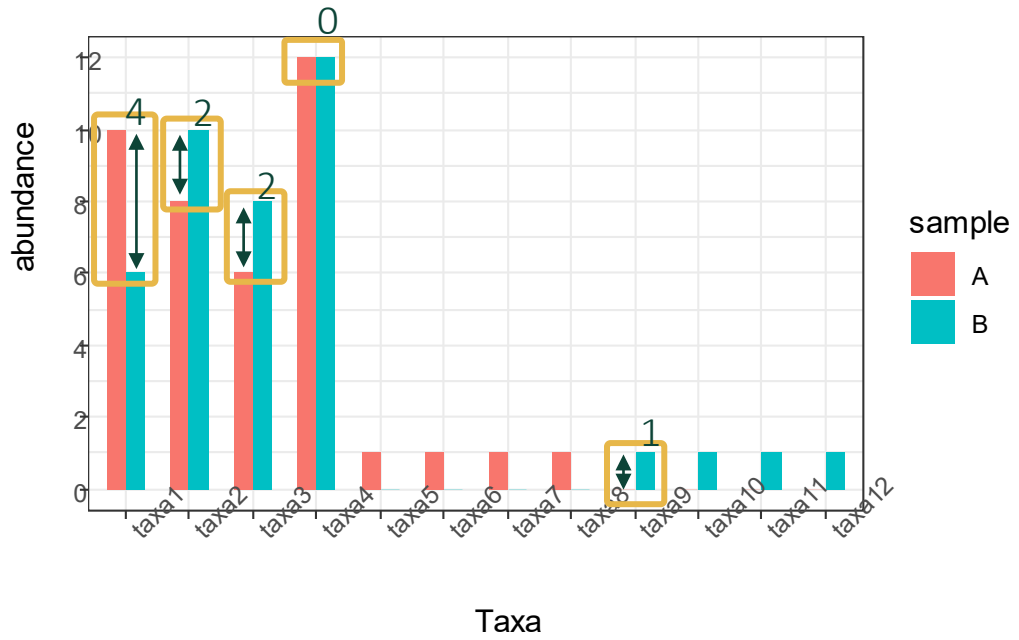
min(Bray) = 0 (all abundances are shared)
max(Bray) = 1 (all abundances are specific)

# Practice session

## Exercise: calculate Jaccard and Bray-Curtis indices




| Taxa | Jaccard | Bray |
|---|---|---|
| taxa12 | 1 | 1 |
| taxa11 | 1 | 1 |
| taxa10 | 1 | 1 |
| taxa9 | 1 | 1 |
| taxa8 | 1 | 1 |
| taxa7 | 1 | 1 |
| taxa6 | 1 | 1 |
| taxa5 | 1 | 1 |
| taxa4 | 0 | 0 |
| taxa3 | 0 | 2 |
| taxa2 | 0 | 2 |
| taxa1 | 0 | 4 |
| | /12 | /80 |




Jaccard: $\frac{0+0+0+0+1+1+1+1+1+1+1+1}{12} = \frac{8}{12} = 0.66$

Bray: $\frac{4+2+2+0+1+1+1+1+1+1+1+1}{16+18+14+24+4+4} = \frac{16}{80} = 0.2$

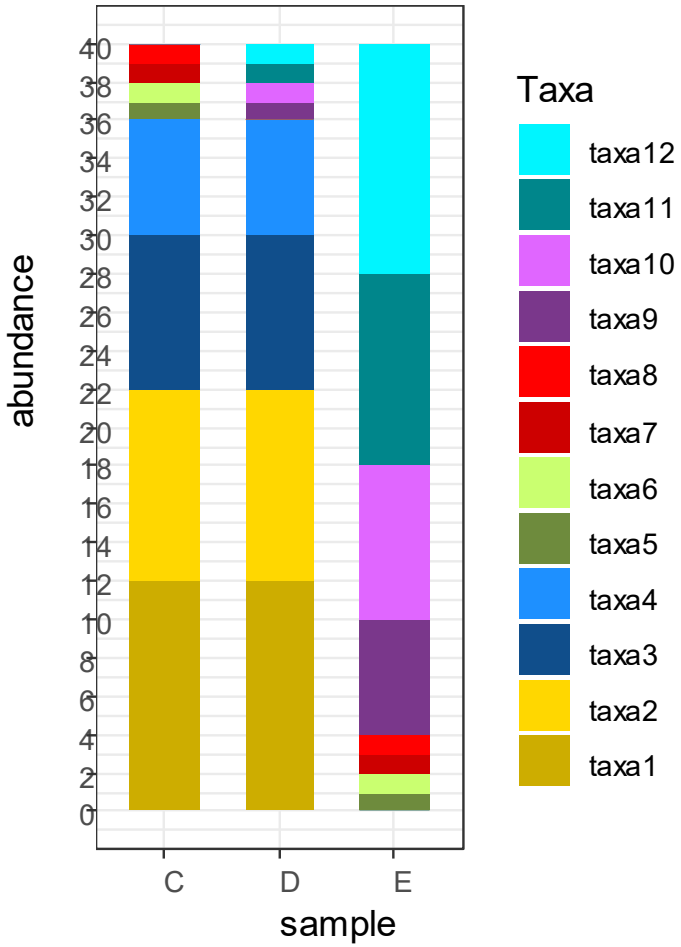
Contribution for each taxa:

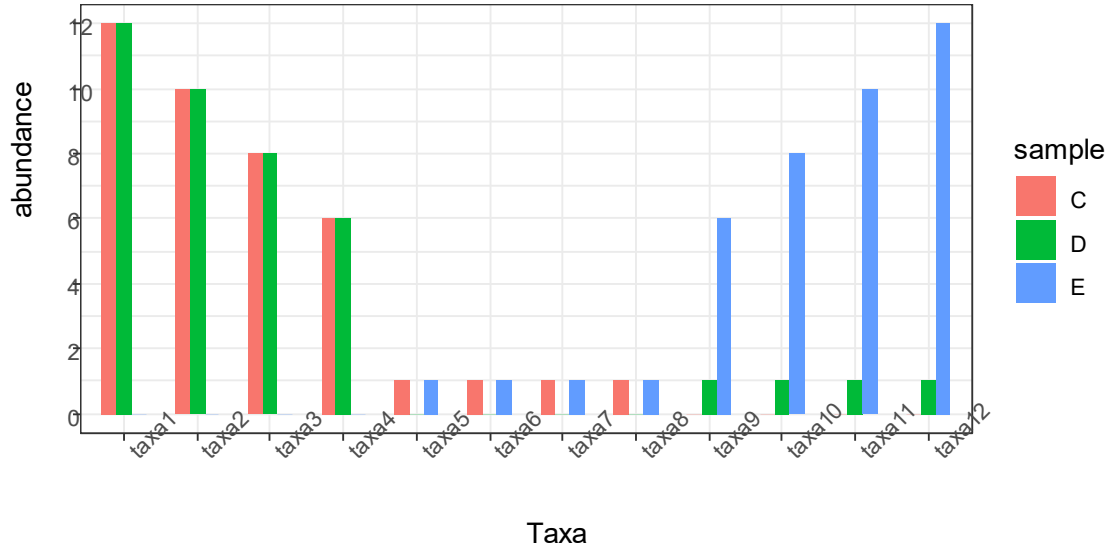
Jaccard: 0 if the taxa is shared, 1 if it is specific

Bray: the absolute value of the differences in abundances for this taxa

# Practice session

## Exercise: calculate Jaccard and Bray-Curtis indices




$$Jaccard_{C>D} = \frac{0+0+0+0+1+1+1+1+1+1+1+1}{12} = \frac{8}{12} = 0.66$$

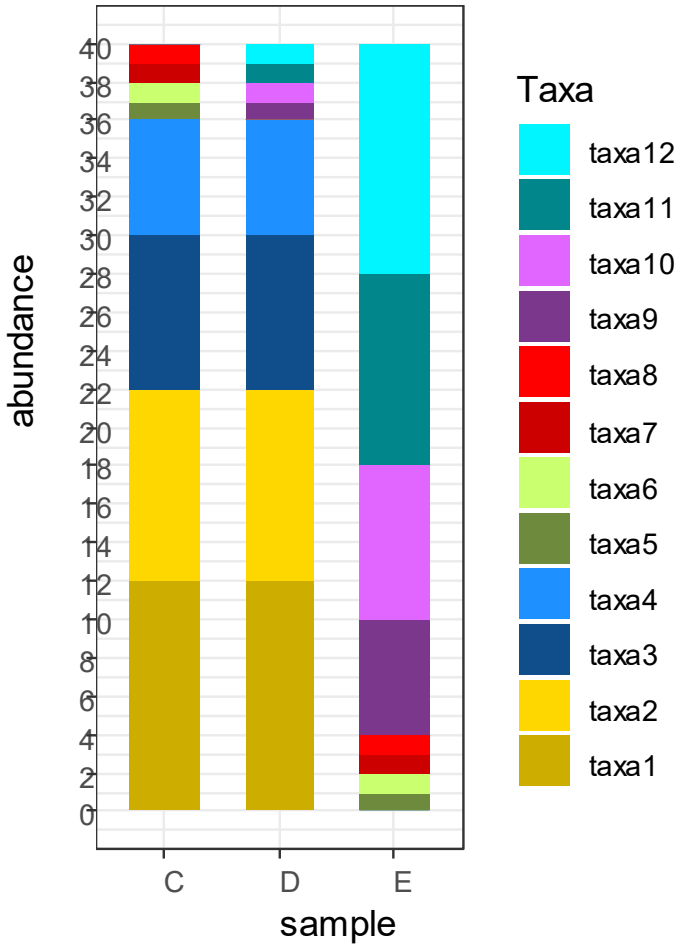
$$Bray_{C>D} = \frac{0+0+0+0+1+1+1+1+1+1+1+1}{24+20+16+12+4+4} = \frac{8}{80} = 0.1$$

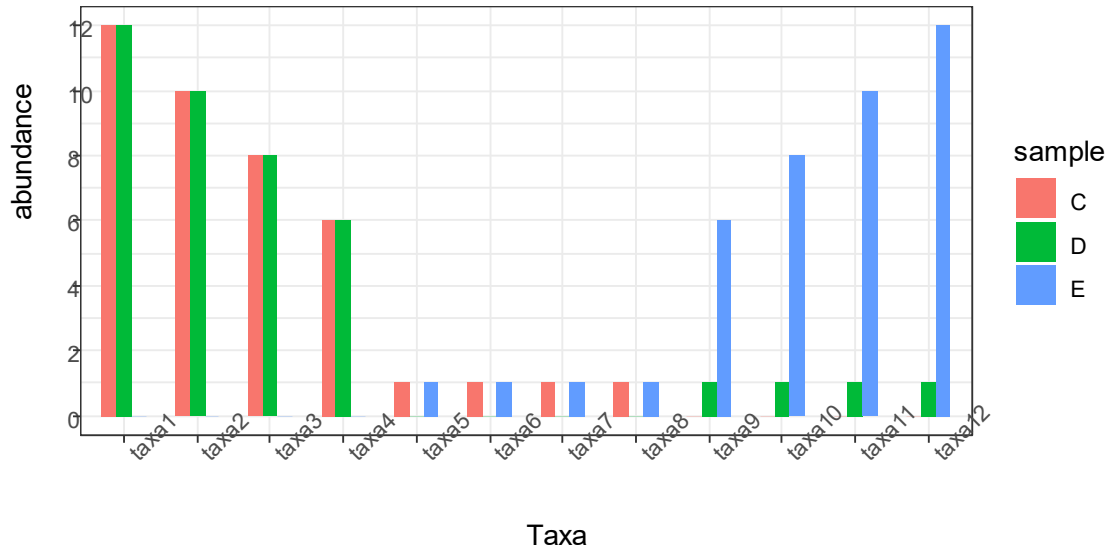
$$Jaccard_{C>E} = \frac{1+1+1+1+0+0+0+0+1+1+1+1}{12} = \frac{8}{12} = 0.66$$

$$Bray_{C>E} = \frac{12+10+8+6+0+0+0+0+6+8+10+12}{24+20+16+12+4+4} = \frac{72}{80} = 0.9$$

# Practice session

## Exercise: calculate Jaccard and Bray-Curtis indices




$Jaccard_{C>D} = 0.66$ High Jaccard -> high proportion of specific ASV

$Bray_{C>D} = 0.1$ Low Bray-Curtis -> high proportion of shared abundances

Combination: *a lot of specific but rare ASV, some shared but abundant*

$Jaccard_{C>E} = 0.66$ High Jaccard -> high proportion of specific ASV

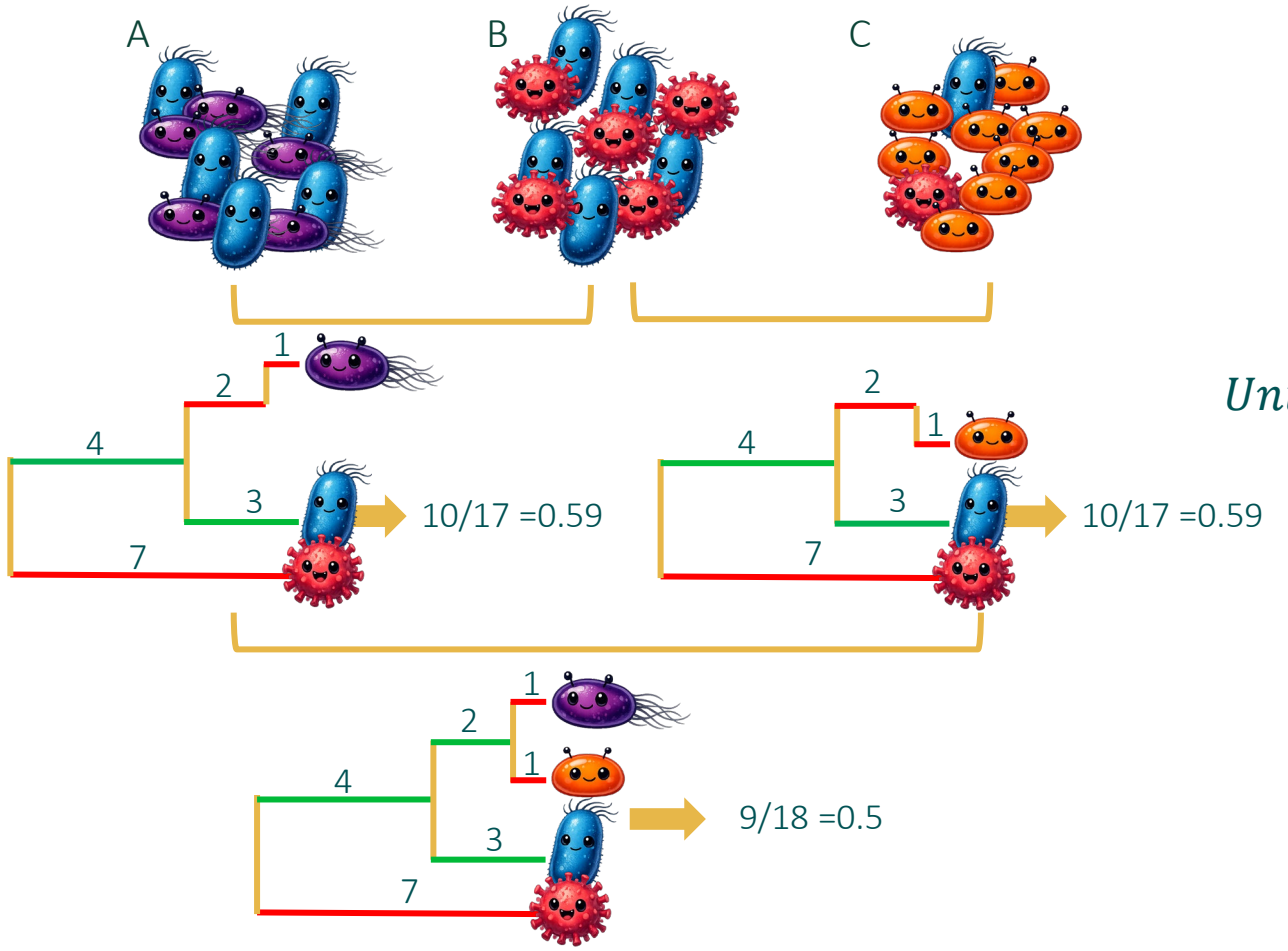
$Bray_{C>E} = 0.9$ High Bray-Curtis -> low proportion of shared abundances

Combination: *a lot of specific and abundant ASV*

# Unifrac index

| | A | B | C |
|---|---|---|---|
| | 5 | 0 | 0 |
| | 5 | 5 | 1 |
| | 0 | 5 | 1 |
| | 0 | 0 | 8 |

The Unifrac index measures
the <u>proportion</u> of ***the length of the phylogenetic tree*** <u>specific</u> to either a sample or the other

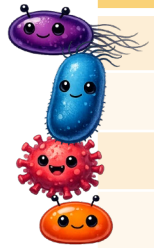



$$Unifrac_A^B = \frac{tree\_length_A^{specific} + tree\_length_B^{specific}}{tree\_length_{total}}$$

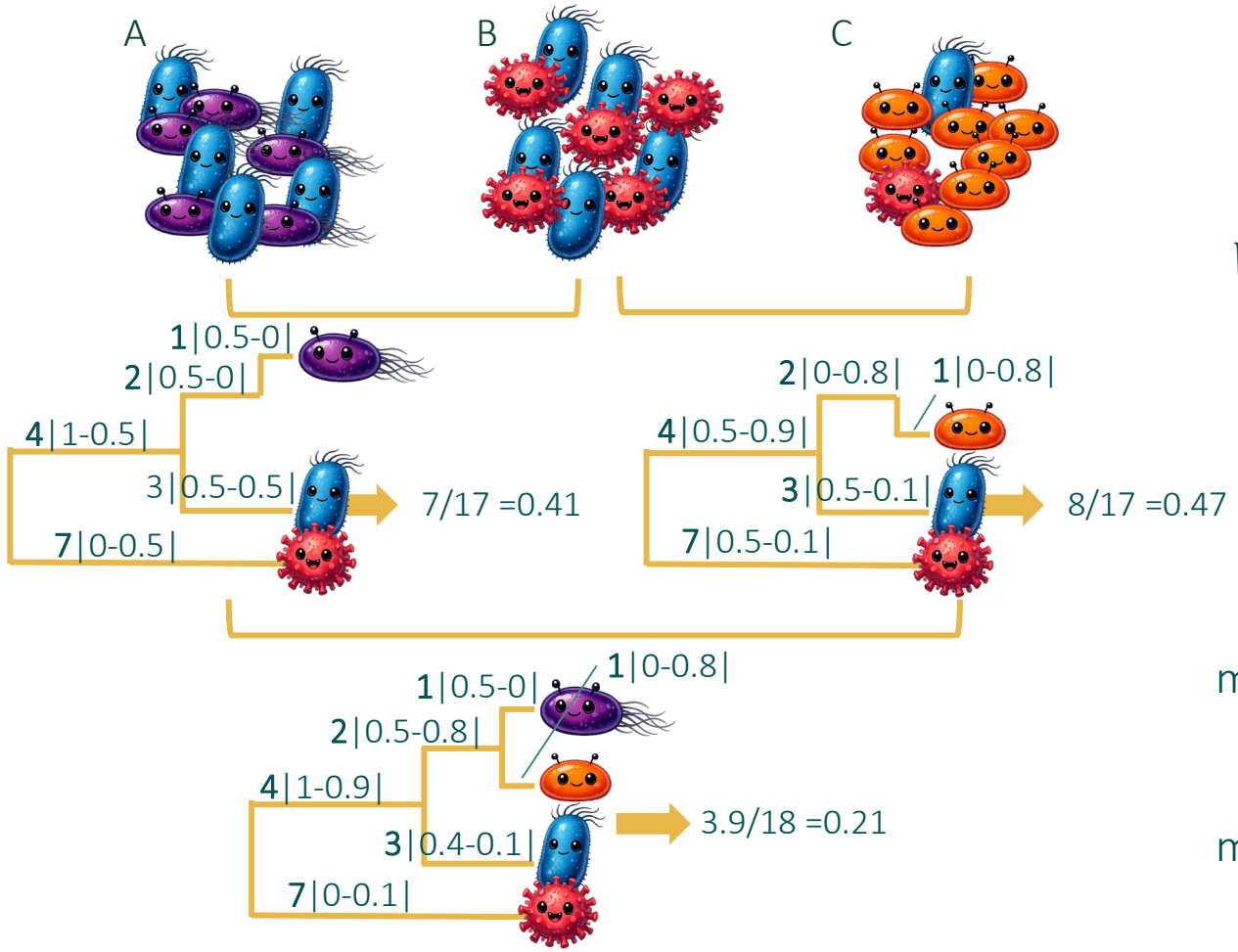
min(Unifrac) = 0
all tree branches are shared (abundances can vary)

max(Unifrac) = 1
all tree branches are specific

# Weighted-Unifrac index

| | A | B | C |
|---|---|---|---|
|  | 5 | 0 | 0 |
| | 5 | 5 | 1 |
| | 0 | 5 | 1 |
| | 0 | 0 | 8 |

The weighted Unifrac index measures
the **proportion** of ***the length of the phylogenetic tree*** **specific** to a sample, ***weighted*** *by the abundance differences*



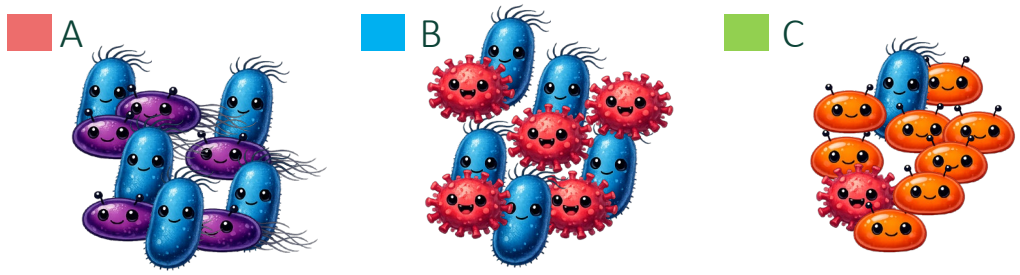

$$WUnifrac_A^B = \frac{\sum reduced\_branch\_length}{\sum non\_reduced\_branch\_length}$$

$reduced\_branch\_length$:
branch length multiplied by the relative abundance of the taxa that descend from it

min(WUnifrac) = 0
all ASV are shared with the same abundances

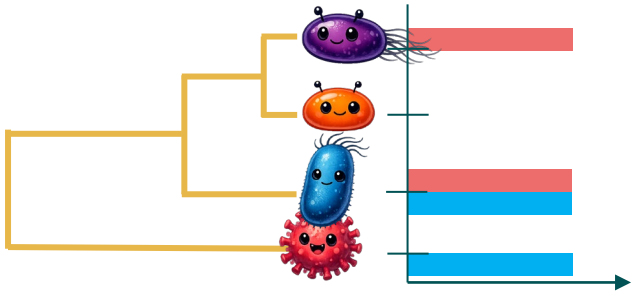
max(WUnifrac) = 1
all ASV are specific and have no tree branch in common

# Summary: β-diversity metrics



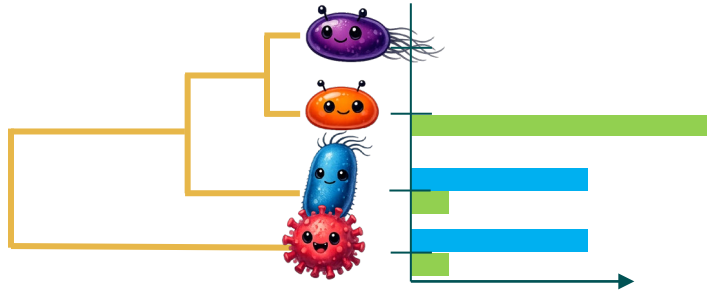

| | Qualitative | Quantitative |
|---|---|---|
| No phylogeny | Jaccard | Bray-Curtis |
| With phylogeny | Unifrac | Weigthed-Unifrac |

A - B



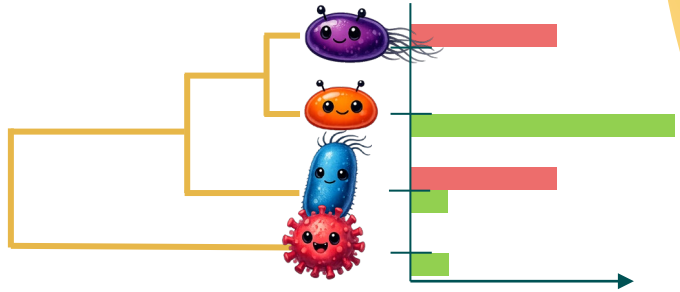
| | |
|---|---|
| Jaccard | 0.66 |
| Bray-Curtis | 0.5 |
| Unifrac | 0.59 |
| Weighted-Unifrac | 0.41 |

