C - Training on Galaxy: Metabarcoding

March 2021 - Webinar

STATISTICS Practice

Lucas Auer, Maria Bernard, Laurent Cauquil, Mahendra Mariadassou, Géraldine Pascal

Goals

- Exploratory Data Analysis
  - α-diversity: how diverse is my community?
  - β-diversity: how different are two communities?
  - Visual assessment of the data
    - Barplots: what is the composition of each community?
    - Multidimensional Scaling: how are communities related?
    - Heatmaps: are there interactions between species and (groups of) communities?
  - Use a distance matrix to study structures:
    - Hierarchical clustering: how do the communities cluster?
    - Permutational ANOVA: are the communities structured by some known environmental factor (pH, height, etc)?
    - Differential abundance analysis: 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

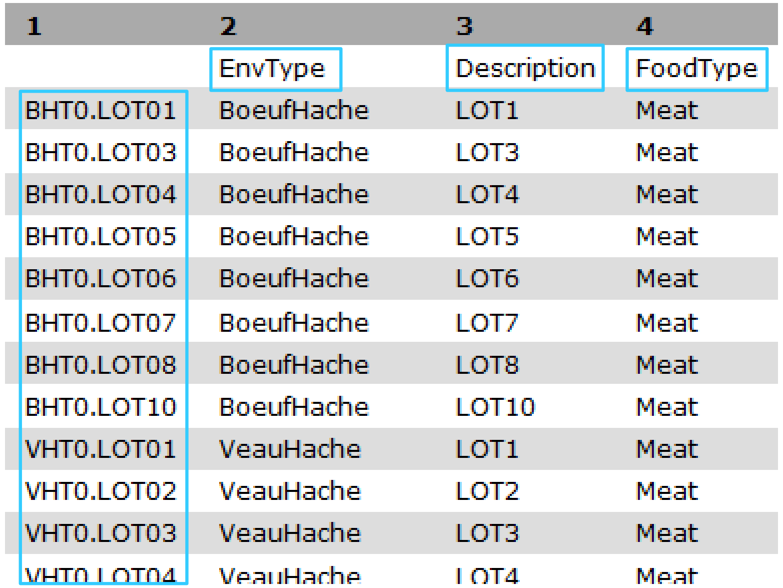
Metadata: sample description in TSV file

Exercise 1

➔ Take a look at the metadata

Exercise 1

➔ Take a look at the metadata



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

FoodType:

Meat or Seafood

EnvType: 8 environment types

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

The FROGS biom format contains:

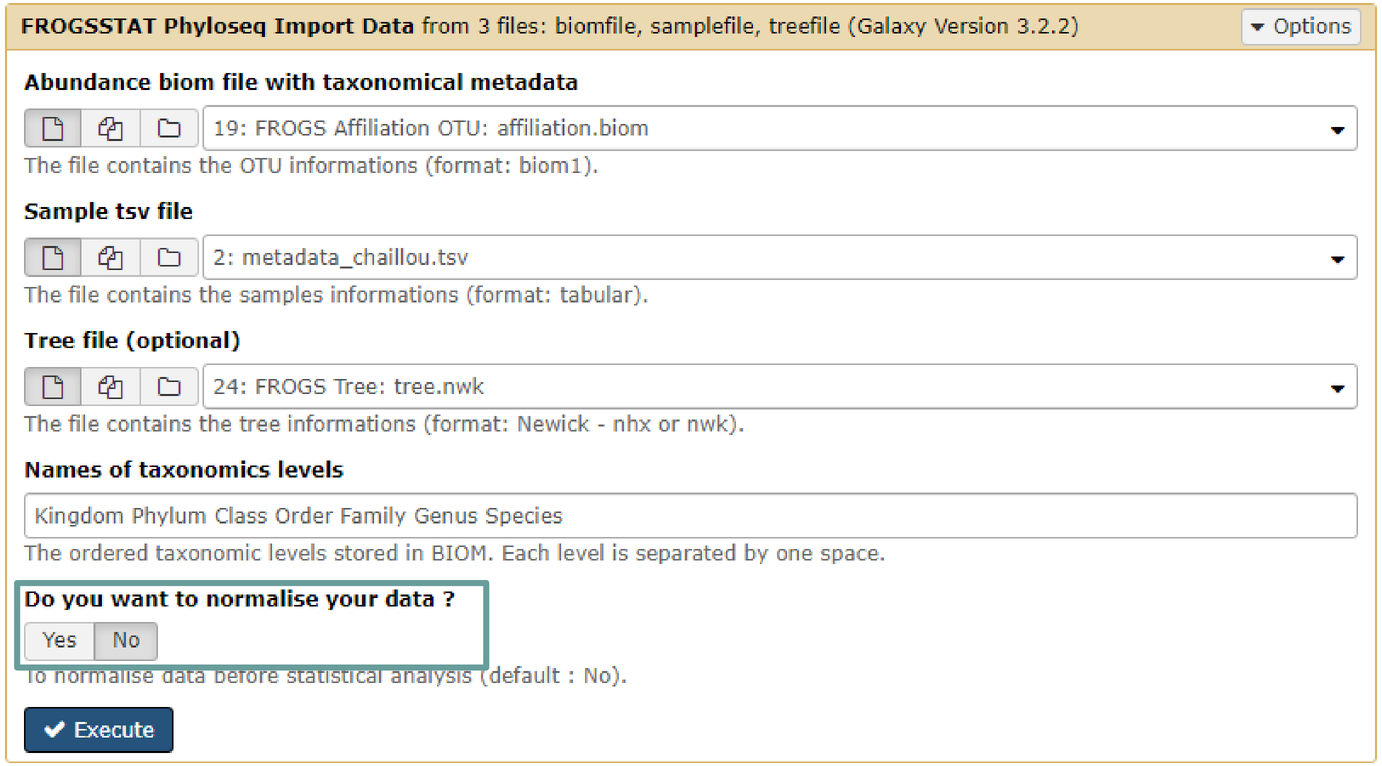
- OTU count tables (required)
- OTU description : taxonomy

Others information used in FROGSSTAT are:

- sample description in TSV file
- phylogenetic tree in Newick format

  (nwk or nhx)

→ Create 2 phyloseq objects, with and without normalisation (rename them)



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

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 normalisation ?

3. Explore the HTML results

Exercise 2

1. What are the resulting datasets ?

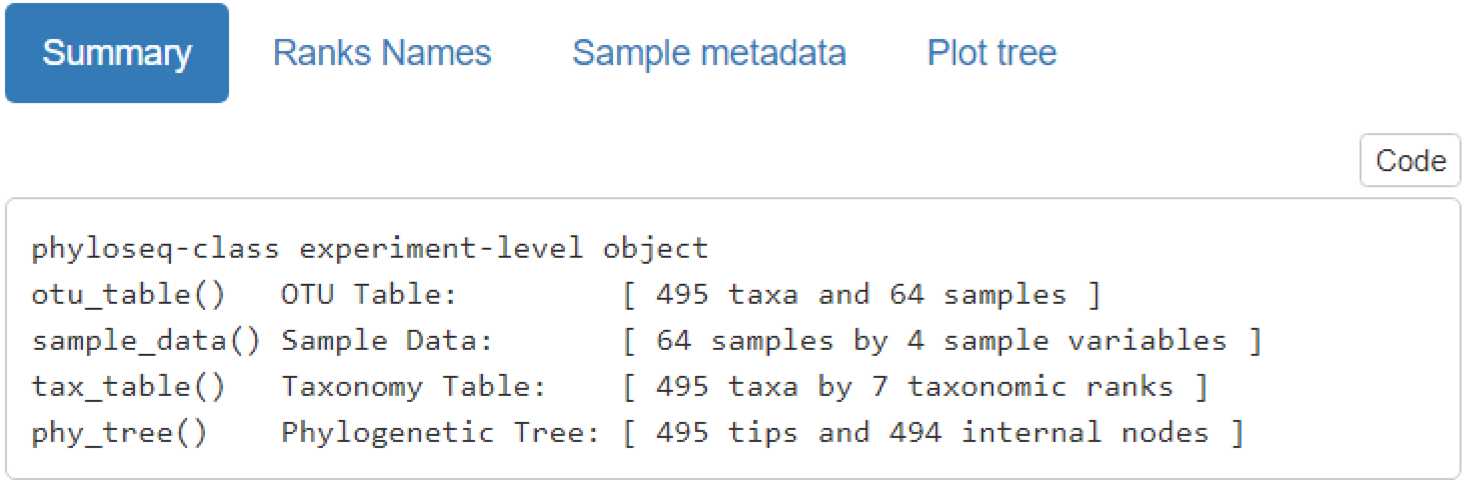
→ Rdata file: R object used by phyloseq package for statistics

→ HTML report: summary of the phyloseq object

Exercise 2

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

Without normalisation

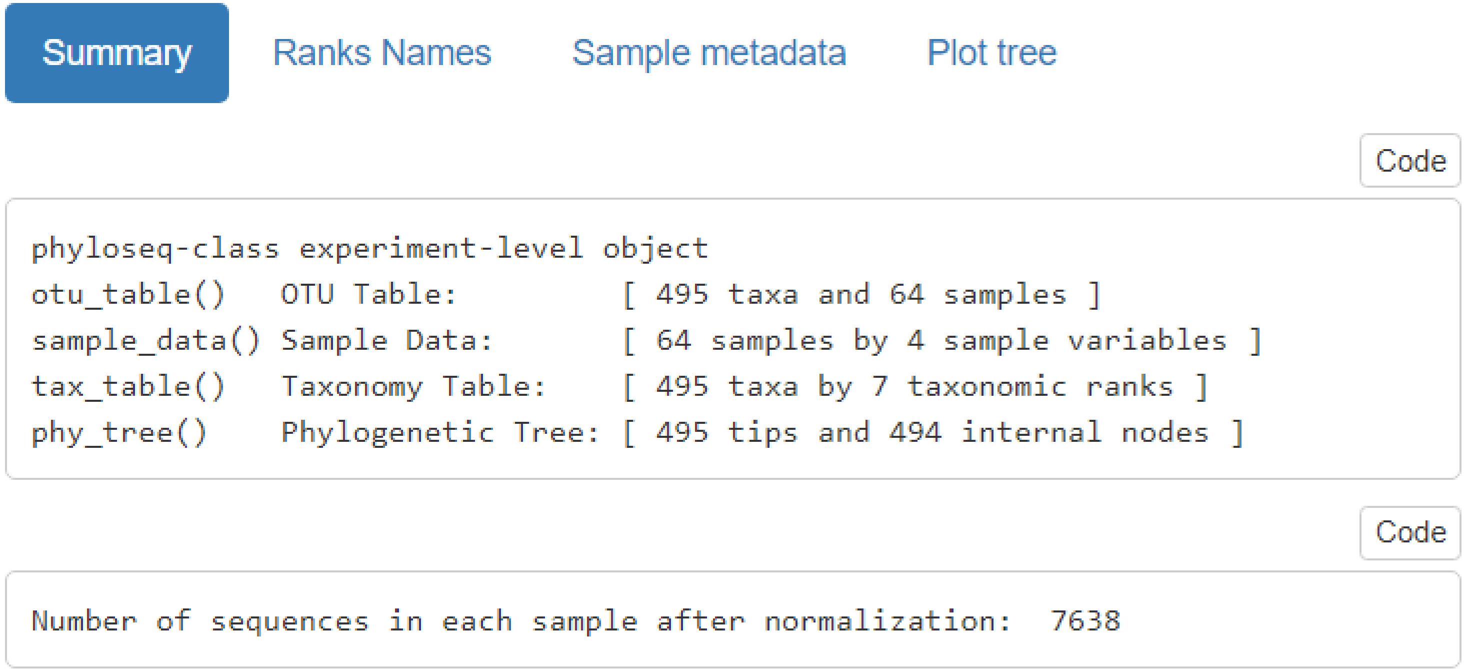


```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 495 taxa and 64 samples ]
sample_data() Sample Data:       [ 64 samples by 4 sample variables ]
tax_table()   Taxonomy Table:    [ 495 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 495 tips and 494 internal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

Exercise 2

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

With normalisation (rarefaction)



```
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 ]
```

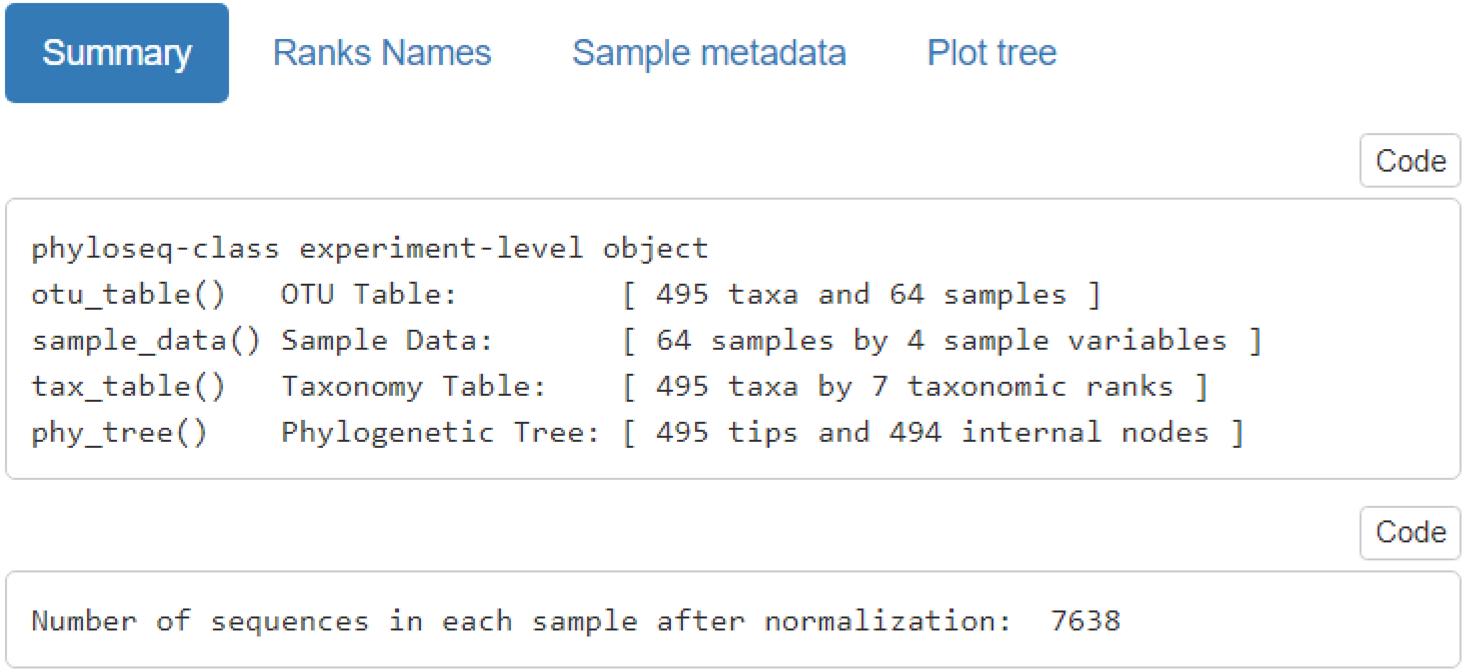
Minimum number of sequences kept in each sample

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

Exercise 2

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

With normalisation (rarefaction)



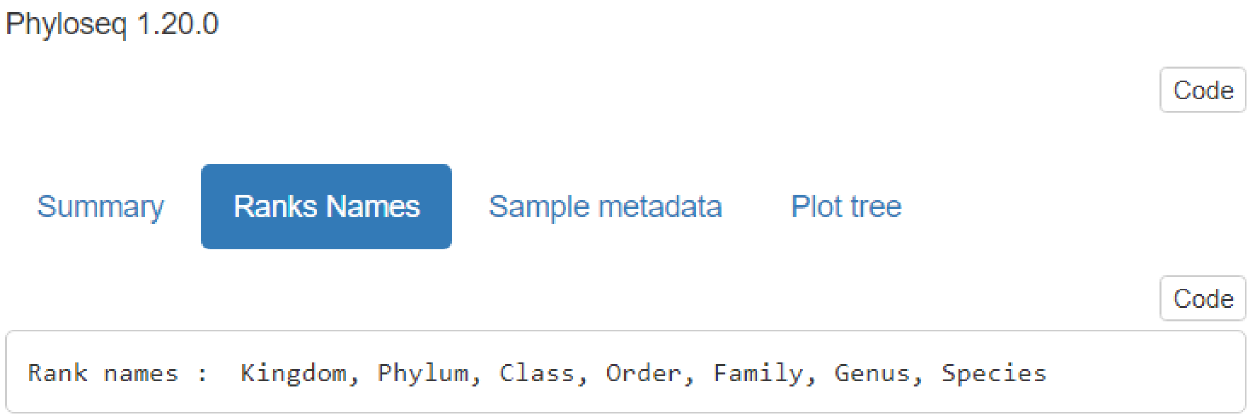
```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 495 taxa and 64 samples ]
sample_data() Sample Data:       [ 64 samples by 4 sample variables ]
tax_table()   Taxonomy Table:    [ 495 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 495 tips and 494 internal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

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

Be aware the number of OTU (taxa) may decrease

Exercise 2

3. Explore the HTML results

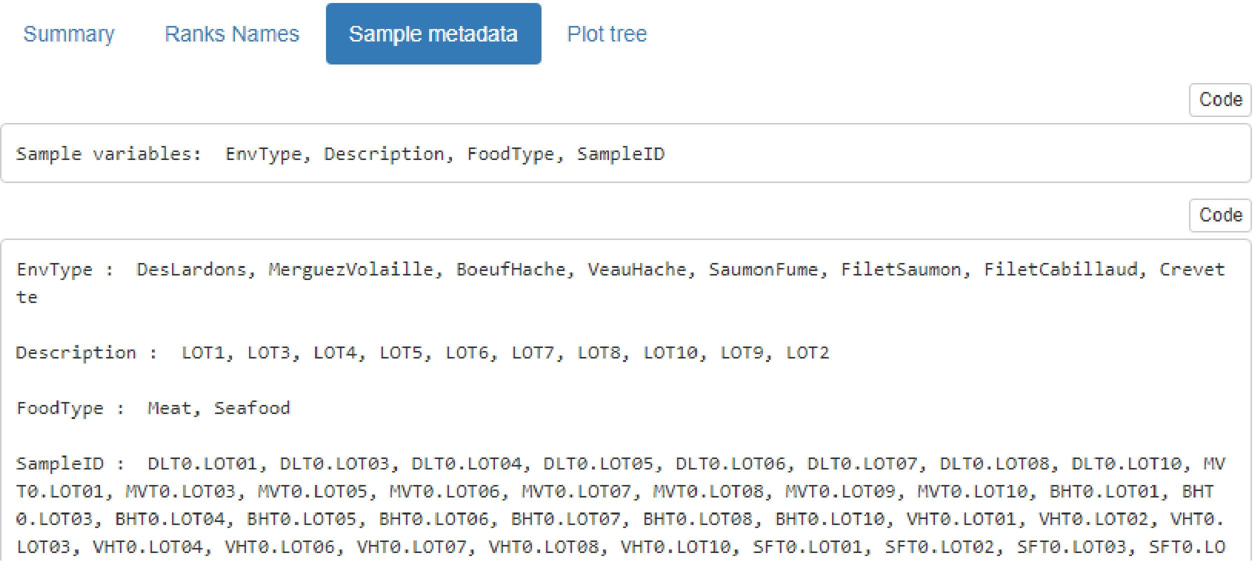


```
Rank names :  Kingdom, Phylum, Class, Order, Family, Genus, Species
```

Taxonomic levels

Exercise 2

3. Explore the HTML results



```
Sample variables:  EnvType, Description, FoodType, SampleID
```

```
EnvType :  DesLardons, MerguezVolaille, BoeufHache, VeauHache, SaumonFume, FiletSaumon, FiletCabillaud, Crevet
te

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

FoodType :  Meat, Seafood

SampleID :  DLT0.LOT01, DLT0.LOT03, DLT0.LOT04, DLT0.LOT05, DLT0.LOT06, DLT0.LOT07, DLT0.LOT08, DLT0.LOT10, MV
T0.LOT01, MVT0.LOT03, MVT0.LOT05, MVT0.LOT06, MVT0.LOT07, MVT0.LOT08, MVT0.LOT09, MVT0.LOT10, BHT0.LOT01, BHT
0.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.LO
```

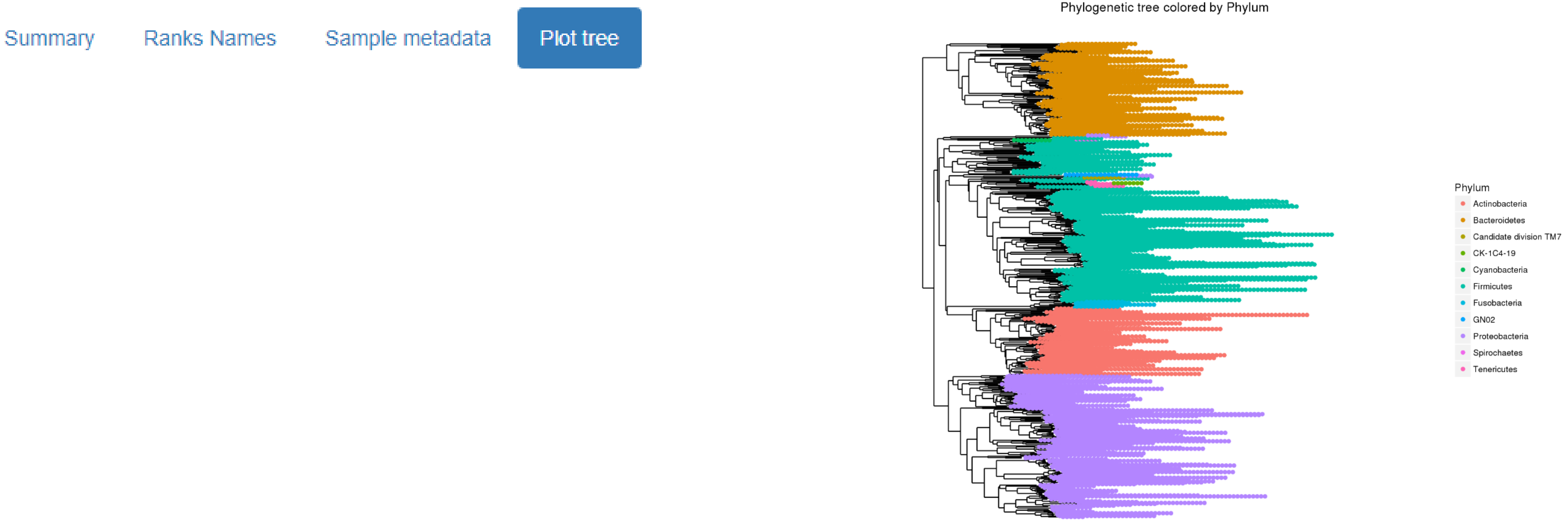
Warning !

Metadata order (in each sample variable) are used to organize graphics.

So take extra care when you construct your sample_metadata file

Exercise 2

3. Explore the HTML results

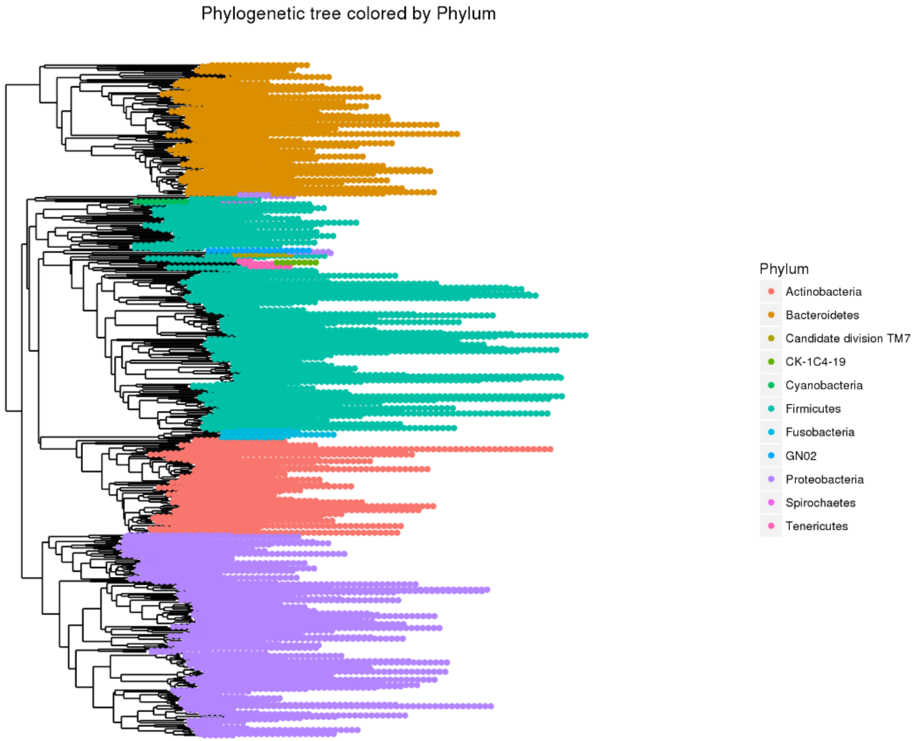


Exercise 2

3. Explore the HTML results

➔Information: Most represented phylum

- Bacteroidota
- Firmicutes
- Actinobacteriota
- Proteobacteria



Biodiversity analysis

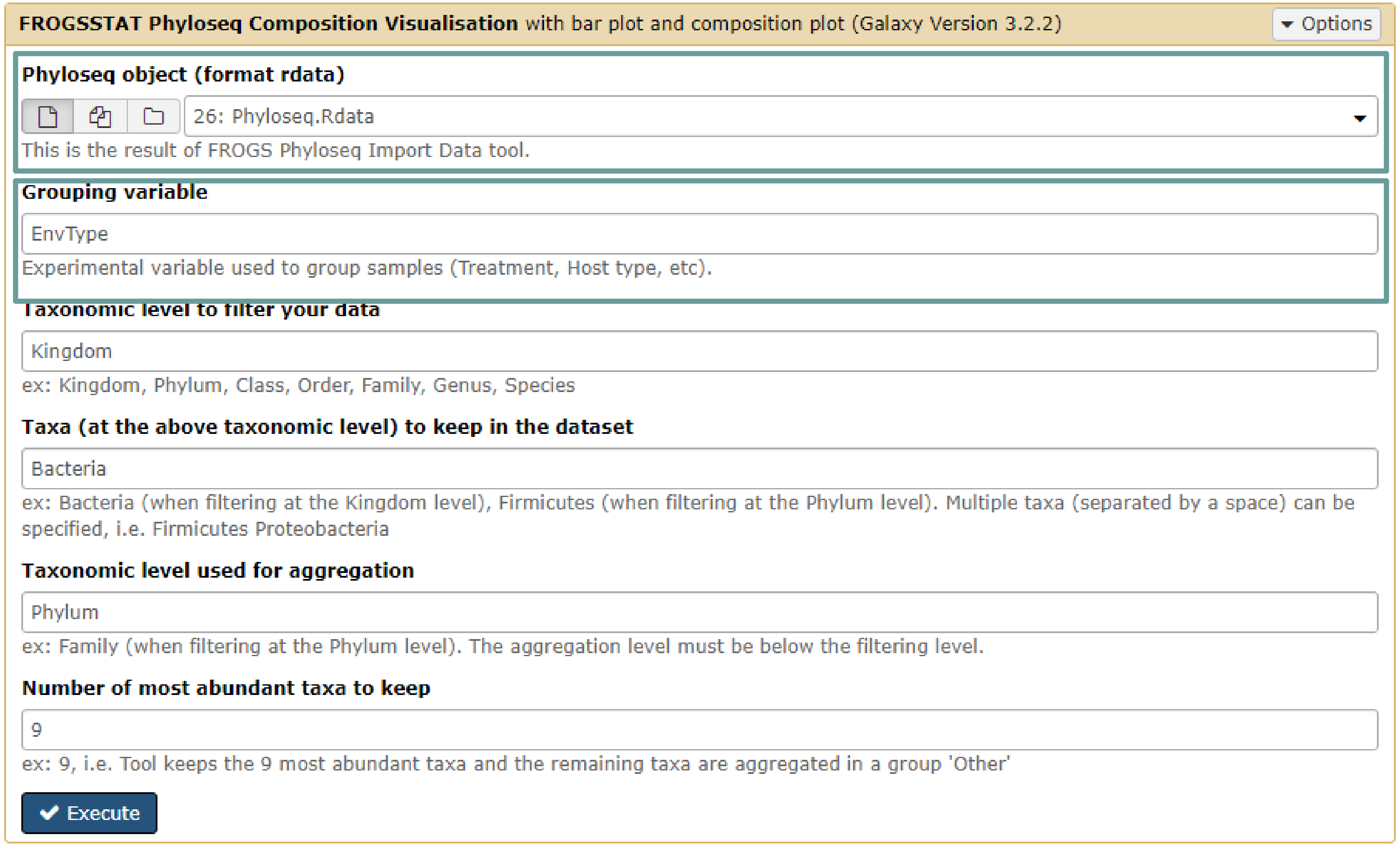
Biodiversity analysis

1. Exploring sample composition
2. Notions of biodiversity
3. α-diversity analysis
4. β-diversity analysis

I. Biodiversity analysis

COMPOSITION VISUALISATION

Exploring biodiversity : visualisation



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

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 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'

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

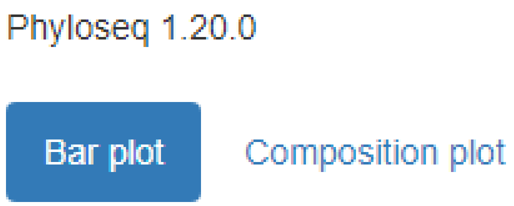
1. What are the resulting datasets ?
2. Difference between Bar plot and Plot composition ?
3. What biological information could you extract? ?
4. Perspectives to go further ?

Exercise 3

1. What are the resulting datasets ?

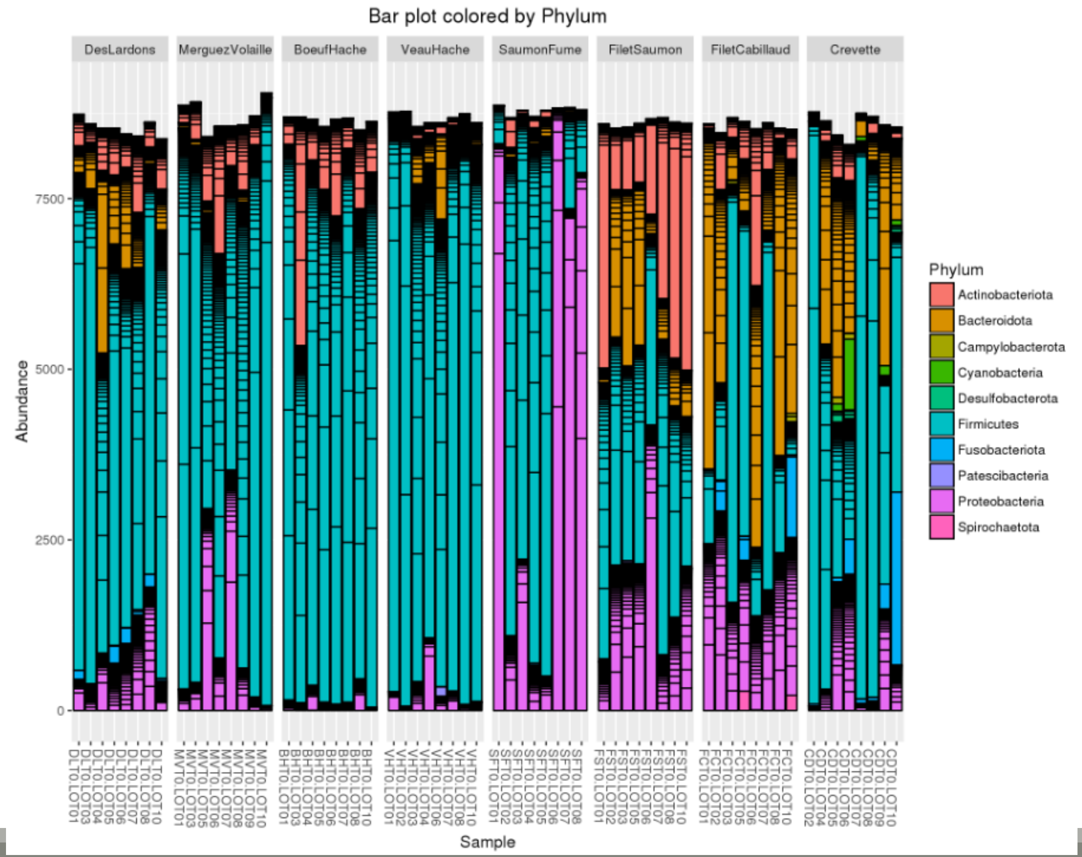
→ HTML report: summary of the phyloseq object

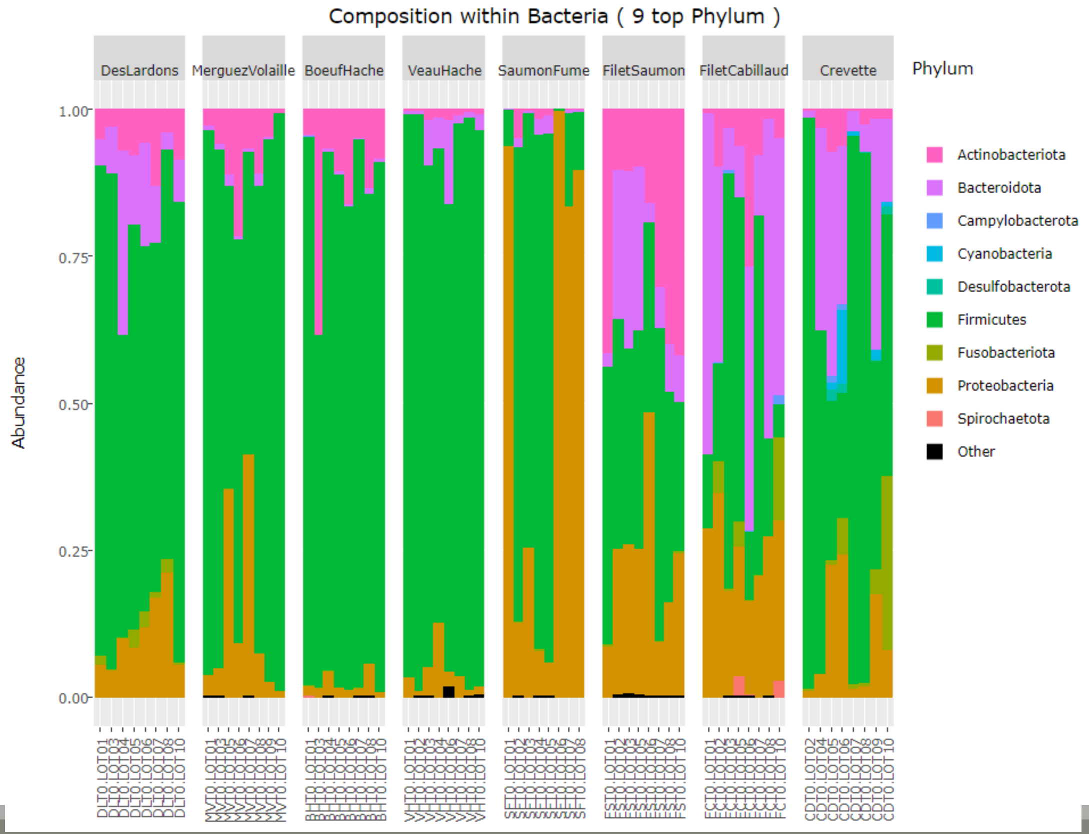
- Bar plot
- Composition plot


Phyloseq 1.20.0

Exercise 3

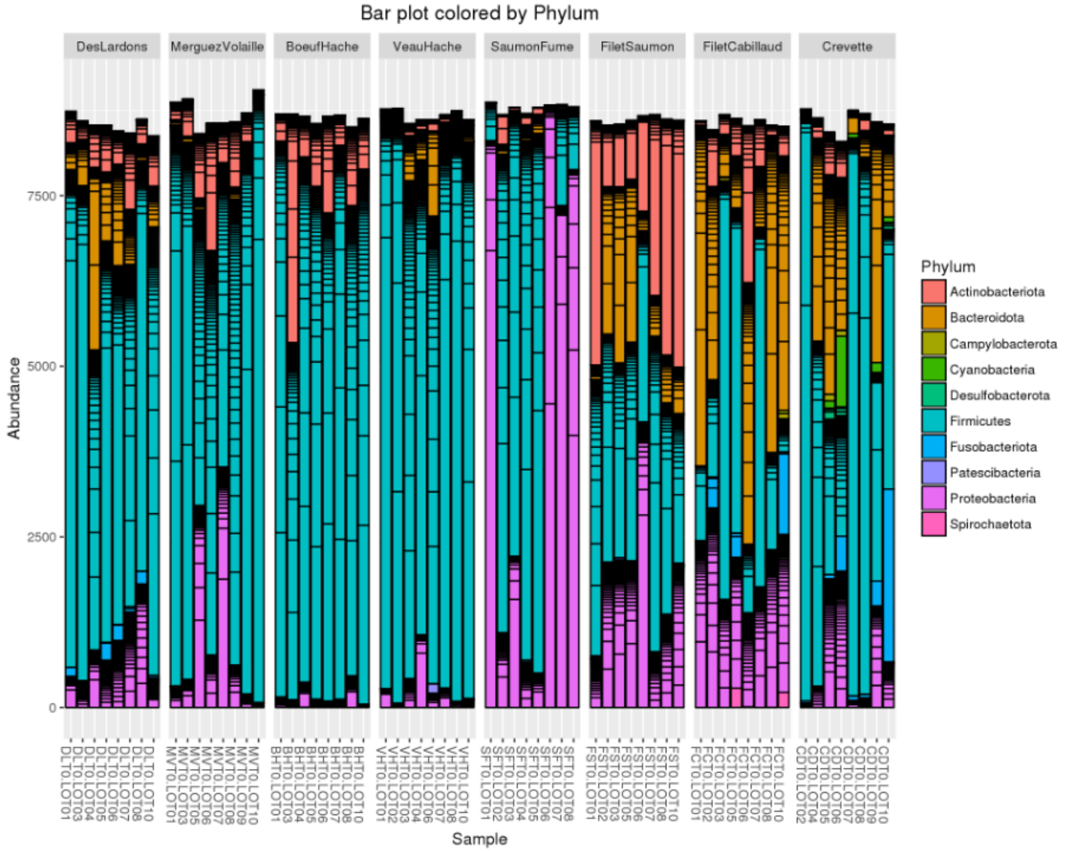
2. Difference between Bar plot and Plot composition ?





Exercise 3

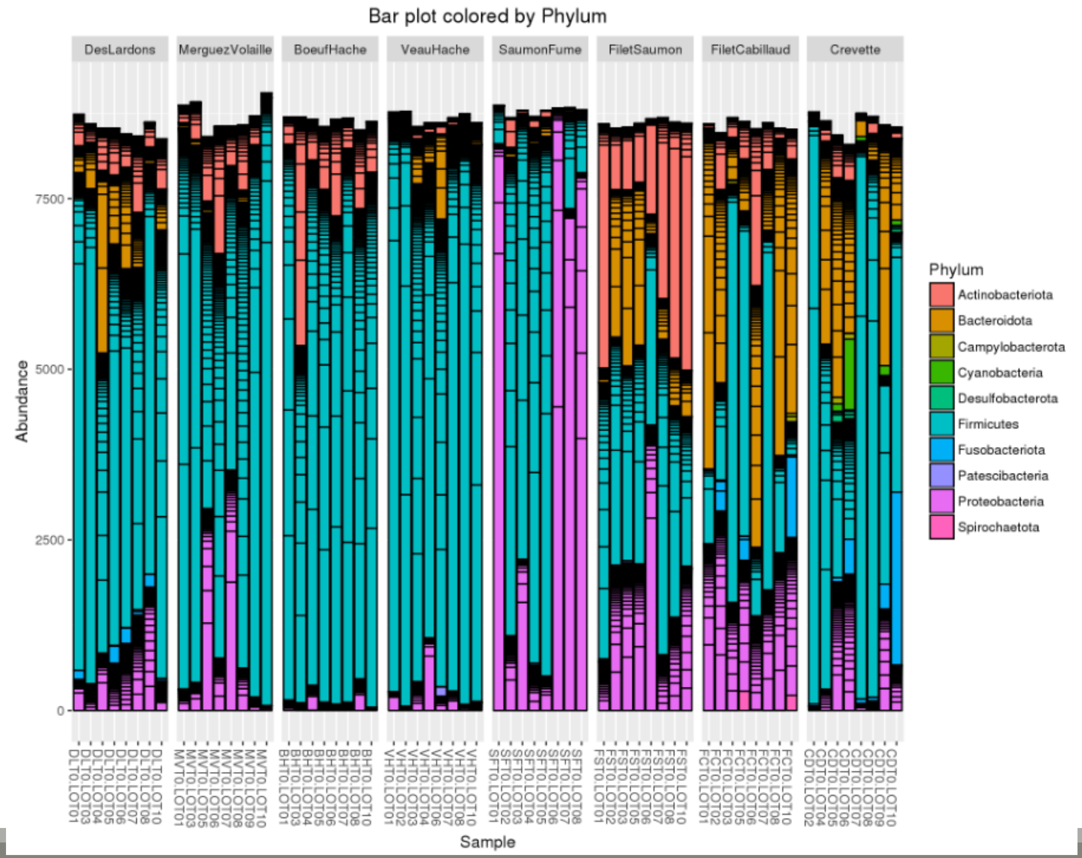
2. Difference between Bar plot and Plot composition ?



