

---

# C - Training on Galaxy: Metabarcoding

December 2021 - Webinar

## STATISTICS Practice

---

LUCAS AUER, MARIA BERNARD, LAURENT CAUQUIL, MAHENDRA MARIADASSOU, GÉRALDINE PASCAL

How different  
are two  
communities?

My samples  
are they  
homogenous  
or diverse?

What is the  
composition of  
each  
community?

Are the communities  
structured by some known  
environmental factor (pH,  
height, etc)?

Are there  
interactions  
between  
species and  
communities?

Are there OTU with  
differential  
abundance between  
conditions?

# FROGSSTAT with Phyloseq R package

---

- R package (McMurdie and Holmes, 2013) to analyse community composition data in a phylogenetic framework

It uses other R packages:

- Community ecology functions from vegan, ade4
- Tree manipulation from ape
- Graphics from ggplot2
- Differential analysis from DESeq2

# Exercise 1

---

→ At the end of FROGS pipeline, what kind of data do we have ?

# Exercise 1

---

→ At the end of FROGS pipeline, what kind of data do we have ?

FROGS biom containing:

- OTU count tables (required)
- OTU description : taxonomy

Phylogenetic tree in Newick format

Metadata: sample description in TSV file

# Exercise 1

---

➔ Take a look at the metadata

# Exercise 1

→ Take a look at the metadata

FoodType:

Meat or Seafood

EnvType: 8 environment types

1	2	3	4
	EnvType	Description	FoodType
BHT0.LOT01	BoeufHache	LOT1	Meat
BHT0.LOT03	BoeufHache	LOT3	Meat
BHT0.LOT04	BoeufHache	LOT4	Meat
BHT0.LOT05	BoeufHache	LOT5	Meat
BHT0.LOT06	BoeufHache	LOT6	Meat
BHT0.LOT07	BoeufHache	LOT7	Meat
BHT0.LOT08	BoeufHache	LOT8	Meat
BHT0.LOT10	BoeufHache	LOT10	Meat
VHT0.LOT01	VeauHache	LOT1	Meat
VHT0.LOT02	VeauHache	LOT2	Meat
VHT0.LOT03	VeauHache	LOT3	Meat
VHT0.LOT04	VeauHache	LOT4	Meat

Meat → Ground Beef, Ground veal, Poultry sausage, Diced bacon

Seafood → Cooked schrimps, Smoked salmon, Salmon filet, Cod filet

---

# Phyloseq Import Data tool

---

PHYLOSEQ OBJECT CREATION



# Phyloseq : Data import

1. Statistical analysis is done in R, so an R object called Rdata must be created.

2. Run PhyloSeq Data import

The FROGS biom format contains:

- OTU count tables (required)
- OTU description : taxonomy

Others information used in FROGSSTAT are:

- sample description in TSV file
- phylogenetic tree in Newick format (nwk or nhx)

3. Create 2 phyloseq objects, with and without normalization (rename them)

**FROGSSTAT Phyloseq Import Data** from 3 files: biomfile, samplefile, treefile (Galaxy Version 3.2.2) Options

**Abundance biom file with taxonomical metadata**  
19: FROGS Affiliation OTU: affiliation.biom  
The file contains the OTU informations (format: biom1).

**Sample tsv file**  
2: metadata\_chaillou.tsv  
The file contains the samples informations (format: tabular).

**Tree file (optional)**  
24: FROGS Tree: tree.nwk  
The file contains the tree informations (format: Newick - nhx or nwk).

**Names of taxonomics levels**  
Kingdom Phylum Class Order Family Genus Species  
The ordered taxonomic levels stored in BIOM. Each level is separated by one space.

**Do you want to normalise your data ?**  
 Yes  No  
To normalise data before statistical analysis (default : No).

# Exercise 2

---

1. What are the resulting datasets ?
2. What is the difference between the resulting objects with and without normalization ?
3. Explore the HTML results

# Exercise 2

---

## 1. What are the resulting datasets ?

- Rdata file: R object used by phyloseq package for statistics
- HTML report: summary of the phyloseq object

# Exercise 2

---

2. What is the difference between the resulting objects with and without normalization ?

Without normalization



**Summary** Ranks Names Sample metadata Plot tree

Code

```
phyloseq-class experiment-level object
otu_table() OTU Table: [ 495 taxa and 64 samples ]
sample_data() Sample Data: [ 64 samples by 4 sample variables ]
tax_table() Taxonomy Table: [ 495 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 495 tips and 494 internal nodes ]
```

# Exercise 2

---

2. What is the difference between the resulting objects with and without normalization ?

Summary

Ranks Names

Sample metadata

Plot tree

Code

```
phyloseq-class experiment-level object
otu_table() OTU Table: [ 495 taxa and 64 samples ]
sample_data() Sample Data: [ 64 samples by 4 sample variables ]
tax_table() Taxonomy Table: [ 495 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 495 tips and 494 internal nodes ]
```

Code

```
Number of sequences in each sample after normalization: 7638
```

With normalization (rarefaction)

Minimum number of sequences kept in each sample



# Exercise 2

2. What is the difference between the resulting objects with and without normalization ?

With normalization (rarefaction)



Be aware that the number of OTUs (taxa) may decrease

Summary

Ranks Names

Sample metadata

Plot tree

Code

```
phyloseq-class experiment-level object
otu_table() OTU Table: [ 495 taxa and 64 samples ]
sample_data() Sample Data: [ 64 samples by 4 sample variables ]
tax_table() Taxonomy Table: [ 495 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 495 tips and 494 internal nodes ]
```

Code

```
Number of sequences in each sample after normalization: 7638
```

# Exercise 2

---

## 3. Explore the HTML results

Phyloseq 1.20.0



Code

Summary

**Ranks Names**

Sample metadata

Plot tree

Code

Taxonomic levels

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

# Exercise 2

## 3. Explore the HTML results

Summary Ranks Names **Sample metadata** Plot tree

Sample variables: EnvType, Description, FoodType, SampleID

Code

EnvType : DesLardons, MerguezVolaille, BoeufHache, VeauHache, SaumonFume, FiletSaumon, FiletCabillaud, Crevette

Description : LOT1, LOT3, LOT4, LOT5, LOT6, LOT7, LOT8, LOT10, LOT9, LOT2

Code

FoodType : Meat, Seafood

SampleID : DLT0.LOT01, DLT0.LOT03, DLT0.LOT04, DLT0.LOT05, DLT0.LOT06, DLT0.LOT07, DLT0.LOT08, DLT0.LOT10, MVT0.LOT01, MVT0.LOT03, MVT0.LOT05, MVT0.LOT06, MVT0.LOT07, MVT0.LOT08, MVT0.LOT09, MVT0.LOT10, BHT0.LOT01, BHT0.LOT03, BHT0.LOT04, BHT0.LOT05, BHT0.LOT06, BHT0.LOT07, BHT0.LOT08, BHT0.LOT10, VHT0.LOT01, VHT0.LOT02, VHT0.LOT03, VHT0.LOT04, VHT0.LOT06, VHT0.LOT07, VHT0.LOT08, VHT0.LOT10, SFT0.LOT01, SFT0.LOT02, SFT0.LOT03, SFT0.LOT04

Variable names

Script R

the different modalities for each qualitative variable

**Warning !**  
Metadata order (in each sample variable) are used to organize graphics.  
So take extra care when you construct your sample\_metadata file

It may make sense to order the metadata file i.e. the meats are together and the seafood together



# Exercise 2

## 3. Explore the HTML results



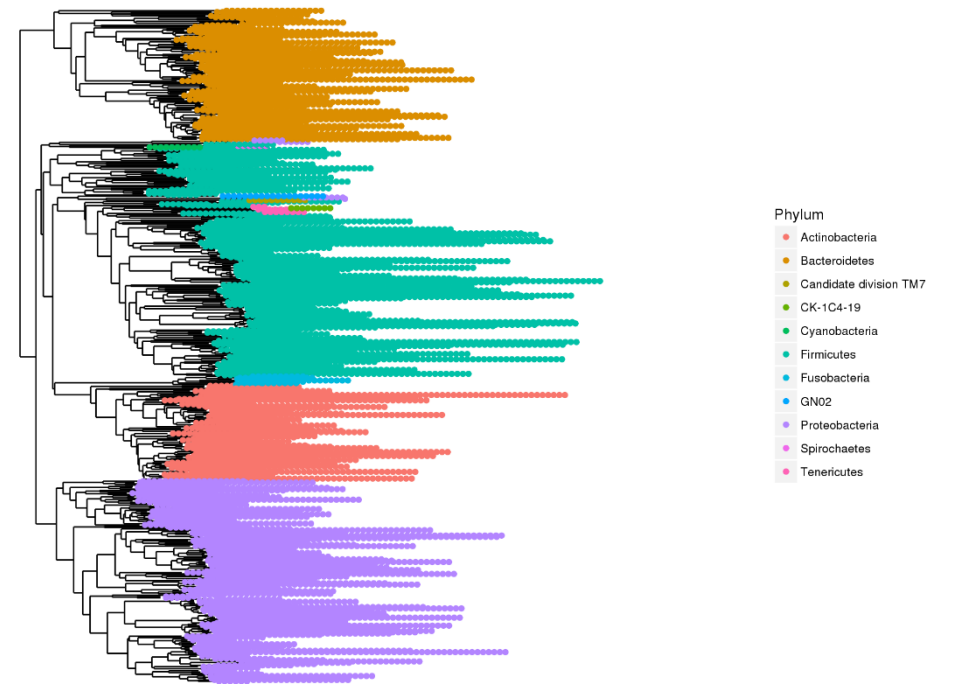
Summary

Ranks Names

Sample metadata

Plot tree

Phylogenetic tree colored by Phylum



# Exercise 2

## 3. Explore the HTML results

Summary

Ranks Names

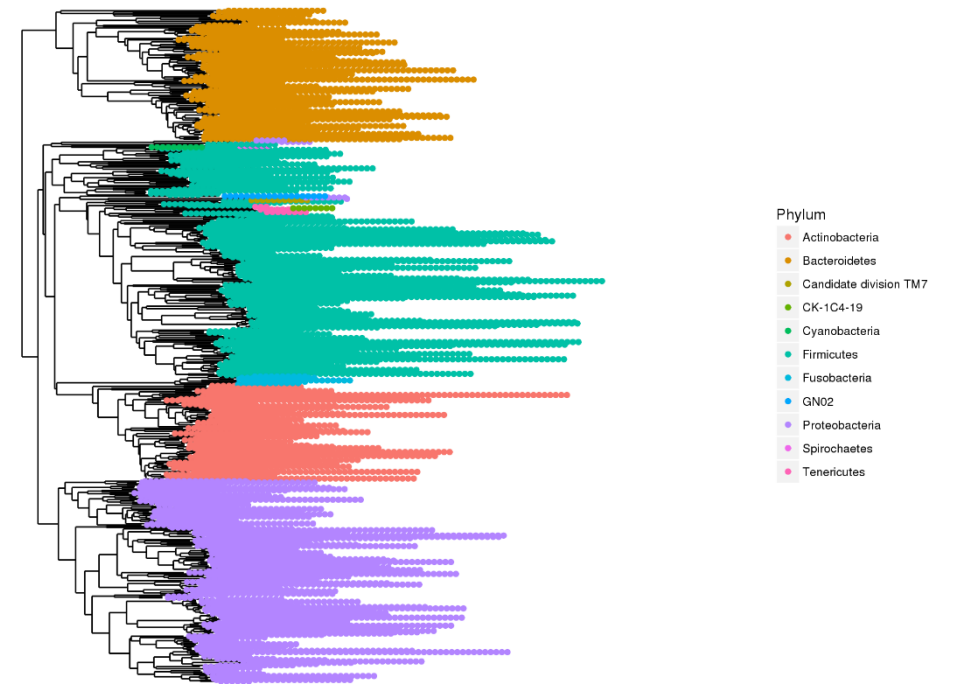
Sample metadata

Plot tree

→ Information: Most represented phylum (in OTUs count)

- Bacteroidota
- Firmicutes
- Actinobacteriota
- Proteobacteria

Phylogenetic tree colored by Phylum



---

# Biodiversity analysis

---

# The points we will cover on biodiversity analysis

---

1. Exploring sample composition
2. Notions of biodiversity
3.  $\alpha$ -diversity analysis
4.  $\beta$ -diversity analysis

---

# I. Biodiversity analysis

---

COMPOSITION VISUALIZATION

# Exploring biodiversity : visualization

**FROGSSTAT Phyloseq Composition Visualisation** with bar plot and composition plot (Galaxy Version 3.2.2) Options

**Phyloseq object (format rdata)**  
26: Phyloseq.Rdata  
This is the result of FROGS Phyloseq Import Data tool.

**Grouping variable**  
EnvType  
Experimental variable used to group samples (Treatment, Host type, etc).

**Taxonomic level to filter your data**  
Kingdom  
ex: Kingdom, Phylum, Class, Order, Family, Genus, Species

**Taxa (at the above taxonomic level) to keep in the dataset**  
Bacteria  
ex: Bacteria (when filtering at the Kingdom level), Firmicutes (when filter specified, i.e. Firmicutes Proteobacteria)

**Taxonomic level used for aggregation**  
Phylum  
ex: Family (when filtering at the Phylum level). The aggregation level

**Number of most abundant taxa to keep**  
9  
ex: 9, i.e. Tool keeps the 9 most abundant taxa and the remaining taxa

Execute

At what taxonomic rank do we want to study?

Inside this taxonomic rank, what are the target group ?

On which rank do we want to group the OTUs?

Number of majority groupings to be displayed

Explore the sample **RAW** or **NORMALISED** count

Choose a sample variable to organize graphics: either EnvType or FoodType



For the first usage, let the default parameters

# Exercise 3

---

1. What are the resulting datasets ?
2. What is the difference between Bar plot and Plot composition ?
3. What biological information could you extract ?
4. What are the perspectives for going further?

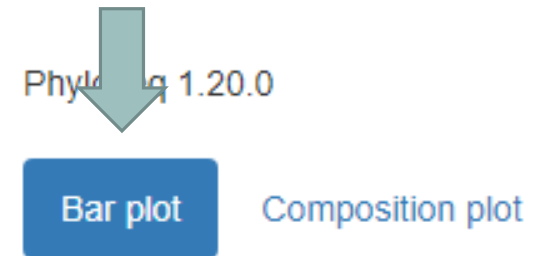
# Exercise 3

---

1. What are the resulting datasets ?

→ HTML report: summary of the phyloseq object

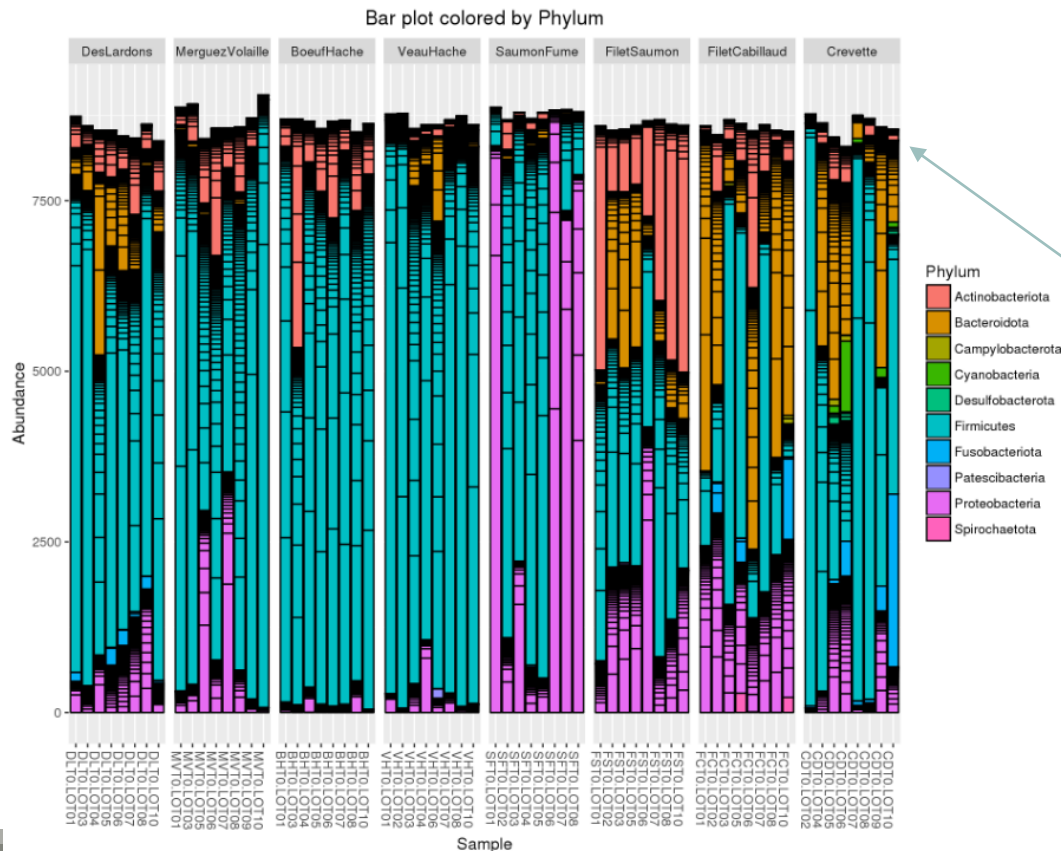
- Bar plot
- Composition plot





# Exercise 3

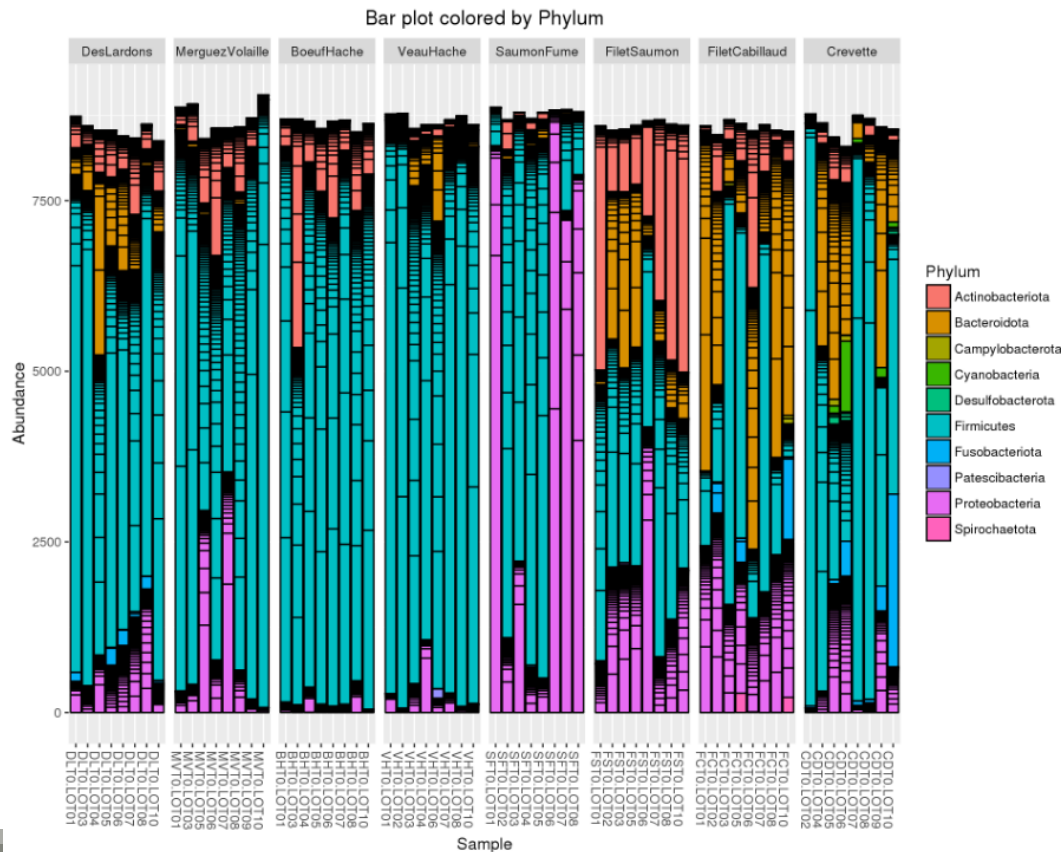
## 2. What is the difference between Bar plot and Plot composition ?



- one rectangle is one OTU
- one color is one phylum
- y axis: number of sequences – these are absolute counts
- size of rectangle depends on number of sequences

# Exercise 3

## 2. What is the difference between Bar plot and Plot composition ?



### Limitations:

- Plot bar works at the OTU-level and displays all the OTU at the specified rank
- This may lead to cluttered graphics and unnecessary legends
- No easy way to look at a subset of the data
- Works with absolute counts (beware of unequal depths or used normalized function)



[load-extra-functions.R](#)



Bar plot

Composition plot

# Exploring biodiversity : visualization

Another graph: `plot_composition` function :

- Works with **relative abundances**
- **Subsets OTUs** at a given taxonomic level
- **Aggregates OTUs** at another taxonomic level
- Shows **only a given number** of taxa

## Taxonomic level to filter your data

Kingdom

ex: Kingdom, Phylum, Class, Order, Family, Genus, Species

## Taxa (at the above taxonomic level) to keep in the dataset

Bacteria

ex: Bacteria (when filtering at the Kingdom level), Firmicutes (when filtering at the Phylum level). Multiple taxa (separated by a space) can be specified, i.e. Firmicutes Proteobacteria

## Taxonomic level used for aggregation

Phylum

ex: Family (when filtering at the Phylum level). The aggregation level must be below the filtering level.

## Number of most abundant taxa to keep

9

ex: 9, i.e. Tool keeps the 9 most abundant taxa and the remaining taxa are aggregated in a group 'Other'

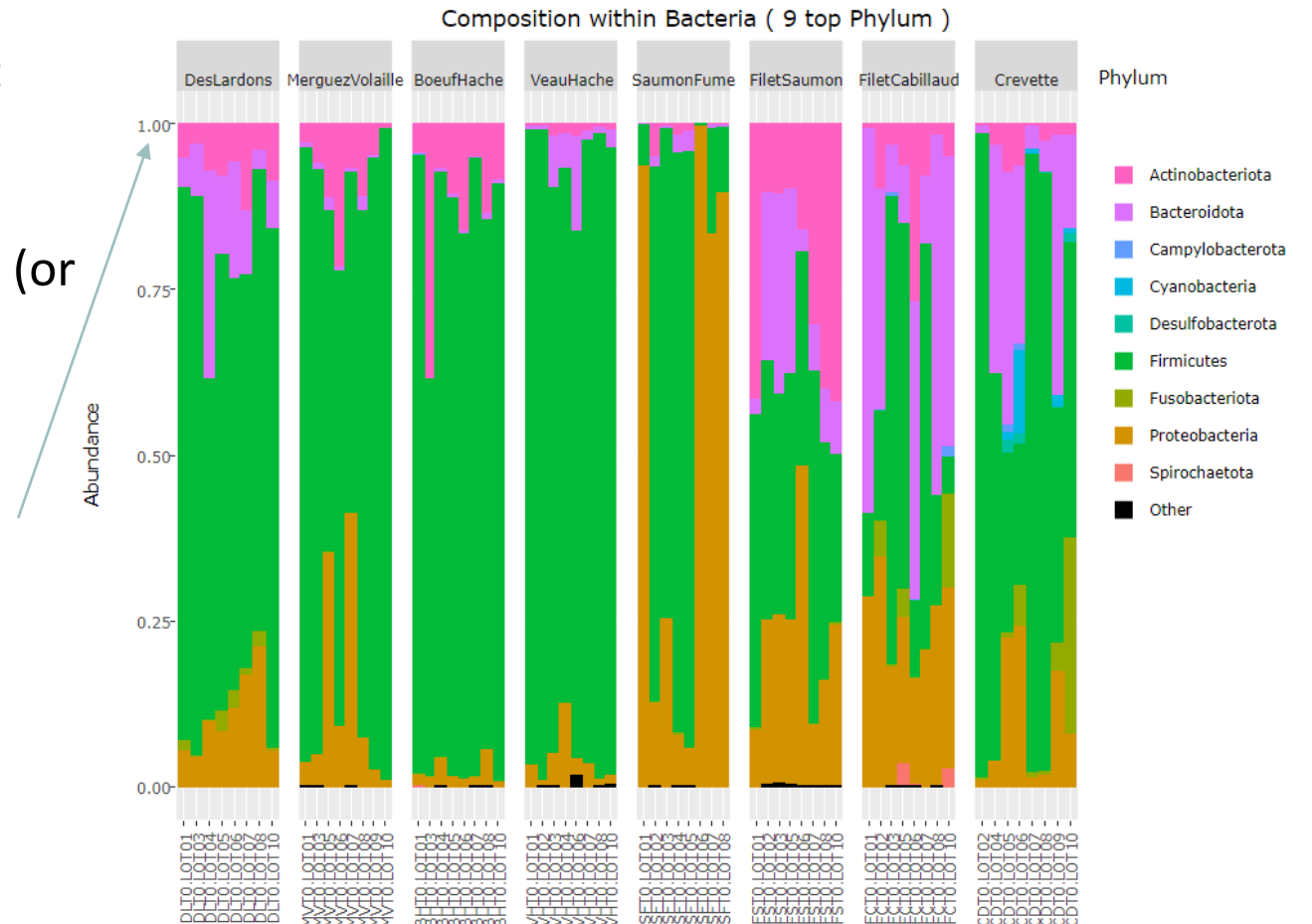
# Exercise 3

Bar plot

Composition plot

2. What is the difference between Bar plot and Plot composition ?

- one rectangle is one phylum (no borderline) (or any other specified taxonomy rank)
- one color is one phylum
- y axis: counts are reduced to 1, so, here, we have relative counts

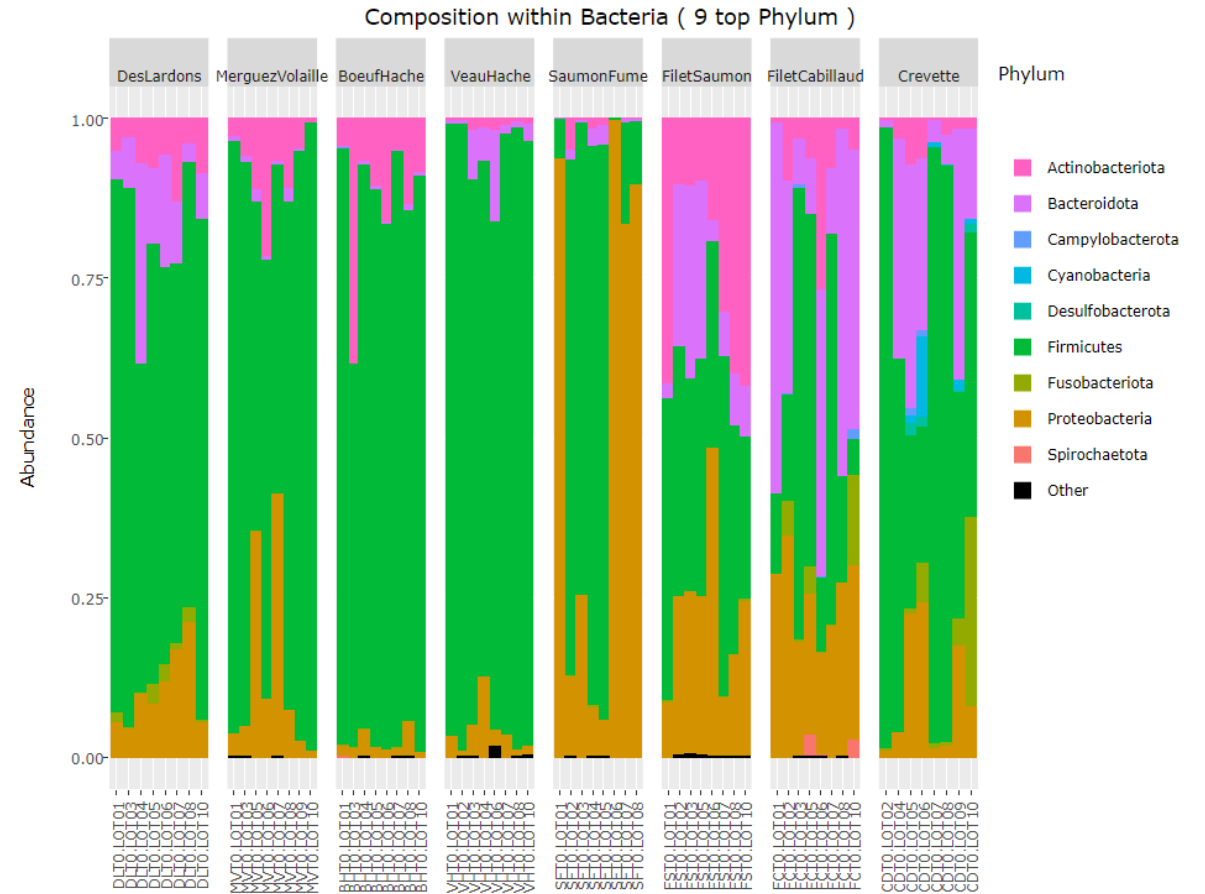


# Exercise 3

Bar plot

Composition plot

3. What biological information could you extract ?



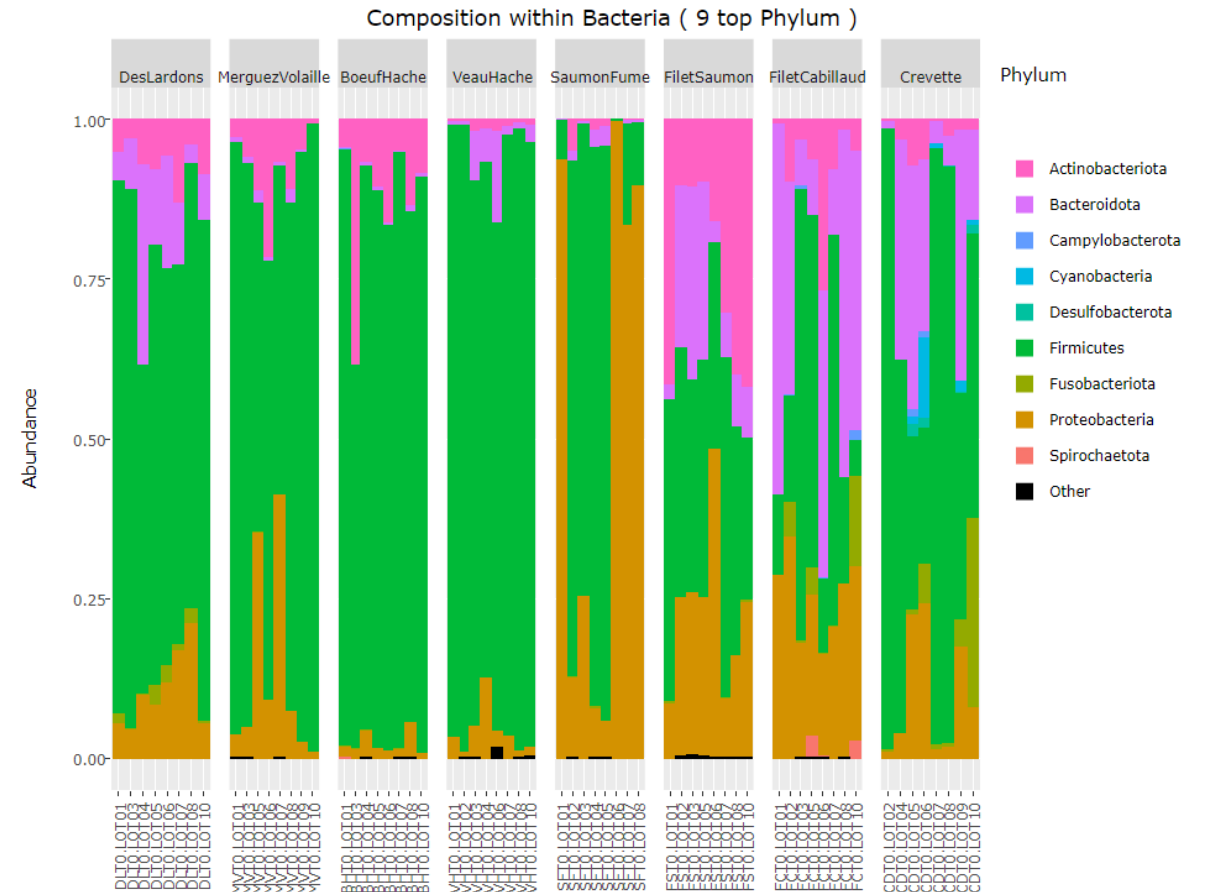
# Exercise 3

Bar plot

Composition plot

## 3. What biological information could you extract ?

- Meat types on the left share common Phylum composition, with a majority of Firmicutes (easy to remark thanks of ordered levels)
- Seafoods seem to be much more variable
- Firmicutes and Proteobacteria are present in all samples, but with a wide range of abundance

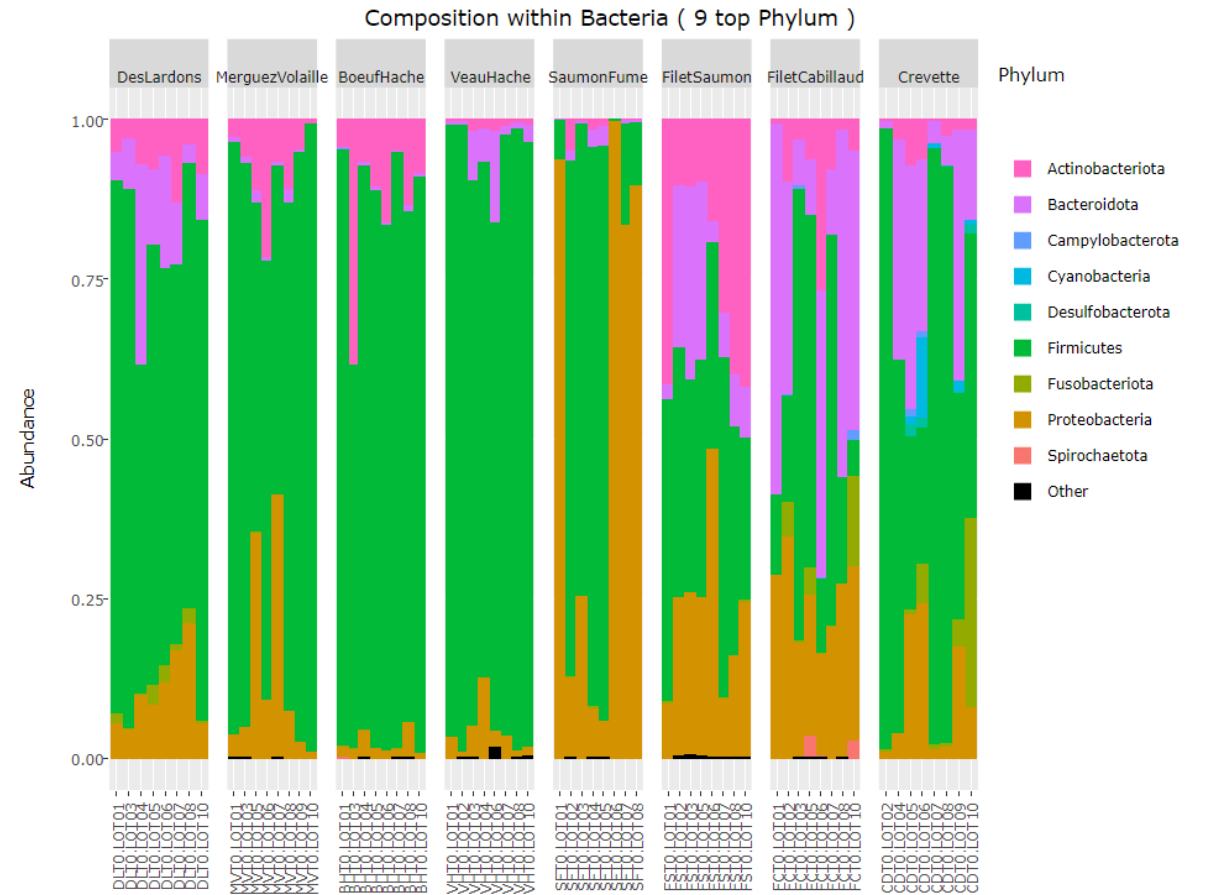


# Exercise 3

Bar plot

Composition plot

4. What are the perspectives for going further?



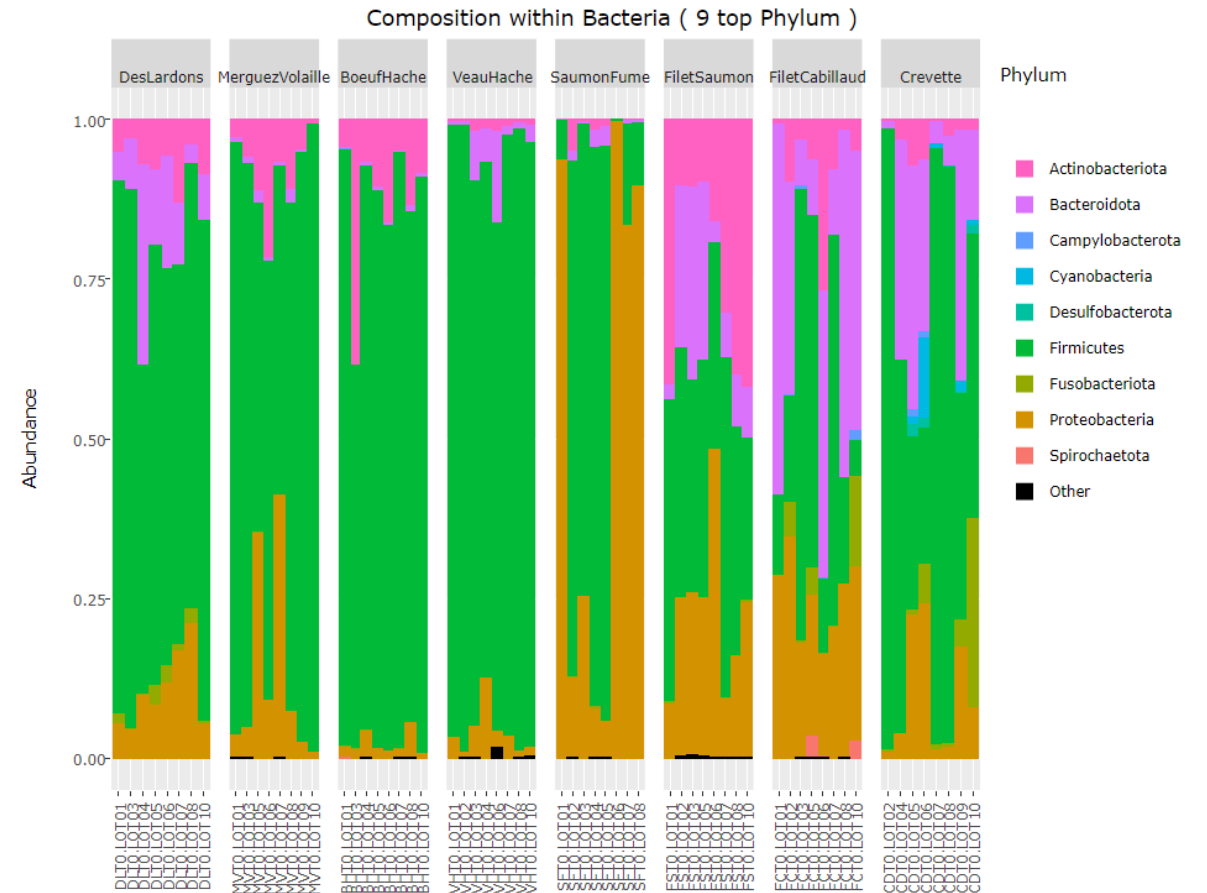
# Exercise 3

Bar plot

Composition plot

## 4. What are the perspectives for going further?

- What is the composition of the 9 most abundant Families of *Firmicutes* ?
- What is the composition of the 9 most abundant Families of *Proteobacteria* ?





# Exercise 4

---

1. What is the composition of the 9 most abundant Families of Firmicutes ?
2. What is the composition of the 9 most abundant Families of Proteobacteria ?

# Exercise 4

## 1. What is the composition of the 9 most abundant Families of Firmicutes ?

### Taxonomic level to filter your data

Phylum

ex: Kingdom, Phylum, Class, Order, Family, Genus, Species

### Taxa (at the above taxonomic level) to keep in the dataset

Firmicutes

ex: Bacteria (when filtering at the Kingdom level), Firmicutes (when filtering at the Phylum level). Multiple taxa (separated by a space) can be specified, i.e. Firmicutes Proteobacteria

### Taxonomic level used for aggregation

Family

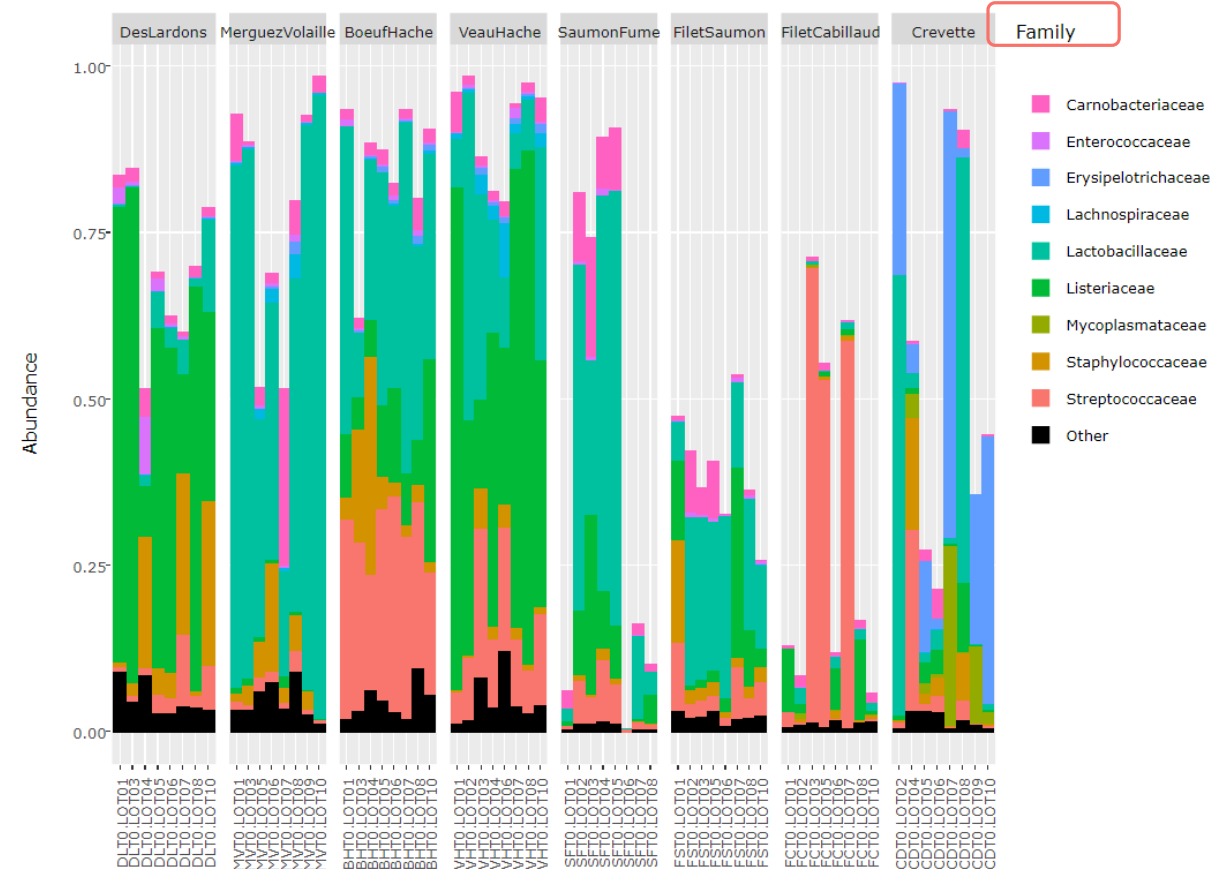
ex: Family (when filtering at the Phylum level). The aggregation level must be below the filtering level.