B - C



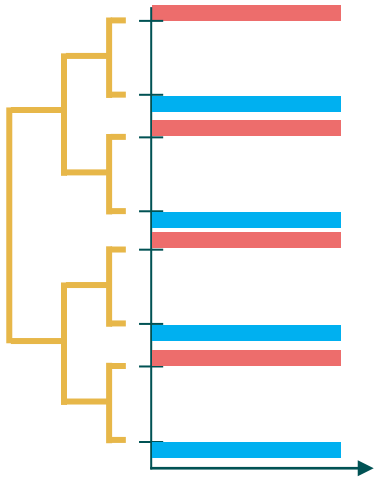
| | |
|---|---|
| Jaccard | 0.33 |
| Bray-Curtis | 0.8 |
| Unifrac | 0.59 |
| Weighted-Unifrac | 0.47 |

A - C



| | |
|---|---|
| Jaccard | 0.75 |
| Bray-Curtis | 0.9 |
| Unifrac | 0.5 |
| Weighted-Unifrac | 0.21 |

# Combining β-diversity indices

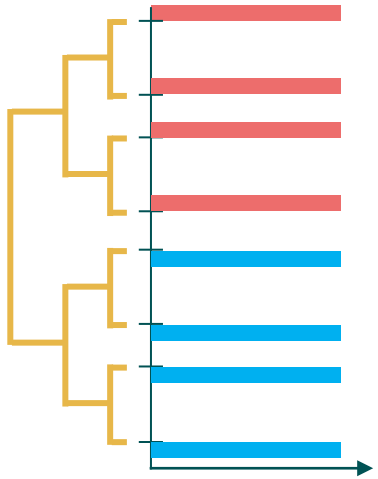


High Jaccard (=1)
all ASV are specific

Low Unifrac:
only the last branches are specific

High Bray (=1)
all abundances are specific

Low wUnifrac
strong abundance differences
but on close branches

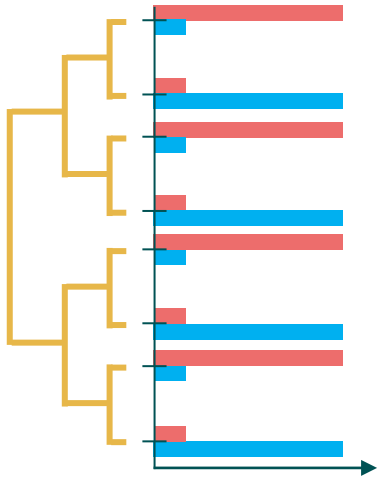


High Jaccard (= 1)
all ASV are specific

High Unifrac (=1)
all the branches are specific

High Bray (=1)
all abundances are specific

High wUnifrac (=1)
strong abundances differences
on independent parts of the tree

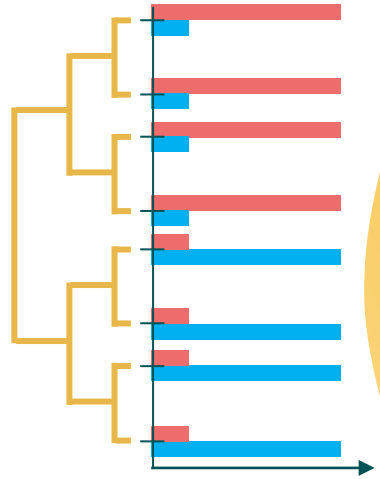


Low Jaccard (=0)
all ASV are shared

Low Unifrac (=0)
all the branches are shared

High Bray
Strong abundances differences

Low wUnifrac
strong abundances differences
but on close branches



Low Jaccard (=0)
all ASV are shared

Low Unifrac (=0)
all the branches are shared

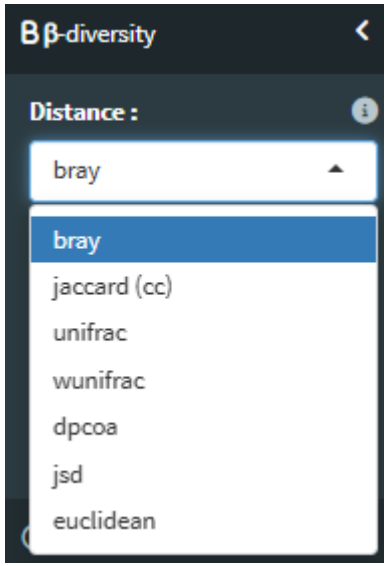
High Bray
Strong abundances differences

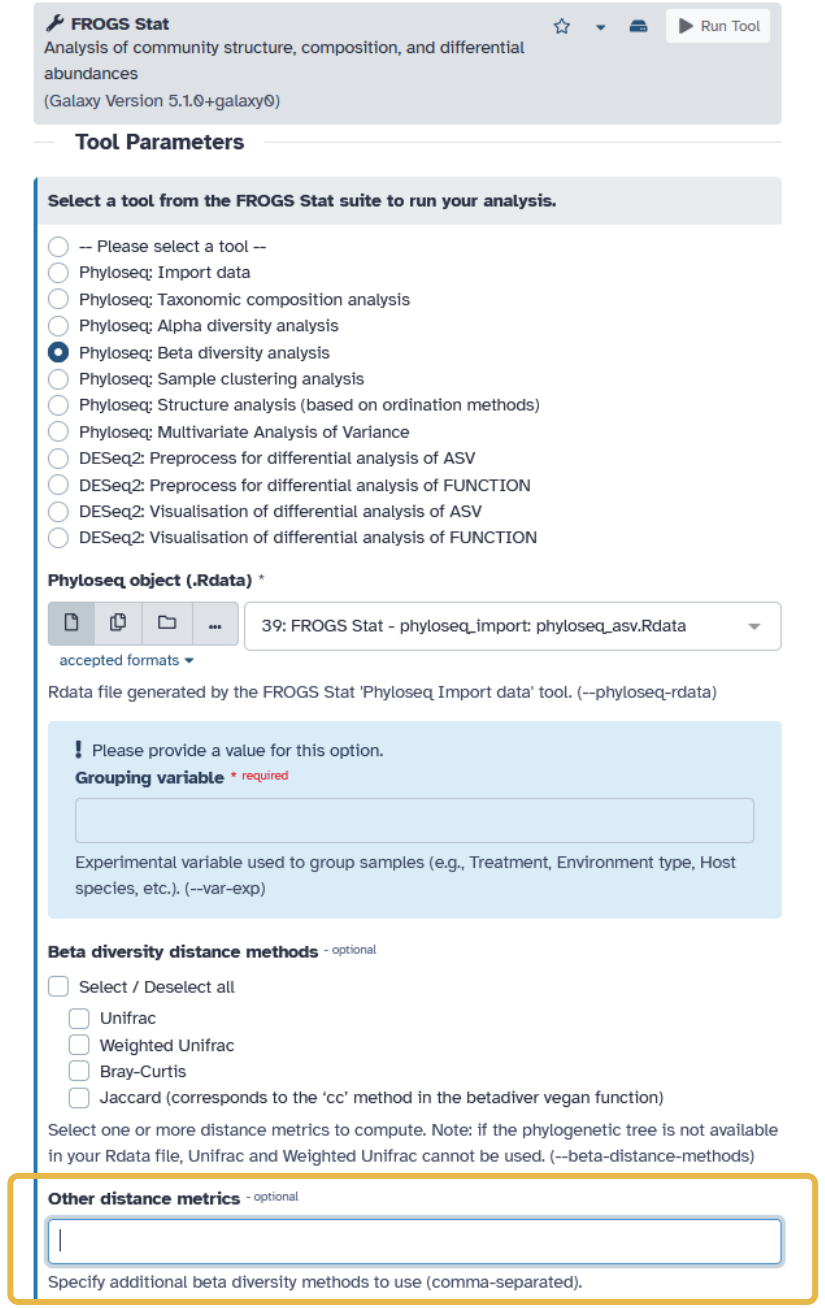
High wUnifrac
strong abundances differences
between the basal branches

# Other β-diversity metrics

Several other distances exist.
Phyloseq currently supports 43 β-diversity distance methods.
They are all accessible through the FROGS Stat Galaxy tool:

On Easy16S, a small selection is avaible, including the 4 most usual, bray, jaccard, unifrac and wunifrac:

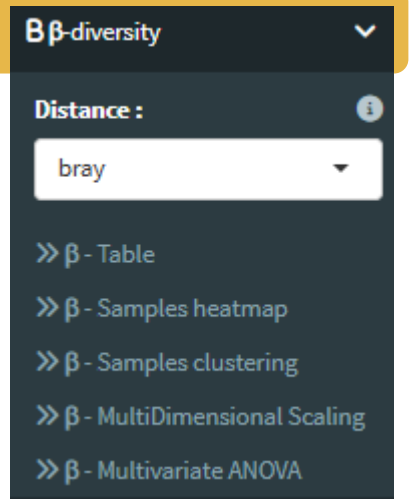


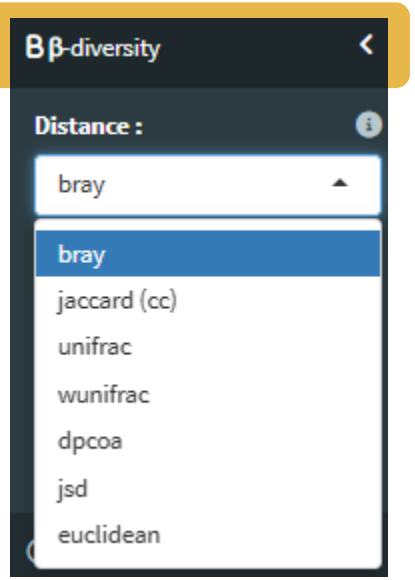
# Practice session

## Analyse β-diversity

In Easy16S select




And choose a distance



# Practice session

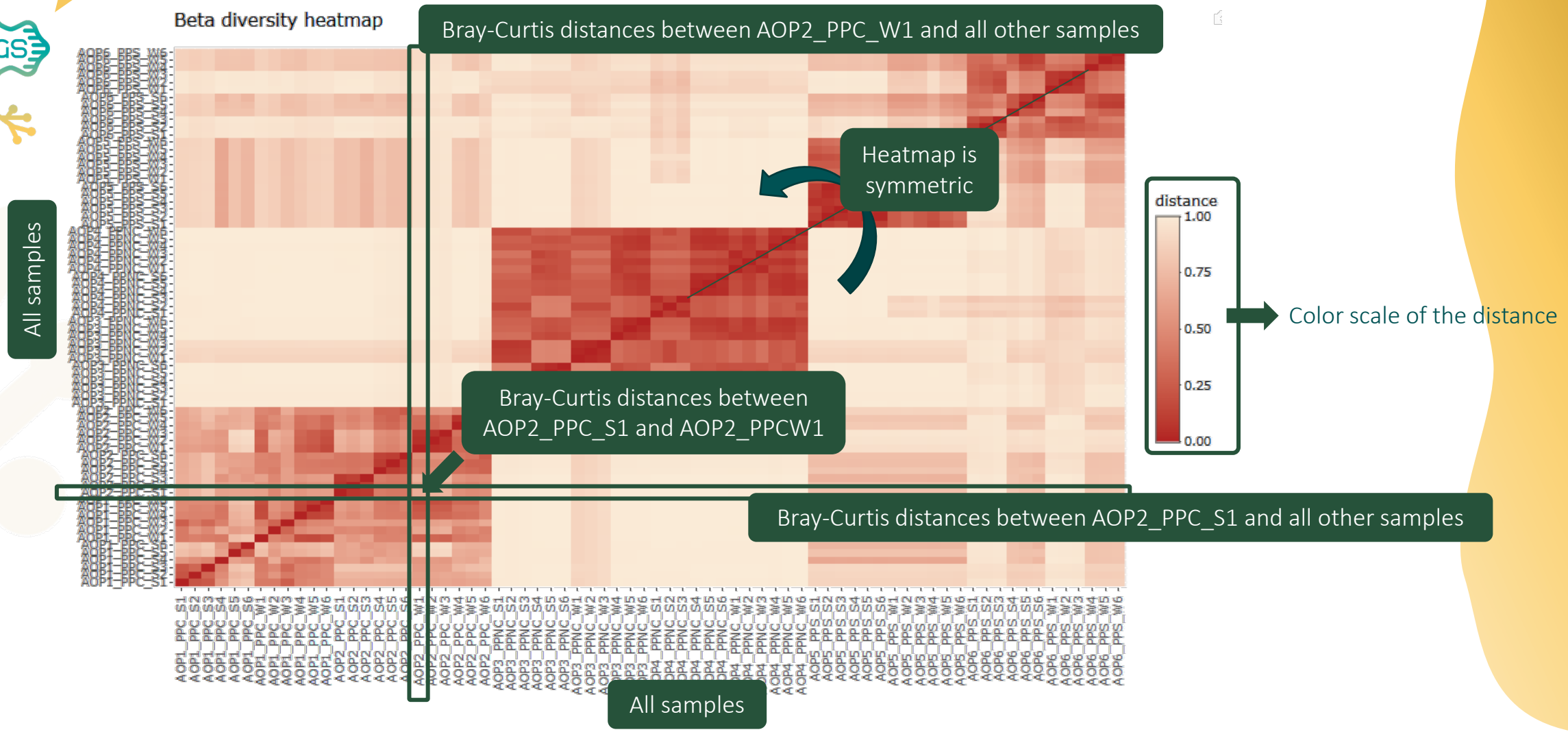
## Draw β-diversity heatmaps

How to read the output?

What biological interpretation can be extracted?

# Draw β-diversity heatmaps

How to read the output?



# Practice session

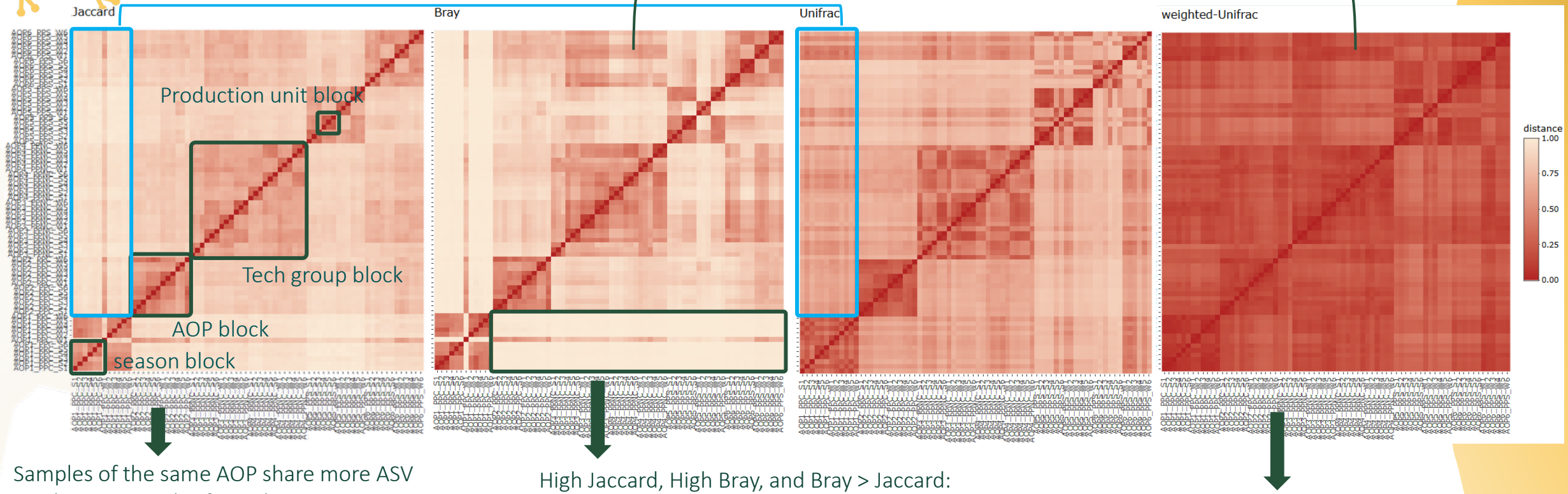
## Draw β-diversity heatmaps

What biological interpretation can be extracted?

16S dada2

High Jaccard, Low Unifrac:
lots of specific ASV, but with close relatives

High Bray, Low weighted-Unifrac:
strong abundances differences but on close branches




Samples of the same AOP share more ASV
Inside AOP samples from the same site
(1-3 & 4-6 replicates) share more ASV
Seasons effect for AOP1 ; PPNC effect.

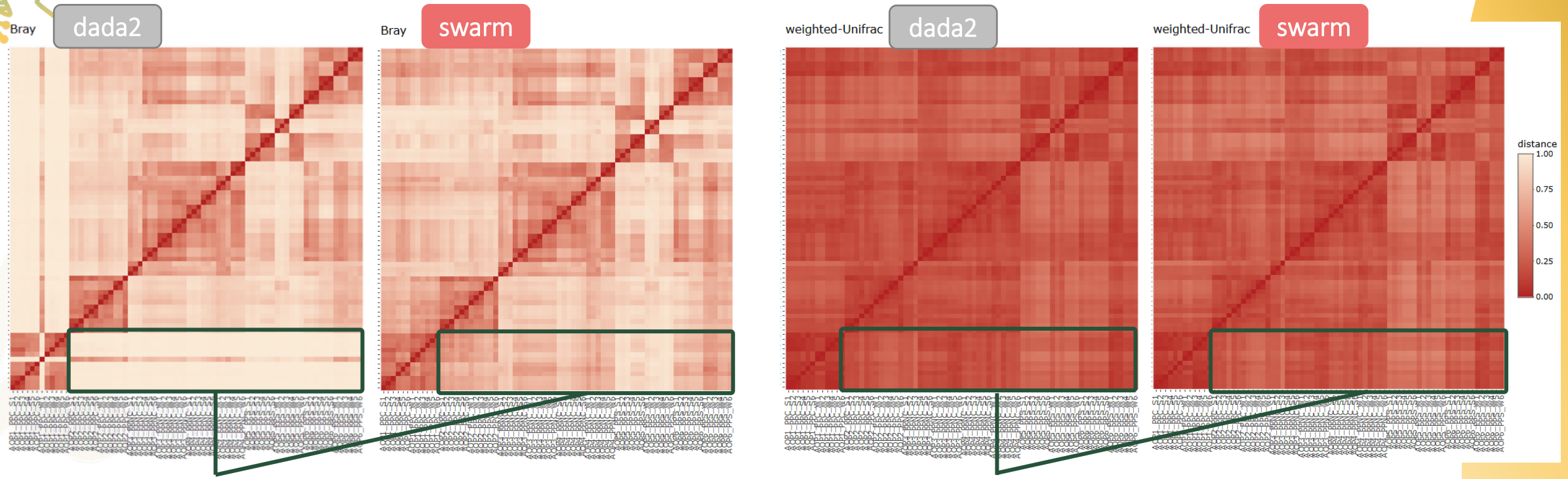
High Jaccard, High Bray, and Bray > Jaccard:
lots of specific ASV with high abundances

Very low wUnifrac, without strong structuration:
All samples are similar in phylogenetic abundance

# Practice session

## Draw β-diversity heatmaps

What biological interpretation can be extracted?




Dada2 shows strong differences where swarm shows weaker ones:
Dada2 split ASVs of the same species into several distinct ASV

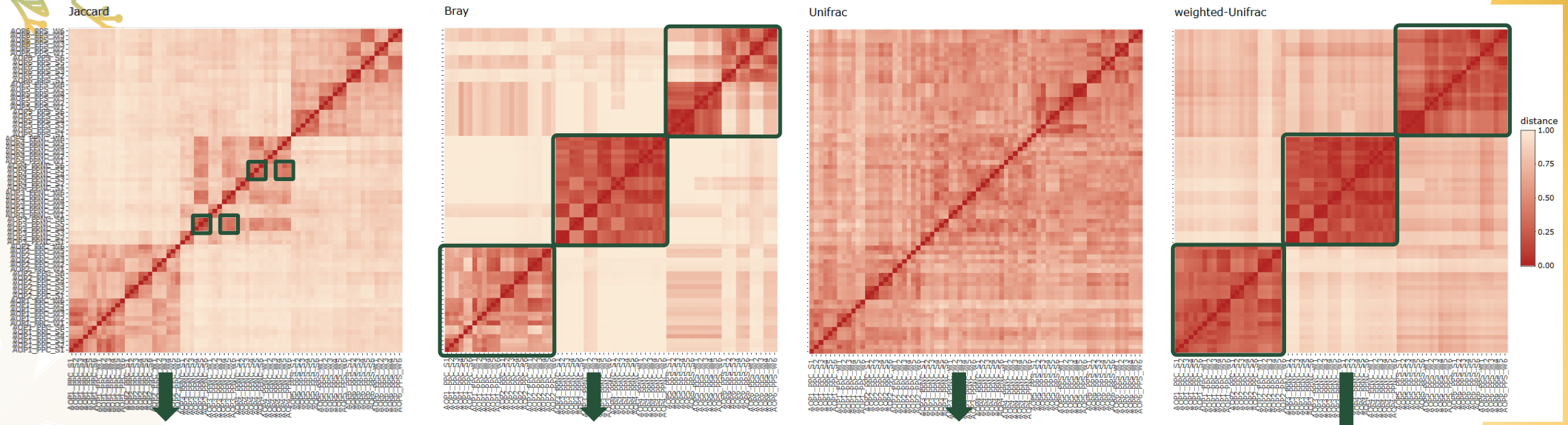
Strong impact on Jaccard and Bray-Curtis

Considering phylogeny, as the split ASVs are very close, there are only small differences between dada2 and swarm

# Practice session

## Draw β-diversity heatmaps

What biological interpretation can be extracted?




Quite High Jaccard even inside AOPs
Strong “production unit” effects

Strong intra/inter families differences
Strong tech families effect

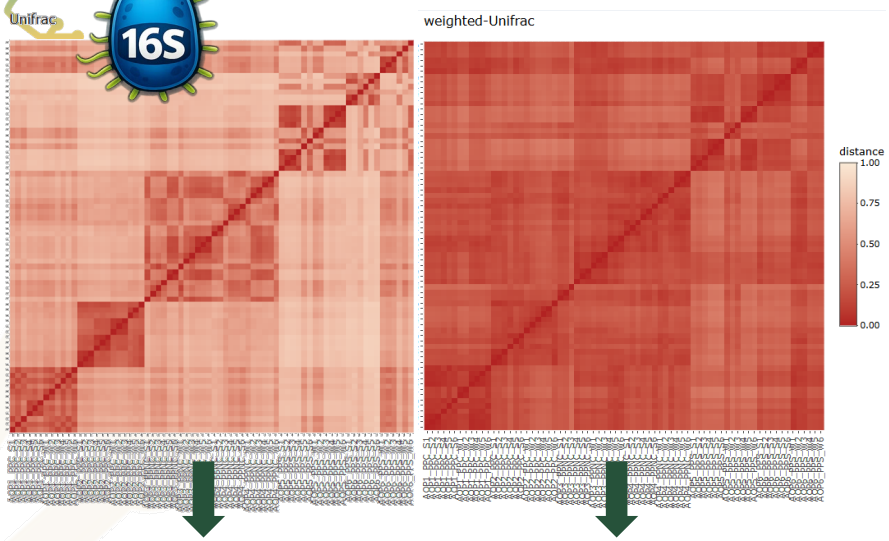
Almost no structuration and Unifrac<Jaccard
Specific but close ASV, random differences

Strong intra/inter families differences
Strong tech families effect

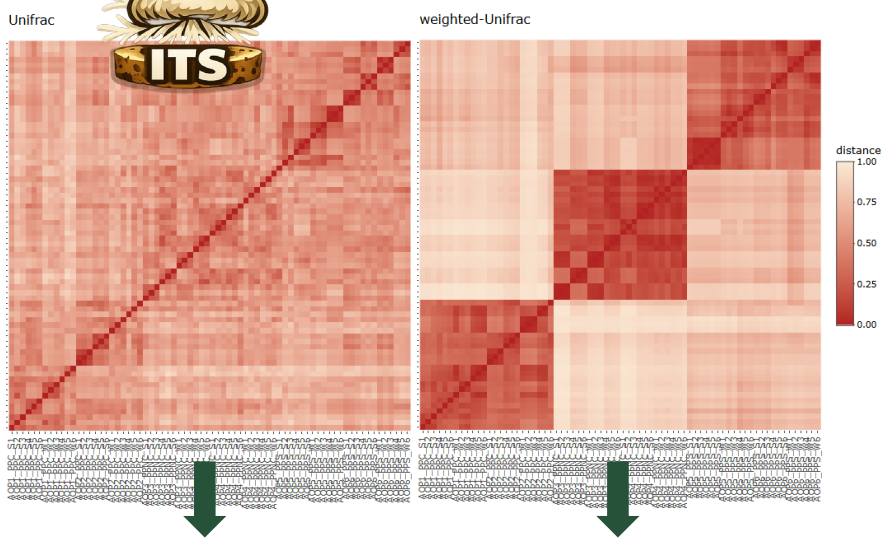
# Practice session

## Draw β-diversity heatmaps

What biological interpretation can be extracted?