- one rectangle is one OTU
- one color is one phylum
- y axis: number of sequences
- size of rectangle depends on number of sequences

Exercise 3

2. Difference between Bar plot and Plot composition ?



Limitations:

- Plot bar works at the OTU-level...
- ...which may lead to graph cluttering and useless legends
- No easy way to look at a subset of the data
- Works with absolute counts (beware of unequal depths or used normalised function)

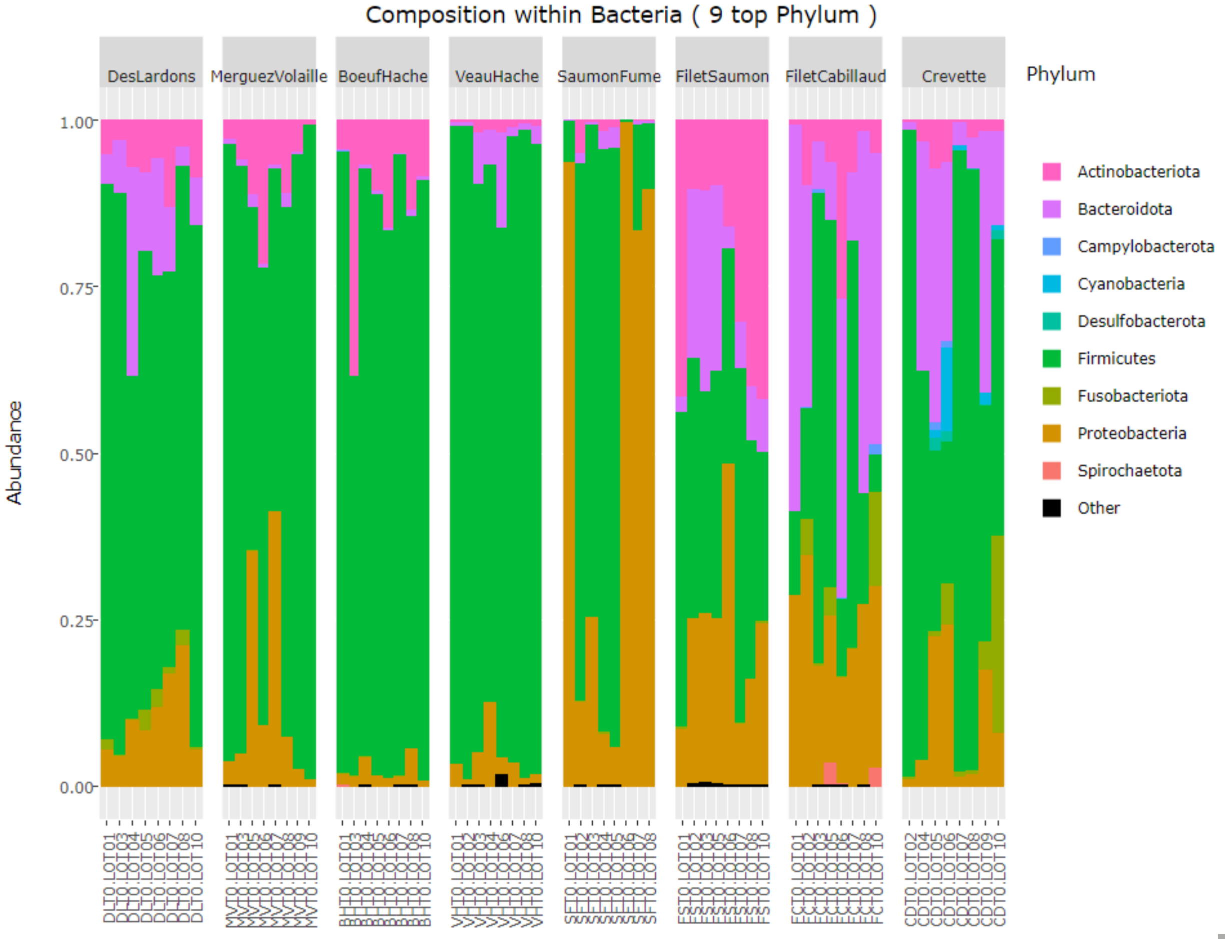
Exercise 3

Bar plot

Composition plot

2. Difference between Bar plot and Plot composition ?

- one rectangle is one phylum (no borderline)
- one color is one phylum
- y axis: normalise to 1 → relative abundance

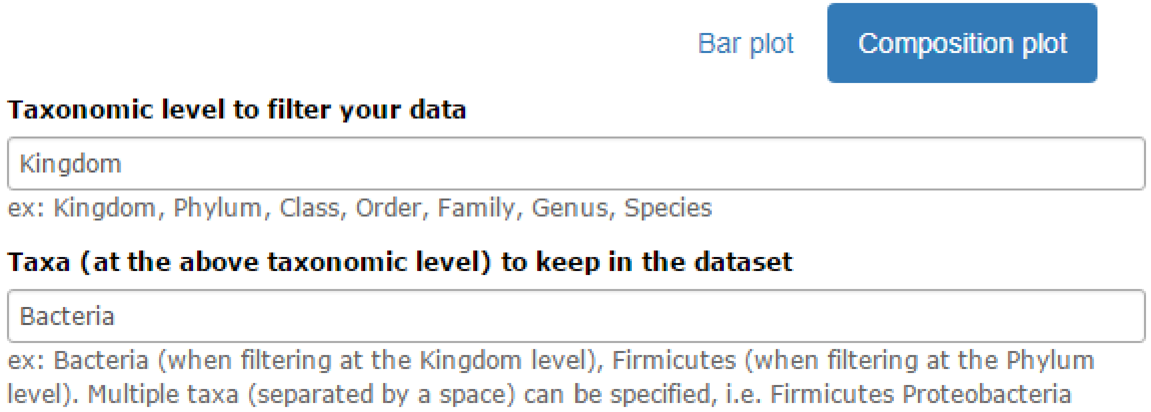


load-extra-functions.R

Exploring biodiversity : visualisation

Customization: plot_composition function :

- Works with relative abundances
- Subsets OTUs at a given taxonomic level



Taxonomic level to filter your data

Kingdom

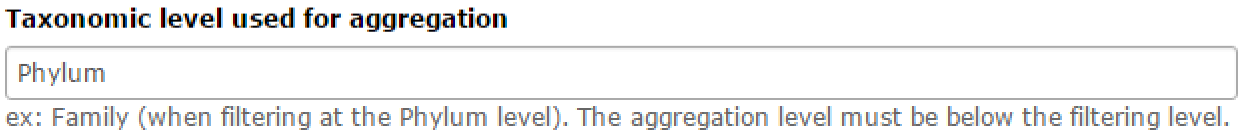
ex: Kingdom, Phylum, Class, Order, Family, Genus, Species

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

- Aggregates OTUs at another taxonomic level

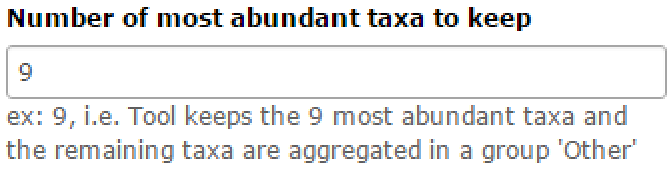


Taxonomic level used for aggregation

Phylum

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

- Shows only a given number of OTUs



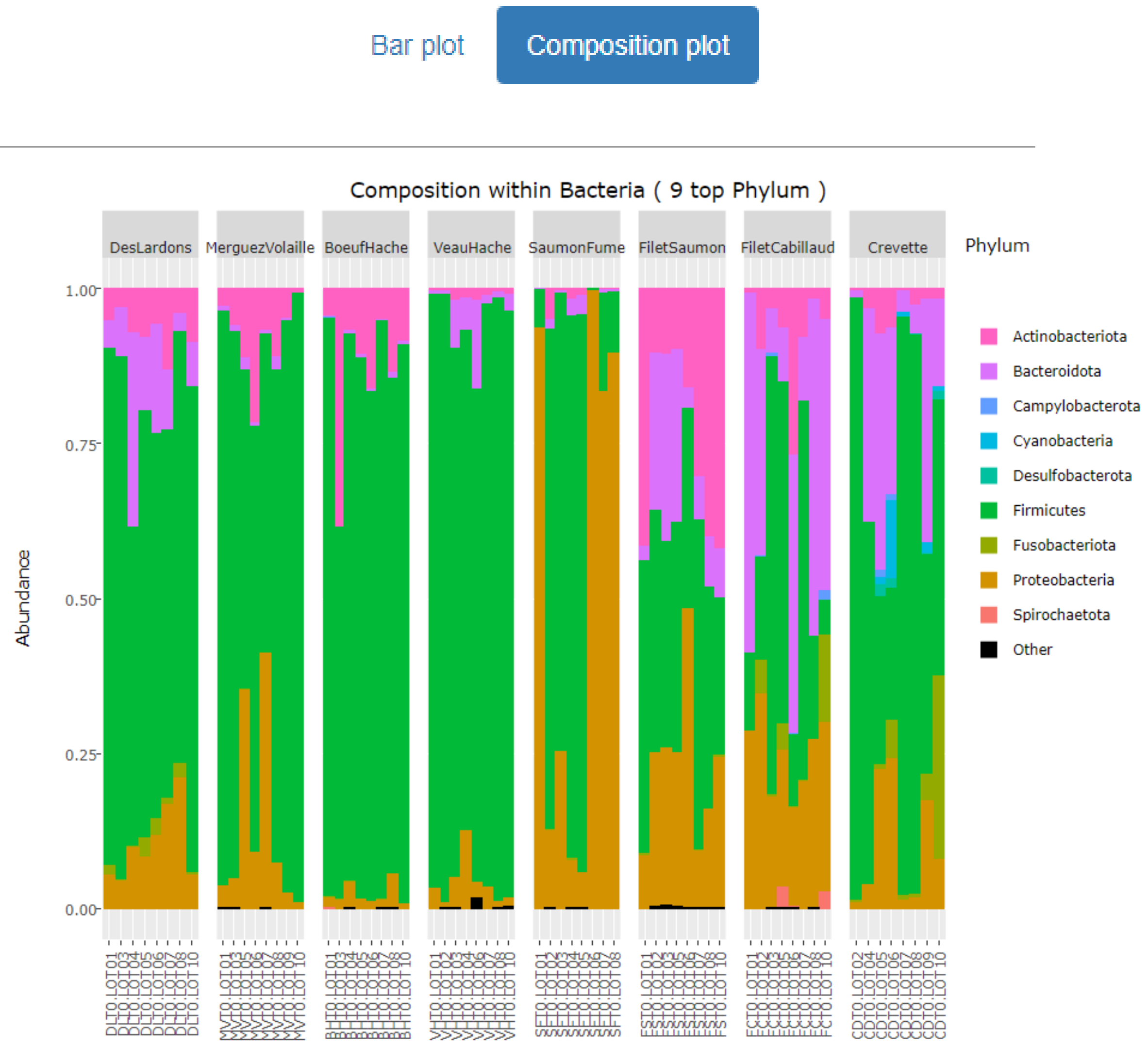
Number of most abundant taxa to keep

9

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

Exercise 3

3. Information ?



Exercise 3

Bar plot

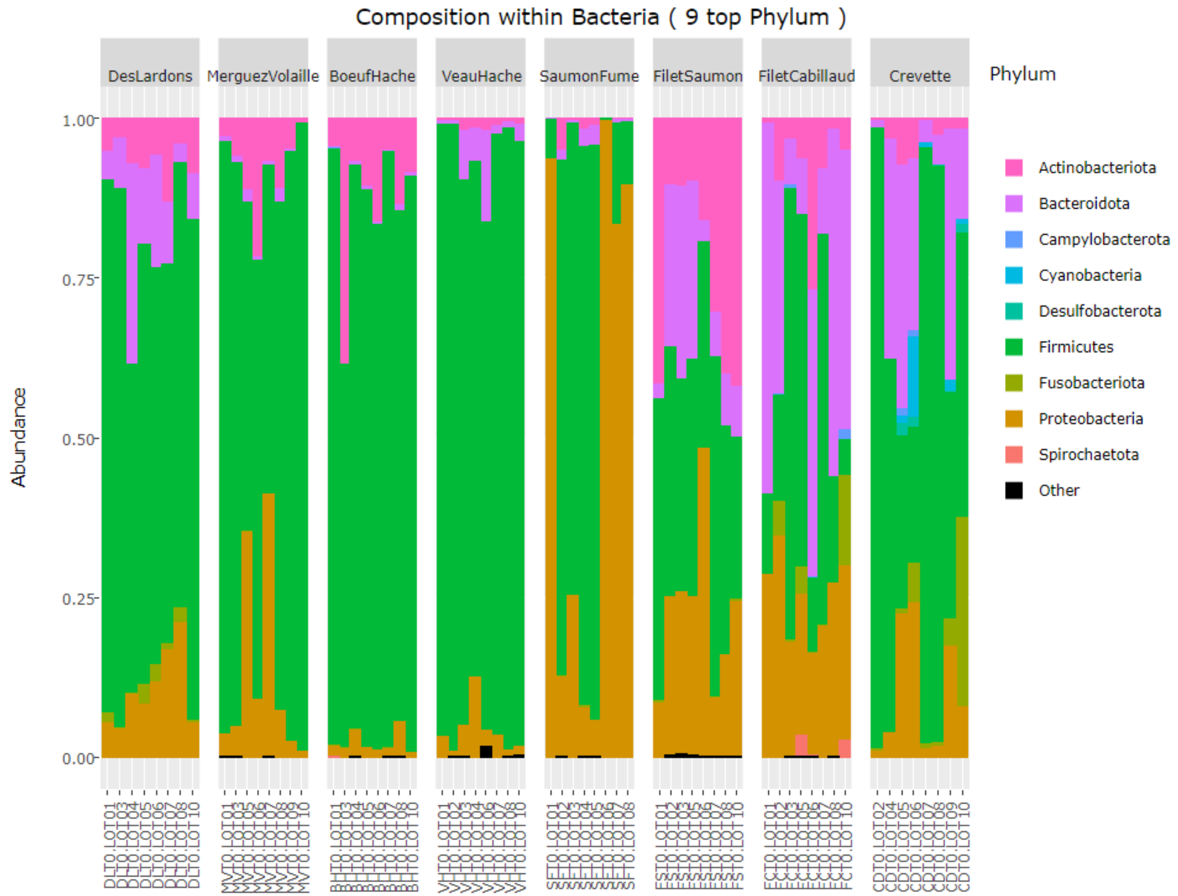
Composition plot

3. Information ?

- Meat type on the left share common Phylum composition, with a majority of Firmicutes
  (easy to remark thanks of ordered levels)

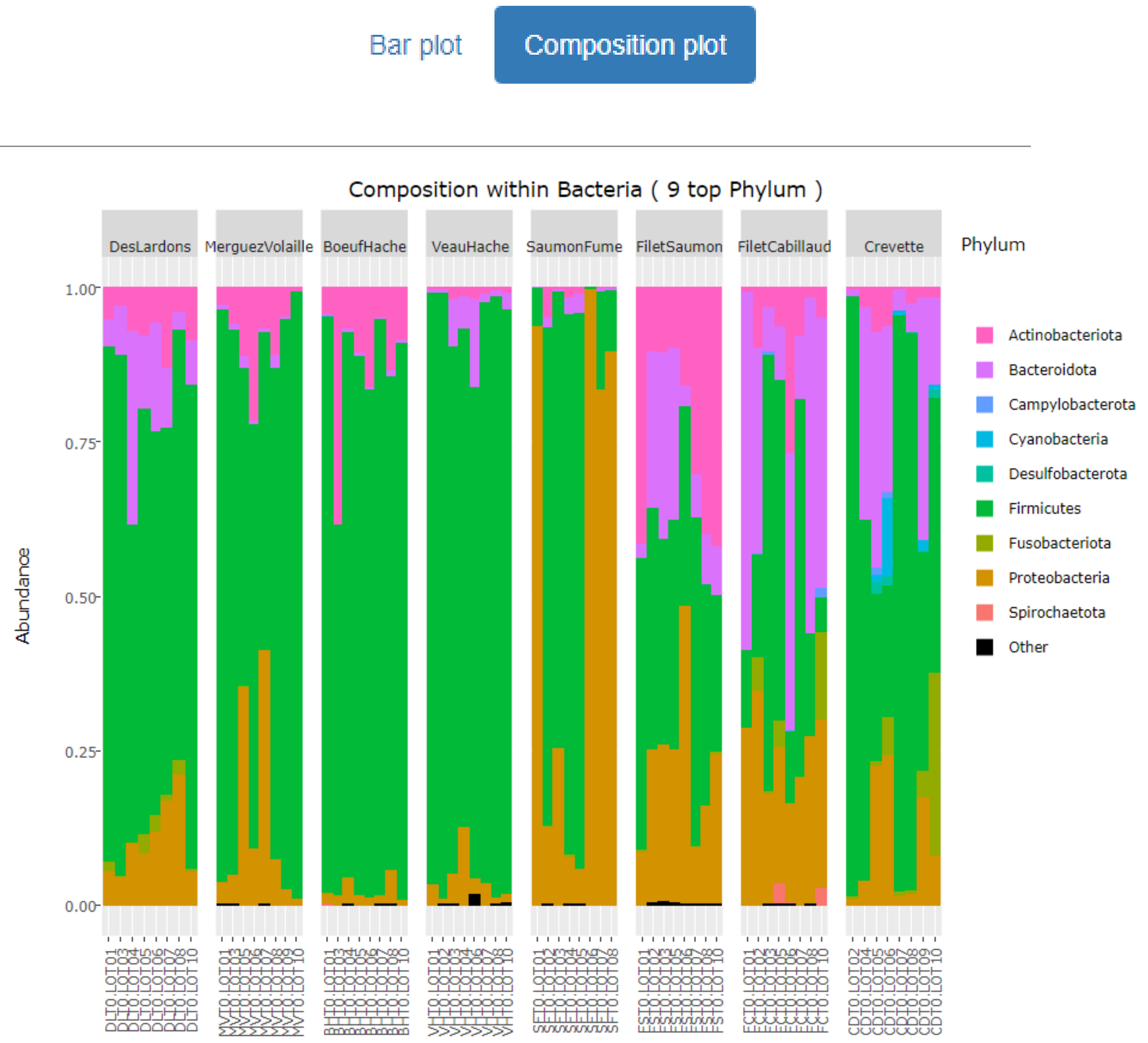
- Seafoods seem to be much more variable

- Firmicutes and Proteobacteria are present in all samples, but with a wide range of abundance



Exercise 3

4. Perspectives to go further ?



Exercise 3

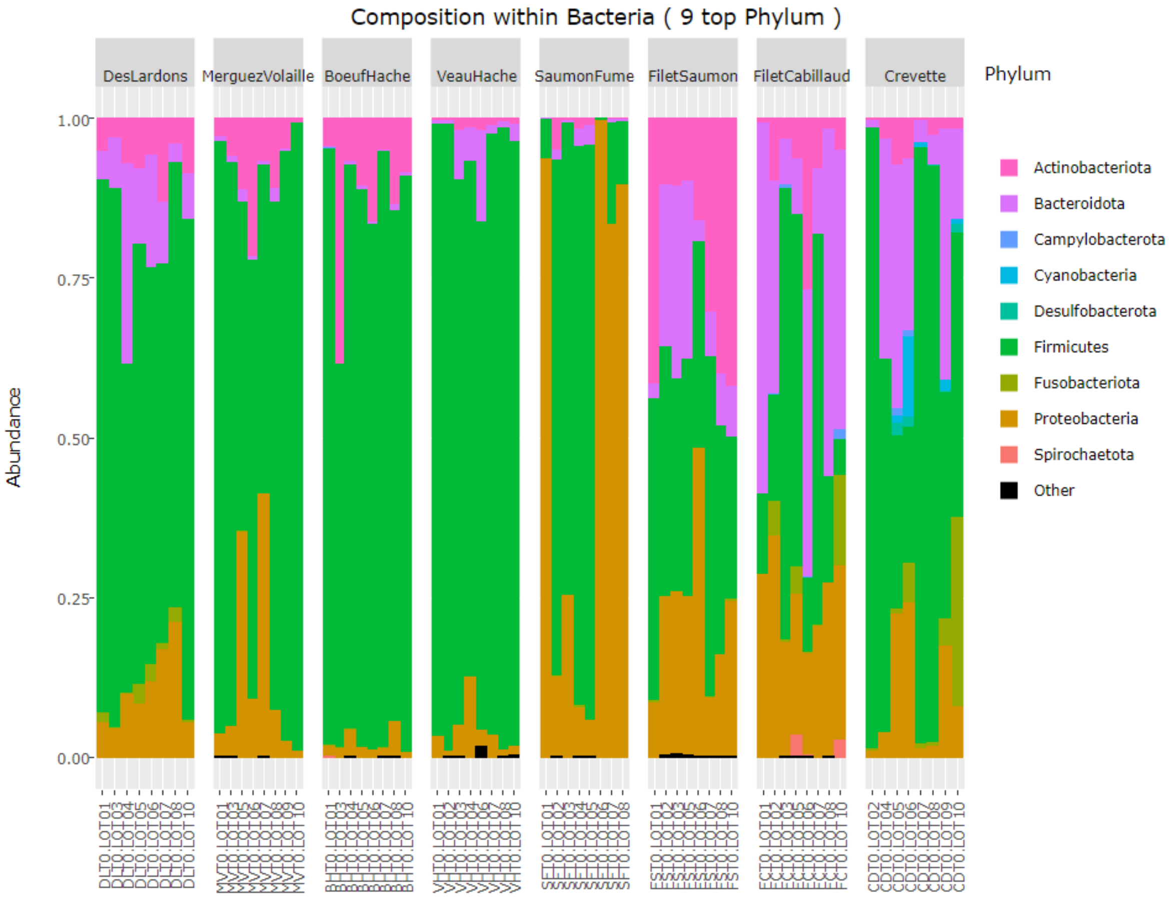
Bar plot

Composition plot

4. Perspectives to go further ?

➔ What are the composition of the 9 most abundant Families of *Firmicutes* ?

➔ What are the composition of the 9 most abundant Families of *Proteobacteria* ?



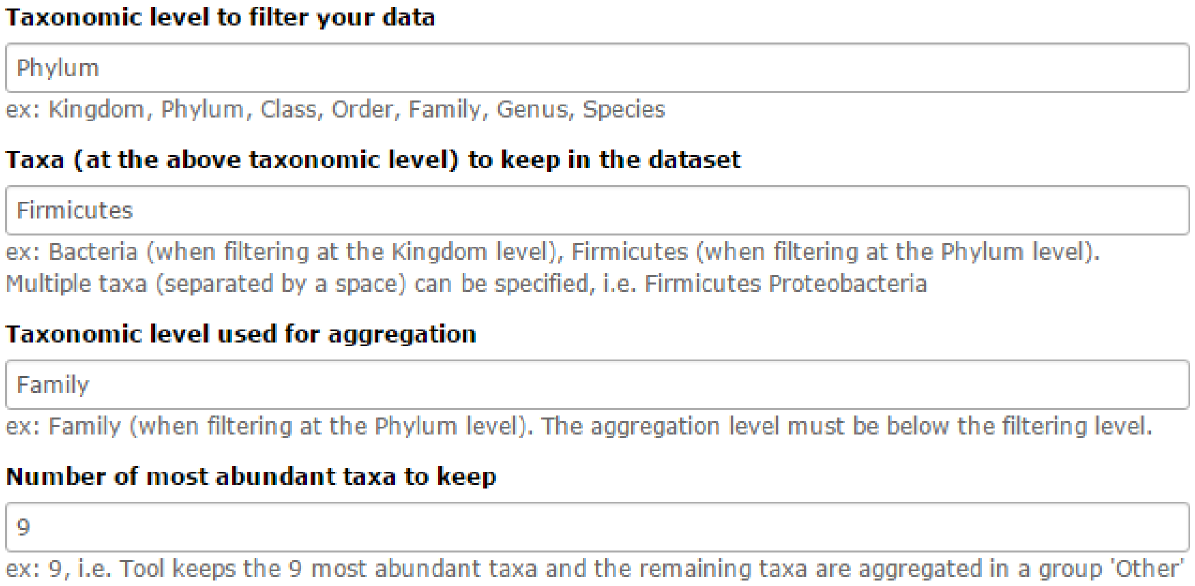
Exercise 4

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

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

Exercise 4

1. What are 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

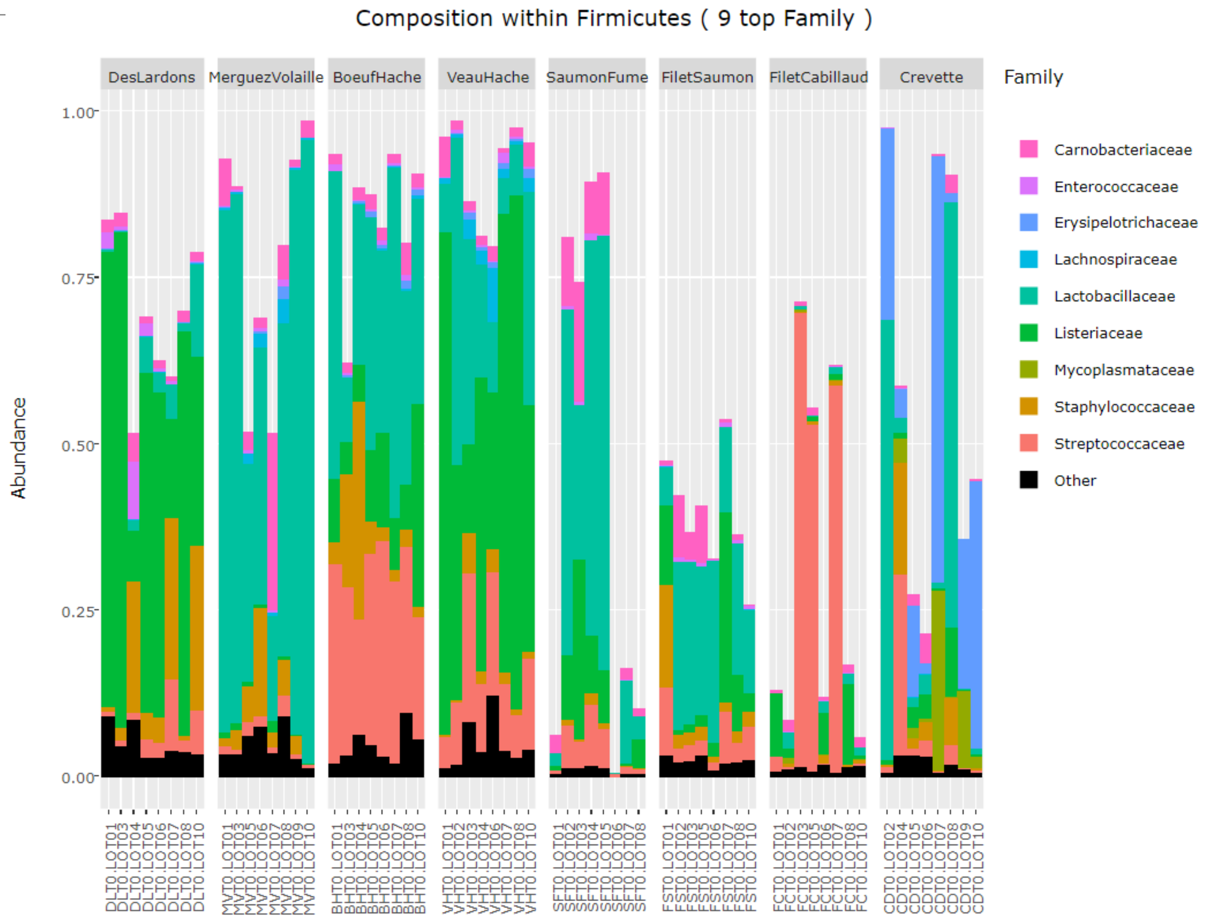
Family

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'

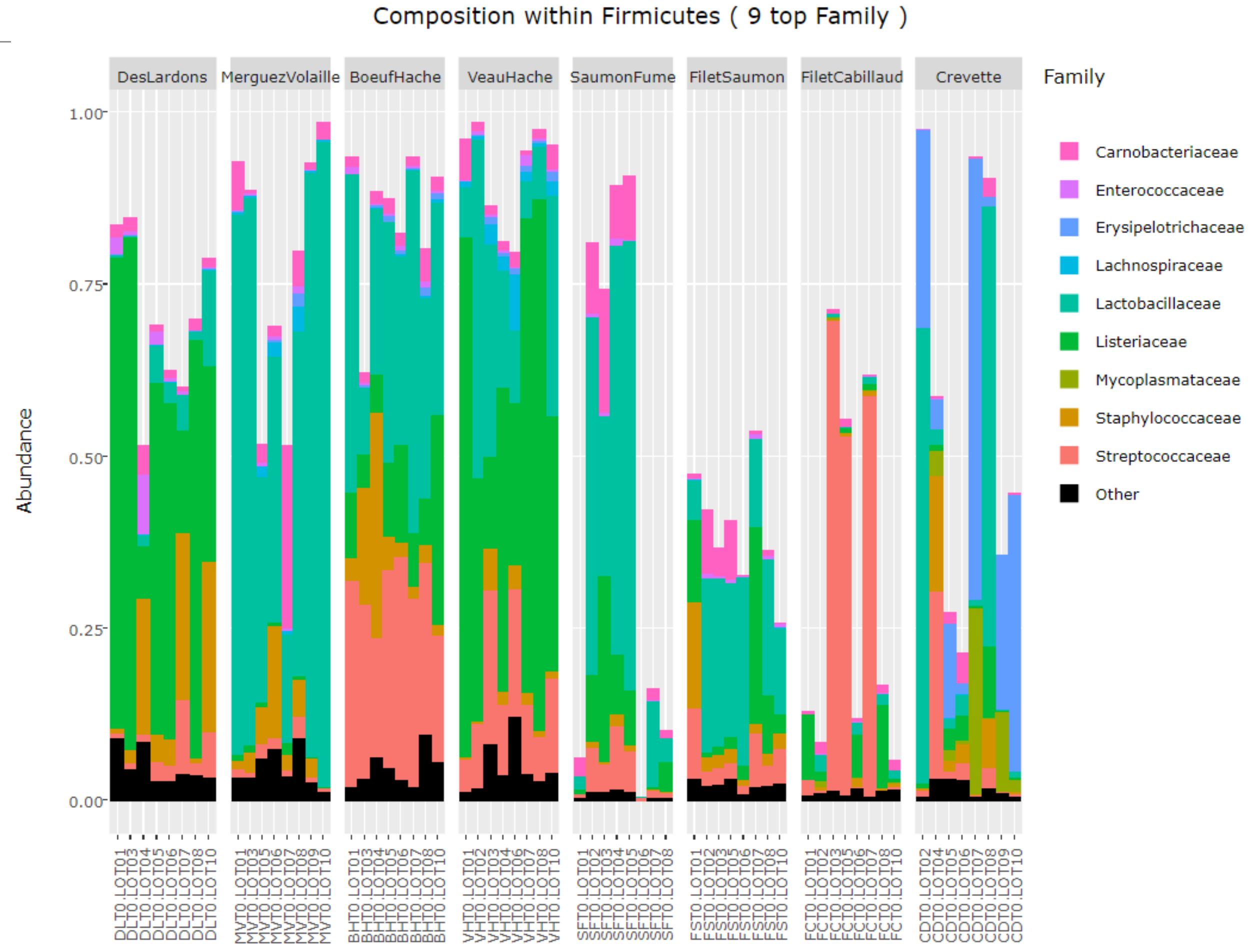


Composition within Firmicutes (9 top Family)

Exercise 4

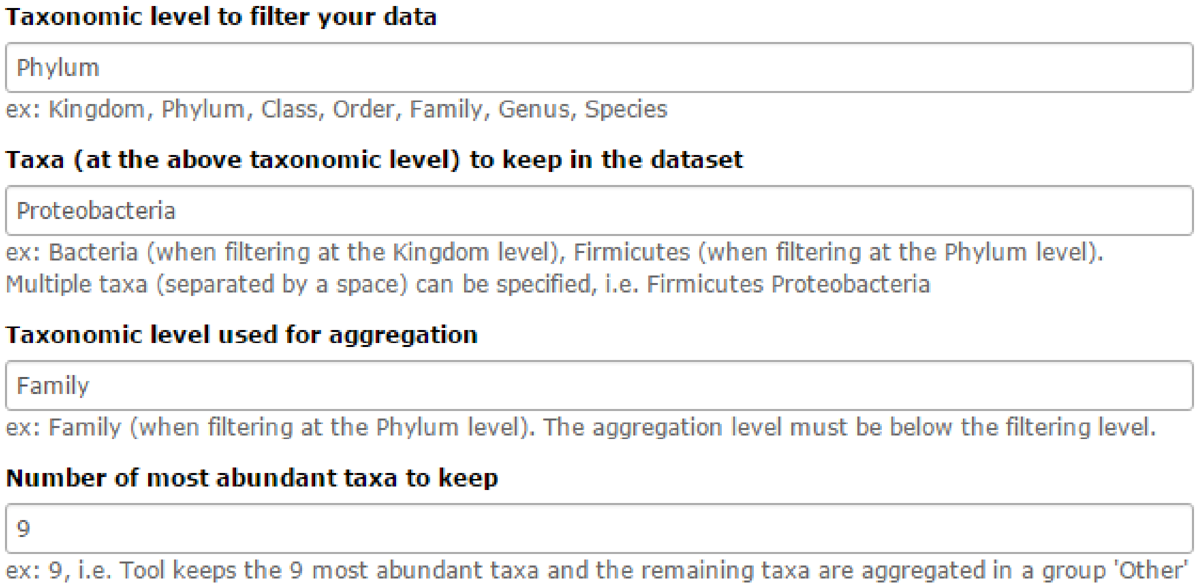
1. What are the composition of the 9 most abundant Families of Firmicutes ?

- top 9 families of Firmicutes are most represented in meat food



Exercise 4

2. What are 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

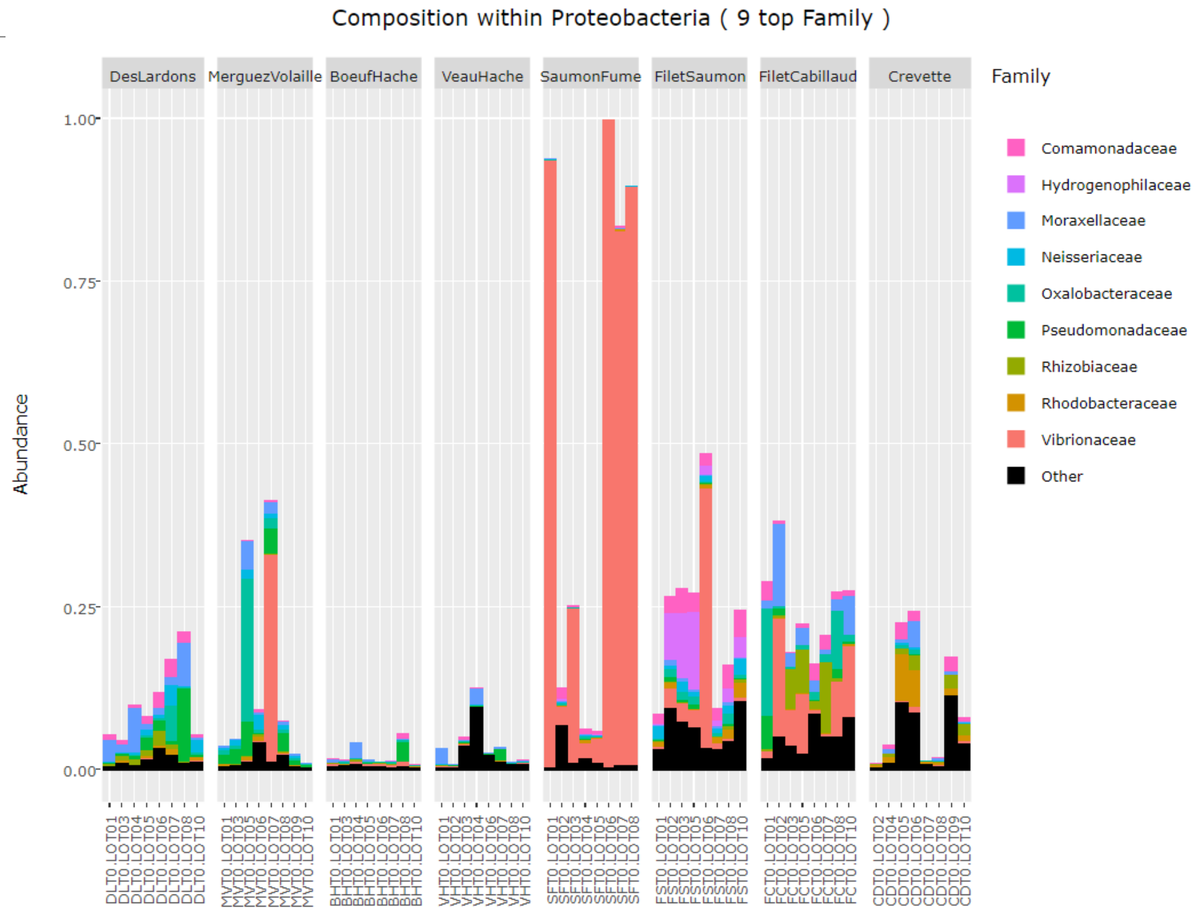
Family

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'

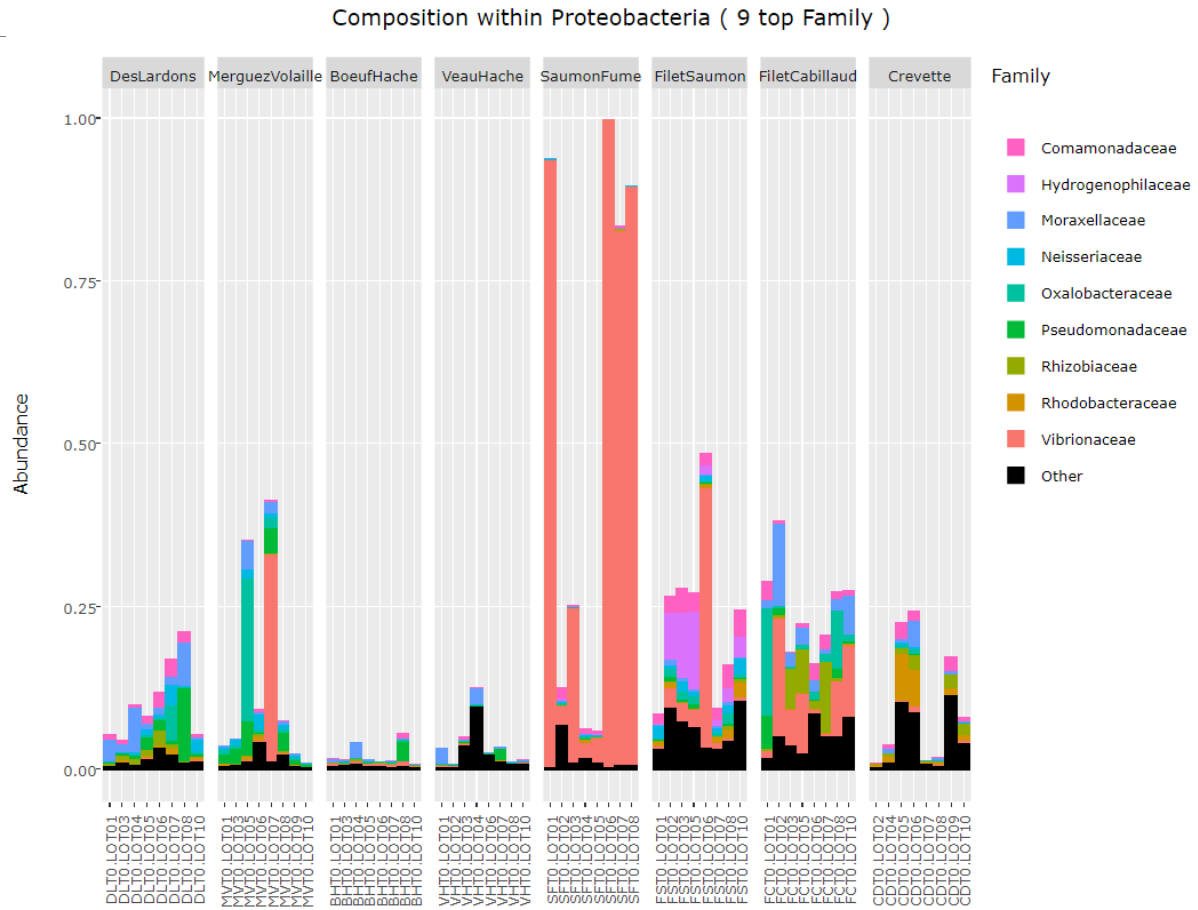


Exercise 4

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

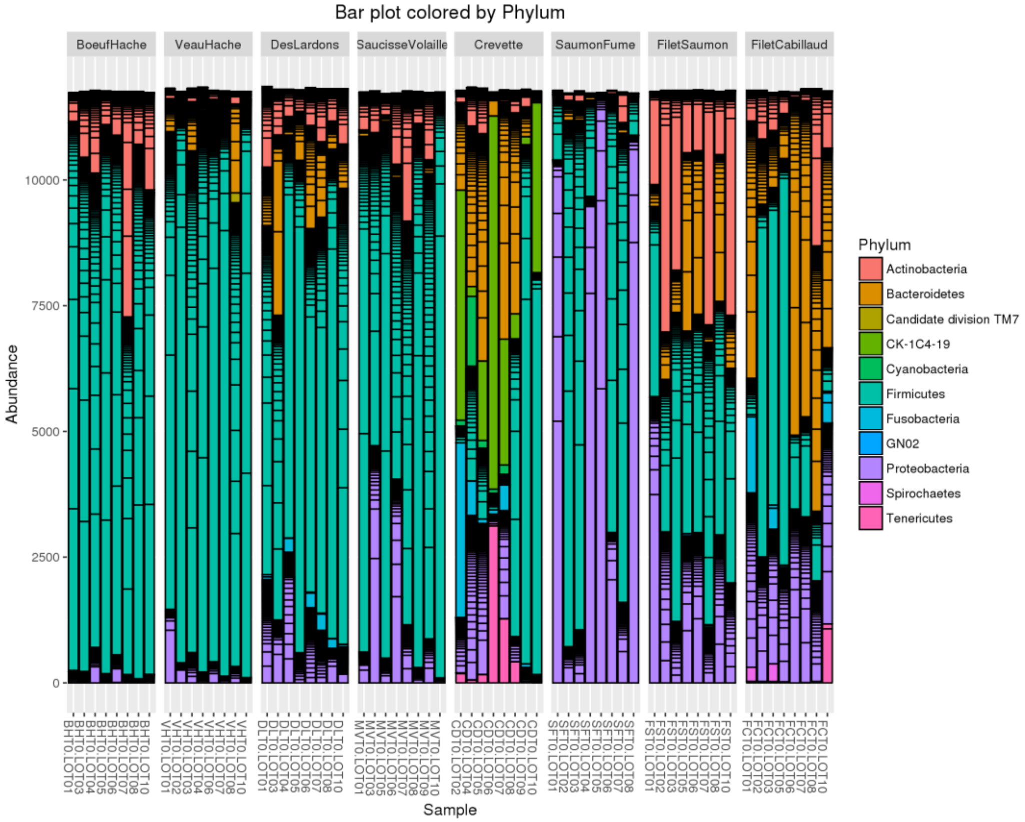
- top 9 families of proteobacteria are most represented in seafood

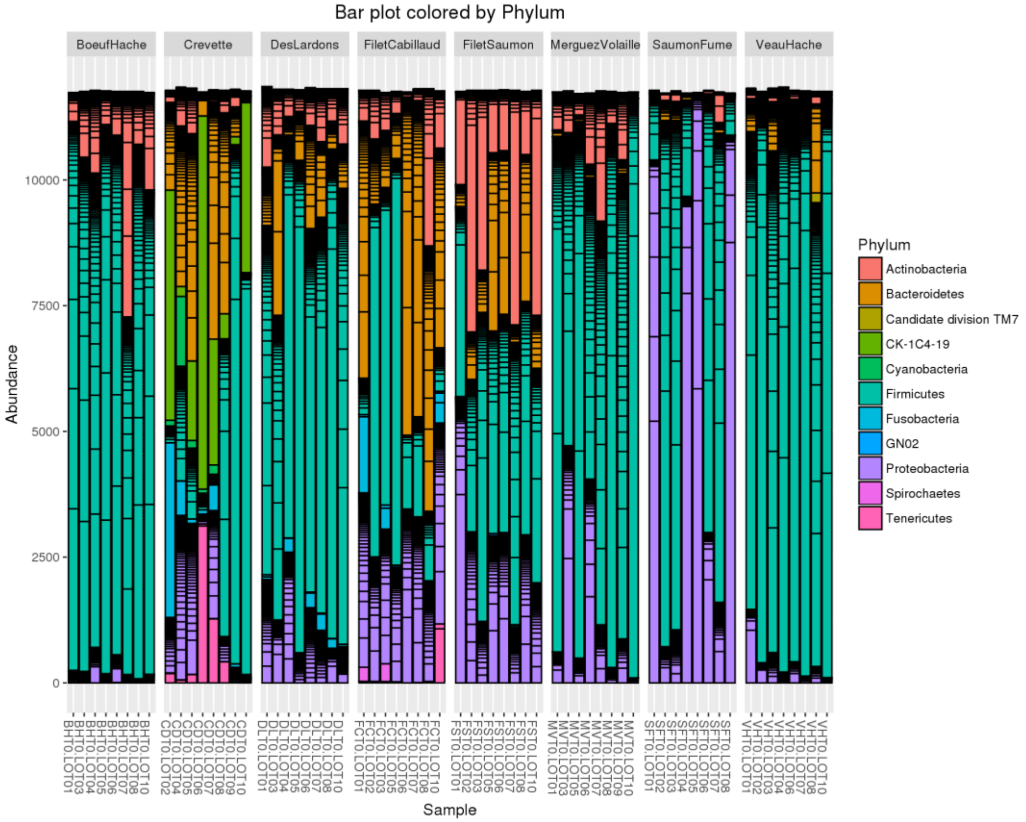
- Vibrionaceae dominate in SaumonFume for 4 samples