### Number of most abundant taxa to keep

9

ex: 9, i.e. Tool keeps the 9 most abundant taxa and the remaining taxa are aggregated in a group 'Other'

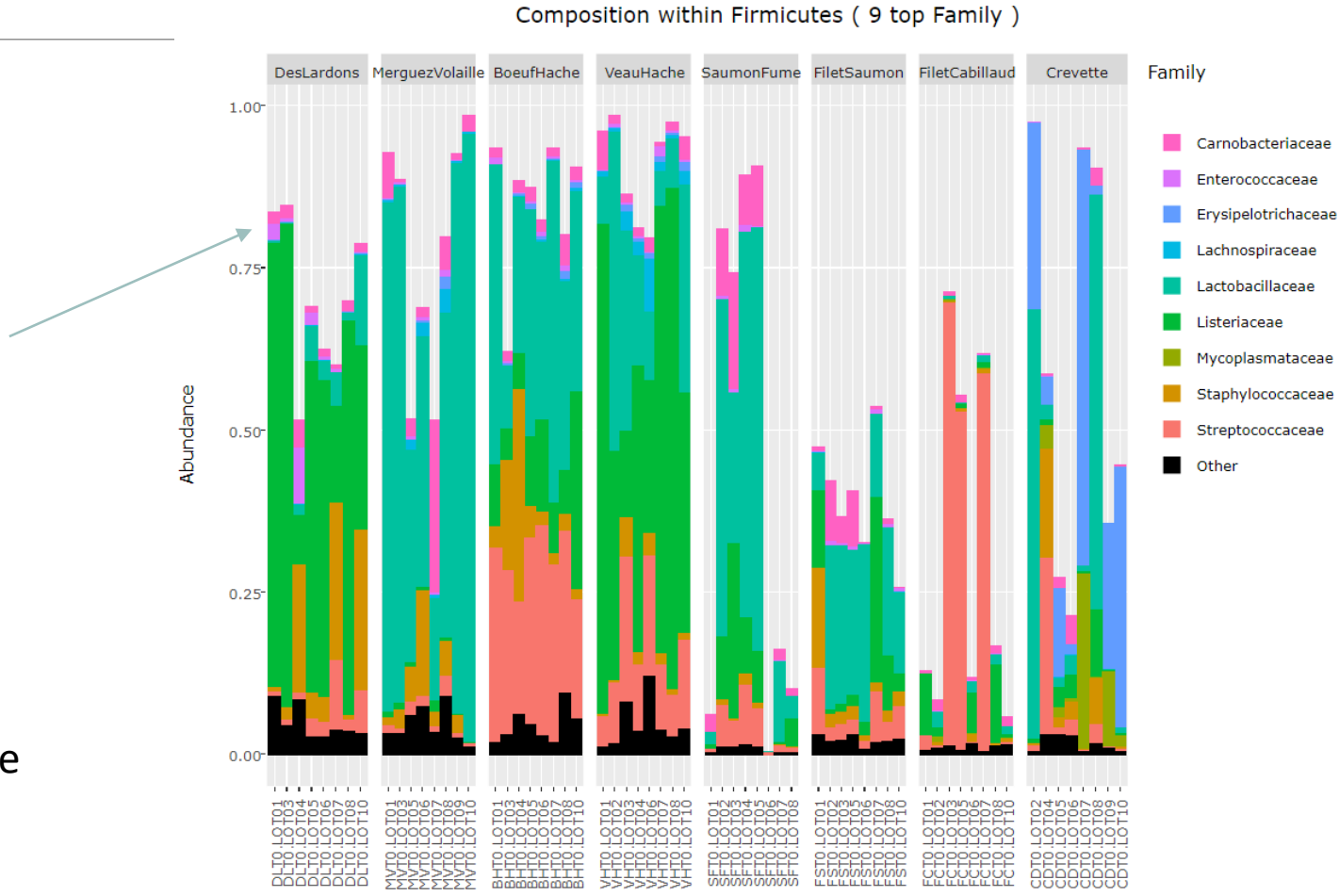
Composition within Firmicutes ( 9 top Family )



# Exercise 4

## 1. What is the composition of the 9 most abundant Families of Firmicutes ?

- Abundance does not reach 1 because only Phylum Firmicutes is displayed, the "missing" abundance is carried by other Phyla.
- As seen at the Phylum level, Firmicutes are more represented in meat types than in seafoods
- Dominant Firmicutes families are not the same in each food type



# Exercise 4

## 2. What is the composition of the 9 most abundant Families of Proteobacteria ?

### Taxonomic level to filter your data

Phylum

ex: Kingdom, Phylum, Class, Order, Family, Genus, Species

### Taxa (at the above taxonomic level) to keep in the dataset

Proteobacteria

ex: Bacteria (when filtering at the Kingdom level), Firmicutes (when filtering at the Phylum level). Multiple taxa (separated by a space) can be specified, i.e. Firmicutes Proteobacteria

### Taxonomic level used for aggregation

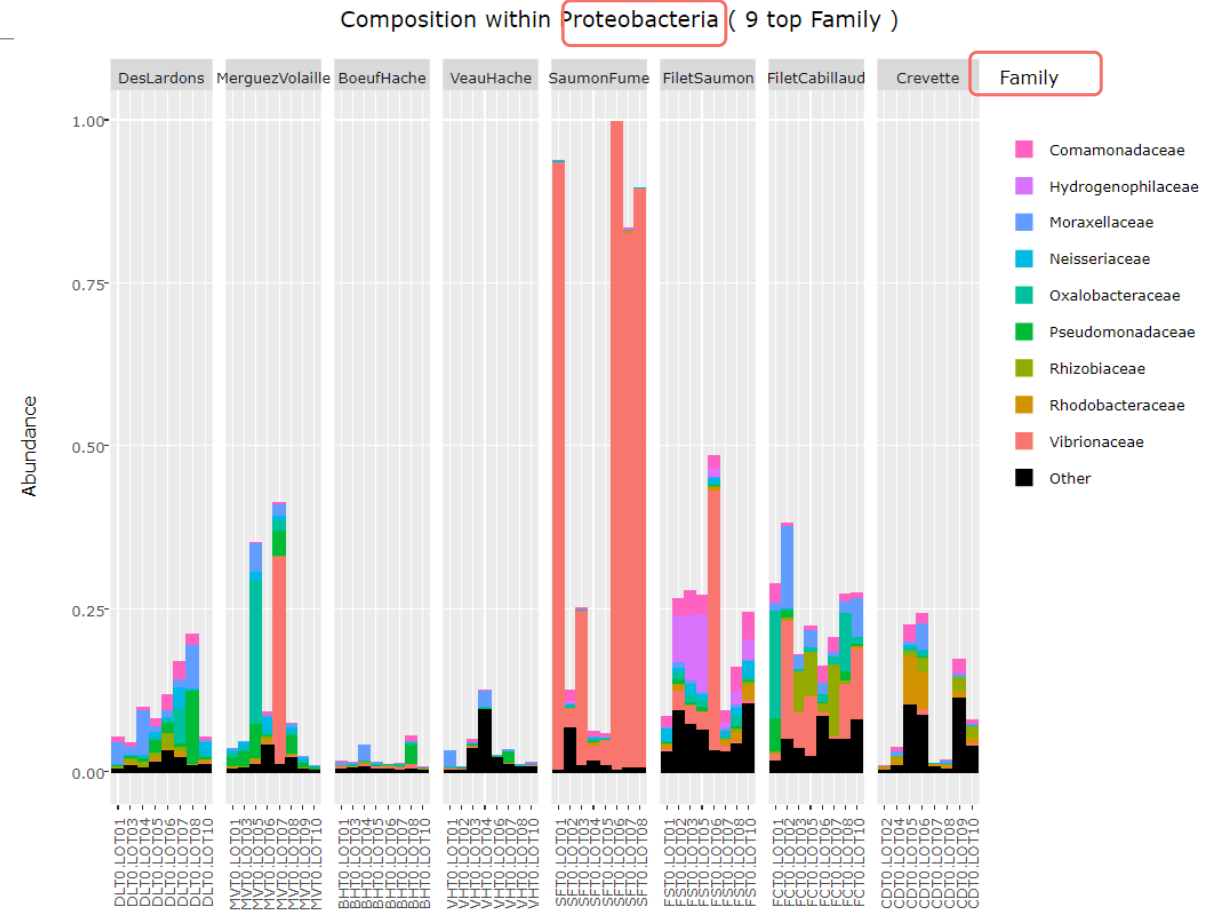
Family

ex: Family (when filtering at the Phylum level). The aggregation level must be below the filtering level.

### Number of most abundant taxa to keep

9

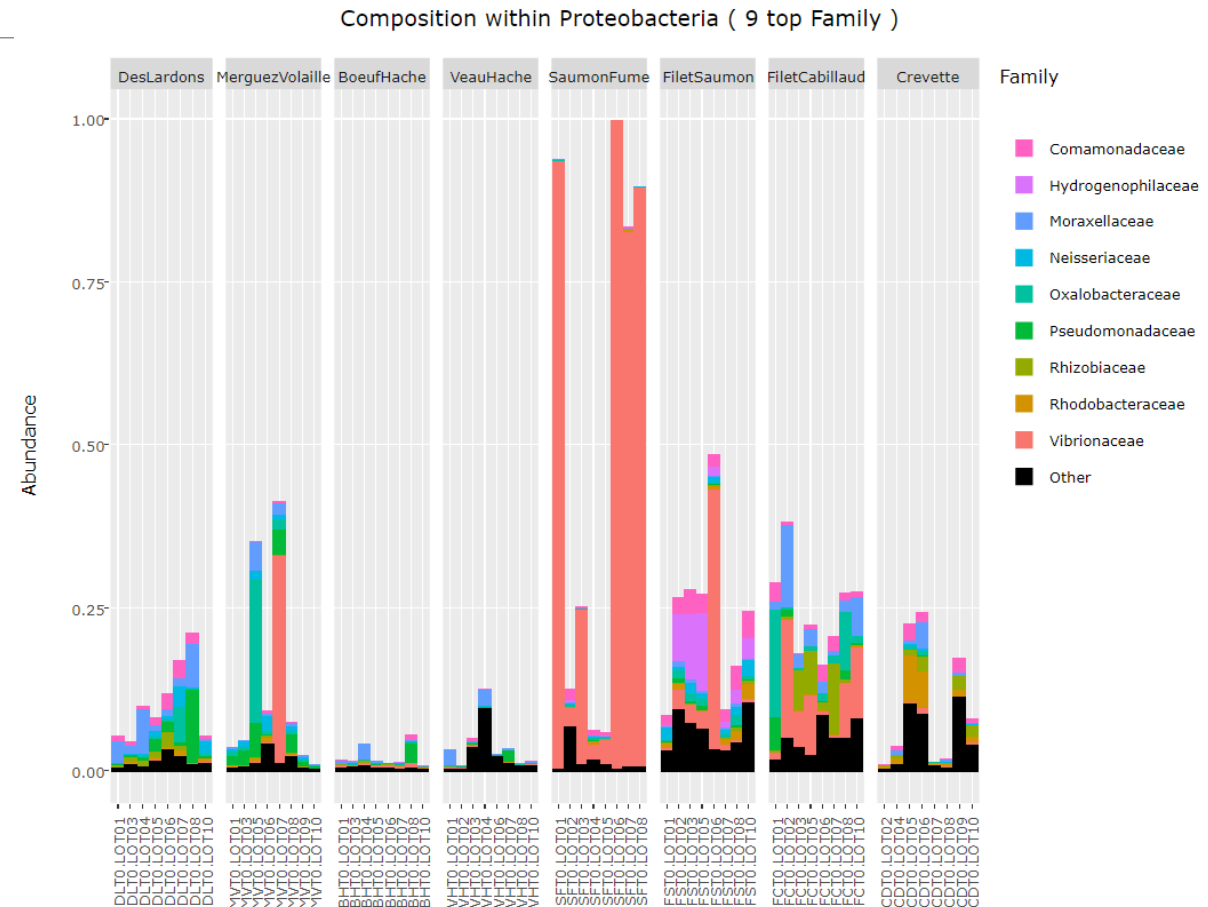
ex: 9, i.e. Tool keeps the 9 most abundant taxa and the remaining taxa are aggregated in a group 'Other'



# Exercise 4

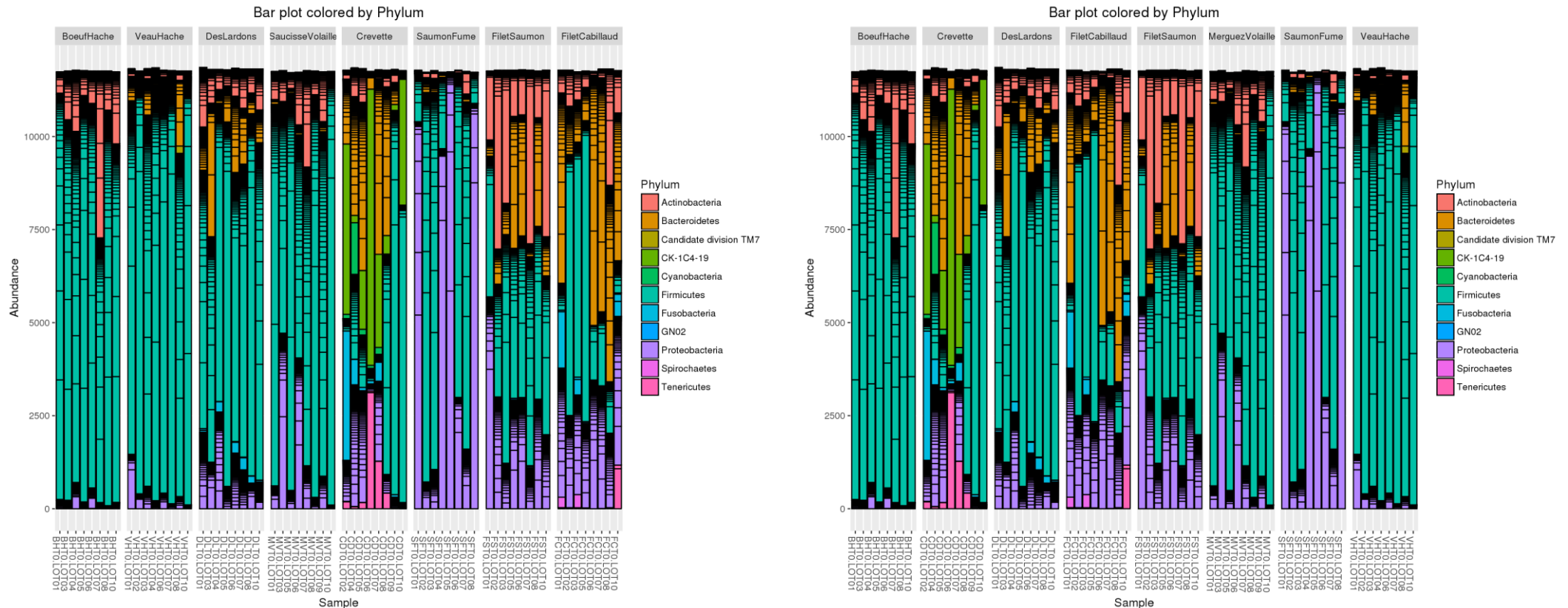
## 2. What is the composition of the 9 most abundant Families of Proteobacteria ?

- As seen at the Phylum level, Proteobacteria are particularly present in seafood samples
- SaumonFume samples with extremely high levels of Proteobacteria are dominated by Vibrionaceae family, while other food types are balanced between several families



# Exploring biodiversity : visualization

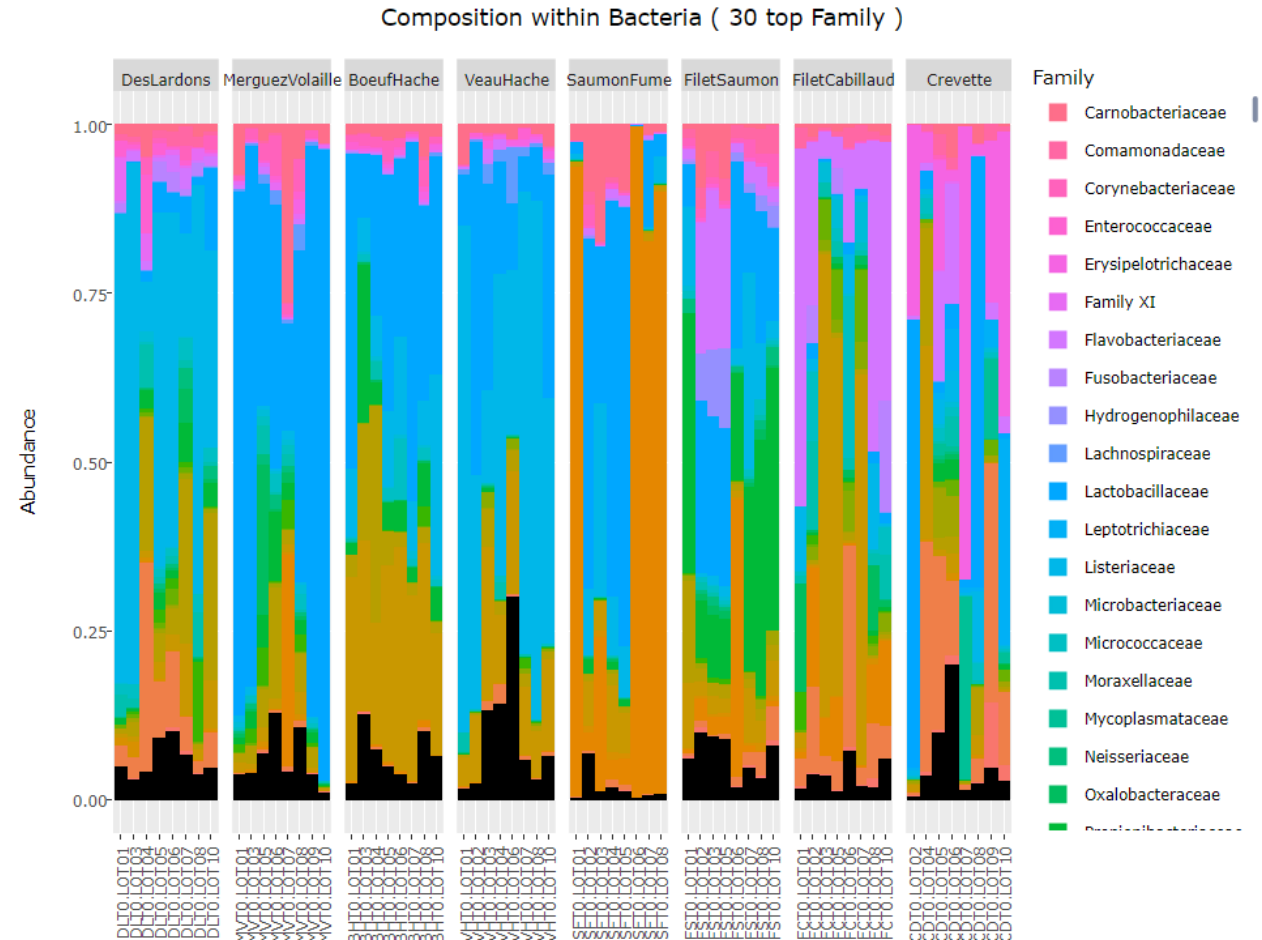
Remark 1 : An example of what happens when sample metadata file is not sorted in a meaningful way



# Exploring biodiversity : visualisation

Remark 2 : Keep in mind that human eye cannot distinguish more than 12 colors at the same time.

Example of the 30 most abundant Families among Bacteria



---

# II. Biodiversity analysis

---

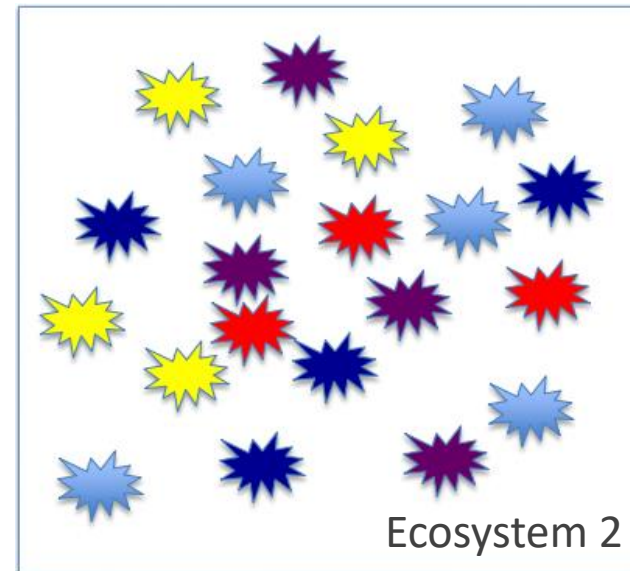
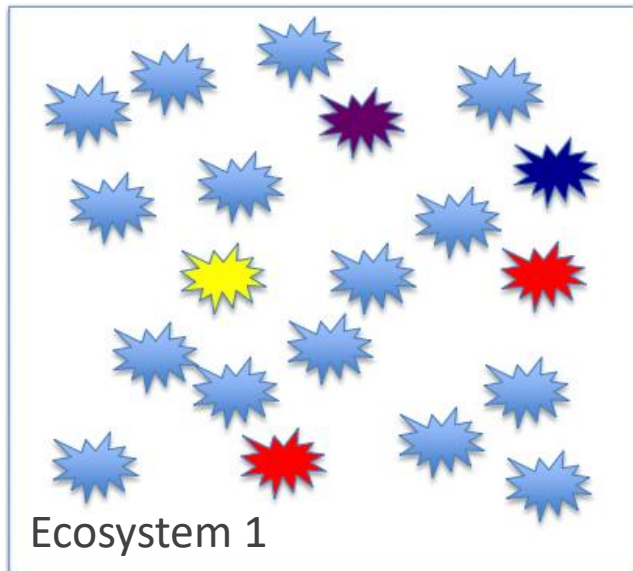
DIVERSITY INDICES



# Exploring biodiversity : descriptors

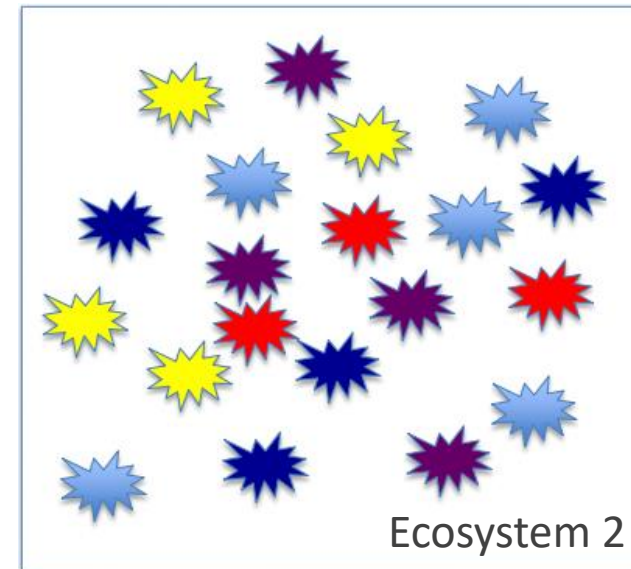
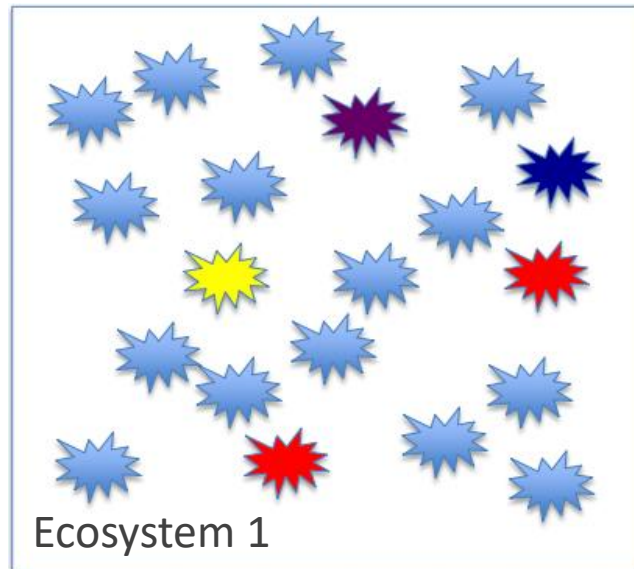
---

- The **richness** corresponds to the number of OTUs or functional groups present in communities. It characterizes the **composition**.
- The **diversity** takes into account the relative abundancy of species. It characterizes the **structure**



# Exploring biodiversity : descriptors

- The **richness** corresponds to the number of OTUs or functional groups present in communities. It characterizes the **composition**.
- The **diversity** takes into account the relative abundance of species. It characterizes the **structure**



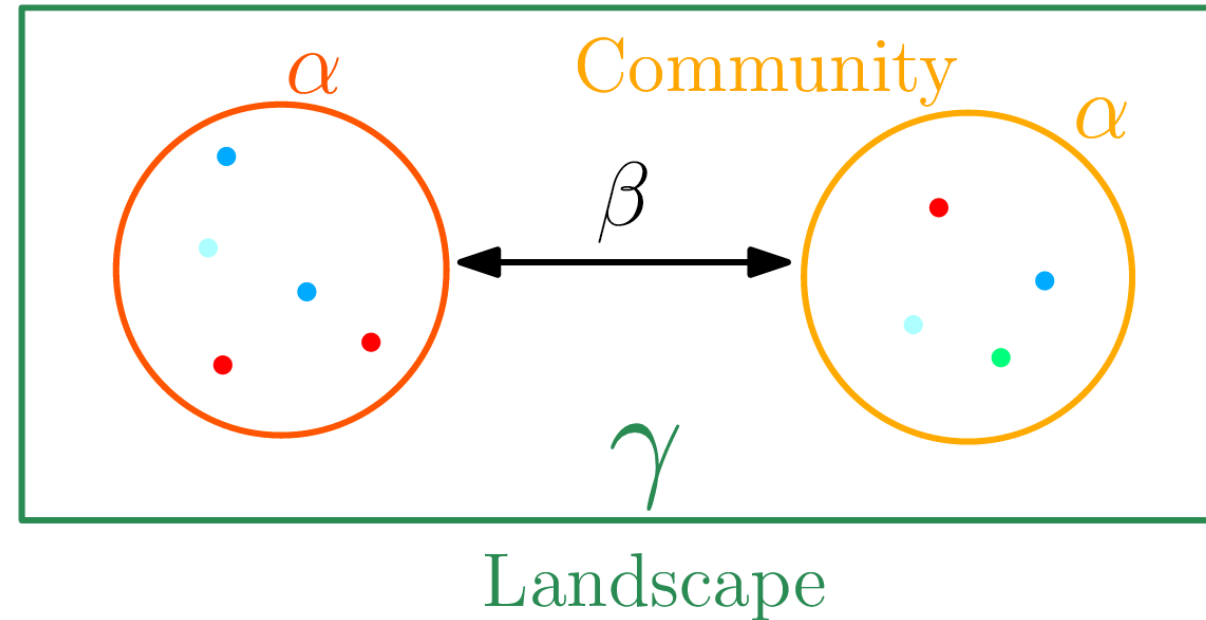
Richness : Eco1 = Eco2

Diversity: Eco2 > Eco1

# Exploring biodiversity : statistical indices

3 levels of diversity:

- **$\alpha$ -diversity**: diversity **within** a community
- **$\beta$ -diversity**: diversity **between** communities
  - $\beta$ -dissimilarities/distances
    - dissimilarities between pairs of communities
    - often used as a first step to compute diversity
- $\gamma$ -diversity: diversity at the landscape scale (blurry for bacterial communities)



# Exploring biodiversity : statistical indices

---

There are qualitative, quantitative and phylogenetic indices:

## Qualitative (Presence/Absence) vs. Quantitative (Abundance )

- Qualitative indices give equal weight to all species, dominant or rare
- Qualitative indices are more sensitive to differences in sampling depths
- Qualitative indices emphasize differences in taxa diversity while quantitative are more sensitive to increases in composition differences

## Phylogenetic indices

- Require a phylogenetic tree
- phylogeny allows to attenuate clustering errors because 2 different OTUs can be phylogenetically close

---

# III. Biodiversity analysis

---

## $\alpha$ -DIVERSITY INDICES

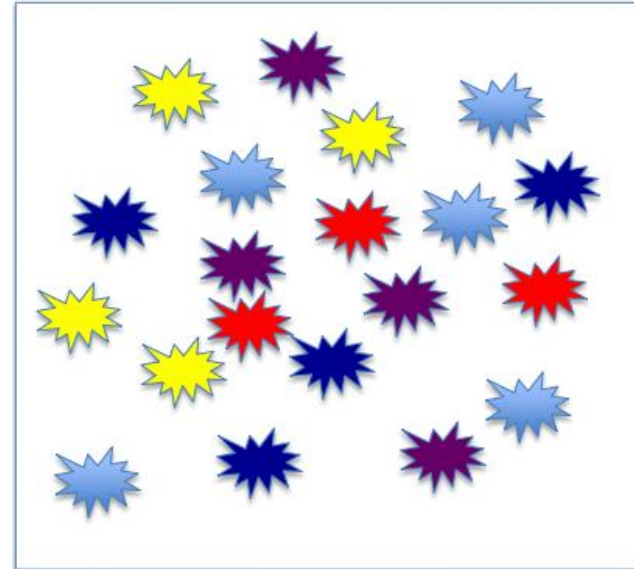
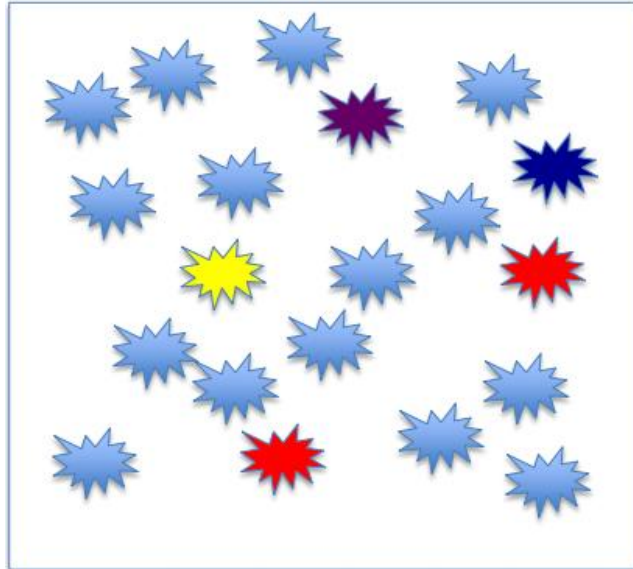
# 4 $\alpha$ -diversity indices

---

1. Richness
2. Chao
3. Shannon
4. Inv-Simpson

# Richness

---

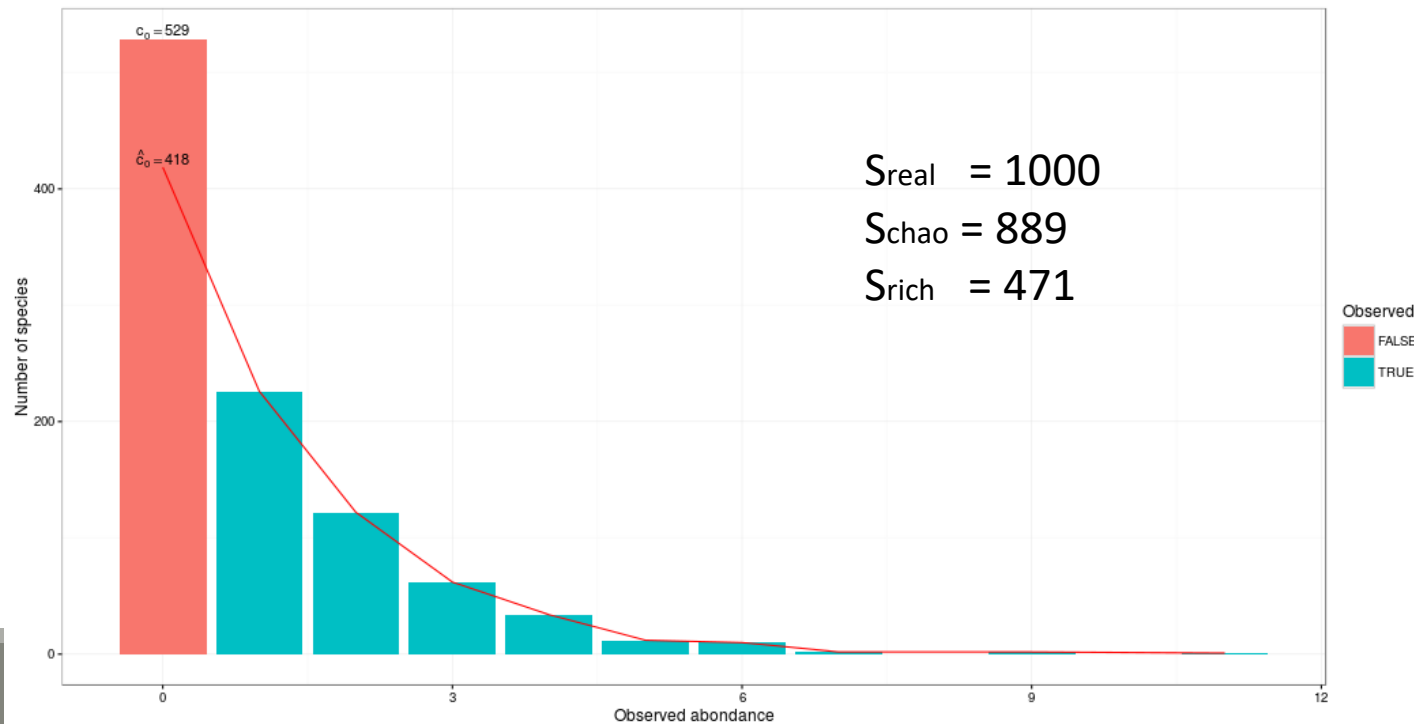


Richness : Eco1 = Eco2

Richness
Number of observed species

# $\alpha$ -diversity: Chao1

Richness	Chao
Number of observed species	Richness + (estimated) number of unobserved species





# $\alpha$ -diversity: Chao1

---

Chao1 is an abundance-based estimator. This means that the data it needs relate to the abundance of taxa in the sample.

This index **estimates the number of unobserved species** from those that have only been **observed once or twice**. This diversity index is a minimum estimator. In order for it to fit the dataset, it is necessary that singletons and duplicates represent a significant part of the information

Many taxa, species, are represented by a few individuals (rare species) and others can be represented by many individuals (abundant species).

Well, **chao1 is based on the rare species**.

So we need to know how many species are represented by **1 individual (singleton)** and how many species are represented by **2 individuals (doubletons)**:

$$S_{\text{est}} = S_{\text{obs}} + F^2/2G$$

$S_{\text{est}}$  (nb of species we want to estimate),  $S_{\text{obs}}$  (nb of species observed), F (nb of singletons) and G (nb of doubletons)

If the **chao1 is close to the richness** → the part of the missed OTUs is low → the sequencing depth is good.

# $\alpha$ -diversity: Chao1

Example of a abundance table, after FROGS processing, with OTUs filtering with 0.005% threshold:

observation_name	observation_sum	complexe-ADN-1	echantillon1-1	echantillon1-2	echantillon1-3	echantillon2-1	echantillon2-2	echantillon2-3
Cluster_1	298637	56	227	234	120	36754	59089	56534
Cluster_2	155012	688	20604	38077	45508	8417	10464	10655
Cluster_3	52753	2469	14	76	68	37	8	19
Cluster_4	34062	3459	5041	11458	12799	0	37	84
Cluster_5	30263	3	10	13	13	570	806	800
Cluster_6	26805	1301	7	51	35	21	6	16
Cluster_7	25237	1015	7	30	34	16	5	14
Cluster_8	20483	893	6	34	19	18	1	16
Cluster_9	26069	2504	32	60	87	26	7	22
Cluster_10	17383	712	5	23	17	19	8	13
Cluster_11	16674	715	6	27	25	26	2	7
Cluster_12	11420	0	37	76	79	19	24	13
Cluster_13	9414	189	0	24	12	6	0	8
Cluster_14	7972	498	3	7	11	7	3	5
Cluster_15	7267	13	0	19	12	11	2	7
Cluster_16	7131	150	3	8	15	11	0	2
Cluster_17	6407	4953	22	7	1	0	13	4
Cluster_18	6538	28	1	10	18	16	0	6
Cluster_19	5633	3	12	12	45	24	0	3
Cluster_20	5223	183	0	5	12	8	1	1
Cluster_21	4078	12	0	6	9	6	0	4
Cluster_22	4507	0	10	13	20	13	0	2
Cluster_23	4232	3	0	10	8	9	0	4
Cluster_24	3404	160	1	4	6	4	1	0
Cluster_25	3857	1	0	3	6	10	0	2
Cluster_26	2616	1926	16	12	9	2	8	9
Cluster_27	2781	2182	7	2	0	0	6	1

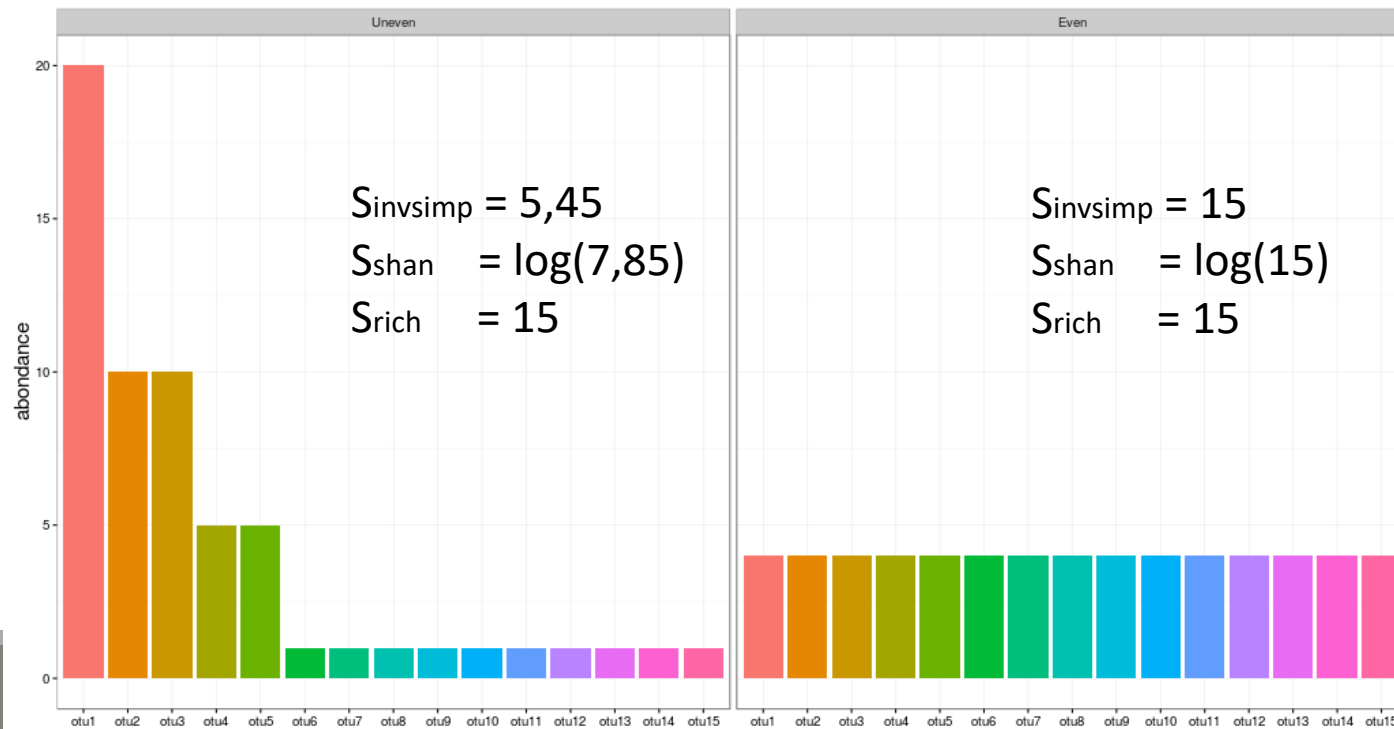
singletons  
and  
doubletons

→ Chao1 computation possible

# $\alpha$ -diversity: Shannon and Inv-Simpson

$\alpha$ -diversity is equivalent to the richness : number of species

Shannon	Inv-Simpson
Evenness of the species abundance distribution	Inverse probability that two sequences sampled at random come from the same species



Interpretation :

15 observed species, but according to Shannon, the uneven community acts like there is 7.85 equally abundant species (5.45 for invSimp)

# $\alpha$ -diversity indices

---

1. Chao1 close to Richness  $\rightarrow$  all species have been detected
2. higher Shannon index  $\rightarrow$  higher homogeneity  $\rightarrow$  greater diversity
3. greater invsimpson index  $\rightarrow$  greater diversity

# Exploring biodiversity : $\alpha$ -diversity

---

$\alpha$ -diversity indices available in phyloseq :

- Species **richness** : number of observed OTU
- **Chao1** : number of observed OTU + estimation of the number of unobserved OTU
- **Shannon** entropy / **Jensen** : the width of the OTU relative abundance distribution. Roughly, it reflects our (in)ability to predict OTU of a randomly picked bacteria.
- **Simpson** :  $1 -$  probability that two bacteria picked at random in the community belong to different OTU
- **Inverse Simpson** : inverse of the probability that two bacteria picked at random belong to the same OTU
- Other estimators of alpha diversity exist (Chao2, ACE, ICE,...), however the indices presented above allow us to understand alpha diversity with sufficient precision

# Exploring biodiversity : $\alpha$ -diversity

**FROGSSTAT Phyloseq Alpha Diversity** with richness plot (Galaxy Version 3.2.2) Options

**Phyloseq object (format rdata)**

28: Phyloseq\_raref.Rdata

This file is the result of FROGS Phyloseq Import Data tool.

**Experiment variable**

EnvType

The experiment variable that you want to analyse.

**The alpha diversity indices to compute**

Select/Unselect all

- Observed
- Chao1
- Shannon
- InvSimpson
- Simpson
- ACE
- Fisher

Explore the sample **NORMALISED** count

Choose a sample variable to organize graphics test on **EnvType**

Choose which  $\alpha$ -diversity indices you want to compute

# Exercise 5

---

1. What are the output files ?
2. Which interpretation could you make on the boxplot results ?
3. Does EnvType has an impact on  $\alpha$ -diversity indices ?