Bacterial phylogenetic composition depends on tech families and AOP for rare ASVs but the abundant taxa are similar for all conditions



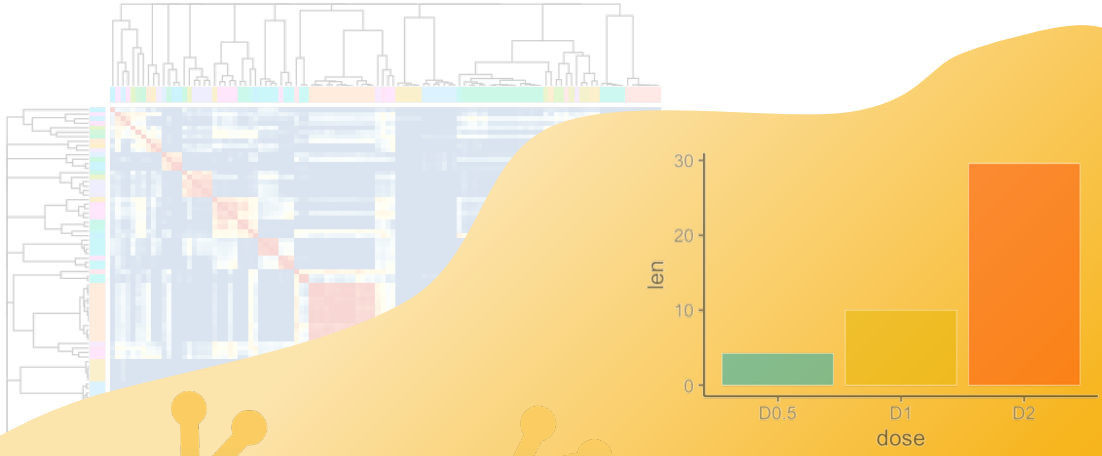

Fungal phylogenetic composition randomly varies between samples, but the abundant taxa are strongly dependent on tech families.

ITS is not a good marker for phylogeny and Unifrac distances may not correctly reflect phylogeny

# FROGS Stats

## β-diversity

## Structure analyses






# How communities are structured based on β-diversity?

Are there better solutions than eye-observations on heatmaps?

# Exploring structures: the hierarchical clustering

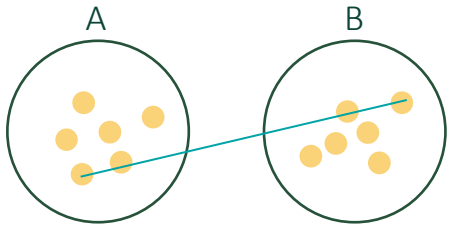
1- **Find the two most similar samples** (the smallest distance)
2- **Merge them into a group linked by a node** (keep the distance between them)
3- **Recalculate the distance between this new group and the remaining samples**
(depending on the linkage method)
4- **Repeat until all samples are merged into a single tree**

Hierarchical clustering produces a tree where it is possible to:

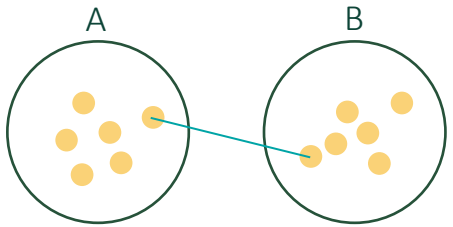
- Visualise the structure of similarities between samples
- Identify natural groups (clusters)
- Visualise hierarchical relationships (who is similar to who, at which level)
- Detect subgroups and subgroups hierarchy

# Exploring structures: the hierarchical clustering

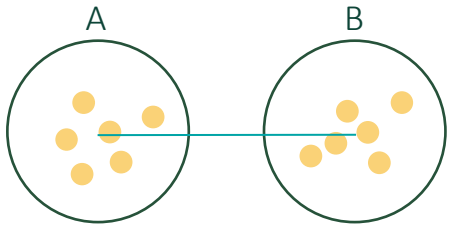
Linkage methods: several exist and are available but the main are




Complete linkage
d(A,B)= longer distance between elements in A and B
Tends to produce compact spherical clusters, can fragment a lot




Single linkage
d(A,B) = shorter distance between elements in A and B
Tends to produce chain-like clusters, noise-sensitive





Average linkage
d(A,B) = average distance between elements in A and B
Preserves the actual structure without structure assumptions

ward.D2 linkage produces visually beautiful clusters but relies on several mathematical hypothesis that are not respected by β-diversity distances.

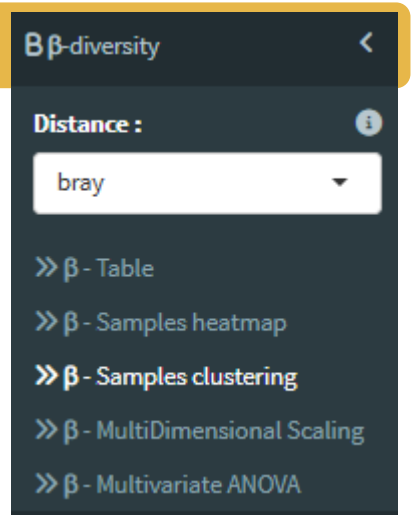
# Practice session

## Explore β-diversity structure with hierarchical clustering

How to read the output?

What biological interpretation can be extracted?

In Easy16S select



# Practice session

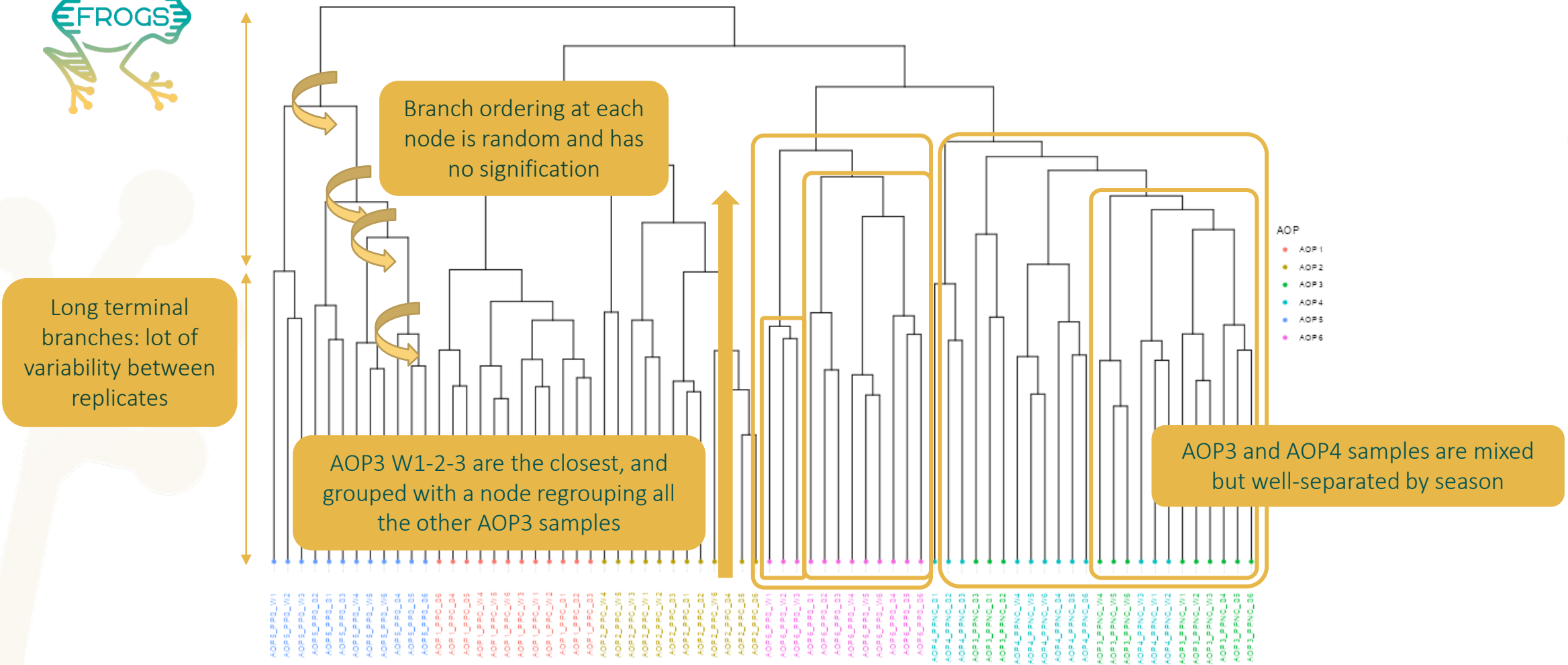
## Explore β-diversity structure with hierarchical clustering

How to read the output?

Bottom-up reading

swarm

Jaccard



# Practice session

## Explore β-diversity structure with hierarchical clustering

How to read the output?

Top-down reading

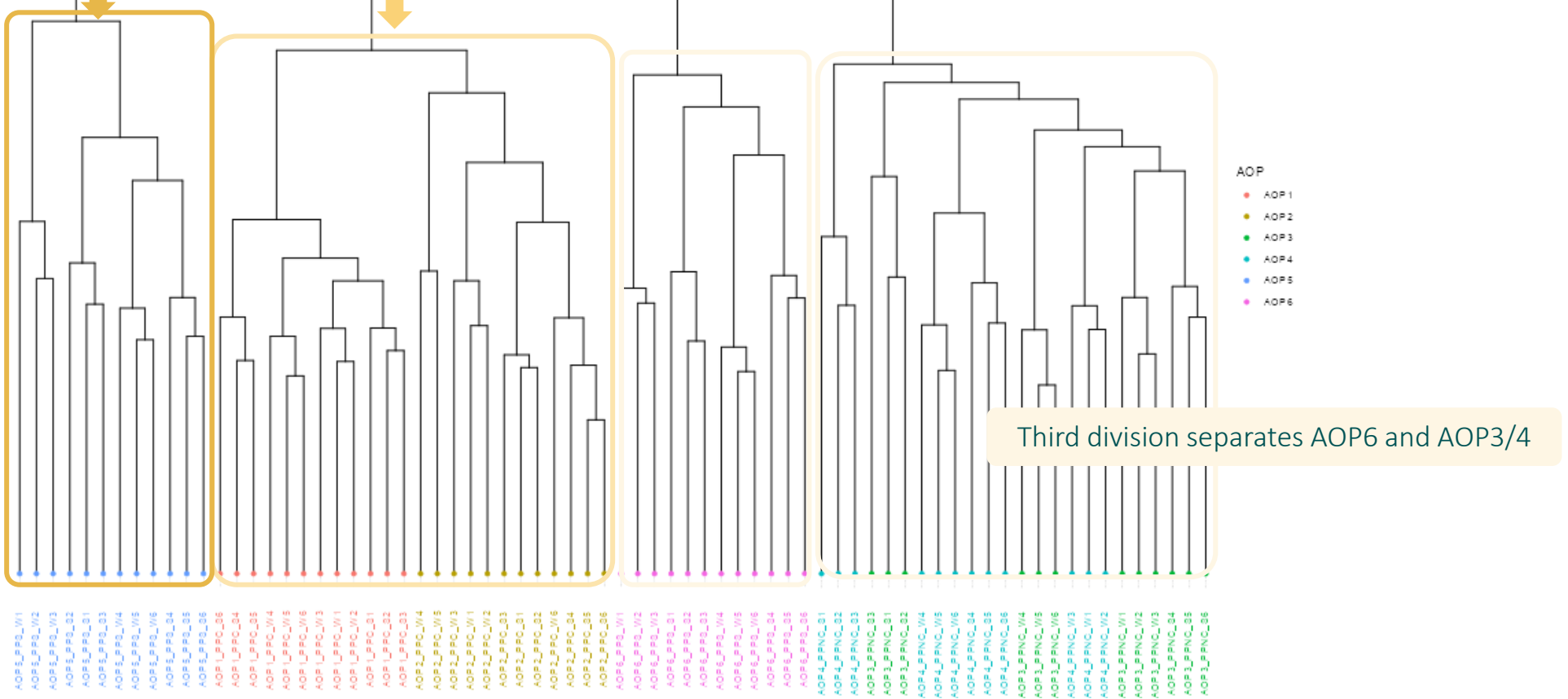
swarm

Jaccard

First separation is between AOP5 and all others

Second division separates PPC and others

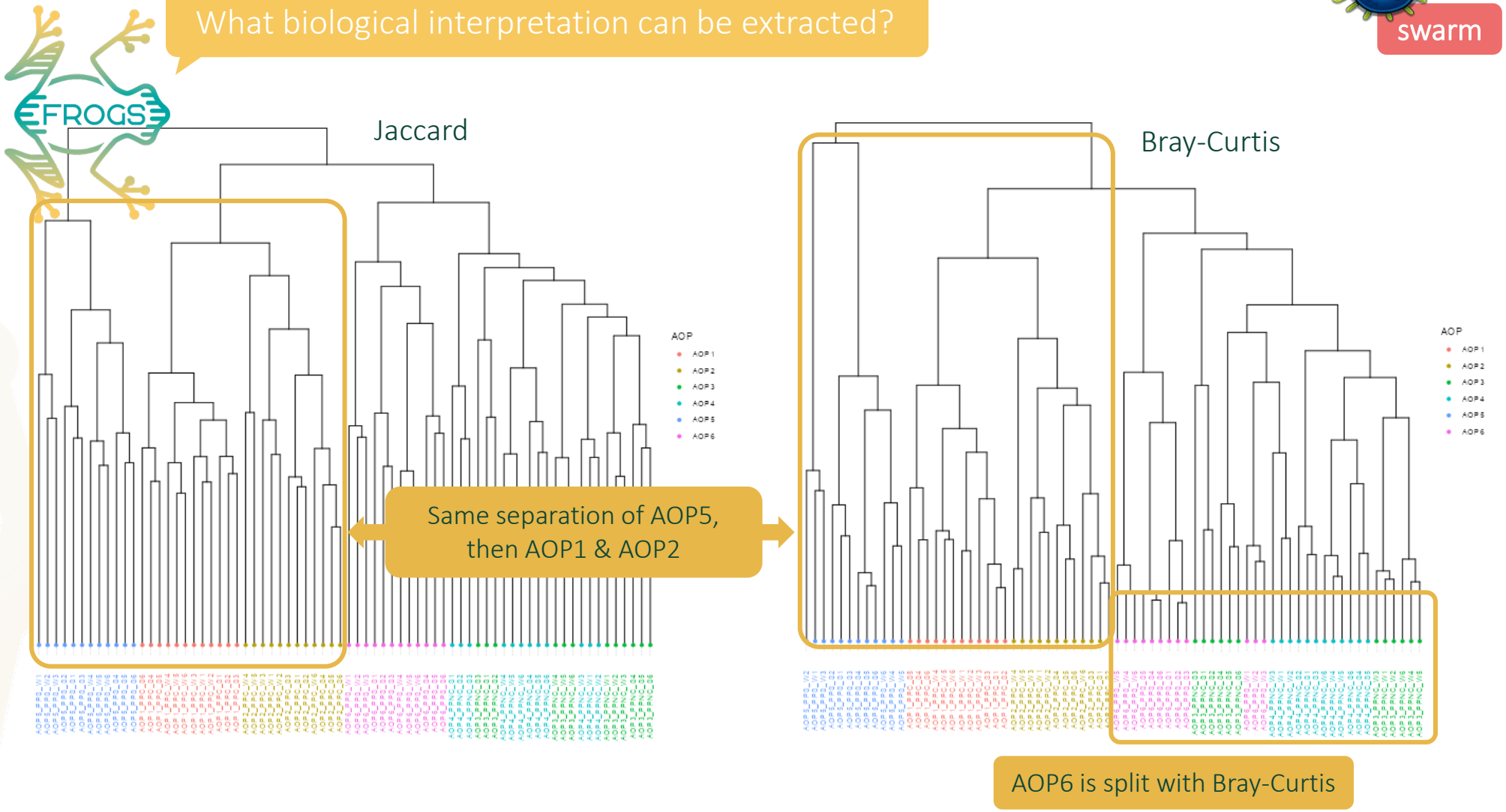
Third division separates AOP6 and AOP3/4



# Practice session

## Explore β-diversity structure with hierarchical clustering

What biological interpretation can be extracted?

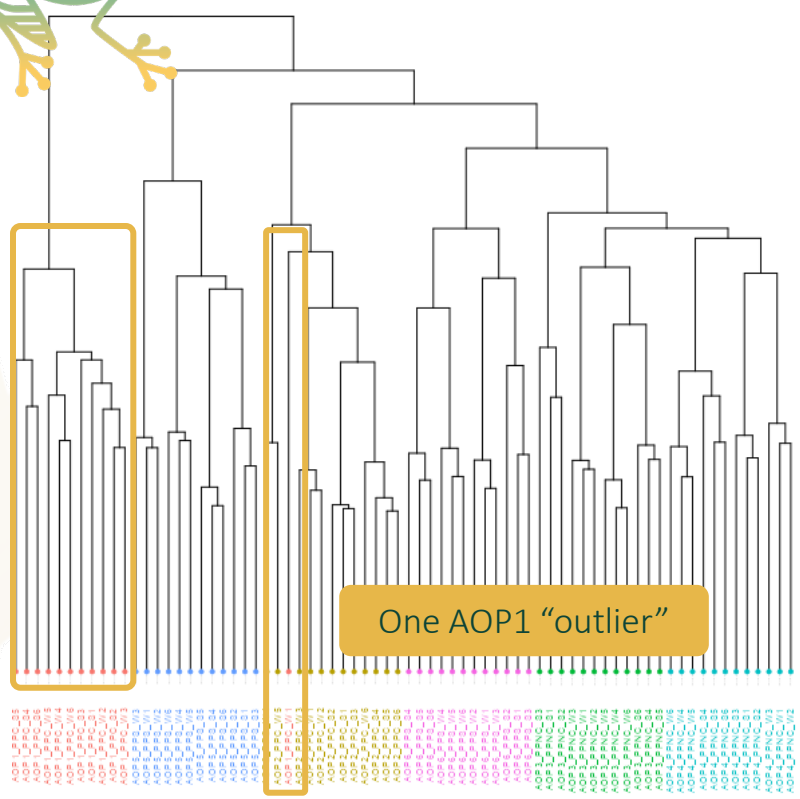


# Practice session

## Explore β-diversity structure with hierarchical clustering

What biological interpretation can be extracted?

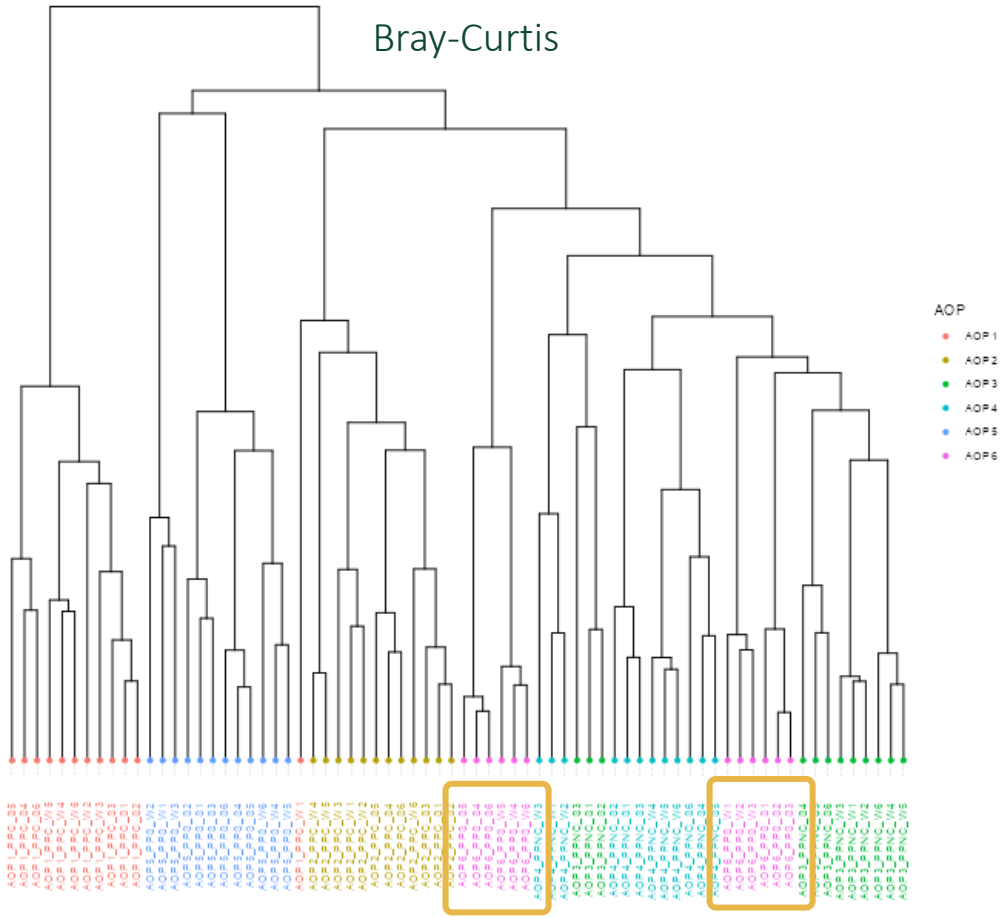
Jaccard



One AOP1 “outlier”

AOP1 is the outgroup instead of AOP5

Bray-Curtis



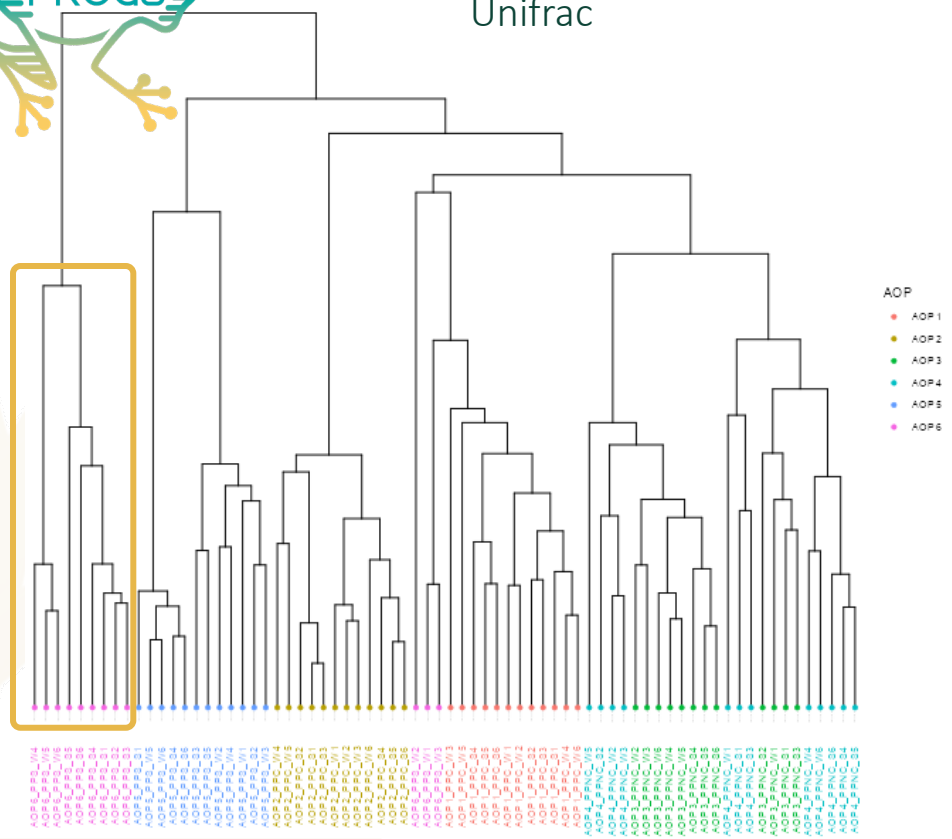

Production units are split for AOP6

# Practice session

## Explore β-diversity structure with hierarchical clustering

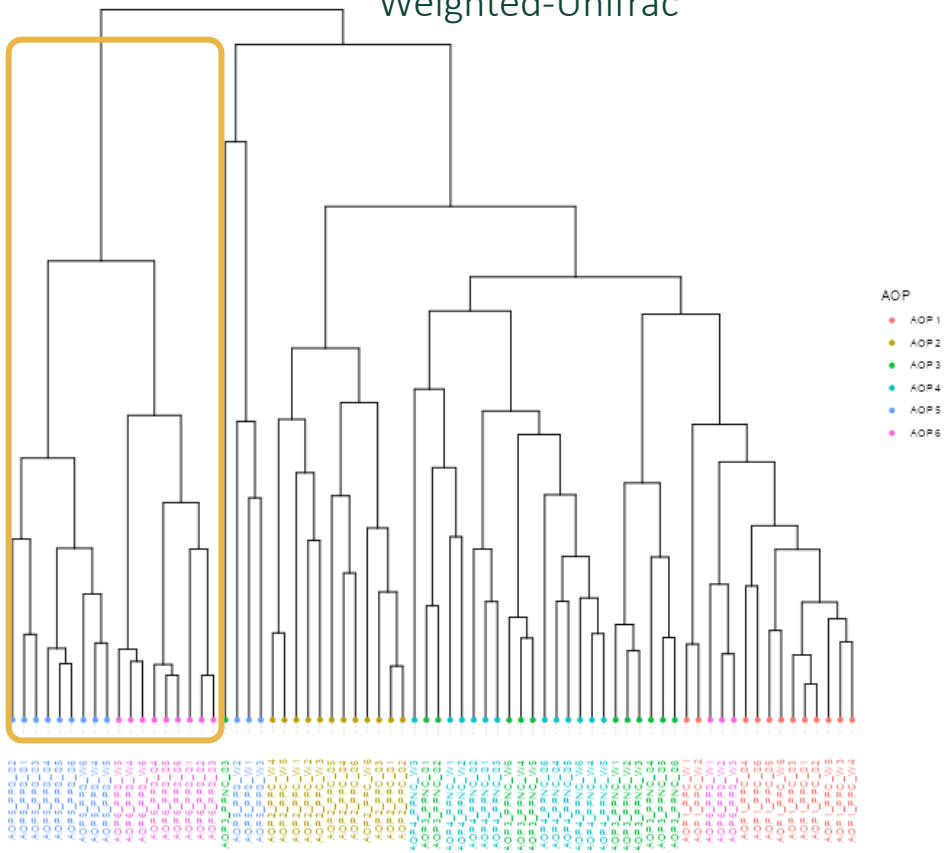
What biological interpretation can be extracted?