Exploring biodiversity : visualisation

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

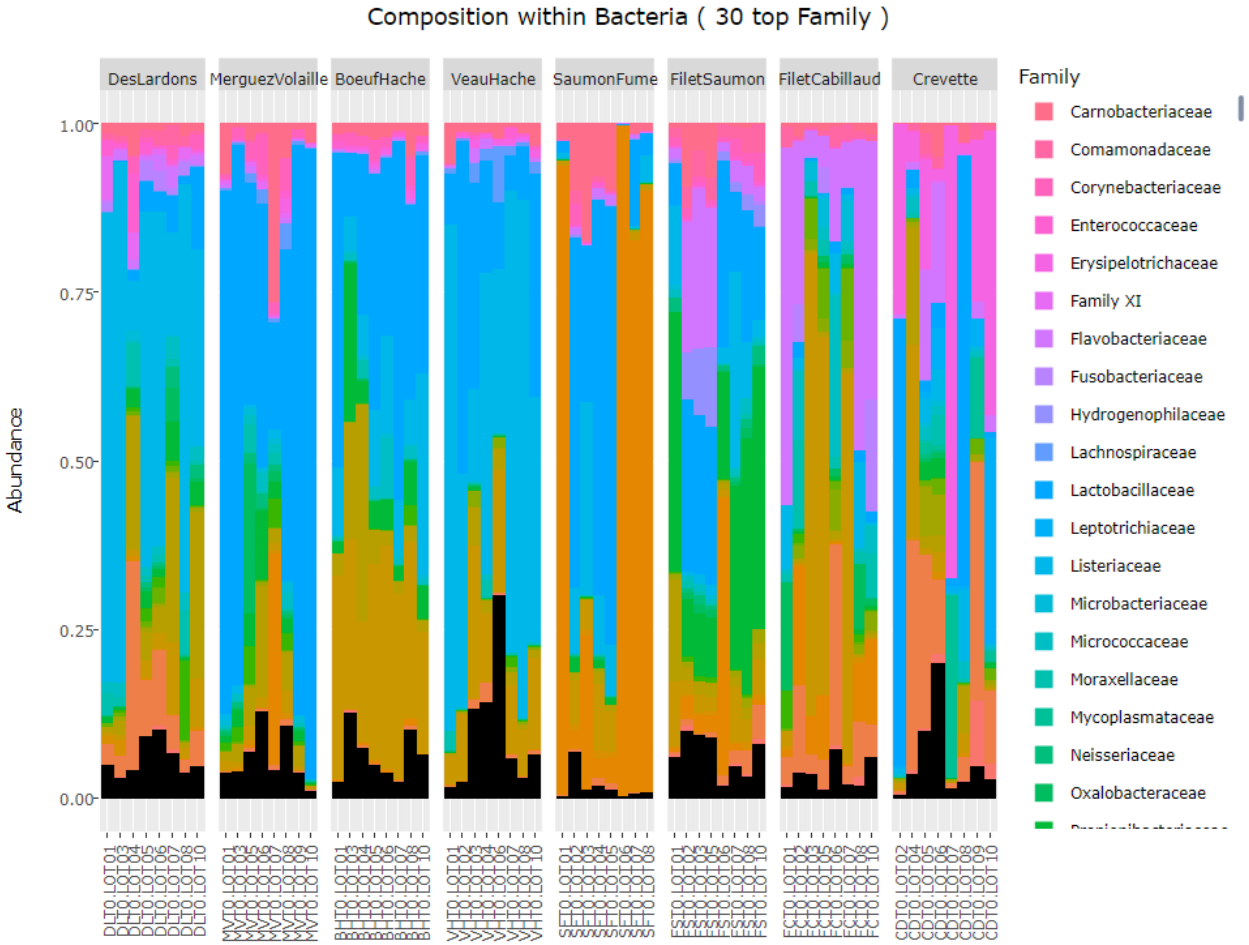




Exploring biodiversity : visualisation

<u>Remark 2</u> : 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



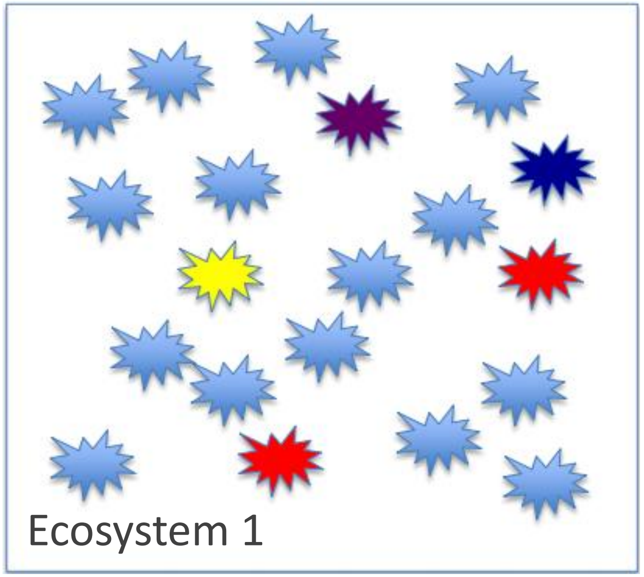
II. Biodiversity analysis

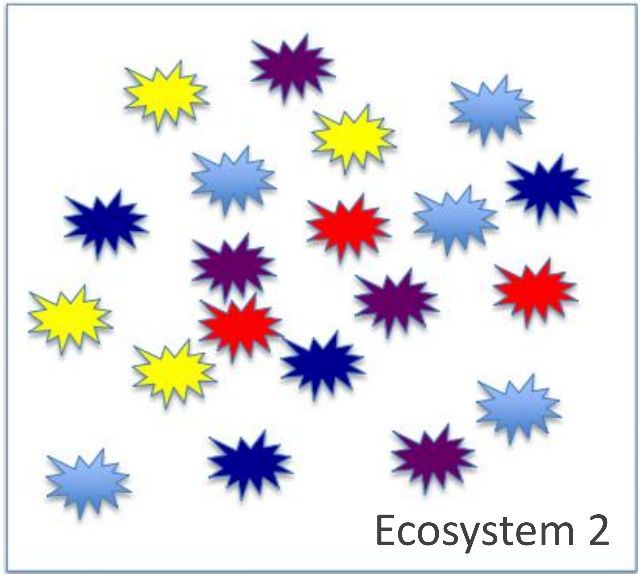
DIVERSITY INDICES

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**

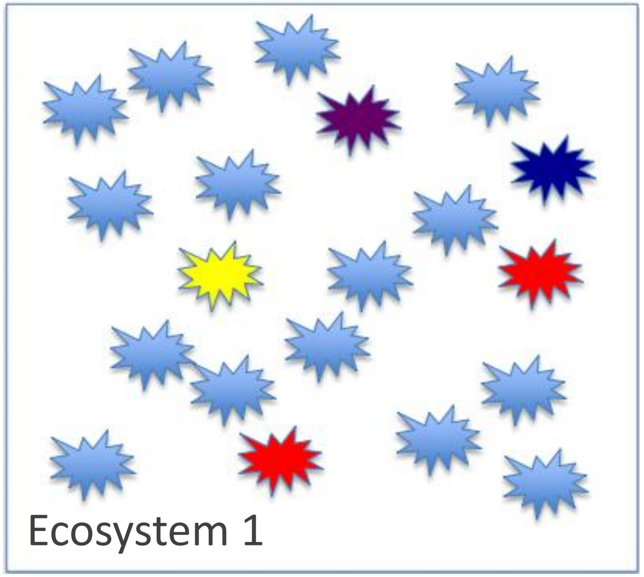


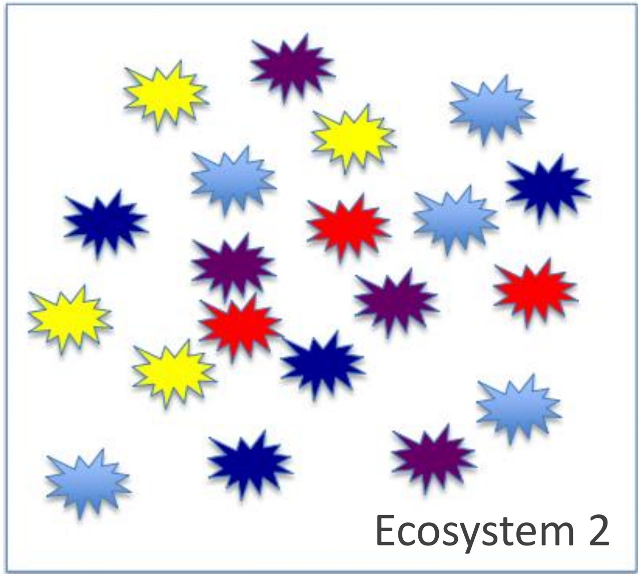


Exploring biodiversity : descriptors

- 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 abundancy of species. It characterizes the **structure**





Richness : Eco1 = Eco2

Diversity: Eco2 > Eco1

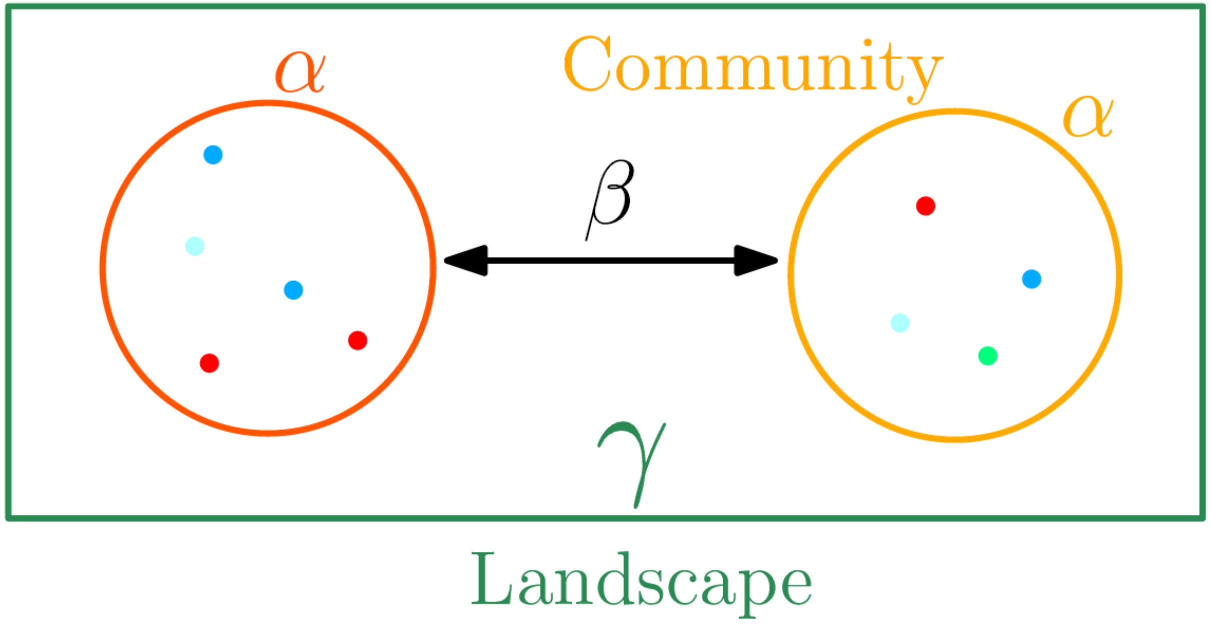
Exploring biodiversity : statistical indices

Compute and compare diversity indices. 3 levels of diversity:

- **α-diversity**: diversity within a community

- **β-diversity**: diversity between communities
  - β-dissimilarities/distances
    - dissimilarities between pairs of communities
    - often used as a first step to compute diversity

- γ-diversity: diversity at the landscape scale (blurry for bacterial communities)



Exploring biodiversity : statistical indices

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

- Qualitative gives less weight to dominant species
- Qualitative is more sensitive to differences in sampling depths
- Qualitative indices emphasize differences in taxa diversity while quantitative are more sensitive to raise differences in composition

Compositional vs. Phylogenetic

- Compositional does not require a phylogenetic tree
- Compositional is more sensitive to erroneous OTU picking
- Compositional gives the same importance to all OTUs

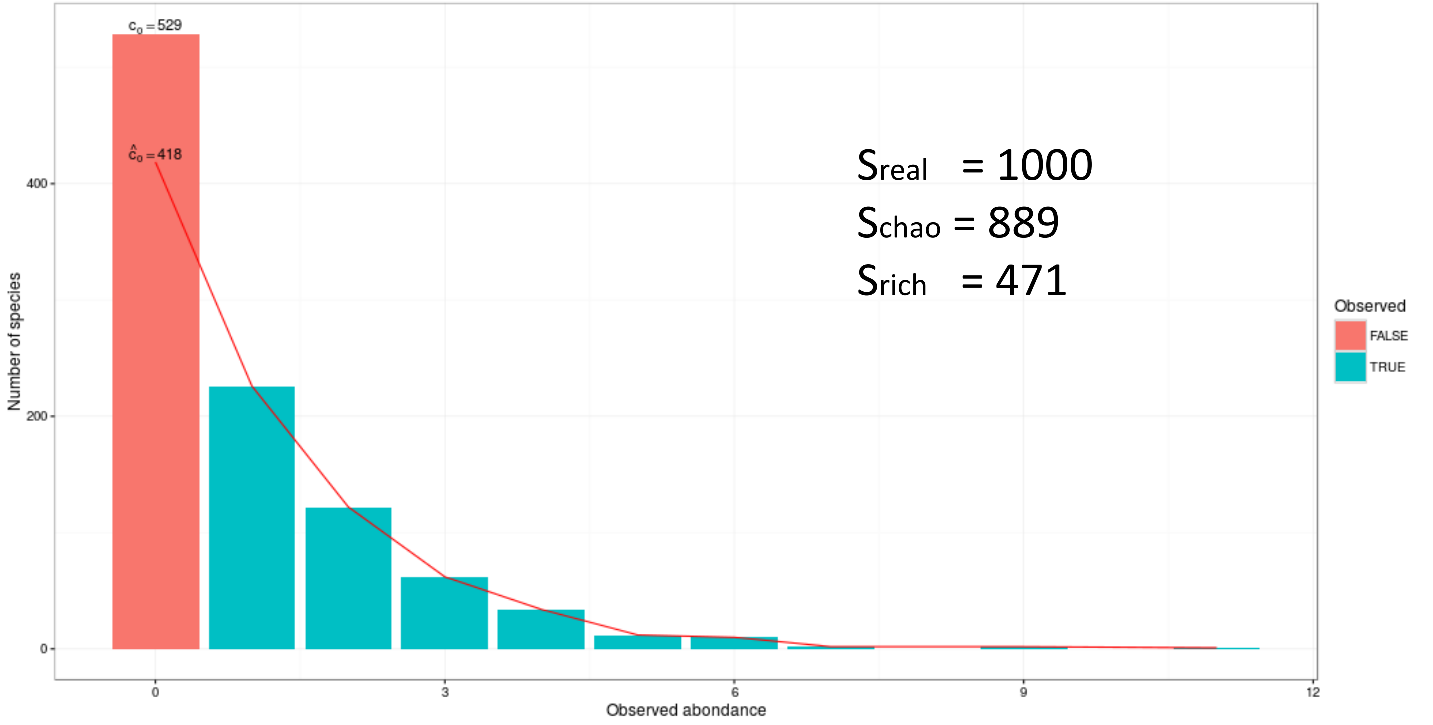
III. Biodiversity analysis

α-DIVERSITY INDICES

Exploring biodiversity : α-diversity

α-diversity is equivalent to the richness : number of species

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

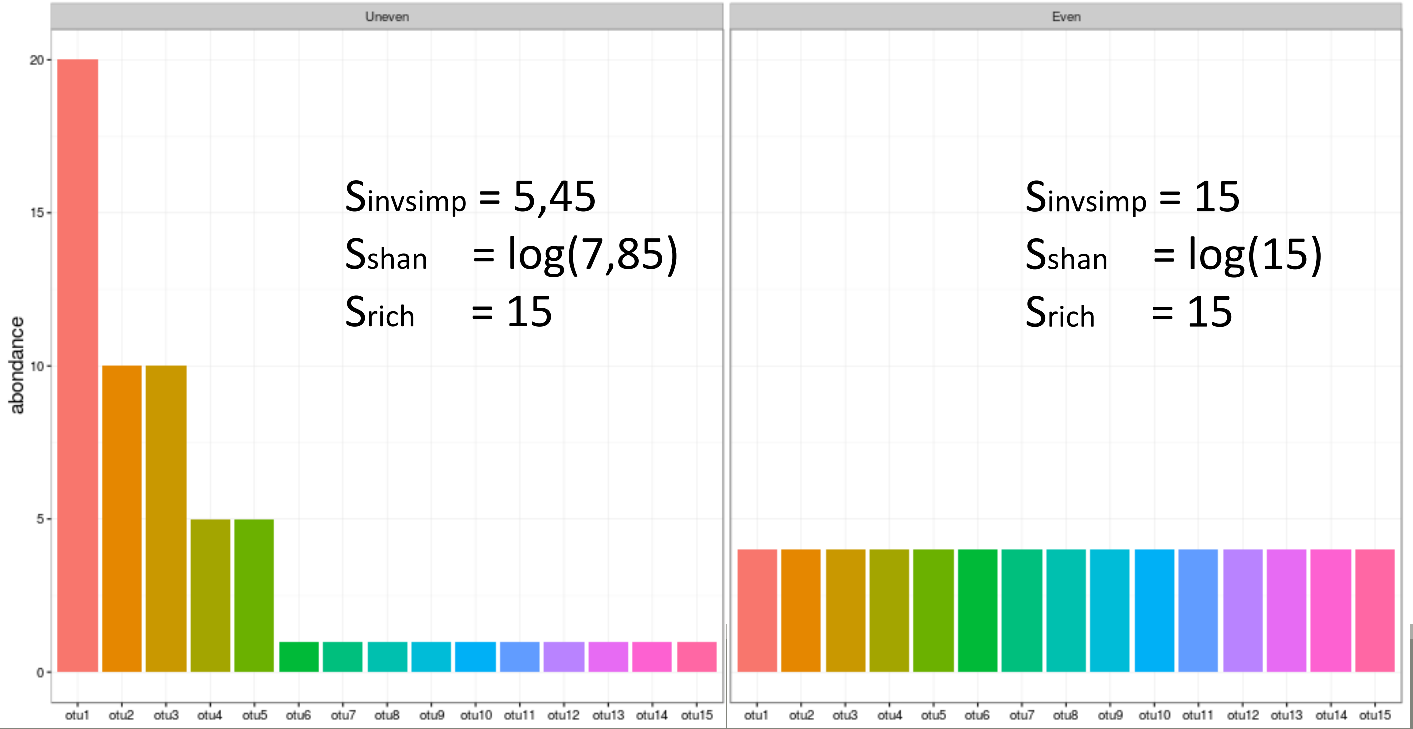


S_{real} = 1000
S_{chao} = 889
S_{rich} = 471

Exploring biodiversity : α-diversity

α-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



$S_{invsimp} = 5,45$
$S_{shan} = \log(7,85)$
$S_{rich} = 15$

$S_{invsimp} = 15$
$S_{shan} = \log(15)$
$S_{rich} = 15$

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

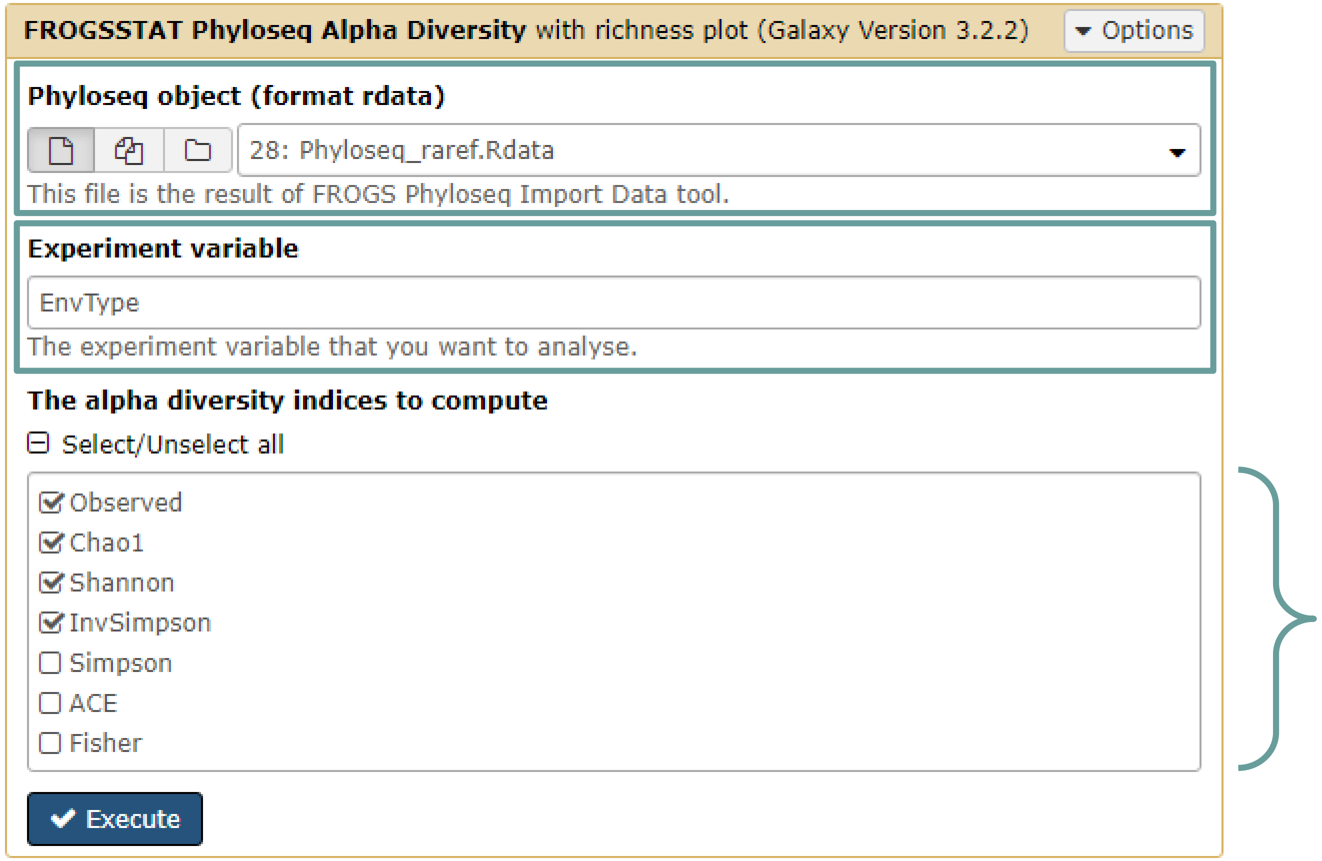
It is called effective diversities

Exploring biodiversity : α-diversity

α-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

Exploring biodiversity : α-diversity



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

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 α-diversity indices you want to compute

Exercise 5

1. What are the resulting datasets ?

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

3. Have EnvType got an impact on α-diversity indices ?

Exercise 5

1. What are the resulting datasets ?

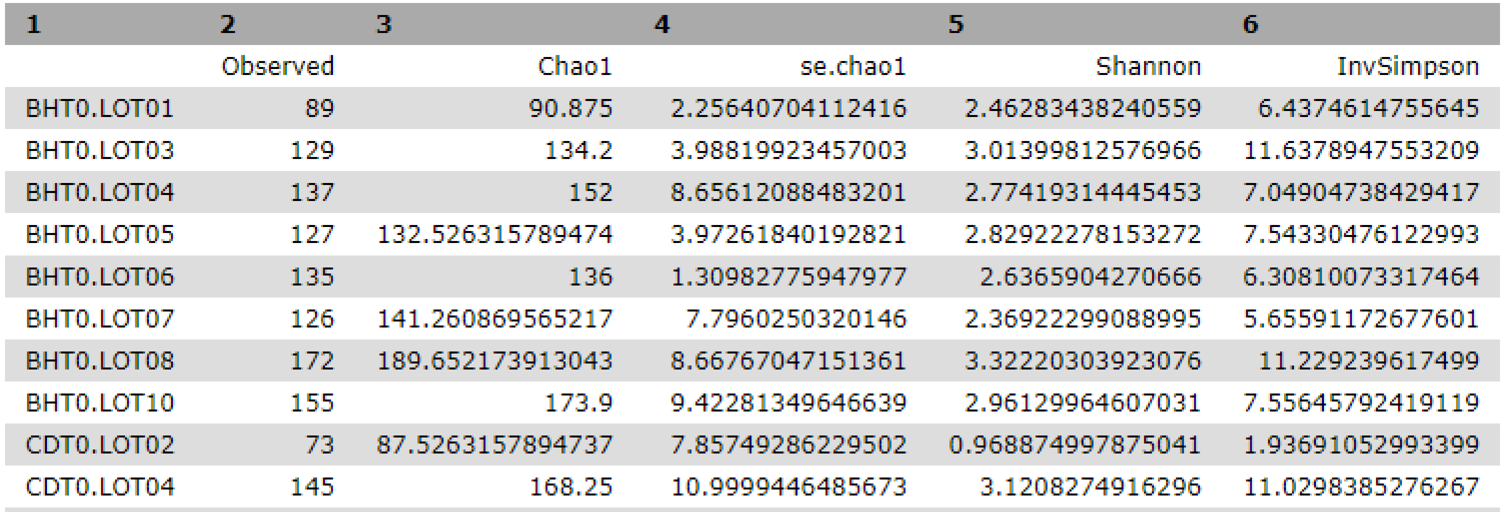
→ Tabular file: contain the detailed value of indices in each sample

→ HTML report: graphical and statistical results

Exercise 5

1. What are the resulting datasets ?

→ **Tabular file:** contain 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

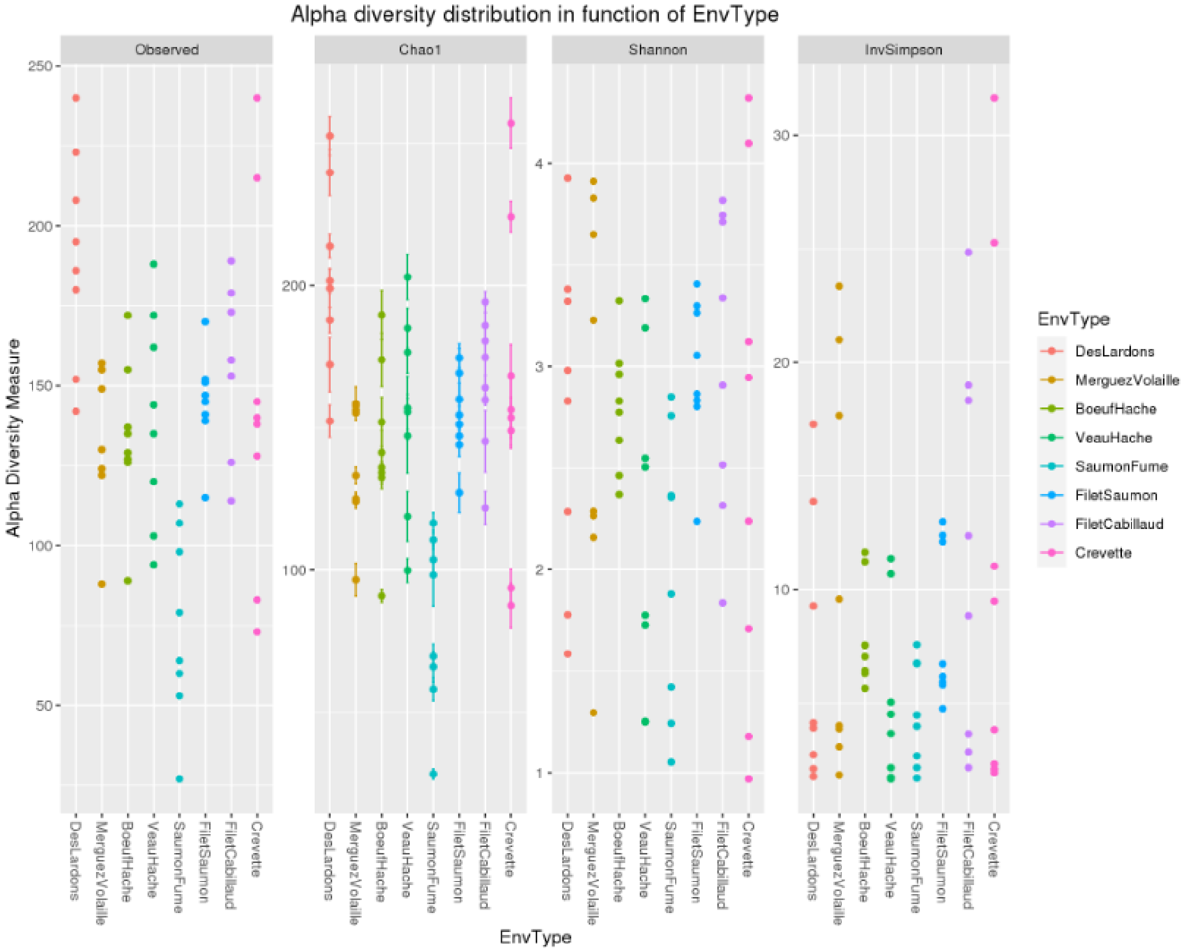
Exercise 5

1. What are the resulting datasets ?

→ HTML report: graphical and statistical results

Exercise 5

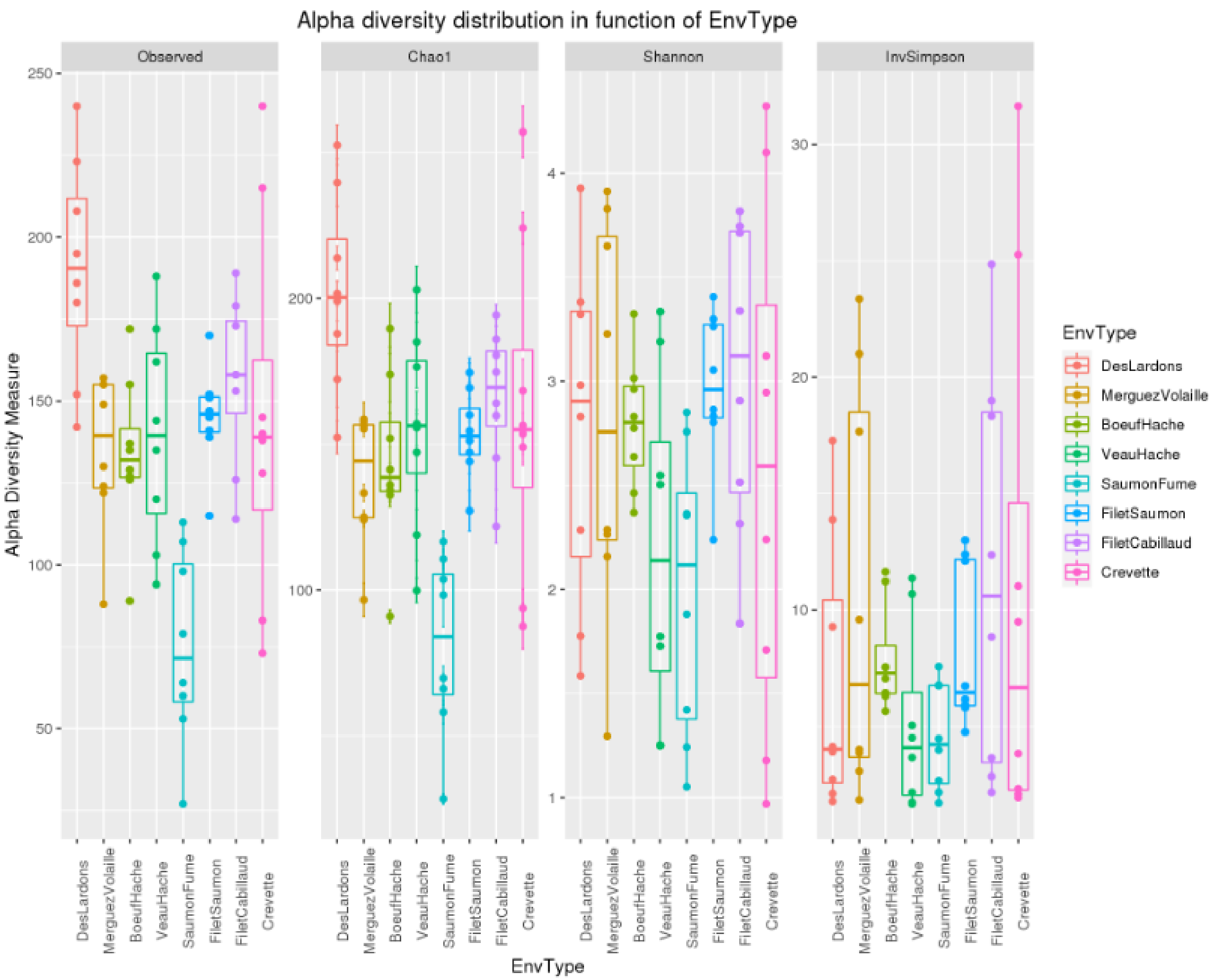
Richness plot
Richness plot with boxplot
Alpha Diversity Indice Anova Analysis
Rarefaction curves


Alpha diversity distribution in function of EnvType
Observed
Chao1
Shannon
InvSimpson
Alpha Diversity Measure
EnvType
250
200
150
100
50
200
100
4
3
2
1
30
20
10
DesLardons
MerguezVolaille
BoeufHache
VeauHache
SaumonFume
FiletSaumon
FiletCabillaud
Crevette
EnvType
DesLardons
MerguezVolaille
BoeufHache
VeauHache
SaumonFume
FiletSaumon
FiletCabillaud
Crevette

Exercise 5

Richness plot | **Richness plot with boxplot** | Alpha Diversity Indice Anova Analysis | Rarefaction curves

Informations ?

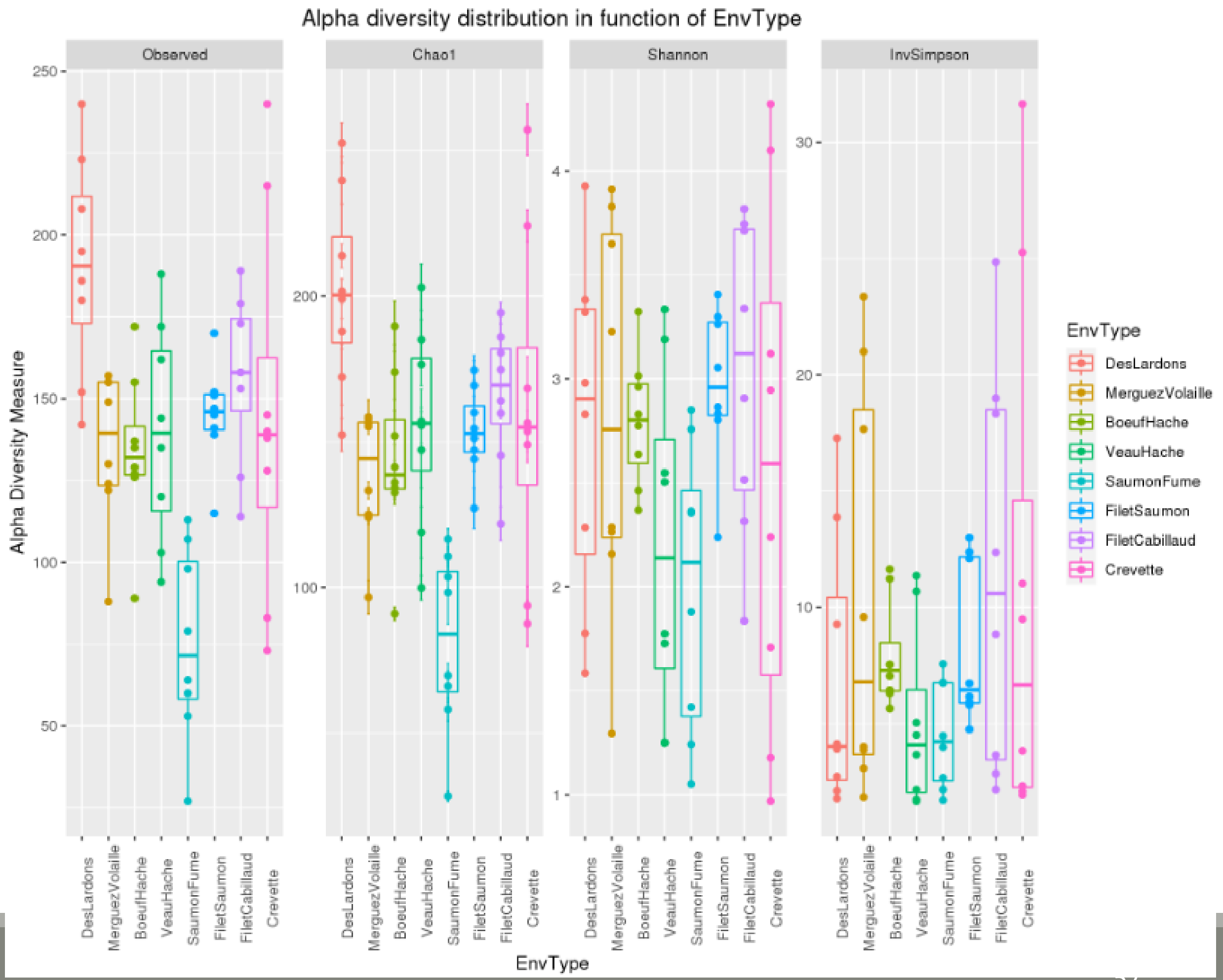


Exercise 5

Richness plot | **Richness plot with boxplot** | Alpha Diversity Indice Anova Analysis | Rarefaction curves

Informations ?

- 4 plots for the 4 indices
- Same legend for all plots
- x axis: 8 boxplot for each EnvType, dots represent samples
- y axis: values of each alpha index
- Scales in y axis are different



Exercise 5

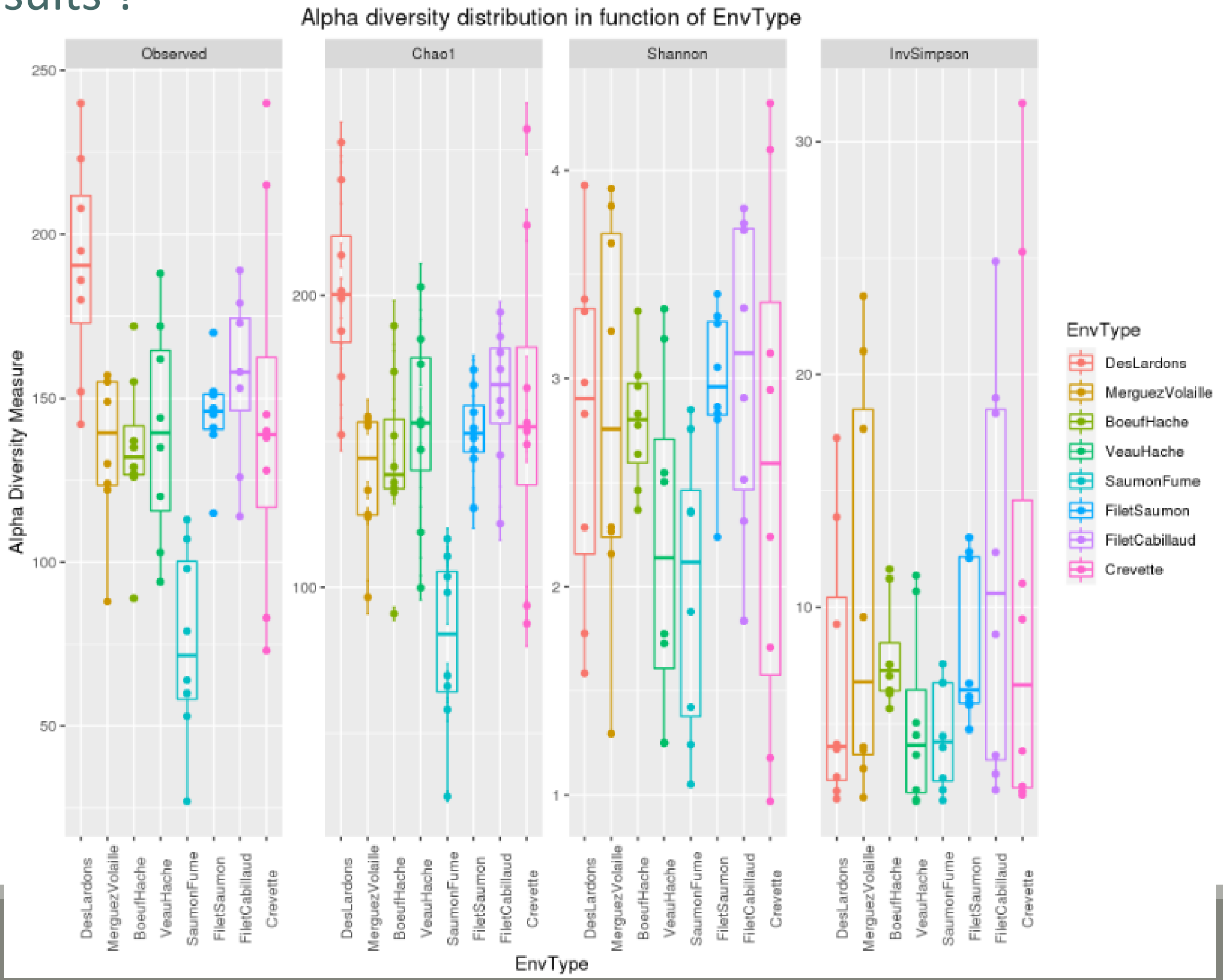
Richness plot

Richness plot with boxplot

Alpha Diversity Indice Anova Analysis

Rarefaction curves

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



Exercise 5

Richness plot | Richness plot with boxplot | Alpha Diversity Indice Anova Analysis | Rarefaction curves

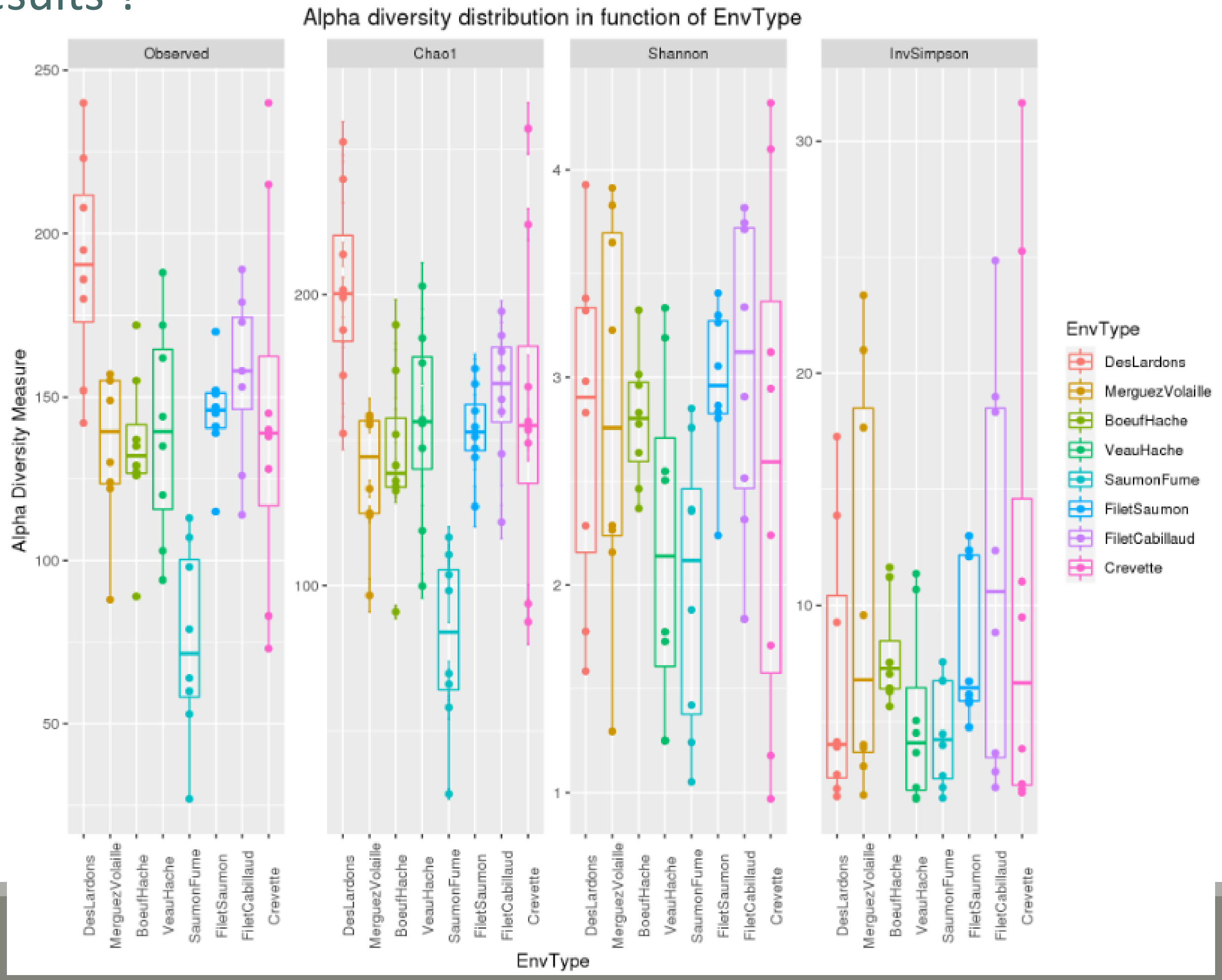
2. Which interpretation could you make on the boxplot results ?