# Exercise 5

---

1. What are the output files ?

→ Tabular file: contains the detailed value of indices in each sample

→ HTML report: graphical and statistical results



# Exercise 5

---

## 1. What are the output files ?

→ Tabular file: contains the detailed value of indices in each sample

1	2	3	4	5	6
	Observed	Chao1	se.chao1	Shannon	InvSimpson
BHT0.LOT01	89	90.875	2.25640704112416	2.46283438240559	6.4374614755645
BHT0.LOT03	129	134.2	3.98819923457003	3.01399812576966	11.6378947553209
BHT0.LOT04	137	152	8.65612088483201	2.77419314445453	7.04904738429417
BHT0.LOT05	127	132.526315789474	3.97261840192821	2.82922278153272	7.54330476122993
BHT0.LOT06	135	136	1.30982775947977	2.6365904270666	6.30810073317464
BHT0.LOT07	126	141.260869565217	7.7960250320146	2.36922299088995	5.65591172677601
BHT0.LOT08	172	189.652173913043	8.66767047151361	3.32220303923076	11.229239617499
BHT0.LOT10	155	173.9	9.42281349646639	2.96129964607031	7.55645792419119
CDT0.LOT02	73	87.5263157894737	7.85749286229502	0.968874997875041	1.93691052993399
CDT0.LOT04	145	168.25	10.9999446485673	3.1208274916296	11.0298385276267

# Exercise 5

---

1. What are the output files ?

→ HTML report: graphical and statistical results

# Exercise 5

Richness plot

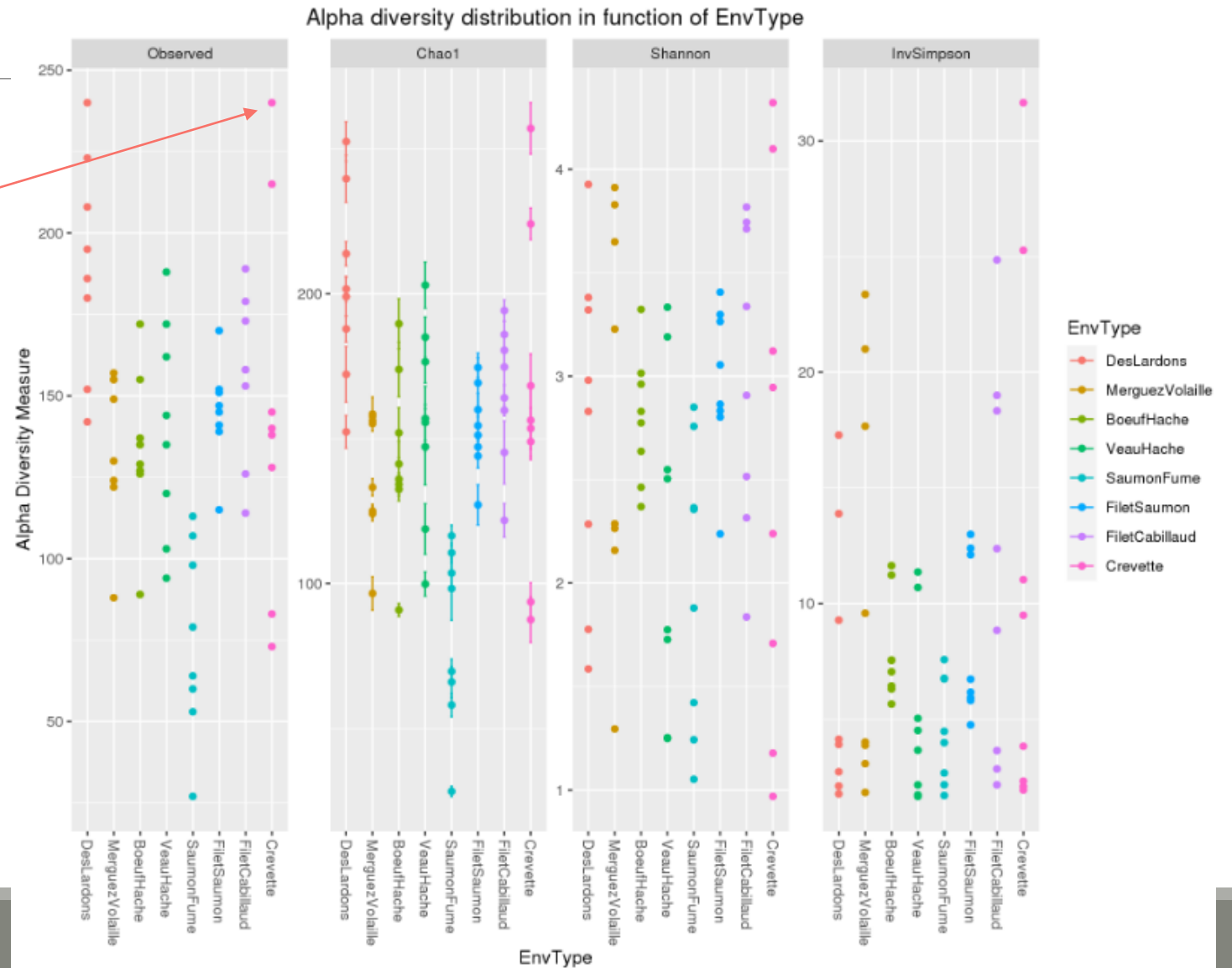
Richness plot with boxplot

Alpha Diversity Indice Anova Analysis

Rarefaction curves

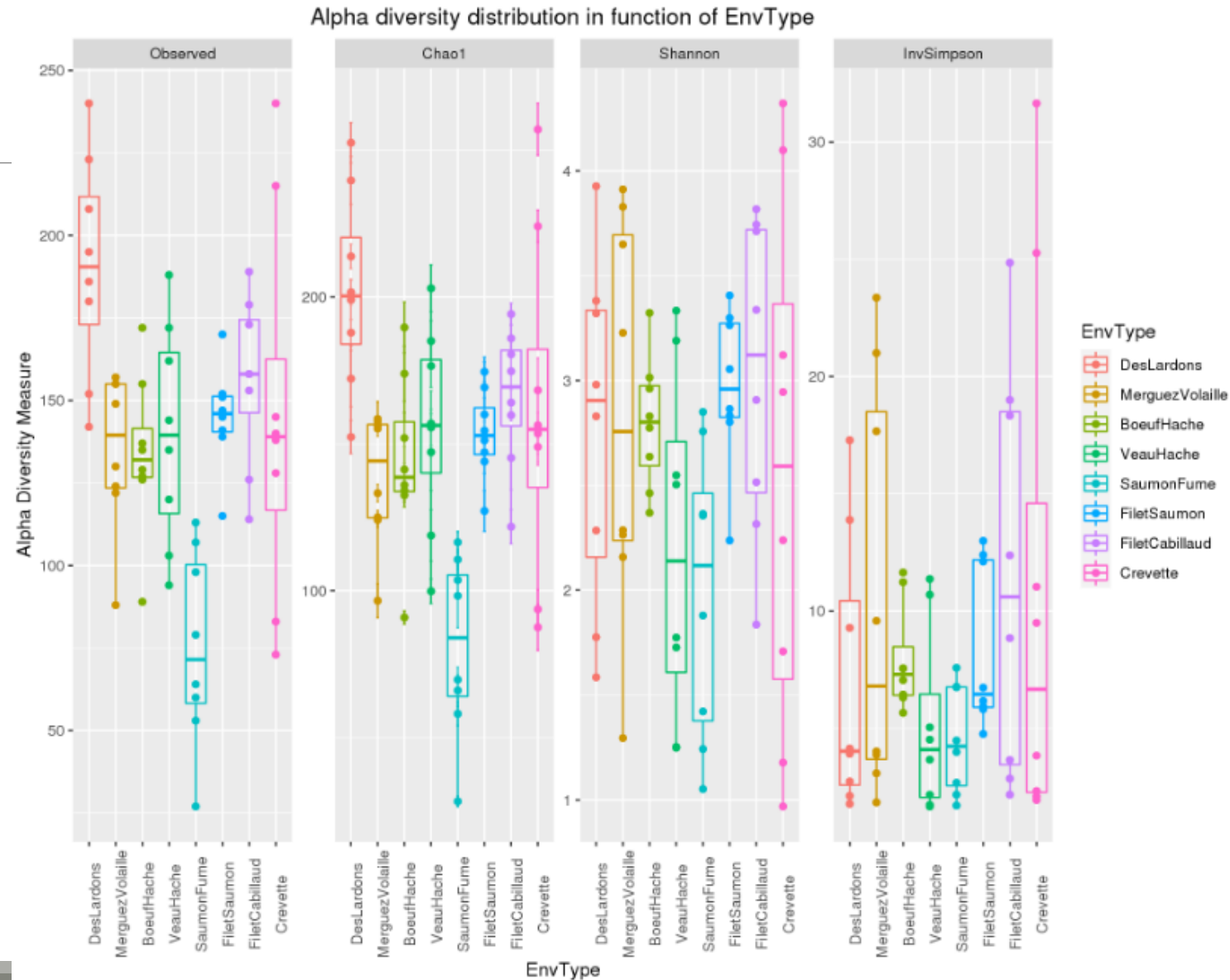
1 dot = 1 sample

One graph per asked indice




# Exercise 5

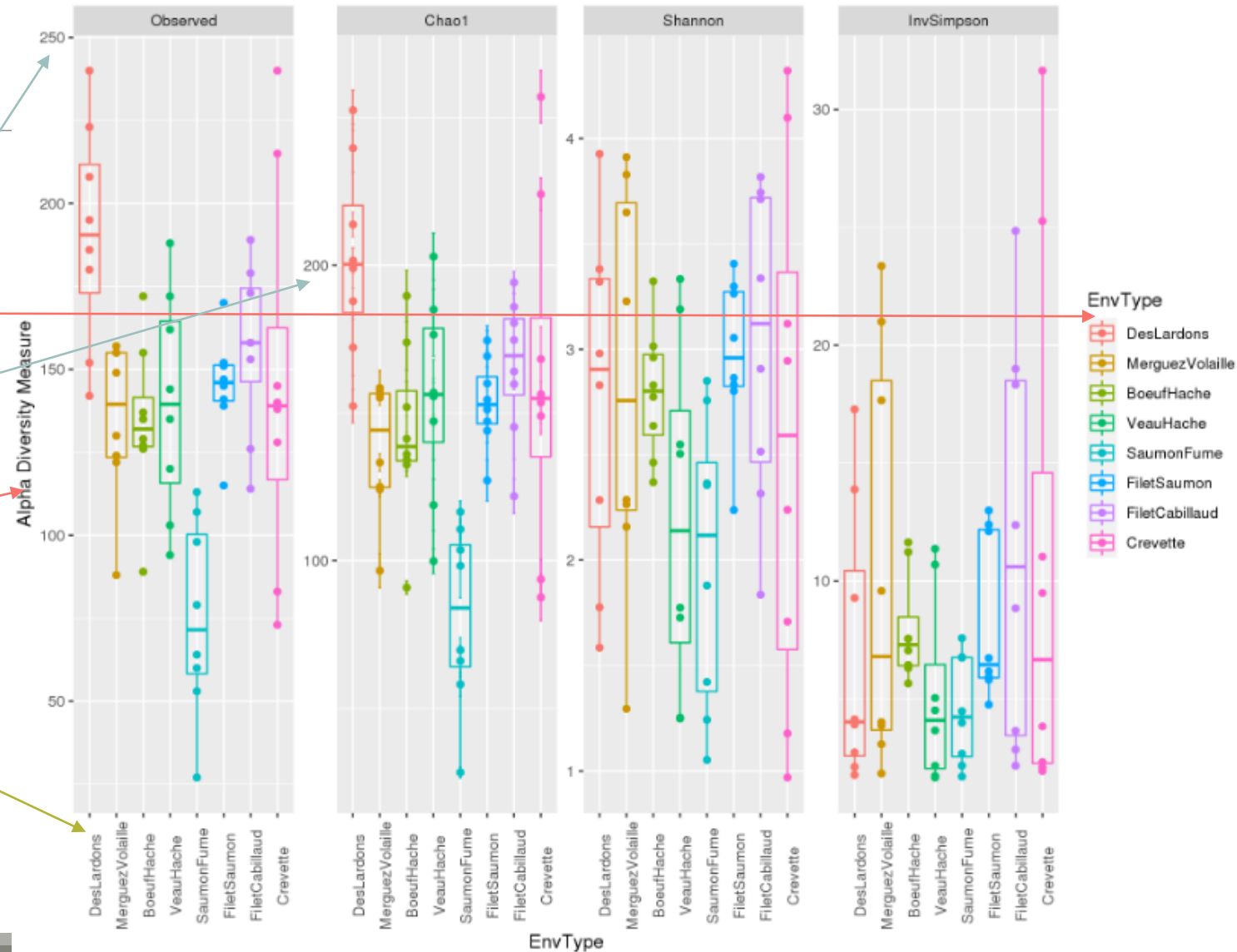
more readable thanks to boxplots



# Exercise 5

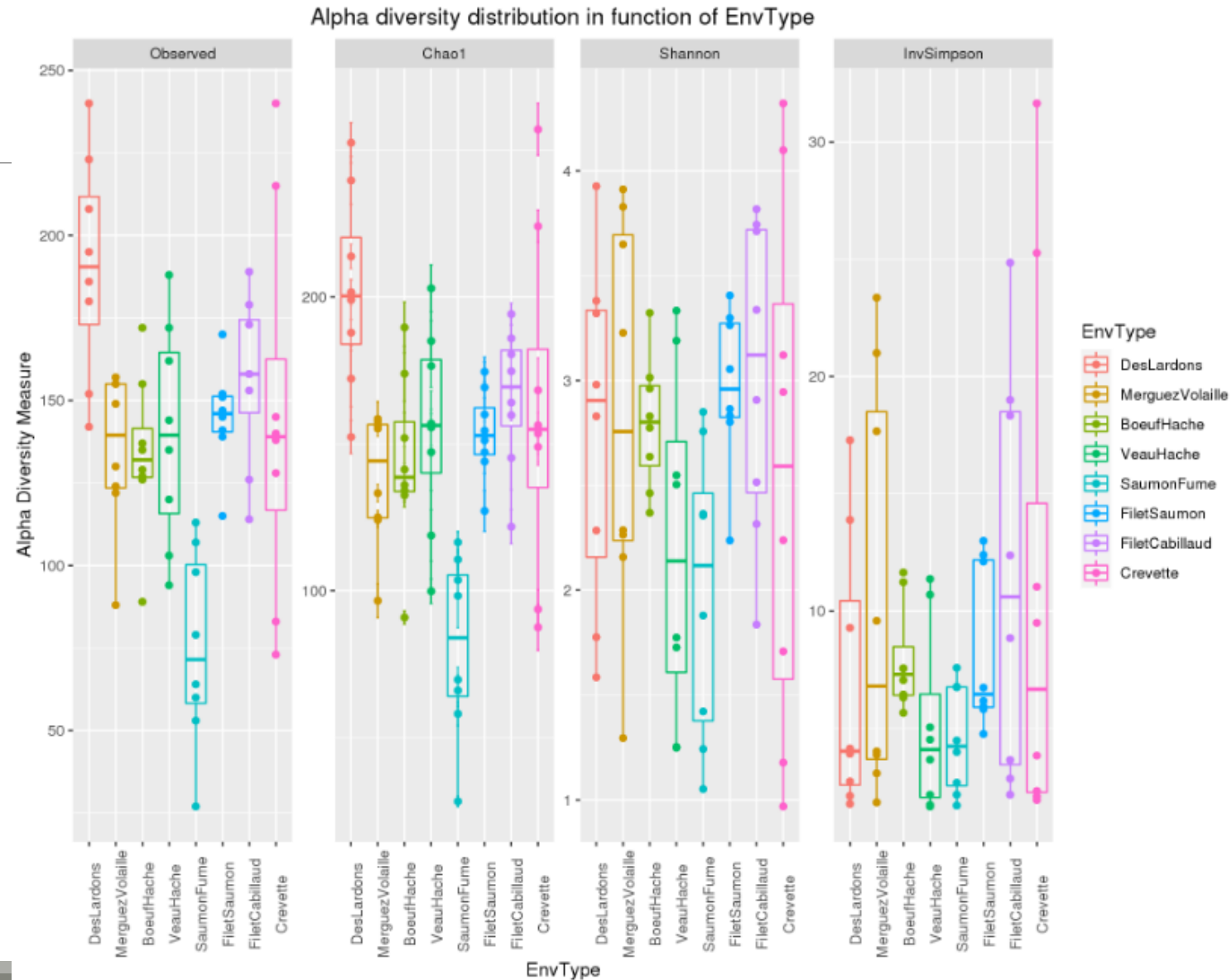
- Same legend for all indices
-  Scales in y axis are different
- y axis: values of each alpha index
- x axis: 8 boxplots for each indices (4 indices, 8 EnvTypes)

Alpha diversity distribution in function of EnvType



# Exercise 5

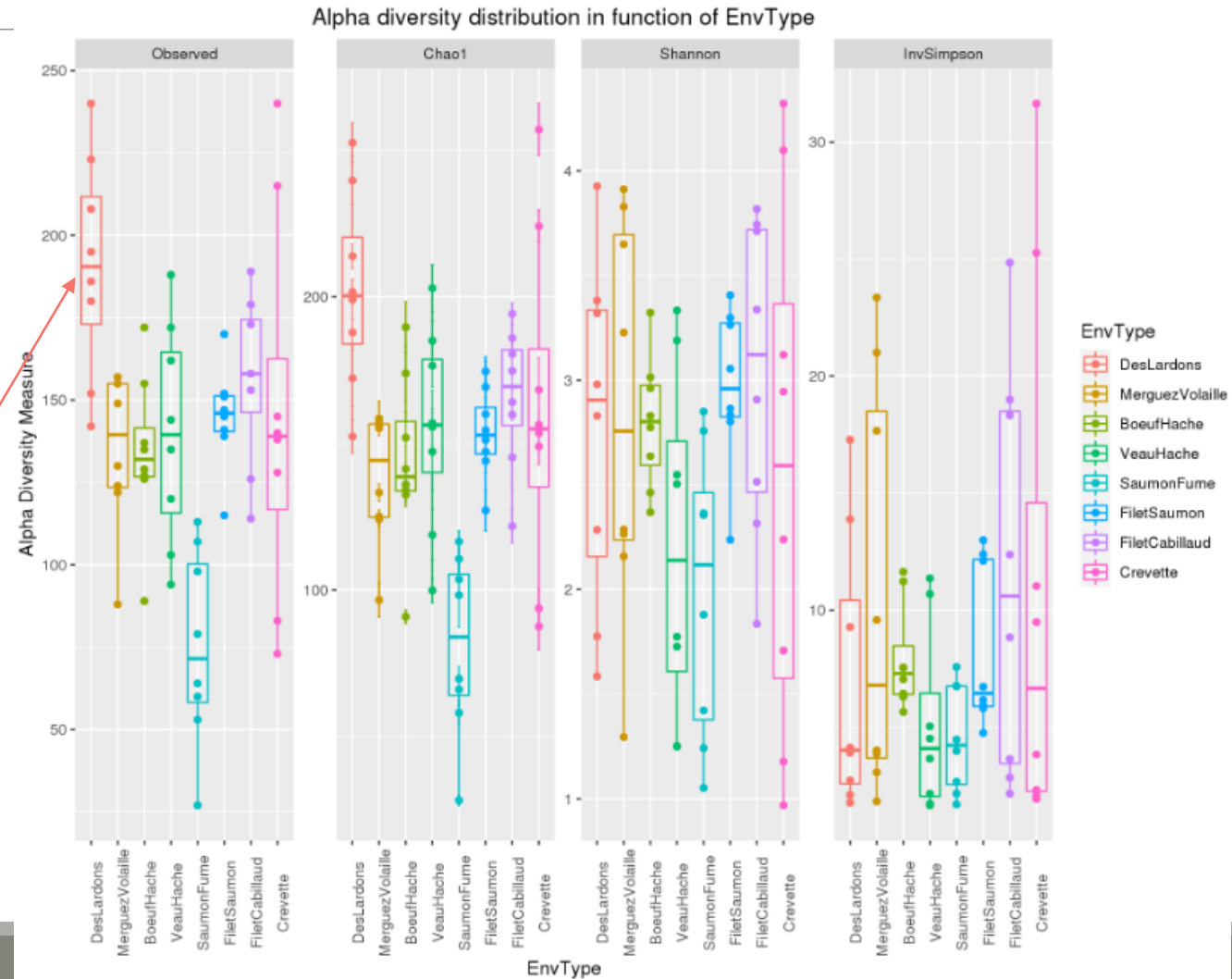
2. Which interpretation could you make on the boxplot results ?



# Exercise 5

2. Which interpretation could you make on the boxplot results ?

- Same image in same scale for Richness and Chao1  
→ all species have been detected
- High variability in the number of OTUs per EnvType
- Many taxa observed in **DesLardons** (highest observed richness)
- Most foods have low effective diversities (Shannon & InvSimpson)  
→ communities are dominated by few abundant taxa



# Exercise 5

Richness plot

Richness plot with boxplot

Alpha Diversity Indices Anova Analysis

Rarefaction curves



3. Does EnvType has an impact on  $\alpha$ -diversity indices ?

- What is an ANOVA used for?

→ Test the significance of the previous observations by performing an ANOVA of alpha-diversity

indices against the covariate of interest (EnvType)



# Exercise 5

## 3. Does EnvType has an impact on $\alpha$ -diversity indices ?

### Anova interpretations

Richness plot

Richness plot with boxplot

Alpha Diversity Indices Anova Analysis

Rarefaction curves

```
#####  
#Perform ANOVA on Observed, which effects are significant  
anova.Observed <-aov( Observed ~ Depth + EnvType, anova_data)  
summary(anova.Observed)  
              Df Sum Sq Mean Sq F value    Pr(>F)  
EnvType      7  57320     8189   7.731 1.61e-06 ***  
Residuals   56  59312     1059  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Chao1, which effects are significant  
anova.Chao1 <-aov( Chao1 ~ Depth + EnvType, anova_data)  
summary(anova.Chao1)  
              Df Sum Sq Mean Sq F value    Pr(>F)  
EnvType      7  64366     9195   8.446 5.14e-07 ***  
Residuals   56  60971     1089  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Shannon, which effects are significant  
anova.Shannon <-aov( Shannon ~ Depth + EnvType, anova_data)  
summary(anova.Shannon)  
              Df Sum Sq Mean Sq F value Pr(>F)  
EnvType      7    7.61  1.0878   1.696  0.129  
Residuals   56   35.92  0.6414
```

```
#####  
#Perform ANOVA on InvSimpson, which effects are significant  
anova.InvSimpson <-aov( InvSimpson ~ Depth + EnvType, anova_data)  
summary(anova.InvSimpson)  
              Df Sum Sq Mean Sq F value Pr(>F)  
EnvType      7   392.4   56.06   1.264  0.285  
Residuals   56 2484.3   44.36
```

# Exercise 5

3. Does EnvType has an impact on  $\alpha$ -diversity indices ?

## Anova interpretations

Does the EnvType have an effect on Observed indice ?

```
#####  
#Perform ANOVA on Observed, which effects are significant  
anova.Observed <-aov( Observed ~ Depth + EnvType, anova_data)  
summary(anova.Observed)  
              Df Sum Sq Mean Sq F value    Pr(>F)  
EnvType      7  57320    8189   7.731 1.61e-06 ***  
Residuals   56  59312    1059  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Chao1, which effects are significant  
anova.Chaol <-aov( Chaol ~ Depth + EnvType, anova_data)  
summary(anova.Chaol)  
              Df Sum Sq Mean Sq F value    Pr(>F)  
EnvType      7  64366    9195   8.446 5.14e-07 ***  
Residuals   56  60971    1089  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Shannon, which effects are significant  
anova.Shannon <-aov( Shannon ~ Depth + EnvType, anova_data)  
summary(anova.Shannon)  
              Df Sum Sq Mean Sq F value Pr(>F)  
EnvType      7    7.61  1.0878   1.696  0.129  
Residuals   56   35.92  0.6414
```

```
#####  
#Perform ANOVA on InvSimpson, which effects are significant  
anova.InvSimpson <-aov( InvSimpson ~ Depth + EnvType, anova_data)  
summary(anova.InvSimpson)  
              Df Sum Sq Mean Sq F value Pr(>F)  
EnvType      7   392.4   56.06   1.264  0.285  
Residuals   56 2484.3   44.36
```

# Exercise 5

## 3. Does EnvType has an impact on $\alpha$ -diversity indices ?

### Anova interpretations

- Environments differ in terms of richness but not in terms of Shannon and InvSimpson diversity
- This means that all EnvTypes have similar structures (equivalent distributions between several minor OTUs and few dominant OTUs). Even if 2 samples of "Crevette" displayed very high invSimpson (their bacteria were thus more homogeneously distributed), these two samples were not sufficient to make "Crevette" significantly different from the others EnvType.

→ There is no significant difference between the EnvType

```
#####  
#Perform ANOVA on Observed, which effects are significant  
anova.Observed <-aov( Observed ~ Depth + EnvType, anova_data)  
summary(anova.Observed)  
          Df Sum Sq Mean Sq F value Pr(>F)  
EnvType   7  57320    8189   7.731 1.61e-06 ***  
Residuals 56  59312    1059  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Chao1, which effects are significant  
anova.Chaol <-aov( Chaol ~ Depth + EnvType, anova_data)  
summary(anova.Chaol)  
          Df Sum Sq Mean Sq F value Pr(>F)  
EnvType   7  64366    9195   8.446 5.14e-07 ***  
Residuals 56  60971    1089  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Shannon, which effects are significant  
anova.Shannon <-aov( Shannon ~ Depth + EnvType, anova_data)  
summary(anova.Shannon)  
          Df Sum Sq Mean Sq F value Pr(>F)  
EnvType   7    7.61  1.0878   1.696  0.129  
Residuals 56   35.92  0.6414
```

```
#####  
#Perform ANOVA on InvSimpson, which effects are significant  
anova.InvSimpson <-aov( InvSimpson ~ Depth + EnvType, anova_data)  
summary(anova.InvSimpson)  
          Df Sum Sq Mean Sq F value Pr(>F)  
EnvType   7   392.4   56.06   1.264  0.285  
Residuals 56 2484.3   44.36
```

# Exercise 5

## 3. Does EnvType has an impact on $\alpha$ -diversity indices ?

### Anova interpretations

- Depth does not appear in the results, so there is no effect of depth.
- This is expected as the sequencing depth is equivalent between samples
- If Depth appears as a significant effect, you should normalize

```
#####  
#Perform ANOVA on Observed, which effects are significant  
anova.Observed <-aov( Observed ~ Depth + EnvType, anova_data)  
summary(anova.Observed)  
              Df Sum Sq Mean Sq F value    Pr(>F)  
EnvType      7  57320    8189   7.731 1.61e-06 ***  
Residuals   56  59312    1059  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Chao1, which effects are significant  
anova.Chaol <-aov( Chaol ~ Depth + EnvType, anova_data)  
summary(anova.Chaol)  
              Df Sum Sq Mean Sq F value    Pr(>F)  
EnvType      7  64366    9195   8.446 5.14e-07 ***  
Residuals   56  60971    1089  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#####  
#Perform ANOVA on Shannon, which effects are significant  
anova.Shannon <-aov( Shannon ~ Depth + EnvType, anova_data)  
summary(anova.Shannon)  
              Df Sum Sq Mean Sq F value Pr(>F)  
EnvType      7   7.61  1.0878   1.696  0.129  
Residuals   56  35.92  0.6414
```

```
#####  
#Perform ANOVA on InvSimpson, which effects are significant  
anova.InvSimpson <-aov( InvSimpson ~ Depth + EnvType, anova_data)  
summary(anova.InvSimpson)  
              Df Sum Sq Mean Sq F value Pr(>F)  
EnvType      7  392.4   56.06   1.264  0.285  
Residuals   56 2484.3   44.36
```

# Exercise 5

Richness plot

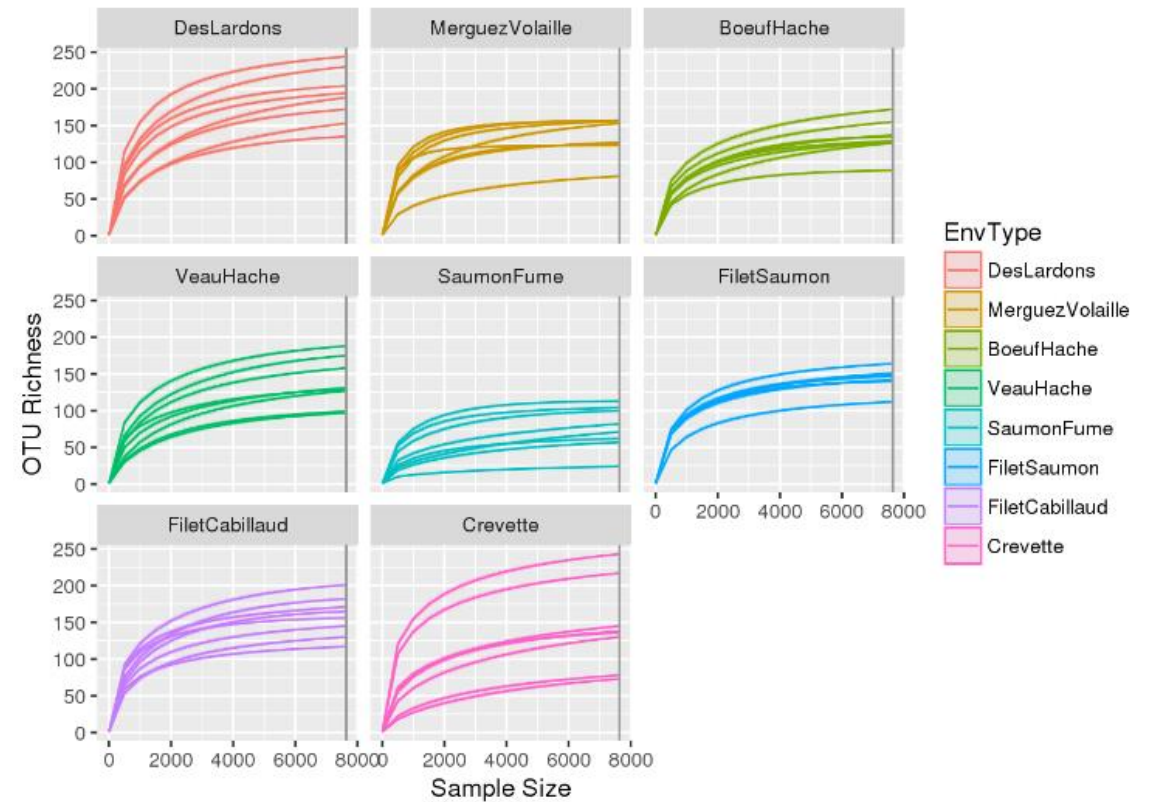
Richness plot with boxplot

Alpha Diversity Indice Anova Analysis

Rarefaction curves



## Rarefaction curve interpretations



# Exercise 5

Richness plot

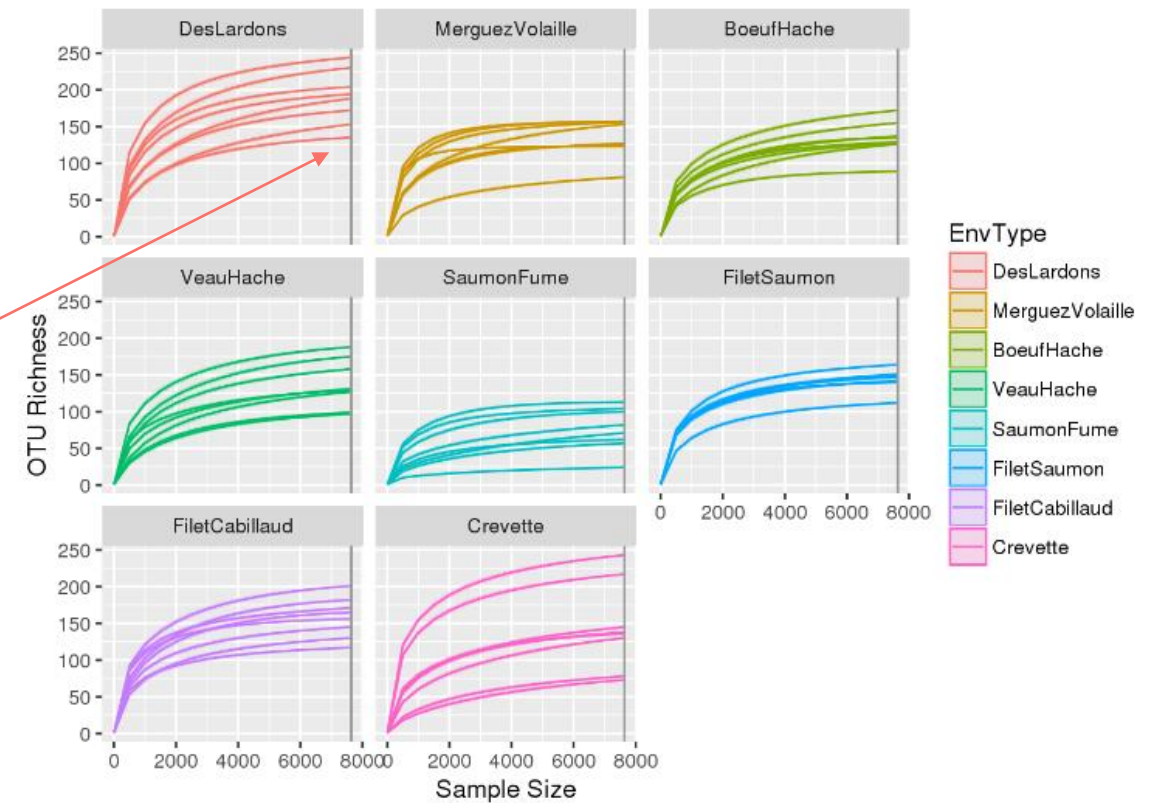
Richness plot with boxplot

Alpha Diversity Indice Anova Analysis

Rarefaction curves

## Rarefaction curve interpretations

- Most of the curves reach a plateau
- A deeper sequencing doesn't add more OTUs
- DesLardons reach the plateau later which correspond to a higher Observed





---

# IV. Biodiversity analysis

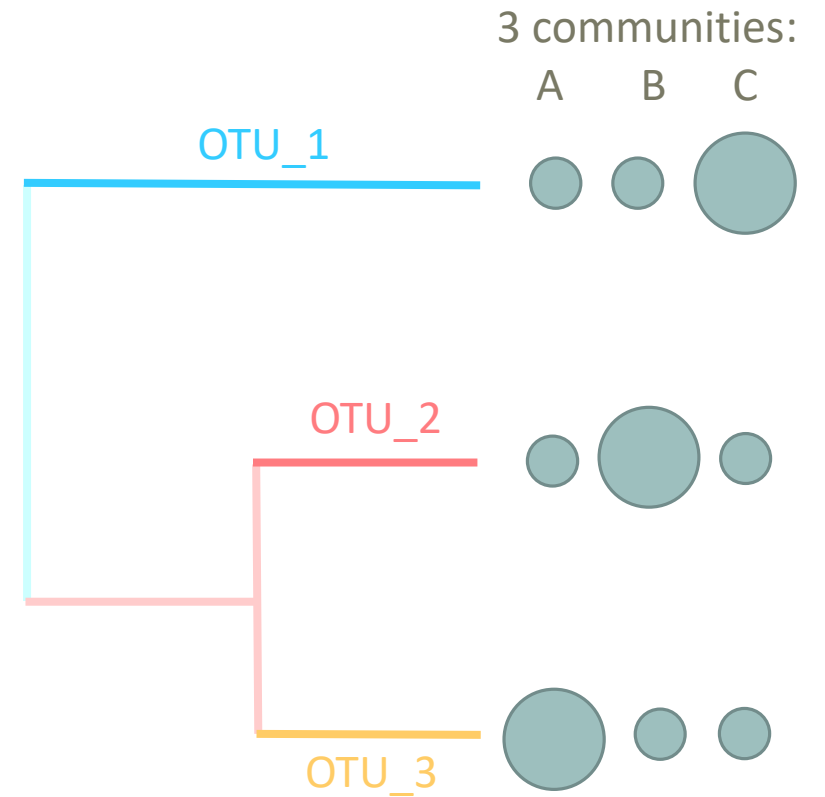
---

## $\beta$ -DIVERSITY INDICES

# Exploring biodiversity : $\beta$ -diversity

Many diversity indices are available with the Phyloseq package through the generic distance function.

Different dissimilarities capture different features of the communities.





# Exploring biodiversity : $\beta$ -diversity

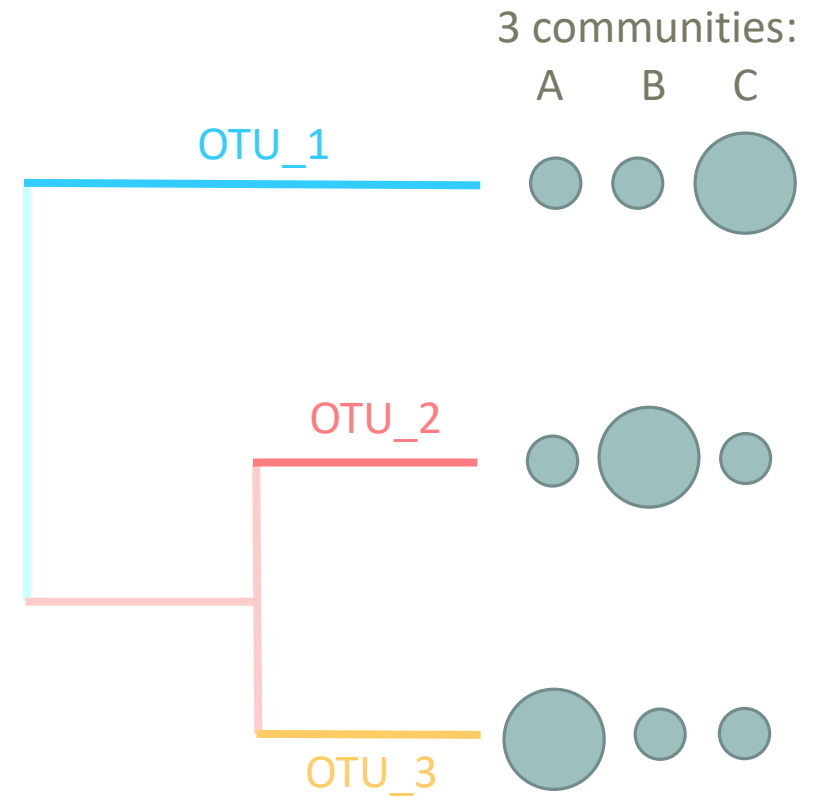
There are different ways to measure beta diversity on a dataset, which give different results.

In this example, 3 ways :

- qualitatively, communities are very similar
- quantitatively, communities are very different
- phylogenetically, two communities seem to be closer than the third one.

Which distance to choose?

- No wrong answer. Each beta-diversity indices will characterize communities differently



# Exploring biodiversity : $\beta$ -diversity

---

If we compare 2 communities A and B:

## Jaccard index:

- Fraction of species specific to either A or B → qualitative index

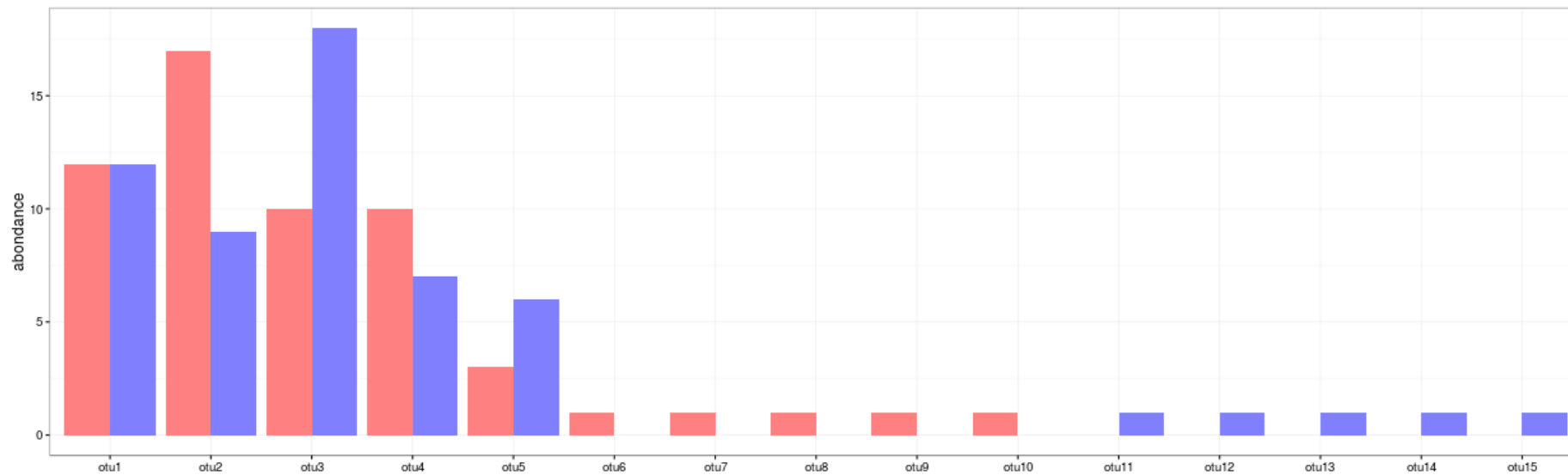
## Bray-Curtis index:

- Fraction of the community specific to either A or B → quantitative index

# Exploring biodiversity : $\beta$ -diversity

---

- 2 communities, Red and Blue
- 15 OTUs with different abundances in Red community and Blue community

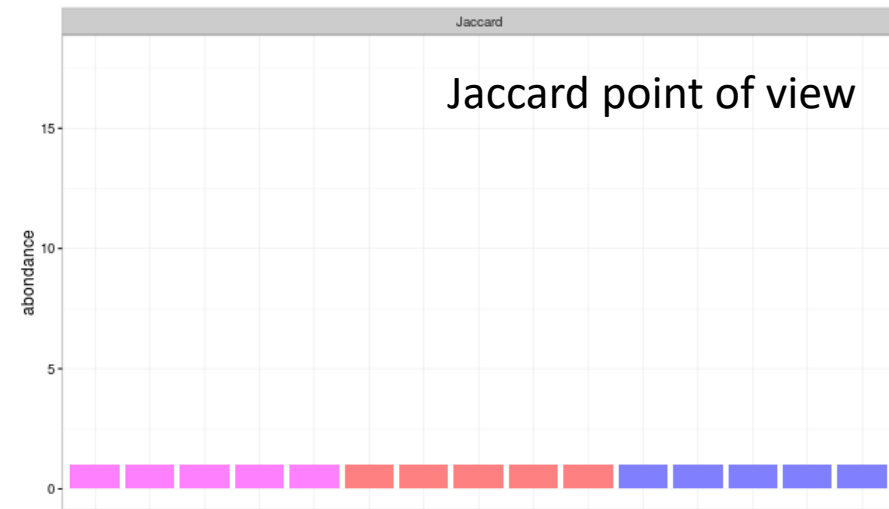
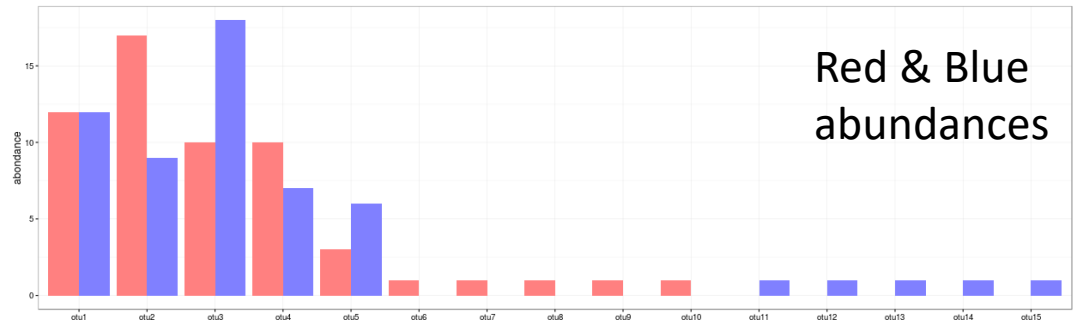


# Exploring biodiversity : $\beta$ -diversity

## Jaccard index:

- Proportion of species/OTUs specific to either Red or Blue  
→ qualitative index
- Pink = common OTUs between the 2 communities (5)
- Red= OTUs specific to Red community (5)
- Blue= OTUs specific to Blue community (5)

$$D_{jac} = 10/15 = 0.667$$

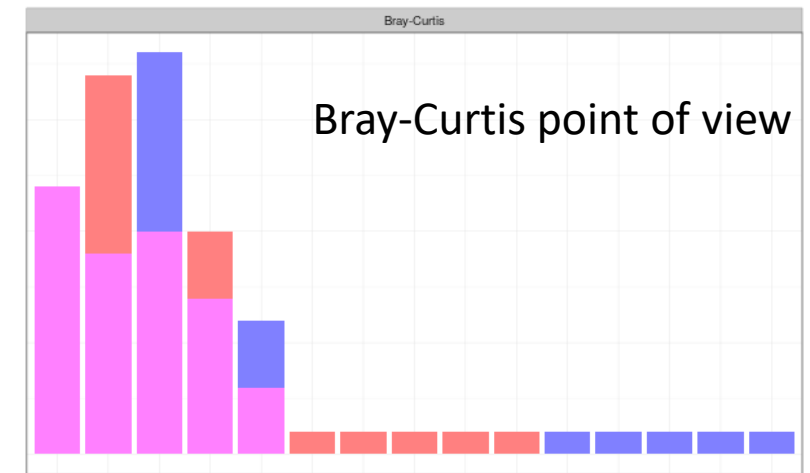
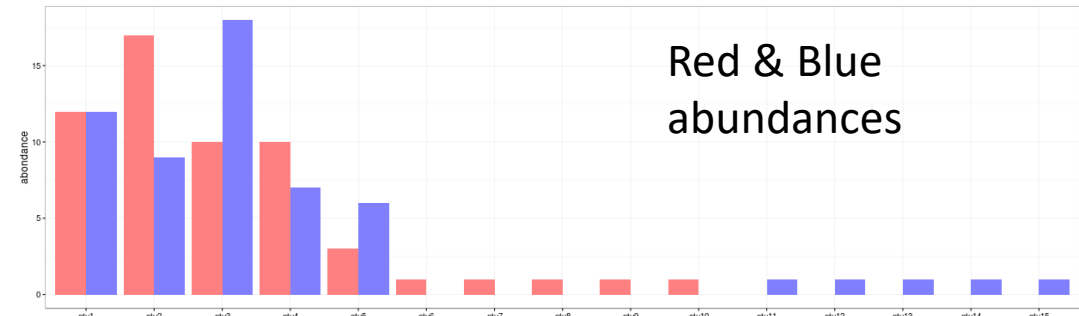