Unifrac




Accounting for phylogeny
AOP6 is the most different

Weighted-Unifrac

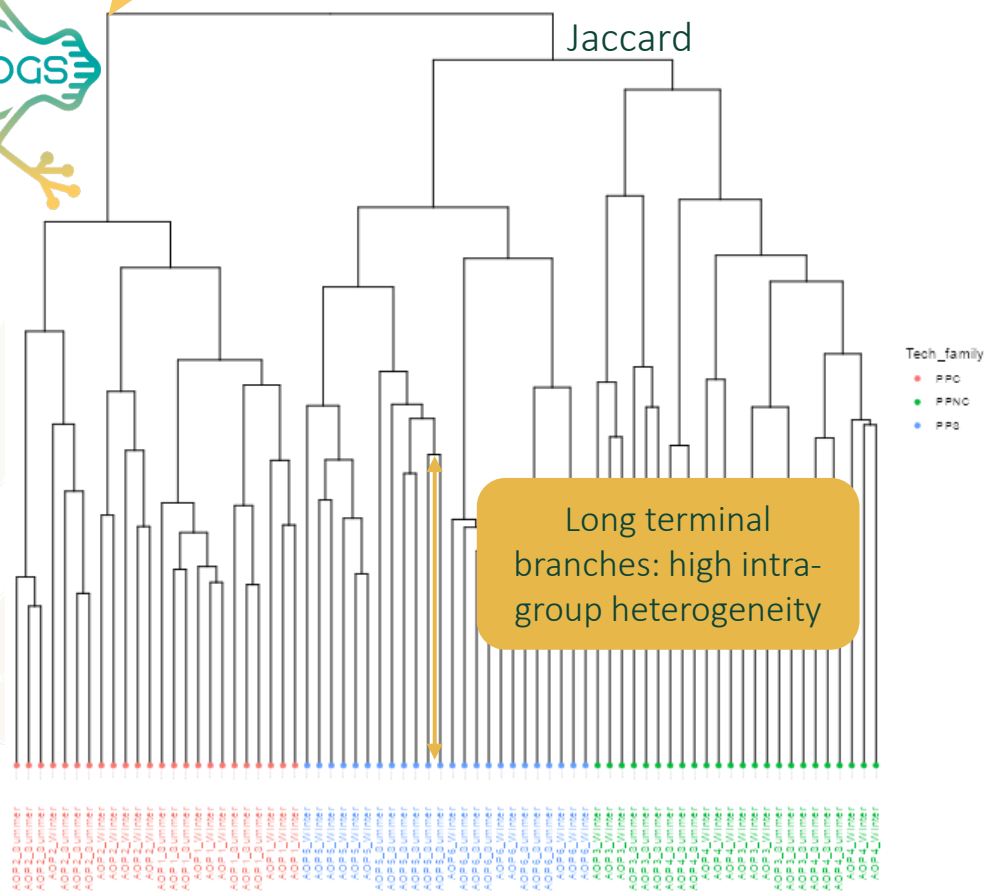



AOP5 and AOP6 are grouped
(but not all AOP5, some are split)

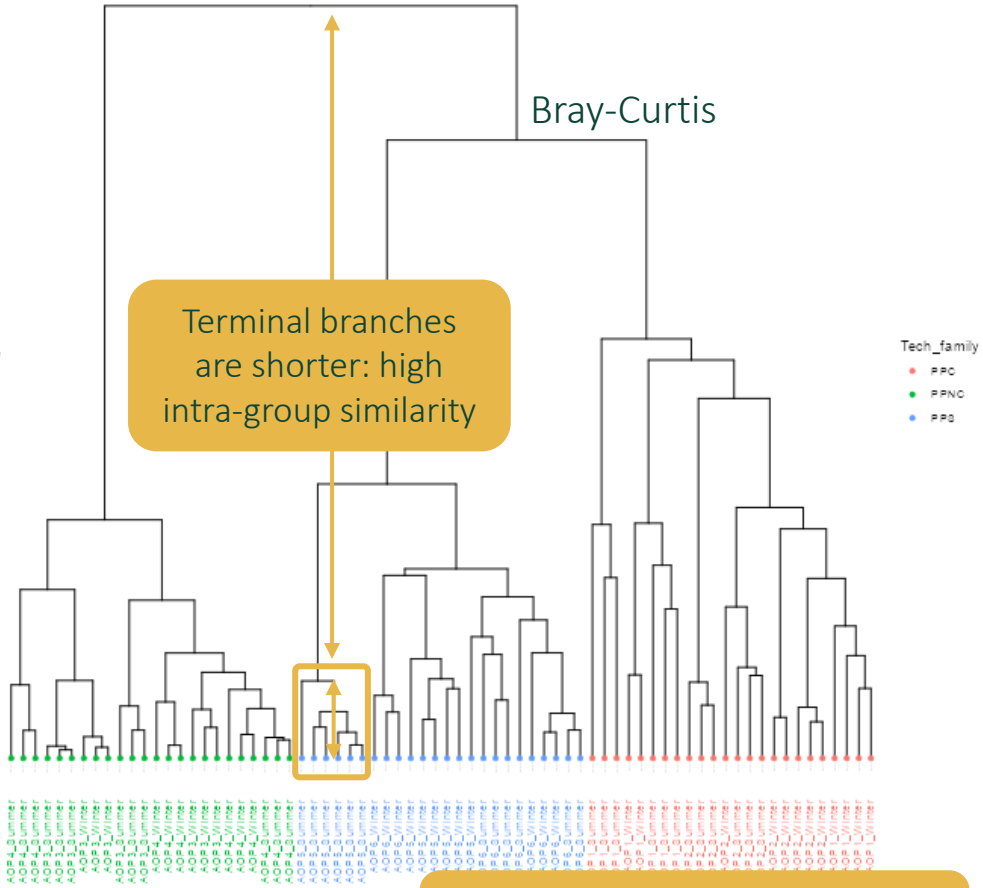
# Practice session

## Explore β-diversity structure with hierarchical clustering

What biological interpretation can be extracted?




Long terminal branches: high intra-group heterogeneity




Terminal branches are shorter: high intra-group similarity

AOP5 Summer samples are separated from AOP5/6

Samples are well-separated by tech family

# Explore β-diversity structure with hierarchical clustering

What biological interpretation can be extracted?

AOP has a strong effect on the structure of bacteria data (stronger than tech family)

Ultimately samples are generally well-grouped by replicates
(same AOP, same production unit, same season)

Intragroup differences are higher with bacteria than with fungi (especially with quantitative metrics)

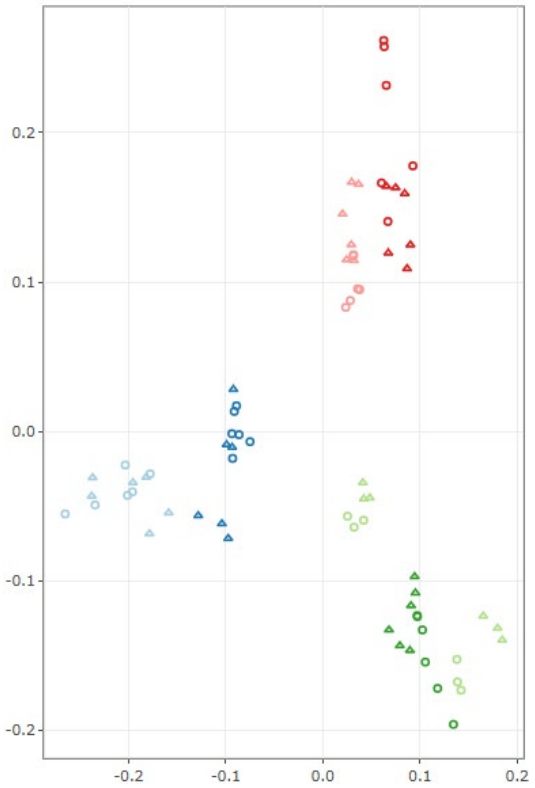
Fungal data are well-separated by technical families, then AOP, then sometimes seasons, with strong differences between groups (stronger than intragroup differences)

# How communities are structured based on β-diversity?

Are there better solutions than eye-observations on heatmaps?

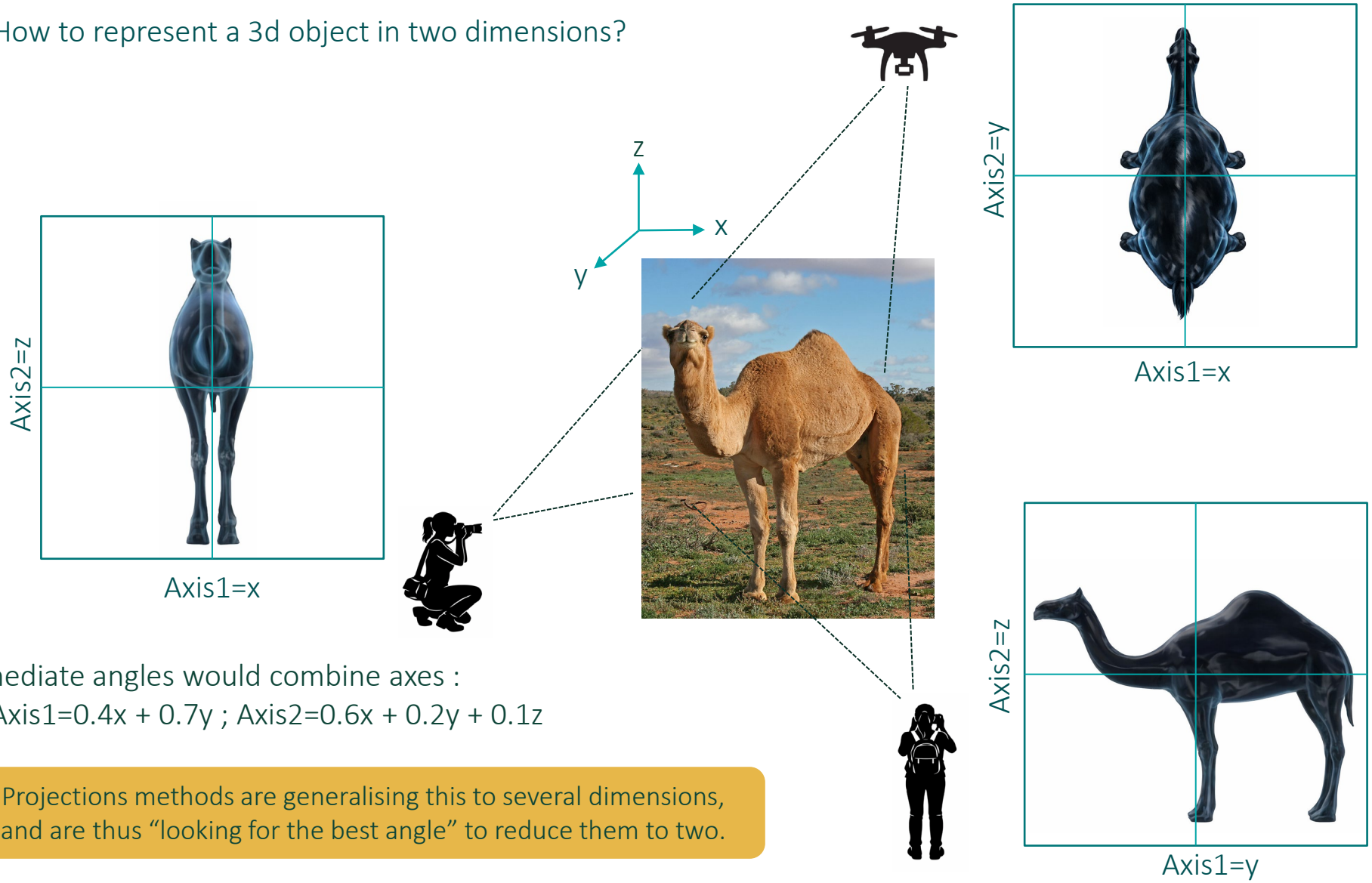
# Exploring structures: Multidimensional Scaling

Another way of visualising data with high dimensions (data with a large number of variables) is to use “projection methods” to reduce the data and represent them on only two dimensions (two variables).



# Dimensionality-reduction methods: principle

How to represent a 3d object in two dimensions?




Intermediate angles would combine axes :

Axis1=0.4x + 0.7y ; Axis2=0.6x + 0.2y + 0.1z

Projections methods are generalising this to several dimensions, and are thus “looking for the best angle” to reduce them to two.

# Dimensionality-reduction methods: PCA

There are many projection methods, which differ in how they define “the optimal viewing angle”, e.g. the combination of variable (“axes”) to used in the resulting component (axis1 & axis2).

**PCA** (Principle Component Analysis, in French ACP Analyse en Composantes Principales) is the most usual one. It assumes Euclidean distances between samples and choose its angle **by maximising variance** (maximising the spreading of the points).

This is good strategy in most of the cases but β-diversity, as PCA would focus on differences between rare ASVs and distort ecological dissimilarities.

Moreover, PCA requires independency between variables (that should not be linked), and this is not the case with β-diversity distance matrix.

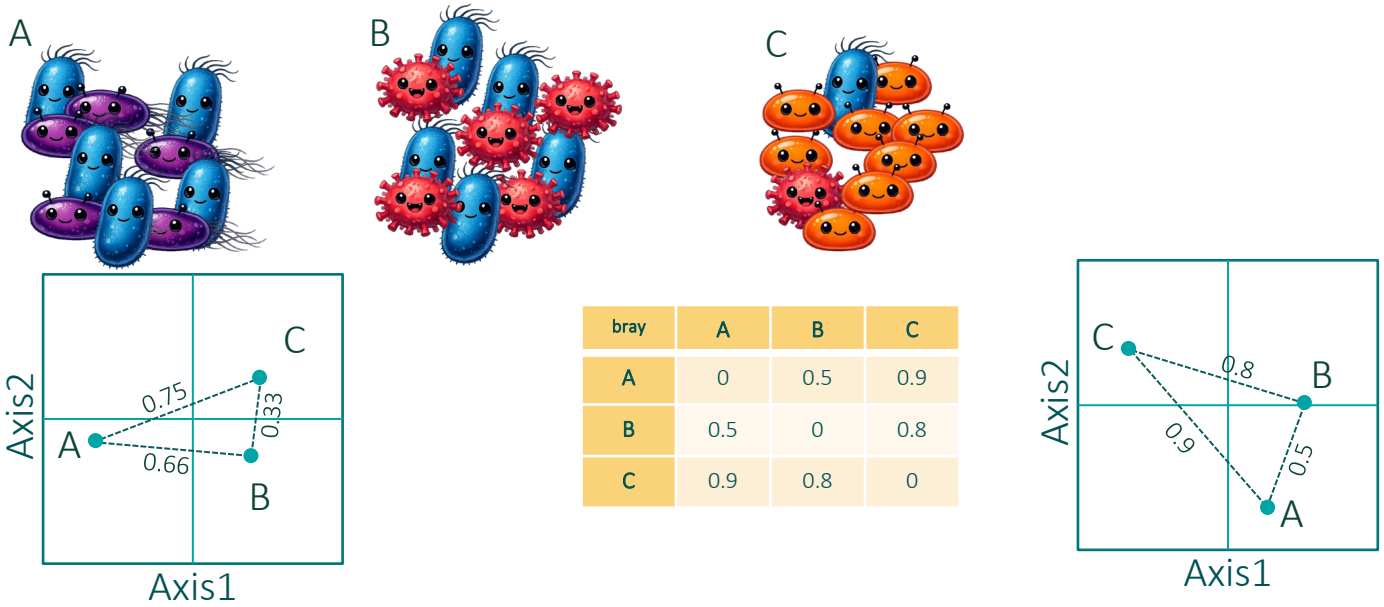
The projection of high-dimensional relationships into a smaller space cannot preserve all the original structure, and thus projection methods lose information.

# Dimensionality-reduction methods: MDS/PCoA

Whereas PCA maximise variance, **MDS** (Multi-Dimensional Scaling, also named **PCoA** Principal Coordinate Analysis) expect a distance matrix as input and find a projection angle such that

**the distances between points match the original multi-dimension distances as closely as possible.**

*That means that if two samples were close in the distance matrix, they end close on the plot (respectively distant).*




| jaccard | A | B | C |
|---|---|---|---|
| A | 0 | 0.66 | 0.75 |
| B | 0.66 | 0 | 0.33 |
| C | 0.75 | 0.33 | 0 |

| bray | A | B | C |
|---|---|---|---|
| A | 0 | 0.5 | 0.9 |
| B | 0.5 | 0 | 0.8 |
| C | 0.9 | 0.8 | 0 |

Several other methods exist with their specificities.
In particular, **NMDS** is similar to MDS but only consider the rank (higher distance, 2nd one, 3rd one etc.) and not the absolute values of distances.

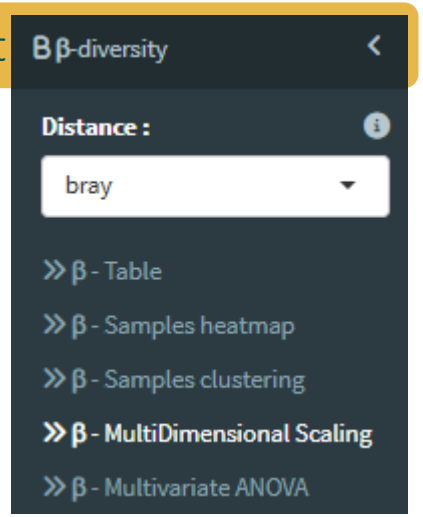
# Practice session

## Explore β-diversity structure with Multi-Dimensional Scaling

How to read the output?

What biological interpretation can be extracted?

In Easy16S select



# Practice session

## Explore β-diversity structure with Multi-Dimensional Scaling

How to read the output?

swarm

Settings :

**Ordination method :** MDS / PCoA (Principal Coordinate Analysis)

**Axis 1** 1

**Axis 2** 2

**Title :**

**Samples color :** AOP

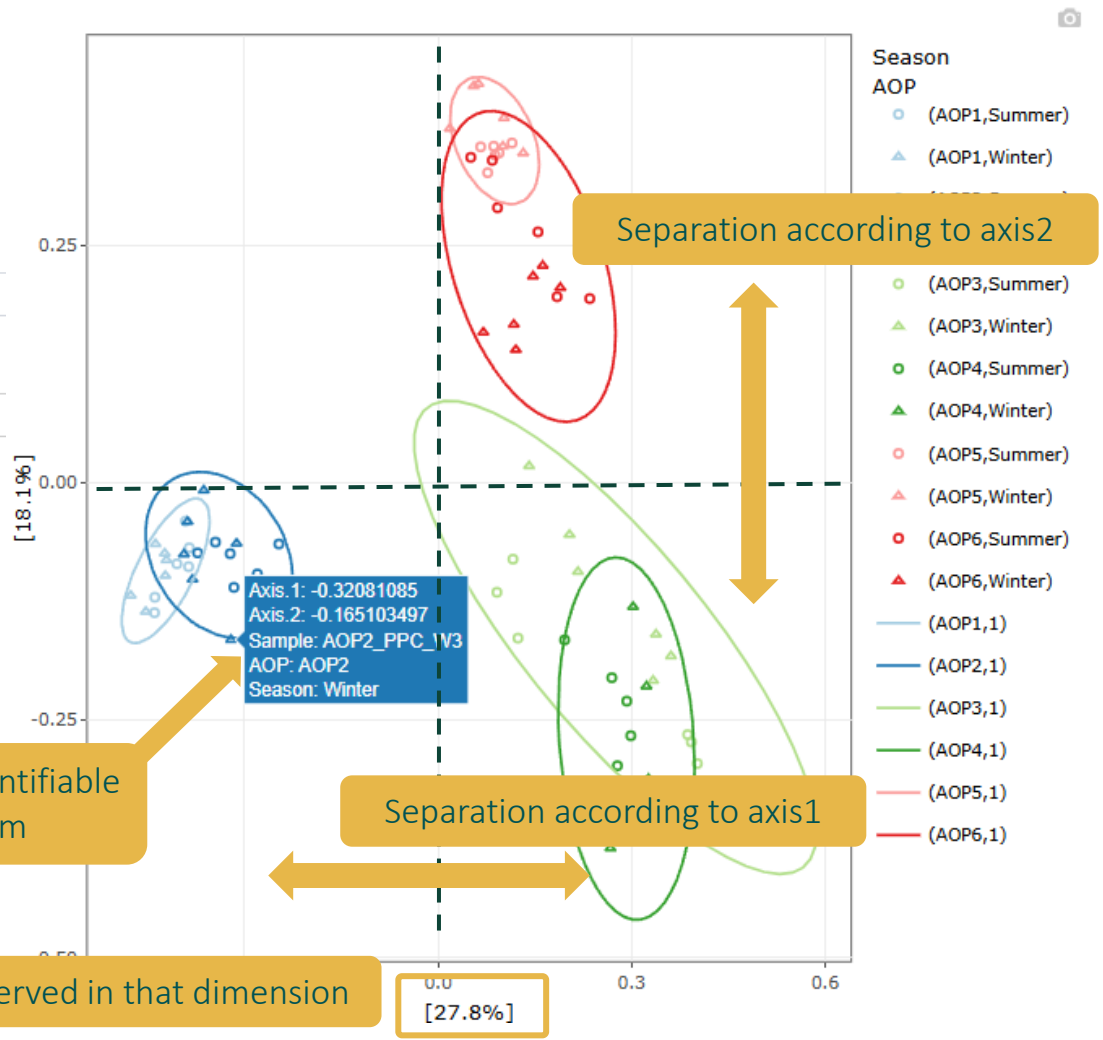
Add ellipse

**Samples shape (ignored if label selected) :** Season

**Samples label :** ...

variable to add shapes

variable to color samples (and draw ellipse if selected)




Separation according to axis2

Dots are samples, identifiable by pointing them

Separation according to axis1

% of the original distances structure preserved in that dimension

# Practice session

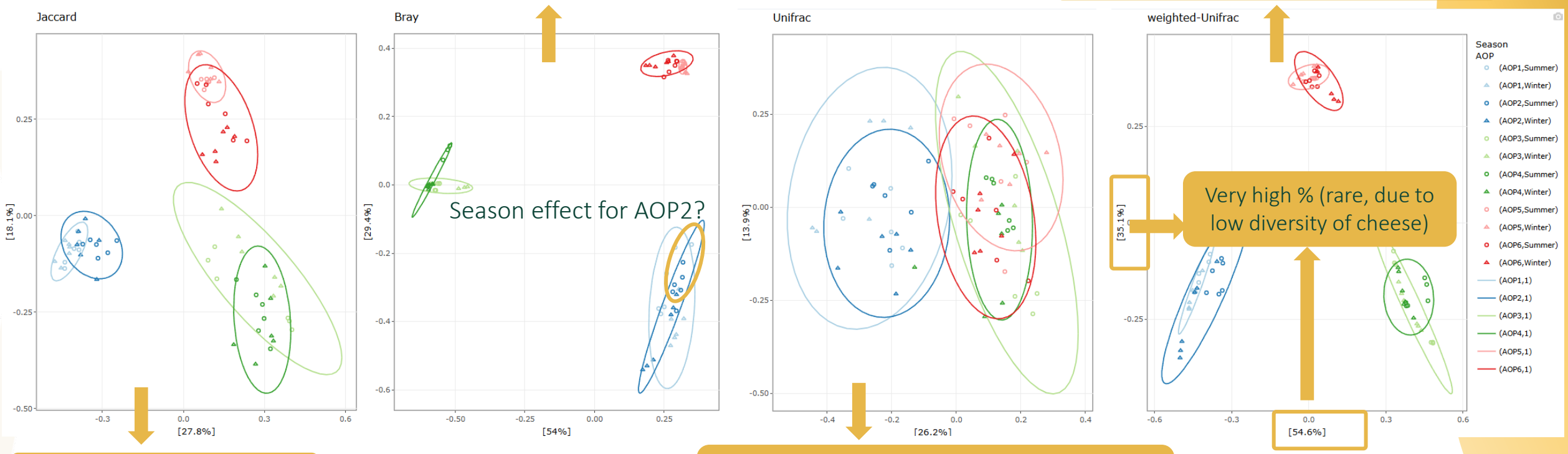
## Explore β-diversity structure with Multi-Dimensional Scaling

What biological interpretation can be extracted?

swarm

Stronger separation with low intragroup differences

A lot of heterogeneity (warning, could be due to bad phylogenetic tree construction)




Season effect for AOP2?