- Observed and Chao have almost the same scale

  → All species have been detected

- Many taxa observed in DesLardons (high Chao1, high Observed)
- Most foods have low effective diversities (InvSimpson)

  → communities are dominated by few abundant taxa



Exercise 5

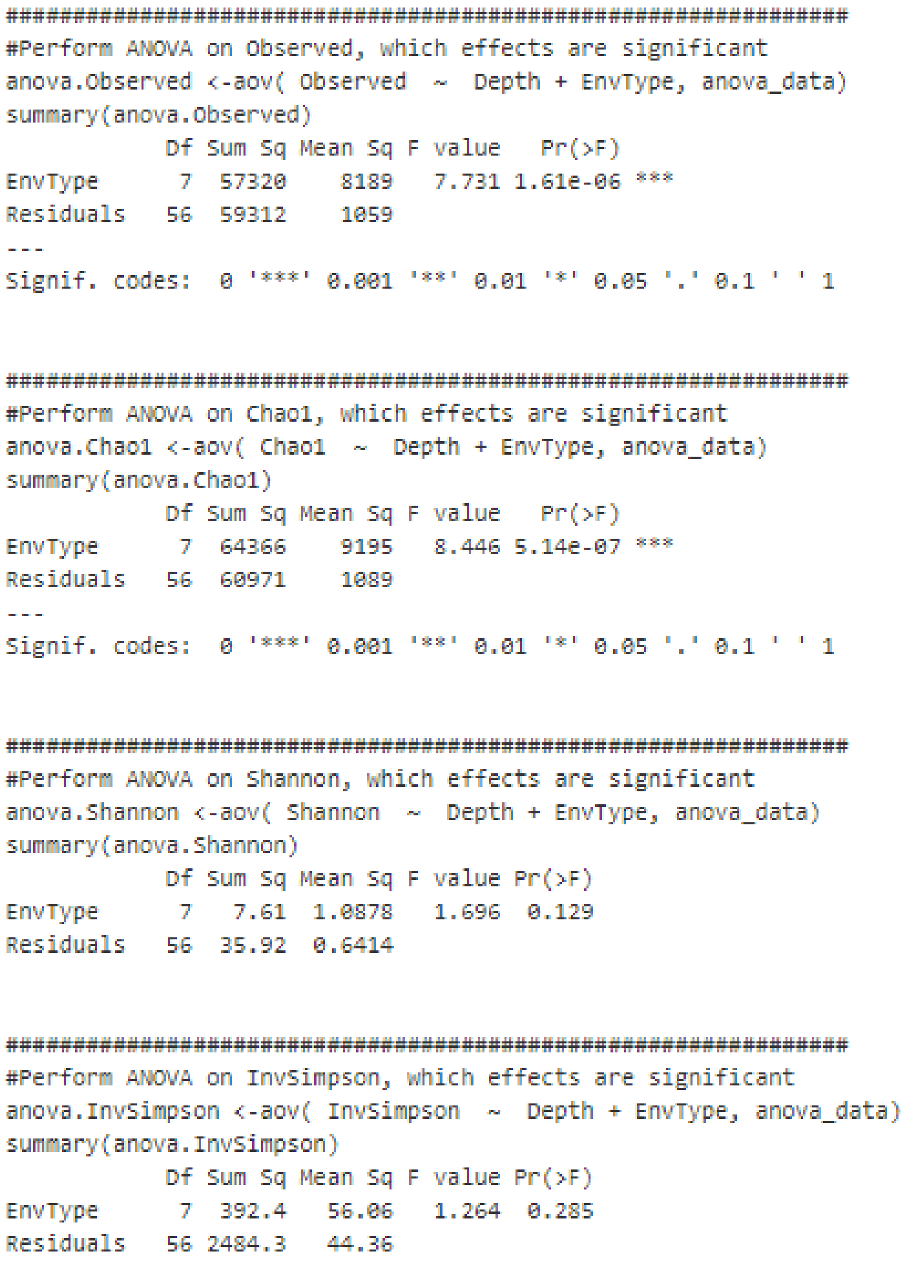
Richness plot
Richness plot with boxplot
Alpha Diversity Indice Anova Analysis
Rarefaction curves

Test the significance of the previous observations by performing an ANOVA of alpha-diversity indices against the covariate of interest (EnvType)

Exercise 5

Richness plot   Richness plot with box

Anova interpretations

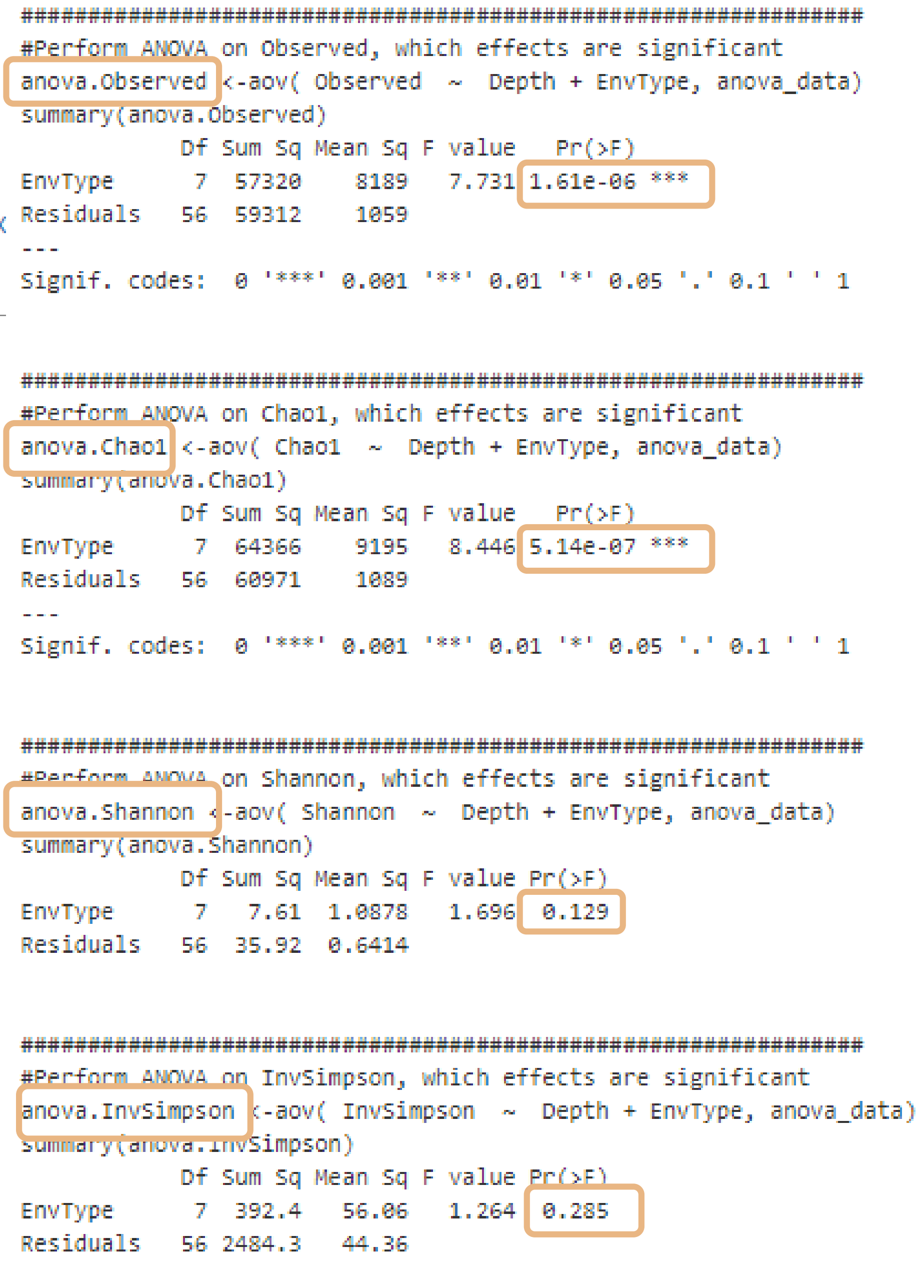


```
###############################################################
#Perform ANOVA on Observed, which effects are significant
anova.Observed <-aov( Observed  ~  Depth + EnvType, anova_data)
summary(anova.Observed)
            Df Sum Sq Mean Sq F value   Pr(>F)
EnvType      7  57320    8189   7.731 1.61e-06 ***
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

Richness plot   Richness plot with box

Anova interpretations



```
################################################################
#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               
```

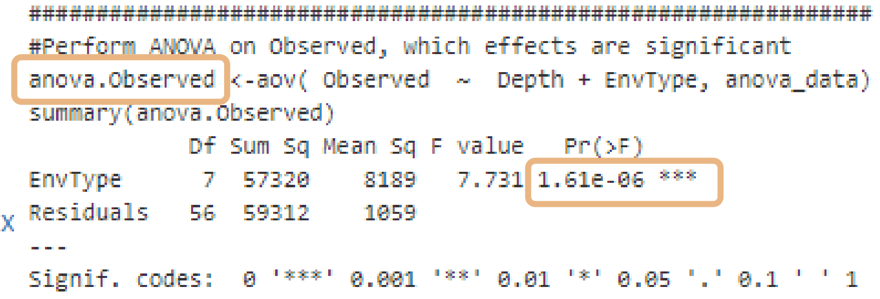
Exercise 5

Richness plot   Richness plot with box

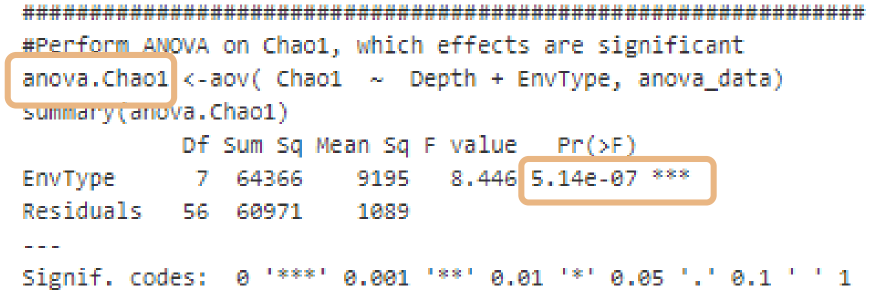
Anova interpretations

- Environments differ a lot in terms of richness...
- ...but not so much in terms of Shannon and InvSimpson diversity

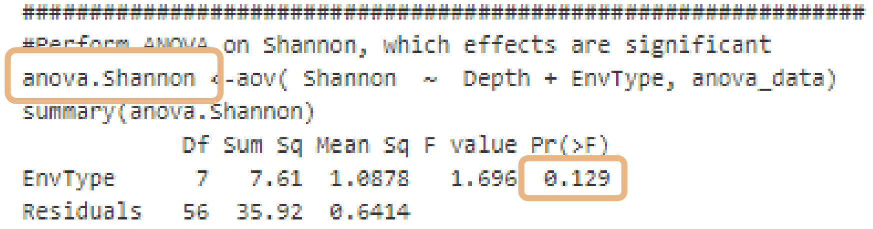
➔ Effective diversities are quite similar



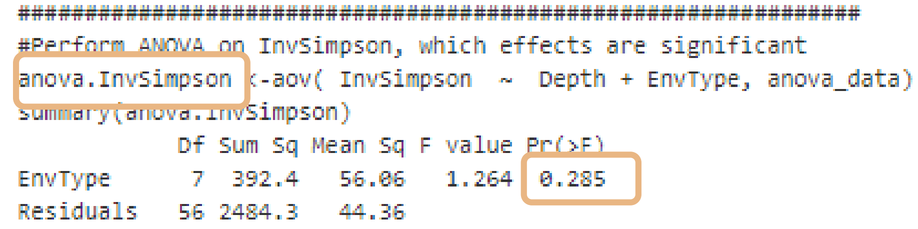
```
################################################################
#Perform ANOVA on Observed, which effects are significant
anova.Observed <-aov( Observed  ~  Depth + EnvType, anova_data)
summary(anova.Observed)
            Df Sum Sq Mean Sq F value   Pr(>F)    
EnvType      7  57320    8189   7.731 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               
```

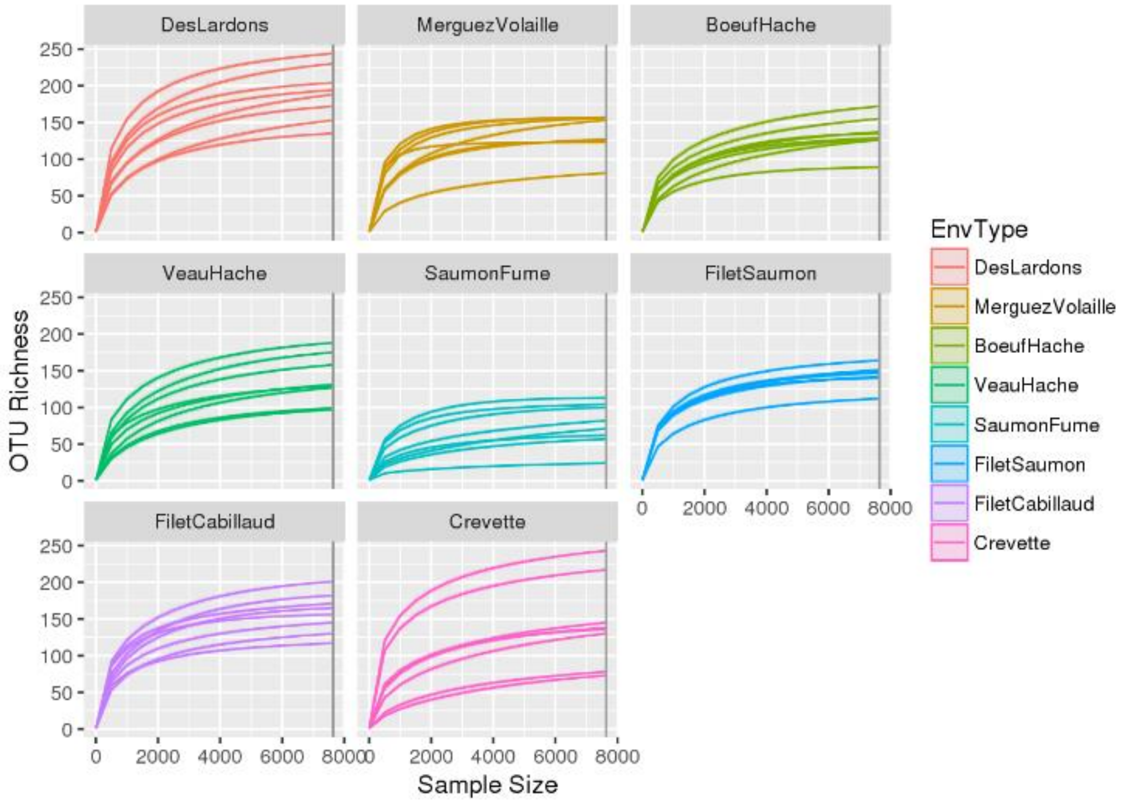


```
################################################################
#Perform ANOVA on InvSimpson, which effects are significant
anova.InvSimpson <-aov( InvSimpson  ~  Depth + EnvType, anova_data)
summary(anova.InvSimpson)
            Df Sum Sq Mean Sq F value Pr(>F)
EnvType      7  392.4   56.06   1.264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

Exercise 5

Richness plot | Richness plot with boxplot | Alpha Diversity Indice Anova Analysis | **Rarefaction curves**

Rarefaction curve interpretations

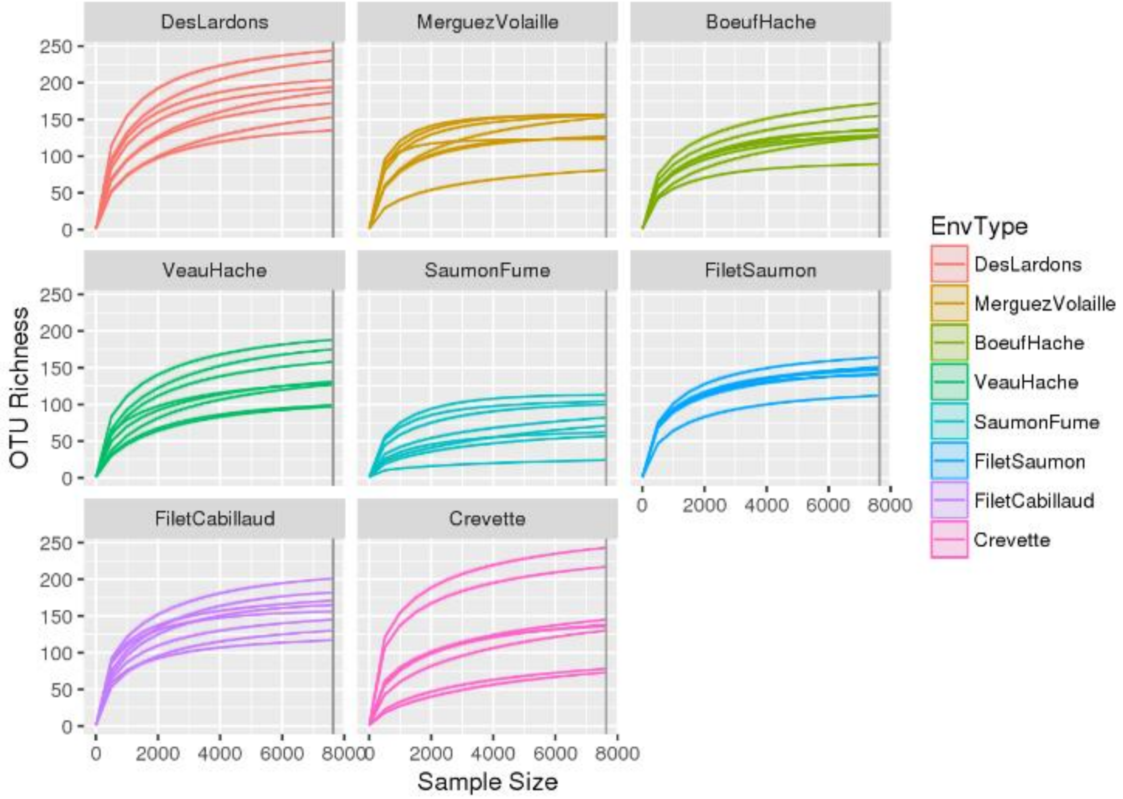


Exercise 5

Richness plot | Richness plot with boxplot | Alpha Diversity Indice Anova Analysis | Rarefaction curves

Rarefaction curve interpretations

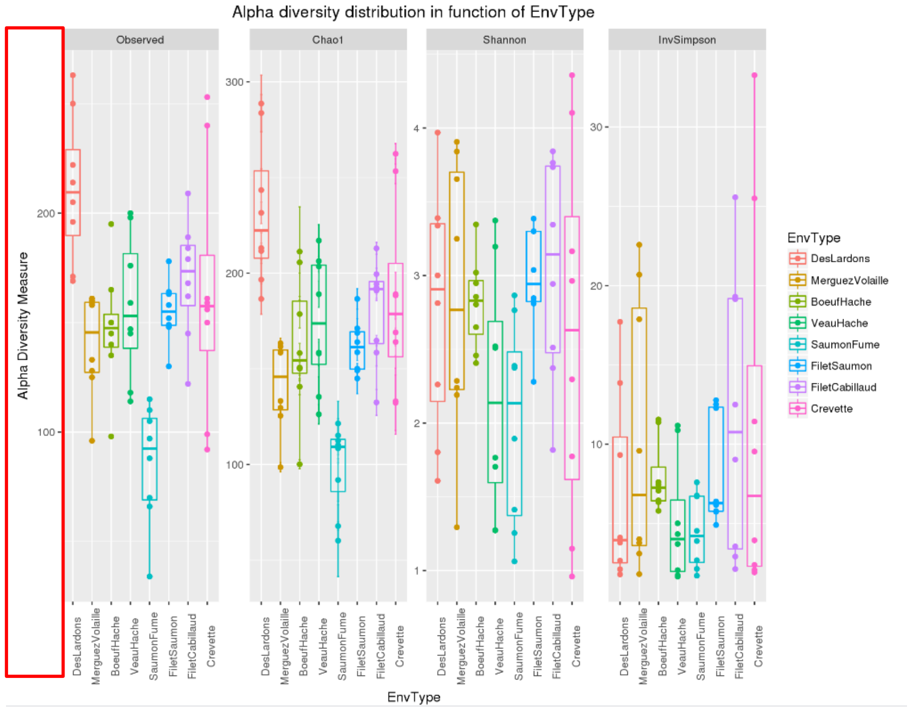
- Most of the curves reach a plateau
- A deeper sequencing doesn't add more OTU
- It confirms the Chao index
- DesLardons reach the plateau later which correspond to a higher Chao

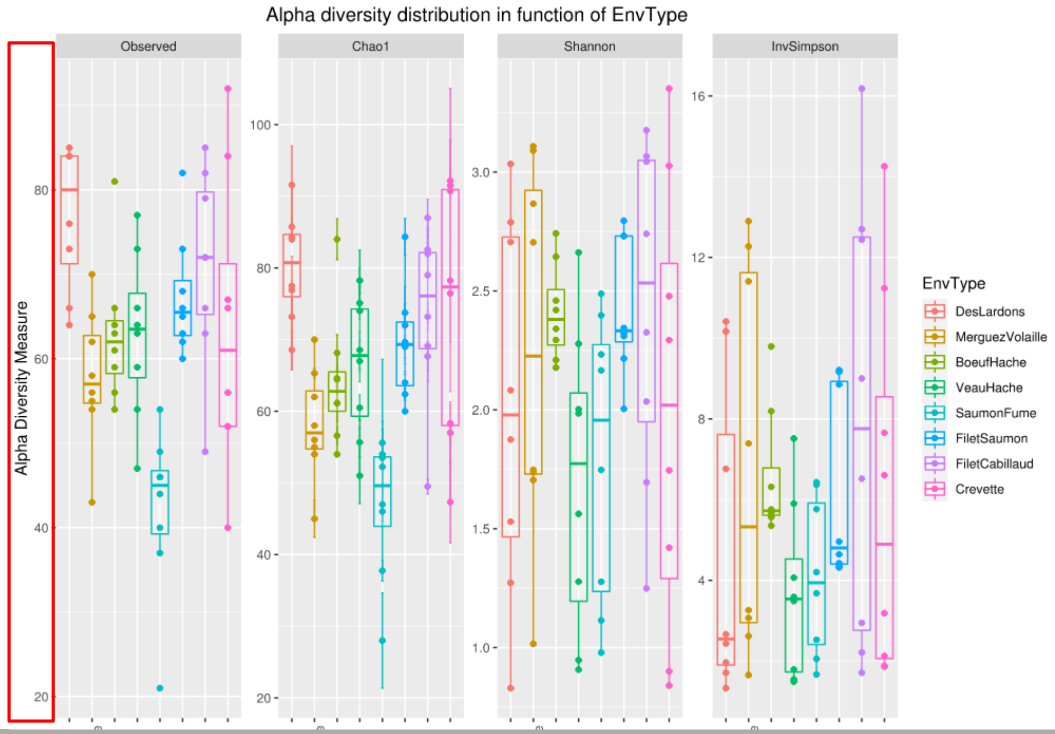


Exploring biodiversity : α-diversity

WARNING : Many diversity indices (richness, Chao) depend a lot on rare OTUs. Do not trim rare OTUs before computing them as it can drastically alter the result.

α-diversity: without (left) and with (right) trimming on rare OTU (total abundance < 500)





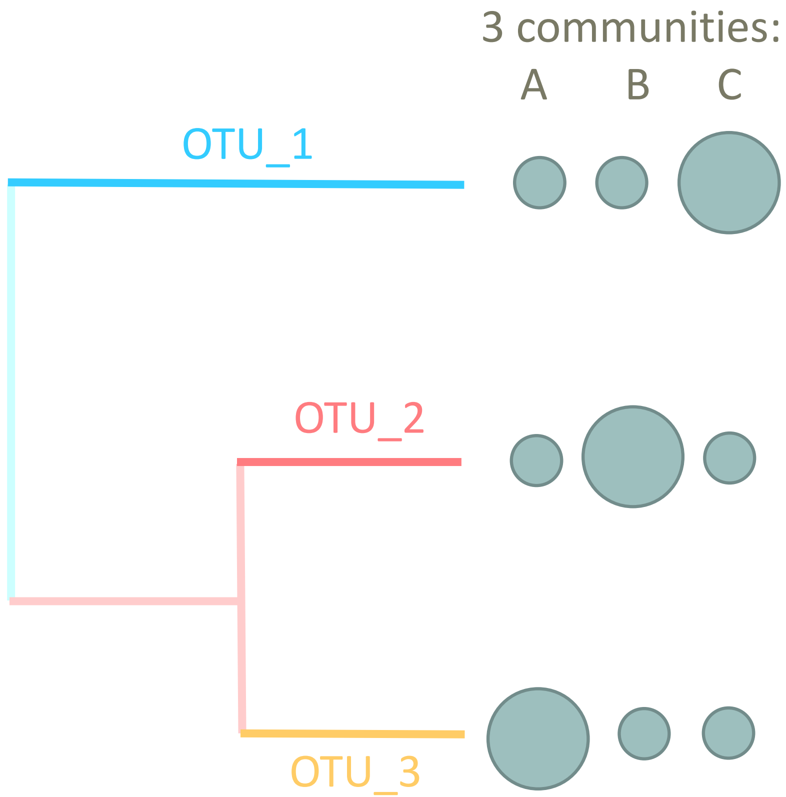
IV. Biodiversity analysis

β-DIVERSITY INDICES

Exploring biodiversity : β-diversity

Many diversity indices (both compositional and phylogenetic) are available with the Phyloseq package through the generic distance function.

Different dissimilarities capture different features of the communities.



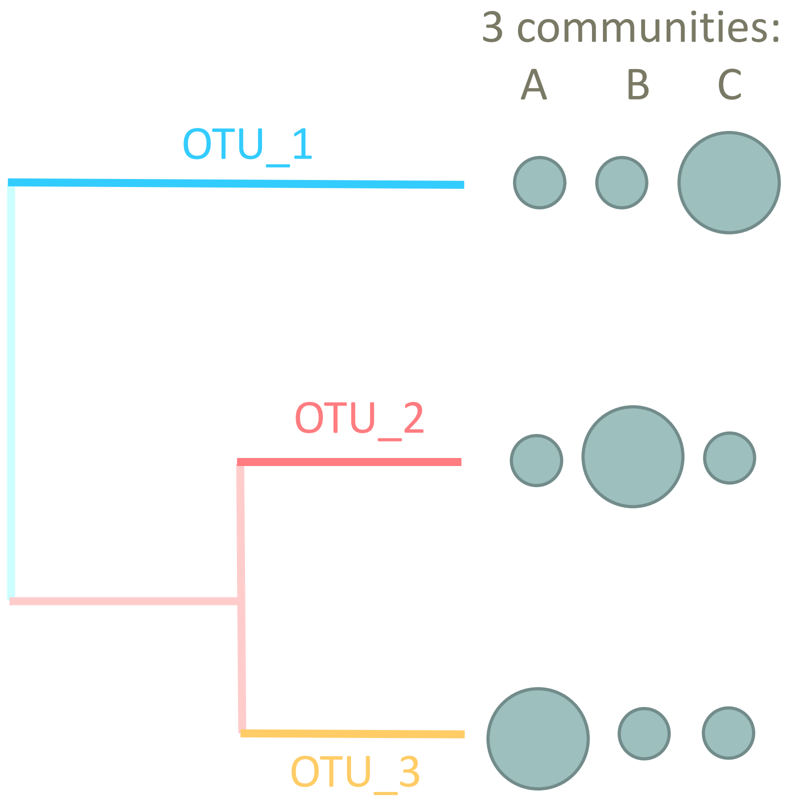
Exploring biodiversity : β-diversity

Many diversity indices (both compositional and phylogenetic) are available with the Phyloseq package through the generic distance function.

Different dissimilarities capture different features of the communities.

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 : β-diversity

Jaccard:

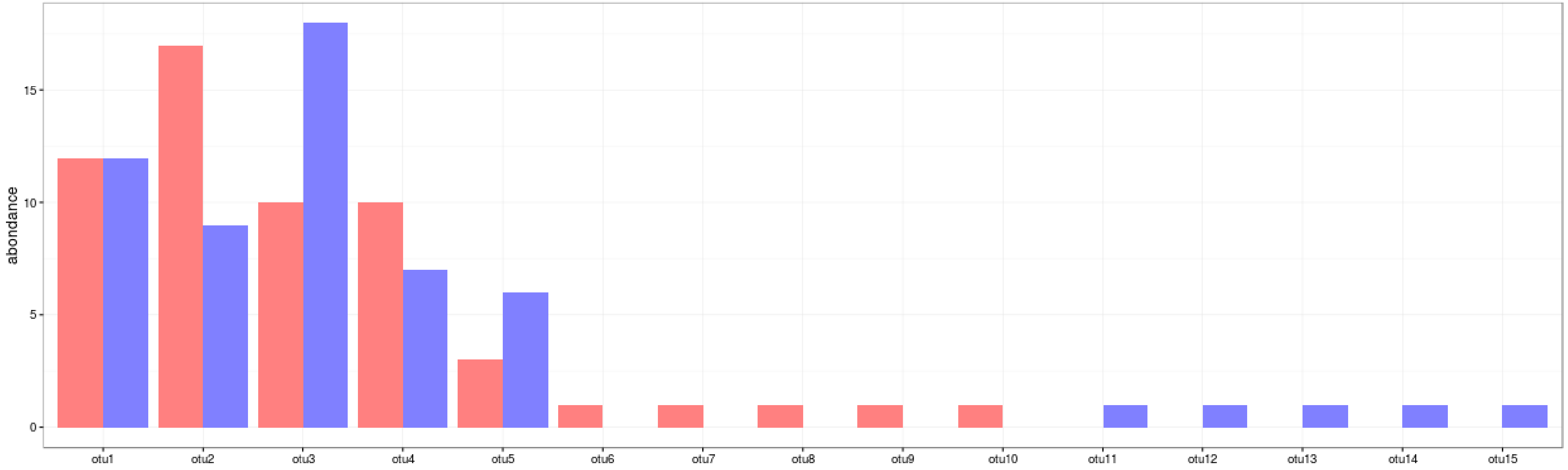
- Fraction of species specific to either 1 or 2

Bray-Curtis:

- Fraction of the community specific to either 1 or 2

Exploring biodiversity : β-diversity

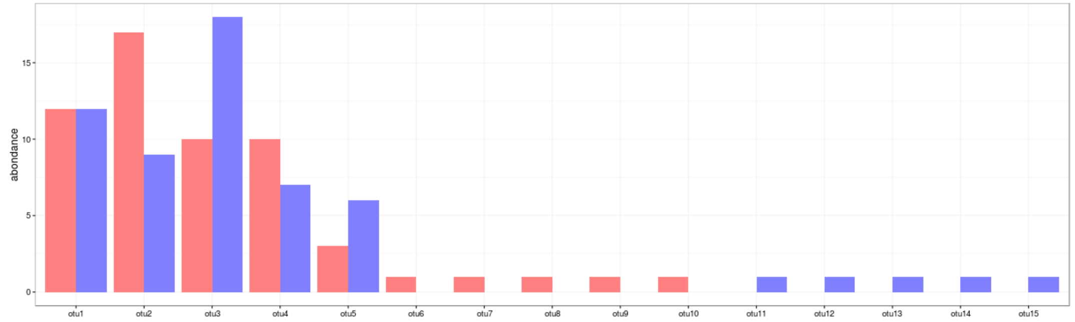
- 2 communities
- 15 OTUs



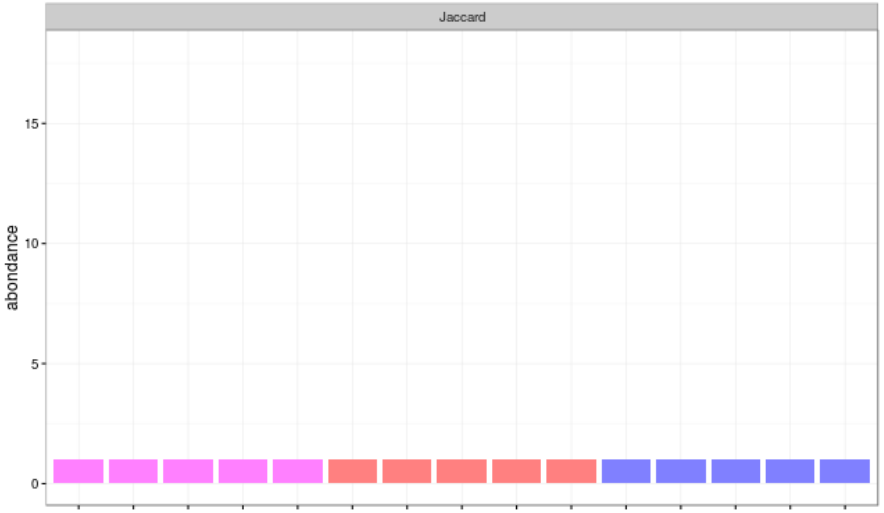
Exploring biodiversity : β-diversity

Jaccard:

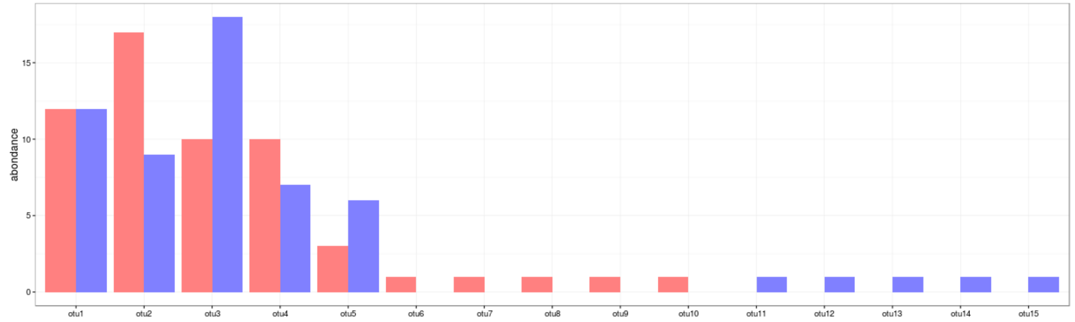
- Fraction of <u>species </u>specific to either 1 or 2



D_{jac} = 10/15 = 0.667

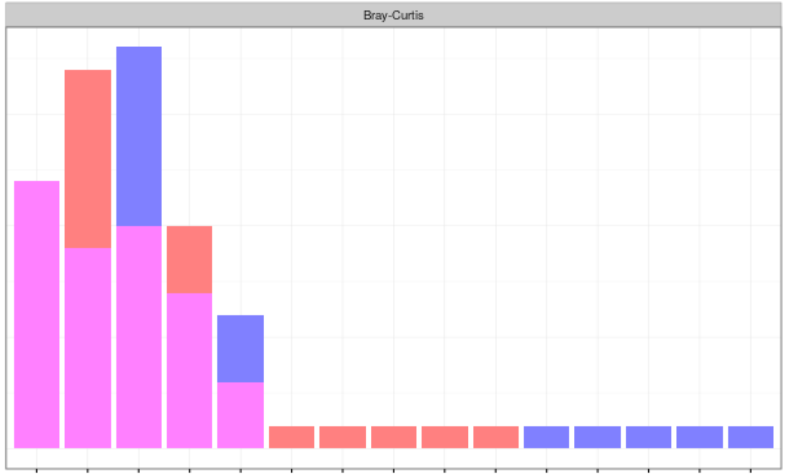


Exploring biodiversity : β-diversity



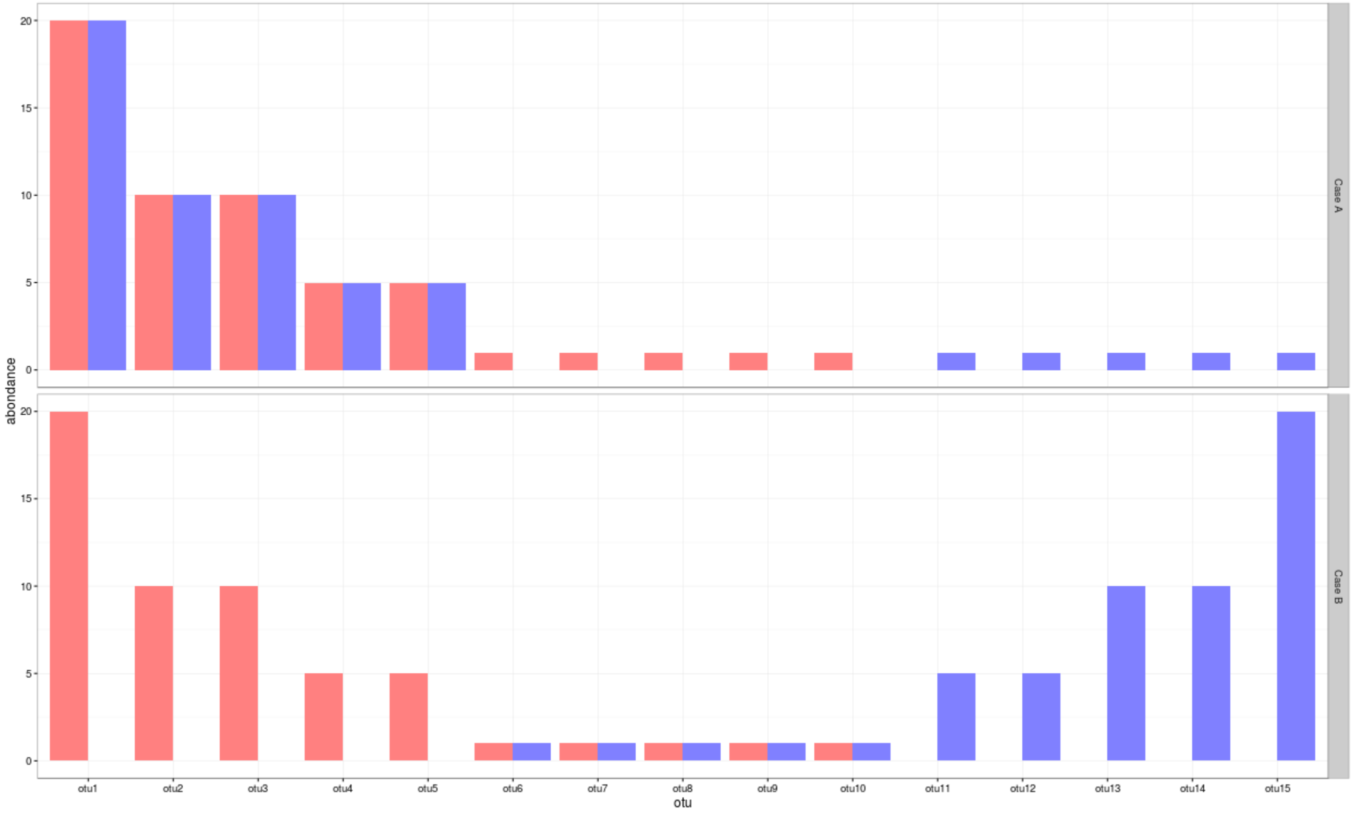
Bray-Curtis:

- Fraction of the community specific to either 1 or 2

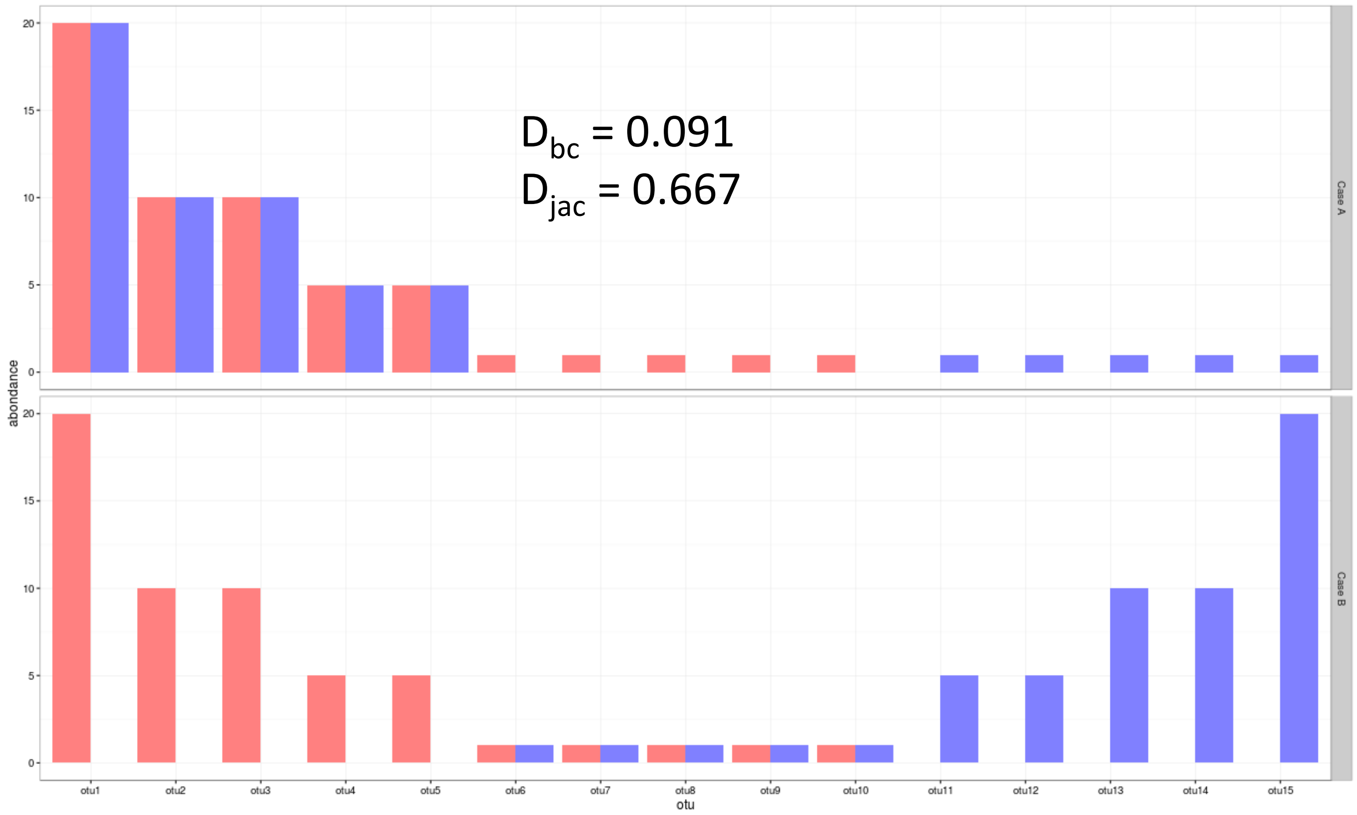


D_{bc} = (8+8+3+3+10) / (24+26+28+17+9+10) = 0.281

Exploring biodiversity : β-diversity



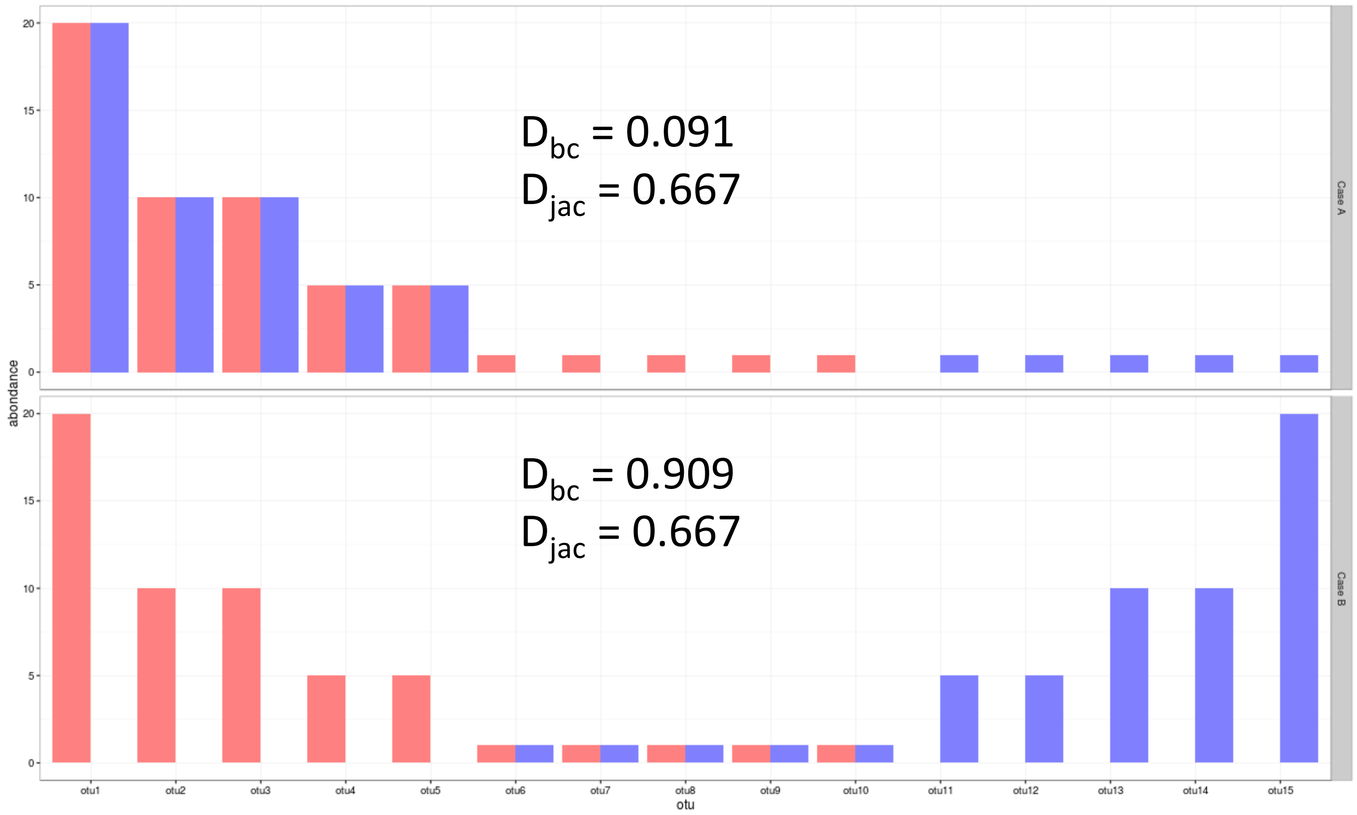
Exploring biodiversity : β-diversity



$D_{bc} = 0.091$

$D_{jac} = 0.667$

Exploring biodiversity : β-diversity



$D_{bc} = 0.091$
$D_{jac} = 0.667$

$D_{bc} = 0.909$
$D_{jac} = 0.667$

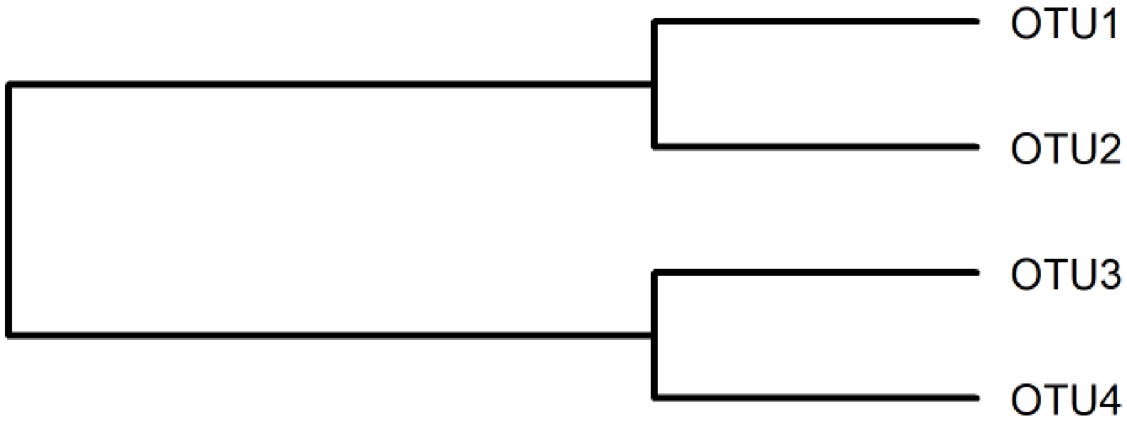
Exploring biodiversity : β-diversity

Unifrac:

- Fraction of the tree specific to either 1 or 2

Weigthed-Unifrac :

- Fraction of the diversity specific to either 1 or 2

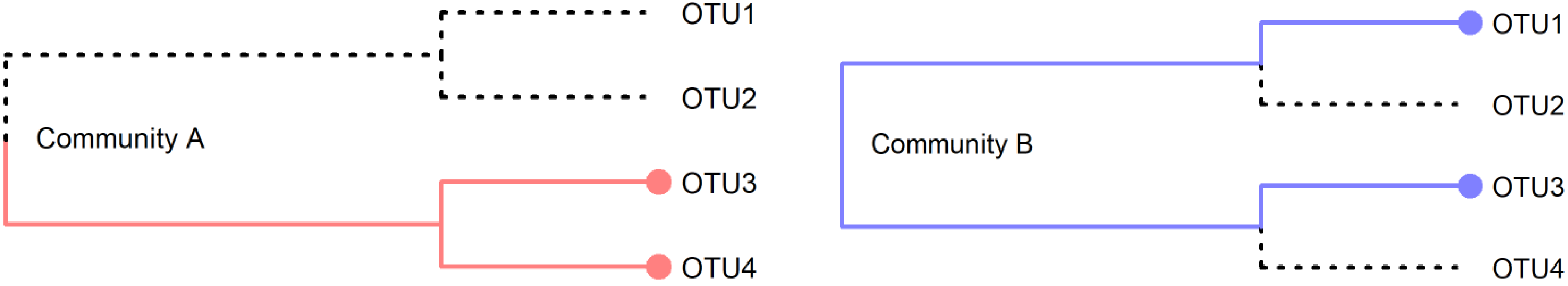


Exploring biodiversity : β-diversity

Unifrac:

- Fraction of the tree specific to either 1 or 2

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



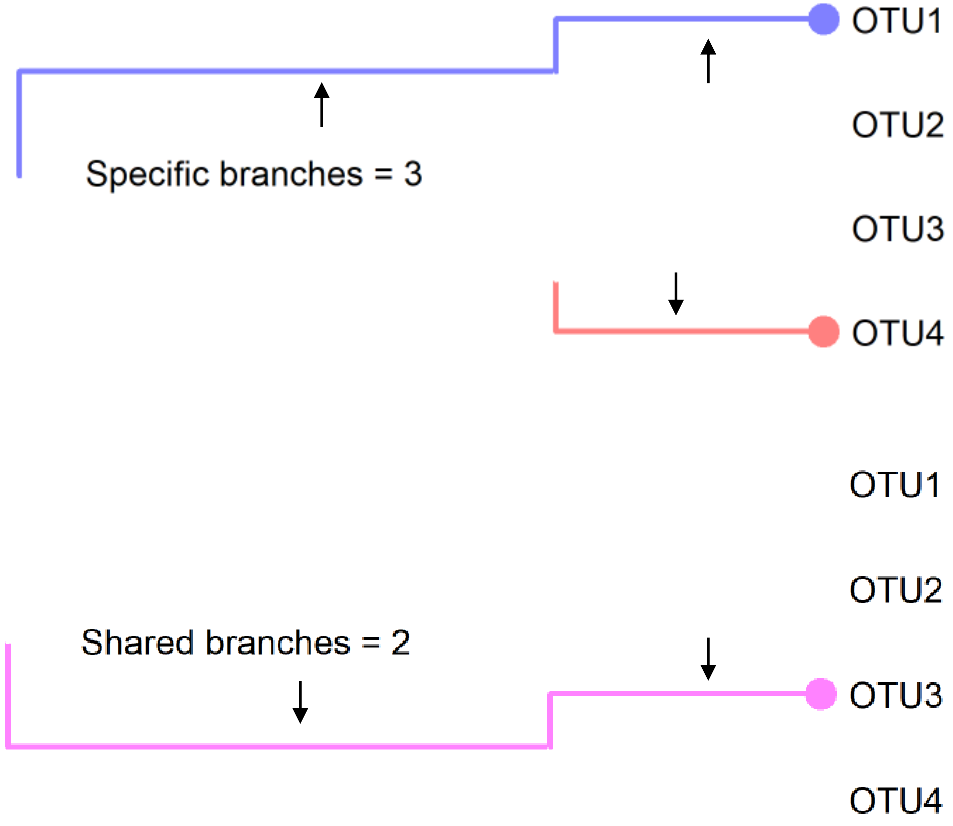
Exploring biodiversity : β-diversity

Unifrac:

- Fraction of the tree specific to either 1 or 2

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

$$Unifrac = \frac{\sum specific\_branch\_length}{\sum all\_branch\_length} = 0.6$$

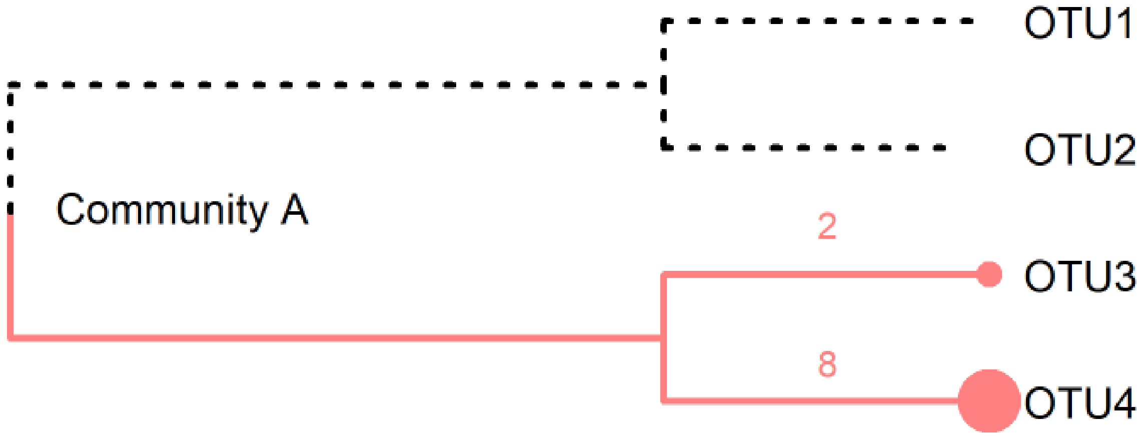


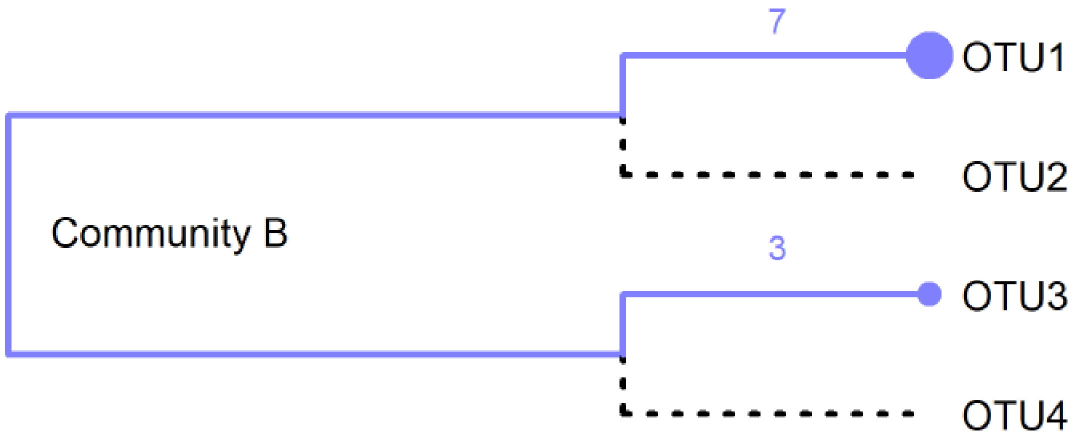
Exploring biodiversity : β-diversity

Weigthed-Unifrac :

- Fraction of the diversity specific to either 1 or 2

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



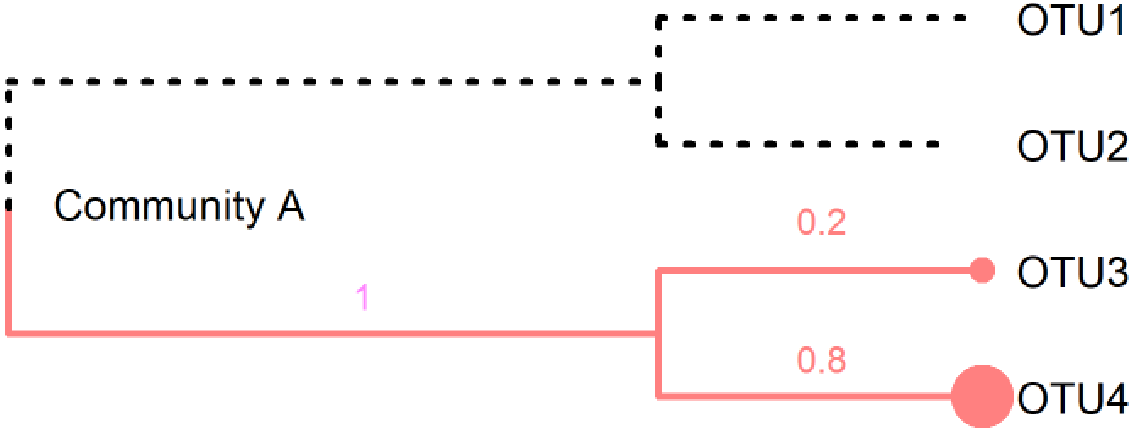


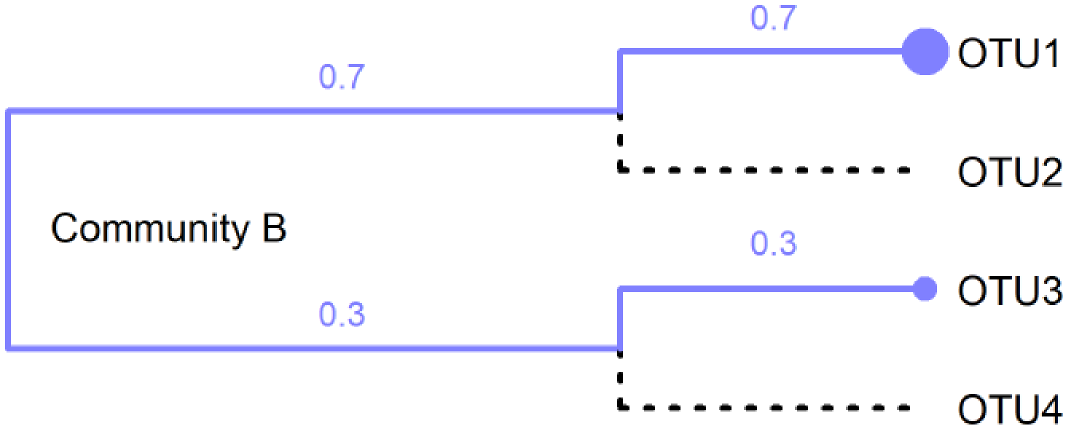
Exploring biodiversity : β-diversity

Weigthed-Unifrac :

- Fraction of the diversity specific to either 1 or 2

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



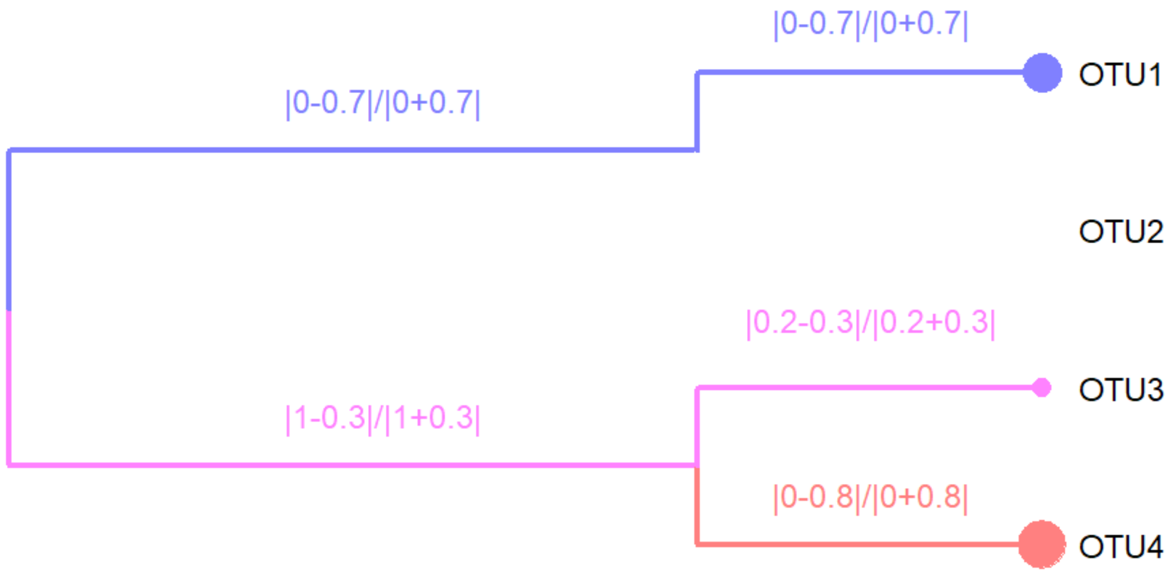


Exploring biodiversity : β-diversity

Weigthed-Unifrac :

- Fraction of the diversity specific to either 1 or 2

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



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

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

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

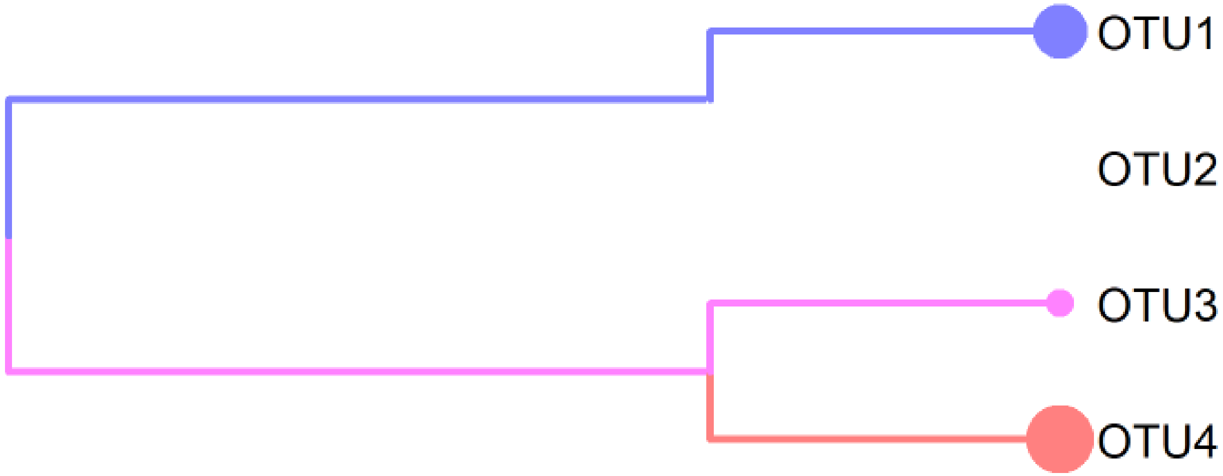
$$\sum reduced\ branch\ length = 3{,}73$$

Exploring biodiversity : β-diversity

Weigthed-Unifrac :

- Fraction of the diversity specific to either 1 or 2

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

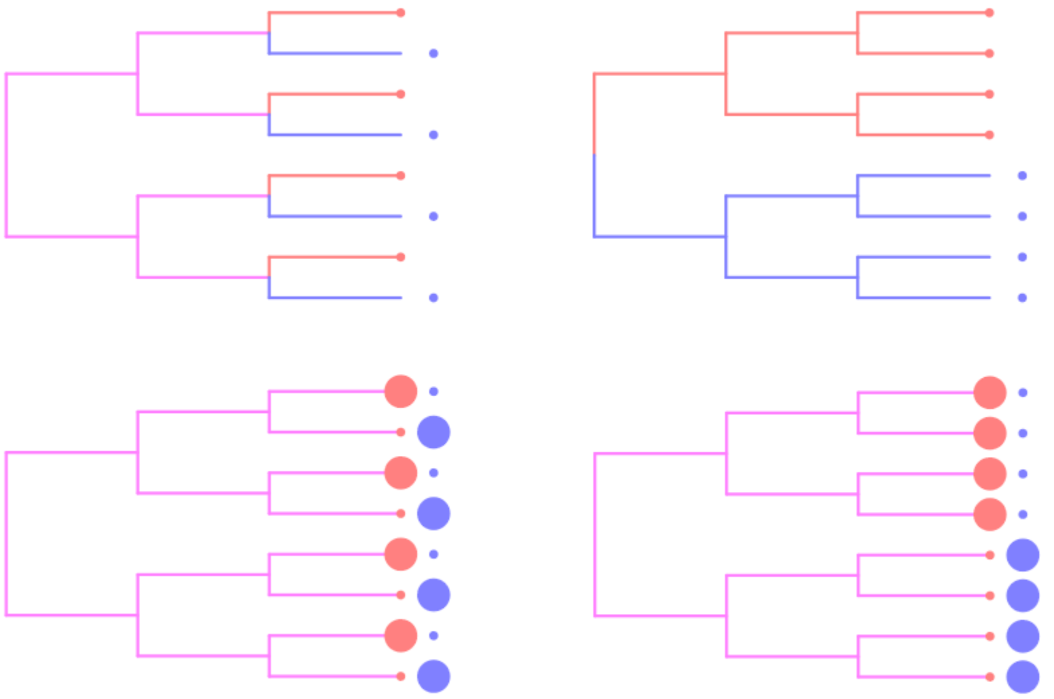


$$\sum non\ reduced\ branch\ length\ = 5$$

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

Exploring biodiversity : β-diversity

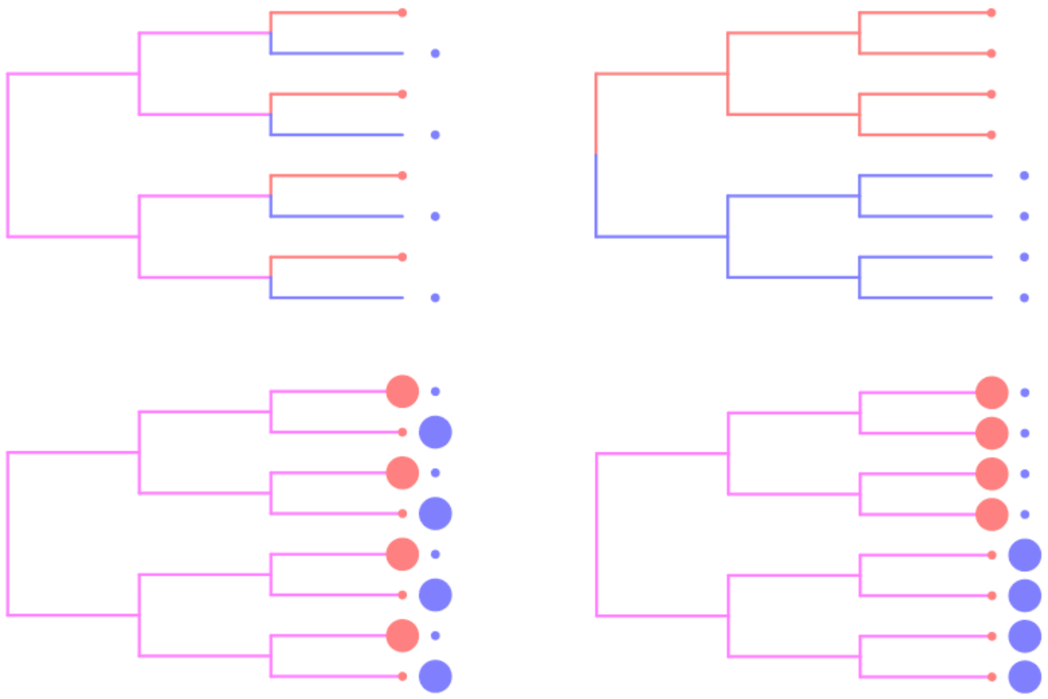
➔ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weigthed Unifrac values?



Exploring biodiversity : β-diversity

➔ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weigthed Unifrac values?

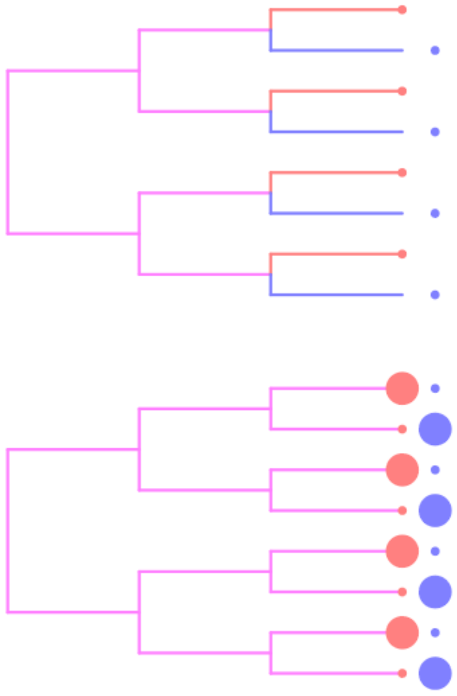
Low Unifrac / High Jaccard



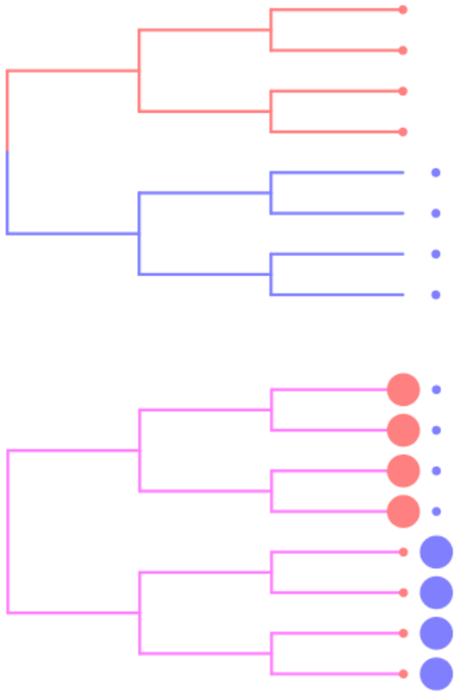
Exploring biodiversity : β-diversity

➔ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weigthed Unifrac values?

Low Unifrac / High Jaccard



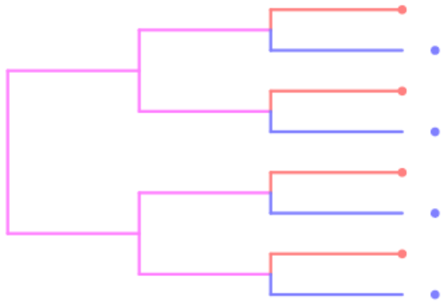
High Unifrac / High Jaccard



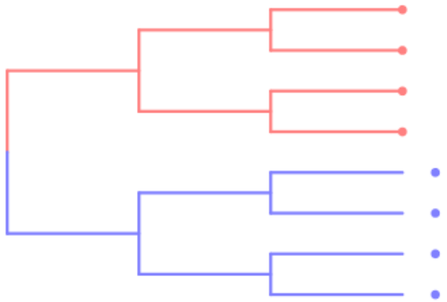
Exploring biodiversity : β-diversity

➔ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weigthed Unifrac values?

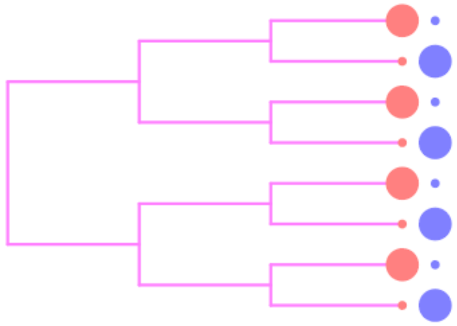
Low Unifrac / High Jaccard

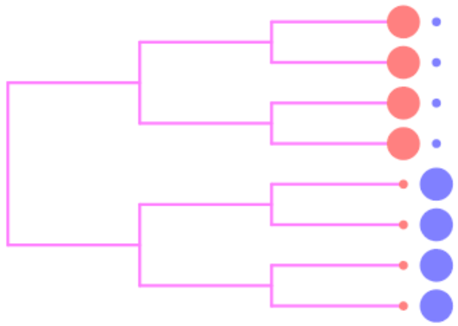


High Unifrac / High Jaccard



Low wUnifrac / High Bray Curtis

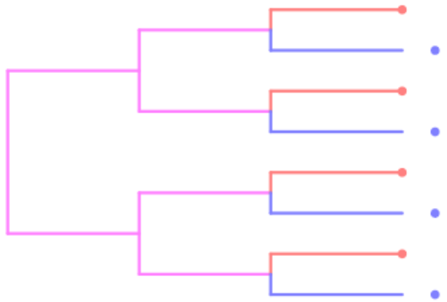




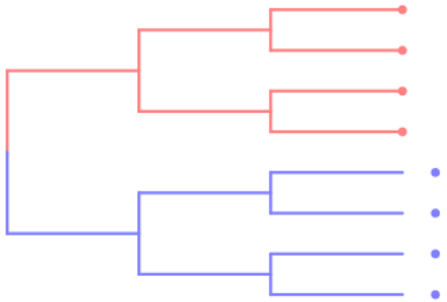
Exploring biodiversity : β-diversity

➔ What do you conclude in terms of Jaccard, Bray Curtis, Unifrac and weigthed Unifrac values?

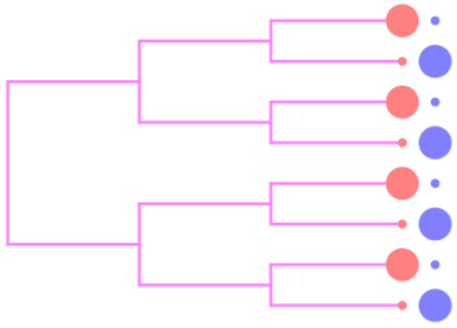
Low Unifrac / High Jaccard



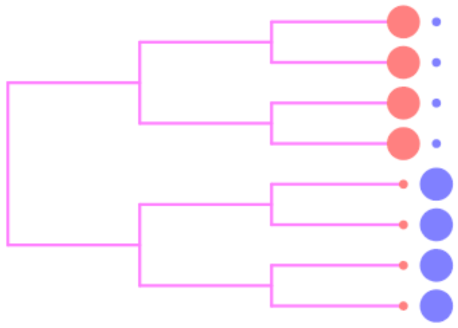
High Unifrac / High Jaccard



Low wUnifrac / High Bray Curtis



High wUnifrac / High Bray Curtis



Exploring biodiversity : β-diversity

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

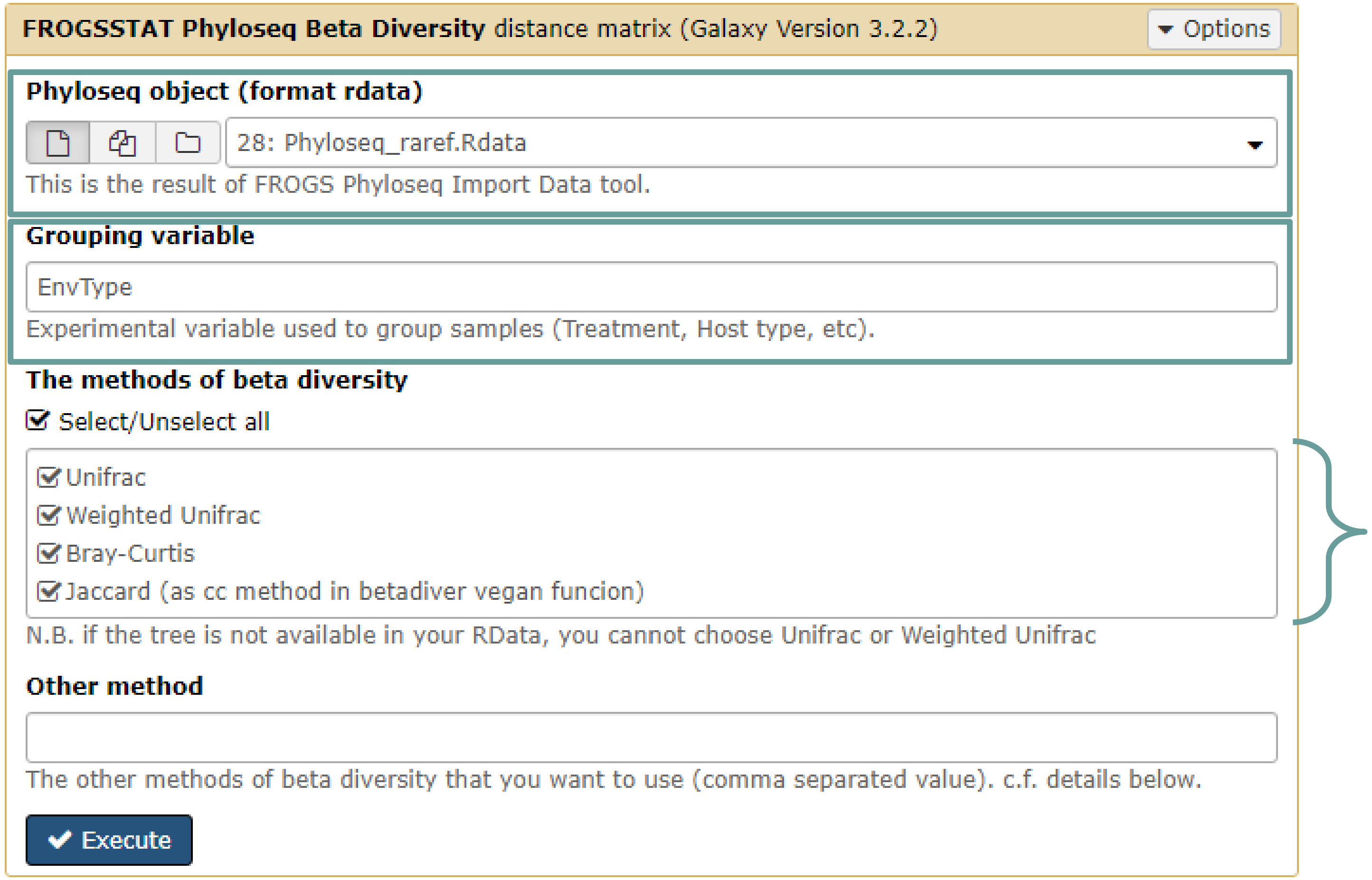
unifrac, wunifrac,

dpcoa, jsd, manhattan, euclidean, canberra,

bray, kulczynski, jaccard, gower, altGower, morisita, horn, mountford, raup, binomial

chao, cao…

Exploring biodiversity : β-diversity



FROGSSTAT Phyloseq Beta Diversity distance matrix (Galaxy Version 3.2.2)

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

Exercise 6

Try it with the 4 most commonly used distance methods

1. What are the output datasets ?
2. *A priori*, abundant OTU are they shared among samples?
3. Considering that Jaccard is higher than Unifrac, what can you conclude ?
4. Considering that Unifrac is higher than weighted Unifrac, what can you conclude ?

Exercise 6

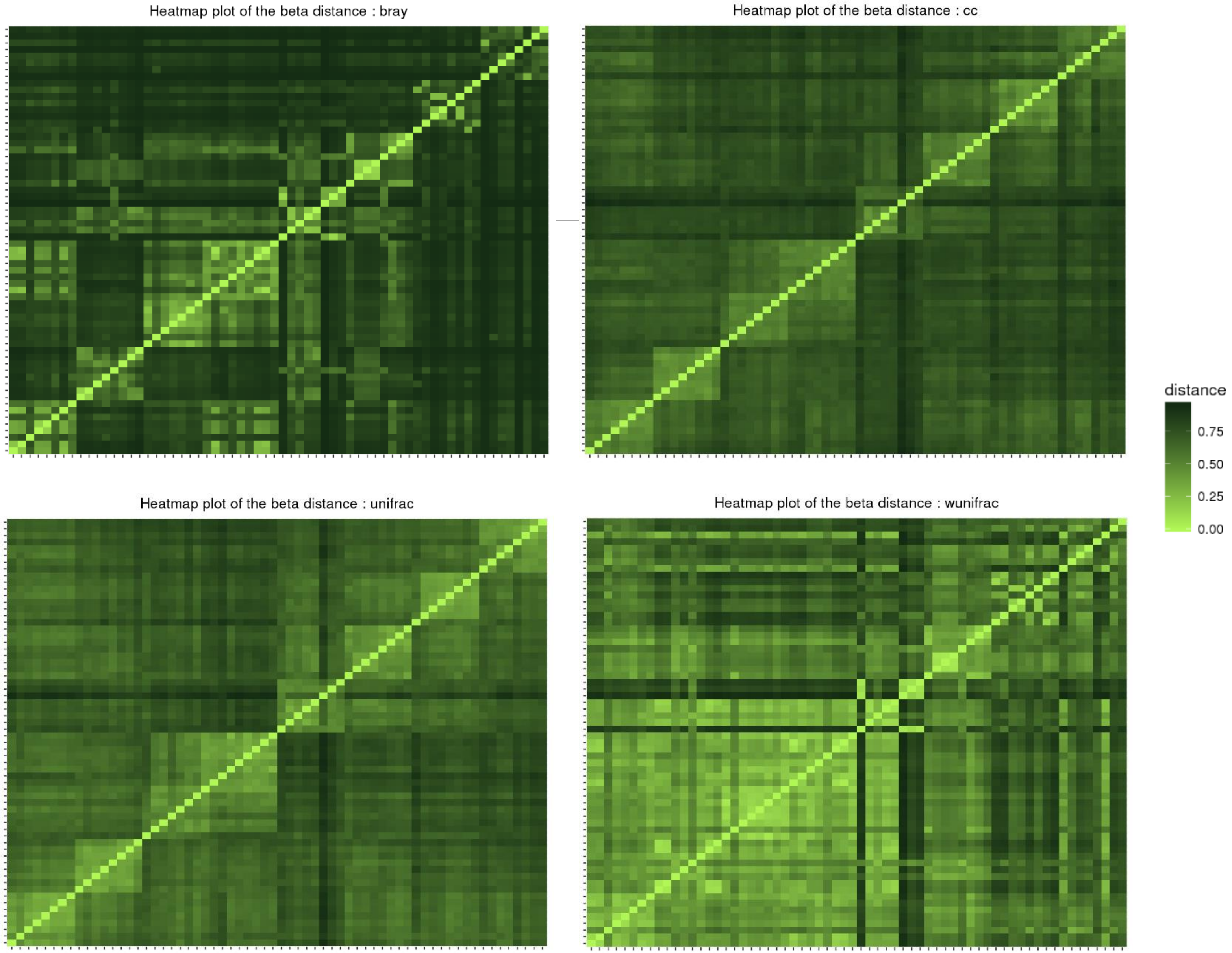
1. What are the output datasets ?

→ 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

Exercise 6

1. What are the output datasets ?



Heatmap plot of the beta distance : bray

Heatmap plot of the beta distance : cc

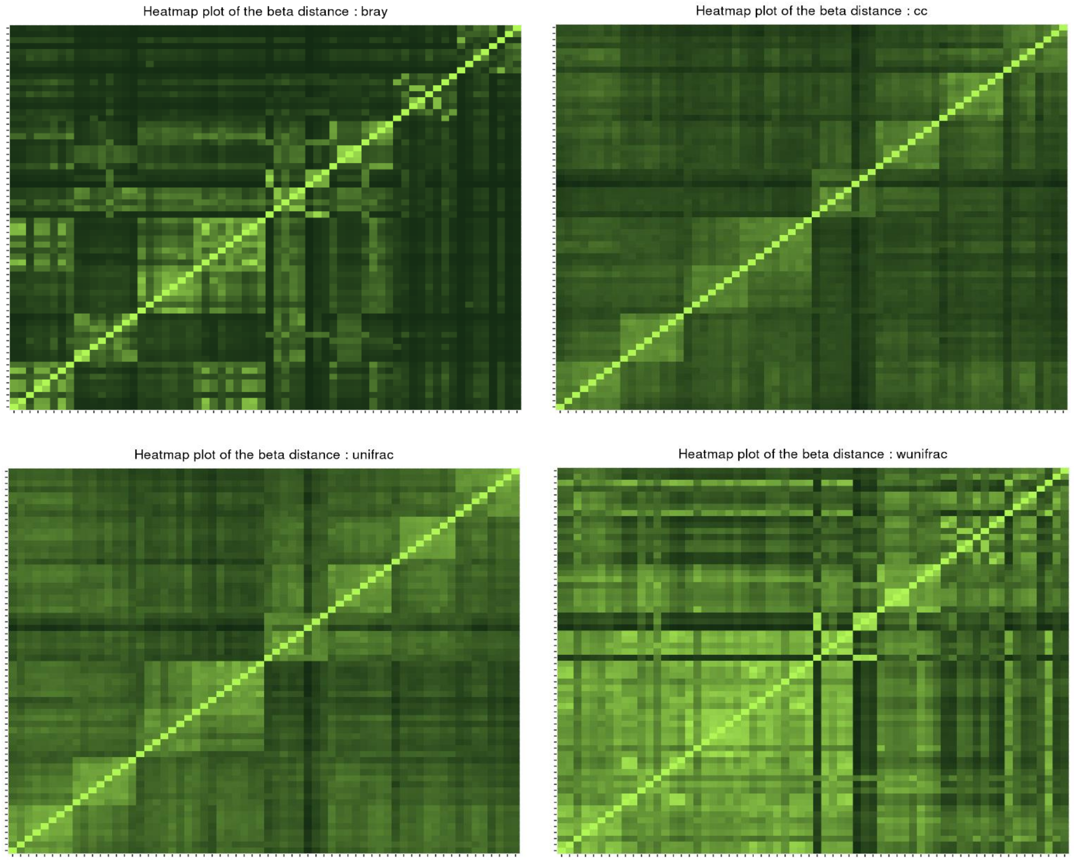
Heatmap plot of the beta distance : unifrac

Heatmap plot of the beta distance : wunifrac

Exercise 6

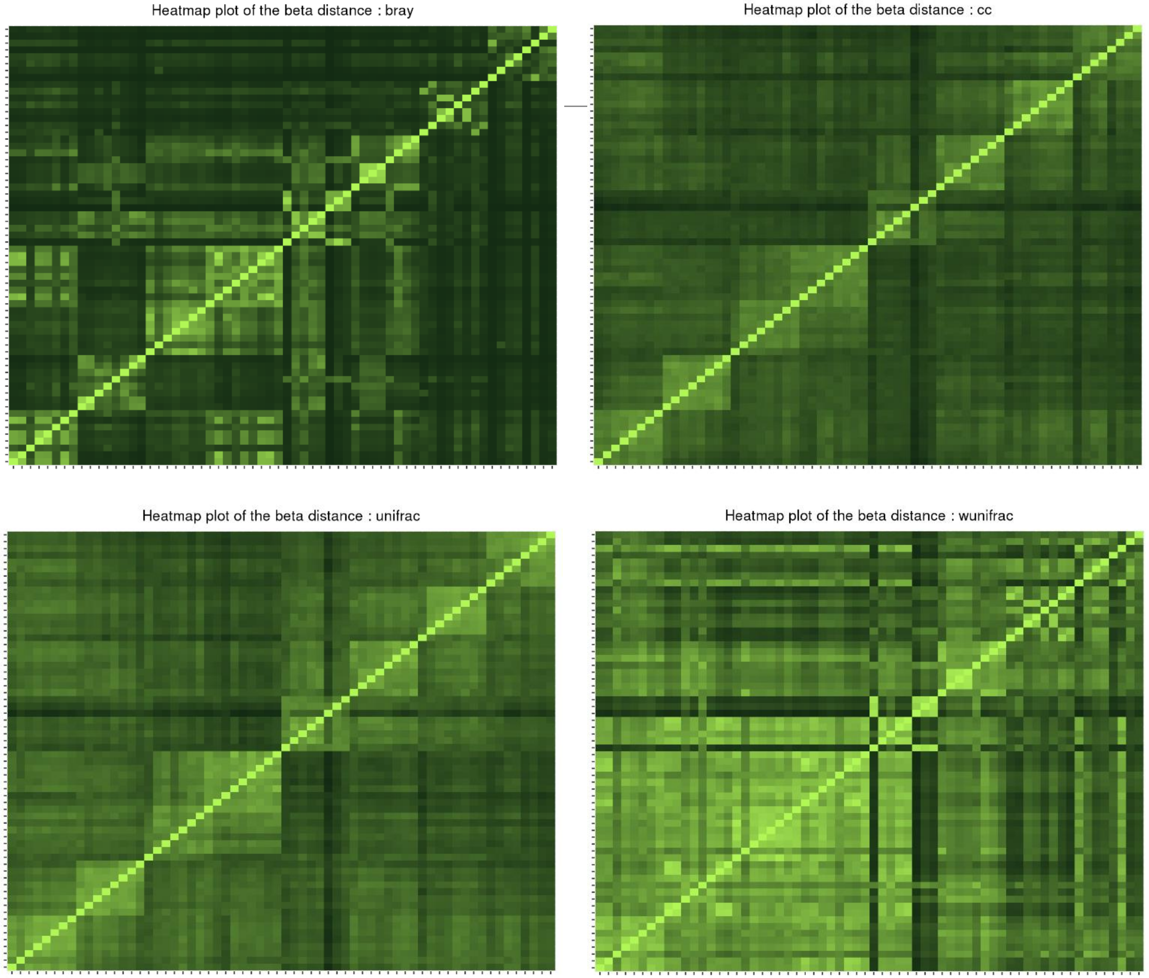
1. What are the output datasets ?