# Exploring biodiversity : $\beta$ -diversity

## Bray-Curtis index:

- Proportion of the abundance specific to either Red or Blue → quantitative index
- Ration (sum of specific abundances)/ (total abundances)
- 1<sup>st</sup> OTU does not contribute (same abundance for Red and Blue communities)
- OTU 2, 3, 4 and 5 contribute up to the excess in one of the communities (8+8+3+3+10) in the sum of specific abundances (Pink is not taken into account in this sum)

$$D_{bc} = (8+8+3+3+10) / (24+26+28+17+9+10) = 0.281$$

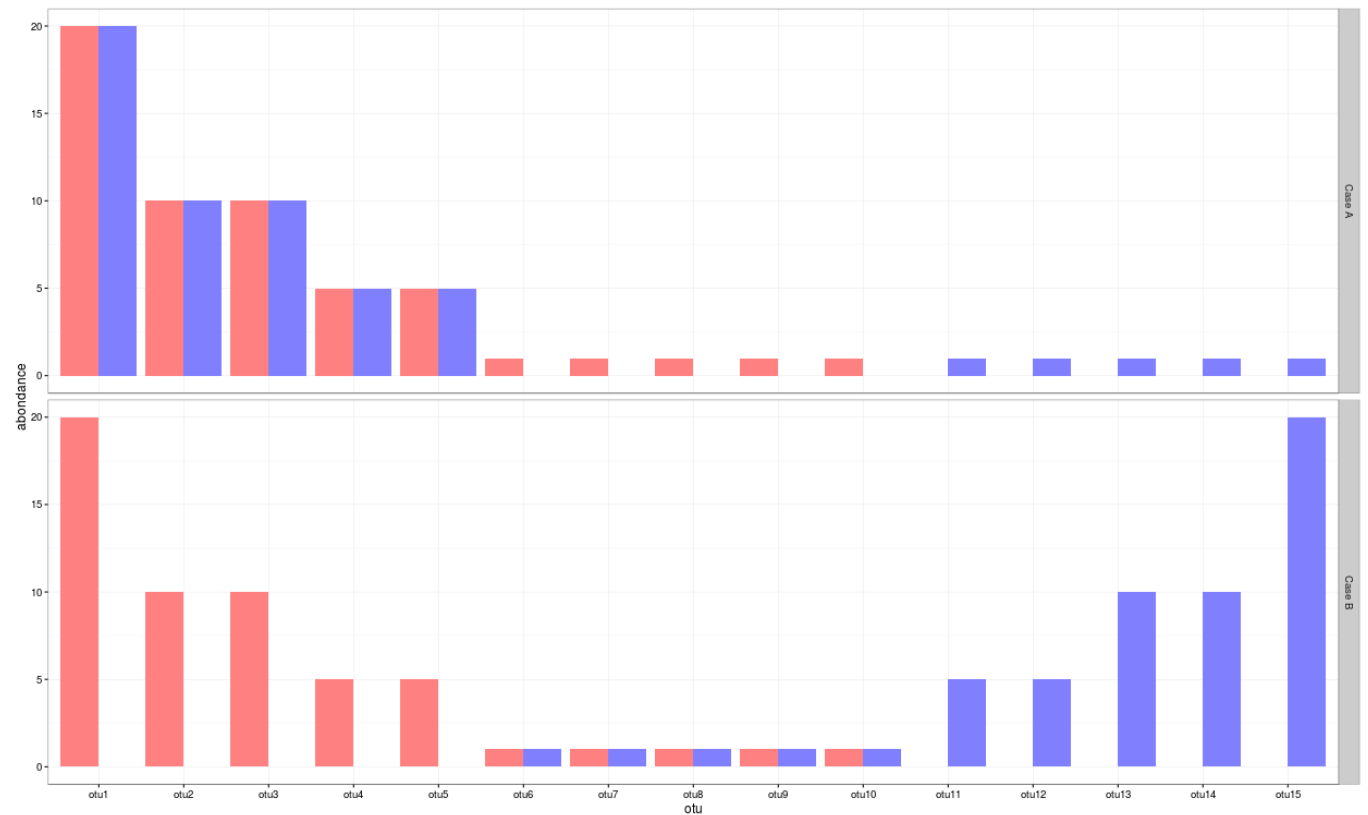


# Exploring biodiversity : $\beta$ -diversity

Indices comparison with different distributions:

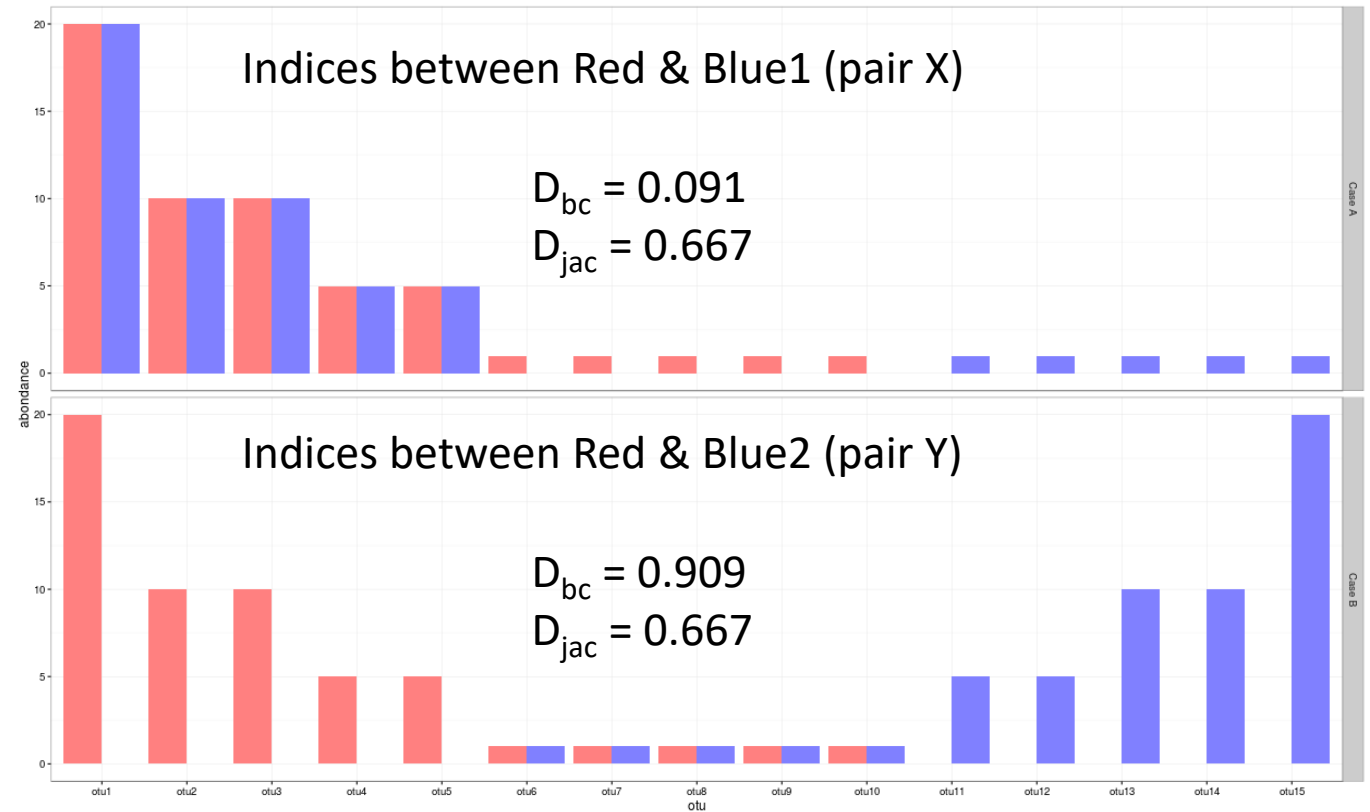
- between Red & Blue1 communities

- between Red & Blue2 communities



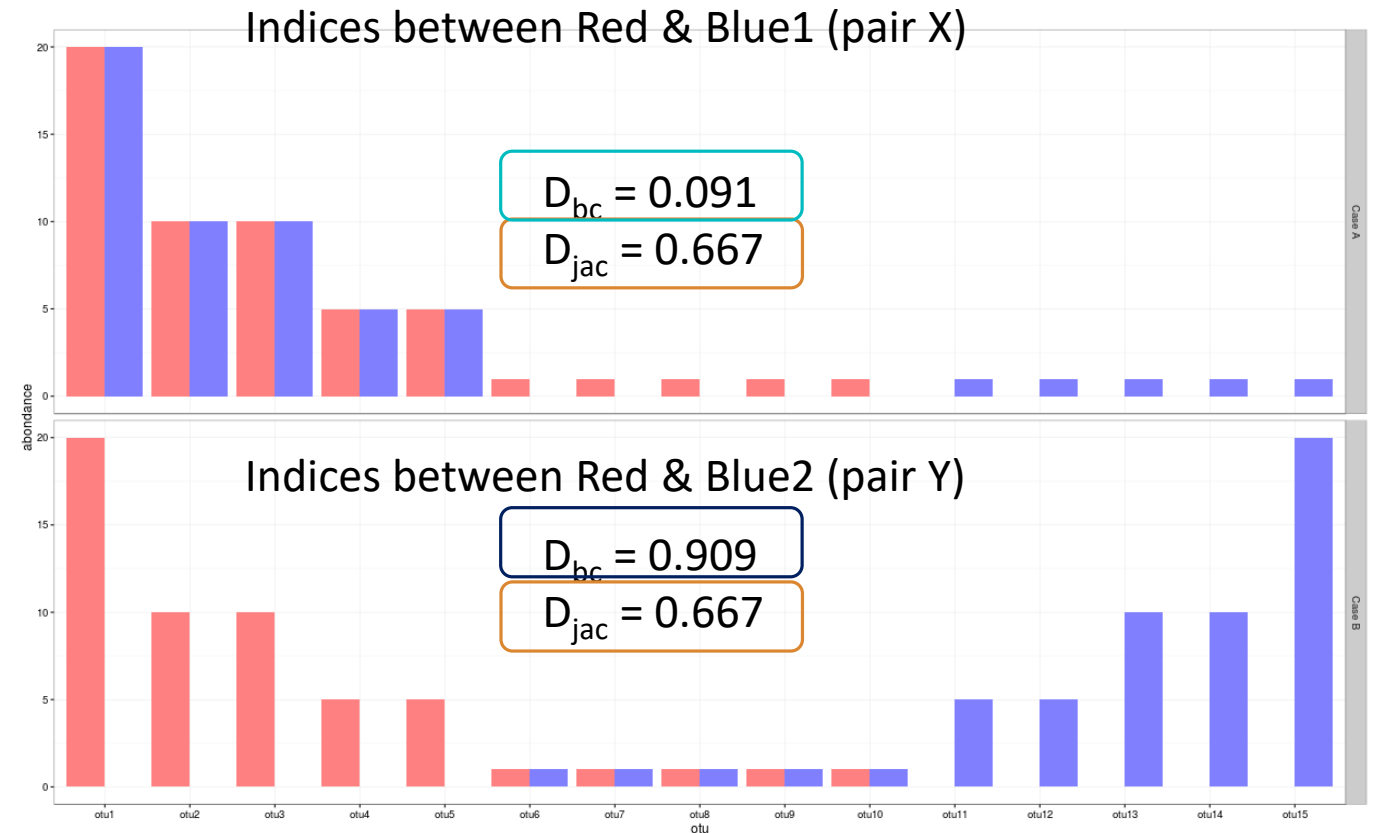
# Exploring biodiversity : $\beta$ -diversity

Jaccard and Bray-Curtis indices are calculated by pairs (in french “deux-à-deux”) so we here compare pair X indices with pair Y indices



# Exploring biodiversity : $\beta$ -diversity

1. Jaccard indices of X and Y are identical  $\rightarrow$  same specific fraction (there are as many OTUs specific to Red or Blue1 in X, as there are OTUs specific to Red or Blue2 in Y).
2. Pair X: Bray-Curtis index is low because shared OTUs between Red and Blue1 communities are abundant and specific OTUs are at low abundance.
3. Pair Y: Bray-Curtis index is high because OTUs specific to Red or Blue2 are abundant and shared OTUs are at low abundance.



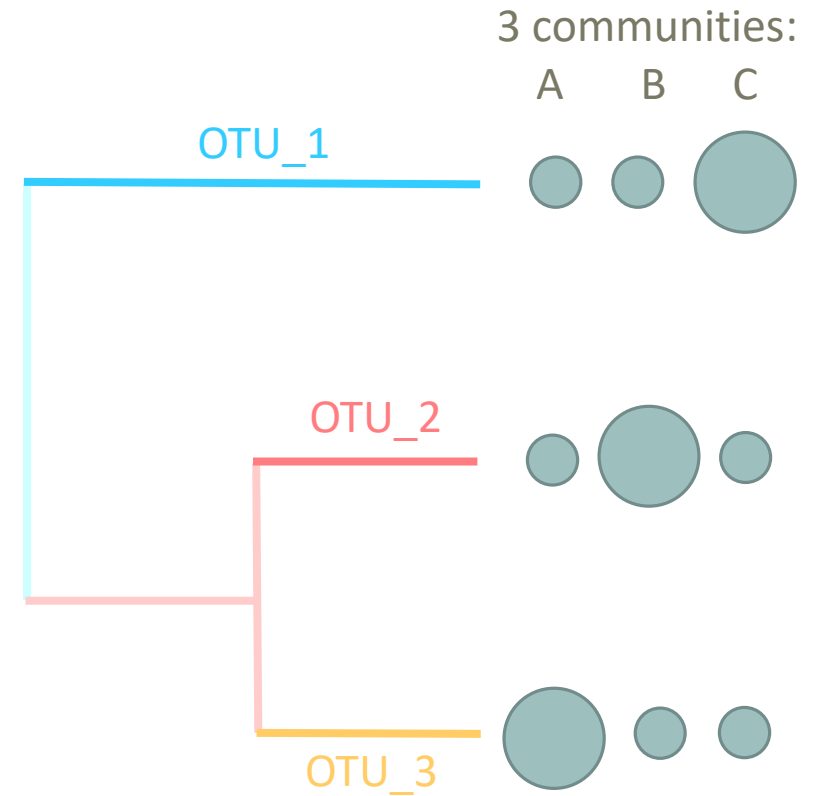


# Exploring biodiversity : $\beta$ -diversity

3 ways to measure beta diversity with the same data set  
→ 3 different results.

In this example :

- ✓ qualitatively, communities are very similar
- ✓ quantitatively, communities are very different
- **phylogenetically**, two communities seem to be closer than the third one.



# Exploring biodiversity : $\beta$ -diversity

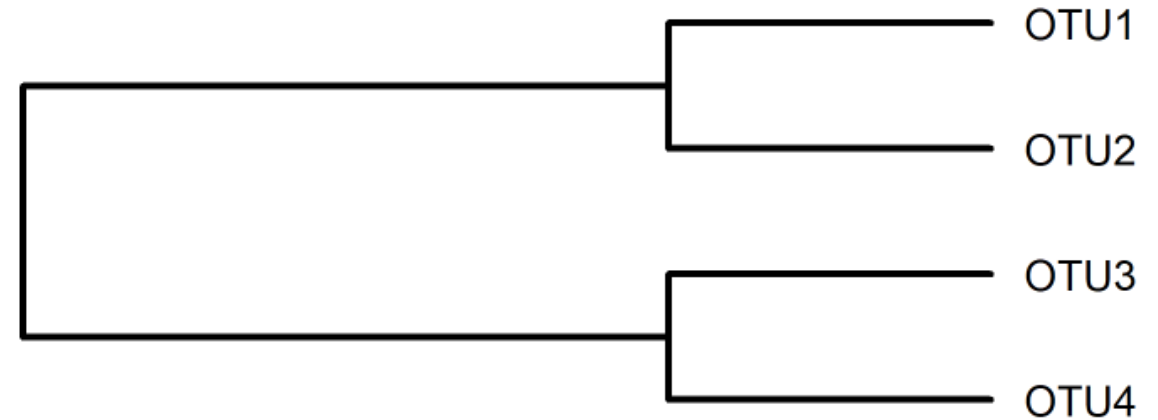
---

## Unifrac index:

- Fraction of the tree specific to either A or B

## Weighted-Unifrac index :

- Fraction of the diversity specific to either A or B

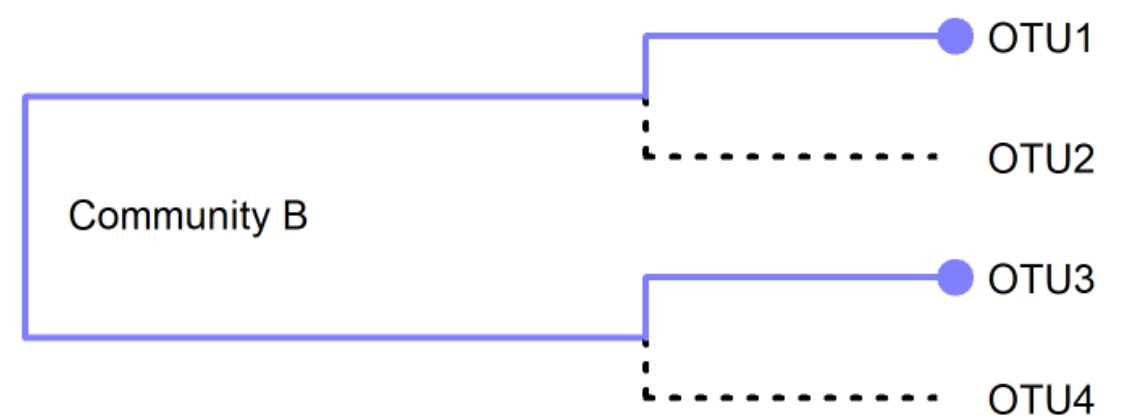
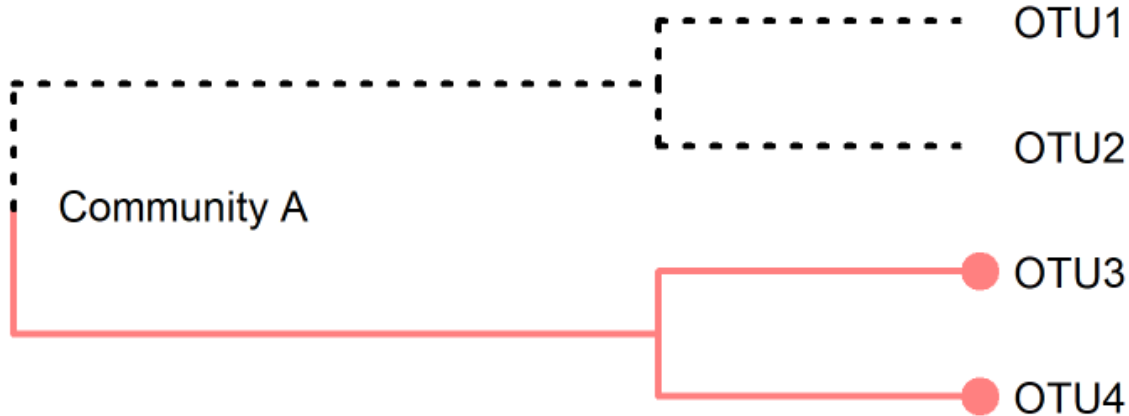


# Exploring biodiversity : $\beta$ -diversity

## Unifrac index:

- Fraction of the tree specific to either A or B

$$Unifrac = \frac{\sum \text{specific\_branch\_length}}{\sum \text{all\_branch\_length}}$$



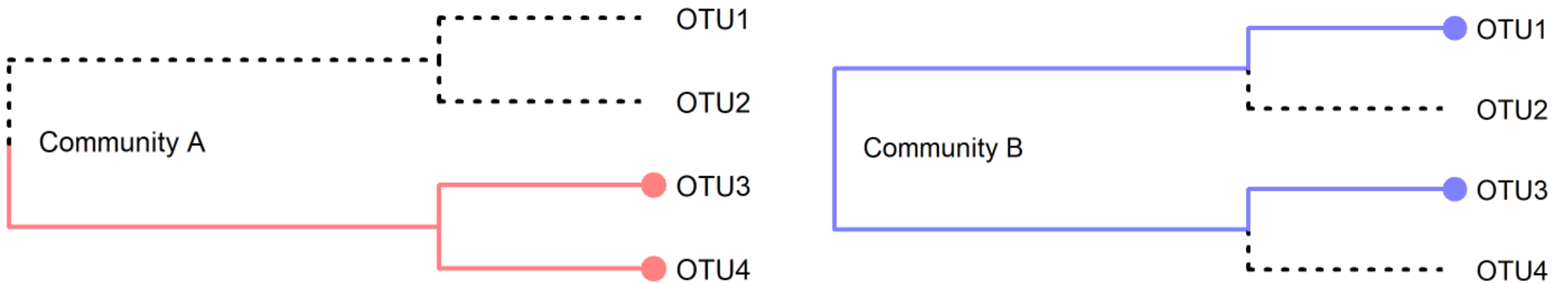
3 OTUs identified by sequencing: OTU3, OTU4 in community A and OTU1, OTU3 in community B

# Exploring biodiversity : $\beta$ -diversity

## Unifrac index:

- Fraction of the tree specific to either A or B

$$Unifrac = \frac{\sum \text{specific\_branch\_length}}{\sum \text{all\_branch\_length}}$$



OTU1 and OTU4 are specific, OTU3 is shared in the 2 communities and OTU2 are absent in the 2 communities

# Exploring biodiversity : $\beta$ -diversity

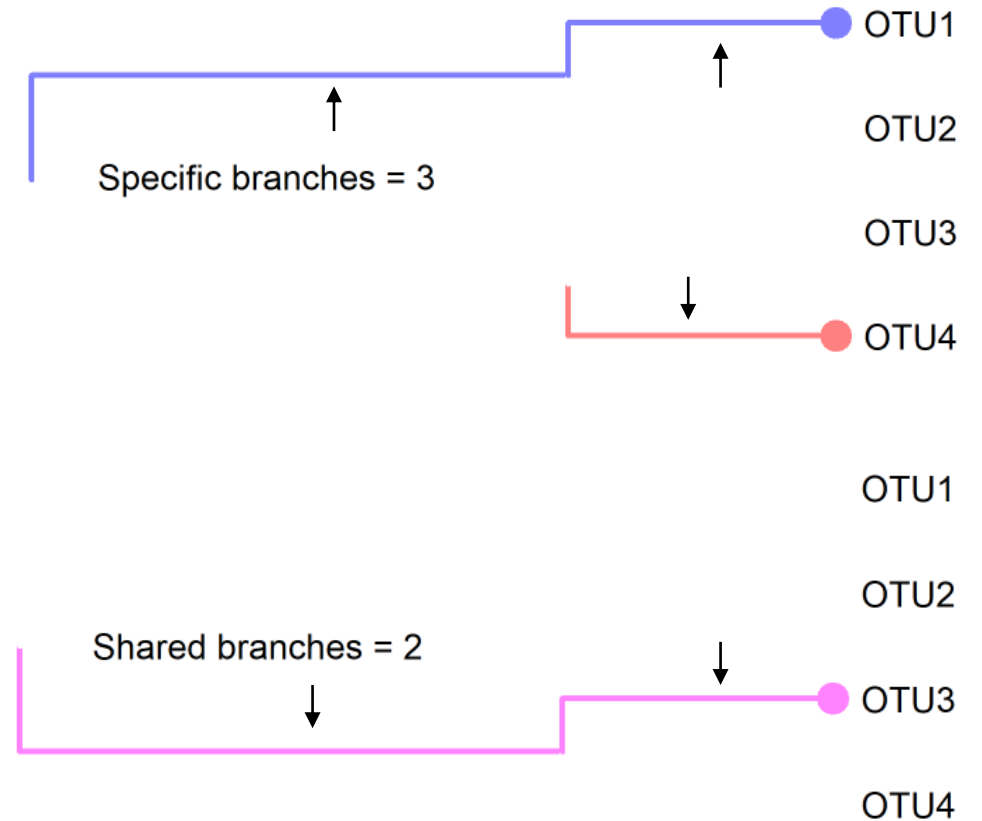
## Unifrac index:

- Fraction of the tree specific to either A or B

If all branch lengths are equal to 1, only branches present in at least one community are taken into account :

$$Unifrac = \frac{\sum \text{specific\_branch\_length}}{\sum \text{all\_branch\_length}} = 3/5 = 0.6$$

- Pink = common OTUs between the 2 communities
- Red = tree branch specific to A
- Blue = tree branch specific to B

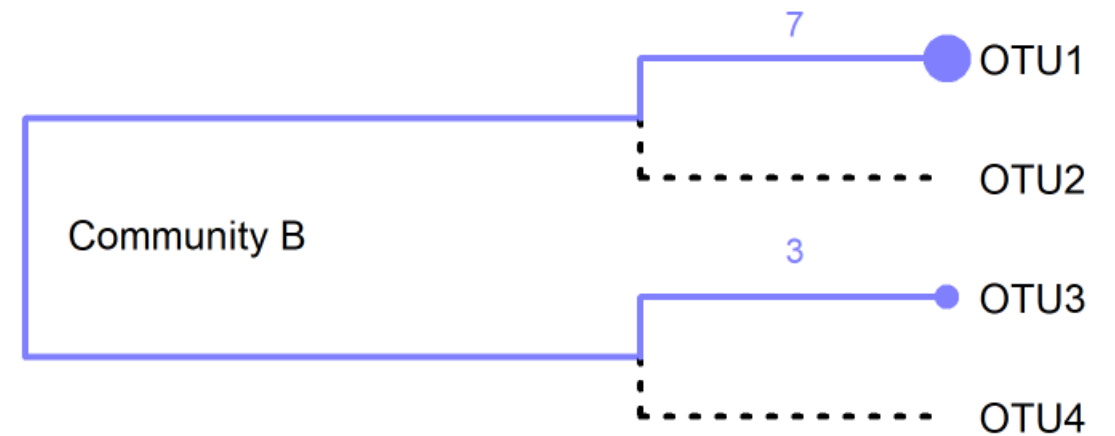
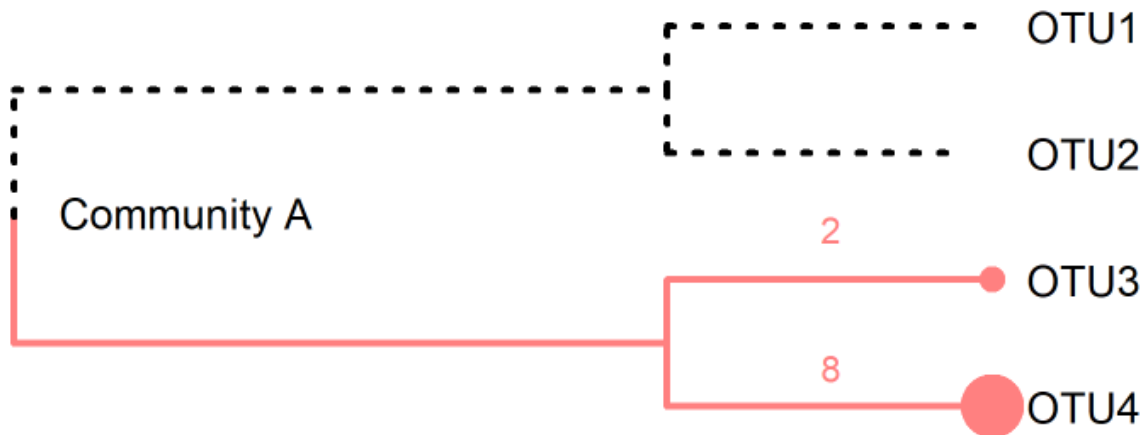


A reduced branch is a branch whose distance is weighted by the relative abundance of the OTU

## Weighted-Unifrac index:

- Fraction of the diversity specific to either A or B

$$WUnifrac = \frac{\sum \text{reduced\_branch\_length}}{\sum \text{non\_reduced\_branch\_length}}$$

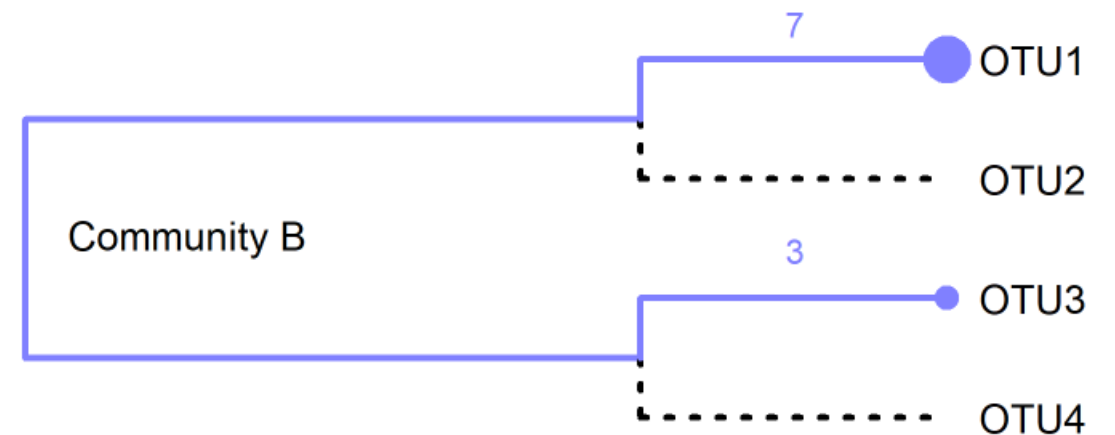
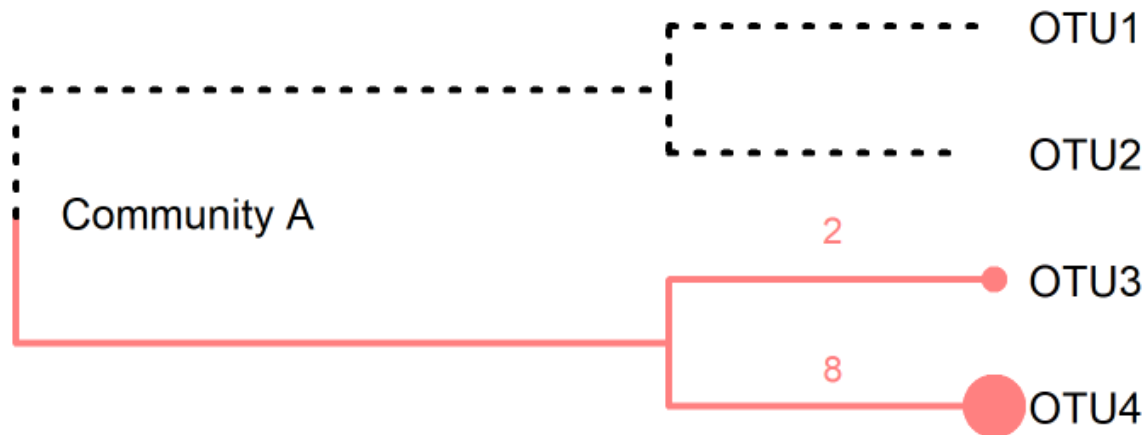


# Exploring biodiversity : $\beta$ -diversity

## Weighted-Unifrac index:

- Fraction of the diversity specific to either A or B

$$WUnifrac = \frac{\sum \text{reduced\_branch\_length}}{\sum \text{non\_reduced\_branch\_length}}$$

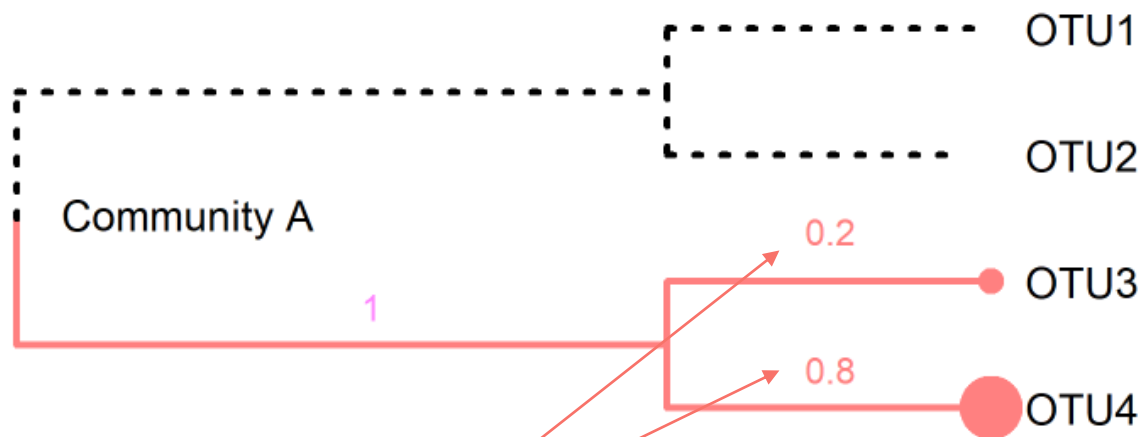


Here the specific OTUs (OTU1 and OTU4) are the most abundant and are also the most phylogenetically distant.

# Exploring biodiversity : $\beta$ -diversity

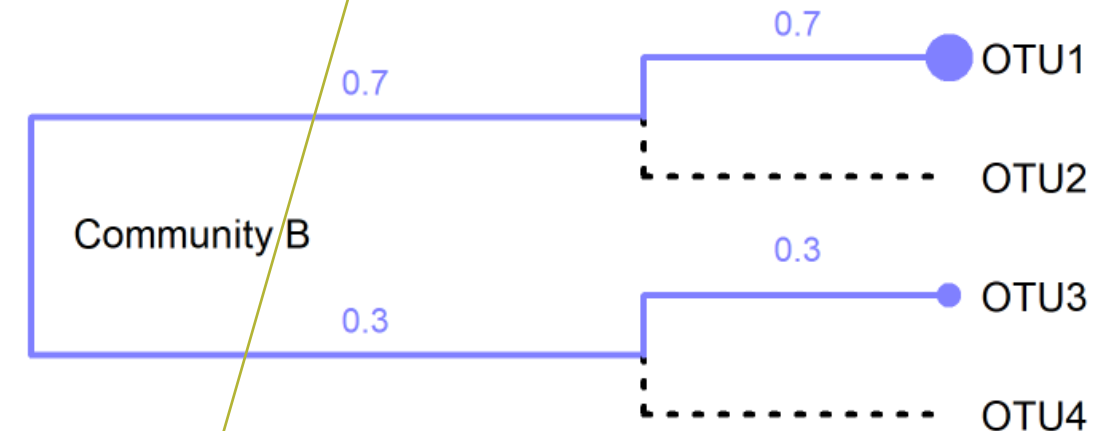
## Weighted-Unifrac index:

- Fraction of the diversity specific to either A or B



ratio of the abundance of each branch

$$WUnifrac = \frac{\sum \text{reduced\_branch\_length}}{\sum \text{non\_reduced\_branch\_length}}$$



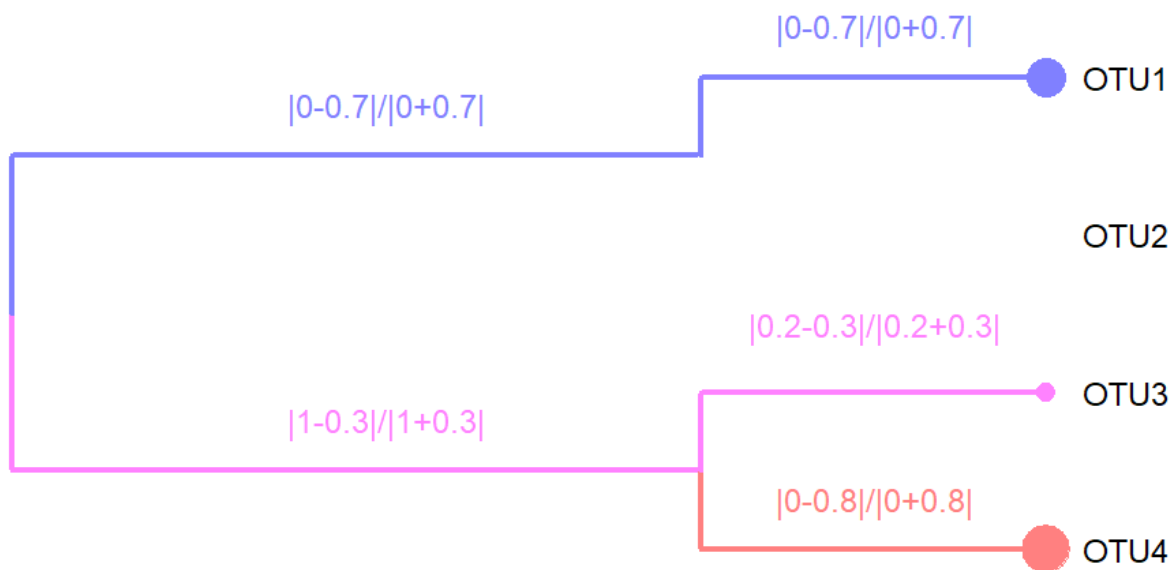
A reduced branch is a branch whose distance is weighted by the relative abundance of the OTU



# Exploring biodiversity : $\beta$ -diversity

## Weighted-Unifrac index:

- Fraction of the diversity specific to either A or B



$$WUnifrac = \frac{\sum \text{reduced\_branch\_length}}{\sum \text{non\_reduced\_branch\_length}}$$

$$\text{Blue branches} = \frac{|0 - 0,7|}{|0 + 0,7|} + \frac{|0 - 0,7|}{|0 + 0,7|} = 1 + 1 = 2$$

$$\text{Red branches} = \frac{|0 - 0,8|}{|0 + 0,8|} = 1$$

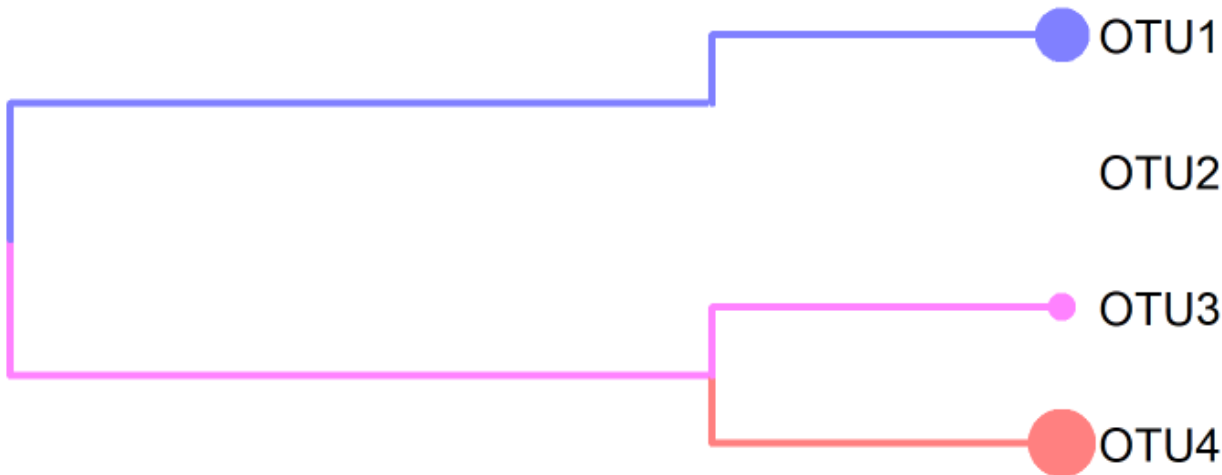
$$\text{Pink branches} = \frac{|1 - 0,3|}{|1 + 0,3|} + \frac{|0,2 - 0,3|}{|0,2 + 0,3|} = \frac{0,7}{0,3} + \frac{0,1}{0,5} = 0,73$$

$$\sum \text{reduced branch length} = 3,73$$

# Exploring biodiversity : $\beta$ -diversity

## Weighted-Unifrac index:

- Fraction of the diversity specific to either A or B



$$WUnifrac = \frac{\sum \text{reduced\_branch\_length}}{\sum \text{non\_reduced\_branch\_length}}$$

$$\sum \text{non reduced branch length} = 5$$

$$WUnifrac = \frac{\sum \text{reduced\_branch\_length}}{\sum \text{non\_reduced\_branch\_length}} = \frac{3,73}{5} = 0,75$$

# Exploring biodiversity : $\beta$ -diversity in brief

---

**qualitative** indices: presence/absence regardless of abundance

**quantitative** indices: compare differences in abundance of OTUs

**phylogenetic** indices: integrate phylogenetic information to qualitative or quantitative indices (weighted or unweighted indices)

**Bray-Curtis** index : to evaluate the dissimilarity between two given samples, in terms of abundance of OTUs present in each sample. When Bray-Curtis index close to 0 means abundant OTUs are shared and in the same quantities between communities.

**Jaccard** index: beta diversity index, qualitative, takes into account the fraction of specific OTUs

**Unifrac** index: beta diversity index, qualitative, takes into account the fraction of specific phylogenetic branches

**Weighted-Unifrac** index: beta diversity index, quantitative, takes into account the relative abundance of OTUs shared between samples

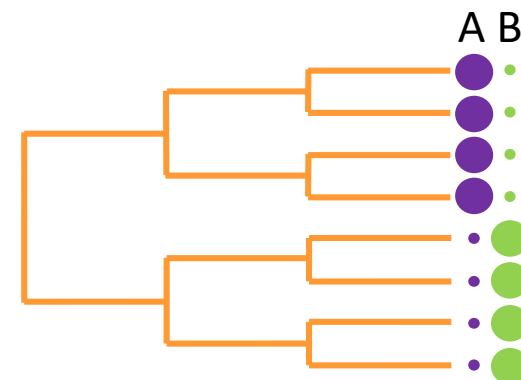
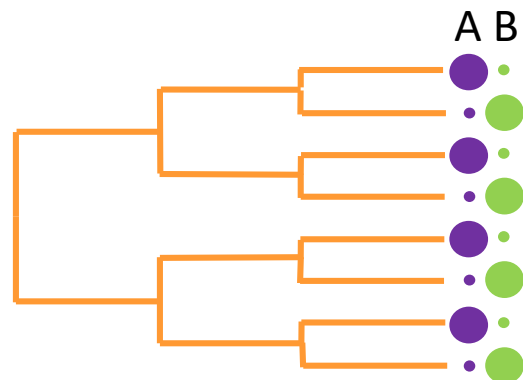
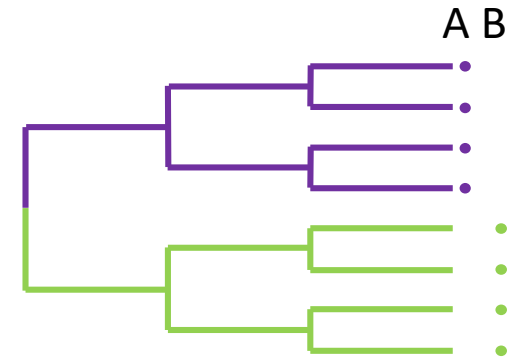
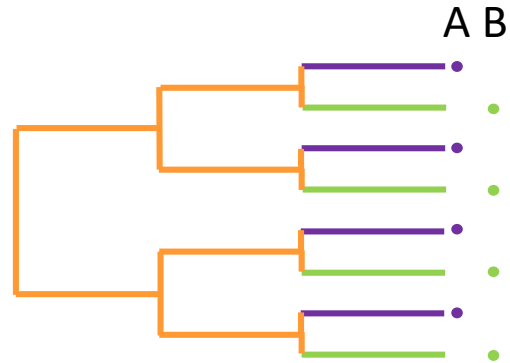
# Exploring biodiversity : $\beta$ -diversity

→ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weighed Unifrac values for these 4 pairs of communities?