Very high % (rare, due to low diversity of cheese)

Strong tech family effects

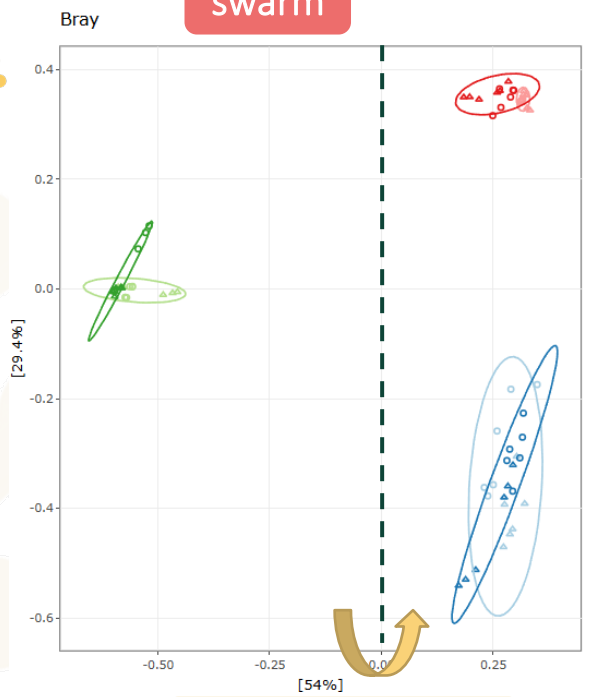
A lot of heterogeneity (warning, could be due to bad phylogenetic ITS tree construction!)

# Practice session

## Explore β-diversity structure with Multi-Dimensional Scaling

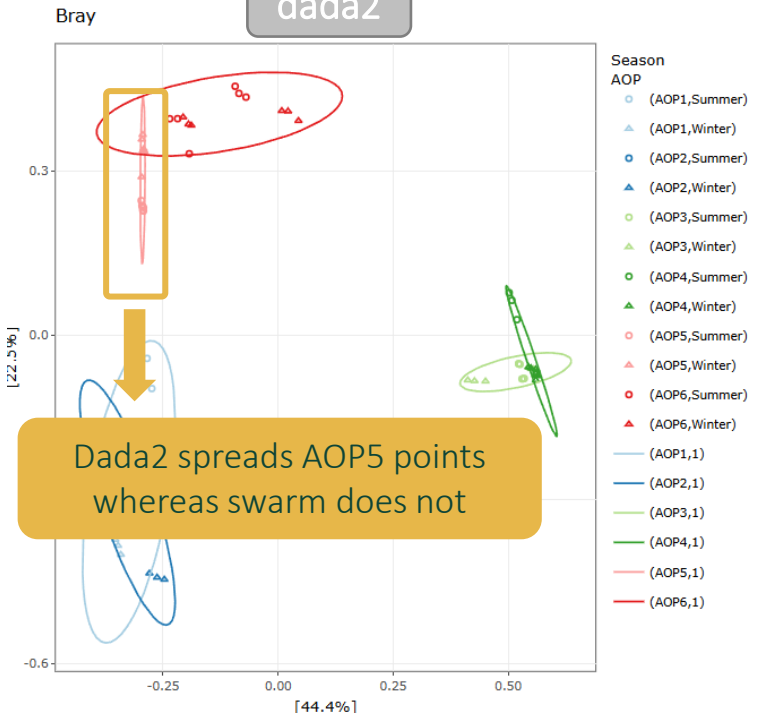
What biological interpretation can be extracted?

swarm



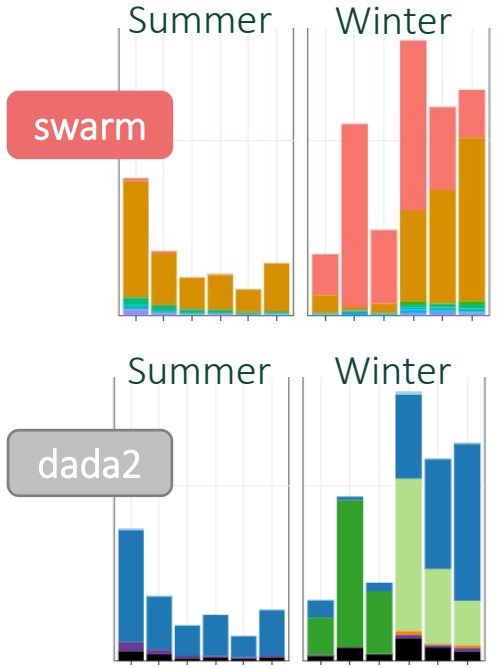

Sign (+/-) is random and not significant

dada2




Dada2 spreads AOP5 points whereas swarm does not

Remember AOP5 Penicillium composition:



# Practice session

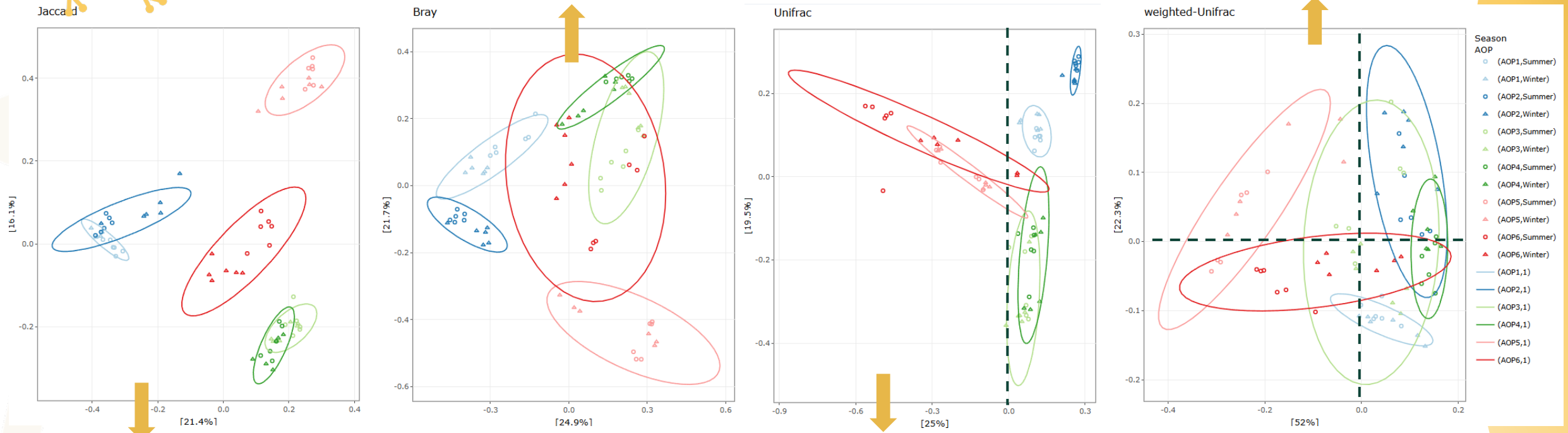
## Explore β-diversity structure with Multi-Dimensional Scaling

What biological interpretation can be extracted?

swarm

Less clear separation considering abundances (abundant ASV are more shared than rare ones)

Unclear separations and effects when integrating abundances and phylogeny




Strong tech family effects, strong AOP effect for PPS

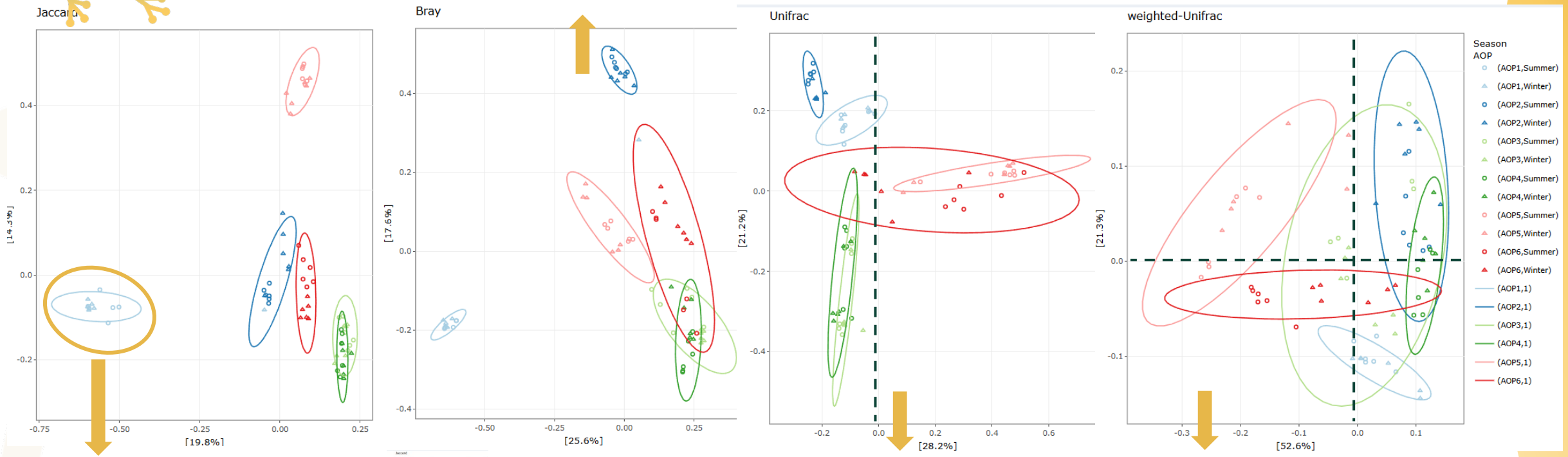
Considering phylogeny, clear separation between PPS and PPC/PPNC
Axis2 separates PPS AOPs and AOP6 seasons

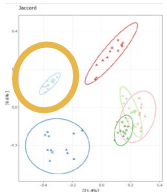
# Practice session

## Explore β-diversity structure with Multi-Dimensional Scaling

What biological interpretation can be extracted?

Clearer AOP separation than with swarm



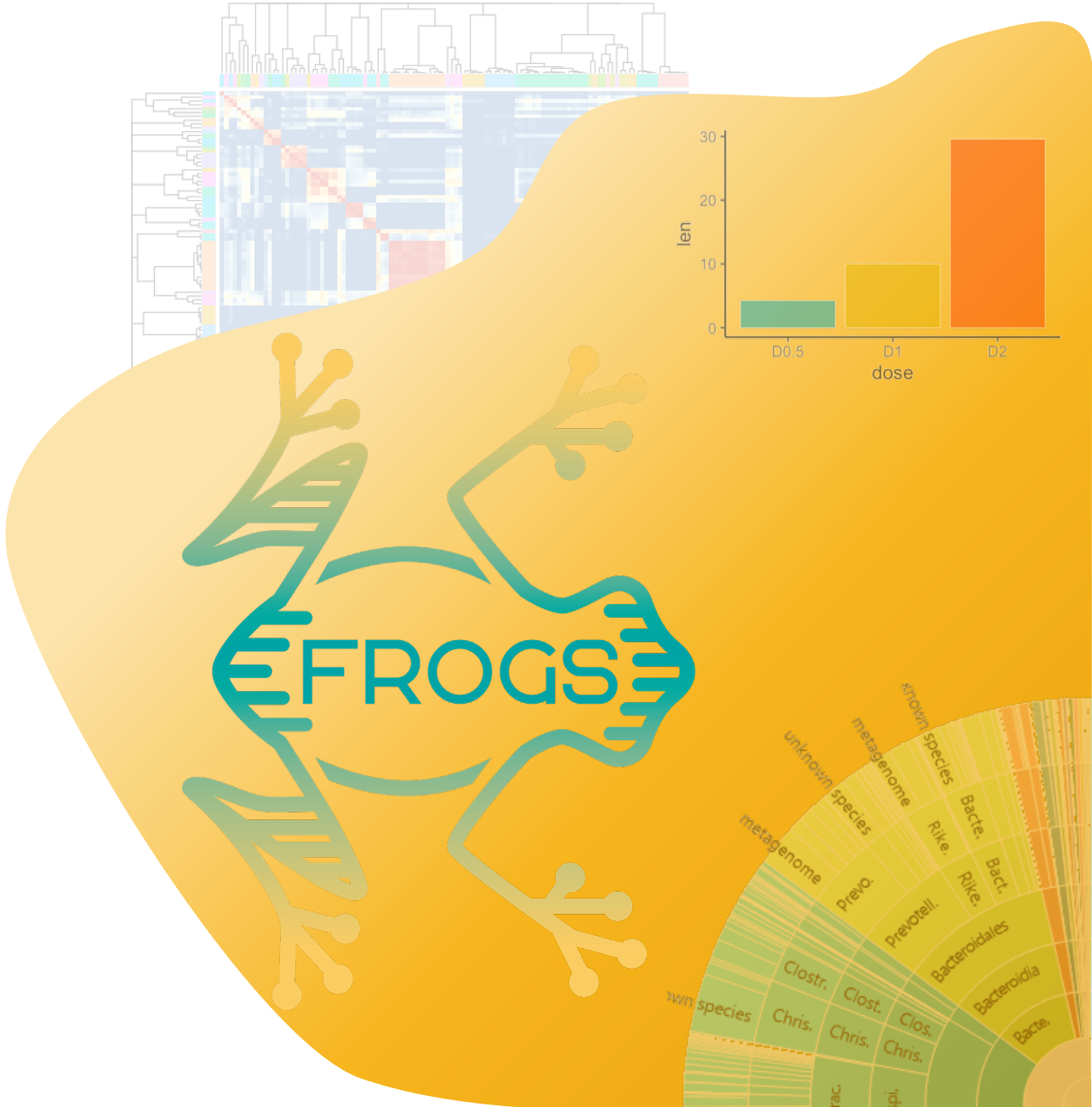
Outgroup not visible with swarm (observable only on axis3)

Considering phylogeny, swarm and dada2 have more similar results (*almost identical with weighted-Unifrac*)

# FROGS Stats

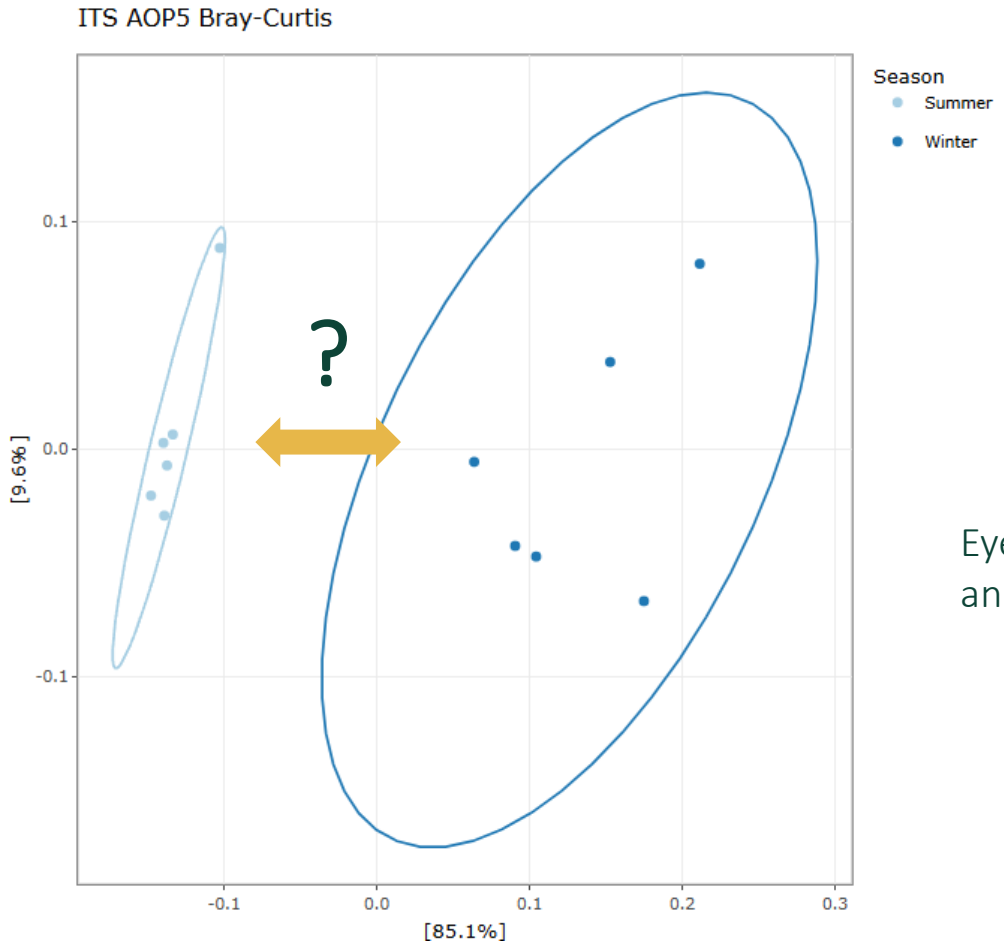
## β-diversity

## Partitioning



# Diversity partitioning

How to measure these effects and there significancy?




Eyes-observations of groups are not measurable and not statistically testable…

# PERMANOVA principle

PERMANOVA tests whether the distances between groups are larger than the distances within groups.

If it is the case, samples of a same group are similar, and more similar than compared to other groups.
This implies that the tested factor (which defines the tested groups) affects the composition.

To answer this question, PERMANOVA:

- Computes the average distance between, and within, groups
- Randomly shuffles the group labels a LOT of times (permutate samples group identity)
- Measures the impact of these permutation on the inter/intra- groups ratio

If the real grouping produces much stronger separation than random permutations, **the effect of the tested factor is significant**.

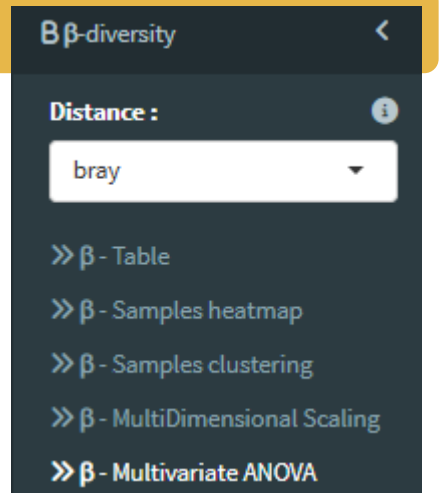
# Practice session

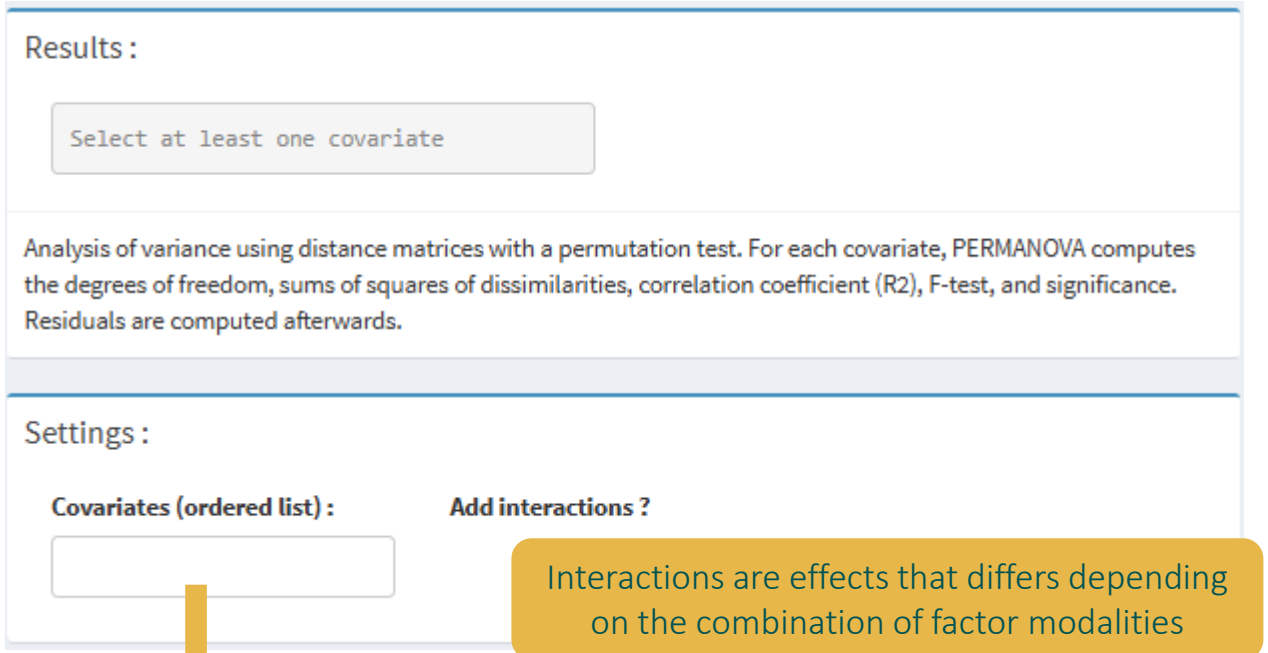
## Evaluate and test metadata factors effects on β-diversity indices

How to read the output?

What biological interpretation can be extracted?

In Easy16S select




Interactions are effects that differs depending on the combination of factor modalities

For example, Season could affect differently each AOP so interaction between seasons and AOP could be tested

Add one ore more variables that will be sequentially tested together

# Practice session

## Evaluate and test metadata factors effects on β-diversity indices

How to read the output?