- 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.



Exercise 6

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



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

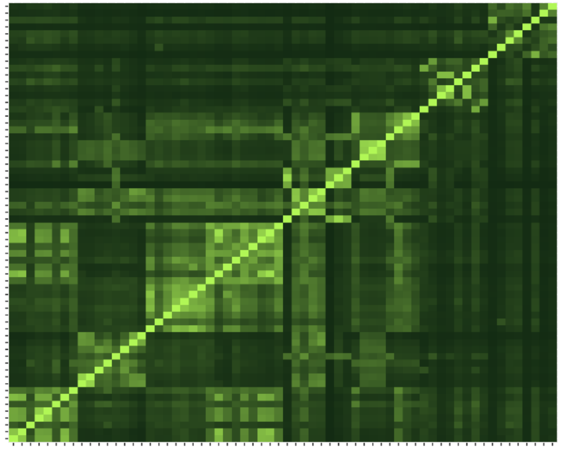
Exercise 6

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

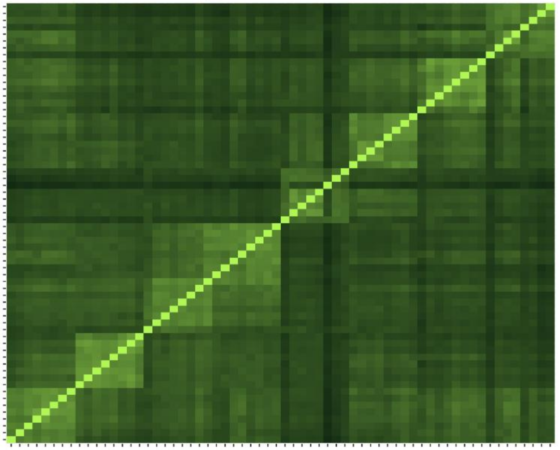
- Jaccard lower than Bray-Curtis

➔abundant taxa are not shared

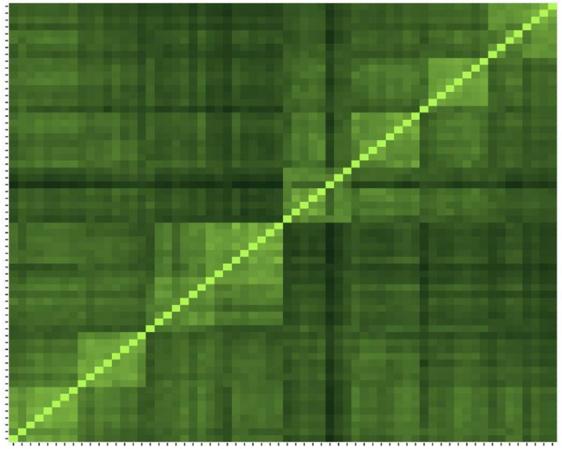
Heatmap plot of the beta distance : bray



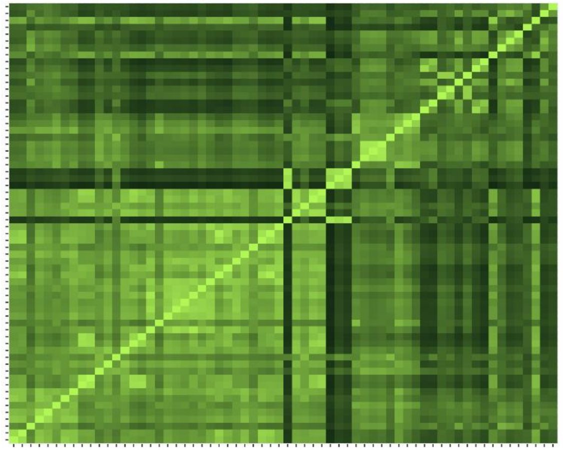
Heatmap plot of the beta distance : cc



Heatmap plot of the beta distance : unifrac

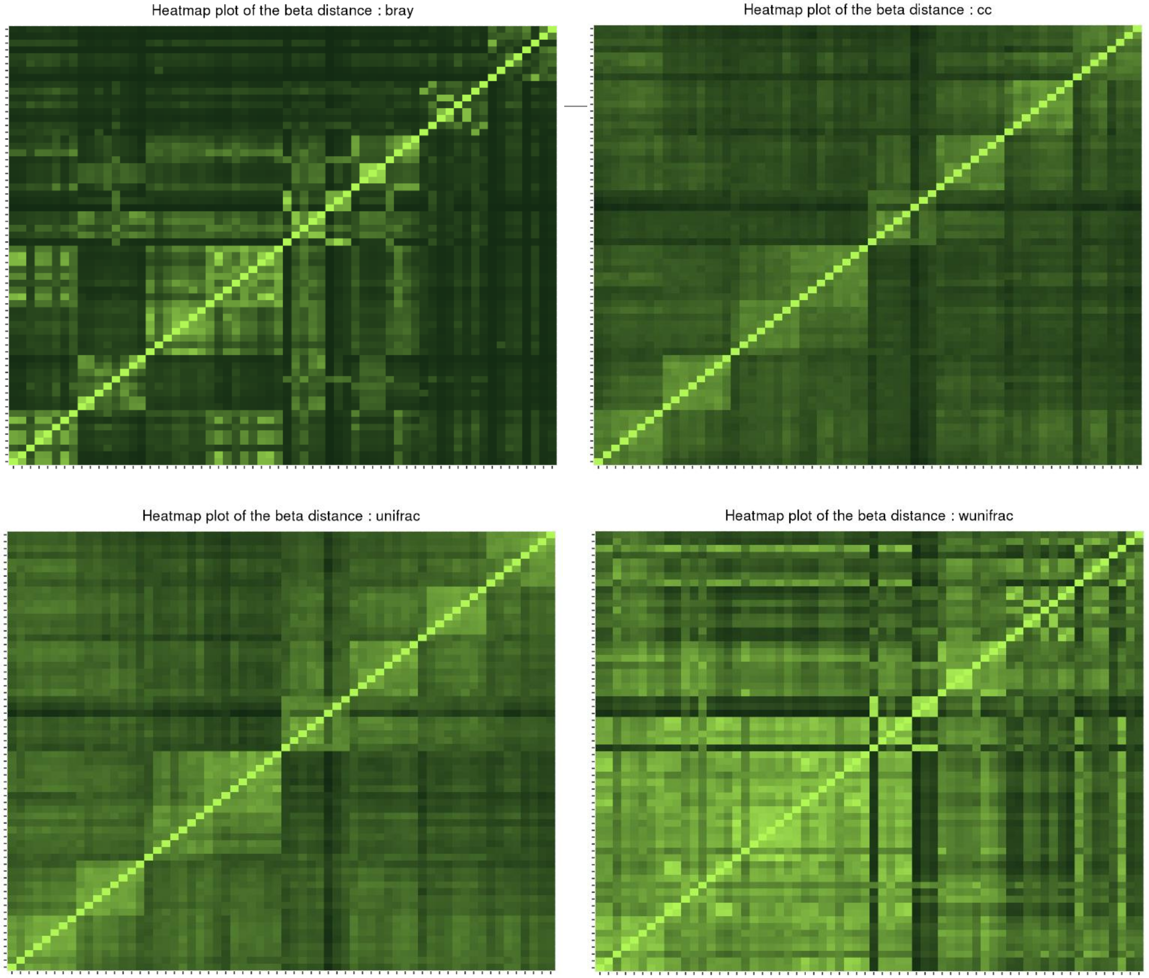


Heatmap plot of the beta distance : wunifrac



Exercise 6

3. Considering that Jaccard is higher than Unifrac, what can you conclude ?



Heatmap plot of the beta distance : bray

Heatmap plot of the beta distance : cc

Heatmap plot of the beta distance : unifrac

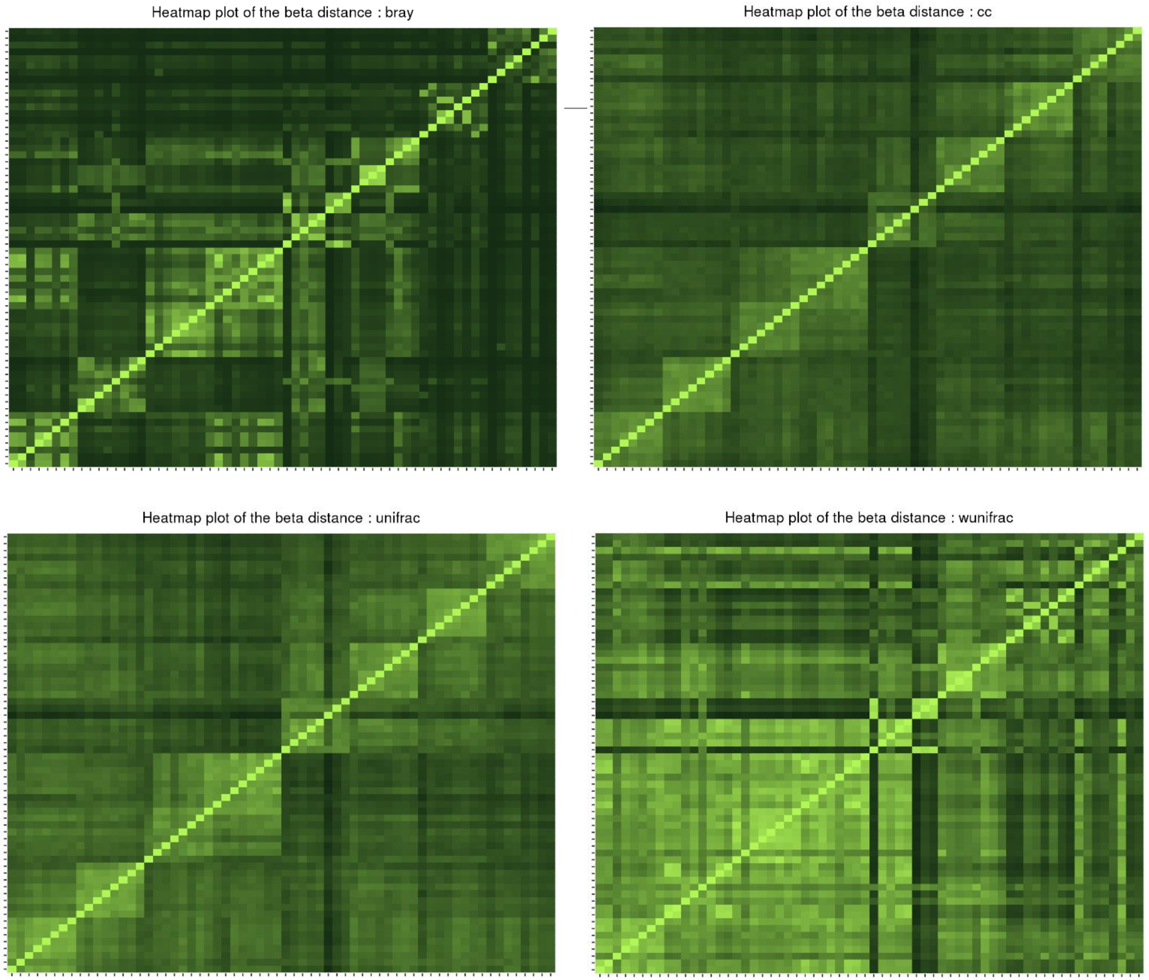
Heatmap plot of the beta distance : wunifrac

Exercise 6

3. Considering that Jaccard is higher than Unifrac, what can you conclude ?

- Jaccard higher than Unifrac

➔ communities' taxa are distinct but phylogenetically related



Heatmap plot of the beta distance : bray

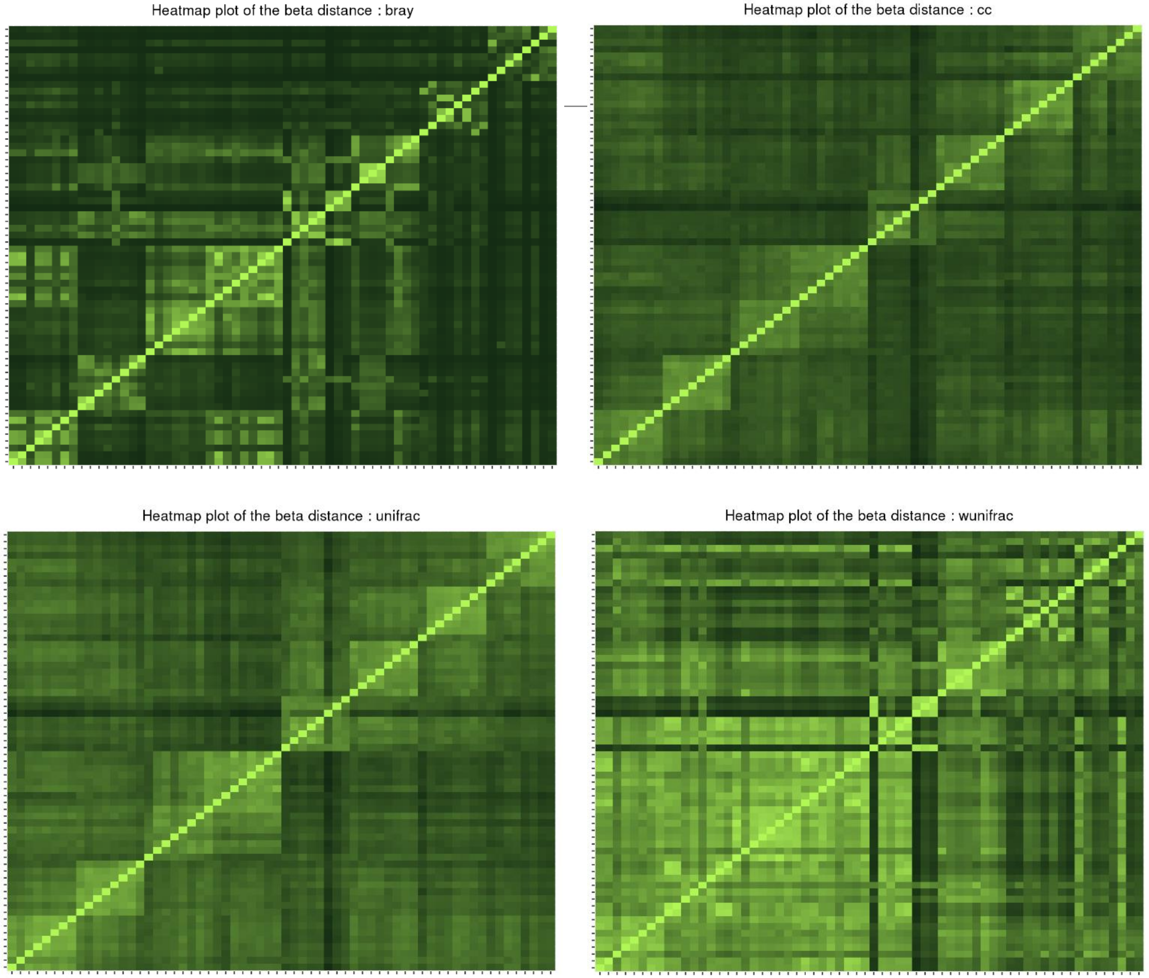
Heatmap plot of the beta distance : cc

Heatmap plot of the beta distance : unifrac

Heatmap plot of the beta distance : wunifrac

Exercise 6

4. Considering that Unifrac is higher than weighted Unifrac, what can you conclude ?



Heatmap plot of the beta distance : bray

Heatmap plot of the beta distance : cc

Heatmap plot of the beta distance : unifrac

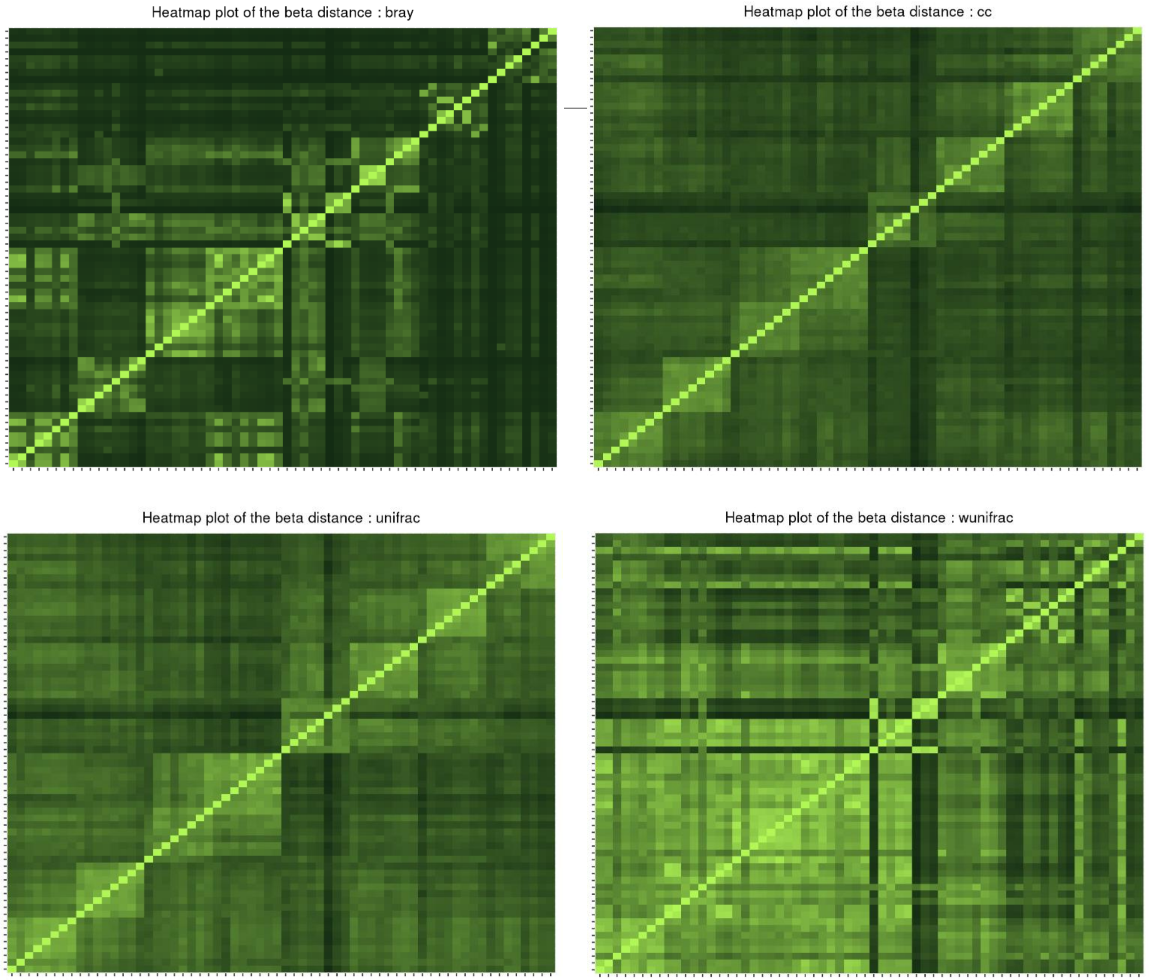
Heatmap plot of the beta distance : wunifrac

Exercise 6

4. Considering that Unifrac is higher than weighted Unifrac, what can you conclude ?

- Unifrac higher than weighted Unifrac

➔abundant taxa in both communities are phylogenetically closed.



Exploring biodiversity : β-diversity

- In general, qualitative diversities are more sensitive to factors that affect presence/absence of organisms (such as pH, salinity, depth, etc) and therefore useful to study and define bioregions (regions with little of no flow between them)...
- ... whereas quantitative distances 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

I. Exploring the structure

ORDINATION AND HEATMAP PLOTS

Exploring the structure : Ordination plot

- Each community is described by OTU abundances
- OTU abundances may be correlated
- PCA finds linear combinations of OTUs that
  - are uncorrelated
  - capture well the variance of community composition

But variance is not a very good measure of β-diversity

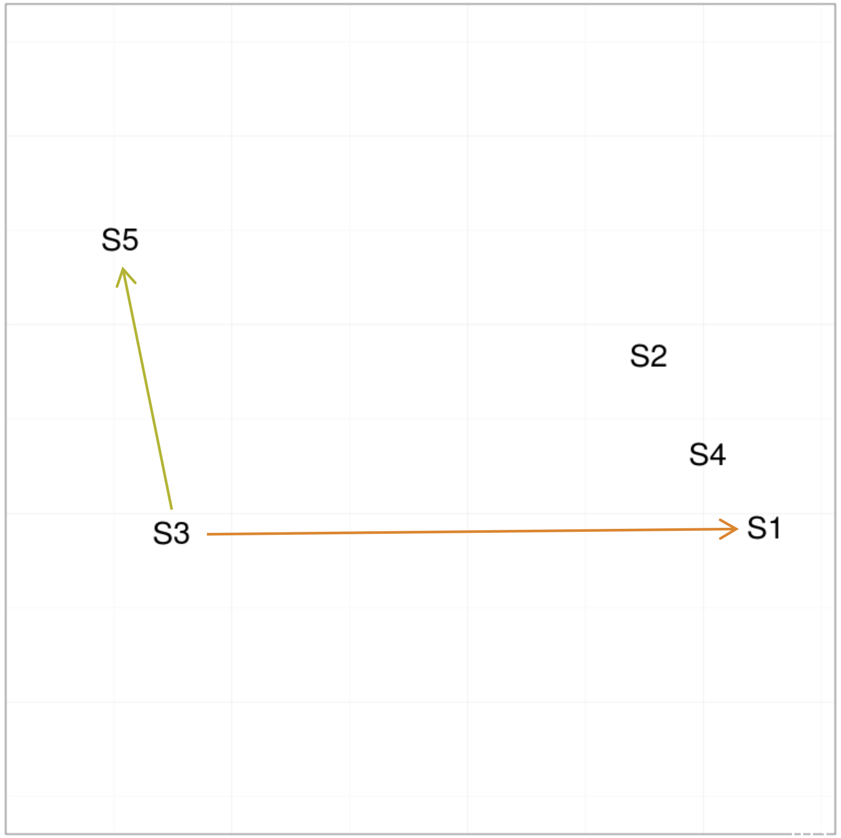
Exploring the structure : Ordination plot

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

The MDS tries to represent samples in two dimensions

➔ The samples ordination.

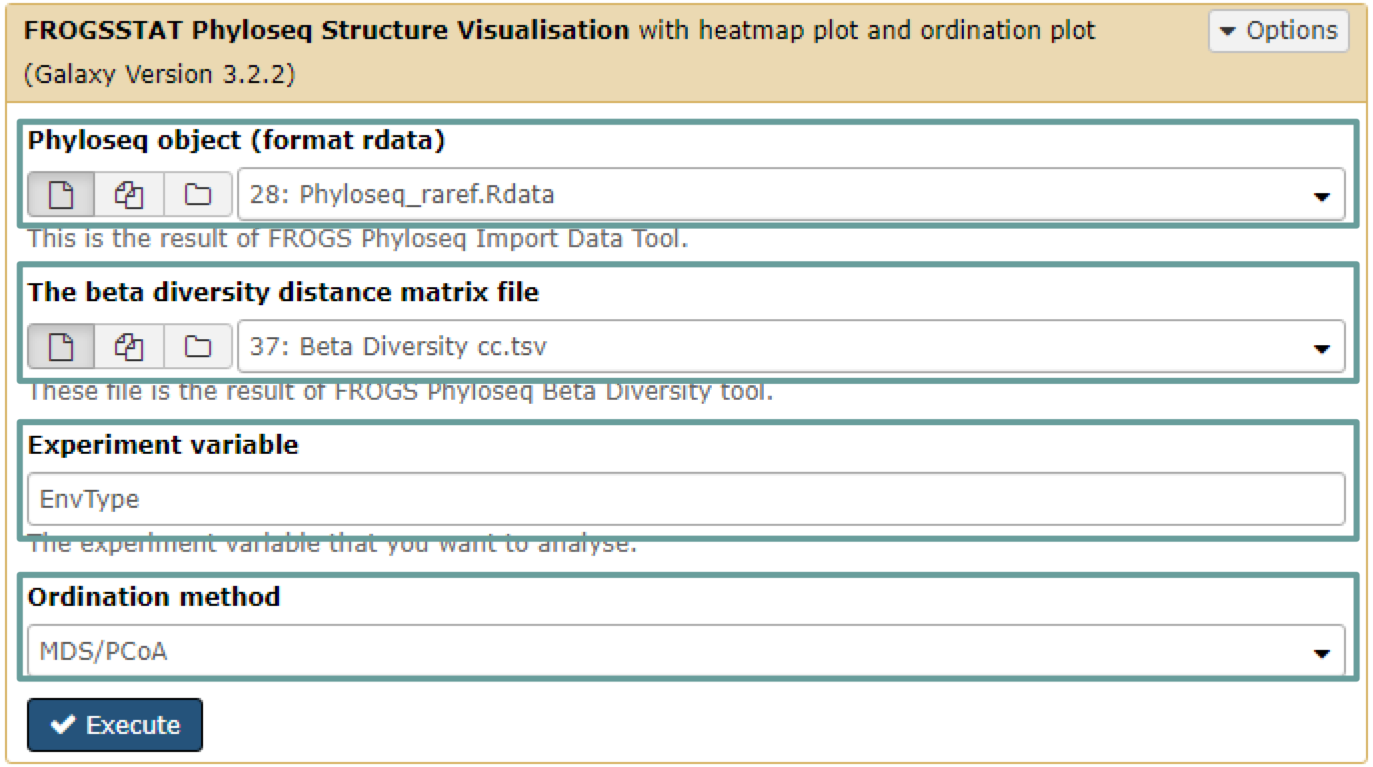
	Distance Matrix				
	S1	S2	S3	S4	S5
S1	0.00	2.21	6.31	0.99	7.50
S2	2.21	0.00	5.40	1.22	5.74
S3	6.31	5.40	0.00	5.75	3.16
S4	0.99	1.22	5.75	0.00	6.64
S5	7.50	5.74	3.16	6.64	0.00



Exploring the structure : 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.
- It
  - 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

Exploring the structure : Ordination plot and Heatmap



FROGSSTAT Phyloseq Structure Visualisation with heatmap plot and ordination plot (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

Explore the sample **NORMALISED** count

Choose the beta diversity distance matrix

Choose a sample variable to organize graphics.

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

Exploring the structure : Ordination plot and Heatmap

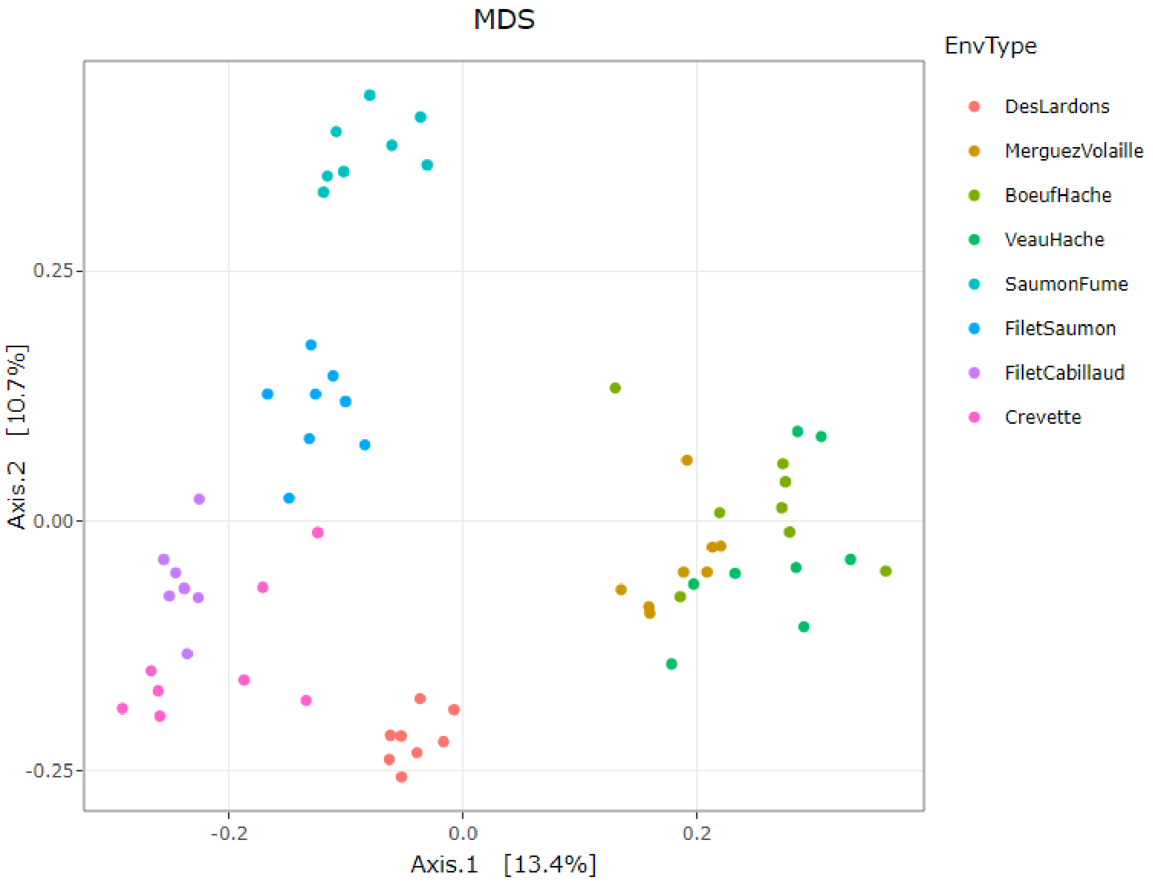
Try it with the 4 distance method matrix

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 ?

Exploring the structure : Ordination plot and Heatmap

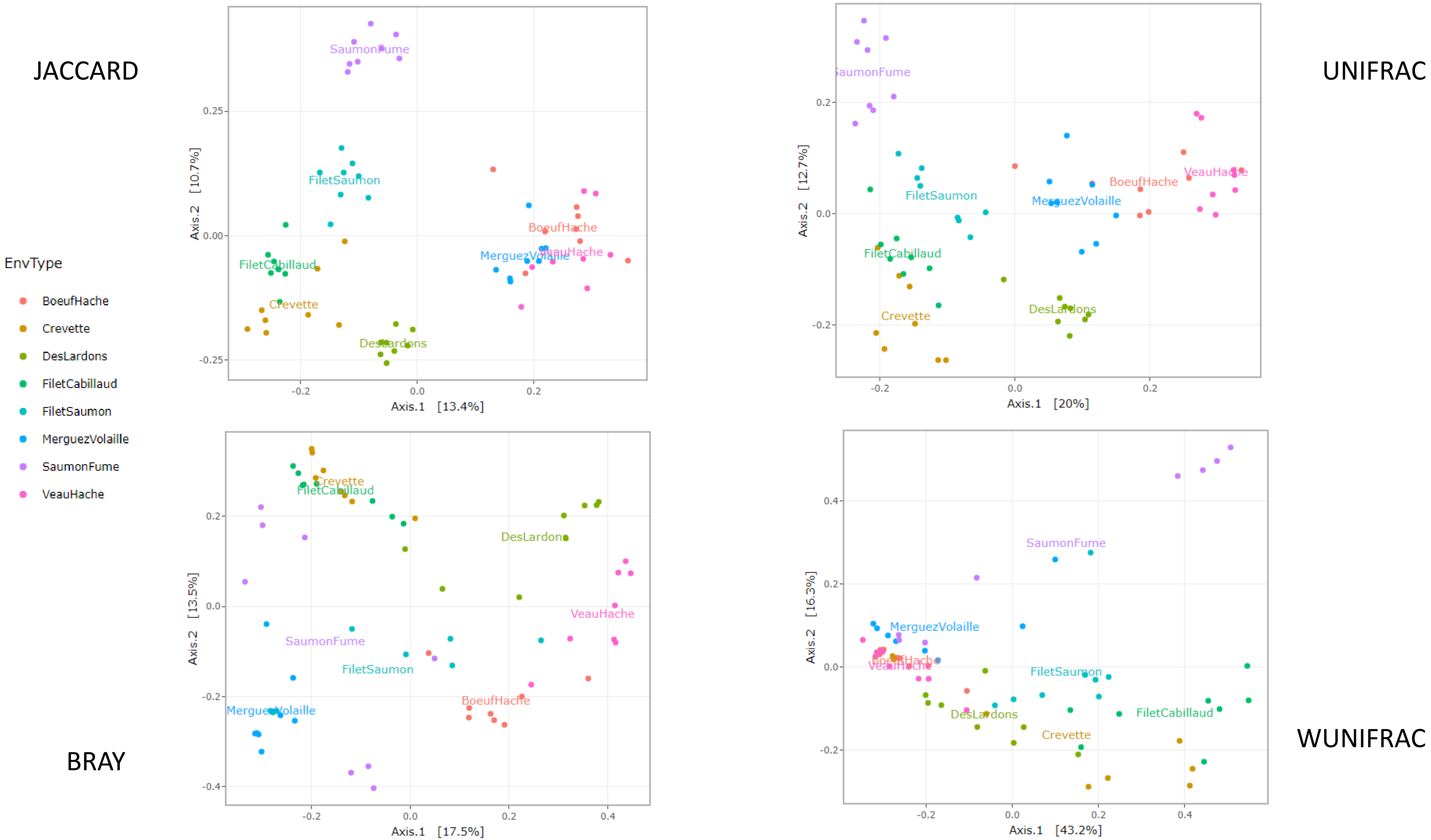
1. What are the output datasets ?

→ HTML report: ordination plot



Exploring the structure : Ordination plot and Heatmap

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



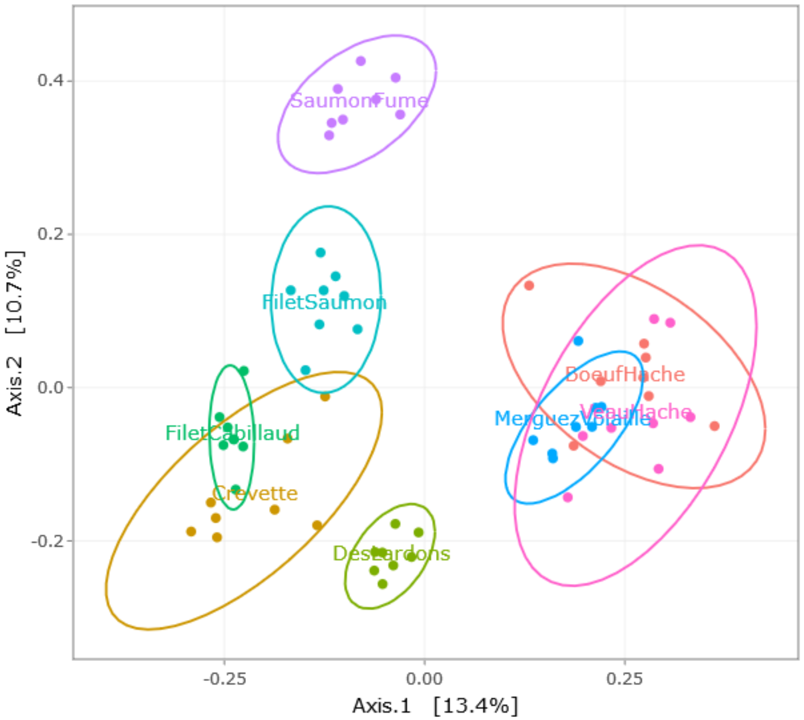
JACCARD

UNIFRAC

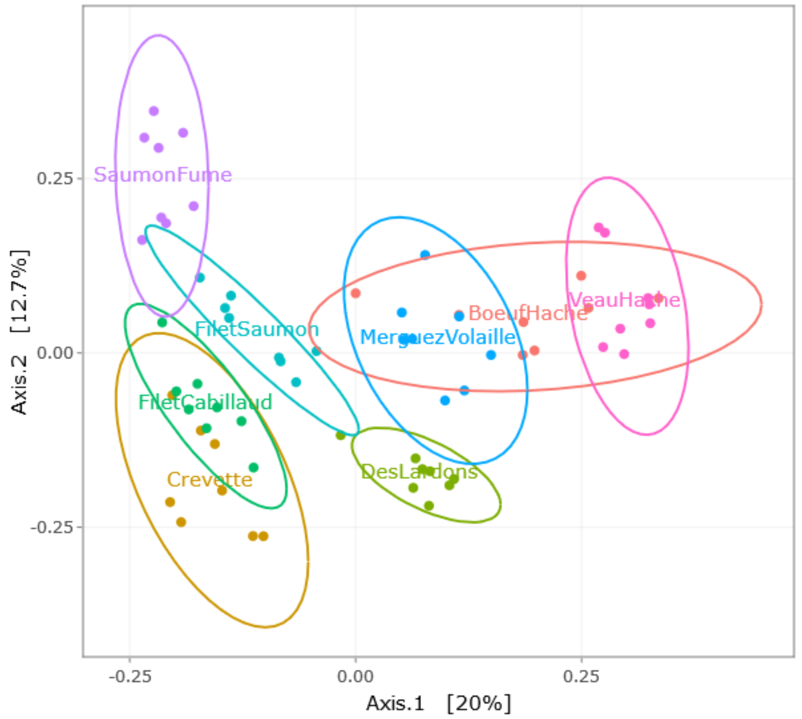
BRAY

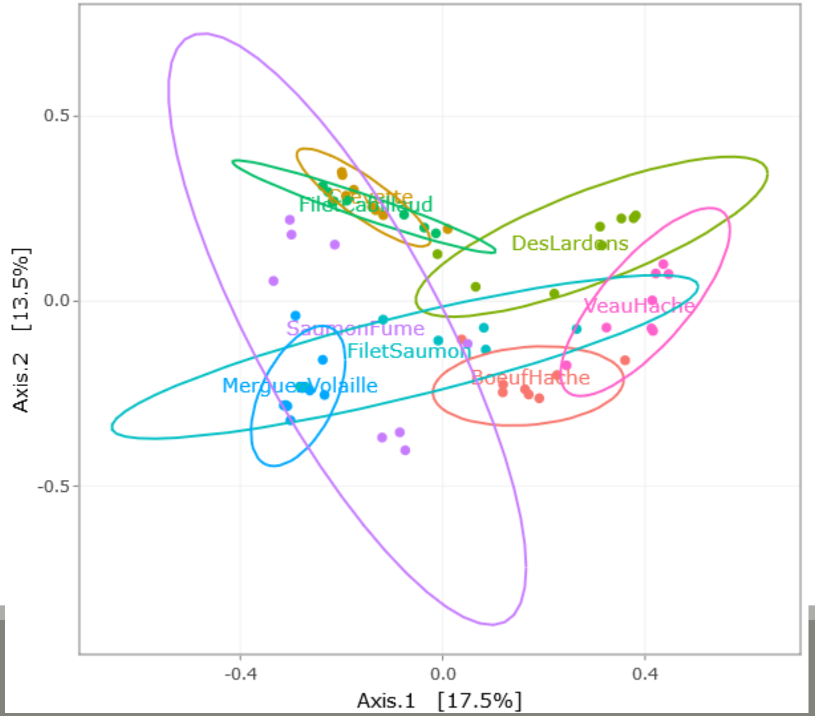
WUNIFRAC

JACCARD



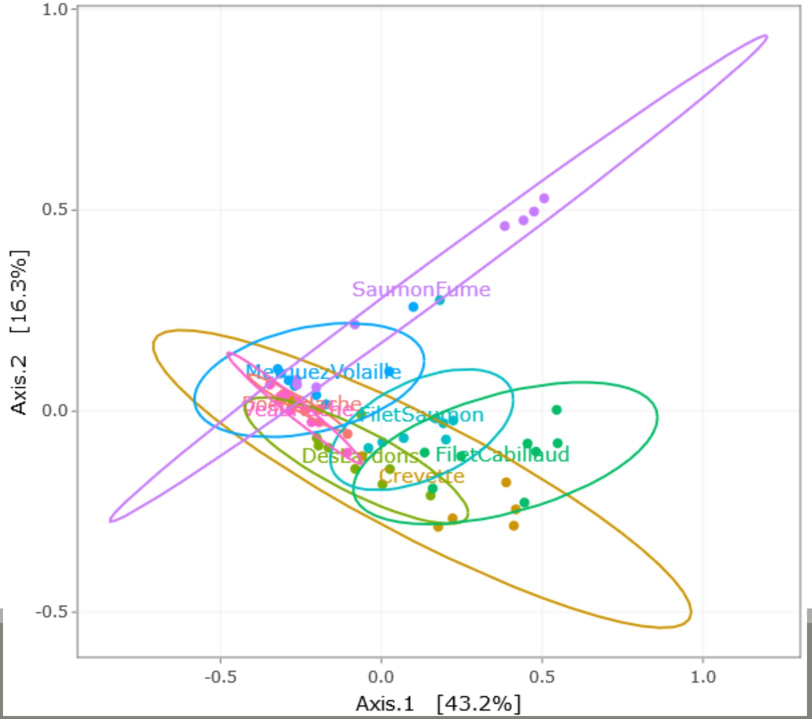
UNIFRAC



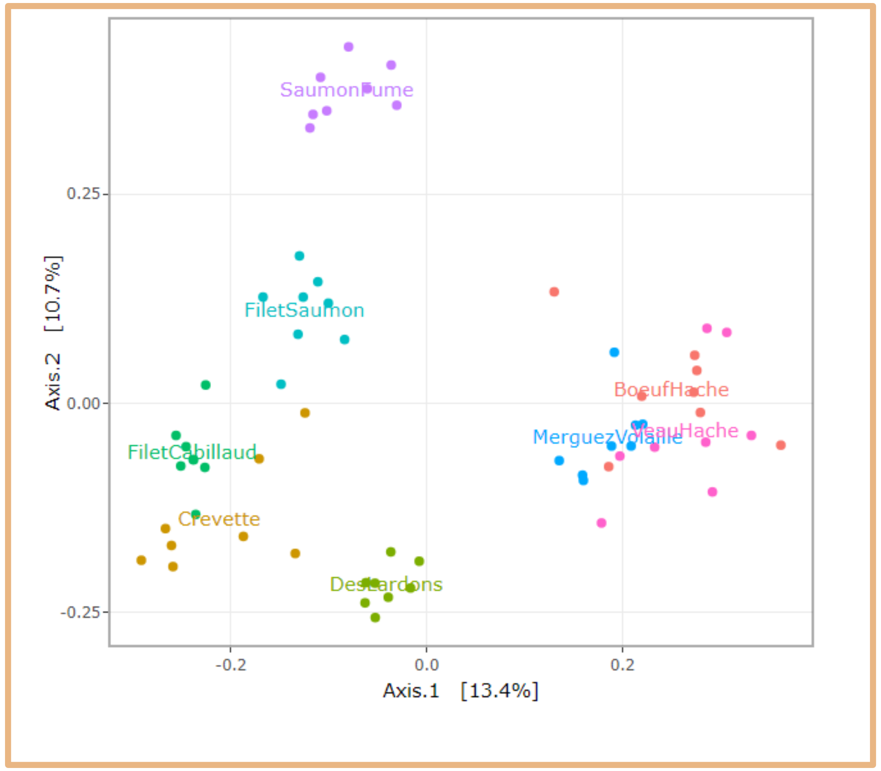


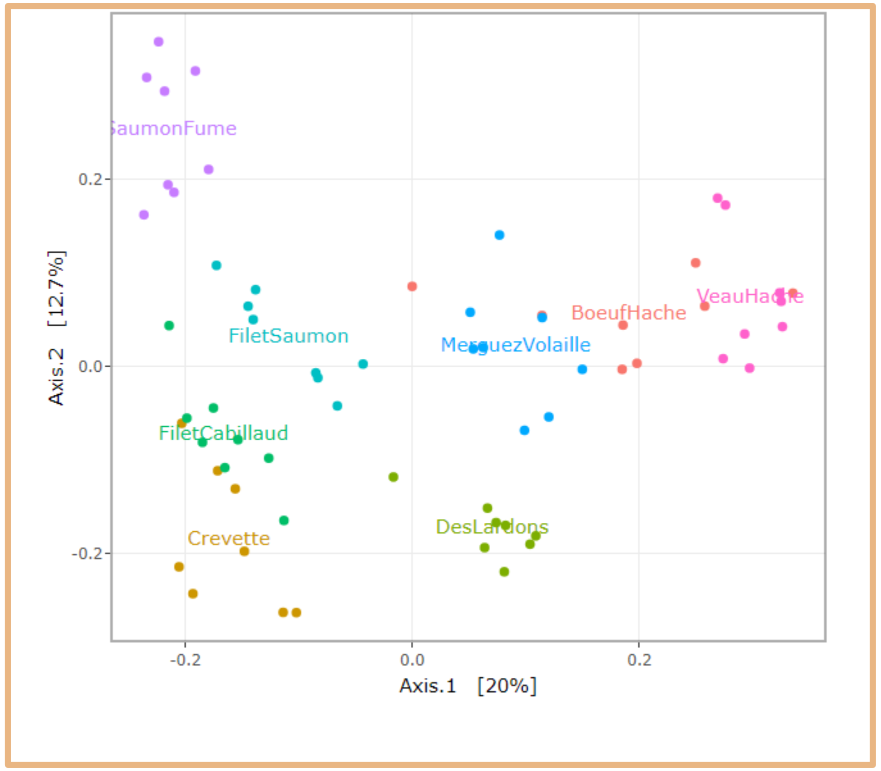
BRAY

WUNIFRAC



JACCARD





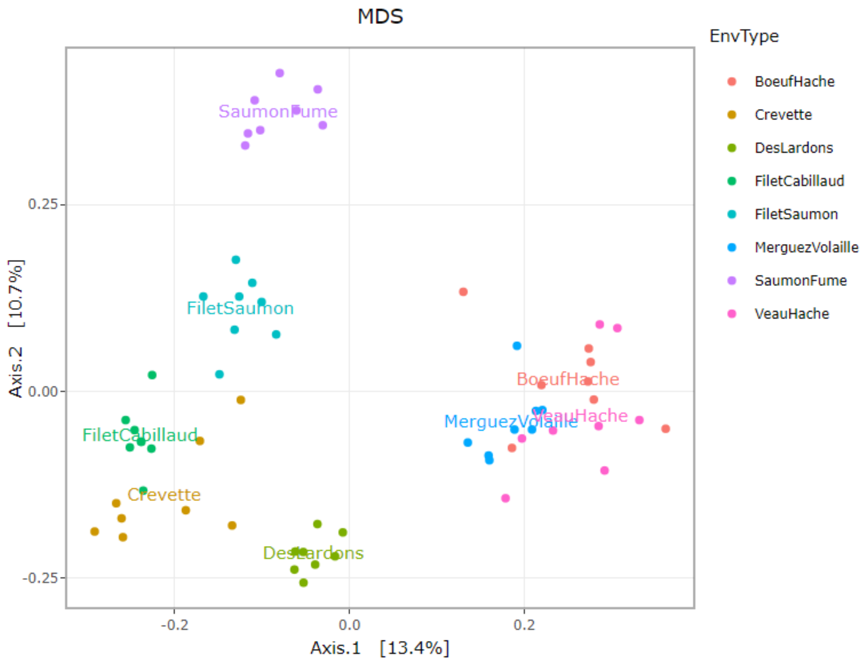
UNIFRAC

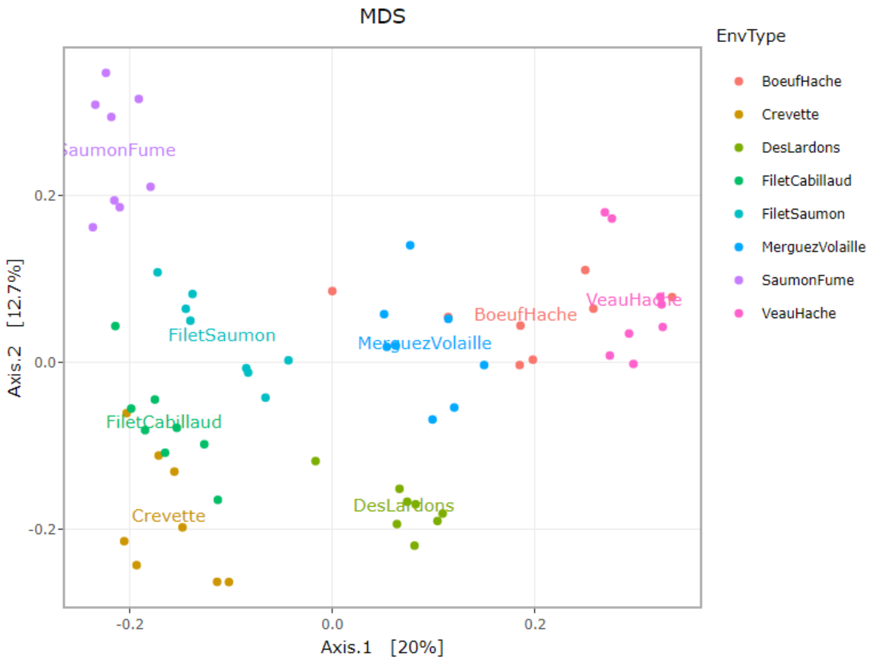
- Qualitative distances (Unifrac, Jaccard) separate meat products from seafood ones