 : in common

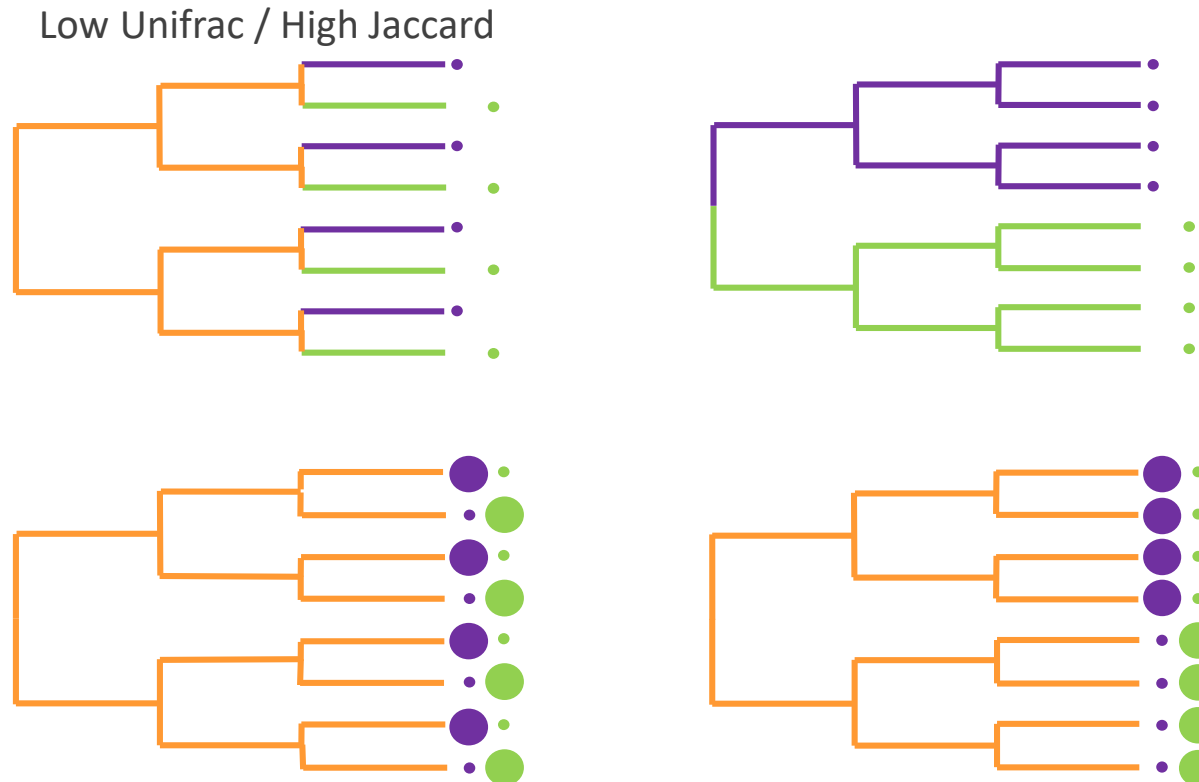
 : specific to A

 : specific to B



# Exploring biodiversity : $\beta$ -diversity

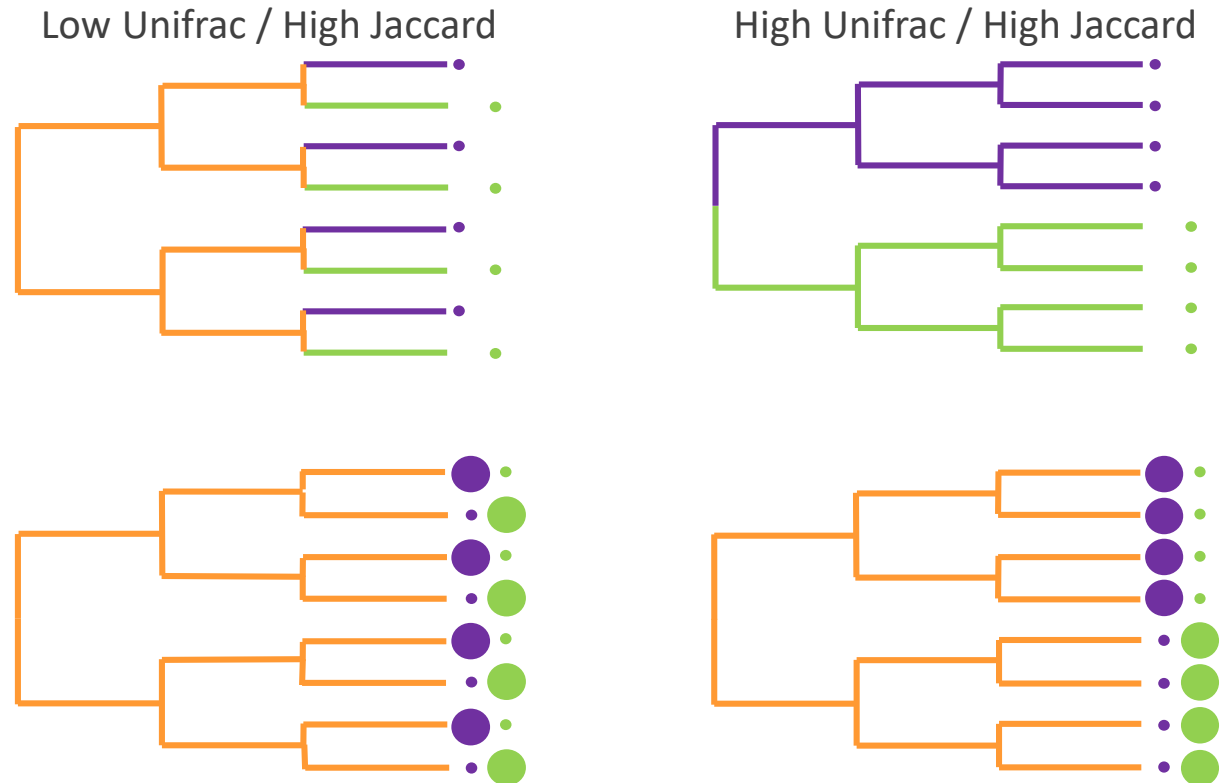
→ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weighed Unifrac values?



High Jaccard: same amount of specific OTUs  
Low Unifrac: small distance between specific branches

# Exploring biodiversity : $\beta$ -diversity

→ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weighed Unifrac values?

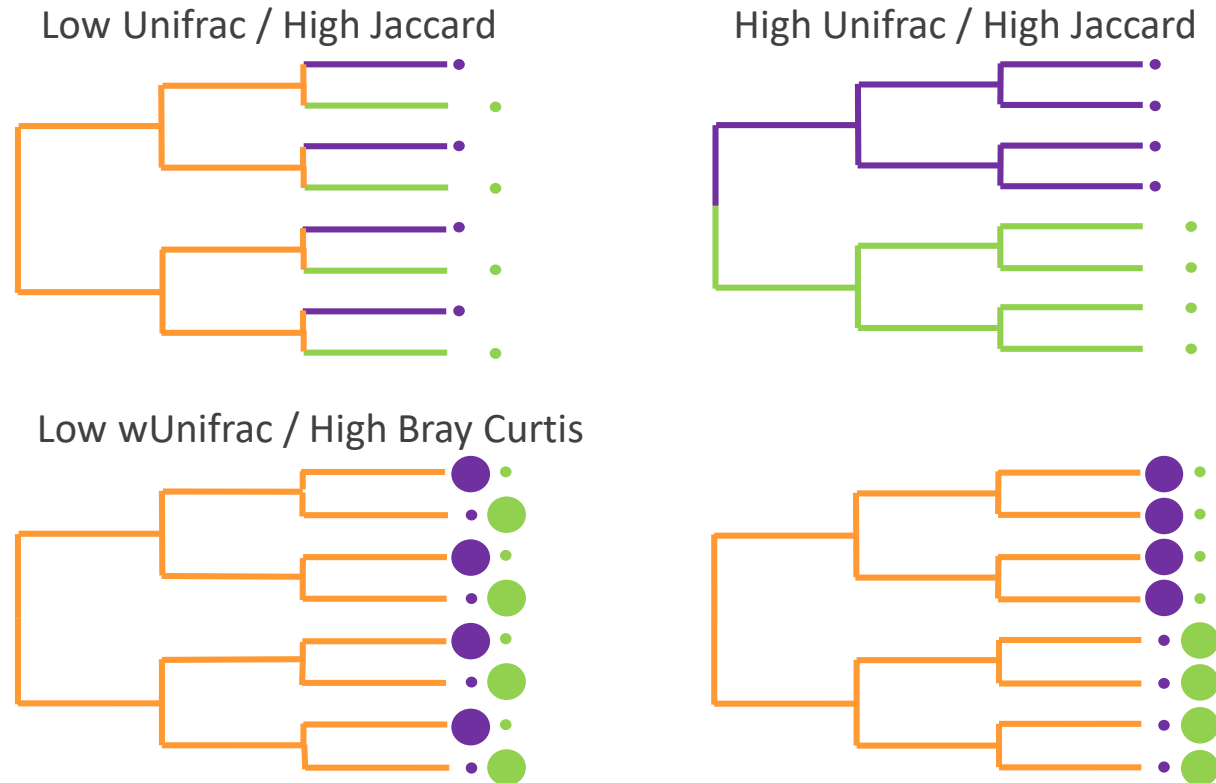


High Jaccard: all OTUs are specific to A or B

High Unifrac: all the branches are specific to A or B

# Exploring biodiversity : $\beta$ -diversity

→ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weighed Unifrac values?

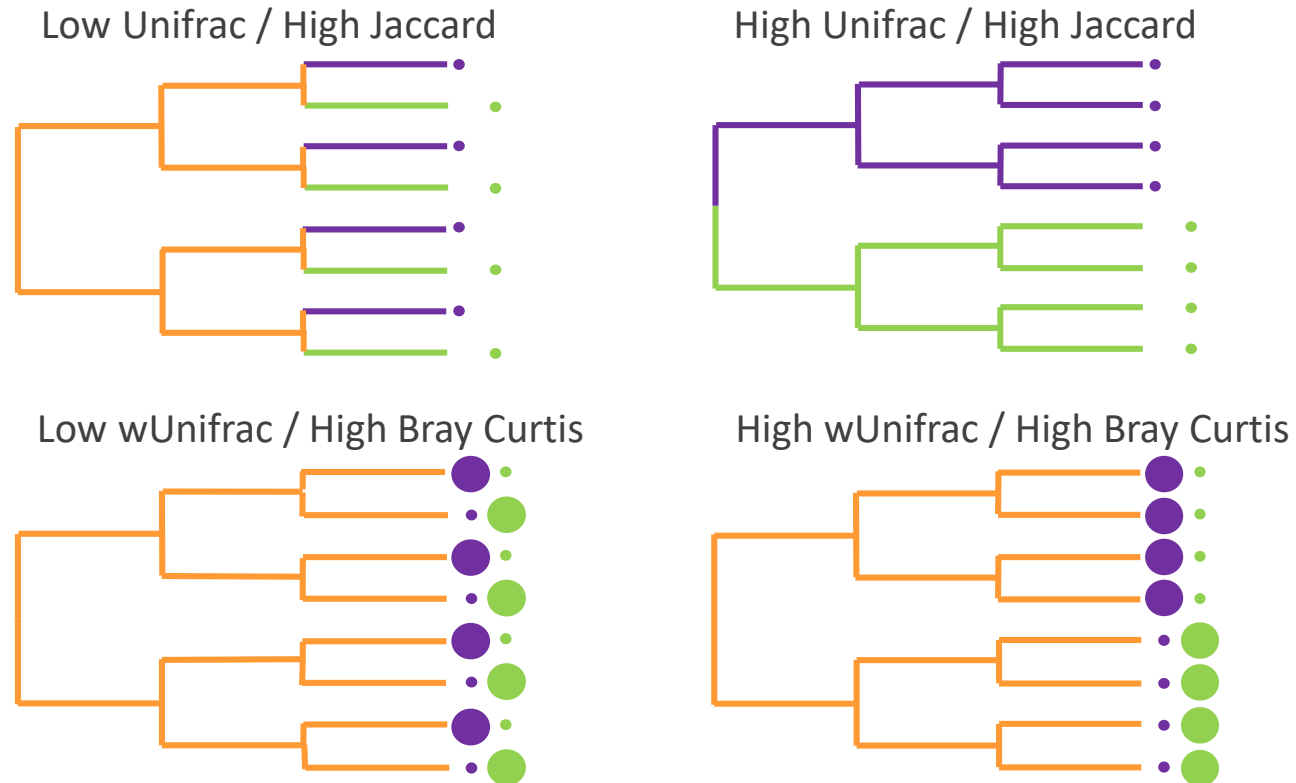


High Bray-Curtis: OTUs are shared but abundant OTUs are not the same in each community

Low weighted-Unifrac: abundant OTUs in a community have a phylogenetically close relative in the other community

# Exploring biodiversity : $\beta$ -diversity

→ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weighed Unifrac values?



High Bray-Curtis: OTUs are shared but abundant OTUs are not the same in each community

High weighted-Unifrac: abundant OTUs in a community are phylogenetically distant to any OTU in the other community



# Exploring biodiversity : $\beta$ -diversity

---

Phyloseq supports currently 43 beta diversity distance methods,  
(see [phyloseq distanceMethodList documentation](#) )

unifrac, wunifrac,

dpcoa, jsd, manhattan, euclidean, canberra,

bray, kulczynski, jaccard, gower, altGower, morisita, horn, mountford, raup, binomial  
chao, cao...

# Exploring biodiversity : $\beta$ -diversity

**FROGSSTAT Phyloseq Beta Diversity** distance matrix (Galaxy Version 3.2.2) Options

**Phyloseq object (format rdata)**

28: Phyloseq\_raref.Rdata

This is the result of FROGS Phyloseq Import Data tool.

**Grouping variable**

EnvType

Experimental variable used to group samples (Treatment, Host type, etc).

**The methods of beta diversity**

Select/Unselect all

Unifrac

Weighted Unifrac

Bray-Curtis

Jaccard (as cc method in betadiver vegan funcion)

N.B. if the tree is not available in your RData, you cannot choose Unifrac or Weighted Unifrac

**Other method**

The other methods of beta diversity that you want to use (comma separated value). c.f. details below.

Explore the sample **NORMALISED** count

Choose a sample variable to organize graphics.

Choose which beta diversity distances you want to compute

You can ask another beta-diversity method

# Exercise 6

---

Try it with the 4 most commonly used distance methods

1. What are the output datasets ?
2. *A priori*, abundant OTUs are they shared among samples?
3. Compare Jaccard and Unifrac, what can you conclude ?
4. Compare Unifrac and weighted Unifrac, what can you conclude ?

# Exercise 6

---

## 1. What are the output files ?

→ Tabular file: a tabular file per distance method containing the “all samples against all” matrix of beta diversity distance

→ HTML report: heatmap representing the distance matrix computed

**FROGSSTAT Phyloseq Beta Diversity: beta\_diversity.nb.html (wunifrac.tsv)**

**FROGSSTAT Phyloseq Beta Diversity: beta\_diversity.nb.html (unifrac.tsv)**

**FROGSSTAT Phyloseq Beta Diversity: beta\_diversity.nb.html (cc.tsv)**

**FROGSSTAT Phyloseq Beta Diversity: beta\_diversity.nb.html (bray.tsv)**

**FROGSSTAT Phyloseq Beta Diversity: beta\_diversity.nb.html**

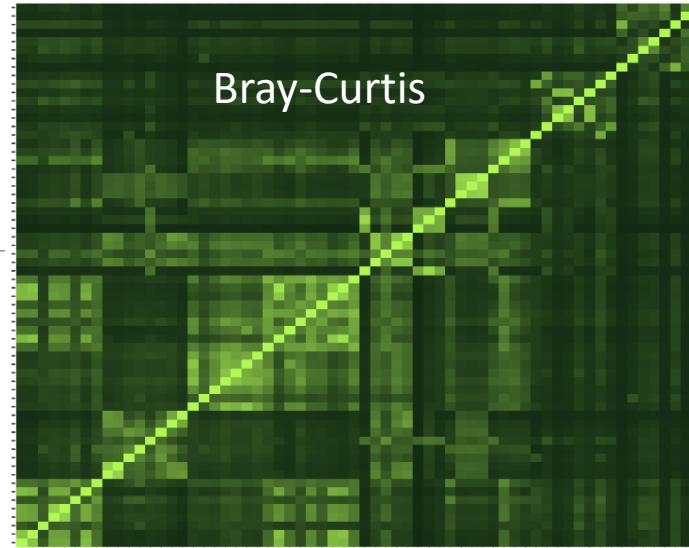
For Jaccard



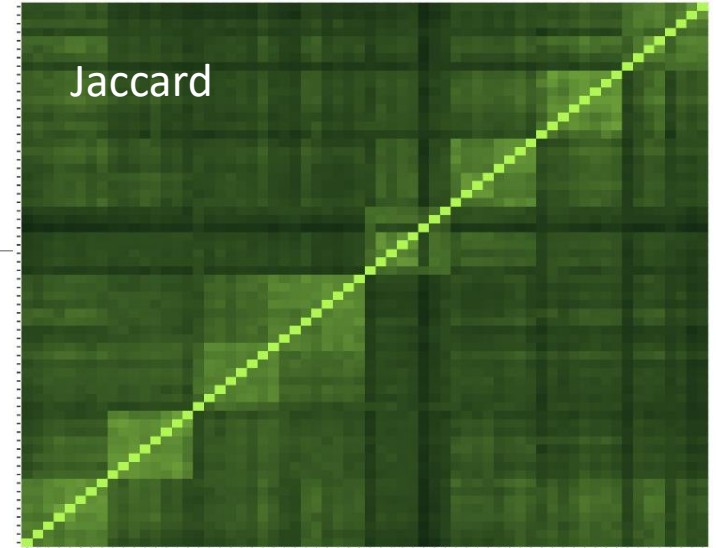
# Exercise 6

[FROGSSTAT Phyloseq Beta Diversity: beta\\_diversity.nb.html](#)

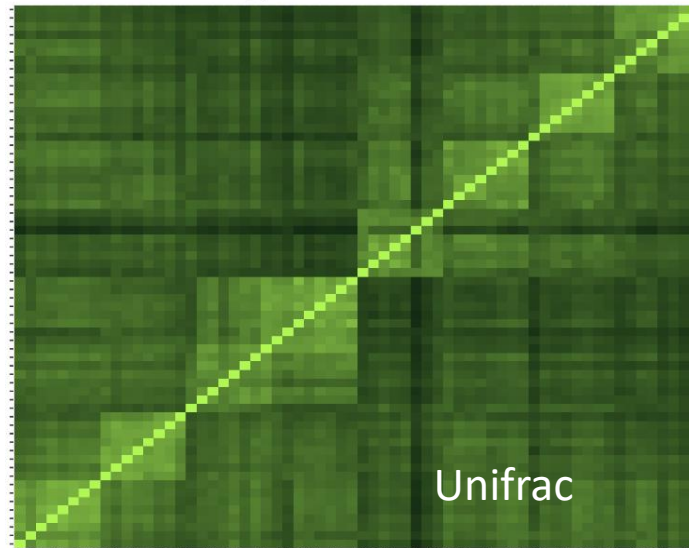
Heatmap plot of the beta distance : bray



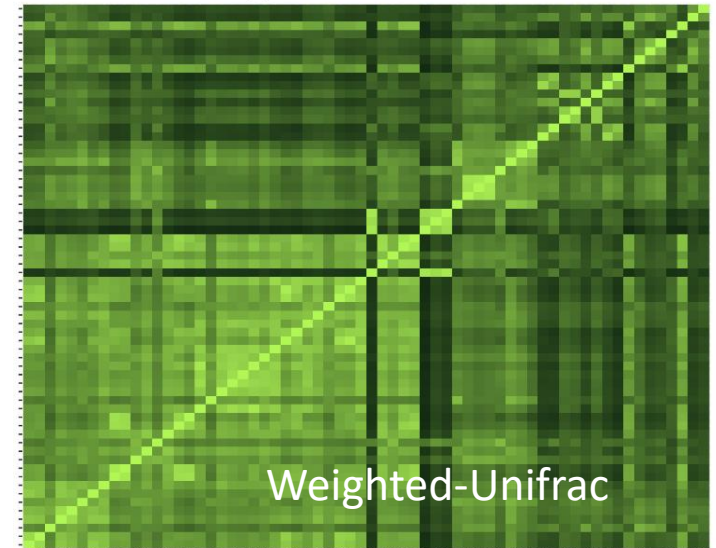
Heatmap plot of the beta distance : cc



Heatmap plot of the beta distance : unifrac



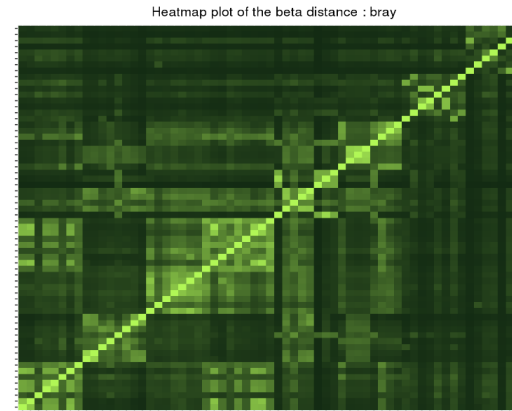
Heatmap plot of the beta distance : wunifrac



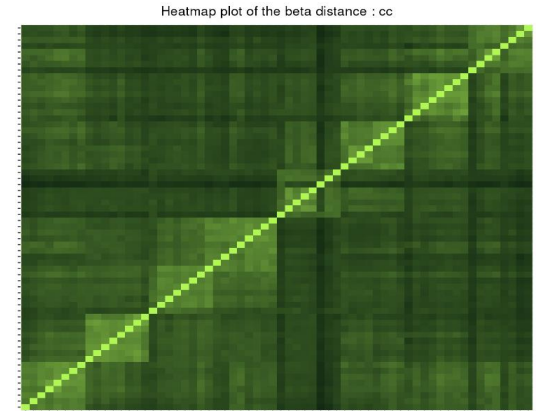
# Exercise 6

- Each square represent a comparison between 2 samples
- Lighter means more similar
- The diagonal represents the comparison of a sample with itself
- Along the diagonal we can spot clearer square structures
- We can assume that these are the different EnvTypes as the samples are ordered.

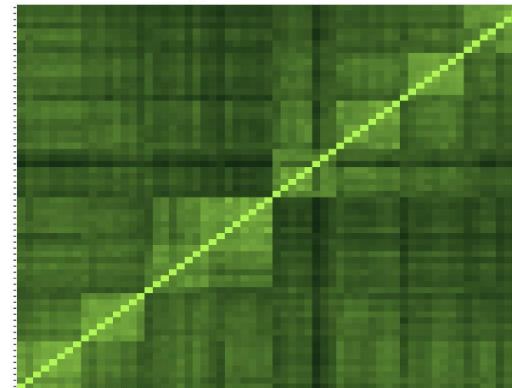
Bray-Curtis



Jaccard

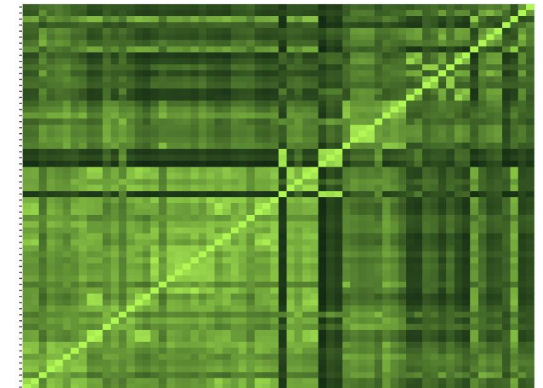


Heatmap plot of the beta distance : unifrac



Unifrac

Heatmap plot of the beta distance : wunifrac



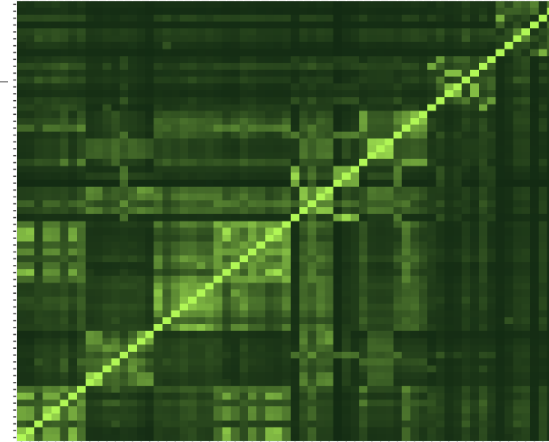
Weighted-Unifrac

# Exercise 6

2. *A priori*, are abundant OTU they shared among samples ?

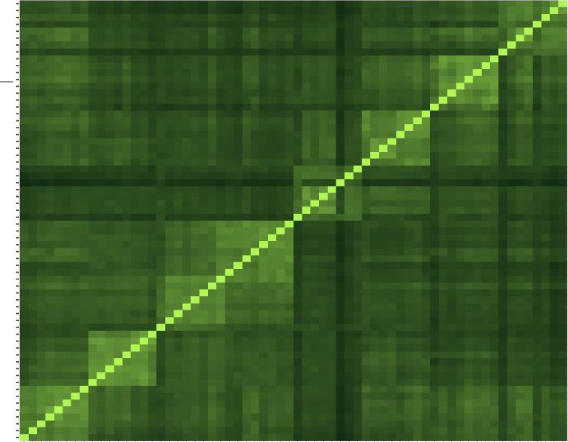
Bray-Curtis

Heatmap plot of the beta distance : Bray

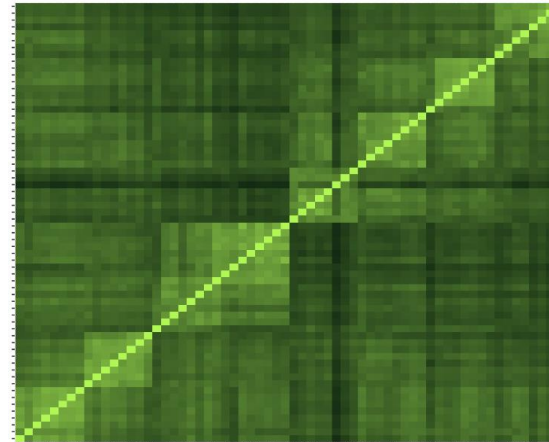


Jaccard

Heatmap plot of the beta distance : cc

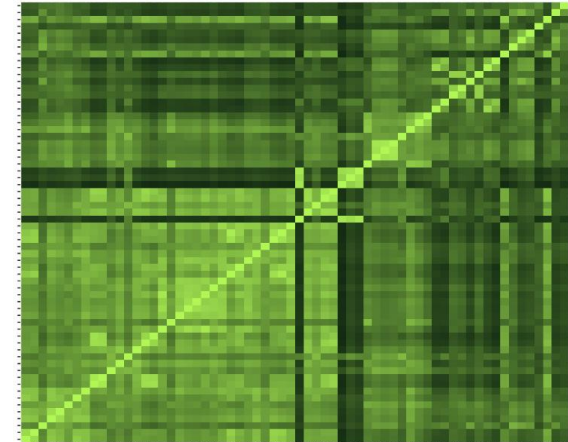


Heatmap plot of the beta distance : unifrac



Unifrac

Heatmap plot of the beta distance : wunifrac



Weighted-Unifrac



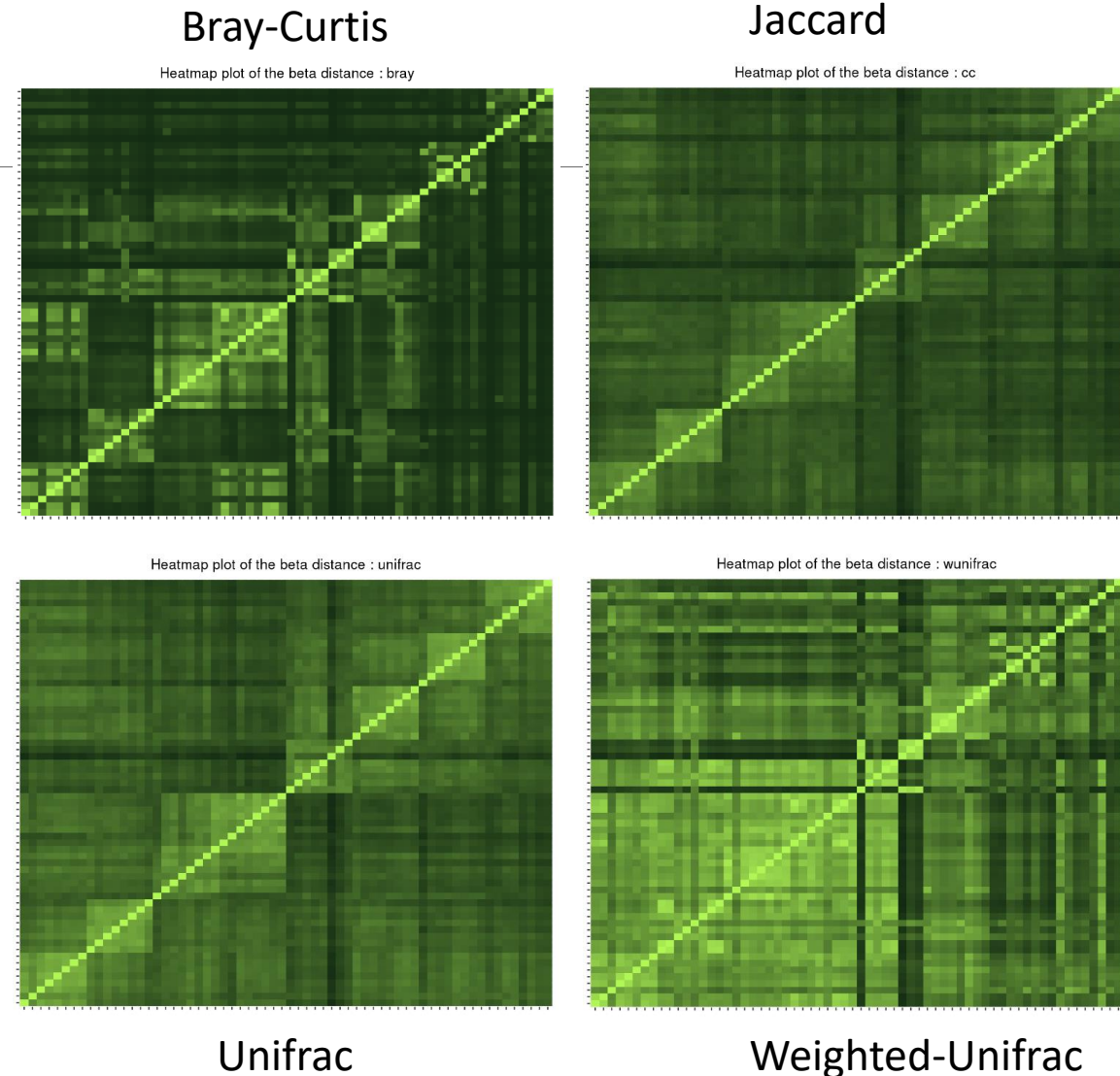
# Exercise 6

2. *A priori*, are abundant OTU they shared among samples ?

- Jaccard lower than Bray-Curtis
- Weighted-Unifrac is lower than Unifrac

→ The abundance accentuates the differences i.e. the distances are greater, i.e. the images are darker

→ abundant OTUs are community specific



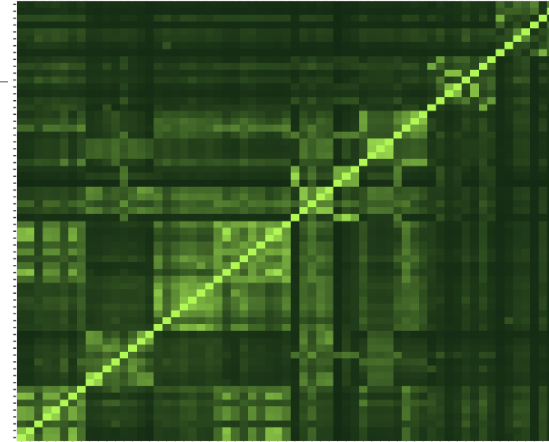


# Exercise 6

3. Compare Jaccard and Unifrac, what can you conclude ?

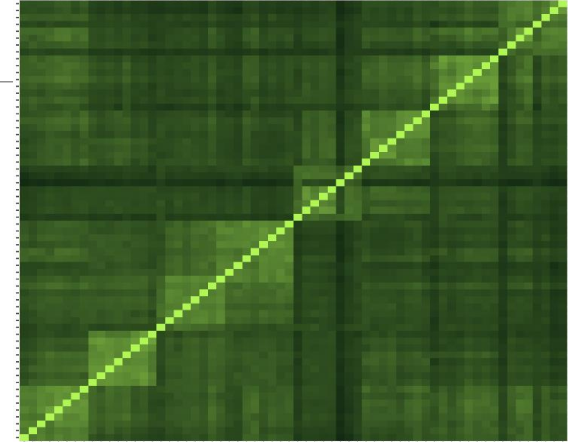
Bray-Curtis

Heatmap plot of the beta distance : bray

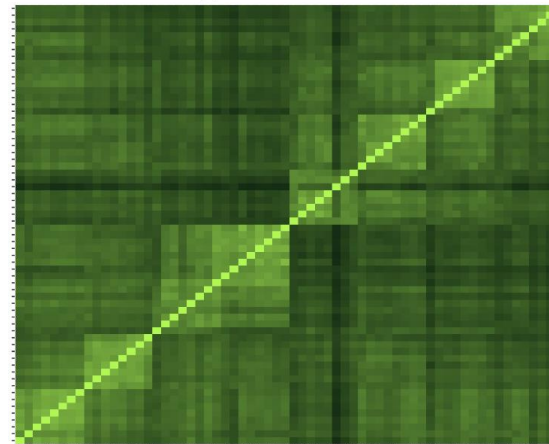


Jaccard

Heatmap plot of the beta distance : cc

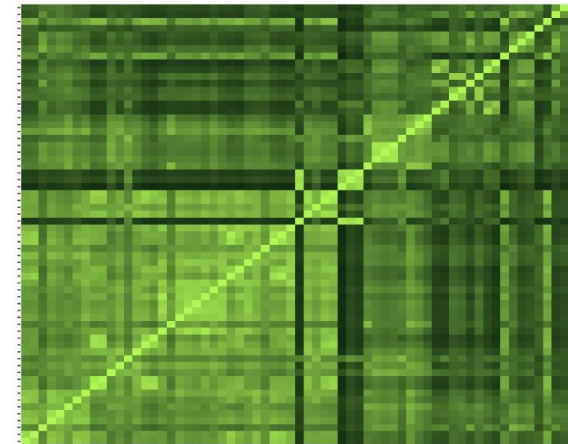


Heatmap plot of the beta distance : unifrac



Unifrac

Heatmap plot of the beta distance : wunifrac



Weighted-Unifrac

# Exercise 6

3. Compare Jaccard and Unifrac, what can you conclude ?

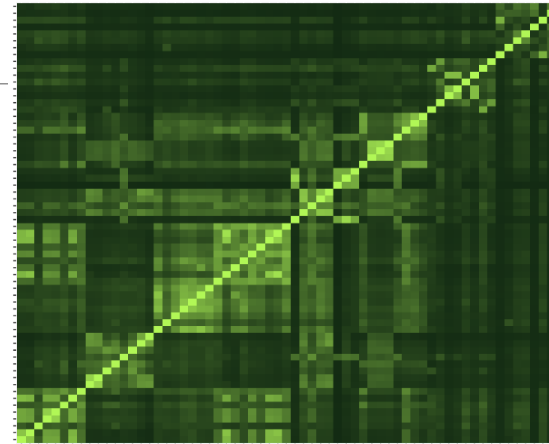
- Jaccard and Unifrac are close.

→ the phylogenetic distances do not accentuate the qualitative data of the Jaccard (neither darker, nor lighter), the species are thus close

→ OTUs are distinct but phylogenetically related

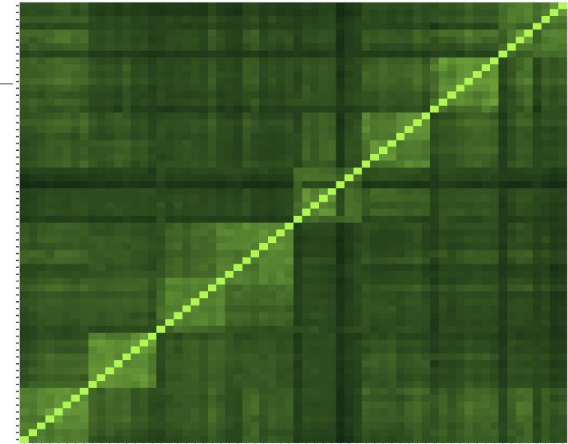
Bray-Curtis

Heatmap plot of the beta distance : bray

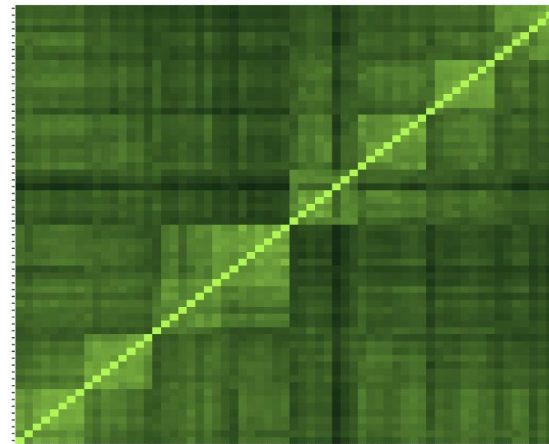


Jaccard

Heatmap plot of the beta distance : cc

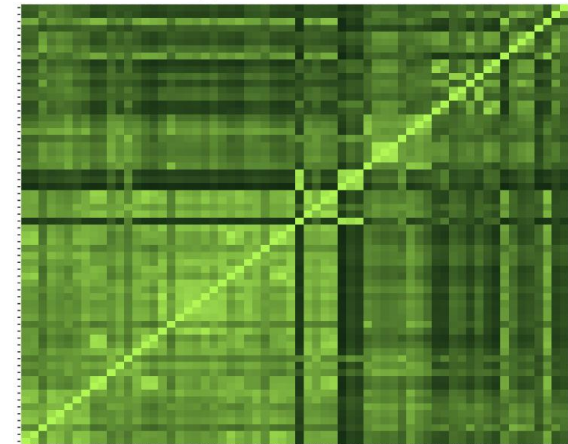


Heatmap plot of the beta distance : unifrac



Unifrac

Heatmap plot of the beta distance : wunifrac



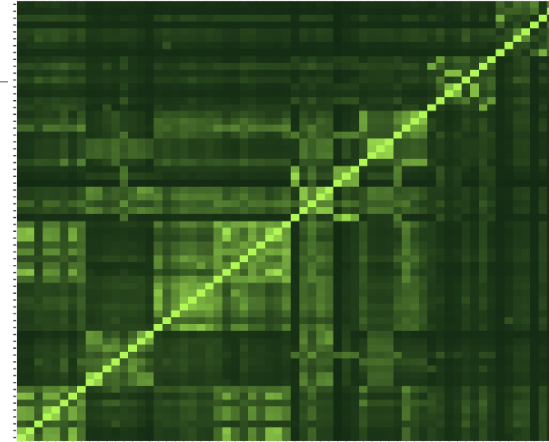
Weighted-Unifrac

# Exercise 6

4. Compare Unifrac and weighted Unifrac, what can you conclude ?

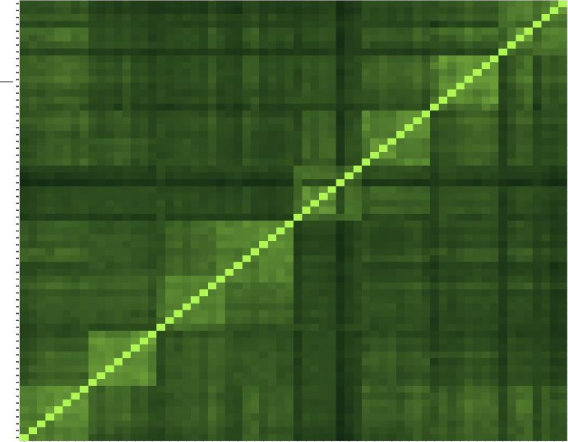
Bray-Curtis

Heatmap plot of the beta distance : bray

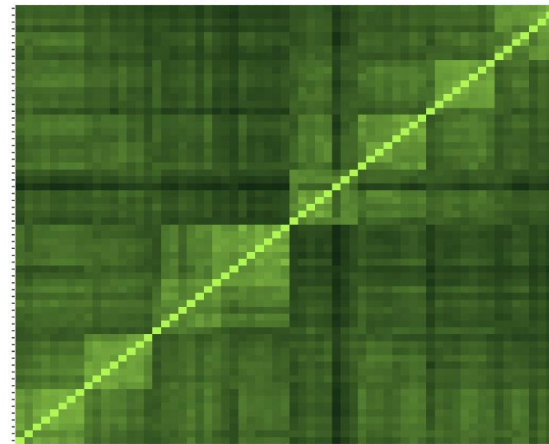


Jaccard

Heatmap plot of the beta distance : cc

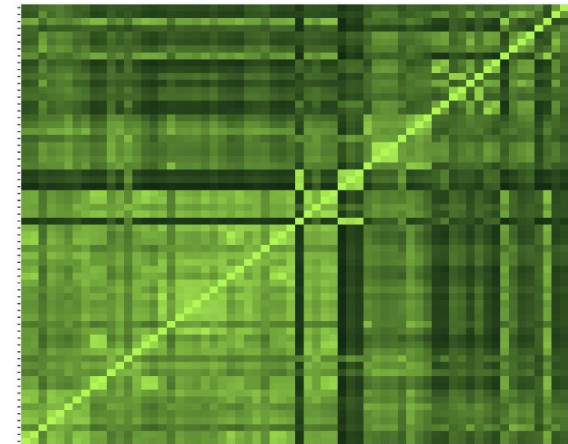


Heatmap plot of the beta distance : unifrac



Unifrac

Heatmap plot of the beta distance : wunifrac



Weighted-Unifrac

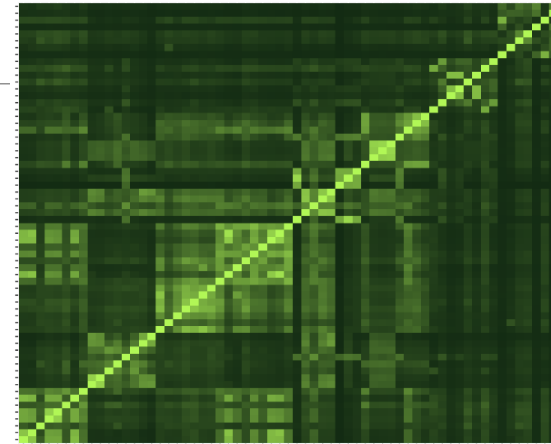
# Exercise 6

4. Compare Unifrac and weighted Unifrac, what can you conclude ?

- Unifrac higher/darker than weighted Unifrac so distance between samples are more important
  - taking into account the abundances makes the samples less distant (lighter)
- abundant OTUs in both communities are phylogenetically closed.

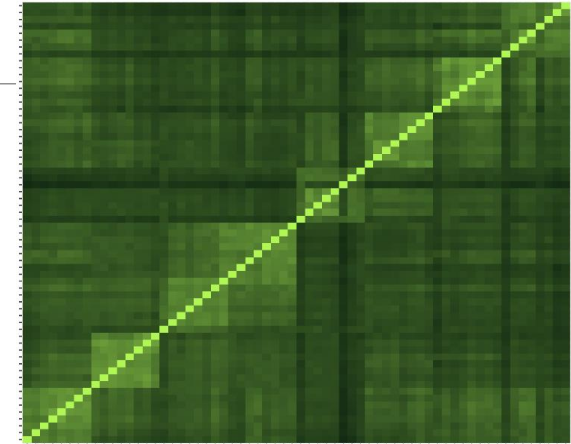
Bray-Curtis

Heatmap plot of the beta distance : bray

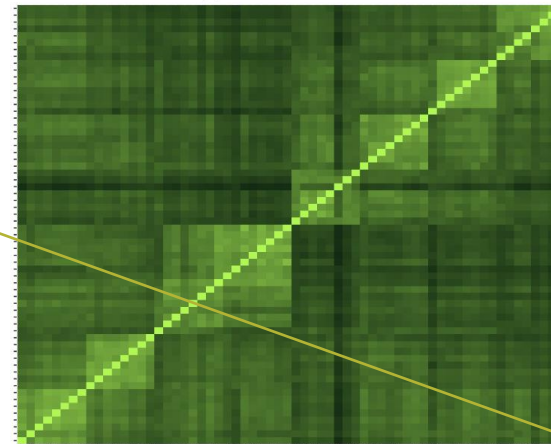


Jaccard

Heatmap plot of the beta distance : cc

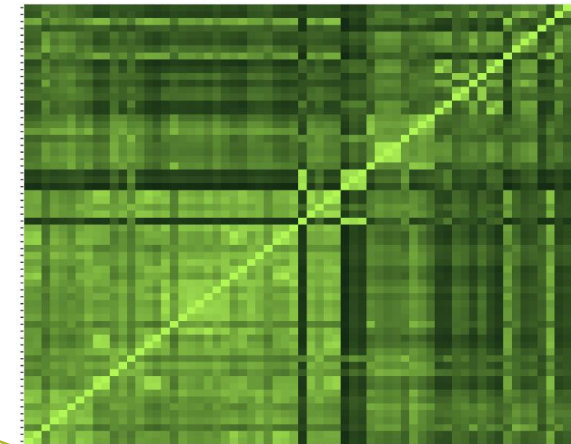


Heatmap plot of the beta distance : unifrac



Unifrac

Heatmap plot of the beta distance : wunifrac



Weighted-Unifrac

# Exploring biodiversity : $\beta$ -diversity

---

- In general, **qualitative** diversities (Jaccard, Unifrac) **are more sensitive to factors that affect presence/absence** of organisms (such as pH, salinity, depth, etc) and therefore are useful to study and define bioregions (regions with little or no flow between them)...
- ... whereas **quantitative** distances (Bray-Curtis, weighted-Unifrac) **focus on factors that affect relative changes** (seasonal changes, nutrient availability, concentration of oxygen, depth, etc.) and therefore useful to monitor communities over time or along an environmental gradient.

Different distances capture different features of the samples.

There is no "one size fits all"



---

# Exploring the structure

---

We will try to identify structures, relationships between samples related to environmental factors

---

# I. Structure Visualisation

---

## ORDINATION AND HEATMAP PLOTS

We have calculated distances now, we will use ordination methods to explore them.

# Structure visualization : with PCA ?

---

- Each community can be described by its OTU abundances, which could be used for a PCA, but high number of OTU make interpretations difficult
- Moreover, PCA maximizes variance and can therefore emphasize differences of rare OTUs between samples instead of giving a good representation of resemblances.  
*Variance is not a very good measure of  $\beta$ -diversity.*
- PCA is not design to use diversity indices and/or distances as it requires independency between variables and does not fit to distance matrix, which is not constructed with samples and variables.  
 $\beta$ -diversity indices thus required dedicated PCA-like methods.

**Purpose of the tool** : ordinate samples based on  $\beta$ -diversity indices and offer tools to visualize it: produce *ordination plots* and *heatmaps*.