Results :

Tested distance

Tested factors

```
vegan::adonis2(formula = bray_dist ~ Tech_family + AOP + Season + AOP:Season,
data = sample_data,
by = 'terms',
perm = 9999)

# Permutation test for adonis under reduced model
# Terms added sequentially (first to last)
# Permutation: free
# Number of permutations: 9999

# Signif. codes:  0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| term | df | SumOfSqs | R2 | statistic | p.value | p.value.signif |
|---|---|---|---|---|---|---|
| Tech_family | 2.00 | 17.37 | 0.82 | 273.18 | 0.00 | **** |
| AOP | 3.00 | 1.06 | 0.05 | 11.07 | 0.00 | **** |
| Season | 1.00 | 0.13 | 0.01 | 4.19 | 0.01 | * |
| AOP:Season | 5.00 | 0.82 | 0.04 | 5.15 | 0.00 | **** |
| Residual | 60.00 | 1.91 | 0.09 | NA | NA | |
| Total | 71.00 | 21.29 | 1.00 | NA | NA | |

Analysis of variance using distance matrices with a permutation test. For each covariate, PERMANOVA computes the degrees of freedom, sum of squares of dissimilarities, correlation coefficient (R2), F-test, and significance. Residuals are computed afterwards.

When p-value < 0.05 the factor has a significant effect on the distance

$R^2$ represents the proportion of the overall distance among samples that can be explained by the tested factor

# Practice session

## Evaluate and test metadata factors effects on β-diversity indices

What biological interpretation can be extracted?

swarm

dada2

### Jaccard

| term | R2 | statistic | p.value | p.value.signif |
|---|---|---|---|---|
| Tech_family | 0.33 | 27.67 | 0.00 | **** |
| AOP | 0.22 | 12.21 | 0.00 | **** |
| Season | 0.02 | 3.06 | 0.00 | ** |
| AOP:Season | 0.09 | 2.92 | 0.00 | **** |
| Residual | 0.35 | NA | NA | |
| Total | 1.00 | NA | NA | |

| term | R2 | statistic | p.value | p.value.signif |
|---|---|---|---|---|
| Tech_family | 0.28 | 23.70 | 0.00 | **** |
| AOP | 0.28 | 15.81 | 0.00 | **** |
| Season | 0.02 | 2.67 | 0.00 | ** |
| AOP:Season | 0.08 | 2.71 | 0.00 | **** |
| Residual | 0.35 | NA | NA | |
| Total | 1.00 | NA | NA | |

Same p-values. Small $R^2$ differences

### Bray-Curtis

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.36 | 37.69 | 0.00 | **** |
| 0.25 | 17.72 | 0.00 | **** |
| 0.02 | 3.32 | 0.00 | ** |
| 0.10 | 4.05 | 0.00 | **** |
| 0.28 | NA | NA | |
| 1.00 | NA | NA | |

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.31 | 33.68 | 0.00 | **** |
| 0.31 | 22.28 | 0.00 | **** |
| 0.02 | 3.29 | 0.00 | ** |
| 0.08 | 3.50 | 0.00 | **** |
| 0.28 | NA | NA | |
| 1.00 | NA | NA | |

### Unifrac

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.37 | 42.50 | 0.00 | **** |
| 0.28 | 21.14 | 0.00 | **** |
| 0.01 | 3.35 | 0.00 | ** |
| 0.07 | 3.05 | 0.00 | **** |
| 0.26 | NA | NA | |
| 1.00 | NA | NA | |

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.41 | 48.91 | 0.00 | **** |
| 0.24 | 19.20 | 0.00 | **** |
| 0.02 | 4.62 | 0.00 | *** |
| 0.08 | 4.09 | 0.00 | **** |
| 0.25 | NA | NA | |
| 1.00 | NA | NA | |

### weighted-Unifrac

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.41 | 53.69 | 0.00 | **** |
| 0.24 | 20.92 | 0.00 | **** |
| 0.02 | 4.59 | 0.01 | ** |
| 0.10 | 4.98 | 0.00 | **** |
| 0.23 | NA | NA | |
| 1.00 | NA | NA | |

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.42 | 54.54 | 0.00 | **** |
| 0.24 | 21.28 | 0.00 | **** |
| 0.02 | 5.09 | 0.00 | ** |
| 0.09 | 4.88 | 0.00 | **** |
| 0.23 | NA | NA | |
| 1.00 | NA | NA | |

Same $R^2$ when integrating phylogeny and abundances

# Practice session

## Evaluate and test metadata factors effects on β-diversity indices

What biological interpretation can be extracted?

swarm

### Jaccard

| term | R2 | statistic | p.value | p.value.signif |
|---|---|---|---|---|
| Tech_family | 0.33 | 27.67 | 0.00 | **** |
| AOP | 0.22 | 12.21 | 0.00 | **** |
| Season | 0.02 | 3.06 | 0.00 | ** |
| AOP:Season | 0.09 | 2. | 0.00 | **** |
| Residual | 0.[illegible]5 | NA | | |
| Total | 1.[illegible]0 | NA | | |

### Bray-Curtis

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.36 | 37.69 | 0.00 | **** |
| 0.25 | 17.72 | 0.00 | **** |
| 0.02 | 3.32 | 0.00 | ** |
| 0.10 | 4.05 | 0.00 | **** |
| | | NA | |
| | | NA | |

33% of Jaccard distance is related with Tech_family, 22% with AOP

### Unifrac

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.37 | 42.50 | 0.00 | **** |
| 0.28 | 21.14 | 0.00 | **** |
| 0.01 | 3.35 | 0.00 | ** |
| 0.07 | 3.05 | 0.00 | **** |
| 0.26 | NA | NA | |
| 1.00 | NA | NA | |

### weighted-Unifrac

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| 0.41 | 53.69 | 0.00 | **** |
| 0.24 | 20.92 | 0.00 | **** |
| 0.02 | 4.59 | 0.01 | ** |
| 0.10 | 4.98 | 0.00 | **** |
| 0.23 | NA | NA | |
| | | NA | |

Balance between Tech and AOP is variable with the tested distance ; Season vs AOP:Season is more stable

dada2

| term | statistic | p.value | p.value.signif |
|---|---|---|---|
| Tech_family | | | |
| AOP | | | |
| Season | | | |
| AOP:Season | 0.08 | 2.71 | 0.00 | **** |
| Residual | 0.35 | NA | NA | |
| Total | 1.00 | NA | NA | |

| R2 | statistic | p.value | p.value.signif |
|---|---|---|---|
| | | | |
| | | | |
| | | | |
| 0.08 | 3.50 | 0.00 | **** |
| 0.28 | NA | NA | |
| 1.00 | NA | NA | |

Season has on small effect (2%) but 9% of Jaccard distance is related to seasons effects **specific to the different AOP**

| R2 | statistic | p.value | |
|---|---|---|---|
| 0.41 | 48.91 | 0.00 | |
| 0.24 | 19.20 | 0.00 | **** |
| 0.02 | 4.62 | 0.00 | *** |
| 0.08 | 4.09 | 0.00 | **** |
| 0.25 | NA | NA | |
| 1.00 | NA | NA | |

| | | | |
|---|---|---|---|
| | | | |
| 0.24 | 21.28 | 0.00 | **** |
| 0.02 | 5.09 | 0.00 | ** |
| 0.09 | 4.88 | 0.00 | **** |
| 0.23 | NA | NA | |
| 1.00 | NA | NA | |

# Practice session

## Evaluate and test metadata factors effects on β-diversity indices

What biological interpretation can be extracted?

With non-independent factors, the order strongly matters!

```
vegan::adonis2(formula = bray_dist ~ AOP + Tech_family + Season + AOP:Season,
data = sample_data,
by = 'terms',
perm = 9999)

# Permutation test for adonis under reduced model
# Terms added sequentially (first to last)
# Permutation: free
# Number of permutations: 9999

# Signif. codes:  0 '****' 1e-04 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

| term | df | SumOfSqs | R2 | statistic | p.value | p.value.signif |
|---|---|---|---|---|---|---|
| AOP | 5.00 | 15.09 | 0.62 | 26.84 | 0.00 | **** |
| Season | 1.00 | 0.37 | 0.02 | 3.29 | 0.00 | ** |
| AOP:Season | 5.00 | 1.97 | 0.08 | 3.50 | 0.00 | **** |
| Residual | 60.00 | 6.75 | 0.28 | NA | NA | |
| Total | 71.00 | 24.17 | 1.00 | NA | NA | |

| term | R2 | statistic | p.value | p.value.signif |
|---|---|---|---|---|
| Tech_family | 0.31 | 33.68 | 0.00 | **** |
| AOP | 0.31 | 22.28 | 0.00 | **** |
| Season | 0.02 | 3.29 | 0.00 | ** |
| AOP:Season | 0.08 | 3.50 | 0.00 | **** |
| Residual | 0.28 | NA | NA | |
| Total | 1.00 | NA | NA | |

When AOP is set first, Tech_Family has no effect… because there are no different Tech families inside a given AOP.
AOP effect becomes the sum of Tech_family and AOP effects

# Practice session

## Evaluate and test metadata factors effects on β-diversity indices

What biological interpretation can be extracted?

ITS swarm

| term | Jaccard R2 | Jaccard statistic | Jaccard p.value | Jaccard p.value.signif | Bray-Curtis R2 | Bray-Curtis statistic | Bray-Curtis p.value | Bray-Curtis p.value.signif | Unifrac R2 | Unifrac statistic | Unifrac p.value | Unifrac p.value.signif | weighted-Unifrac R2 | weighted-Unifrac statistic | weighted-Unifrac p.value | weighted-Unifrac p.value.signif |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tech_family | 0.43 | 38.53 | 0.00 | **** | 0.82 | 273.18 | 0.00 | **** | 0.28 | 18.51 | 0.00 | **** | 0.87 | 415.75 | 0.00 | **** |
| AOP | 0.13 | 7.94 | 0.00 | **** | 0.05 | 11.07 | 0.00 | **** | 0.14 | 6.29 | 0.00 | **** | 0.03 | 9.69 | 0.00 | **** |
| Season | 0.02 | 4.33 | 0.00 | *** | 0.01 | 4.19 | 0.01 | * | 0.04 | 4.66 | 0.00 | *** | 0.00 | 2.11 | 0.11 | |
| AOP:Season | 0.08 | 2.95 | 0.00 | **** | 0.04 | 5.15 | 0.00 | **** | 0.09 | 2.31 | 0.00 | *** | 0.03 | 6.55 | 0.00 | **** |
| Residual | 0.33 | NA | NA | | 0.09 | NA | NA | | 0.45 | NA | NA | | 0.06 | NA | NA | |
| Total | 1.00 | NA | NA | | 1.00 | | | | | | | | 1.00 | NA | NA | |

Strong Tech_family effects with abundances

ITS dada2

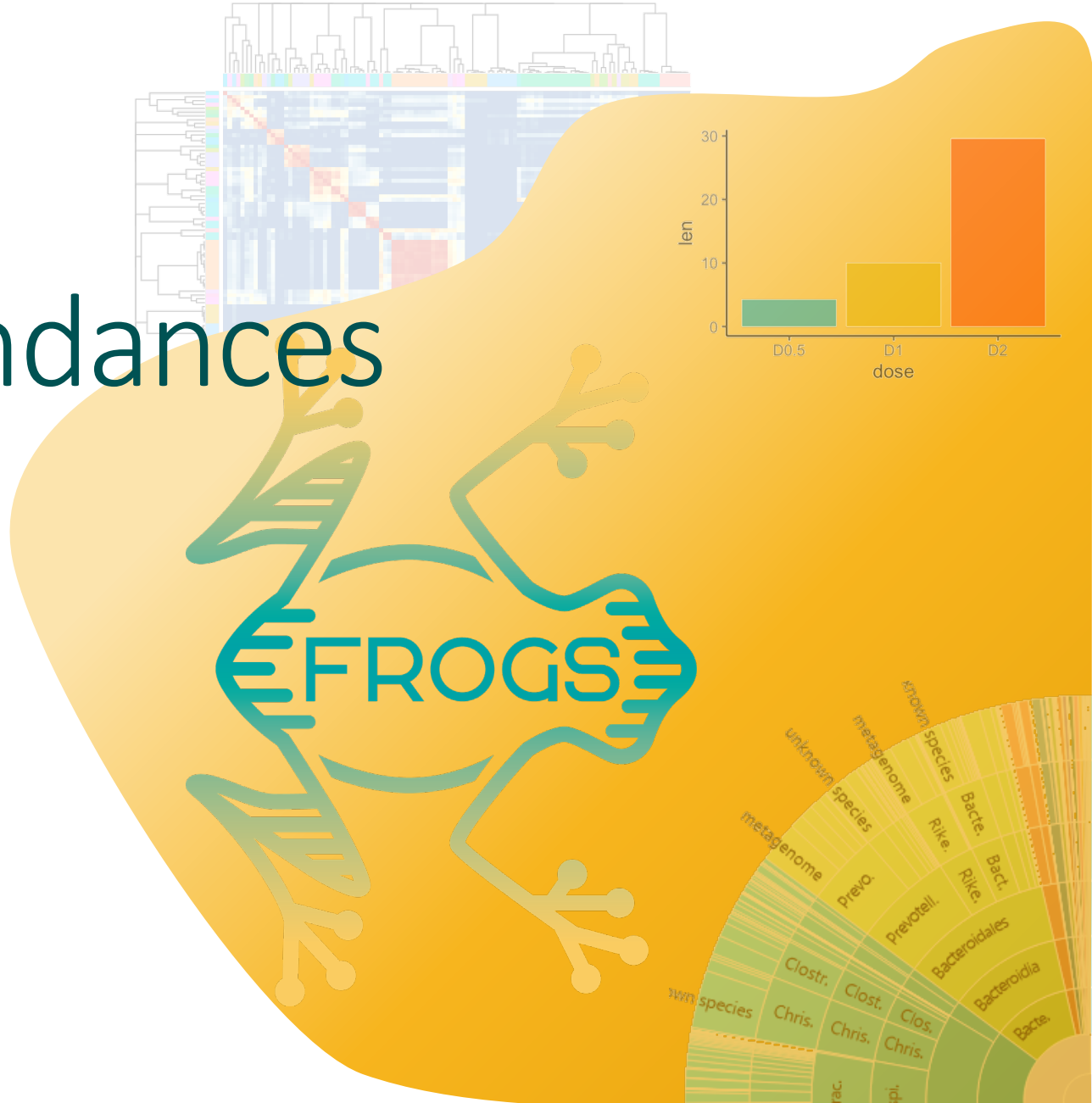
| term | Jaccard R2 | Jaccard statistic | Jaccard p.value | Jaccard p.value.signif | Bray-Curtis R2 | Bray-Curtis statistic | Bray-Curtis p.value | Bray-Curtis p.value.signif | Unifrac R2 | Unifrac statistic | Unifrac p.value | Unifrac p.value.signif | weighted-Unifrac R2 | weighted-Unifrac statistic | weighted-Unifrac p.value | weighted-Unifrac p.value.signif |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tech_family | 0.39 | 33.08 | 0.00 | **** | 0.64 | 151.21 | 0.00 | **** | 0.32 | 24.00 | 0.00 | **** | 0.87 | 463.75 | 0.00 | **** |
| AOP | 0.15 | 8.34 | 0.00 | **** | 0.18 | 27.94 | 0.00 | **** | 0.15 | 7.56 | 0.00 | **** | 0.04 | 12.97 | 0.00 | **** |
| Season | 0.02 | 3.47 | 0.00 | ** | 0.01 | 4.21 | 0.00 | ** | 0.03 | 4.40 | 0.00 | *** | 0.00 | 3.06 | 0.04 | * |
| AOP:Season | 0.09 | 3.16 | 0.00 | **** | 0.04 | 3.91 | 0.00 | **** | 0.10 | 2.86 | 0.00 | **** | 0.03 | 7.19 | 0.00 | **** |
| Residual | 0.35 | NA | NA | | 0.13 | NA | NA | | 0.40 | NA | NA | | 0.06 | NA | NA | |
| Total | 1.00 | NA | NA | | 1.00 | NA | NA | | 1.00 | NA | NA | | 1.00 | NA | NA | |

Stronger Tech families effects on fungi than on Bacteria (opposite with AOP effect).
Similar very low global seasonal effects, but stronger AOP:Season effects.

Same $R^2$ when integrating phylogeny and abundances

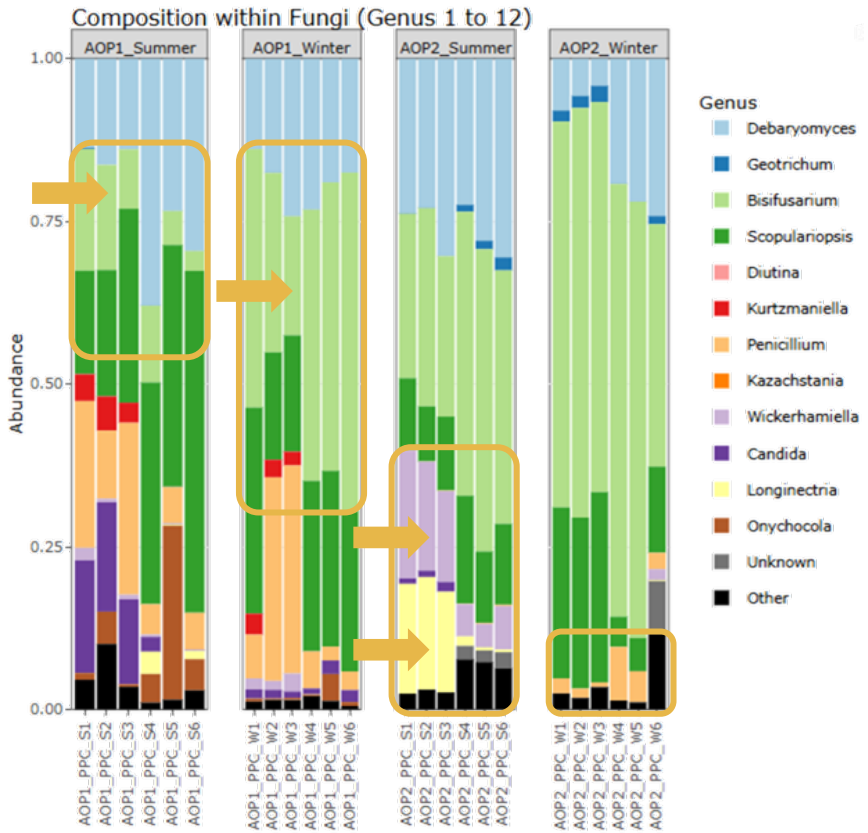
# FROGS Stats

# Differential abundances analyses



# Differential abundances analyses

Are there ASV with differential abundance between 2 conditions?




Eyes-observations of groups are not measurable and not statistically testable…

# DESeq2 principle

Several methods exist to identify ASV with significantly different abundances between conditions, including DESeq2, which is integrated in FROGS.

DESeq2 :

- Calculates size factors that correct for differences in sequencing depth,
  *so it does not require normalised/rarefied abundances* (it is even recommended to use raw data)

- Estimates dispersion for each taxon (across replicates)

- Use a “generalised linear model” to transform data and test whether normalised counts differ between conditions

It has strong hypothesis and related limitations:

- “Sparse” tables with a lot of 0 values prevent DESeq to correctly estimate dispersion
- Highly unbalanced groups (different number of samples per conditions) / low numbers of replicates can provide unreliable results
- DESeq2 is sensitive to compositionality (the normalisation model does not correct it)

**DESeq2 should only be used when comparing samples with similar community compositions, where most taxa are shared and only a small subset is expected to differ in abundance**

# Practice session

## Identify abundance differences between conditions with DESeq2

How to read the output?

What biological interpretation can be extracted?

In Easy16S select Differential abundance


Choose a factor

Compute!

All factors are not pre-calculated
(as it is quite heavy computation)

# Practice session

## Identify abundance differences between conditions with DESeq2

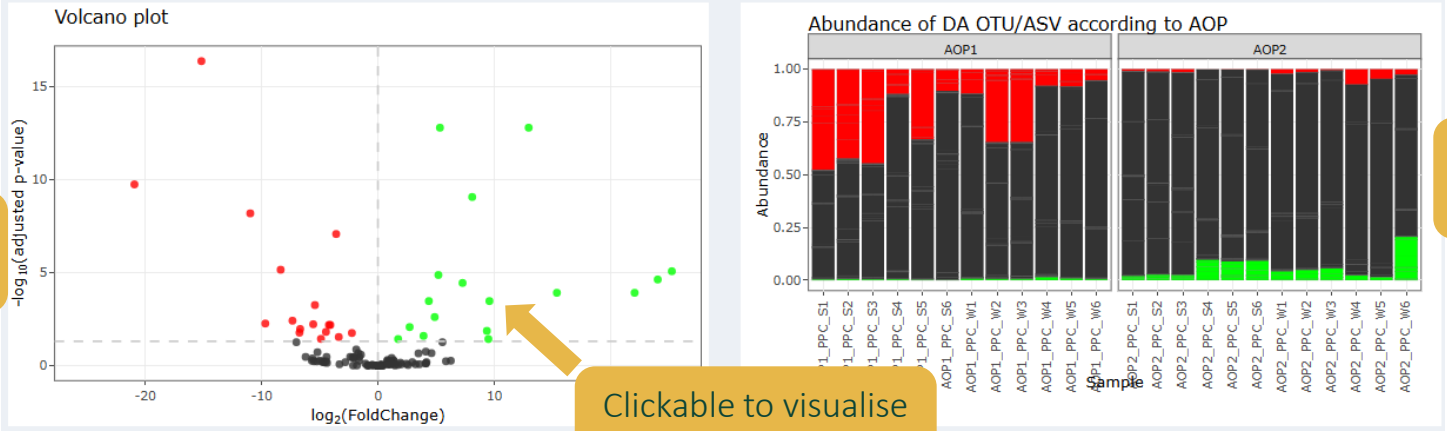
How to read the output?

Contrast left : AOP1

Contrast right : AOP2

Modalities are compared by pairs only

"Volcano plot" (log2FC x pvalue)







Corresponding raw abundances

Clickable to visualise specific abundances

You compare AOP2 and AOP1 of the binary variable AOP.
A positive log2FoldChange means more abundant in AOP2 than in AOP1.

volcano plot"

Table of OTUs/ASVs with significant effect (padj <= 0.05)

CSV | Copy | Excel | Search:

| OTU | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Kingdom | Phylum | Class | Order | Family |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ID_203 | 1.421 | -4.897 | 1.849 | -2.648 | 0.008089 | 0.03716 | Fungi | Ascomycota | Sordariomycetes | Microascales | Microascacea |
| ID_107 | 6.997 | -3.367 | 1.230 | -2.738 | 0.006189 | 0.02935 | Fungi | Ascomycota | Sordariomycetes | Microascales | Microascacea |
| ID_58 | 7.158 | -20.93 | 3.028 | -6.910 | 4.833e-12 | 1.776e-10 | Fungi | Ascomycota | Sordariomycetes | Hypocreales | |
| ID_87 | 2.874 | -6.685 | 2.147 | -3.113 | 0.001851 | 0.01088 | Fungi | Ascomycota | Sordariomycetes | Hypocreales | |
| ID_62 | 7.105 | 15.38 | 3.508 | 4.385 | 0.00001162 | 0.0001225 | Fungi | Ascomycota | Sordariomycetes | Hypocreales | |

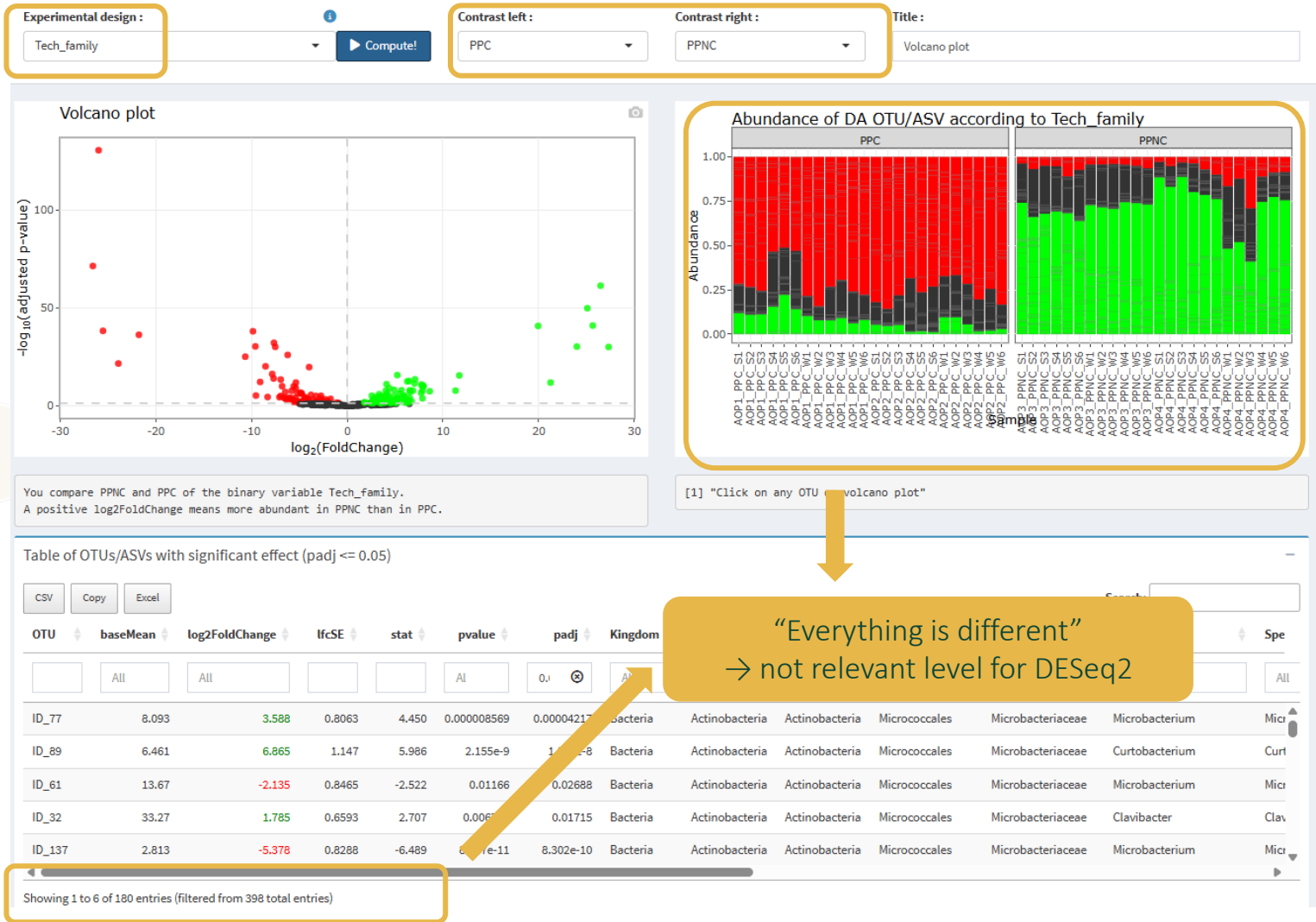
Showing 1 to 6 of 34 entries (filtered from 148 total entries)

Details on all the differentially abundant ASV

# Practice session

## Identify abundance differences between conditions with DESeq2

What biological interpretation can be extracted?