➔ detected taxa segregate by origin

Exploring the structure : Ordination plot and Heatmap

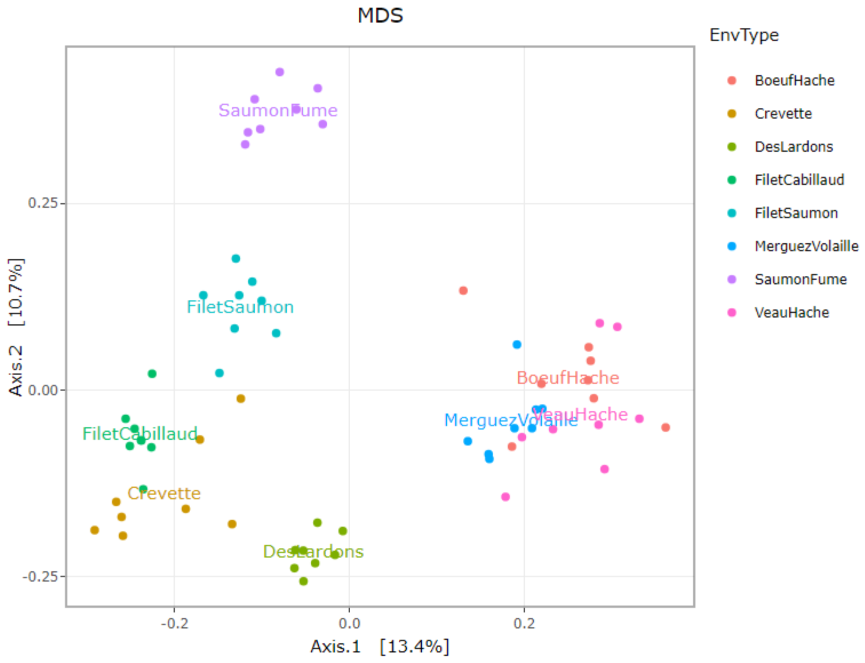
3. Guess why Lardon are somewhere between Meat and Seafood ?

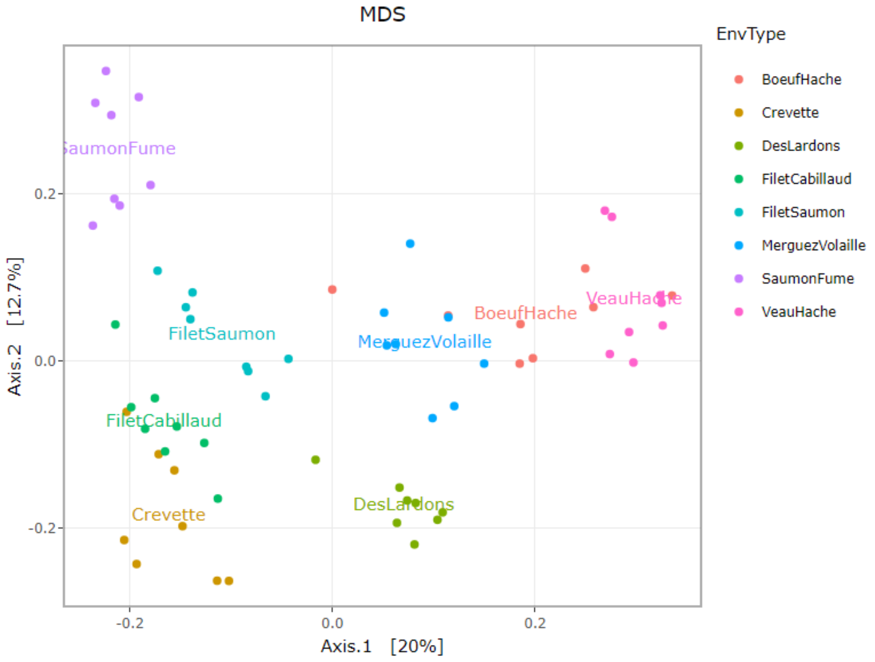




Exploring the structure : Ordination plot and Heatmap

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





- DesLardons is somewhere in between ➔ contamination induced by sea salt

Exploring the structure : Ordination plot and Heatmap

Other conclusions ?

Exploring the structure : Ordination plot and Heatmap

Other conclusions ?

- 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.
- Note the difference between weighted UniFrac and Bray-Curtis for the distances between BoeufHache and VeauHache.
- Warning
  - The 2-D representation captures only part of the original distances.
  - Ellipse are not always an advantage for visualisation

Exploring the structure : Ordination plot and Heatmap

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

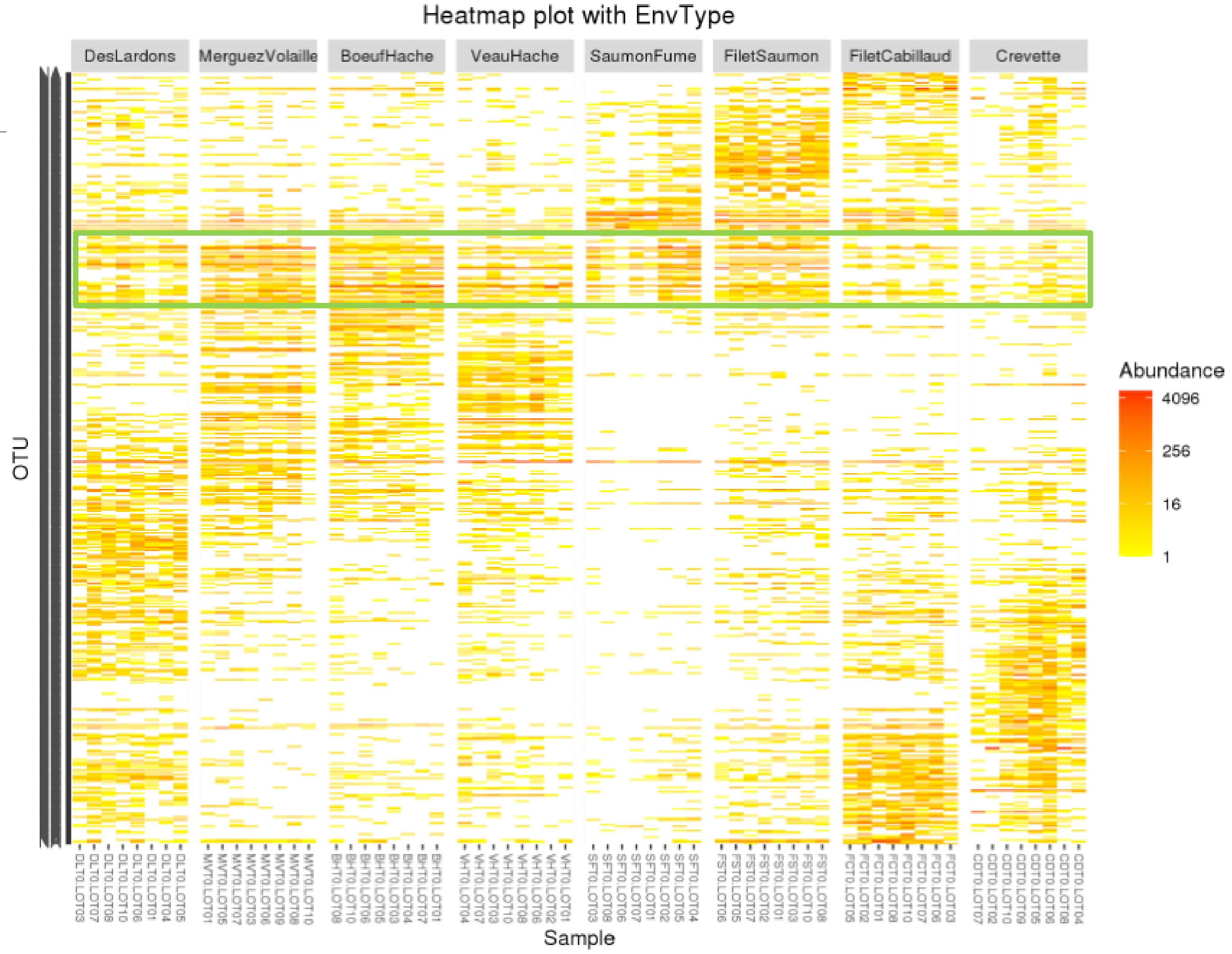
Try to identify:

- 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 ?

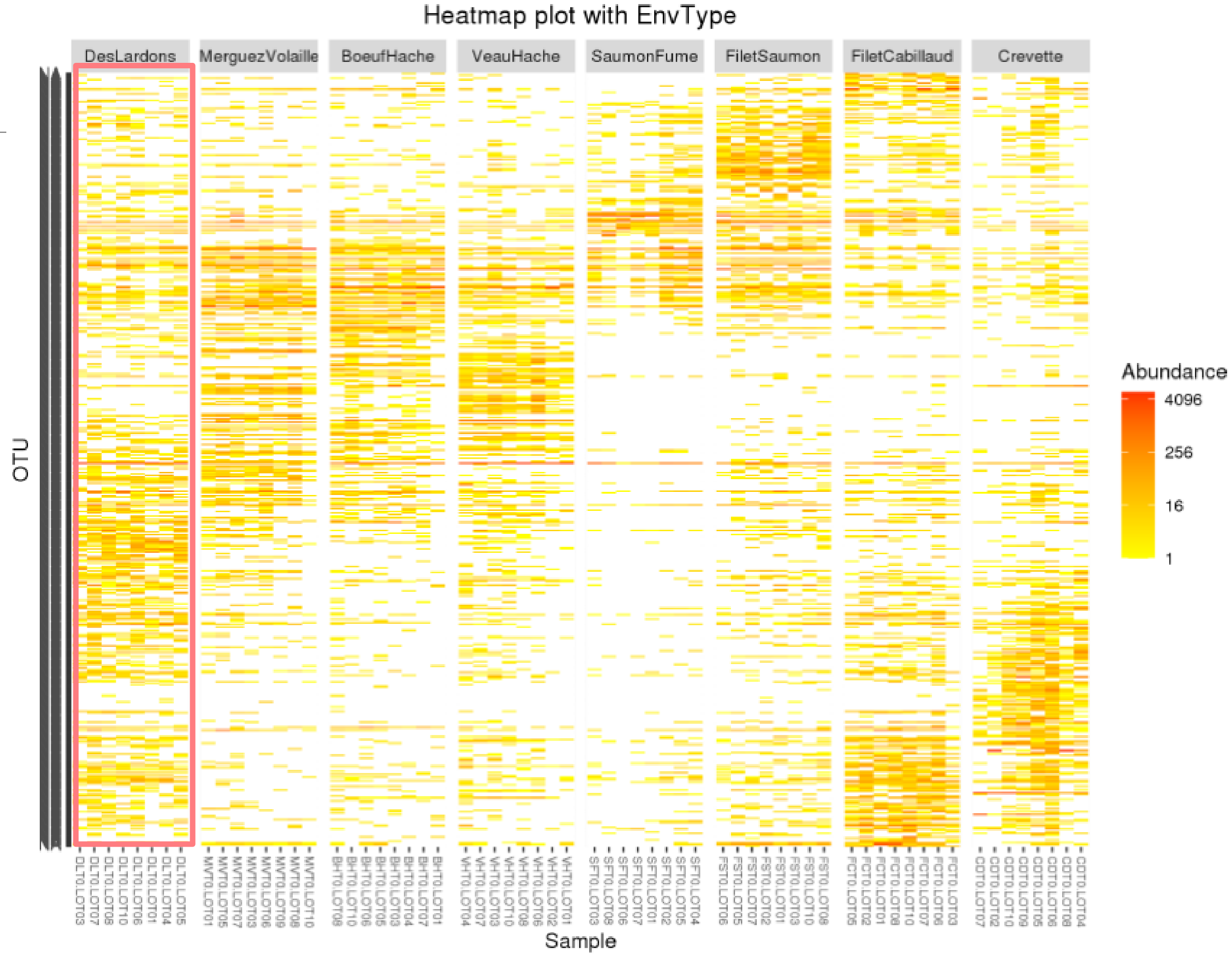
OTU shared by all samples



Exercise 7

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

DesLardon have a lot of OTU in common with seafoods



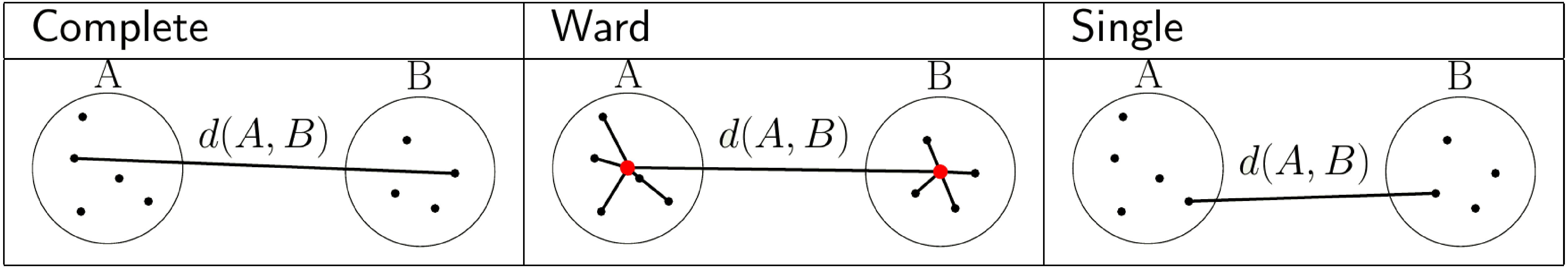
II. Exploring the structure

HIERARCHICAL CLUSTERING

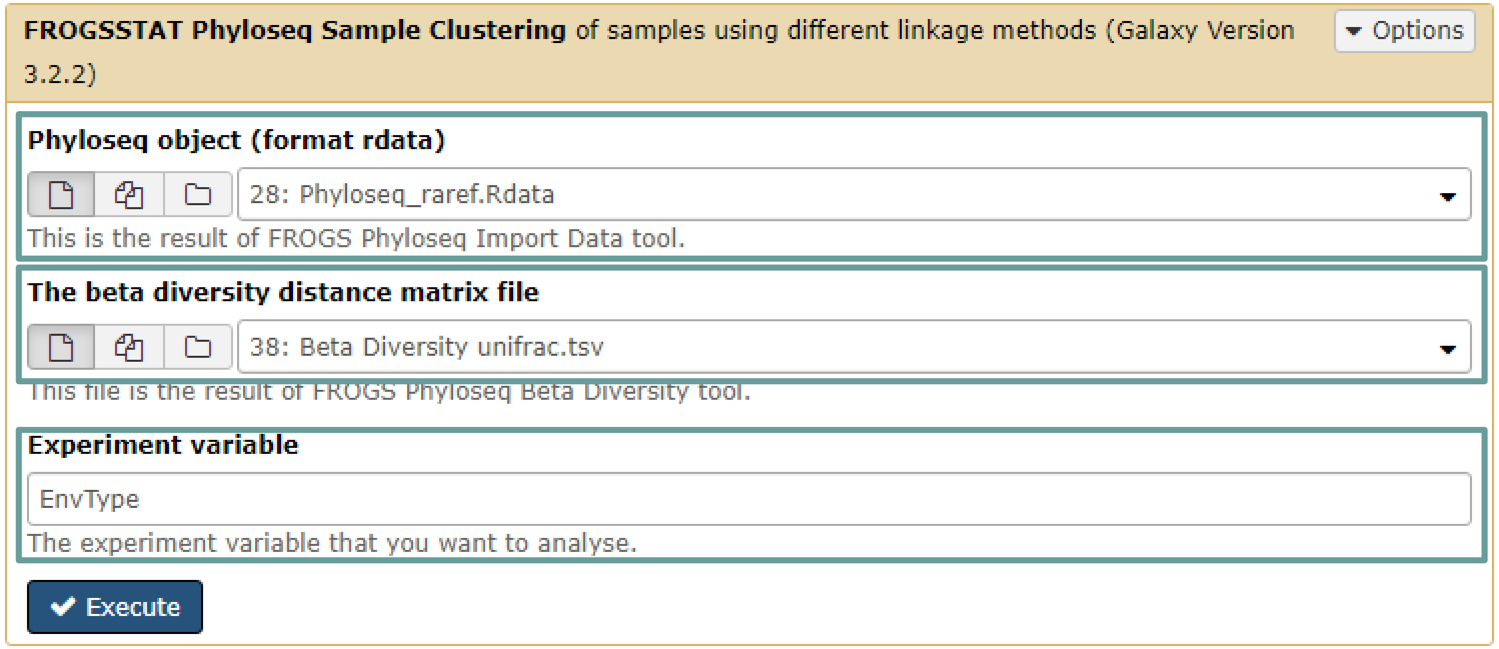
Exploring the structure : clustering

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

- 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)

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.

Explore the sample **NORMALISED** count

Choose the beta diversity distance matrix

Choose a sample variable to organize graphics

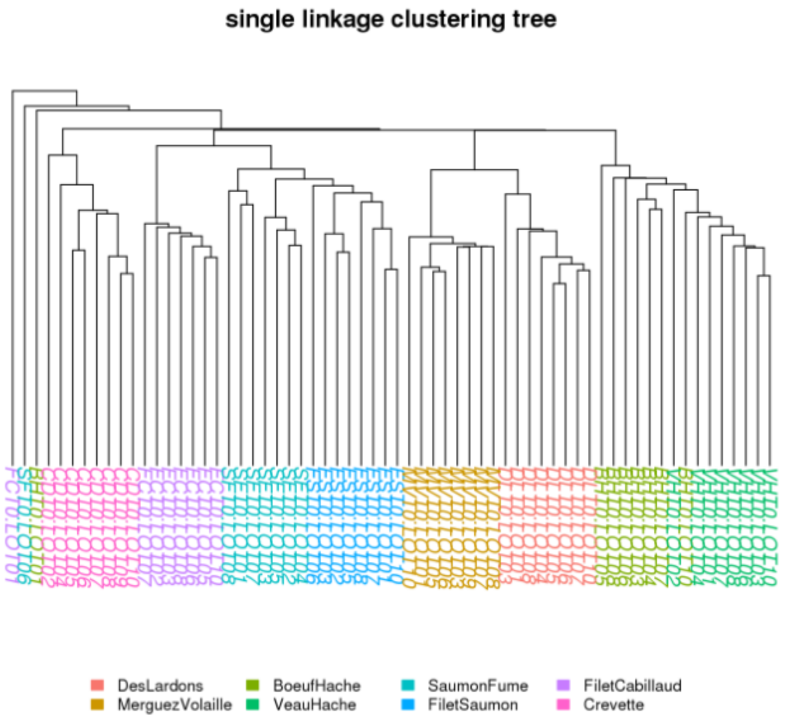
The three different linkage functions will be used, generating three different dendrograms

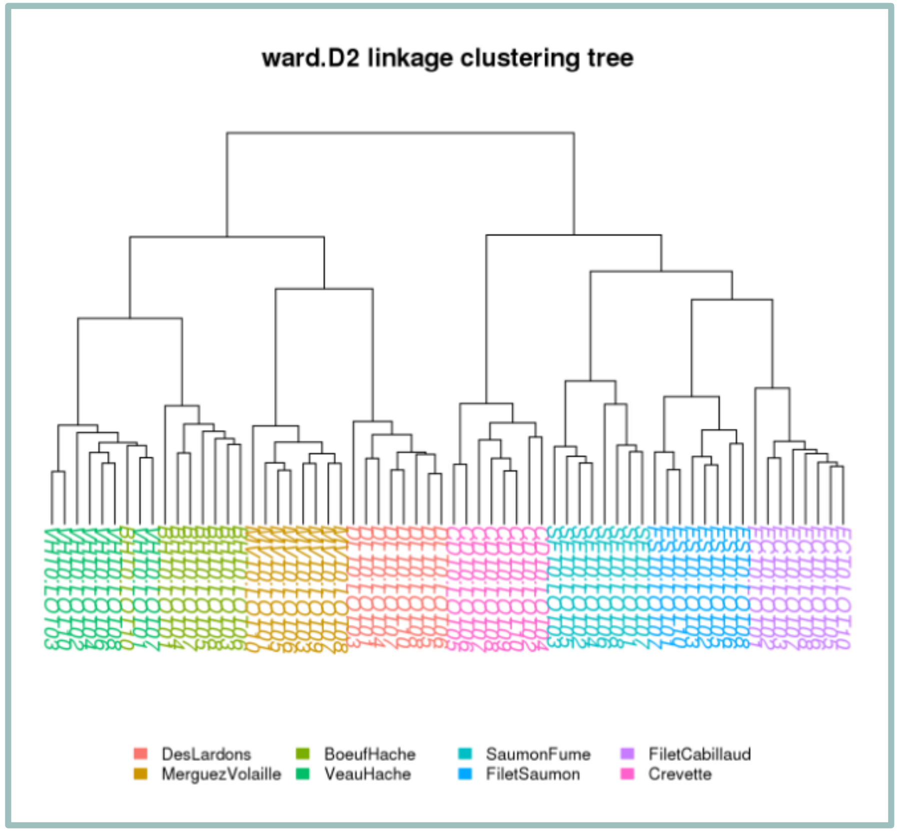
Exercise 8

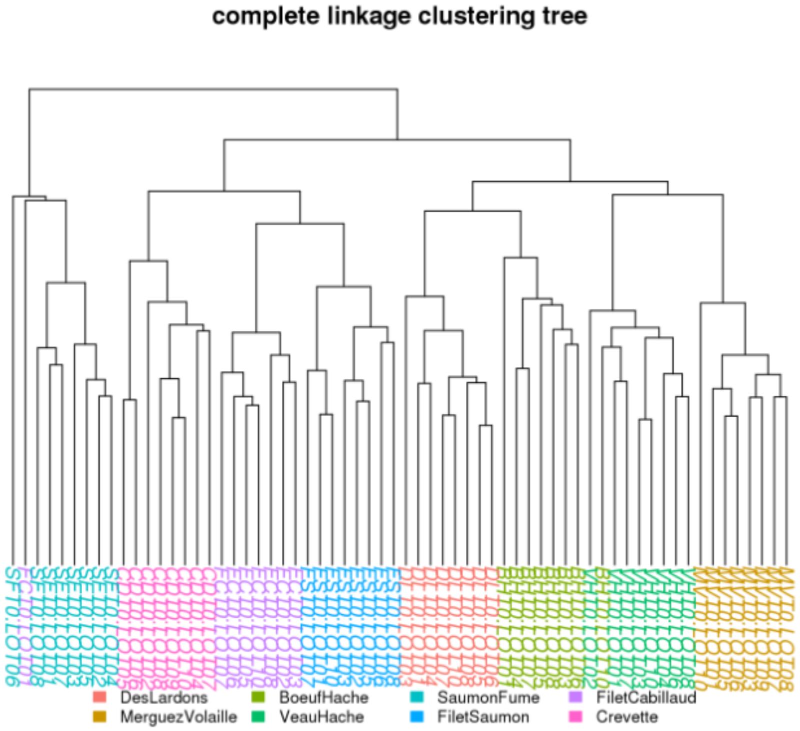
Try it with « a good » distance method matrix on EnvType and on FoodType

➔ Which linkage method seems better to fit the data ?

Exercise 8

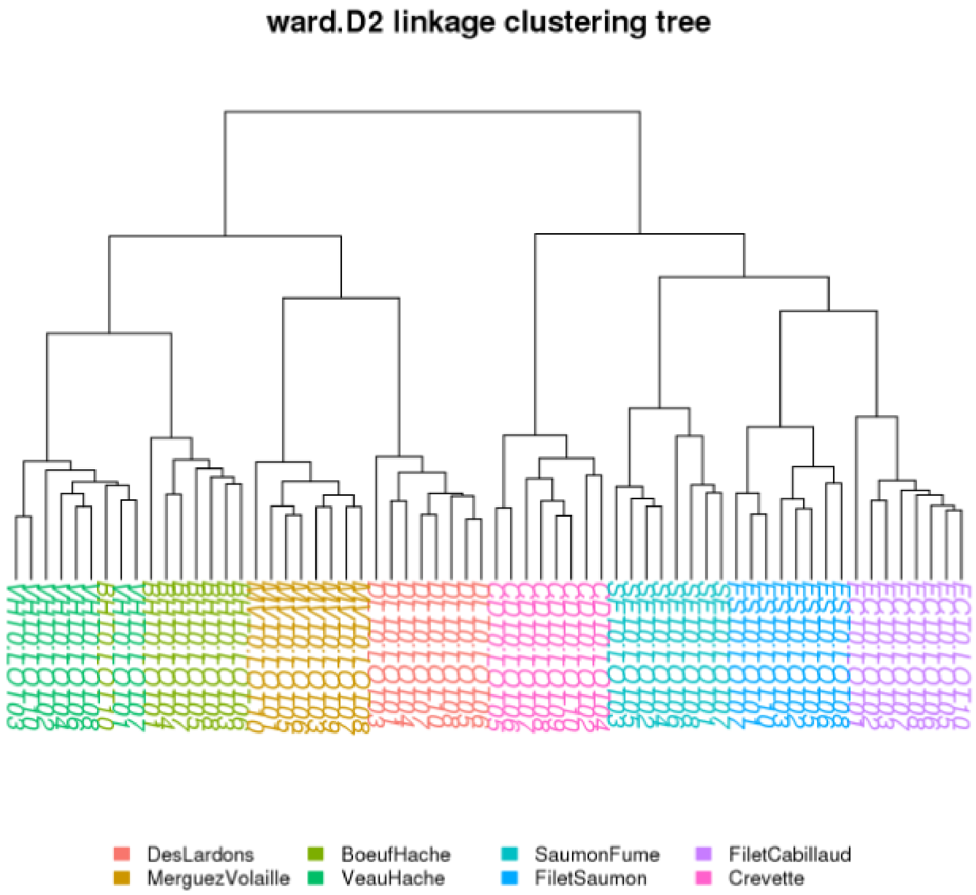

single linkage clustering tree
DesLardons
MerguezVolaille
BoeufHache
VeauHache
SaumonFume
FiletSaumon
FiletCabillaud
Crevette


ward.D2 linkage clustering tree
DesLardons
MerguezVolaille
BoeufHache
VeauHache
SaumonFume
FiletSaumon
FiletCabillaud
Crevette


complete linkage clustering tree
DesLardons
MerguezVolaille
BoeufHache
VeauHache
SaumonFume
FiletSaumon
FiletCabillaud
Crevette

Exercise 8

- Consistently 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)

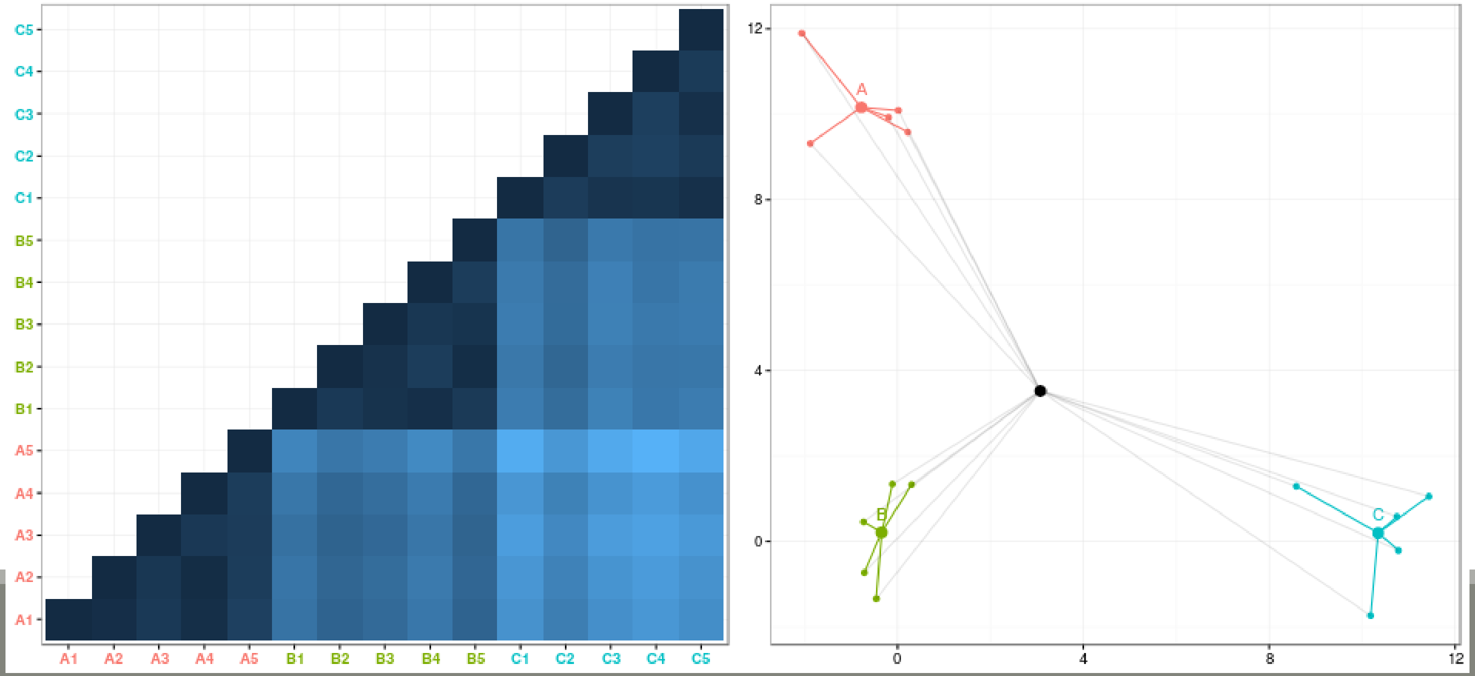
Diversity partitioning

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

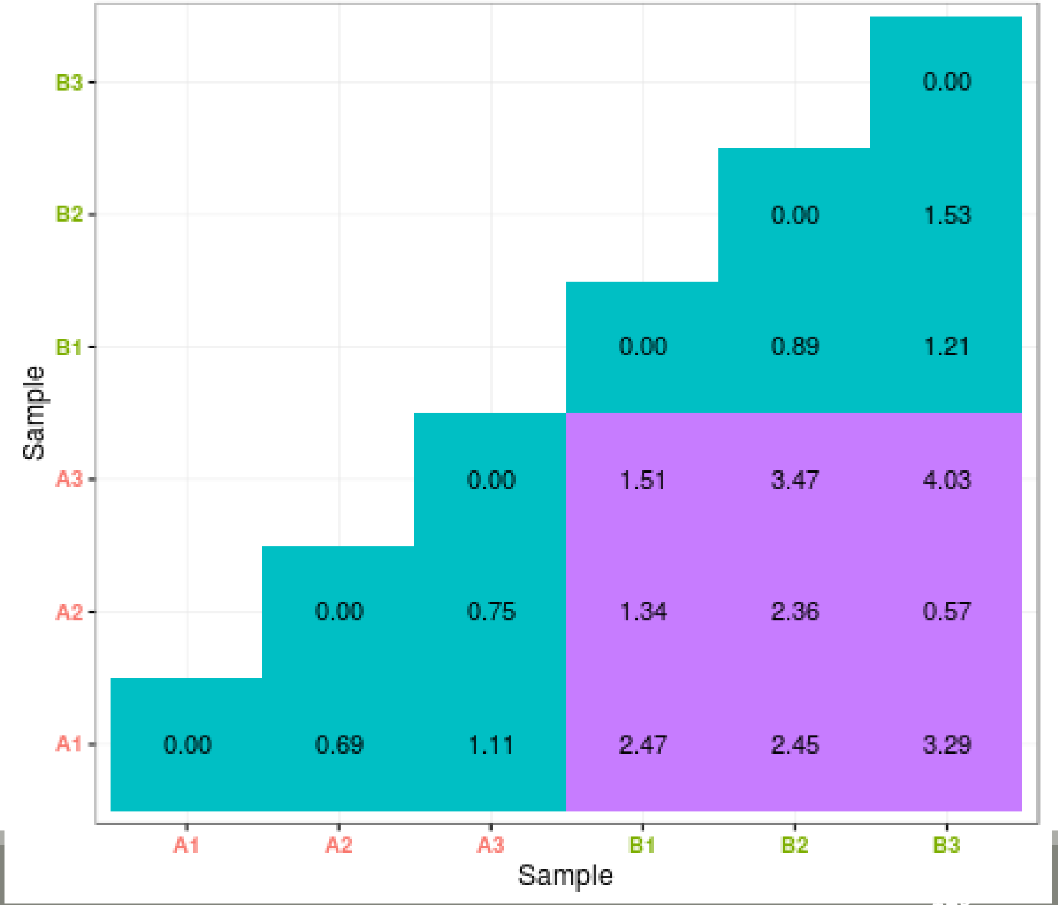


Diversity partitioning : Multivariate ANOVA

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

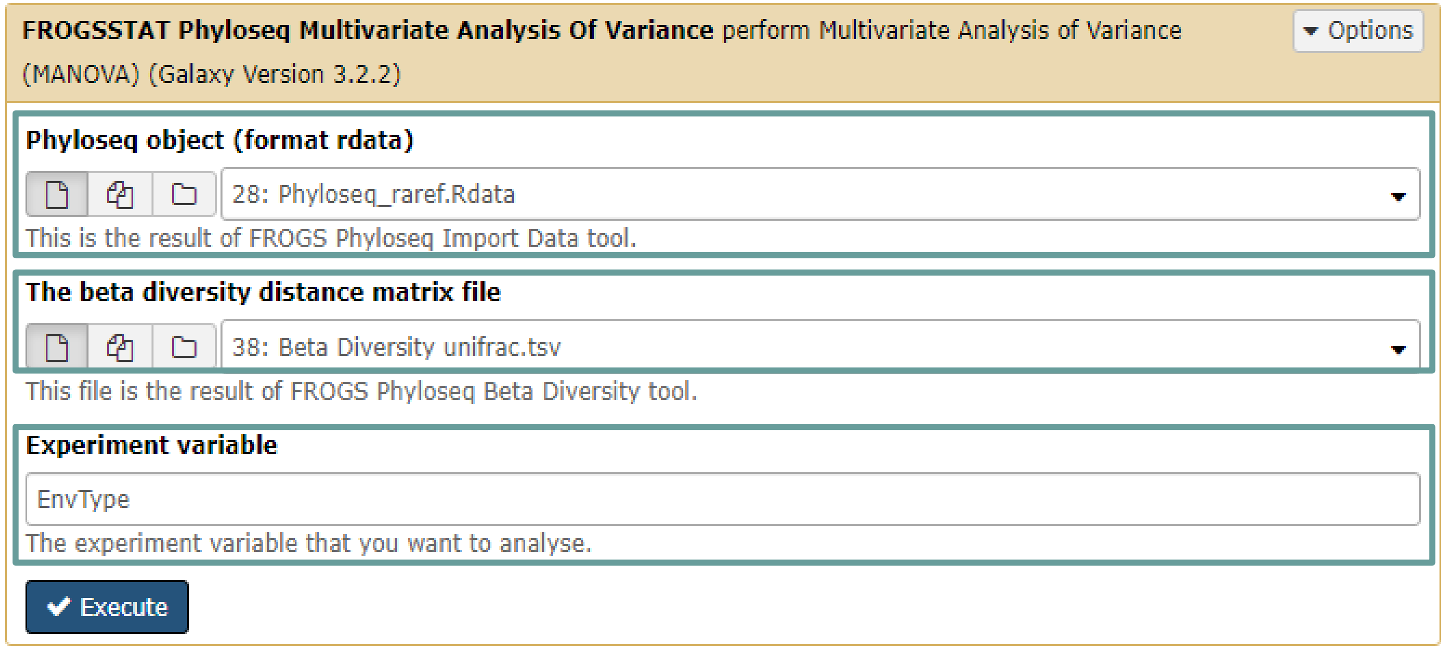
How it works ?

- Computes sum of square distance
- Variance analysis



Sample	A1	A2	A3	B1	B2	B3
B3						0.00
B2					0.00	1.53
B1				0.00	0.89	1.21
A3			0.00	1.51	3.47	4.03
A2		0.00	0.75	1.34	2.36	0.57
A1	0.00	0.69	1.11	2.47	2.45	3.29

Diversity partitioning : Multivariate ANOVA



FROGSSTAT Phyloseq Multivariate Analysis Of Variance perform Multivariate Analysis of Variance (MANOVA) (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

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.

Explore the sample **NORMALISED** count

Choose the beta diversity distance matrix

Choose the variable to explain the variability between samples

Exercise 9

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

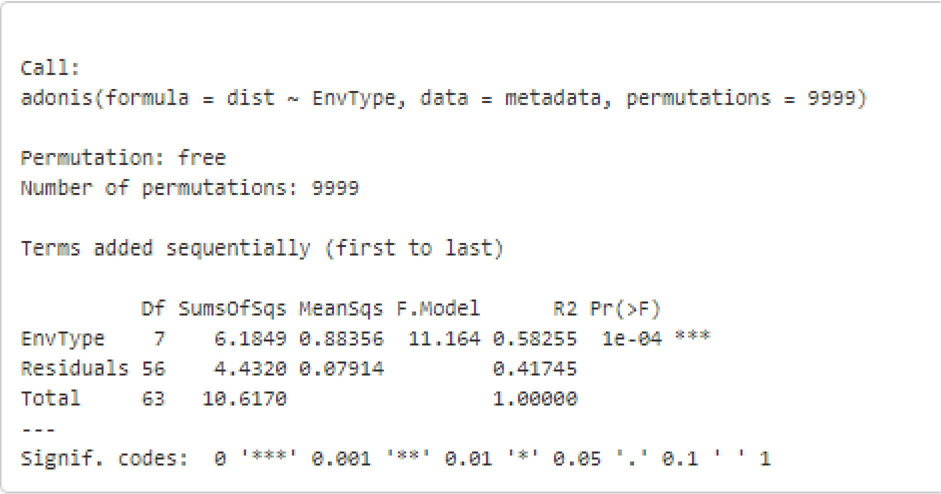
1. Does EnvType have an influence on the beta diversity variance ?