# Structure visualization : Ordination plot

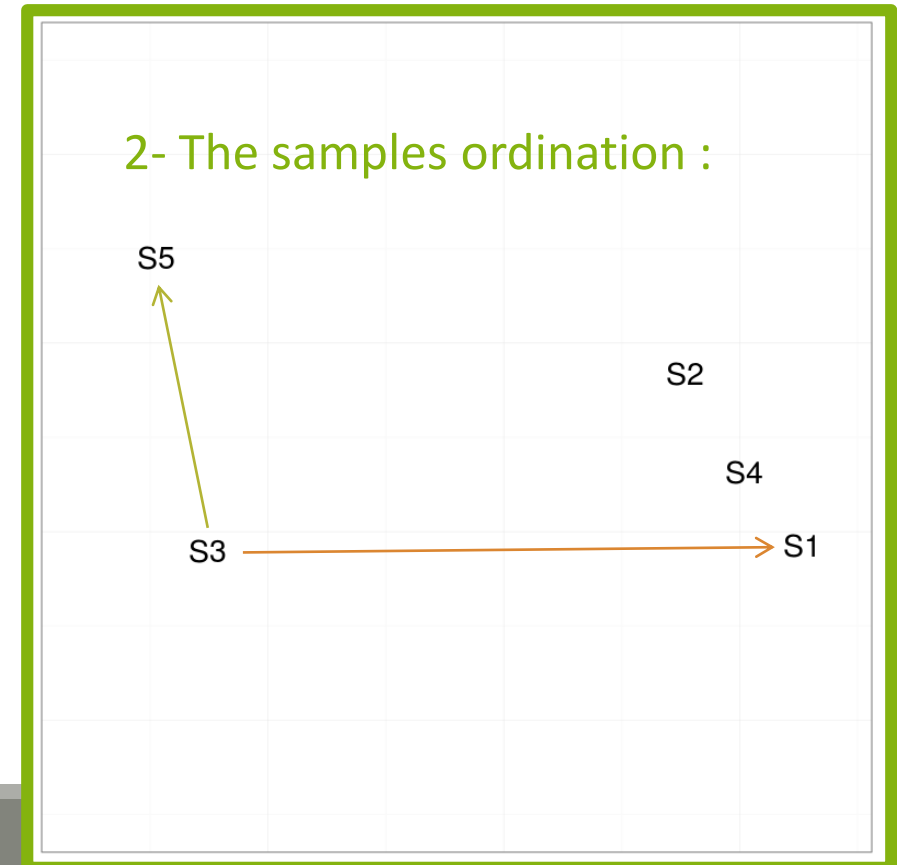
The Multidimensional Scaling (**MDS** or **PCoA**) is equivalent to a Principal Component Analysis (PCA) but preserves the  $\beta$ -diversity instead of the variance.

The MDS tries to represent samples in two dimensions while preserving the distances

1- calculation of a distance matrix.

Distance Matrix					
	S1	S2	S3	S4	S5
S1	0.00	2.21	6.31	0.99	7.50
S2	2.21	0.00	5.40	1.22	5.74
S3	6.31	5.40	0.00	5.75	3.16
S4	0.99	1.22	5.75	0.00	6.64
S5	7.50	5.74	3.16	6.64	0.00

2- The samples ordination :



# Structure visualization : Heatmap

---

- Heatmap is an other representation of the abundance table.
- It tries to reveal if there is a structure between a group of OTUs and a group of samples.
- Heatmap
  - Finds a meaningful order of the samples and the OTUs
  - Allows the user to choose a custom order (in R)
  - Allows the user to change the colour scale (in R)
  - Produces a ggplot2 object, easy to manipulate and customize

# Structure visualization : Ordination plot and Heatmap

**FROGSSTAT Phyloseq Structure Visualisation** with heatmap plot and ordination plot Options  
(Galaxy Version 3.2.2)

**Phyloseq object (format rdata)**  
28: Phyloseq\_raref.Rdata  
This is the result of FROGS Phyloseq Import Data Tool.

**The beta diversity distance matrix file**  
37: Beta Diversity cc.tsv  
These file is the result of FROGS Phyloseq Beta Diversity tool.

**Experiment variable**  
EnvType  
The experiment variable that you want to analyse.

**Ordination method**  
MDS/PCoA

Execute

Explore the sample **NORMALISED** count

To see all, launch **once per distance to ordinate** (Bray, Jaccard, Unifrac and Weighted-Unifrac matrices)

Choose a sample variable to organize graphics

Choose the ordination method (most commonly used is MDS/PCoA)

# Structure visualization : Ordination plot and Heatmap

---

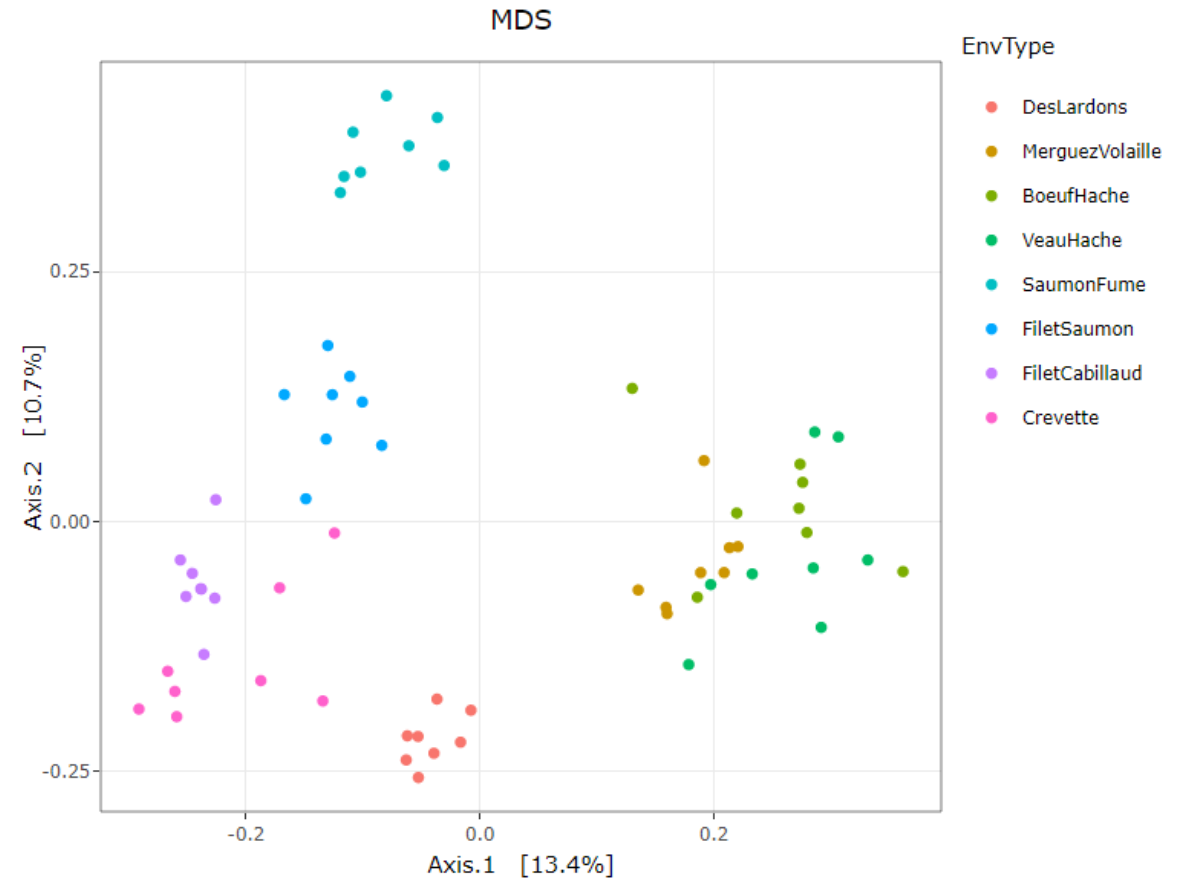
Try it with the 4 distance matrices

1. What are the output datasets ?
2. What is the best distance matrix to use to better separate samples ?
3. Guess why Lardon are somewhere between Meat and Seafood ?
4. Based on your favourite distance matrix, what can you conclude on the heatmap ?

# Structure visualization : Ordination plot and Heatmap

1. What are the output datasets ?

→ HTML report: ordination plot

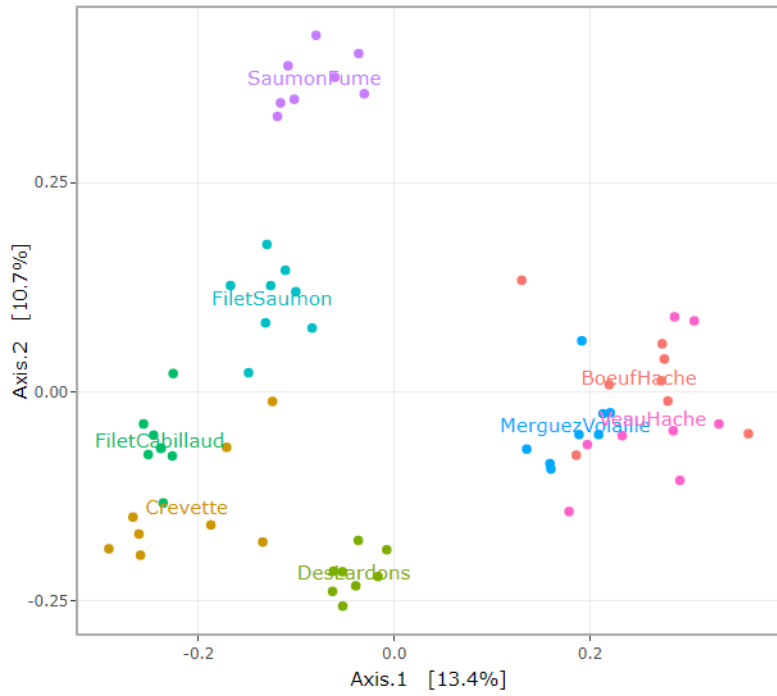


# Structure visualization : Ordination plot and Heatmap

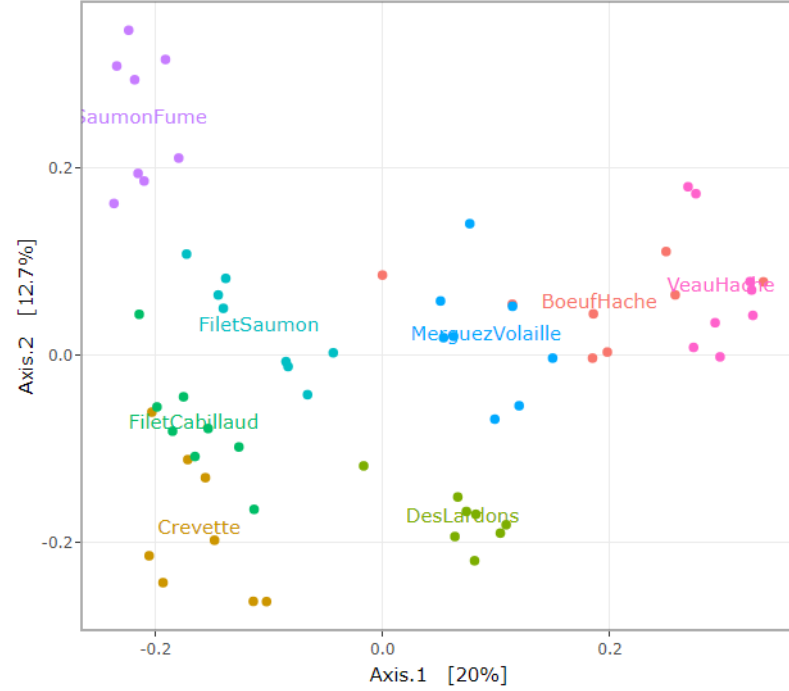
---

2. What is the best distance matrix to use to better separate samples ?

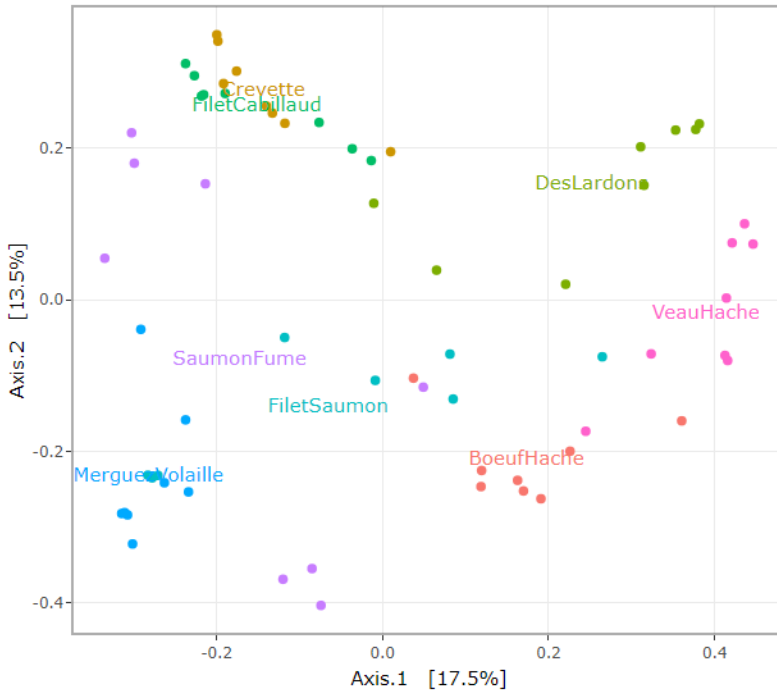
JACCARD



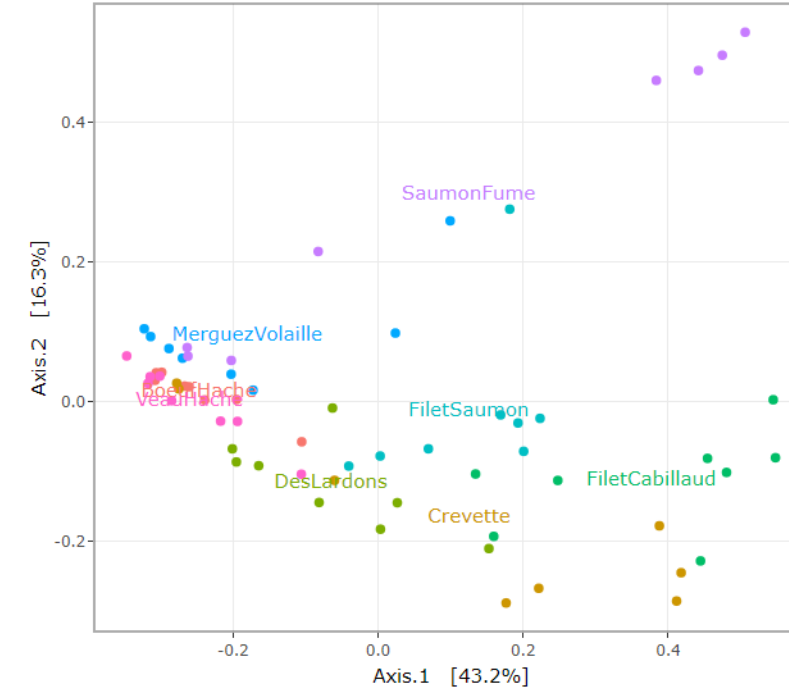
UNIFRAC



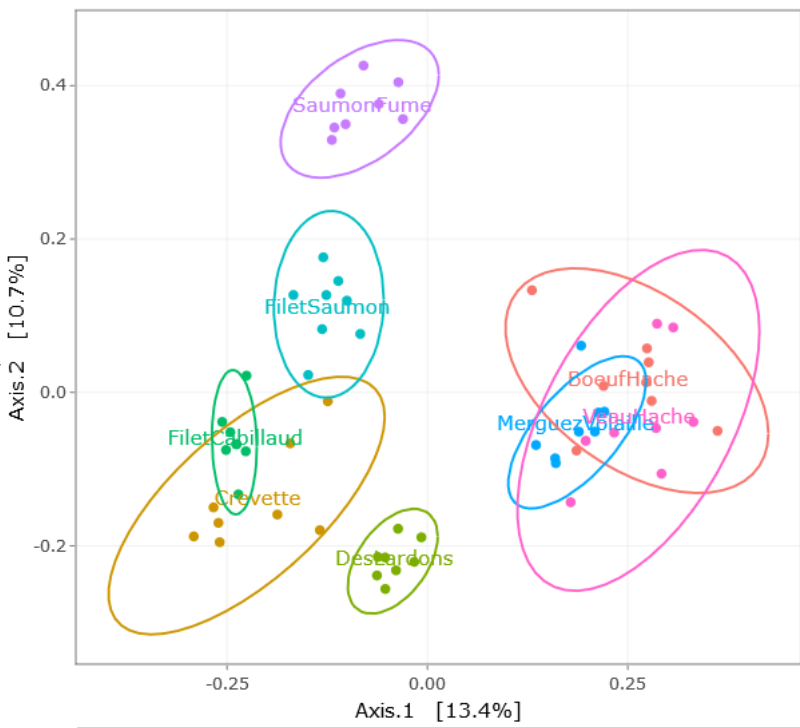
BRAY



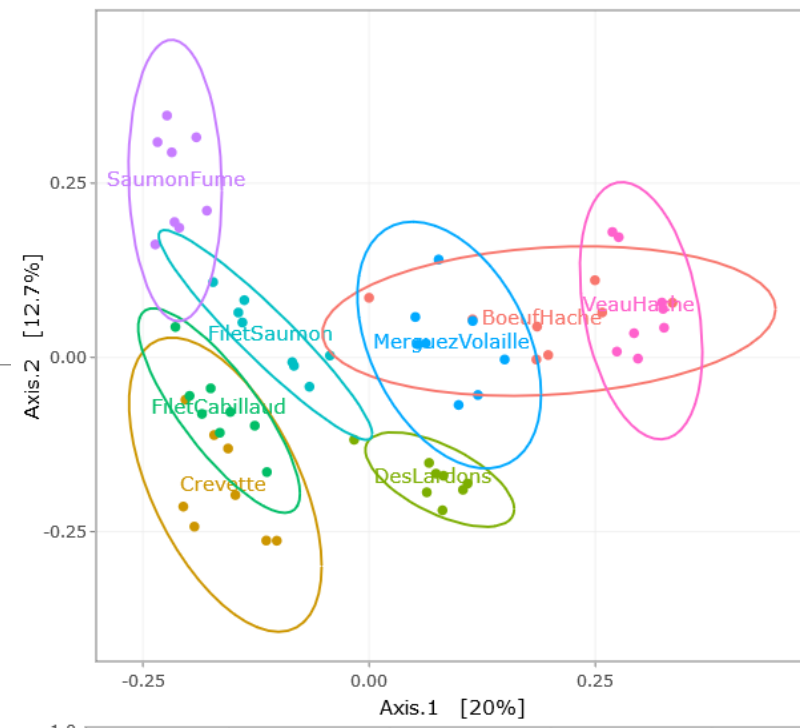
WUNIFRAC



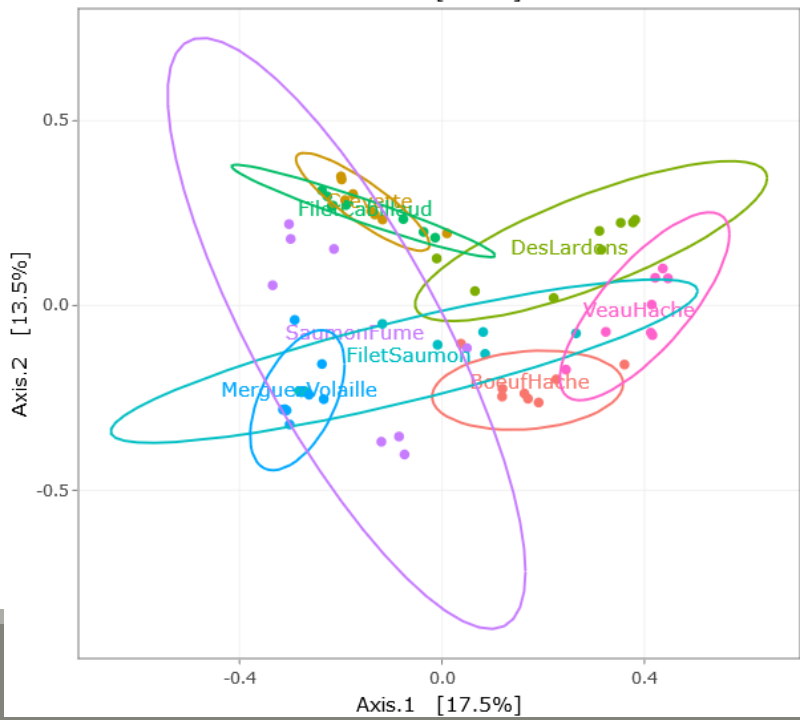
JACCARD



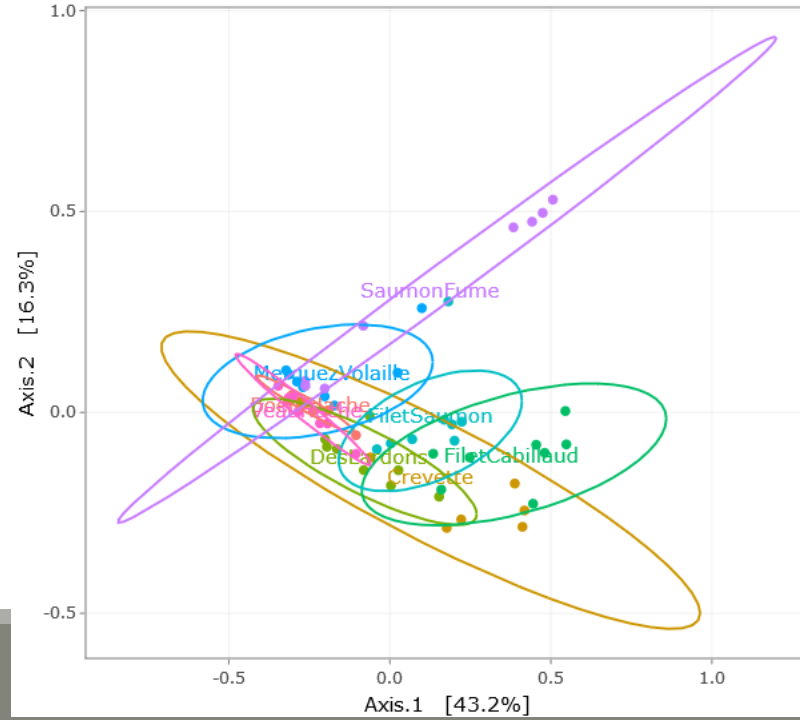
UNIFRAC



BRAY

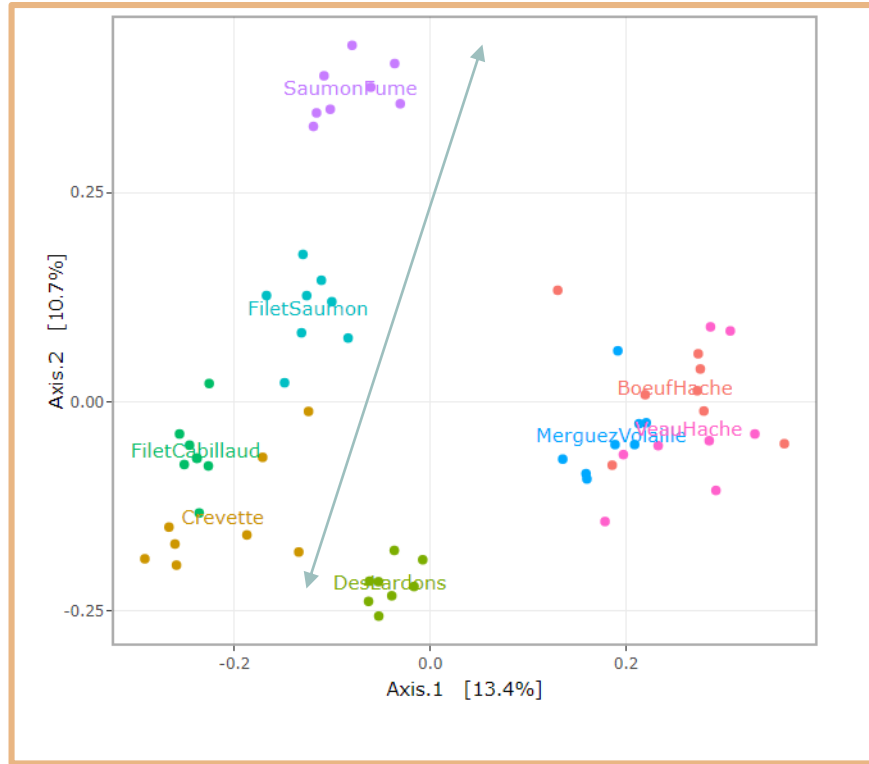


WUNIFRAC

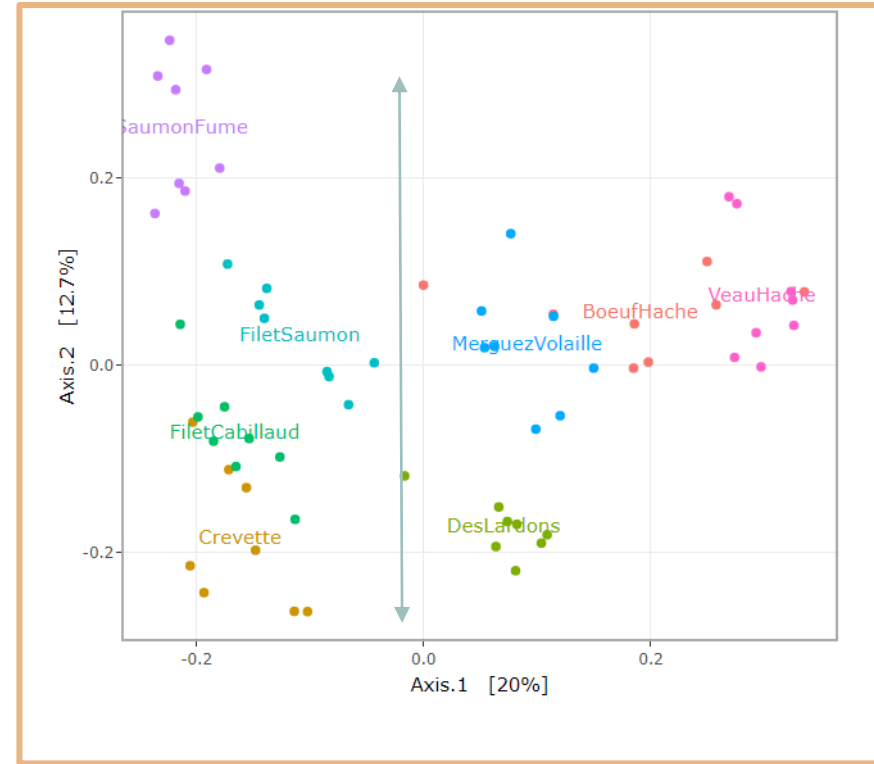




JACCARD



UNIFRAC

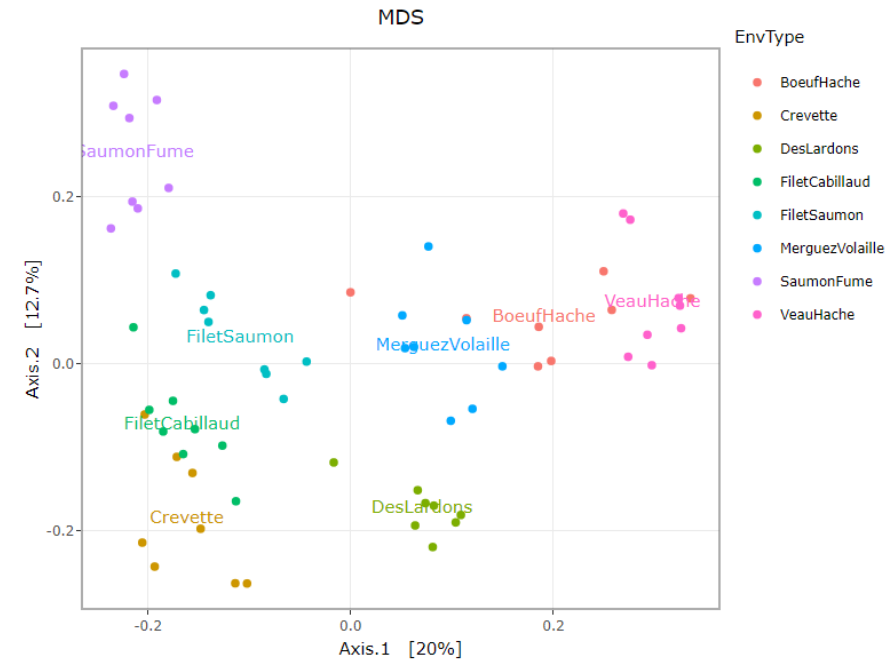
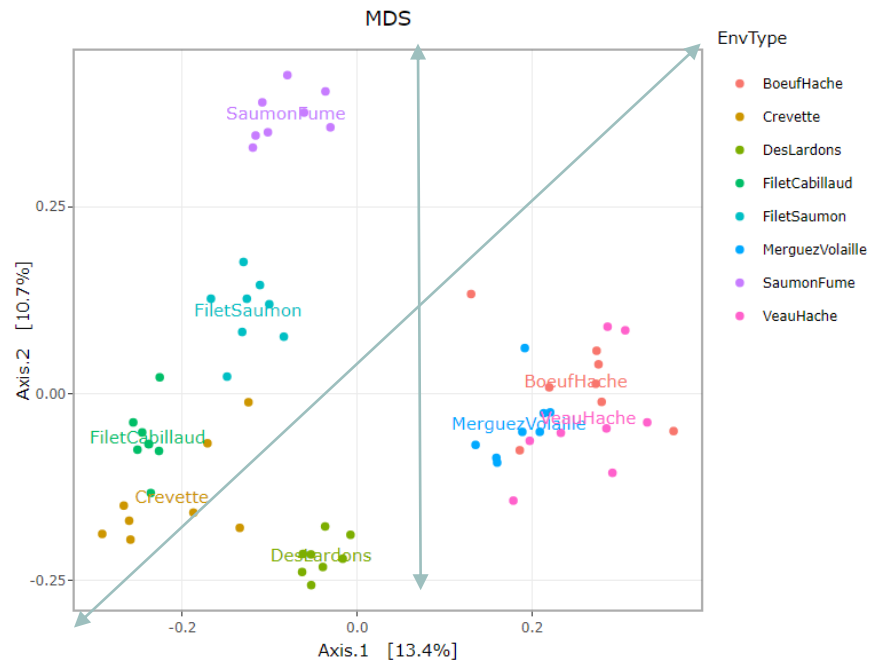


- Qualitative distances (Unifrac, Jaccard) separate meat products from seafood ones
- ➔ detected taxa segregate by origin

# Structure visualization : Ordination plot and Heatmap

## 3. Guess why Lardon are somewhere between Meat and Seafood ?

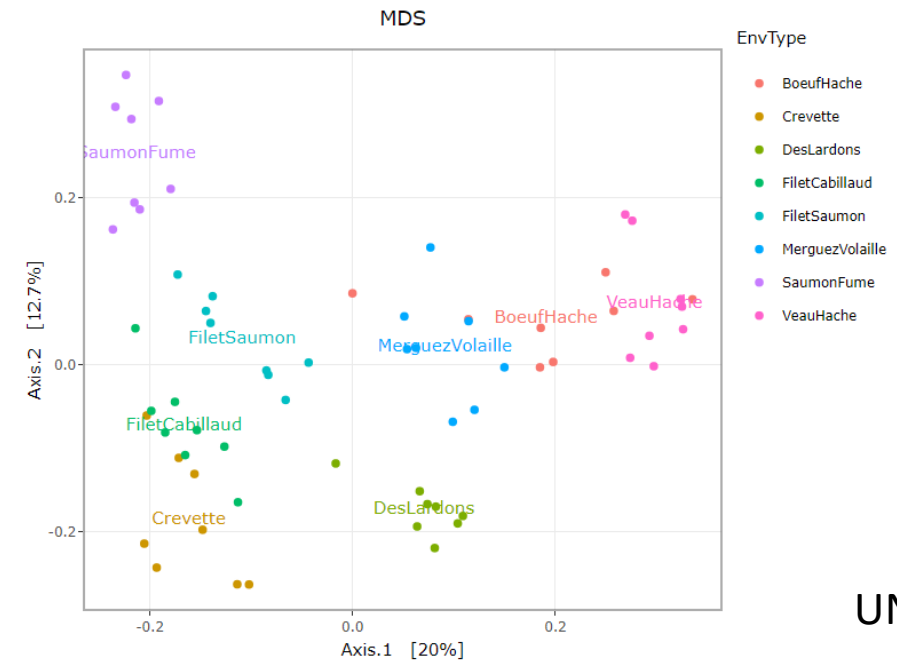
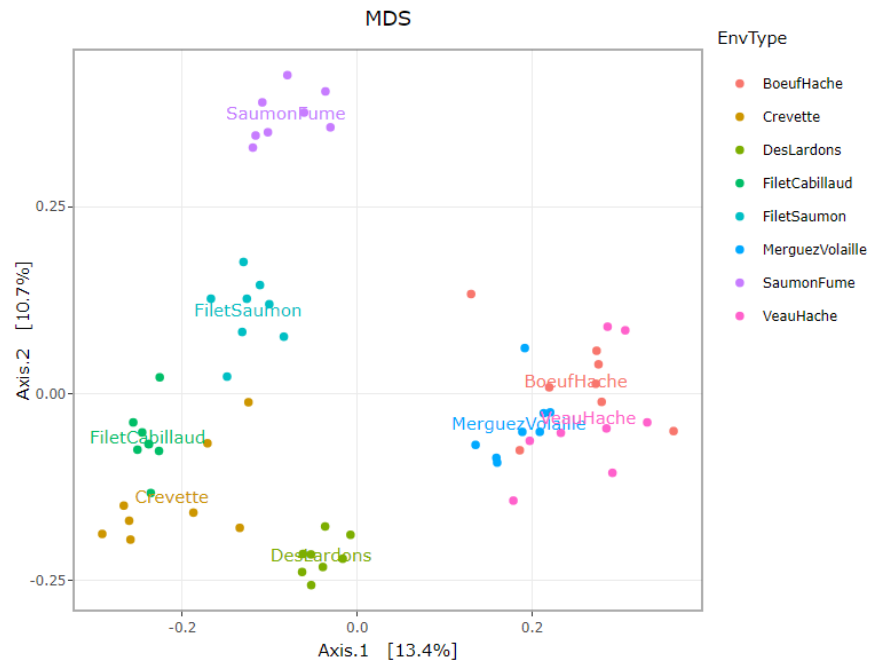
JACCARD



UNIFRAC

# Structure visualization : Ordination plot and Heatmap

## 3. Guess why Lardon are somewhere between Meat and Seafood ?



■ DesLardons is somewhere in between

➔ contamination induced by sea salt

# Structure visualization : Ordination plot and Heatmap

---

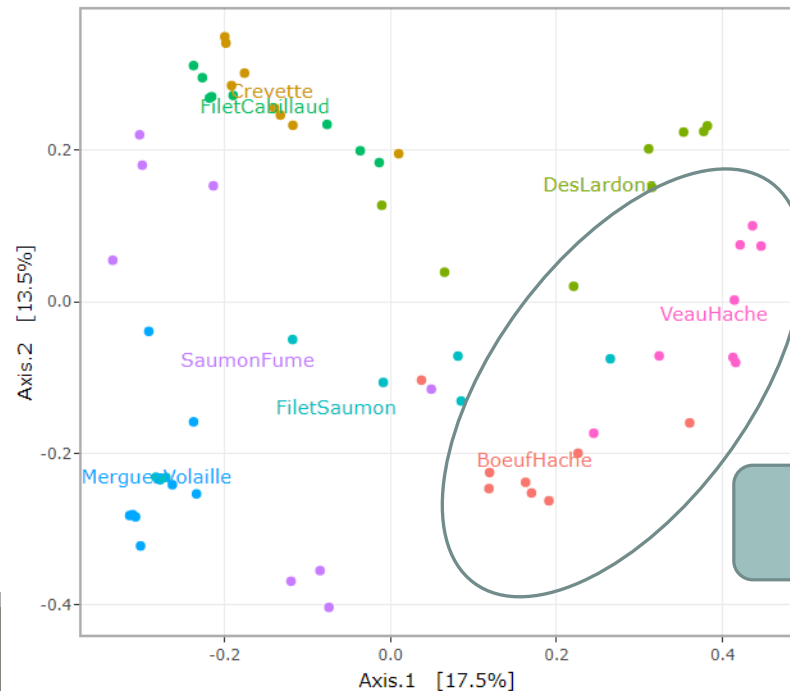
Other conclusions ?

1. Quantitative distances (weighted Unifrac ) exhibit a 'meat – seafood' gradient (on axis 1) with DesLardons in the middle and a 'SaumonFume - everything else' gradient on axis 2.

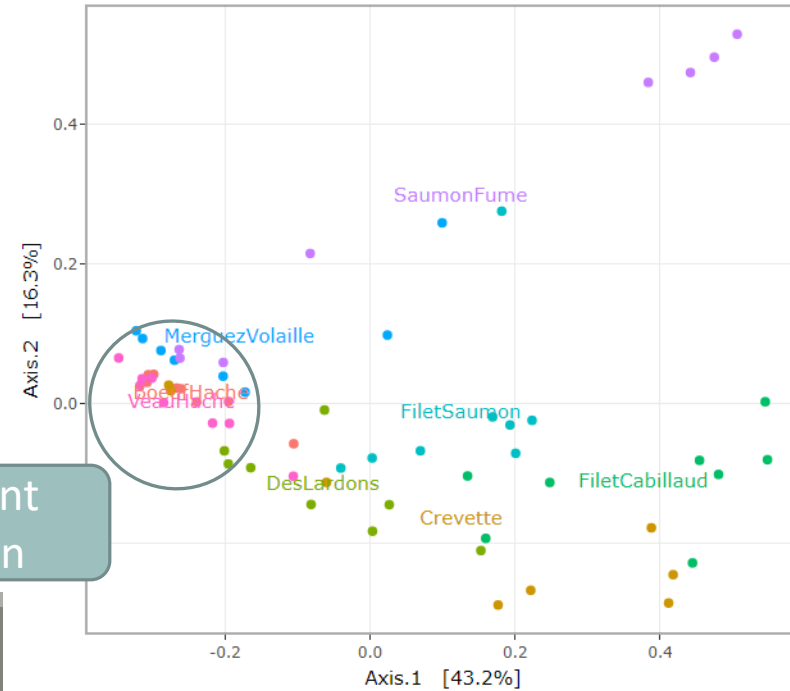
# Structure visualization : Ordination plot and Heatmap

Other conclusions ?

- Note the difference between weighted-UniFrac and Bray-Curtis (2 quantitative indices) for the distances between BoeufHache and VeauHache.



Very different visualization



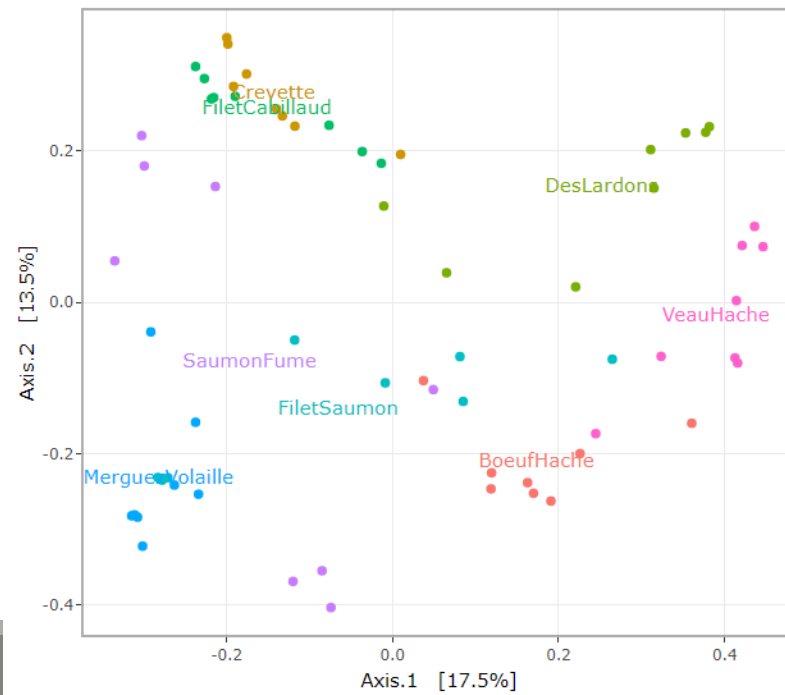
BRAY

WUNIFRAC

# Structure visualization : Ordination plot and Heatmap

Other conclusions ?

3. On Bray-Curtis, on axis 2, we can observe the distribution of Saumon Fumé samples. Axis 1 shows the distribution of MerguezdeVolaille samples



BRAY

# Structure visualization : Ordination plot and Heatmap

---

Other conclusions ?



The 2D representation captures only parts of the original distances

Ellipse are not always an advantage for visualization because it accentuates the 2D effect

# Structure visualization : Ordination plot and Heatmap

---

4. Based on your favourite distance matrix, what can you conclude on the **heatmap** ?

Try to identify:

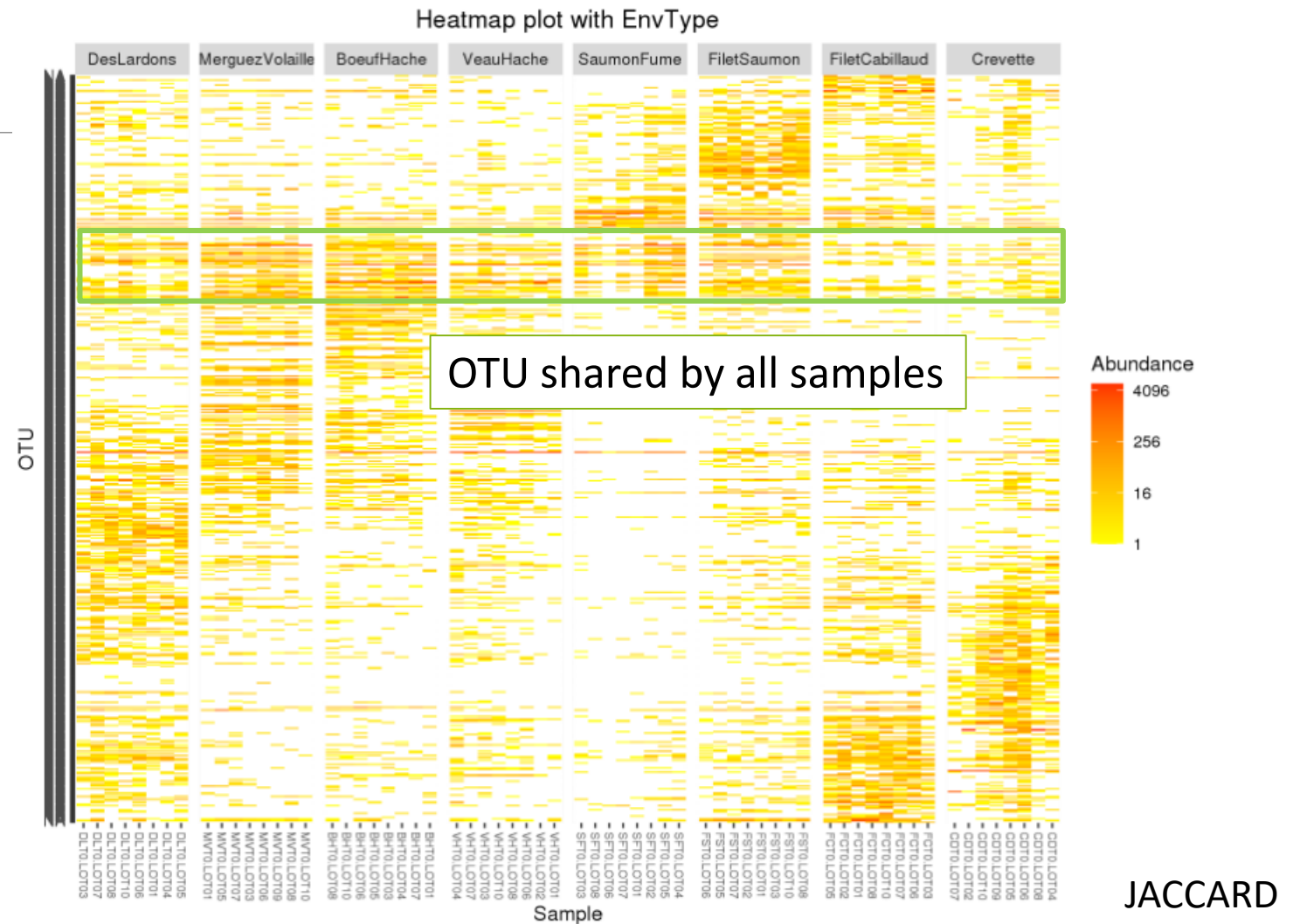
- Block-like structure of the abundance table
- Interaction between (groups of) taxa and (groups of) samples
- Core and condition-specific microbiota



# Exercise 7

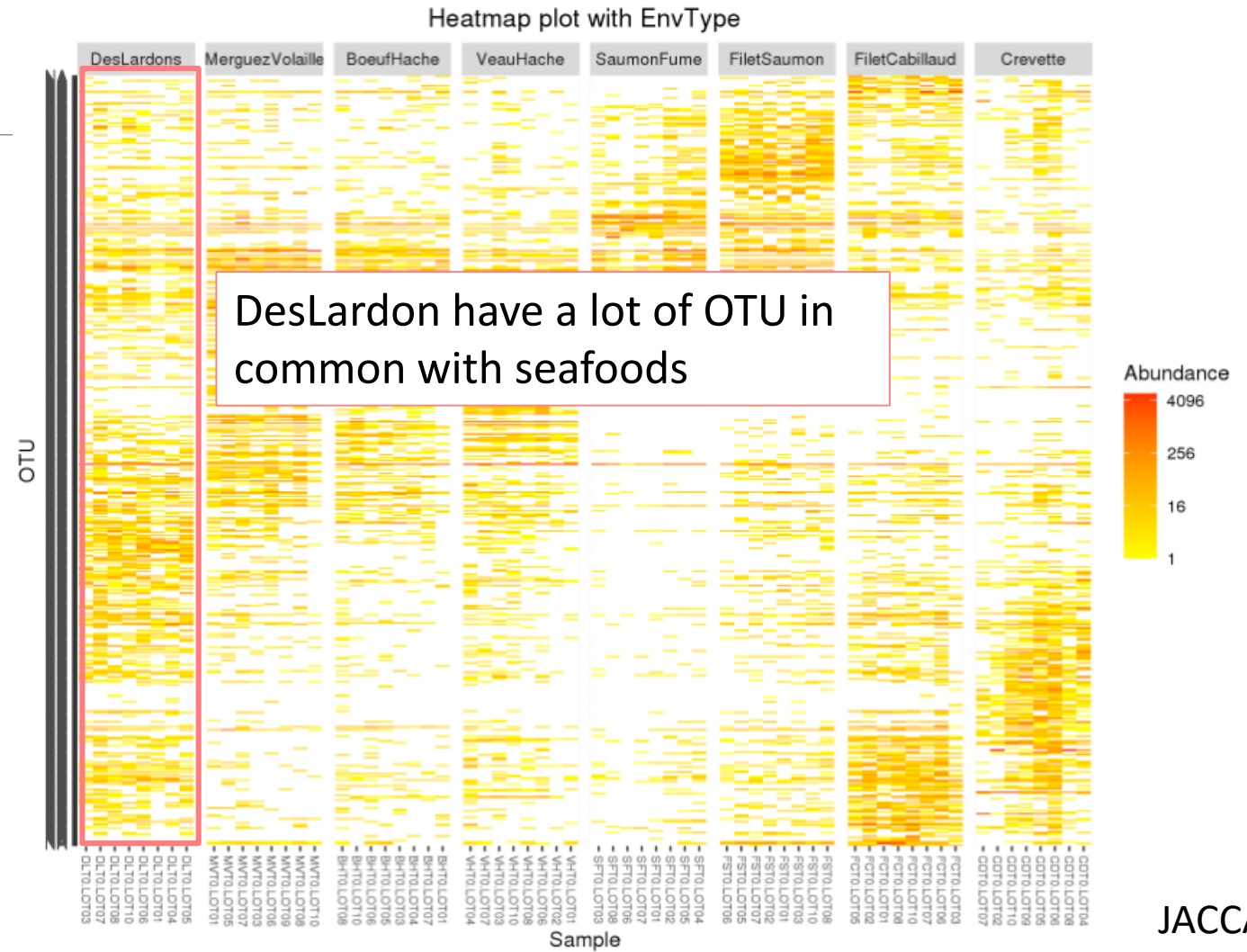
4. Based on your favourite distance matrix, what can you conclude on the heatmap ?

matrix based on Jaccard distance (qualitative) which "sorts" the OTUs. Then a color is applied according to the abundance of OTUs (yellow to red).