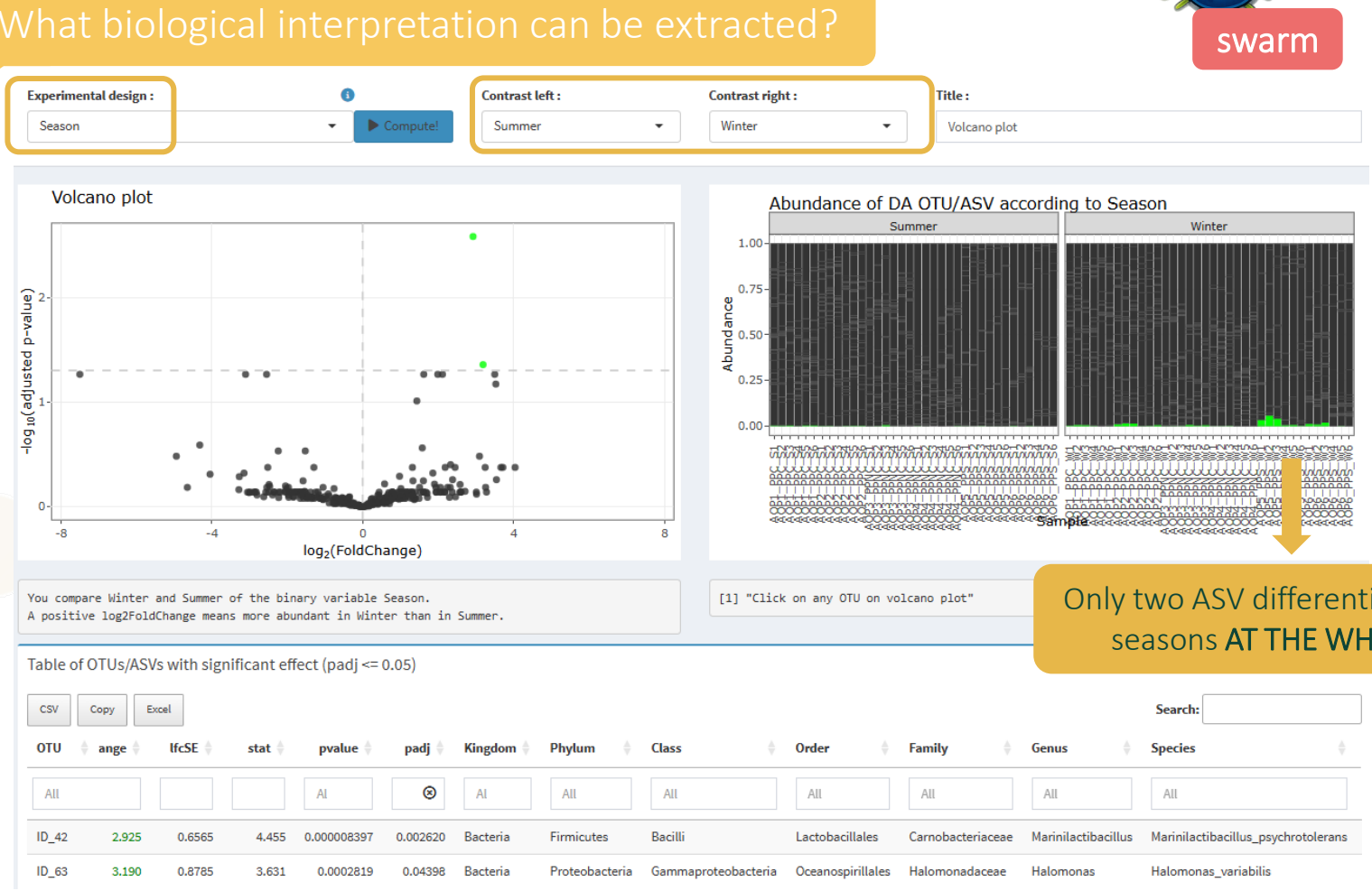



"Everything is different"
→ not relevant level for DESeq2

# Practice session

## Identify abundance differences between conditions with DESeq2

What biological interpretation can be extracted?




Only two ASV differentially abundant between seasons **AT THE WHOLE DATASET SCALE**

ASV composition is strongly AOP-dependant

Are higher taxonomic level more comparable?

Preprocess data

# Practice session

## Identify abundance differences between conditions with DESeq2

RAW DATA
(not subsampled)

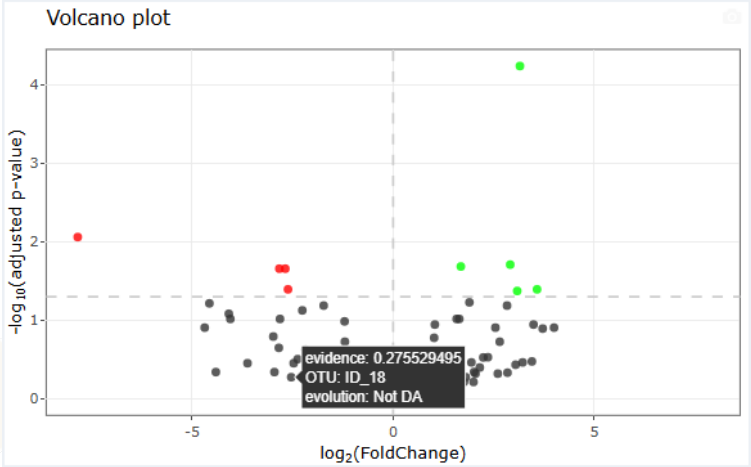
swarm

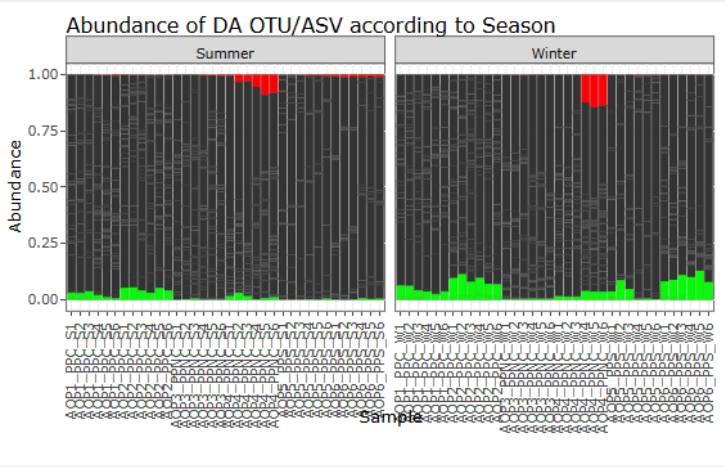
What biological interpretation can be extracted?

Experimental design : Season — Compute!

Contrast left : Summer — Contrast right : Winter

Title : Volcano plot




| OTU | pvalue | padj | Kingdom | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|---|---|---|
| ID_10 | 0.0004529 | 0.02061 | Bacteria | Actinobacteria | Actinobacteria | Corynebacteriales | Corynebacteriaceae | Corynebacterium | Corynebacterium_casei |
| ID_42 | 3.163e-7 | 0.00005757 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Carnobacteriaceae | Marinilactibacillus | Marinilactibacillus_psychrotolerans |
| ID_107 | 0.0007241 | 0.02196 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Lactobacillaceae | Lactobacillus | Lactobacillus_delbrueckii |
| ID_17 | 0.00009557 | 0.008697 | Bacteria | Proteobacteria | Gammaproteobacteria | Enterobacteriales | Enterobacteriaceae | Proteus | Proteus_vulgaris |
| ID_71 | 0.001726 | 0.04028 | Bacteria | Firmicutes | Bacilli | Bacillales | Staphylococcaceae | Staphylococcus | Staphylococcus_vitulinus |

Showing 1 to 6 of 9 entries (filtered from 398 total entries)

9 differentially abundant ASV when using raw data!

# Practice session

## Identify abundance differences between conditions with DESeq2

RAW DATA (not subsampled)

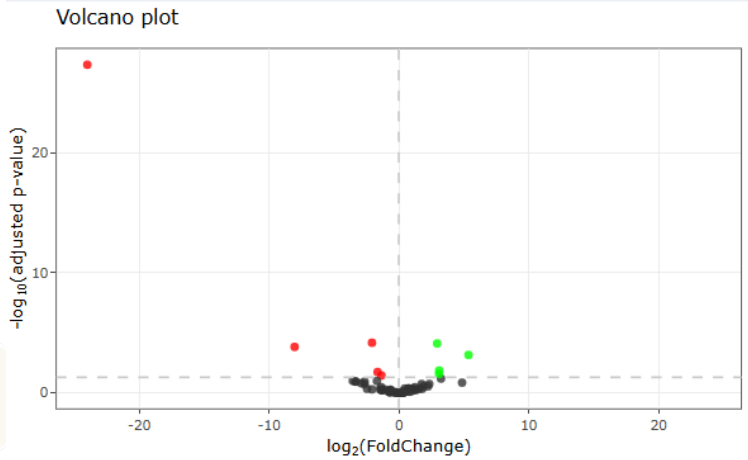
swarm

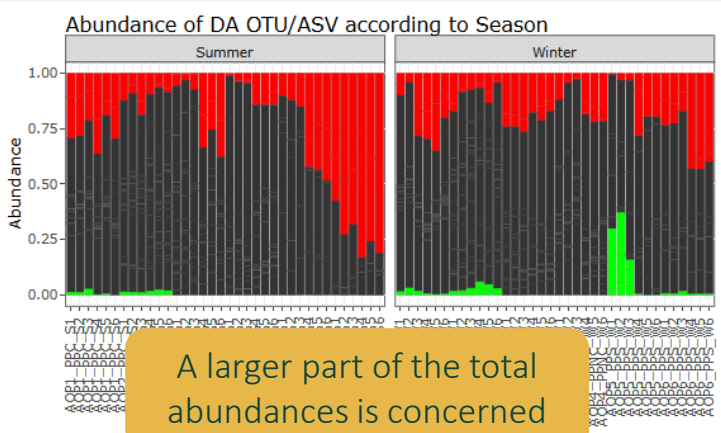
What biological interpretation can be extracted?

Experimental design : Season | Compute!

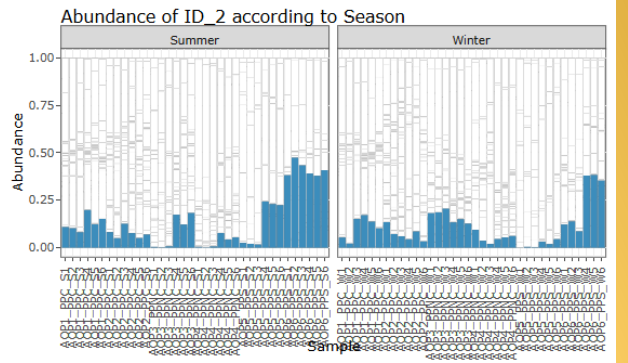
Contrast left : Summer | Contrast right : Winter

Title : Volcano plot

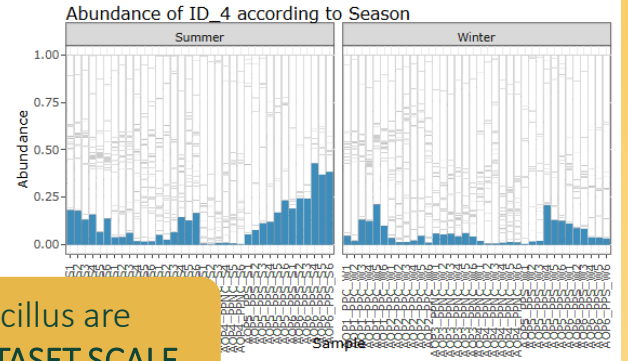



A larger part of the total abundances is concerned




| OTU | Kingdom | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|---|
| ID_2 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Streptococcaceae | Streptococcus | <NA> |




| OTU | Kingdom | Phylum | Class | Order | Family | Genus | Species |
|---|---|---|---|---|---|---|---|
| ID_4 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Lactobacillaceae | Lactobacillus | <NA> |

| OTU | ıge | lfcSE | stat | pvalue | padj | Kingdom | Phylum | Class | Order | Family | Genus |
|---|---|---|---|---|---|---|---|---|---|---|---|
| All | | | | Al | c | Al | All | All | All | All | All |
| ID_2 | -1.326 | 0.4692 | -2.827 | 0.004699 | 0.03603 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Streptococcaceae | Streptococcus |
| ID_4 | -2.061 | 0.4331 | -4.760 | 0.000001938 | 0.00006685 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Lactobacillaceae | Lactobacillus |
| ID_23 | -1.632 | 0.5239 | -3.114 | 0.001844 | 0.01818 | Bacteria | Actinobacteria | Actinobacteria | Micrococcales | Promicromonosporaceae | Cellulosimicrobium |
| ID_13 | 3.104 | 1.048 | 2.962 | 0.003055 | 0.02635 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Carnob | |
| ID_42 | 2.953 | 0.6355 | 4.648 | 0.000003358 | 0.00007722 | Bacteria | Firmicutes | Bacilli | Lactobacillales | Carnob | |

Showing 1 to 6 of 9 entries (filtered from 90 total entries)

Streptococcus and Lactobacillus are decreasing **AT THE WHOLE DATASET SCALE**

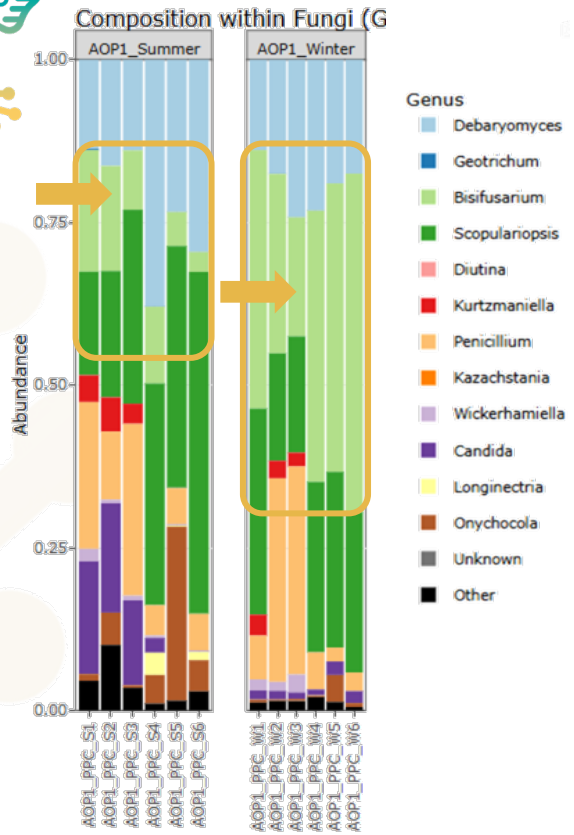
At the Genus level, 9 genera identified

# Practice session

## Identify abundance differences between conditions with DESeq2

RAW DATA
(not subsampled)

What biological interpretation can be extracted?




?


Size factors and dispersion are computed on the complete dataset, differences that are too specific to a small fraction of samples are not well identified

# Practice session

## Identify abundance differences between conditions with DESeq2

RAW DATA
(not subsampled)

Preprocess data

Transformations are applied iteratively, starting from raw data.

Please refer to the documentation to learn more about the transformation modules.

**Transformation module_1**: AOP

**To keep :** AOP1

**Transformation module_2**: Agglomerate taxa

**Rank to agglomerate over :** Kingdom / Phylum / Class / Order / Family / Genus (Genus selected)

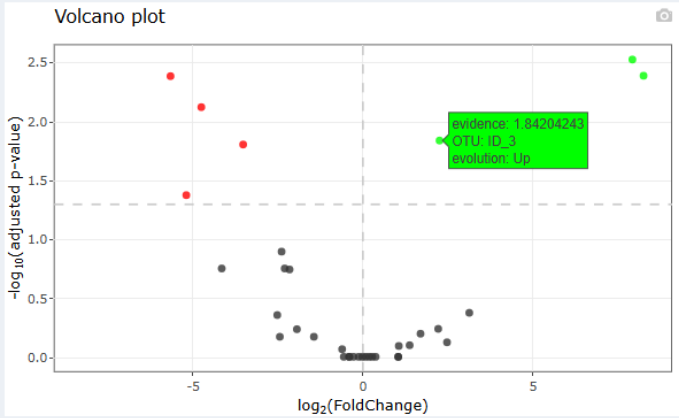
What biological interpretation can be extracted?

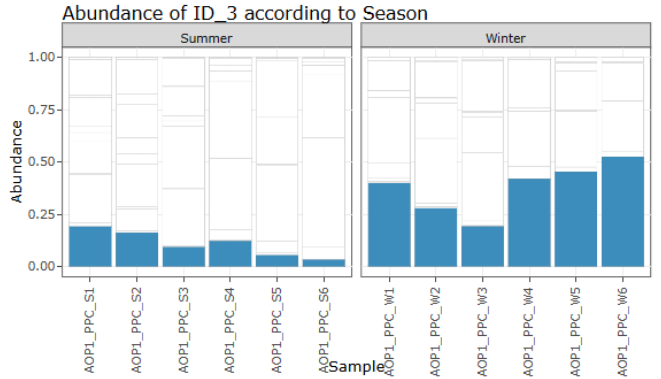
**Experimental design :** Season — Compute!

**Contrast left :** Summer

**Contrast right :** Winter

**Title :** Volcano plot




```
You compare Winter and Summer of the binary variable Season.
A positive log2FoldChange means more abundant in Winter than in Summer.
```

```
  OTU Kingdom     Phylum          Class       Order       Family       Genus Species Ra
1 ID_3   Fungi Ascomycota Sordariomycetes Hypocreales Nectriaceae Bisifusarium    <NA> <
```

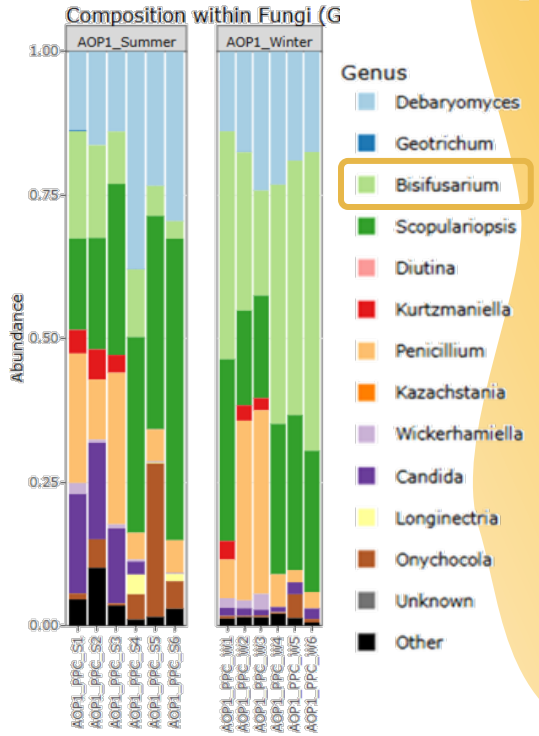
Table of OTUs/ASVs with significant effect (padj <= 0.05)

CSV | Copy | Excel | Search:

| OTU | hange | lfcSE | stat | pvalue | padj | Kingdom | Phylum | Class | Order | Family | Genus | Sp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID_3 | 2.255 | 0.7337 | 3.073 | 0.002116 | 0.01439 | Fungi | Ascomycota | Sordariomycetes | Hypocreales | Nectriaceae | Bisifusarium | |
| ID_9 | -5.661 | 1.587 | -3.567 | 0.0003614 | 0.004095 | Fungi | Ascomycota | Sordariomycetes | Hypocreales | Nectriaceae | Longinectria | |
| ID_84 | -4.754 | 1.430 | -3.326 | 0.0008824 | 0.007501 | Fungi | | | | | Stephanonectria | |
| ID_28 | -3.524 | 1.177 | -2.995 | 0.002745 | 0.01556 | Fungi | | | | | Pichia | |
| ID_83 | 7.928 | 2.020 | 3.924 | 0.00008715 | 0.002963 | Fungi | | | | | Apiotrichum | |

Showing 1 to 6 of 7 entries (filtered from 36 total entries)

Bisifusarium is significantly differentially abundant between seasons in AOP1 **AT AOP1 SCALE**



# DESeq2 summary

Differential abundance analyses **are not relevant in all cases**
(biological questions and/or object, experimental designs…)

When factors are “nested” (e.g. seasons within AOP), DESeq2 works with all samples, not within each subgroups (even if you compare two conditions). This global estimation can distort results if the nested groups differ strongly.
→ use subgroups with Preprocess

Overall, DESeq2 (and other differential analysis methods) are very sensitive to input parameters: slight changes can have a strong impact on the output. *Handle with care!*

**Elementary methods provide more replicable results in microbial differential abundance analysis**

Juho Pelto ✉, Kari Auranen, Janne V Kujala, Leo Lahti

*Briefings in Bioinformatics*, Volume 26, Issue 2, March 2025, bbaf130,

https://doi.org/10.1093/bib/bbaf130

# DESeq2 in FROGS Stat

Two steps:

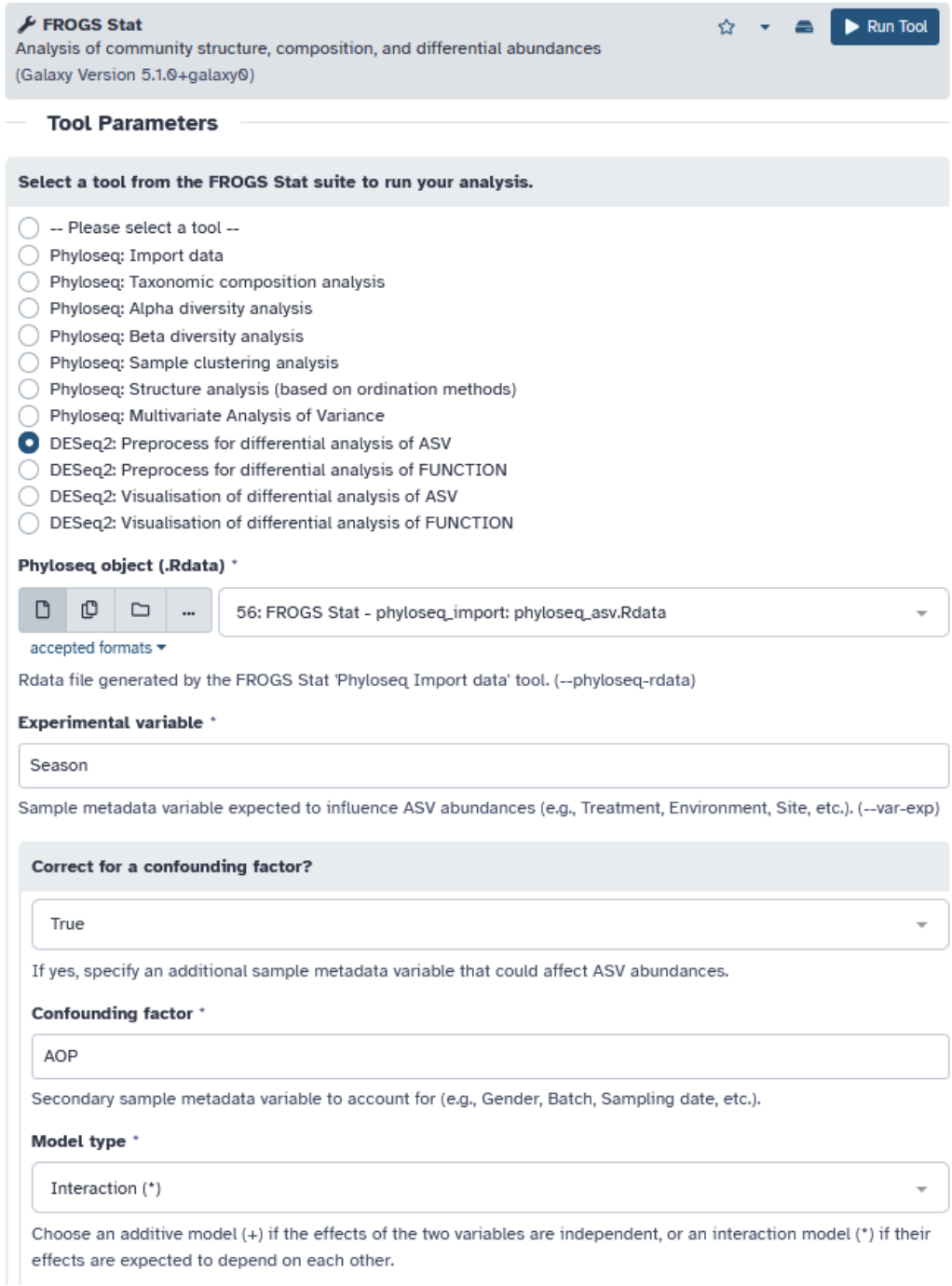
In FROGS tools, select FROGS Stat and choose DESeq2: Preprocess for differential analysis of ASV

Fill with a phyloseq import output

Using raw data (not subsampled) is recommended

Set your experimental variable

Confounding factors are available with FROGS Stat:
Relevant when you have mixed or nested effects and are interested in a single factor, despite the variation of the second one



# DESeq2 in FROGS Stat

2nd step:

In FROGS tools, select FROGS Stat and choose DESeq2: Visualisation of differential analysis of ASV

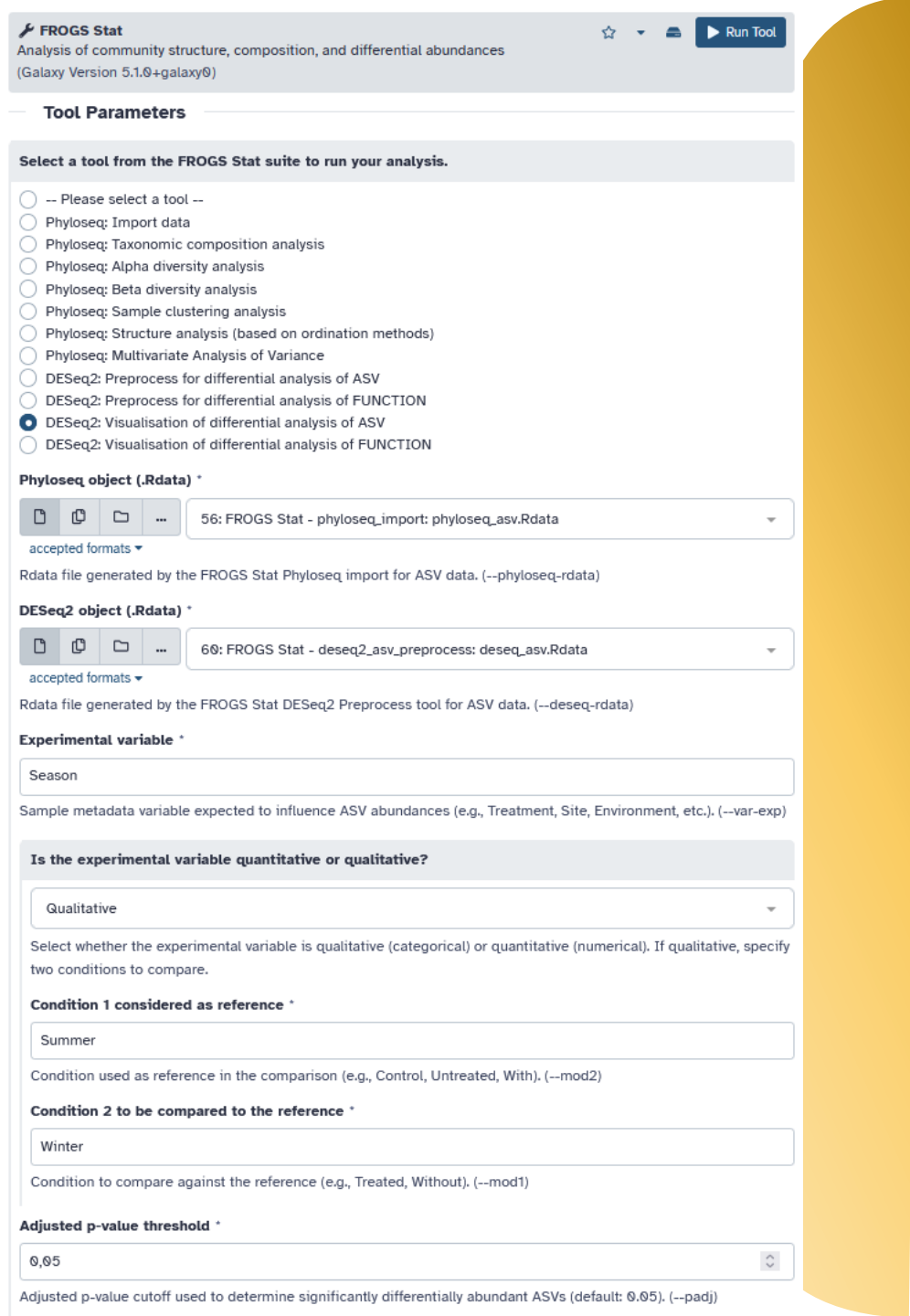
Fill with the *phyloseq import output*

Fill with the *DESeq2 preprocess output*

Set your experimental variable

Choose your comparison

With Galaxy it is required to run as many *DESeq2: Visualisation of differential analysis of ASV* tools as you have comparisons to visualise



# DESeq2 in FROGS Stat

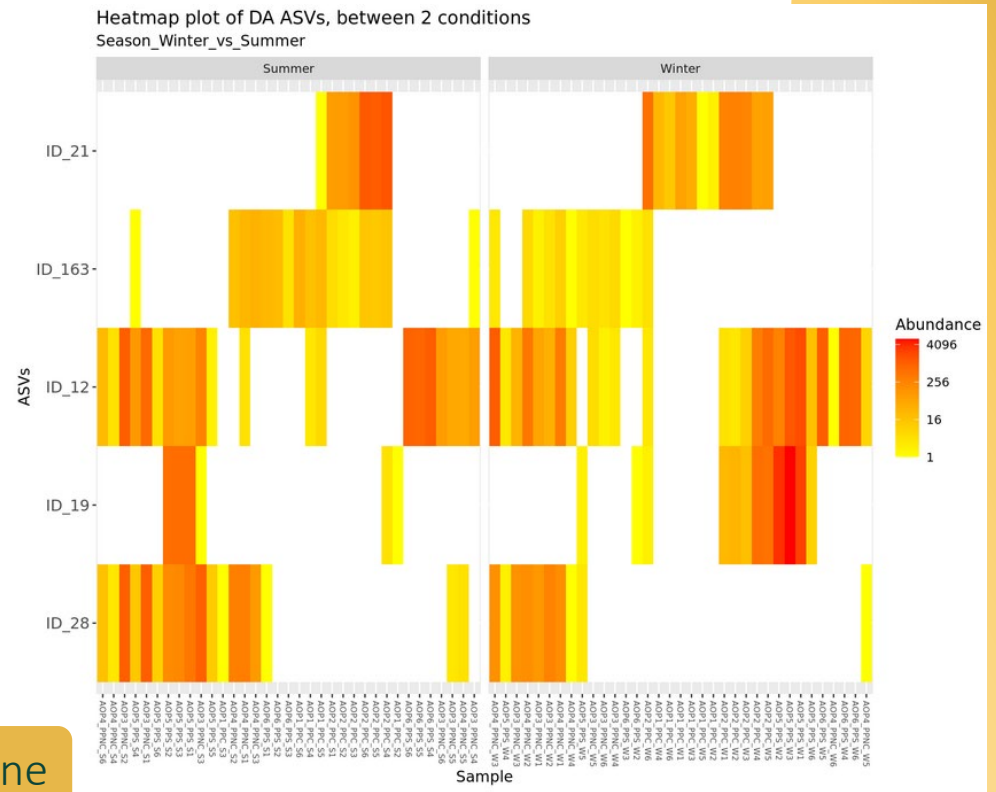
| ID | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Kingdom | Phylum | Class |
|---|---|---|---|---|---|---|---|---|---|
| | All | All | | | All | All | | All | All |
| ID_12 | 238.407 | -11.1072 | 1.97600 | -5.62103 | 1.89823e-8 | 0.00000377747 | Bacteria | Actinobacteria | Actinobacteria |
| ID_19 | 202.334 | -12.2087 | 3.12822 | -3.90276 | 0.0000951009 | 0.0126167 | Bacteria | Proteobacteria | Gammaproteobacteri |
| ID_28 | 90.1279 | -13.9309 | 2.30097 | -6.05437 | 1.40969e-9 | 5.61058e-7 | Bacteria | Proteobacteria | Gammaproteobacteri |
| ID_21 | 58.7431 | 6.59175 | 1.79567 | 3.67090 | 0.000241695 | 0.0240486 | Bacteria | Firmicutes | Clostridia |
| ID_163 | 5.05320 | -6.30783 | 1.81597 | -3.47353 | 0.000513660 | 0.0408873 | Bacteria | Actinobacteria | Actinobacteria |

Show 10 entries

Showing 1 to 5 of 5 entries

Previous 1 Next

Same outputs than Easy16S




DESeq2 here lead to identify only 5 ASVs while not considering confounding factor allowed to identify more

And a specific one

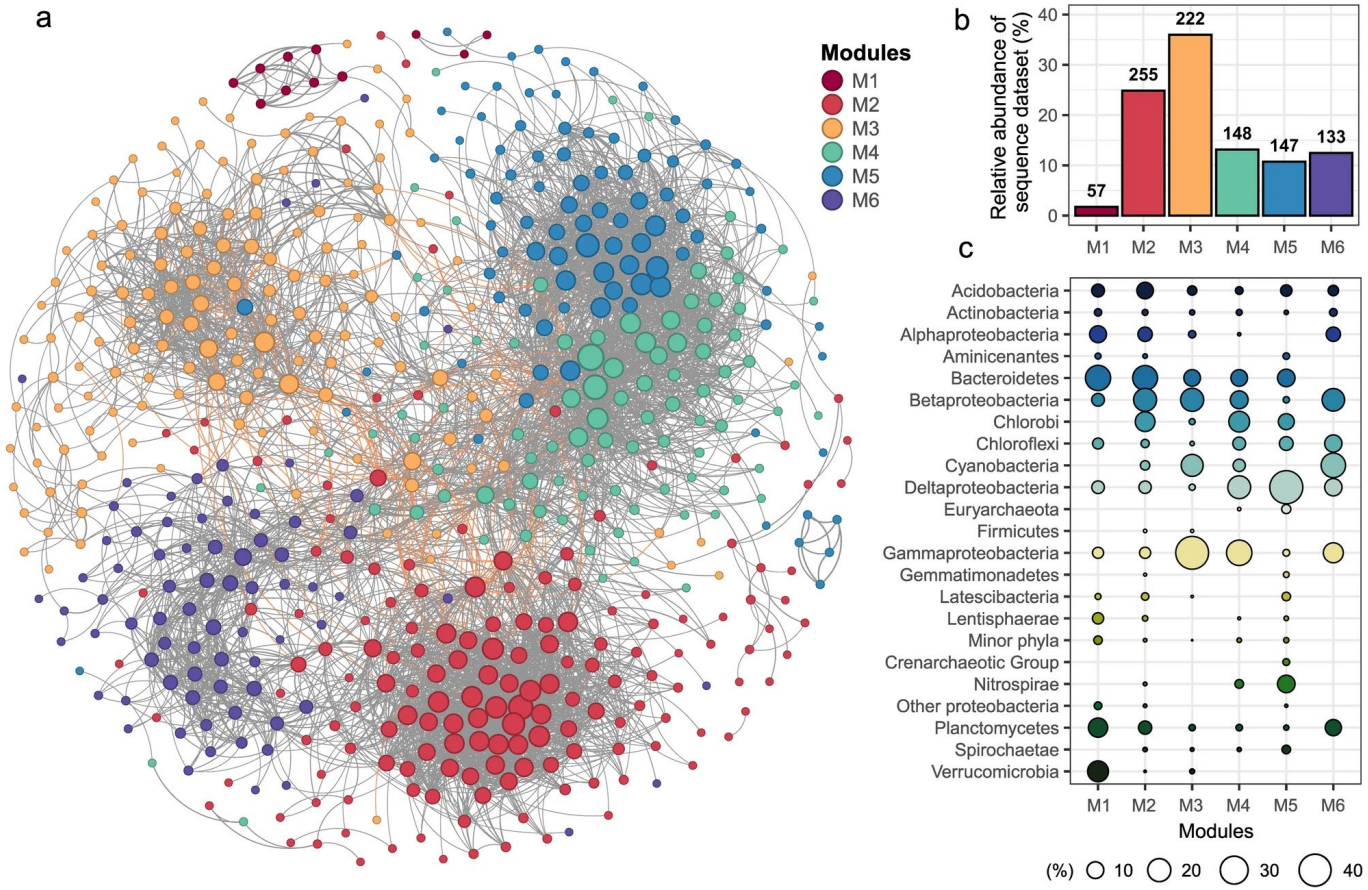
# FROGS Stats

# Networks analyses

# Network analyses principle

Network analyses aim to infer relationships between taxa by examining how their abundances vary across samples

A network represent taxa as nodes and their association (co-occurrences or correlations) as edges.



# Why there are no network analyses in FROGS?

Network analyses are based on **co-occurrences** or **correlations**.

These two metrics are highly biased in metabarcoding data!

Metabarcoding are **sparse** and **compositional**

Tables with many 0 values:

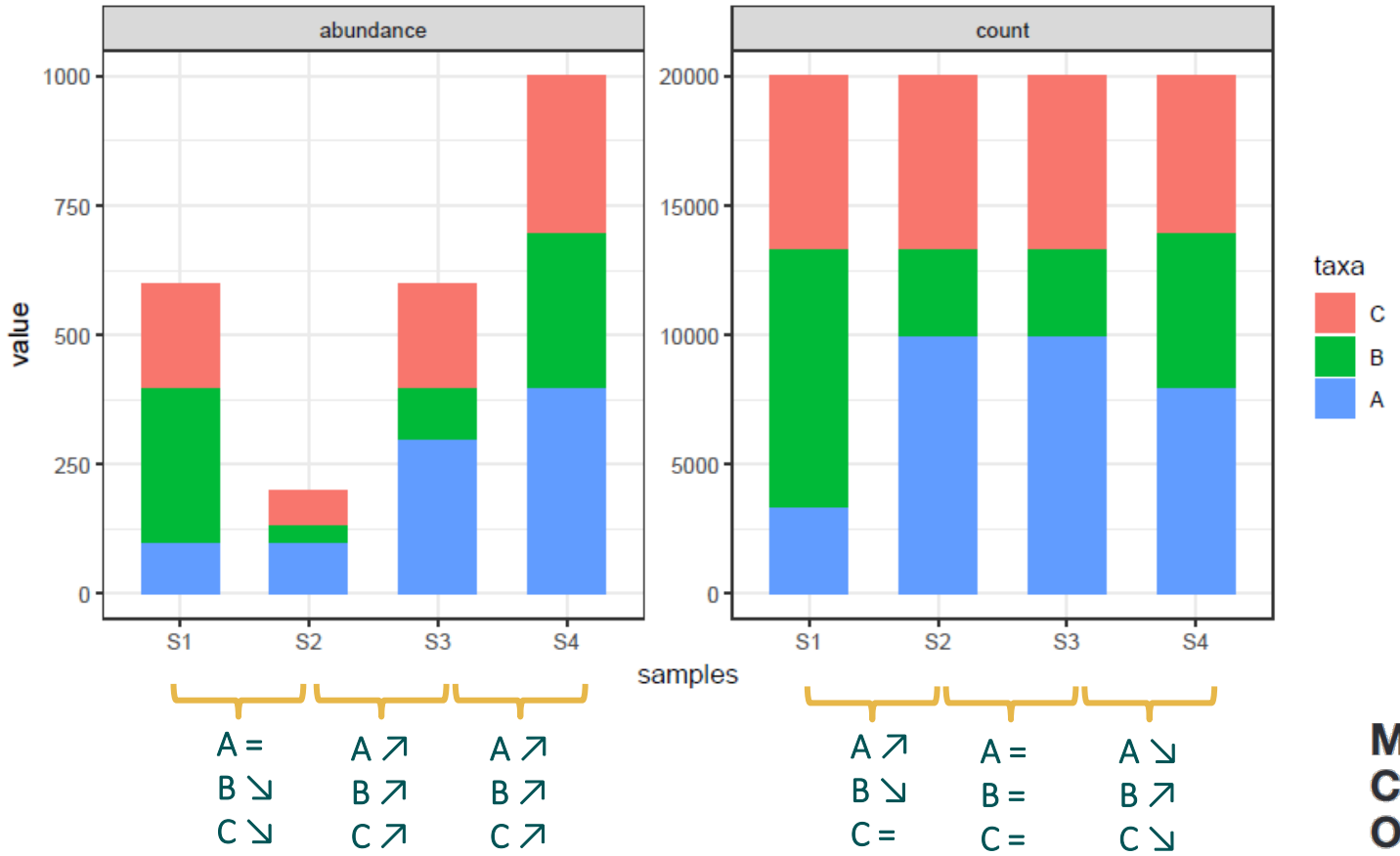
- Strong depth and random effects on ASV detection for low abundances
- Can creates artificial co-absence patterns
- Difficulty to distinguish “true abundance” from “undetected”

Methodologically (through equimolar pooling), sequencing produces **relative**, not absolute abundances

Total read counts is constrained.
An **increase in one taxa** (especially if it is abundant) **forces decreases in all others**, even if it is not biologically true

# Data compositionality

Methodologically (through equimolar pooling), sequencing produces relative, not absolute abundances.




**Microbiome Datasets Are Compositional: And This Is Not Optional**

*Gregory B. Gloor[1]*, Jean M. Macklaim[1], Vera Pawlowsky-Glahn[2] and Juan J. Egozcue[3]*

https://doi.org/10.3389/fmicb.2017.02224

# Data compositionality

Methodologically (through equimolar pooling), sequencing produces relative, not absolute abundances.

Total read counts is constrained.
➡ An **increase in one taxa** (especially if it is abundant)
**forces decreases in all others**, even if it is not biologically true

➡ **Spurious correlations are unavoidable**

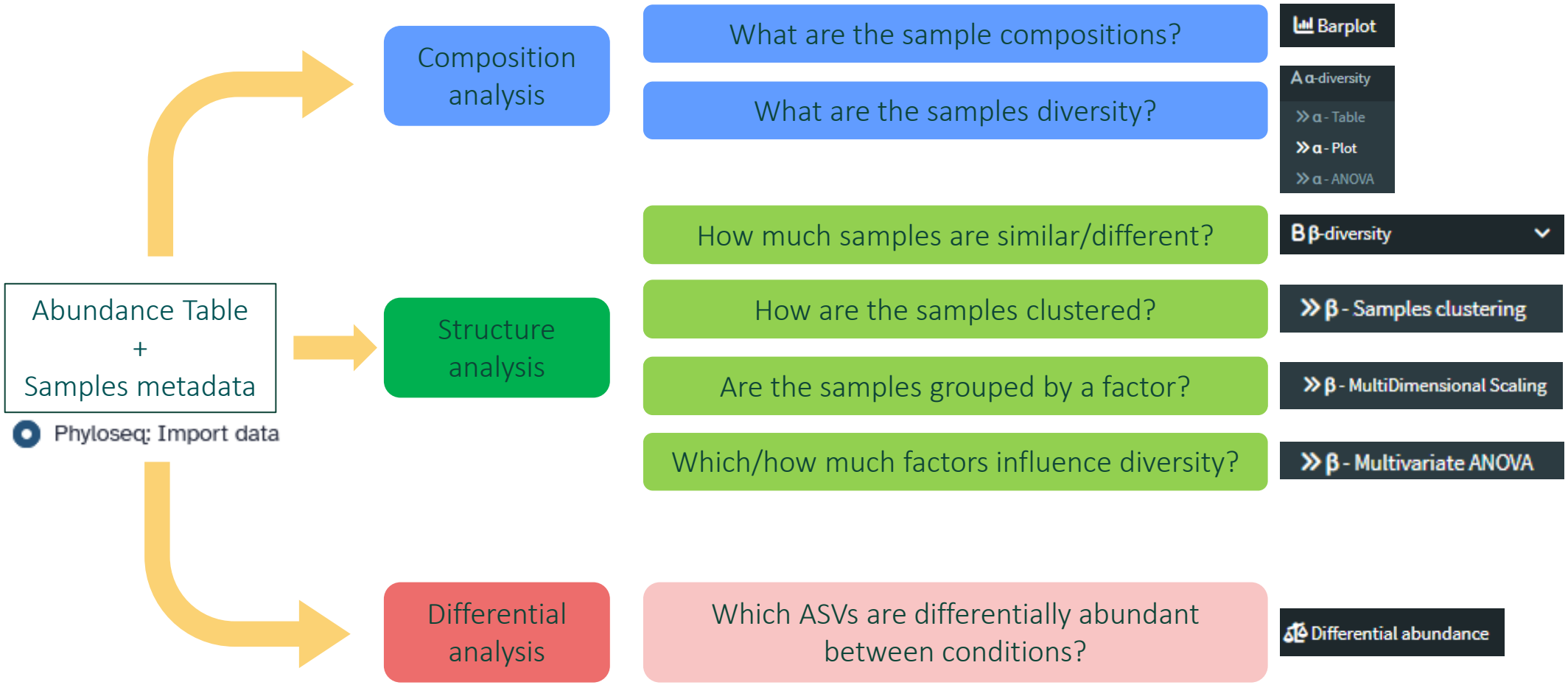
Several data-transformation methods are designed to mitigate compositional effects that sometimes work well,
but their effectiveness is variable, and it is very difficult to know whether or not they are effective for a given dataset.

Correlations are particularly sensitive to compositional effects.
Correlation analyses and especially correlation networks should therefore be handled with great caution (or avoided…)

It would theoretically be possible to convert counts and proportions into absolute abundances if:

- Total biomass is known
- DNA extraction is performed under non-saturating conditions
- Copy-number correction is effective
- Extraction and amplification biases are negligible (or neglected)

# Exploratory Data Analysis in summary



# A few pieces of advice

Carefully build your metadata .tsv file with as much information as possible
Don’t hesitate to include variables that combine several others, or to turn quantitative variables into categories

Exploratory data analysis **need to be performed on normalised** (e.g. subsampled / rarefied) **data**
(to correct for sequencing depth biases)

**DESeq2** is neither magical **nor suitable for every situations**. It is recommended to use it on raw counts.

No distance is universally “best”. Each provides different information, sometimes more relevant in a given context, but more often complementary.

Above all, **exploring metabarcoding data takes time – a lot of it.**
As with omics in general, following a real scientific workflow (question → hypothesis → test → answer)
is what keeps you moving in a direction rather than sinking in the ocean of omics data.