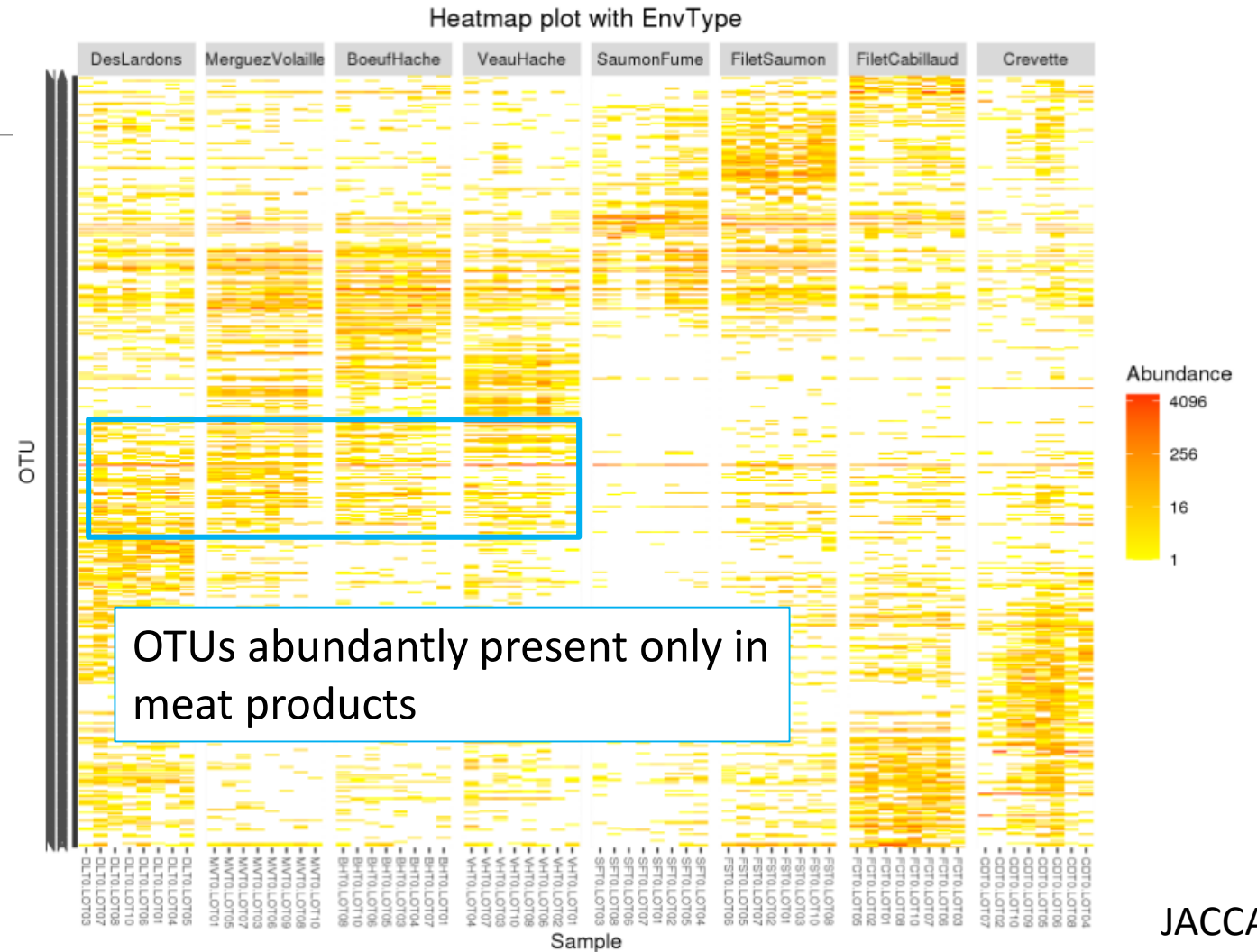
# Exercise 7

4. Based on your favourite distance matrix, what can you conclude on the heatmap ?



# Exercise 7

4. Based on your favourite distance matrix, what can you conclude on the heatmap ?



Note: no evidence for seafood.

---

# II. Exploring the structure

---

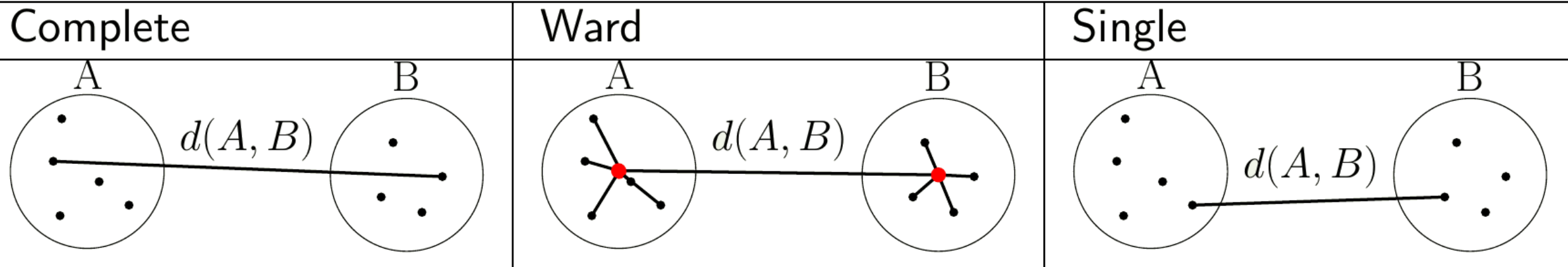
HIERARCHICAL CLUSTERING

# Exploring the structure : clustering

Clustering aims to represent samples in a tree based on a distance matrix and a linkage function:

3 clustering algorithms:

- **Complete linkage:** tends to produce compact, spherical clusters and guarantees that all samples in a cluster are similar to each other.
- **Ward:** tends to also produce spherical clusters but has better theoretical properties than complete linkage.
- **Single:** friend of friend approach, tends to produce banana-shaped or chains-like clusters.



# Exploring the structure : clustering

---

**FROGSSTAT Phyloseq Sample Clustering** of samples using different linkage methods (Galaxy Version 3.2.2) Options

**Phyloseq object (format rdata)**  
28: Phyloseq\_raref.Rdata  
This is the result of FROGS Phyloseq Import Data tool.

**The beta diversity distance matrix file**  
38: Beta Diversity unifrac.tsv  
This file is the result of FROGS Phyloseq Beta Diversity tool.

**Experiment variable**  
EnvType  
The experiment variable that you want to analyse.

Execute

Explore the sample **NORMALISED** count

Choose the beta diversity distance matrix: i.e. Unifrac

Choose a sample variable to organize graphics: i.e. EnvType

The three different linkage functions will be used, generating three different dendrograms

# Exercise 8

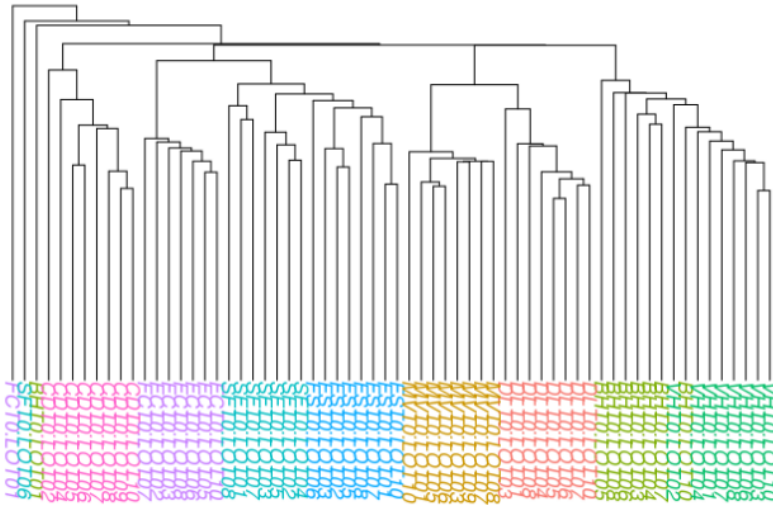
---

Try it with « a good » distance method matrix on EnvType and on FoodType

→ Which linkage method seems to better fit the data ?

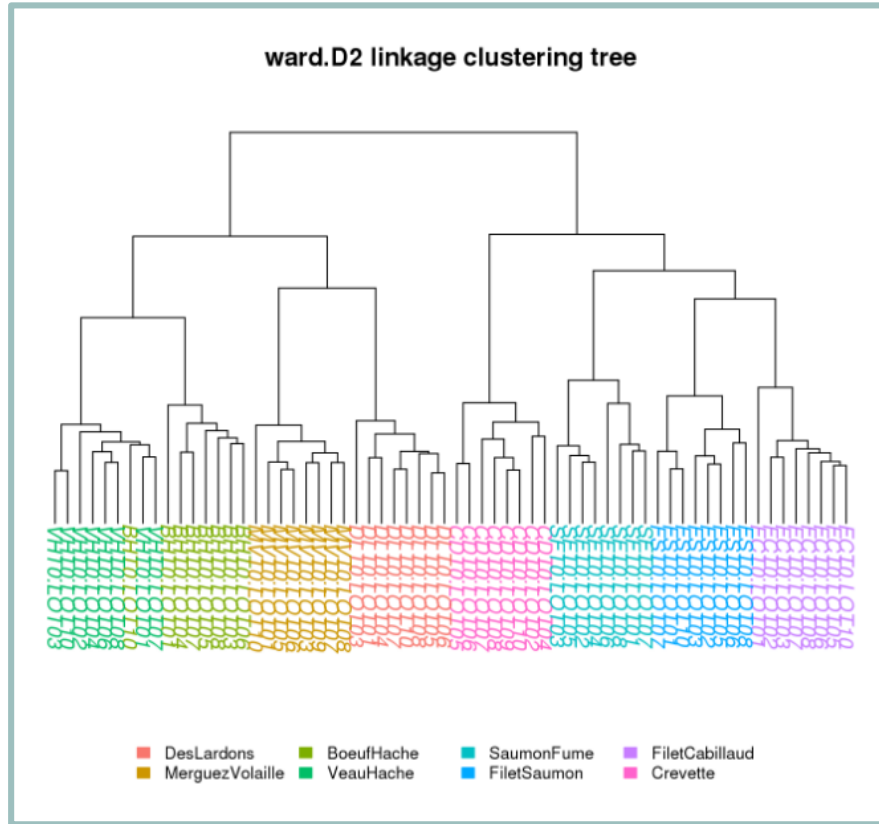
# Exercise 8

single linkage clustering tree



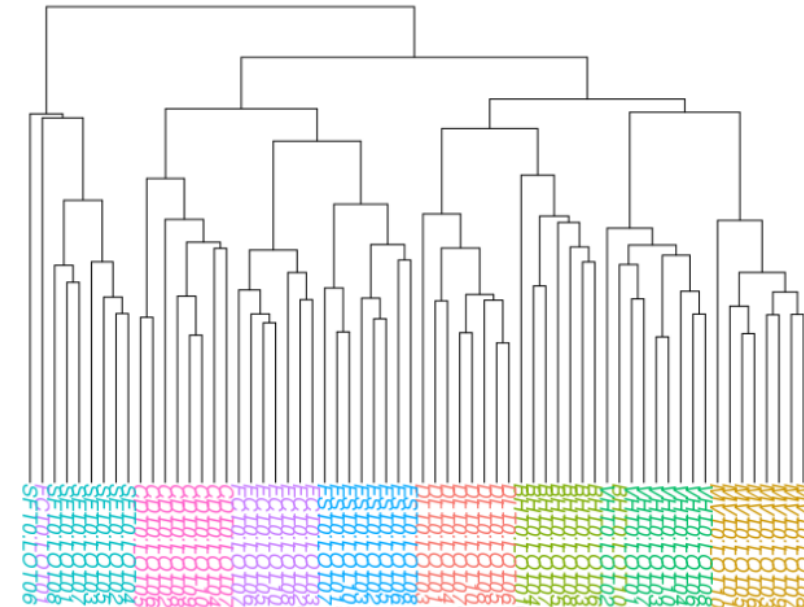
■ DesLardons    ■ BoeufHache    ■ SaumonFume    ■ FiletCabillaud  
■ MerguezVolaille    ■ VeauHache    ■ FiletSaumon    ■ Crevette

ward.D2 linkage clustering tree



■ DesLardons    ■ BoeufHache    ■ SaumonFume    ■ FiletCabillaud  
■ MerguezVolaille    ■ VeauHache    ■ FiletSaumon    ■ Crevette

complete linkage clustering tree



■ DesLardons    ■ BoeufHache    ■ SaumonFume    ■ FiletCabillaud  
■ MerguezVolaille    ■ VeauHache    ■ FiletSaumon    ■ Crevette

the Ward clustering allows to classify the communities according to the EnvType groups



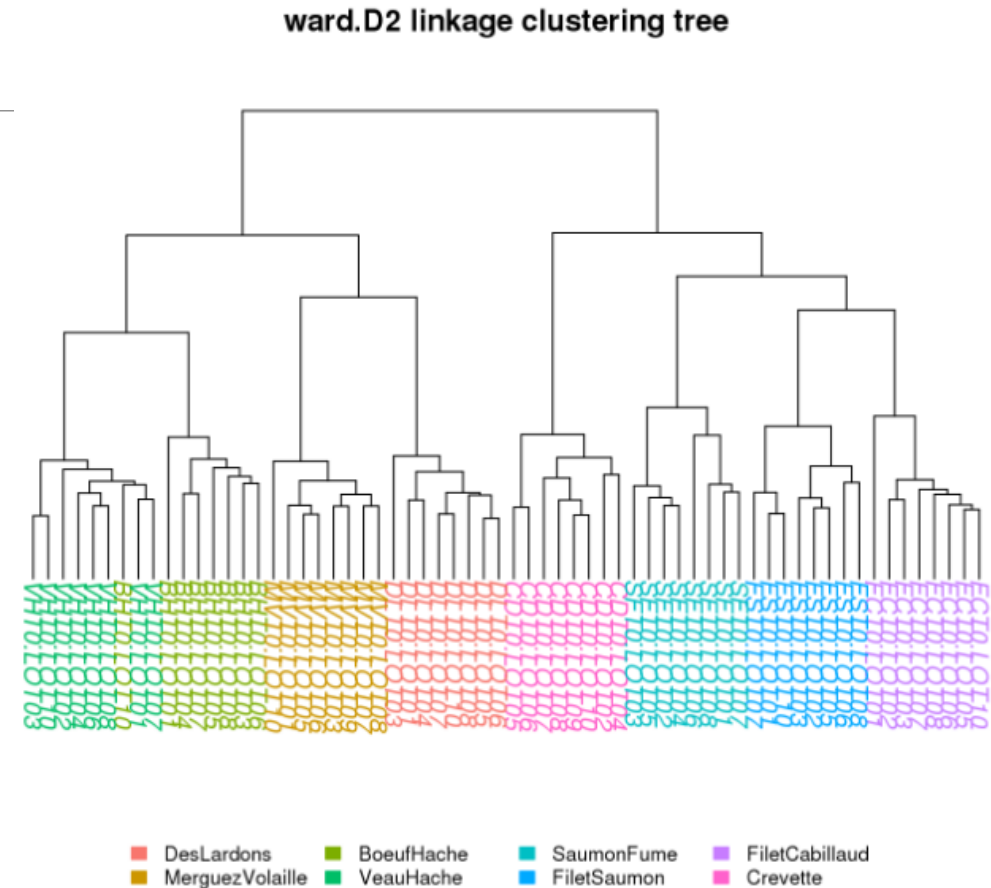
# Exercise 8

---

- Consistently, for these datasets, with the ordination plots, clustering works quite well for the **UniFrac** distance
- The method (Ward.D2) give almost a perfect separation between the different type of food

## Remarks

Clustering is based on the whole distance whereas ordination represents parts of the distance (the most it can with 2 dimensions)



Ward D2

Complete

Single

# Exercise 8

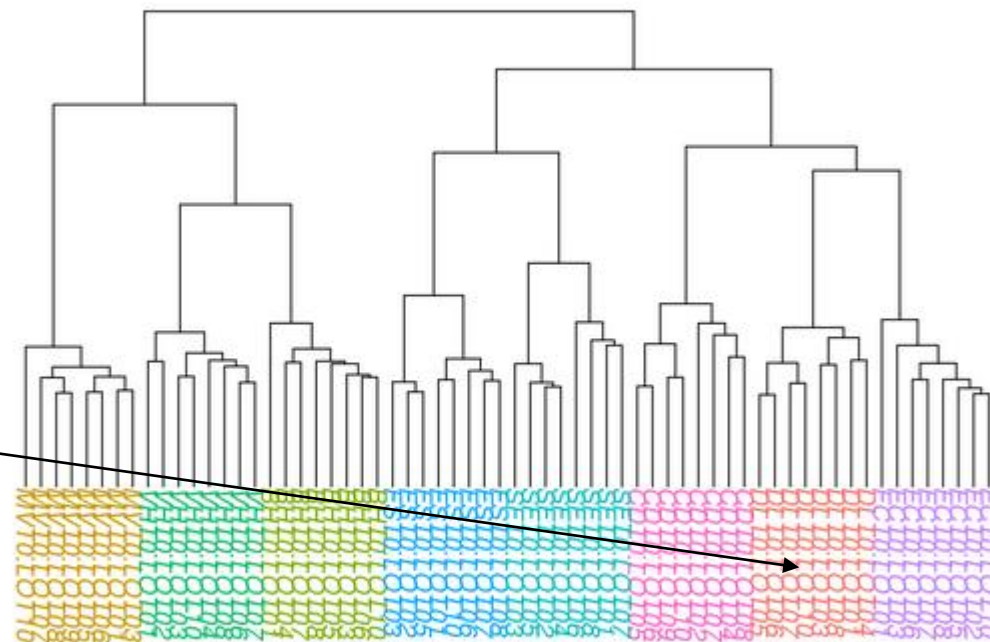
- Not as well clustered with **Jaccard** indices
- DesLardons is in the middle of seafood.

Once again,

Different distances capture different features  
of the samples.

There is no "one solution fits all"

ward.D2 linkage clustering tree



■ DesLardons	■ BoeufHache	■ SaumonFume	■ FiletCabillaud
■ MerguezVolaille	■ VeauHache	■ FiletSaumon	■ Crevette

---

# Diversity partitioning

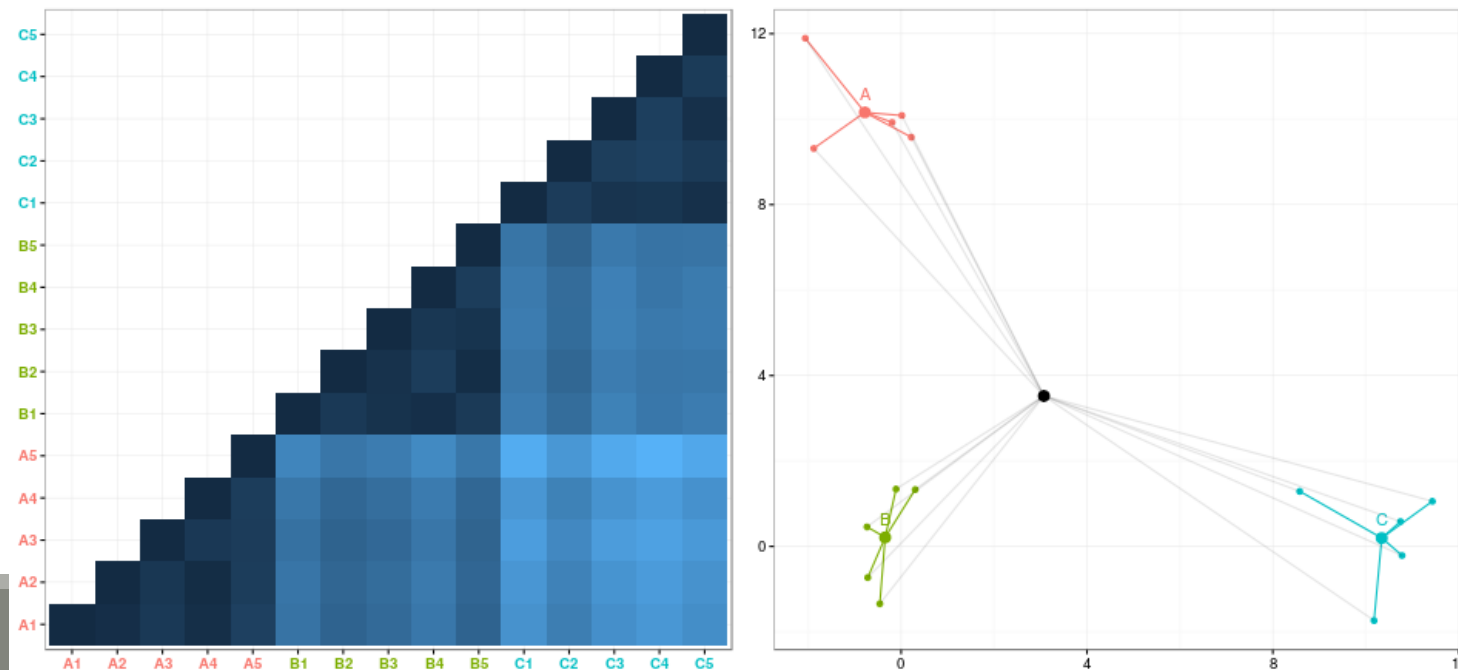
---

# Diversity partitioning

Do the structures seem linked to metadata ? Does the metadata have an effect on the composition of our communities ?

To answer these questions, **multivariate analyses** :

- test **composition differences** of communities from different groups **using a distance matrix**
- compare **within-group** to **between-group** distances

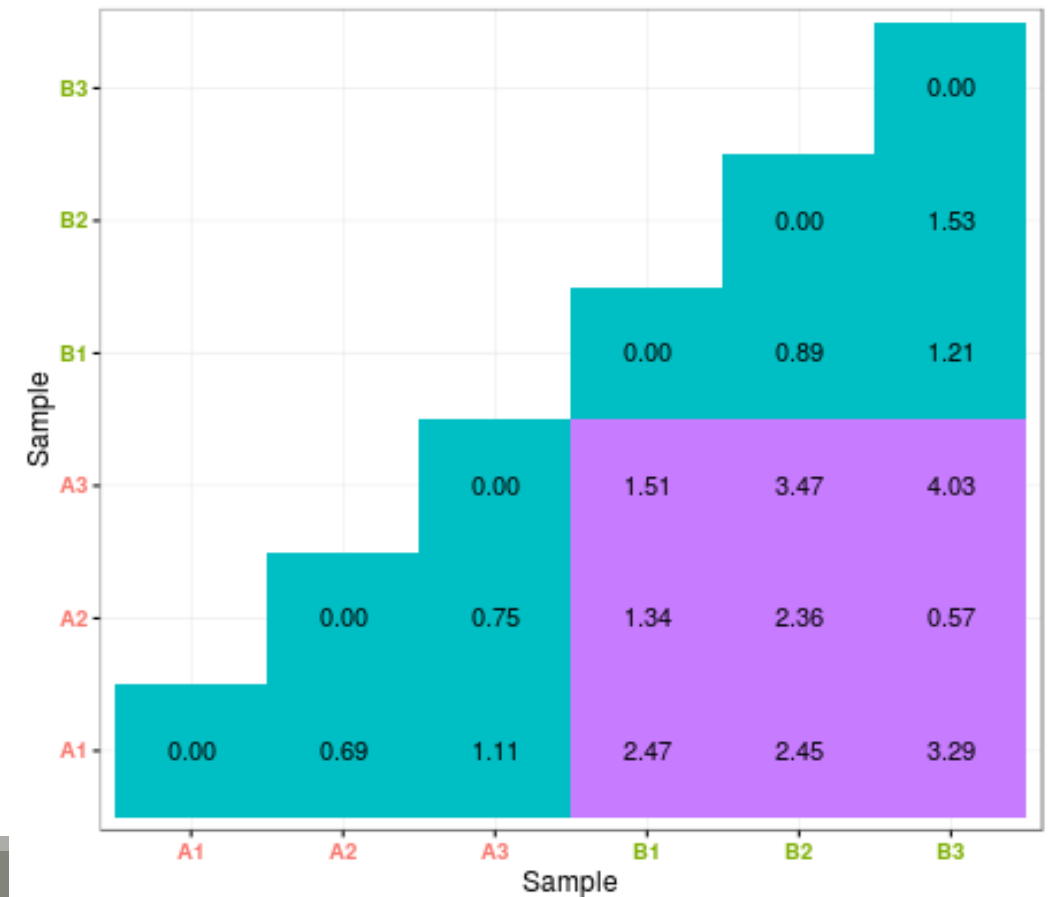


# Diversity partitioning : Multivariate ANOVA

Idea : Test **differences** in the community composition **from different groups** using a **distance matrix**.

## How it works ?

- Computes sum of square distance
- Variance analysis



# Diversity partitioning : Multivariate ANOVA

**FROGSSTAT Phyloseq Multivariate Analysis Of Variance** perform Multivariate Analysis of Variance (MANOVA) (Galaxy Version 3.2.2) Options

**Phyloseq object (format rdata)**  
28: Phyloseq\_raref.Rdata  
This is the result of FROGS Phyloseq Import Data tool.

**The beta diversity distance matrix file**  
38: Beta Diversity unifracs.tsv  
This file is the result of FROGS Phyloseq Beta Diversity tool.

**Experiment variable**  
EnvType  
The experiment variable that you want to analyse.

Execute

Explore the sample **NORMALISED** count

Choose the beta diversity distance matrix: Unifrac

Choose the variable to explain the variability between samples: EnvType

To simultaneously test several variables, you can use “+” symbol as “EnvType+FoodType” to test only additive effects or “\*” symbol as “EnvType\*FoodType” to test for additive effects and interactions between variables

# Exercise 9

---

Try it with a good beta distance matrix with EnvType and FoodType

1. Does EnvType have an influence on the beta diversity variance ?
2. What about FoodType ?

# Exercise 9

---

1. Does EnvType have an influence on the beta diversity variance ?

With Unifrac distance

```
Call:
adonis(formula = dist ~ EnvType, data = metadata, permutations = 9999)

Permutation: free
Number of permutations: 9999

Terms added sequentially (first to last)

              Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
EnvType       7     6.1849 0.88356  11.164 0.58255 1e-04 ***
Residuals    56     4.4320 0.07914           0.41745
Total        63    10.6170           1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# Exercise 9

---

1. Does EnvType have an influence on the beta diversity variance ?

Environment type explains roughly **58%** of the total variation, which is very high

With Unifrac distance

```
Call:
adonis(formula = dist ~ EnvType, data = metadata, permutations = 9999)

Permutation: free
Number of permutations: 9999

Terms added sequentially (first to last)

      Df SumsOfSqs MeanSqs F.Model    R2 Pr(>F)
EnvType  7    6.1849  0.88356  11.164 0.58255 1e-04 ***
Residuals 56    4.4320  0.07914    0.41745
Total    63   10.6170          1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Exercise 9

---

## 2. What about FoodType ?

```
Call:
adonis(formula = dist ~ FoodType, data = metadata, permutations = 9999)
```

```
Permutation: free
Number of permutations: 9999
```

With Unifrac distance

```
Terms added sequentially (first to last)
```

	Df	SumsOfSqs	MeanSqs	F.Model	R2	Pr(>F)	
FoodType	1	1.7858	1.78579	12.537	0.1682	1e-04	***
Residuals	62	8.8312	0.14244		0.8318		
Total	63	10.6170			1.0000		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Exercise 9

---

## 2. What about FoodType ?

Food type explains only **17 %** of the total variation

With Unifrac distance

```
Call:
adonis(formula = dist ~ FoodType, data = metadata, permutations = 9999)

Permutation: free
Number of permutations: 9999

Terms added sequentially (first to last)

      Df SumsOfSqs MeanSqs F.Model    R2 Pr(>F)
FoodType  1    1.7858 1.78579  12.537 0.1682 1e-04 ***
Residuals 62    8.8312 0.14244    0.8318
Total    63   10.6170          1.0000
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '.' 0.1 ' ' 1
```

---

# Differential abundance analysis

---

# Differential abundance analysis

---

Are there OTU with differential abundance between 2 conditions ? And which are they ?

To answer these questions, we perform a differential abundance analysis using DESeq2 on the phyloseq object

The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models

# Differential abundance analysis

---

Are there OTU with differential abundance between 2 conditions ? And which are they ?

To answer these questions, we perform a differential abundance analysis using DESeq2 on the phyloseq object

The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models



Be aware to use data *without normalisation*

**DESeq** has its own normalisation method suited to this kind of data.

It uses the poscount function optimised for metagenomic count table

# Differential abundance analysis

→ 1<sup>st</sup> step: launch *DESeq2 Preprocess* tool to create the **dds object** – the DESeq2 object

**FROGSSTAT DESeq2 Preprocess** import a Phyloseq object and prepare it for DESeq2 differential abundance analysis (Galaxy Version 3.2.2) Options

**Phyloseq object (format rdata)**

26: Phyloseq.Rdata

This is the result of FROGSSTAT Phyloseq Import Data with normalise option set to NO (DESeq2 is more powerful on unnormalised counts).

**Experimental variable**

EnvType

The factor suspected to have an effect on OTUs' abundances. Ex: Treatment, etc.

**Do you want to correct for a confounding factor?**

If yes, specify confounding factor.

Explore the sample **RAW** count

Choose the factor on which the differential abundances will be compared

Specify a confounding factor if necessary  
(*example : testing antibiotic treatment effect with 2 different mice phenotypes, or testing drought effect on soil microbiome with two soil compositions*)

# Differential abundance analysis

---

→ What are the output datasets ?

→ Rdata file: dds object with results of the DESeq analysis

→ 2<sup>nd</sup> step: launch *DESeq2 visualization* tool to explore the dds object



# Differential abundance visualization

**FROGSTAT Deseq2 Visualisation** to extract and visualise differentially abundant OTUs (Galaxy Version 3.2.2) Options

**Phyloseq object (format rdata)**  
26: Phyloseq.Rdata  
This is the result of FROGS Phyloseq Import Data, used in FROGSSTAT DESeq2 Preprocess tool

**DESeq2 object (format rdata)**  
50: DESeq2 dds.Rdata  
This is the result of FROGSSTAT DESeq2 Preprocess tool.

**Experimental variable**  
EnvType  
The factor suspected to have an effect on OTUs' abundances (one of the variables used in FROGS DESeq2 Preprocess tool). Ex : Treatment

**Is your Variable quantitative or qualitative?**  
qualitative  
If qualitative, choose 2 conditions to compare.

**Condition 1 considered as reference**  
BoeufHache  
One condition of the experimental variable (e.g. with).

**Condition 2 to be compared to the reference**  
VeauHache  
Another condition of the experimental variable (e.g. without).

**Adjusted p-value threshold**  
0.05  
Threshold used for statistical significance of the differentially abundant OTUs analysis

Execute

Explore the sample **RAW** count

Result of FROGSSTAT DESeq2 preprocess

Factor on which the differential abundances have been tested

Specify qualitative or quantitative

Precise the two conditions to compare

**Compare BoeufHache vs VeauHache**

Statistical significance threshold (default 0.05)

# Differential abundance visualization

---

What are the output datasets ?

→ HTML report: result table and several plots

Differentially abundant OTU table

Pie chart

Volcano plot

MA plot

Heatmap plot

# Differential abundance visualization

---

Differentially abundant OTU table

Pie chart

Volcano plot

MA plot

Heatmap plot

Since we only have a binary factor we can use the following syntax to format the log2 fold change from the fitted model if not, we will use the other syntax with `contrast=c()`

Code

```
You chose to compare BoeufHache to the reference modality VeauHache. This implies that a positive log2FoldChange means more abundant in BoeufHache than in VeauHache.
```

Then we extract significant OTUs at the p-value adjusted threshold level (after correction) and enrich results with taxonomic informations and sort taxa by pvalue.

# Differential abundance visualization

Differentially abundant OTU table

	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	<input type="text" value="/"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="."/>	<input type="text" value="."/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="/"/>
1	Cluster_53	16.7845	7.93954	1.21935	6.51127	7.45192e-11	2.61562e-8	Bacteria
2	Cluster_43	10.4196	-15.6431	2.48659	-6.29099	3.15453e-10	5.53619e-8	Bacteria
3	Cluster_120	7.49645	-5.21487	0.842194	-6.19200	5.94038e-10	6.95024e-8	Bacteria
4	Cluster_4	284.010	4.46973	0.730032	6.12265	9.20306e-10	8.07569e-8	Bacteria
5	Cluster_85	5.25312	14.8546	2.69005	5.52204	3.35084e-8	0.00000235229	Bacteria
6	Cluster_174	2.99262	17.3671	3.27384	5.30481	1.12788e-7	0.00000659812	Bacteria
7	Cluster_44	22.0406	6.03398	1.14995	5.24715	1.54472e-7	0.00000677746	Bacteria
8	Cluster_141	9.26135	-5.96649	1.13629	-5.25083	1.51415e-7	0.00000677746	Bacteria

Only significantly differentially abundant OTU are displayed (with an adjusted p-value < previously defined threshold - set here to 0.05)

p-value are adjusted using the Benjamini-Hochberg method

# Differential abundance visualization

Differentially abundant OTU table

	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kin
	<input type="text" value="/"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="."/>	<input type="text" value="."/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="/"/>
1	Cluster_53	16.7845	7.93954	1.21935				
				More abundant in BoeufHaché than VeauHaché				
2	Cluster_43	10.4196	-15.6431	2.48659	-6.29099	3.15453e-10	5.53619e-8	Bacteria
3	Cluster_120	7.49645	-5.21487	0.842194				
				More abundant in VeauHaché than BoeufHaché				
4	Cluster_4	284.010	4.46973	0.730032	6.12265	9.20306e-10	8.07569e-8	Bacteria
5	Cluster_85	5.25312	14.8546	2.69005	5.52204	3.35084e-8	0.00000235229	Bacteria
6	Cluster_174	2.99262	17.3671	3.27384	5.30481	1.12788e-7	0.00000659812	Bacteria
7	Cluster_44	22.0406	6.03398	1.14995	5.24715	1.54472e-7	0.00000677746	Bacteria
8	Cluster_141	9.26135	-5.96649	1.13629	-5.25083	1.51415e-7	0.00000677746	Bacteria

# Differential abundance visualization

---

Differentially abundant OTU table

Why  $\log_2$ Foldchange ?

Foldchange:

It's the ratio of the normalized counts between VeauHache and BoeufHache

$\log_2$  is used for interpret and scale reasons:

- Positive values denote an increase, and negative a decrease of abundance
- $\log_2FC = 1$  means a doubling
- $\log_2FC = 2$  means a quadrupling
- $\log_2FC = -1$  means a halving
- $\log_2FC = -2$  means a quartering
- ...

# Differential abundance visualization

Differentially abundant OTU table

	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	<input type="text" value="/"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="."/>	<input type="text" value="."/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="/"/>
1	Cluster_53	16.7845	7.93954	1.21935	6.51127	7.45192e-11	2.61562e-8	Bacteria
2	Cluster_43	10.4196	-15.6431	2.48659	-6.29099	3.15453e-10	5.53619e-8	Bacteria
3	Cluster_120	7.49645	-5.21487	0.842194	-6.19200	5.94038e-10	6.95024e-8	Bacteria
4	Cluster_4	284.010	4.46973	0.730032	6.12265	9.20306e-10	8.07569e-8	Bacteria
5	Cluster_85	5.25312	14.8546	2.69005	5.52204	3.35084e-8	0.00000235229	Bacteria
6	Cluster_174	2.99262	17.3671	3.27384	5.30481	1.12788e-7	0.00000659812	Bacteria
7	Cluster_44	22.0406	6.03398	1.14995	5.24715	1.54472e-7	0.00000677746	Bacteria
8	Cluster_141	9.26135	-5.96649	1.13629	-5.25083	1.51415e-7	0.00000677746	Bacteria

You can sort by log2FoldChange and filter on taxonomy criteria

# Differential abundance visualization

Differentially abundant OTU table

→ Which species have the highest negative log2Foldchange ?

	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	<input type="text" value="/"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="."/>	<input type="text" value="."/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="/"/>
1	Cluster_53	16.7845	7.93954	1.21935	6.51127	7.45192e-11	2.61562e-8	Bacteria
2	Cluster_43	10.4196	-15.6431	2.48659	-6.29099	3.15453e-10	5.53619e-8	Bacteria
3	Cluster_120	7.49645	-5.21487	0.842194	-6.19200	5.94038e-10	6.95024e-8	Bacteria
4	Cluster_4	284.010	4.46973	0.730032	6.12265	9.20306e-10	8.07569e-8	Bacteria
5	Cluster_85	5.25312	14.8546	2.69005	5.52204	3.35084e-8	0.00000235229	Bacteria
6	Cluster_174	2.99262	17.3671	3.27384	5.30481	1.12788e-7	0.00000659812	Bacteria
7	Cluster_44	22.0406	6.03398	1.14995	5.24715	1.54472e-7	0.00000677746	Bacteria
8	Cluster_141	9.26135	-5.96649	1.13629	-5.25083	1.51415e-7	0.00000677746	Bacteria



# Differential abundance visualization

→ Which species have the highest negative log2Foldchange (more present in VeauHaché than BoeufHaché) ?

	OTU	baseMean	log2FoldChange
	<input type="text" value="/"/>	<input type="text" value="All"/>	<input type="text" value="All"/>
9	Cluster_9	150.302	-28.4432

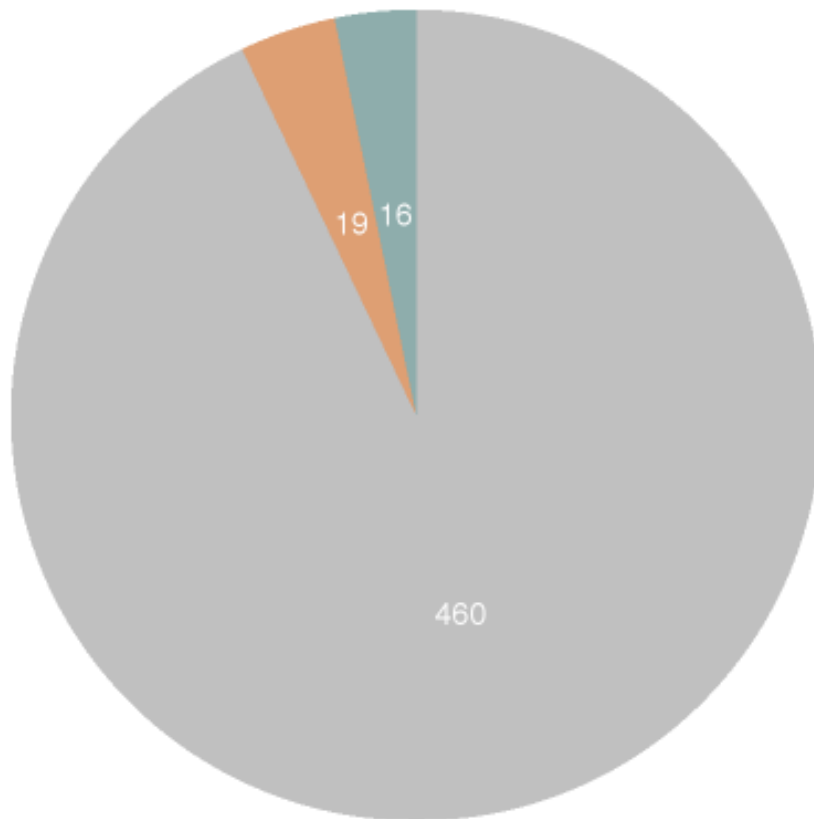
It's the Cluster\_9 which is a *Weissella ceti*

Phylum	Class	Order	Family	Genus	Species
<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="All"/>
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Weissella	Weissella ceti

# Differential abundance visualization

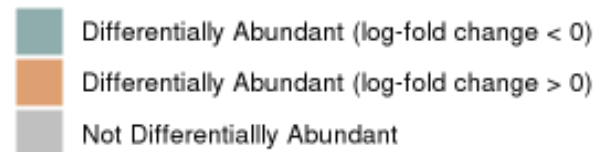
Pie chart to view OTUs number of Differential Abundance test

Pie chart

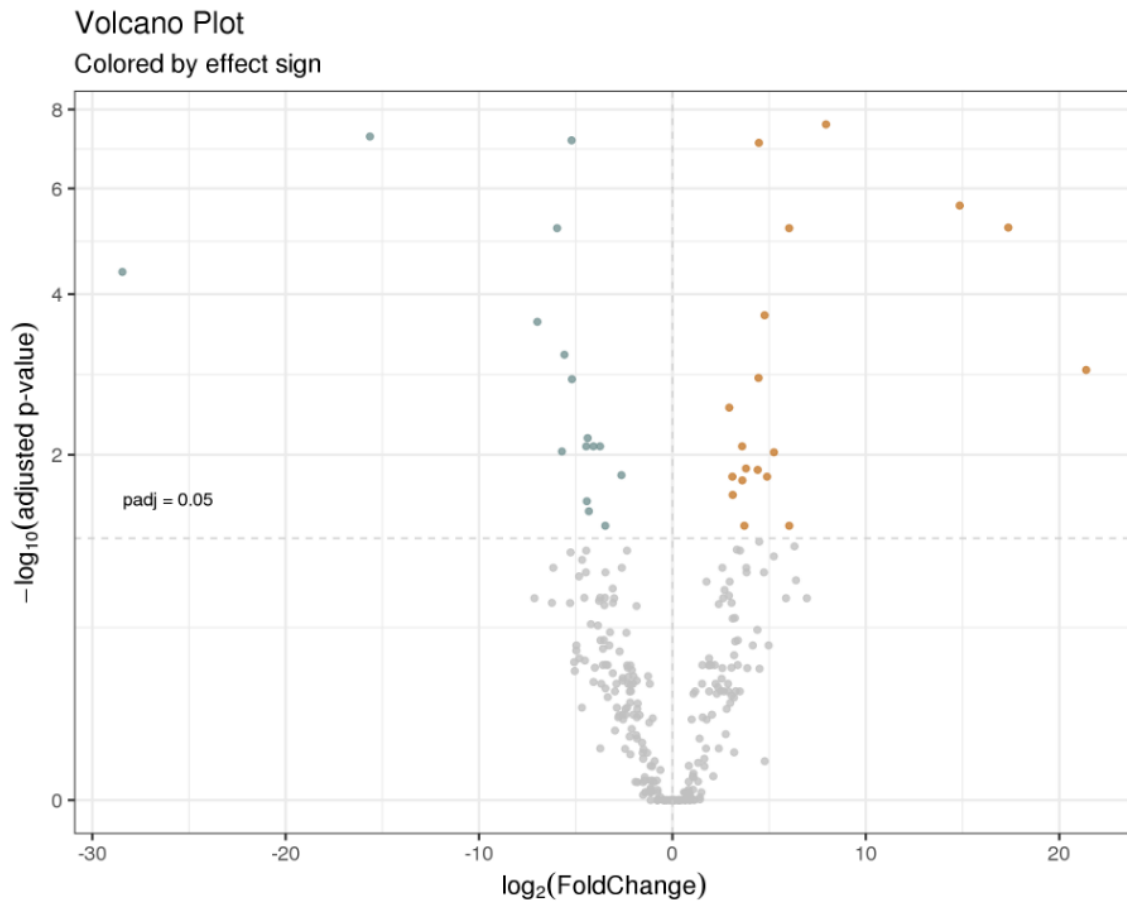


Most of the OTUs are not significantly affected between the conditions

35 OTUs are significantly affected between conditions



# Differential abundance visualization



Volcano plot

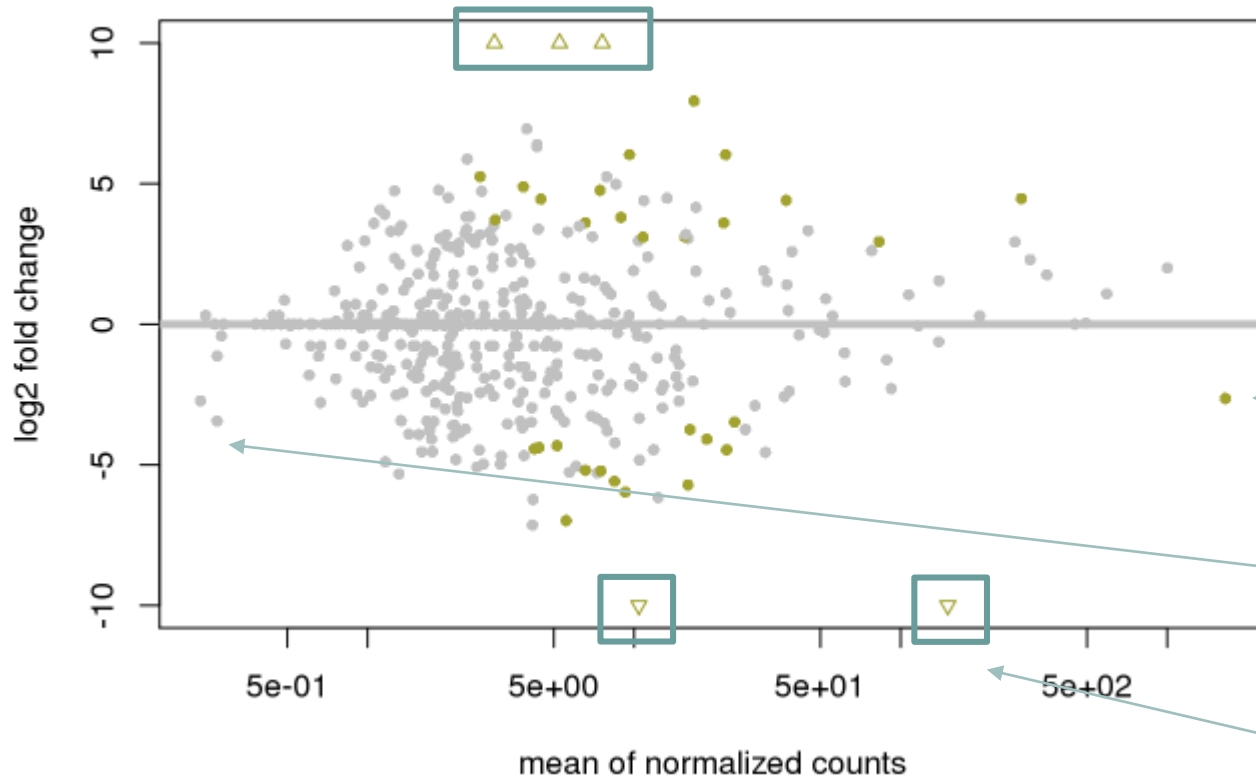
visualization of OTUs  $\log_2\text{FoldChange}$  and their associated adjusted p-values

Only OTUs with a significant adjusted p-value are colored

# Differential abundance visualization

MA plot

Post Normalisation DESeq2: MA plot of log2FoldChange



visualization of the relation between log2foldchange between conditions, and mean abundance of OTUs (significantly affected OTUs are colored)

Colored OTUs on the right : abundant OTUs affected by the conditions

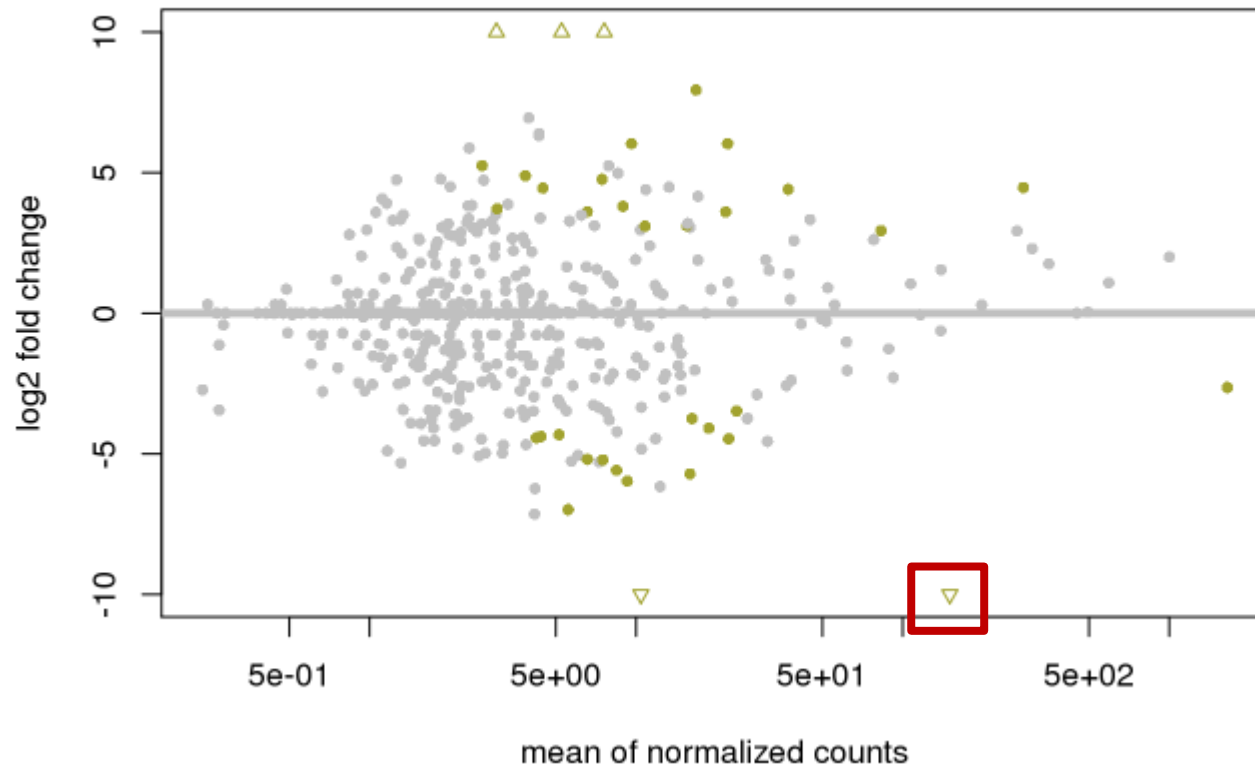
Colored OTUs on the left : affected rare OTUs

Triangles represent OTU out of scale

# Differential abundance visualization

MA plot

Post Normalisation DESeq2: MA plot of log2FoldChange

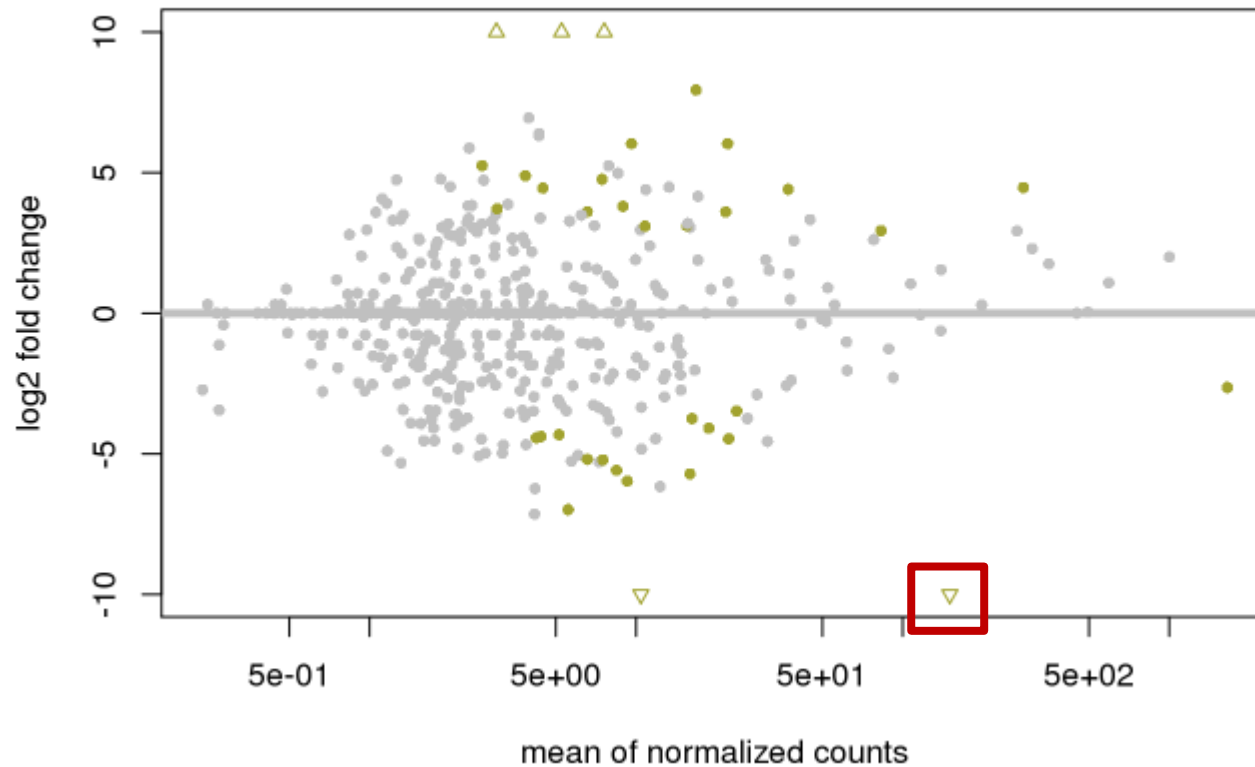


→ Which Cluster is the triangle spotted?

# Differential abundance visualization

MA plot

Post Normalisation DESeq2: MA plot of log2FoldChange



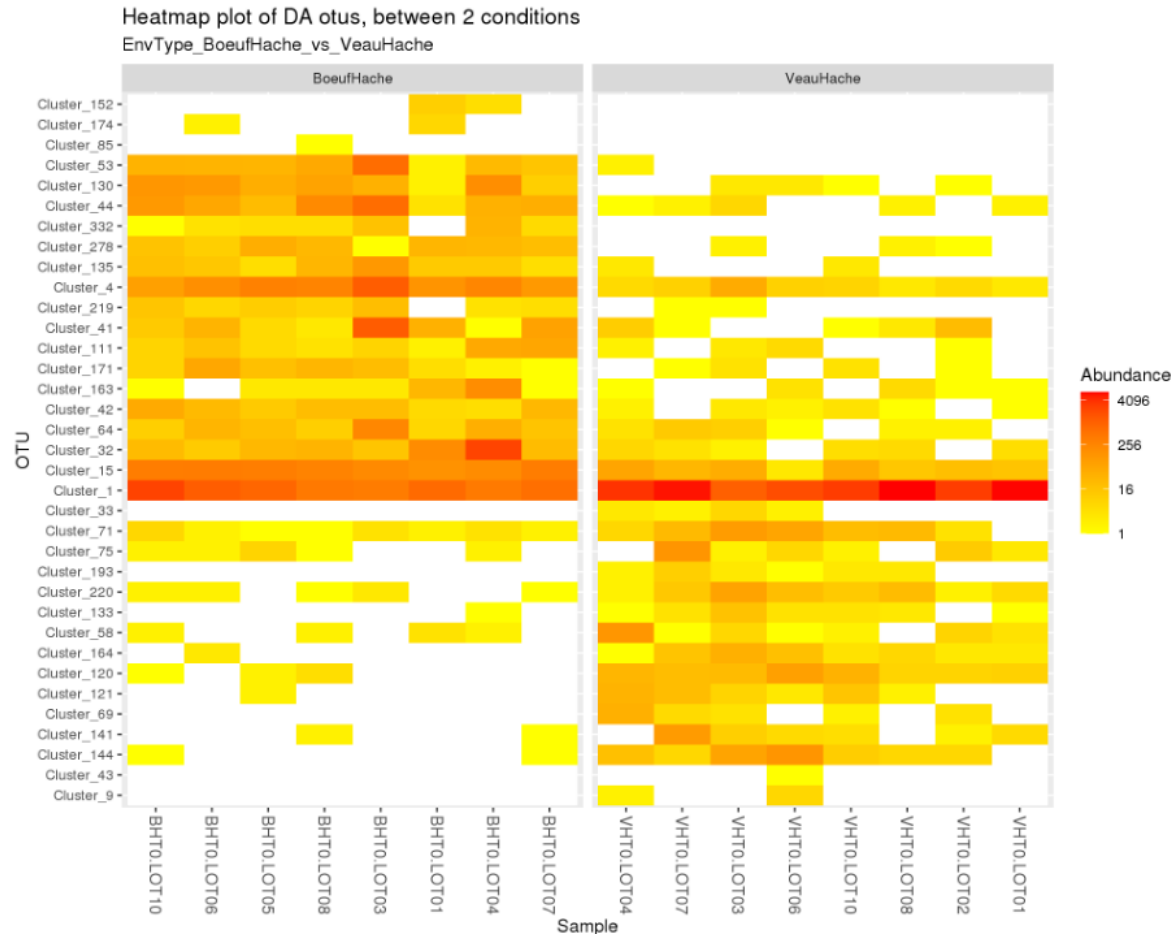
➔ Which Cluster is the triangle spotted?

It's Cluster\_9 !

Medium abundance

	OTU	baseMean	log2FoldChange
	/	All	All
9	Cluster_9	150.302	-28.4432
2	Cluster_43	10.4196	-15.6431

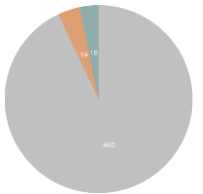
# Differential abundance visualization



Heatmap plot

visualization of the DESeq2 normalised abundances of differentially abundant OTUs grouped by condition

Here, we observe only the significant 35 OTU that are differential abundant



OTUs are ordered from top to bottom in descending order

# Differential abundance visualization

**FROGSTAT Deseq2 Visualization** to extract and visualize differentially abundant OTUs (Galaxy Version 3.2.1) Options

**Phyloseq object (format rdata)**  
17: Phyloseq.Rdata  
This is the result of FROGS Phyloseq Import Data, used in FROGSSTAT Deseq2 Preprocess tool

**DESeq2 object (format rdata)**  
35: FROGSSTAT Deseq2 Preprocess: dds.Rdata  
This is the result of FROGSSTAT Deseq2 Preprocess tool.

**Experimental variable**  
EnvType  
The factor suspected to have an effect on OTUs' abundances (one of the variables used in FROGS Deseq2 Preprocess tool). Ex : Treatment

**Is your Variable quantitative or qualitative?**  
qualitative  
If qualitative, choose 2 conditions to compare.

**Condition 1 considered as reference**  
FiletSaumon  
One condition of the experimental variable (e.g. with).

**Condition 2 to be compared to the reference**  
SaumonFume  
Another condition of the experimental variable (e.g. without).

**Adjusted p-value threshold**  
0.05  
Threshold used for statistical significance of the differentially abundant OTUs analysis.

Execute

Compare FiletSaumon vs SaumonFume



# Differential abundance visualization

---

Differentially abundant OTU table

Pie chart

Volcano plot

MA plot

Heatmap plot

Since we only have a binary factor we can use the following syntax to format the log2 fold change from the fitted model if not, we will use the other syntax with `contrast=c()`

Code

```
You chose to compare FiletSaumon to the reference modality SaumonFume. This implies that a positive log2FoldChange means more abundant in FiletSaumon than in SaumonFume.
```

Then we extract significant OTUs at the p-value adjusted threshold level (after correction) and enrich results with taxonomic informations and sort taxa by pvalue.

# Differential abundance visualization

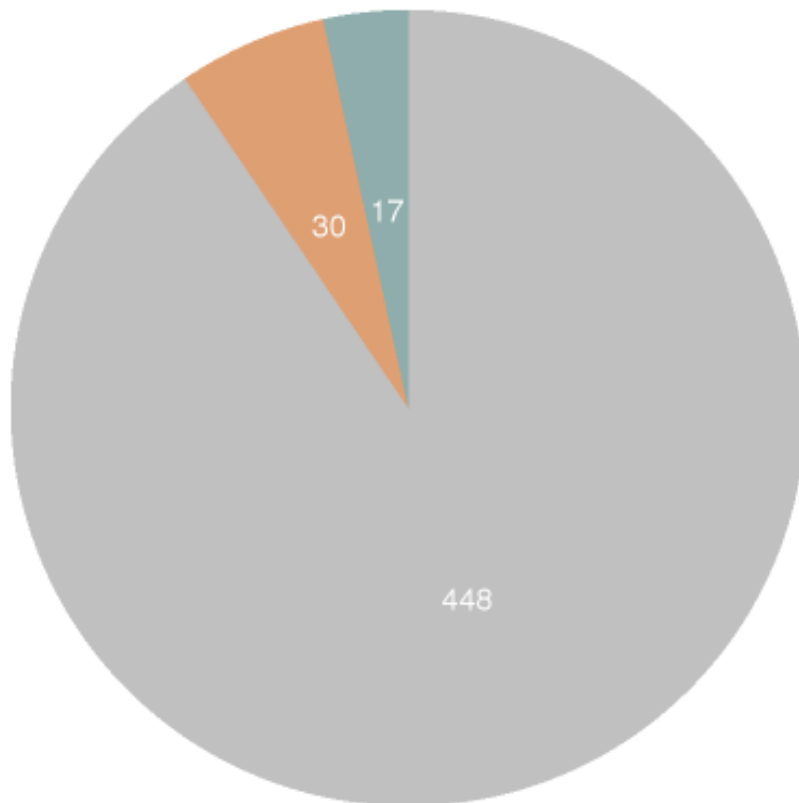
	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	<input type="text" value="/"/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value=""/>	<input type="text" value=""/>	<input type="text" value="All"/>	<input type="text" value="All"/>	<input type="text" value="/"/>
1	Cluster_4	284.010	4.97034	0.718373	6.91888	4.55217e-12	2.25333e-9	Bacteria
2	Cluster_85	5.25312	17.5013	2.66091	6.57718	4.79461e-11	1.18667e-8	Bacteria
3	Cluster_55	19.0634	4.83859	0.825830	5.85906	4.65500e-9	7.68075e-7	Bacteria
4	Cluster_123	10.3886	-7.90236	1.39576	-5.66171	1.49873e-8	0.00000185468	Bacteria
5	Cluster_31	37.4358	5.51672	1.04587	5.27478	1.32917e-7	0.0000131588	Bacteria
6	Cluster_13	139.041	-4.03643	0.838190	-4.81565	0.00000146724	0.000121047	Bacteria

Diferentially abundant OTU table

# Differential abundance visualization

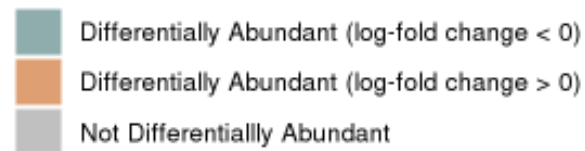
Pie chart

Pie chart to view OTUs number of Differential Abundance test



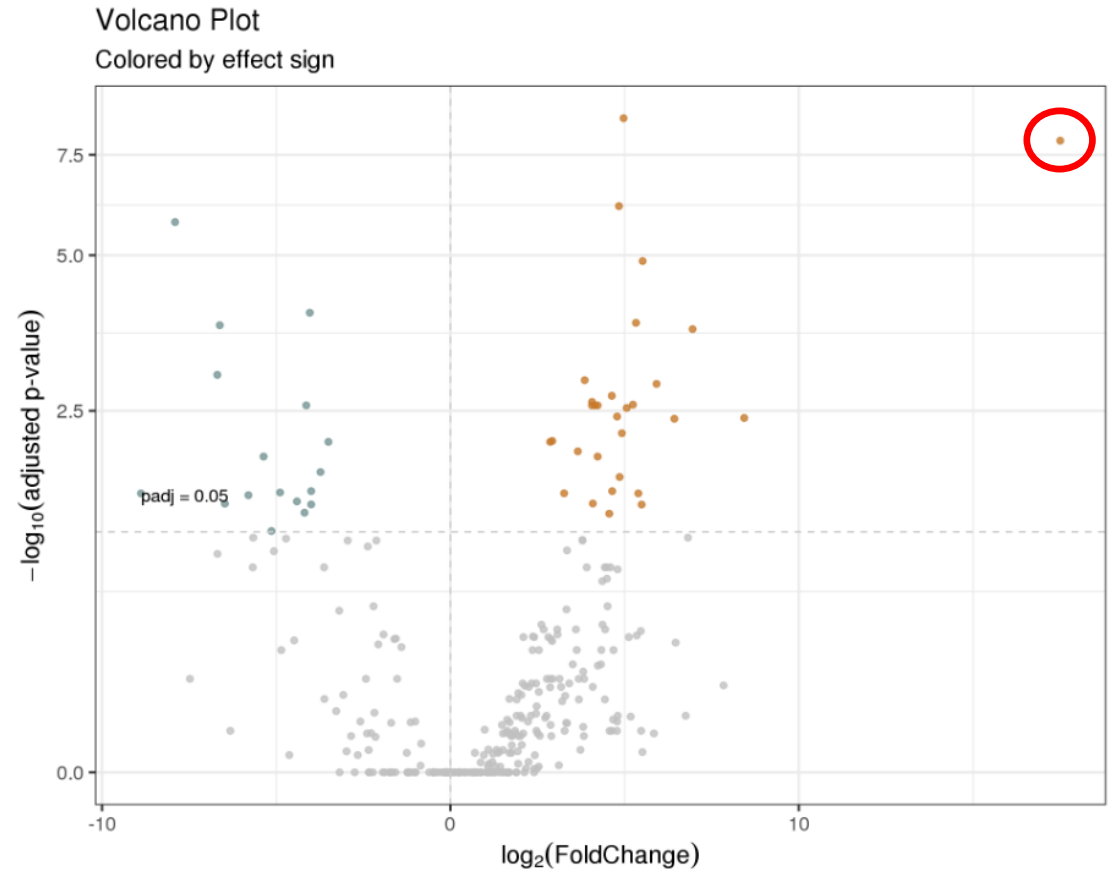
Most of the OTU are not significantly affected between your conditions

Only 47 OTUs are significantly affected between conditions



# Differential abundance visualization

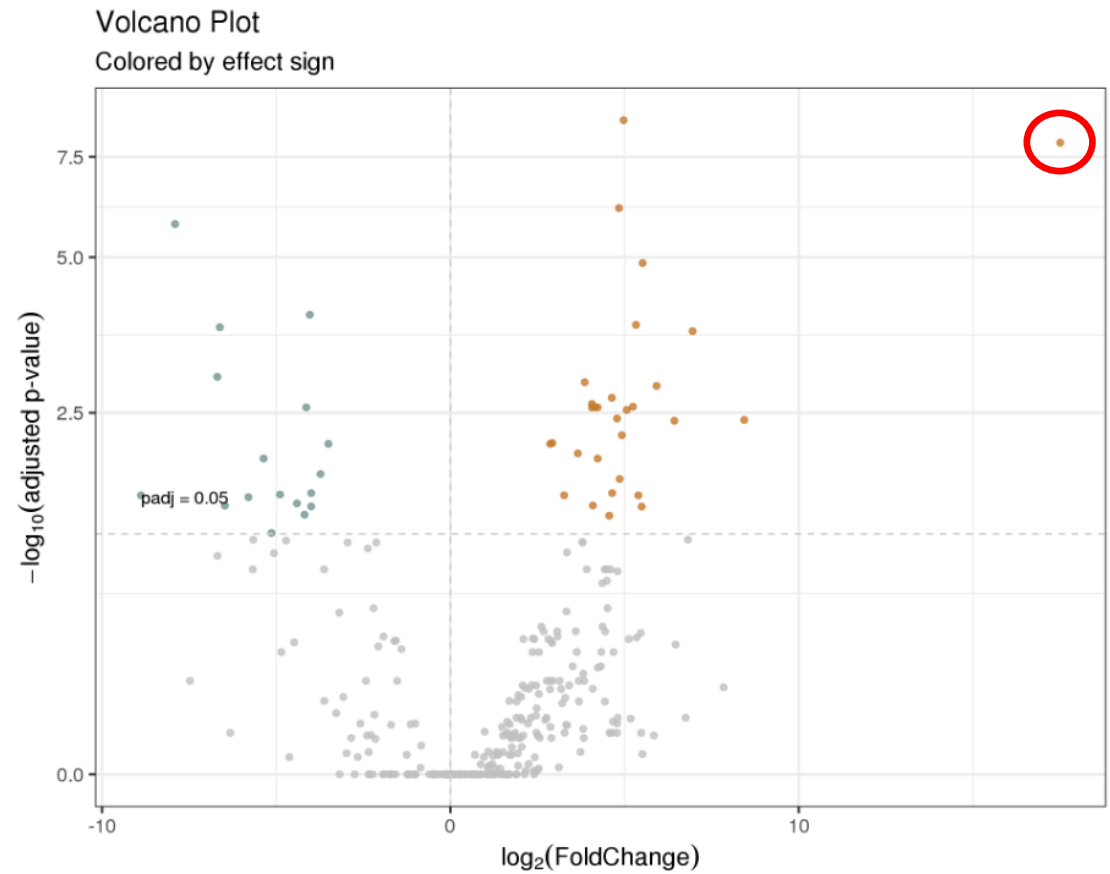
Volcano plot



→ Which Cluster is it ?

# Differential abundance visualization

Volcano plot



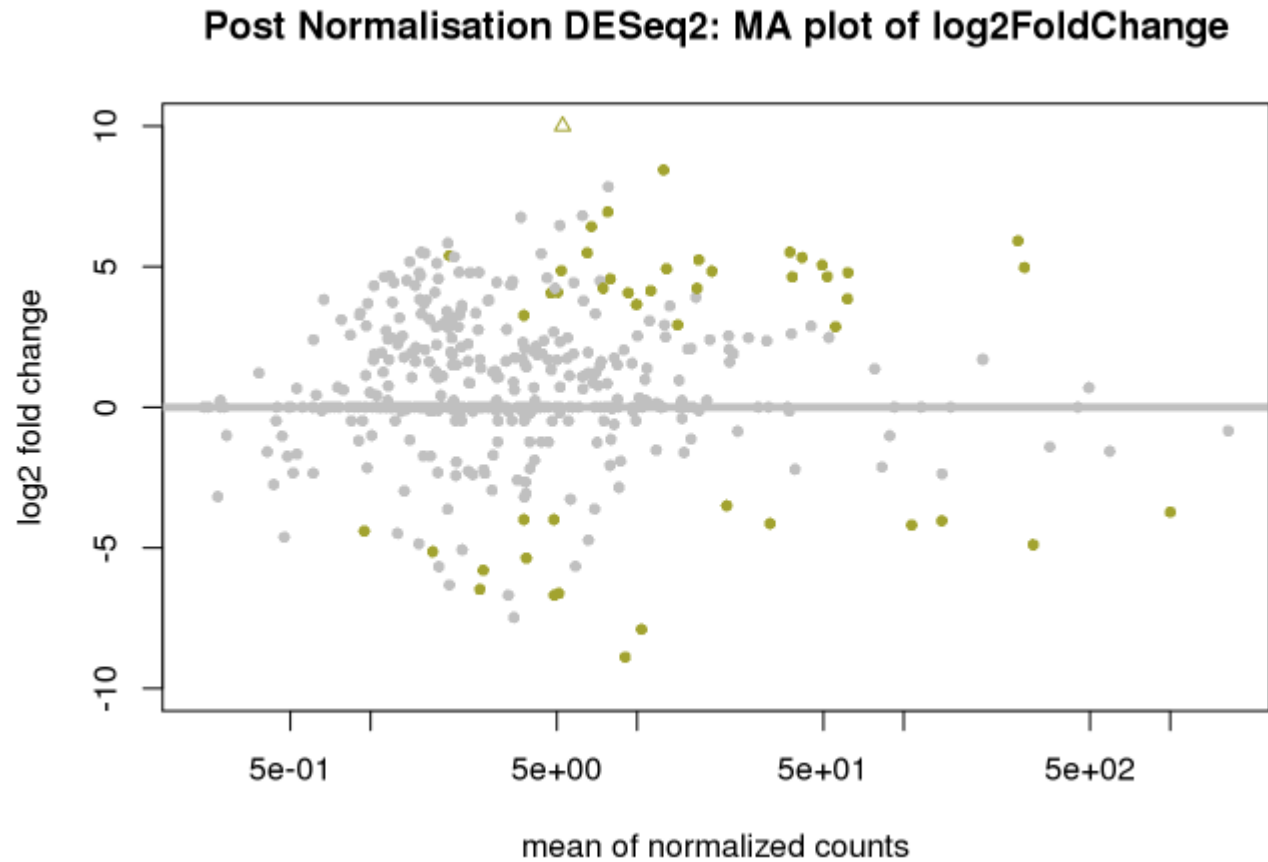
→ Which Cluster is it ?

Cluster\_85: *Flavobacterium omnivorum*

	OTU	baseMean	log2FoldChange
	/	AI	All
2	Cluster_85	5.25312	17.5013
22	Cluster_76	12.5611	8.43272
9	Cluster_73	7.76604	6.95033

# Differential abundance visualization

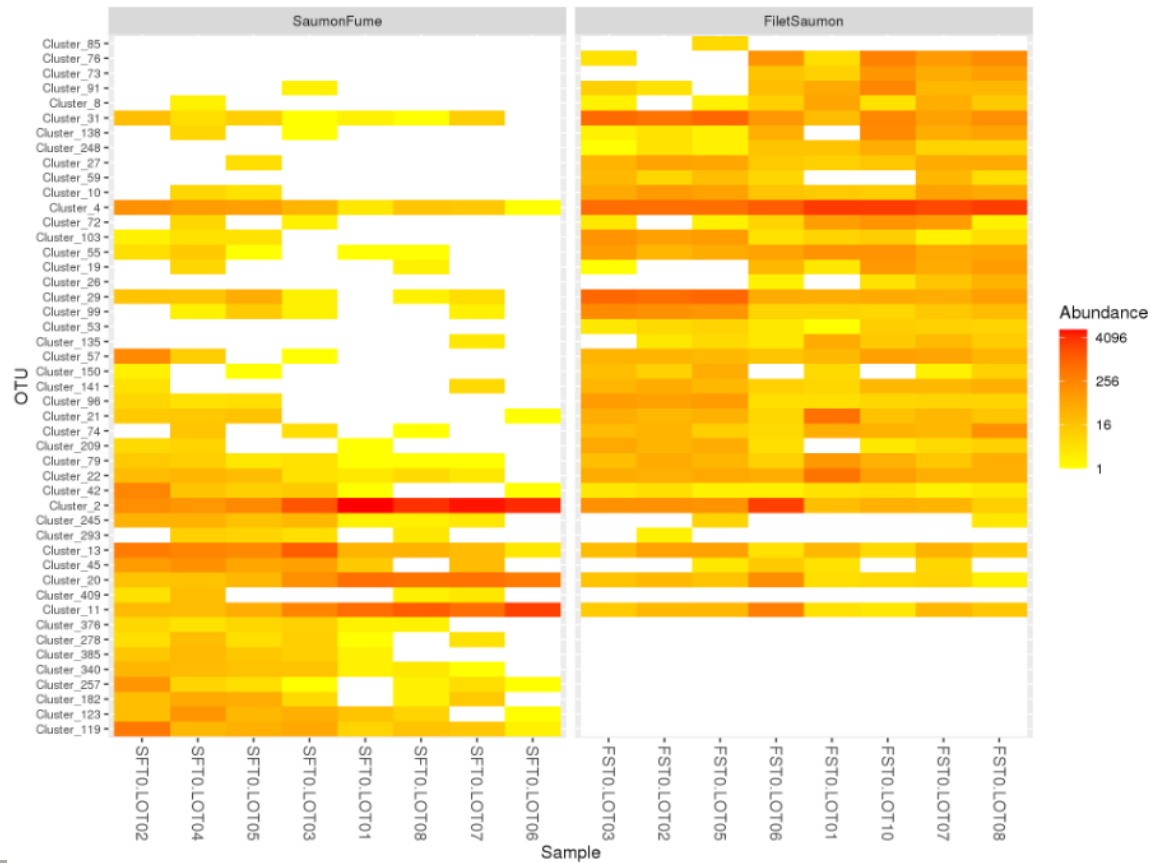
MA plot



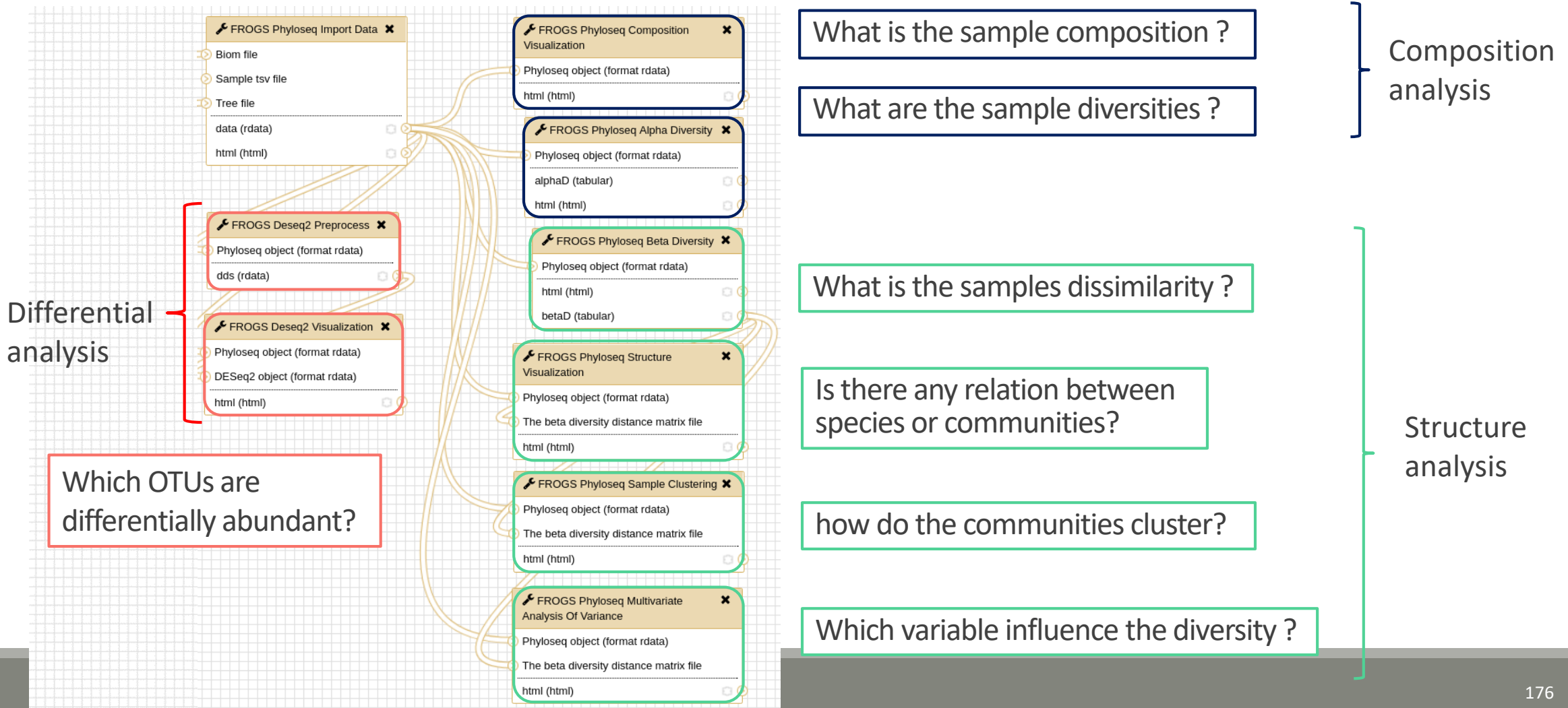
# Differential abundance visualization

Heatmap plot

Heatmap plot of DA otus, between 2 conditions  
EnvType\_FiletSaumon\_vs\_SaumonFume

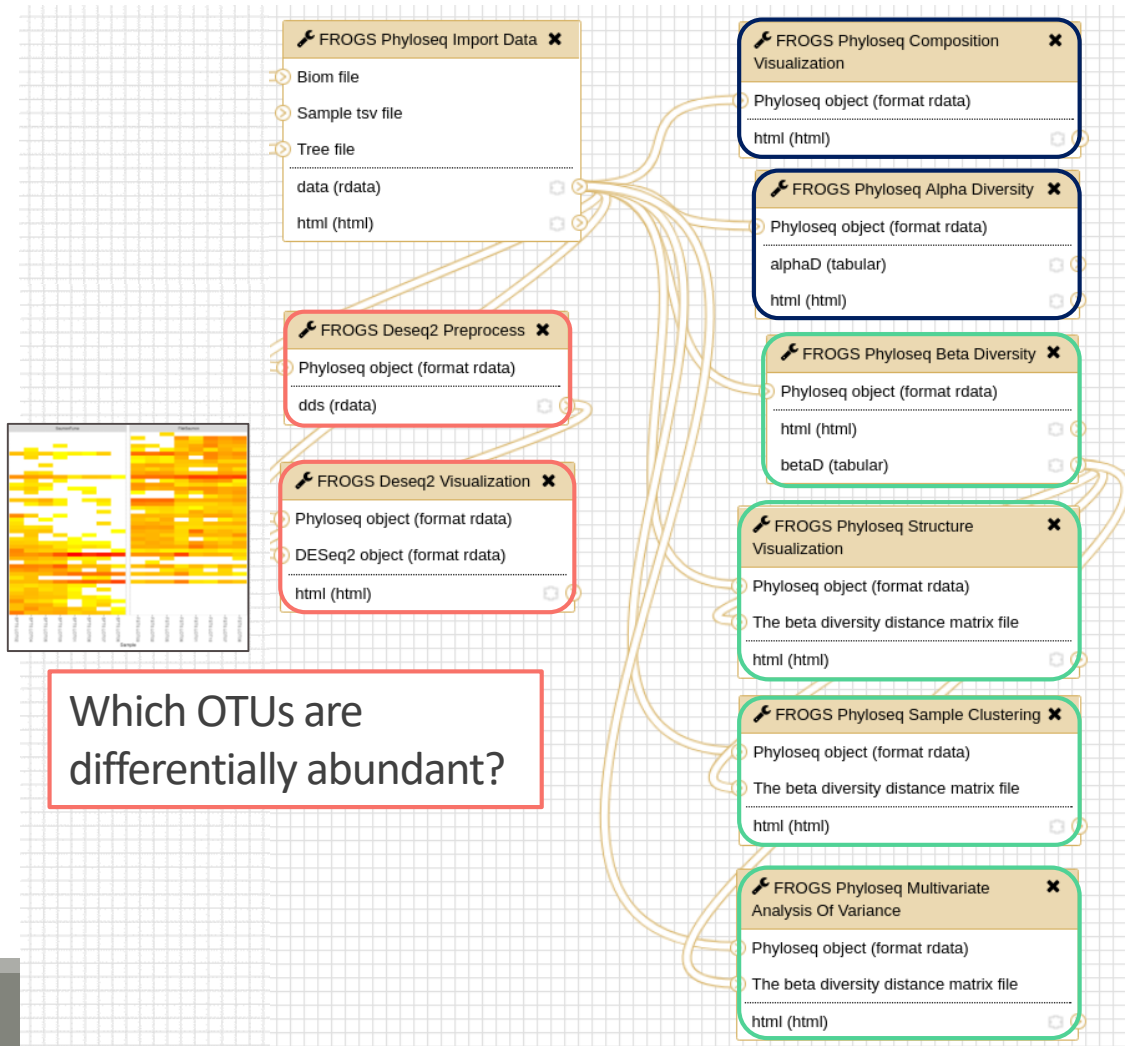


# FROGSStat Summary

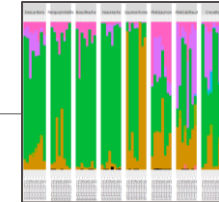




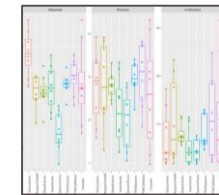
# FROGSStat Summary



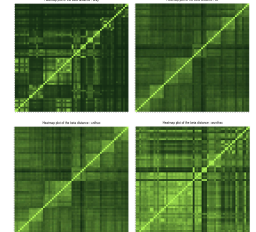
What is the sample composition ?



What are the sample diversities ?



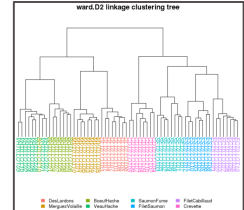
What is the samples dissimilarity ?



Is there any relation between species or communities?



how do the communities cluster?



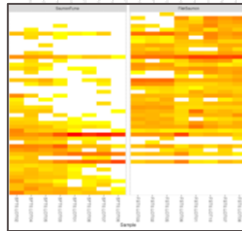
Which variable influence the diversity ?

```

adonis(formula = dist ~ EnvType, data = metadata, permutations = 9999)
Permutation: free
Number of permutations: 9999
Terms added sequentially (first to last)

Df SumOfSqs MeanSqs F.Model    R2 Pr(>F)
EnvType  7  6.1849 0.88356 11.164 0.58255 1e-04 ***
Residuals 56  4.4328 0.77374      0.42745
Total    63 10.6178      1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    
```

Which OTUs are differentially abundant?



---

# Conclusion and advices reminder

---

# FROGSTAT advices

---

- Before starting, **check taxonomy format** : how many levels? What are their names ?
- Carefully construct your **sample\_metadata** TSV file, and after its import, check that your variable order is smart
- Keep in mind that :
  - **Phyloseq composition** and **structure analyses** need to be performed on **normalised** (=rarefied) counts
  - **DESeq** analysis needs to be performed on counts **without normalisation**
  - Different indices or distance methods will give **different but complementary** information
  - **Test different distances and choose which one fits better your data**

# References

---

- Chaillou, S., Chaulot-Talmon, A., Caekebeke, H., Cardinal, M., Christieans, S., Denis, C., Desmonts, M. H., Dousset, X., Feurer, C., Hamon, E., Joraud, J.-J., La Carbona, S., Leroi, F., Leroy, S., Lorre, S., Mace, S., Pilet, M.-F., Prevost, H., Rivollier, M., Roux, D., Talon, R., Zagorec, M., and Champomier-Verges, M.-C. (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. *ISME J*, 9(5):1105{1118.
- McMurdie, P. J. and Holmes, S. (2013). phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, 8(4):e61217.
- Shade, A., Jones, S. E., Caporaso, J. G., Handelsman, J., Knight, R., Fierer, N., and Gilbert, J. A. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *MBio*, 5(4):e01371{e01314.