2. What about FoodType ?

Exercise 9

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

With Unifrac distance



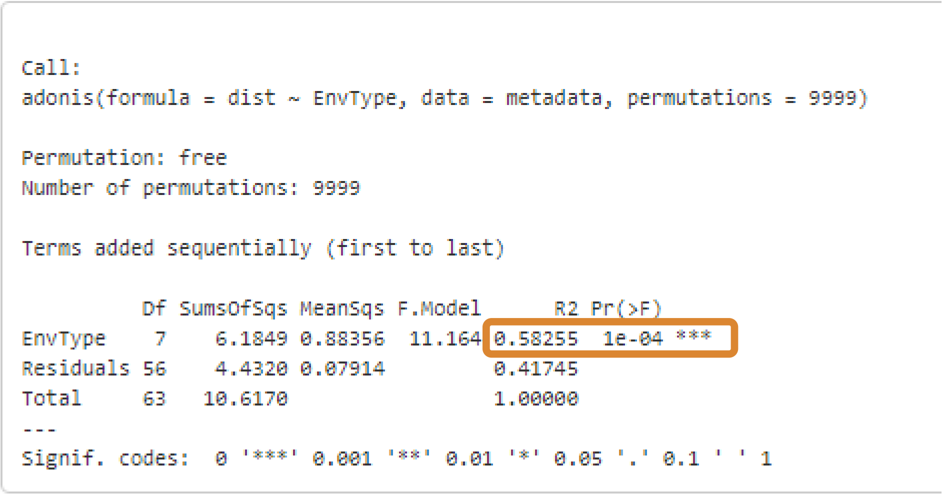
```
Call:
adonis(formula = dist ~ EnvType, data = metadata, permutations = 9999)

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

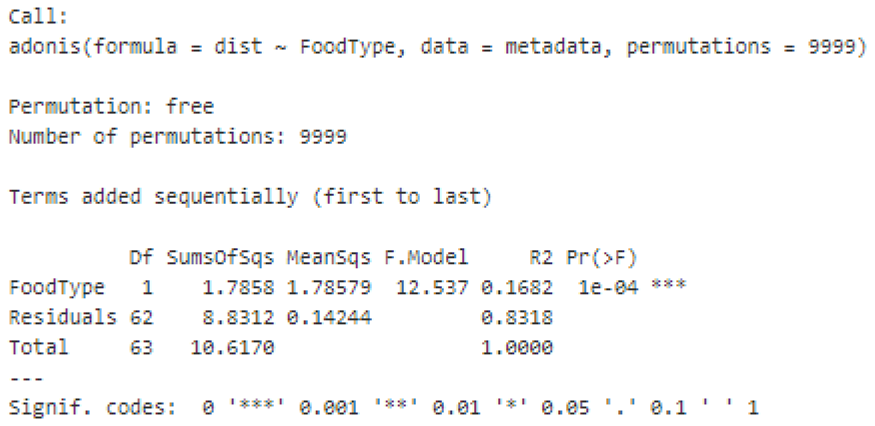


```
Call:
adonis(formula = dist ~ EnvType, data = metadata, permutations = 9999)

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 ?

With Unifrac distance



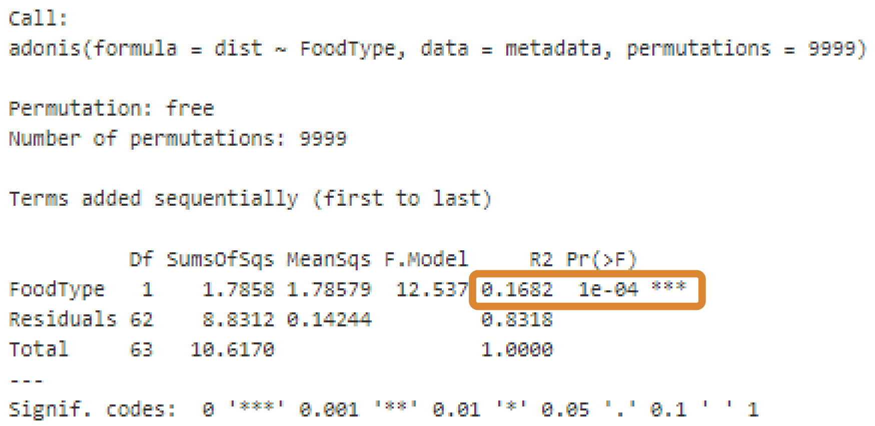
```
Call:
adonis(formula = dist ~ FoodType, data = metadata, permutations = 9999)

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
```

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

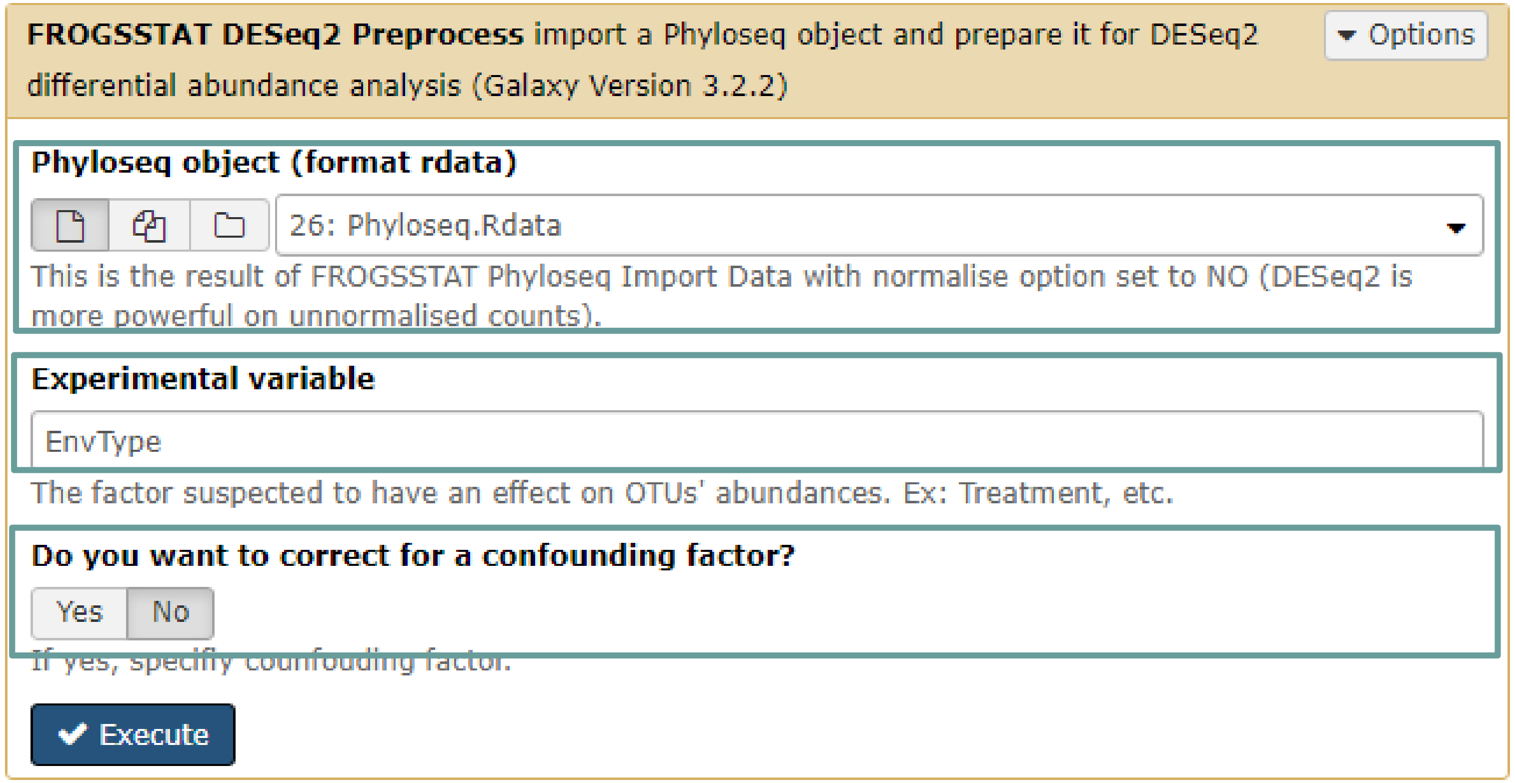
The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models

Be aware to use data without normalisation

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

It uses the poscount function optimised for metagenomic count table

Differential abundance analysis



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

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?

Yes [✓ No]

If yes, specify counfouding factor.

Explore the sample **RAW** count

Choose the factor on which the differential abundances will be compared

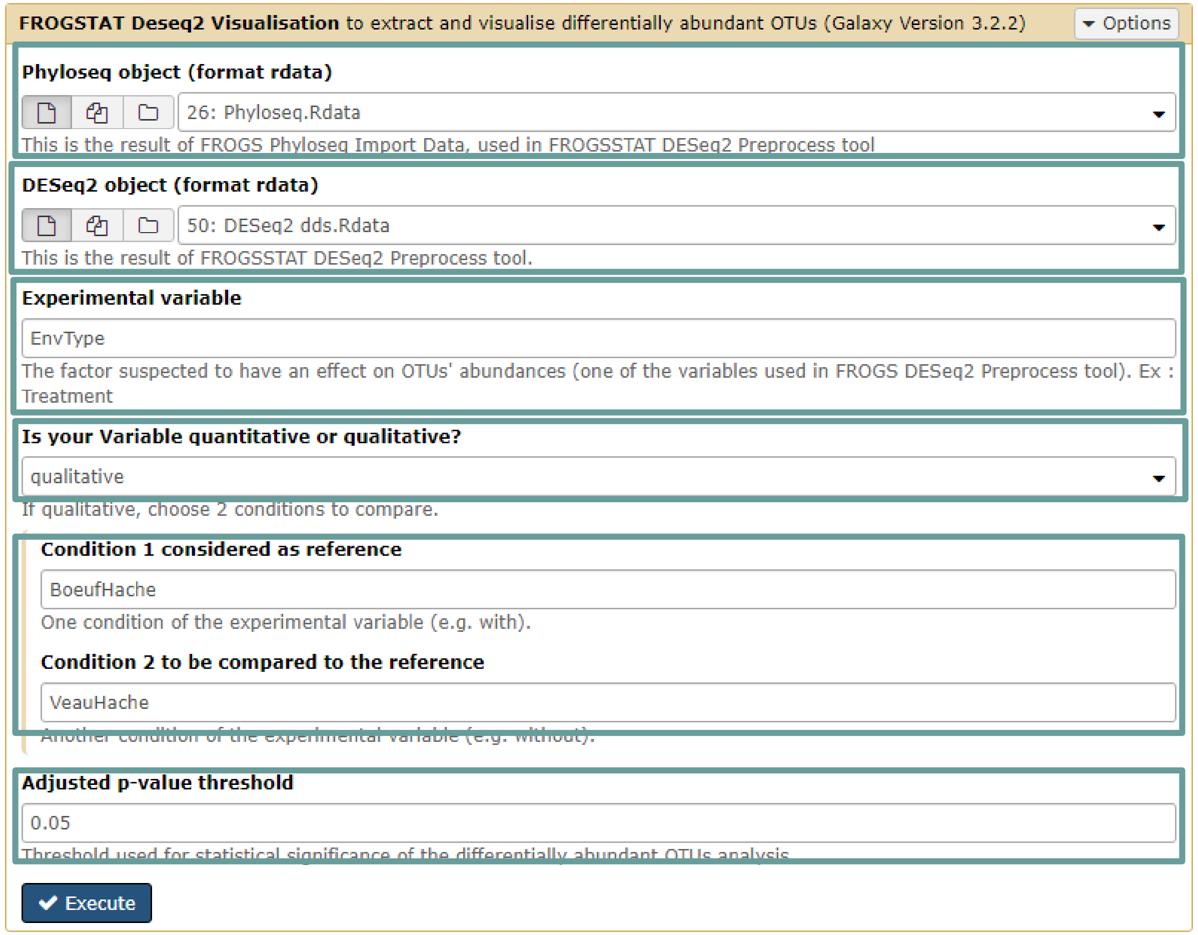
Specify a confounding factor if necessary *(example : testing antibiotic treatment effect with 2 different mice phenotypes, or testing drought effect on soil microbiome with two soil compositions)*

Differential abundance analysis

➔ What are the output datasets ?

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

Differential abundance visualisation



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

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

Explore the sample **RAW** count

Result of FROGSSTAT DESeq2 preprocess

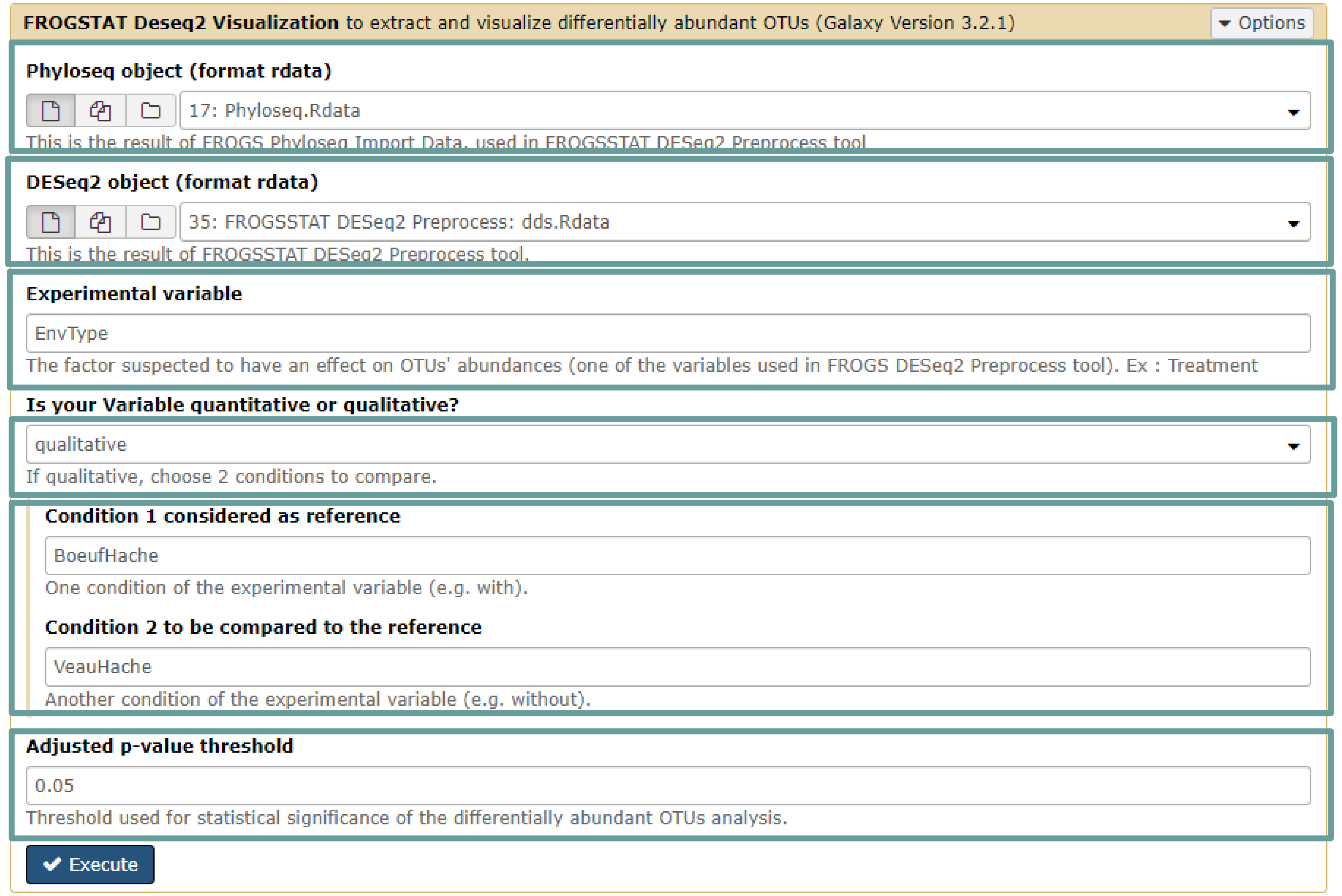
Factor on which the differential abundances have been tested

Specify qualitative or quantitative

Precise the two conditions to compare

Statistical significance threshold (default 0.05)

Differential abundance visualisation



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

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

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.

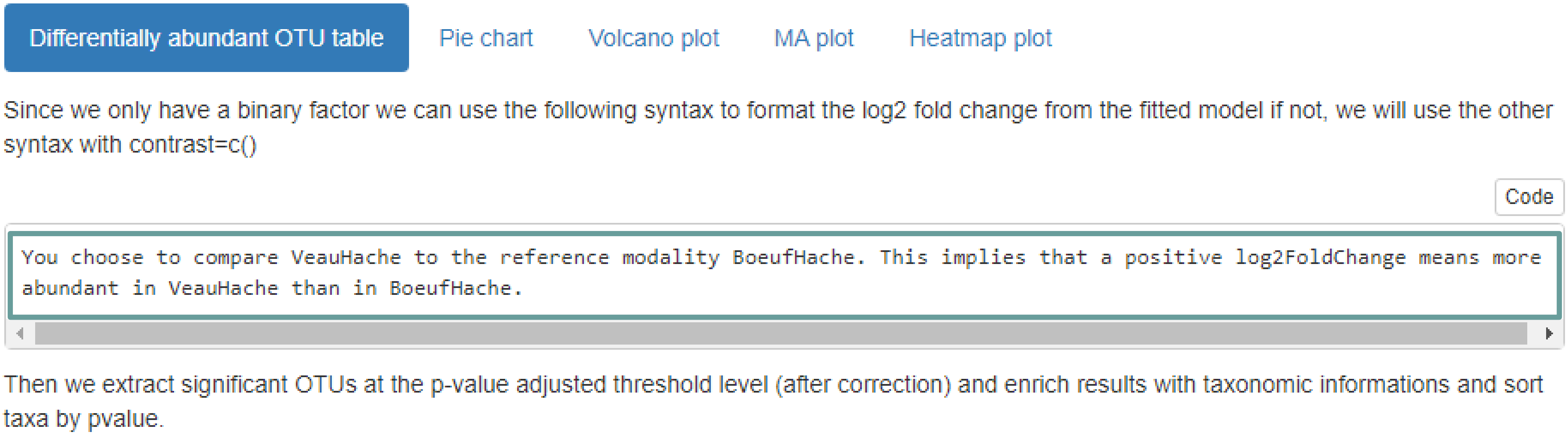
Compare BoeufHache vs VeauHache

Differential abundance visualisation

What are the output datasets ?

→ HTML report: result table and several plot

Differential abundance visualisation

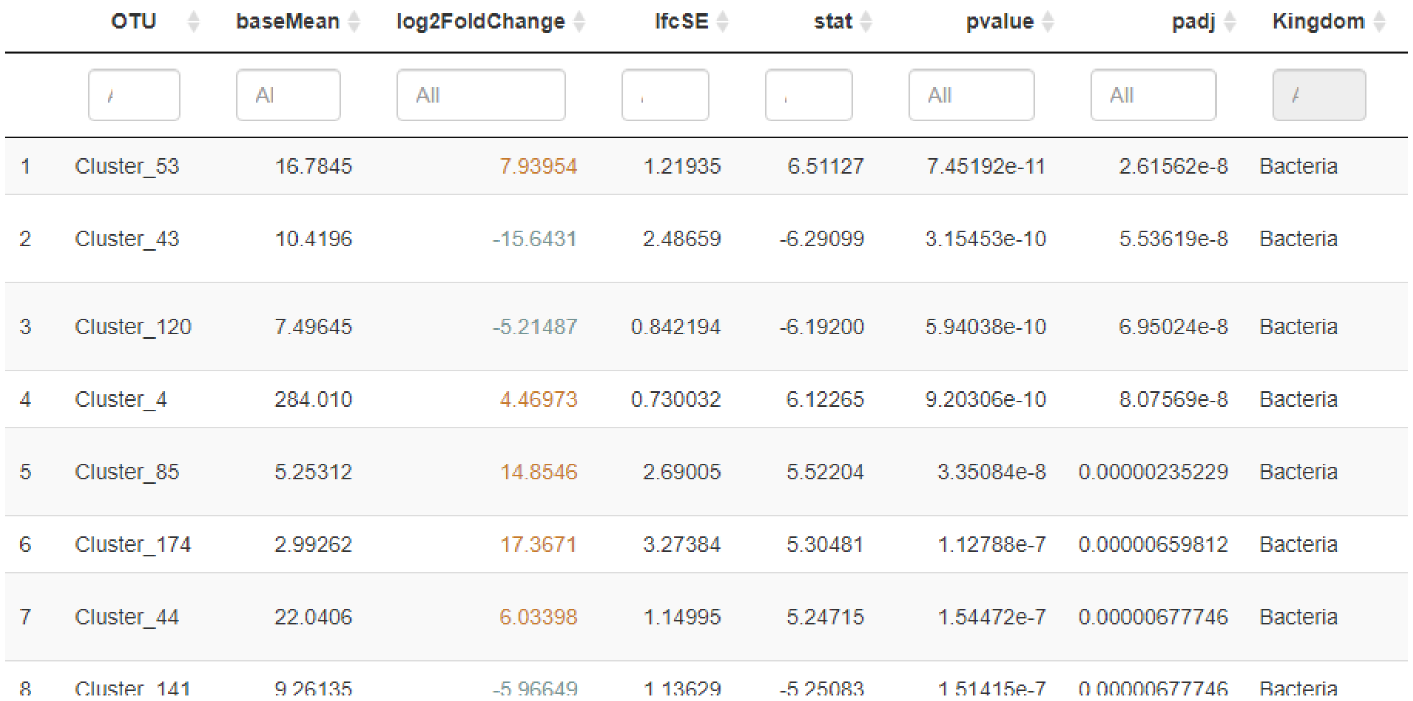


Since we only have a binary factor we can use the following syntax to format the log2 fold change from the fitted model if not, we will use the other syntax with contrast=c()

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

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 visualisation



	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	/	Al	All	.	.	All	All	/
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)

p-value are adjusted using the Benjamini-Hochberg method

Differential abundance visualisation

Why log2Foldchange ?

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

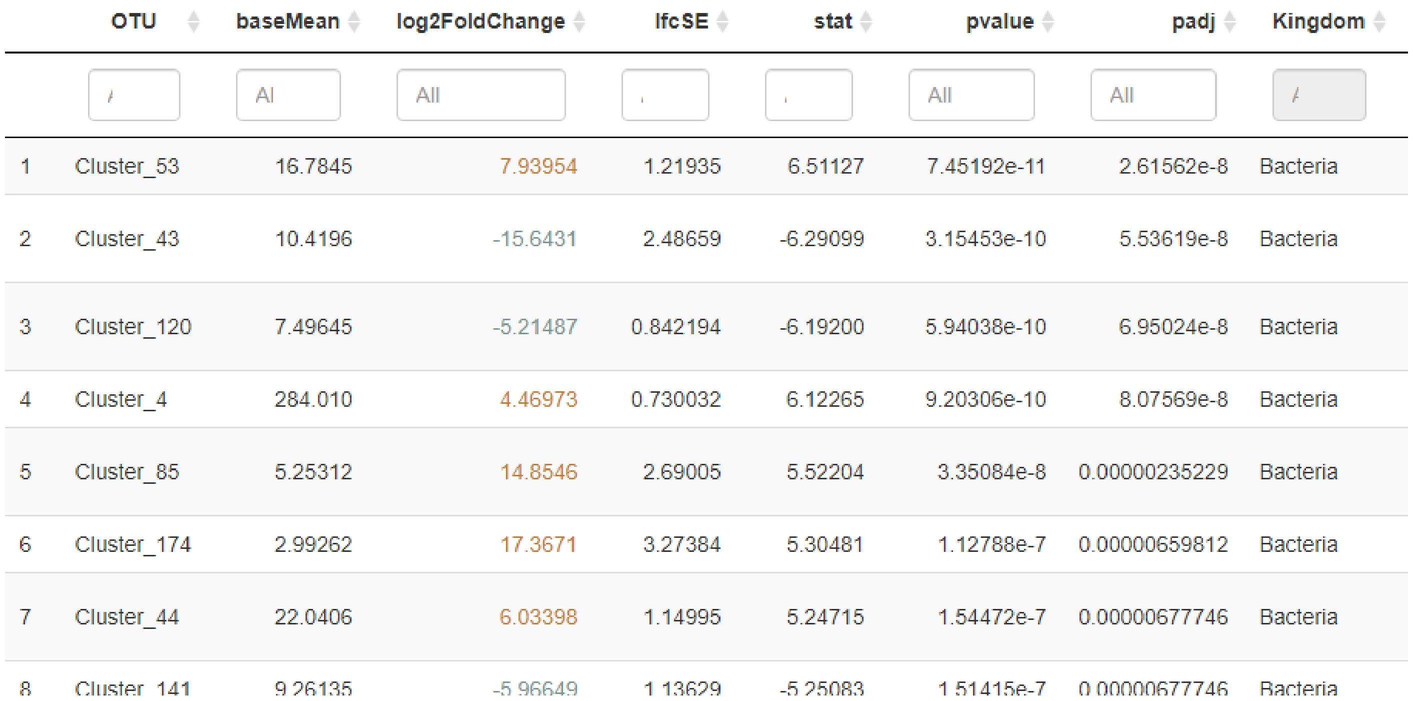
log2 is used for interpret and scale reasons:

- Positive values denote an increase, and negative a decrease of abundance

- log2FC = 1 means a **doubling**
- log2FC = 2 means a **quadrupling**
- log2FC = -1 means a halving
- log2FC = -2 means a **quartering**
- …

Differential abundance visualisation

Differentially abundant OTU table



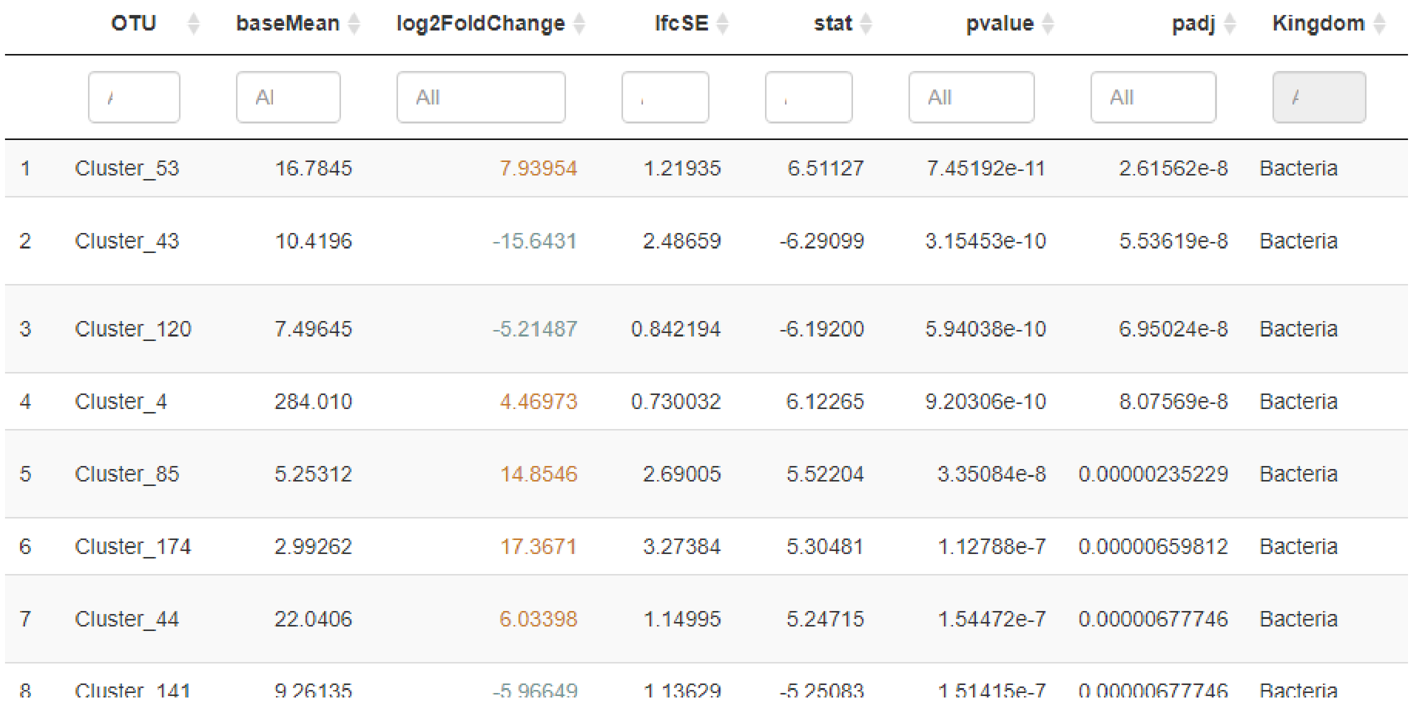
	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	/	Al	All	.	.	All	All	/
1	Cluster_53	16.7845	7.93954	1.21935	6.51127	7.45192e-11	2.61562e-8	Bacteria
2	Cluster_43	10.4196	-15.6431	2.48659	-6.29099	3.15453e-10	5.53619e-8	Bacteria
3	Cluster_120	7.49645	-5.21487	0.842194	-6.19200	5.94038e-10	6.95024e-8	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 visualisation

Differentially abundant OTU table

➔ Significance of the sign of the log2Foldchange ?

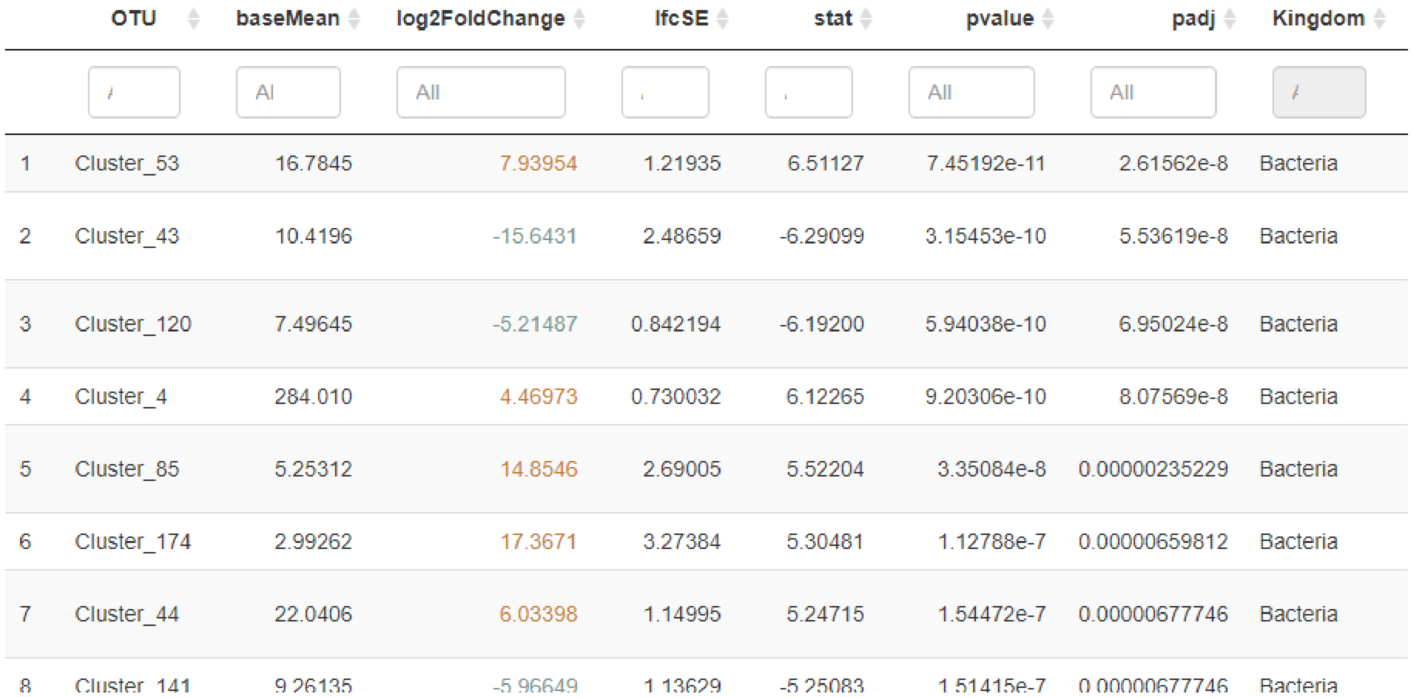


	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	/	Al	All	,	,	All	All	/
1	Cluster_53	16.7845	7.93954	1.21935	6.51127	7.45192e-11	2.61562e-8	Bacteria
2	Cluster_43	10.4196	-15.6431	2.48659	-6.29099	3.15453e-10	5.53619e-8	Bacteria
3	Cluster_120	7.49645	-5.21487	0.842194	-6.19200	5.94038e-10	6.95024e-8	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 visualisation

Differentially abundant OTU table

➔ Significance of the sign of the log2Foldchange ?



	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	/	Al	All	,	,	All	All	/
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

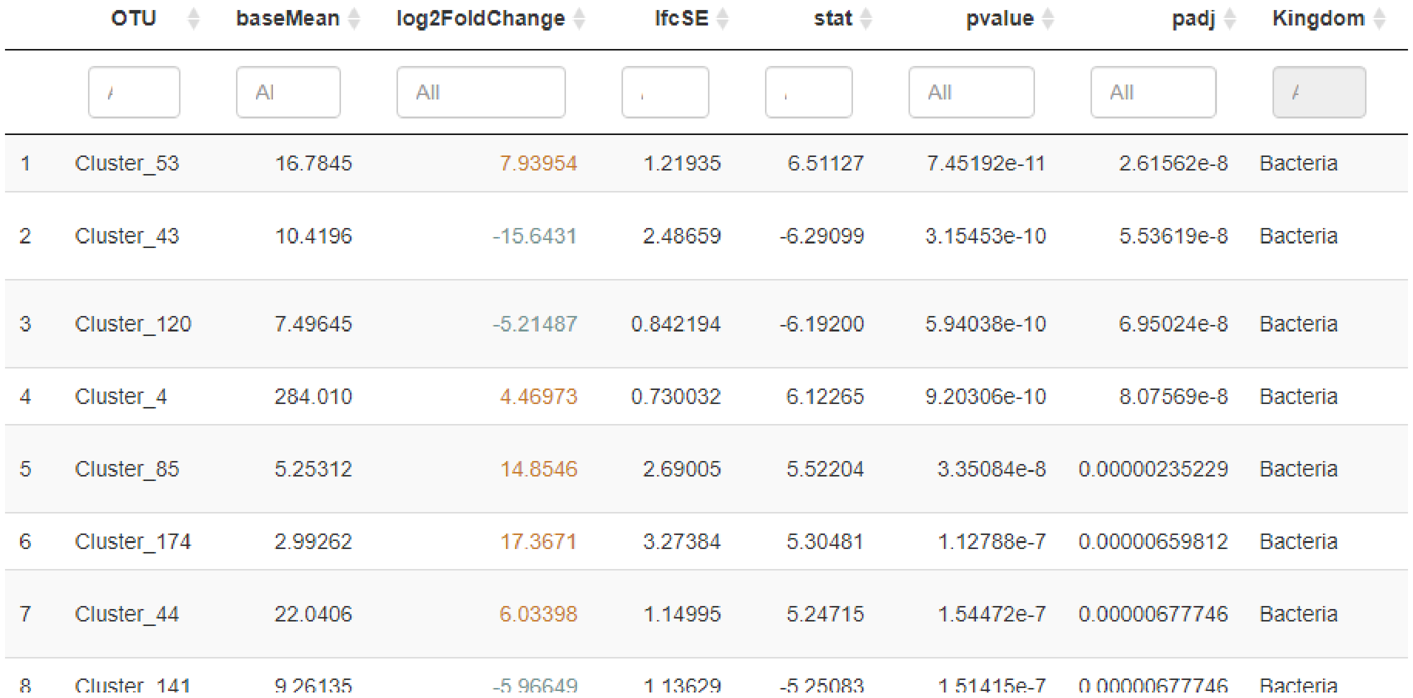
Positive log2FoldChange means more abundant in VeauHache than in BoeufHache

Cluster_53 is more abundant in VeauHache than in BoeufHache

Differential abundance visualisation

Differentially abundant OTU table

➔ Which species have the highest negative log2Foldchange ?

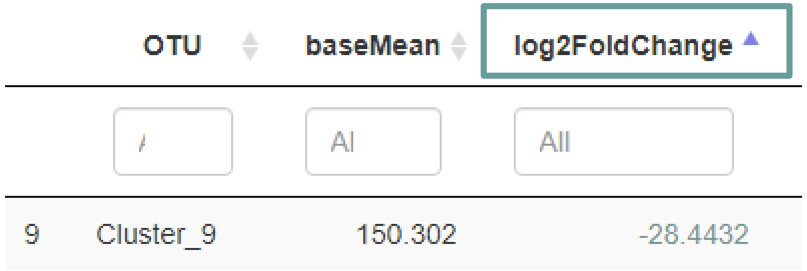


	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	/	Al	All	'	'	All	All	/
1	Cluster_53	16.7845	7.93954	1.21935	6.51127	7.45192e-11	2.61562e-8	Bacteria
2	Cluster_43	10.4196	-15.6431	2.48659	-6.29099	3.15453e-10	5.53619e-8	Bacteria
3	Cluster_120	7.49645	-5.21487	0.842194	-6.19200	5.94038e-10	6.95024e-8	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 visualisation

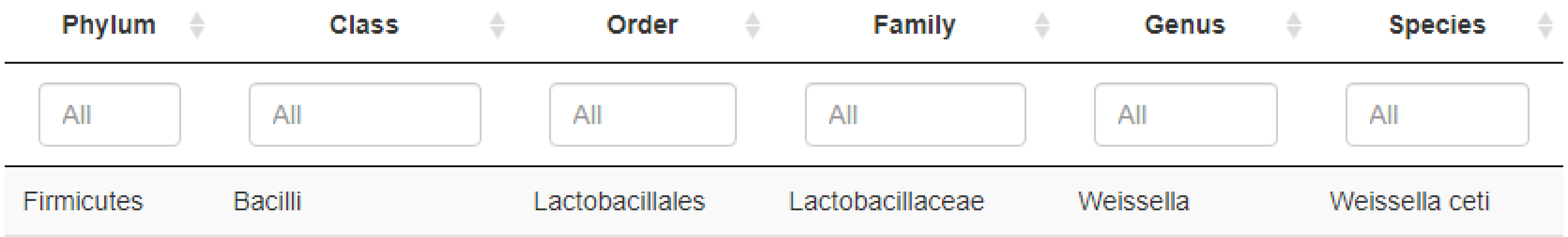
Differentially abundant OTU table

➔ Which species have the highest negative log2Foldchange ?



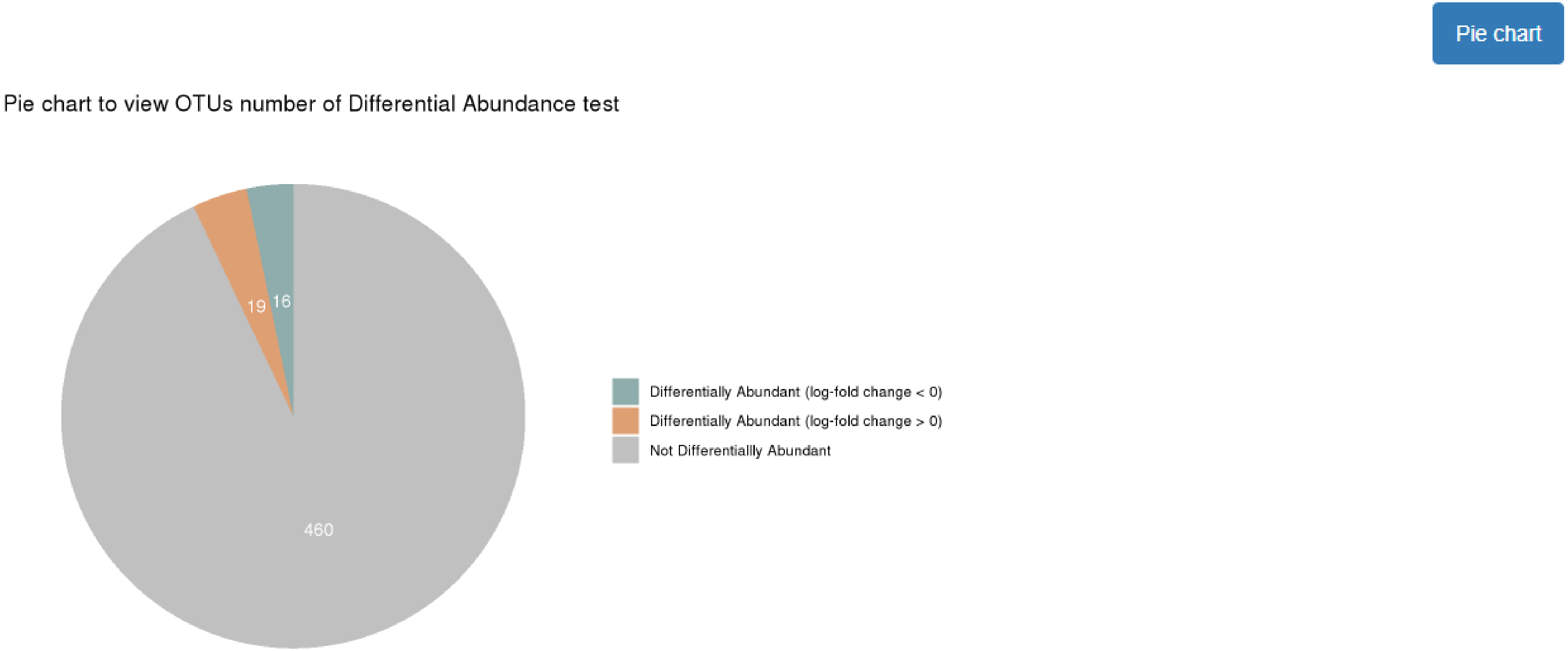
	OTU	baseMean	log2FoldChange
	/	Al	All
9	Cluster_9	150.302	-28.4432

It's the Cluster_9 which is a *Weissella ceti*



Phylum	Class	Order	Family	Genus	Species
All	All	All	All	All	All
Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Weissella	Weissella ceti

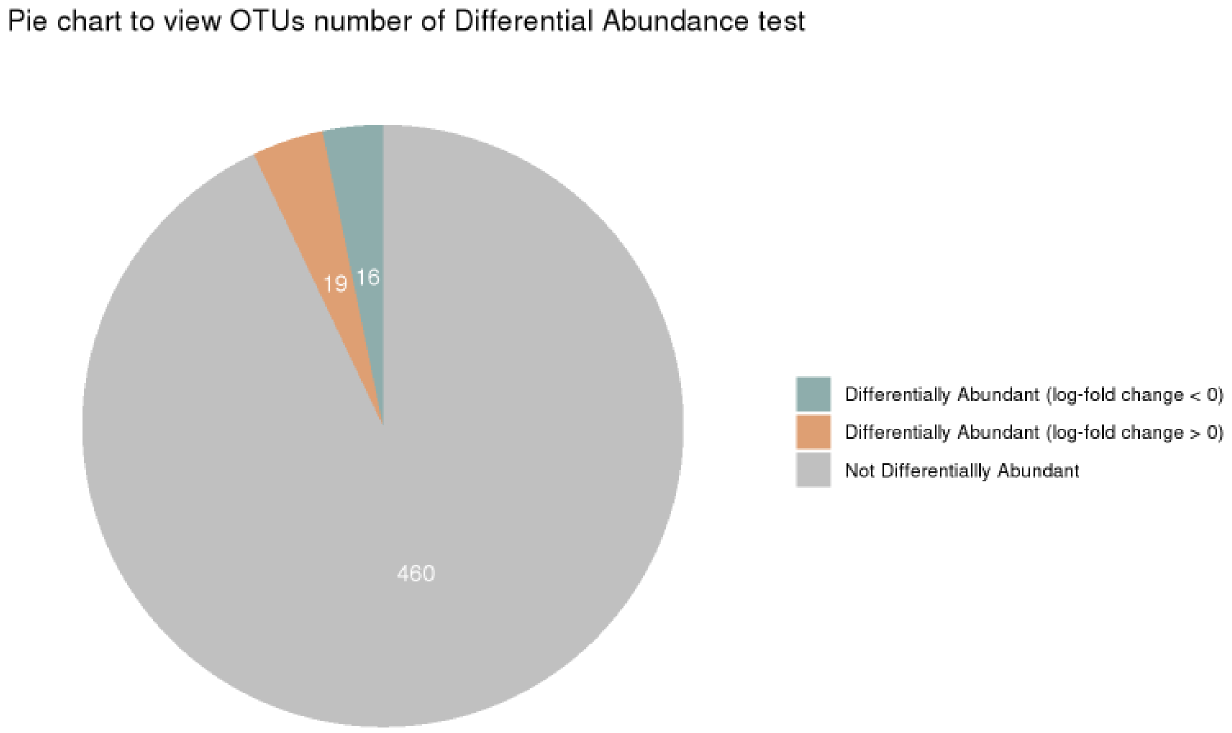
Differential abundance visualisation



Pie chart to view OTUs number of Differential Abundance test

Differential abundance visualisation

Pie chart

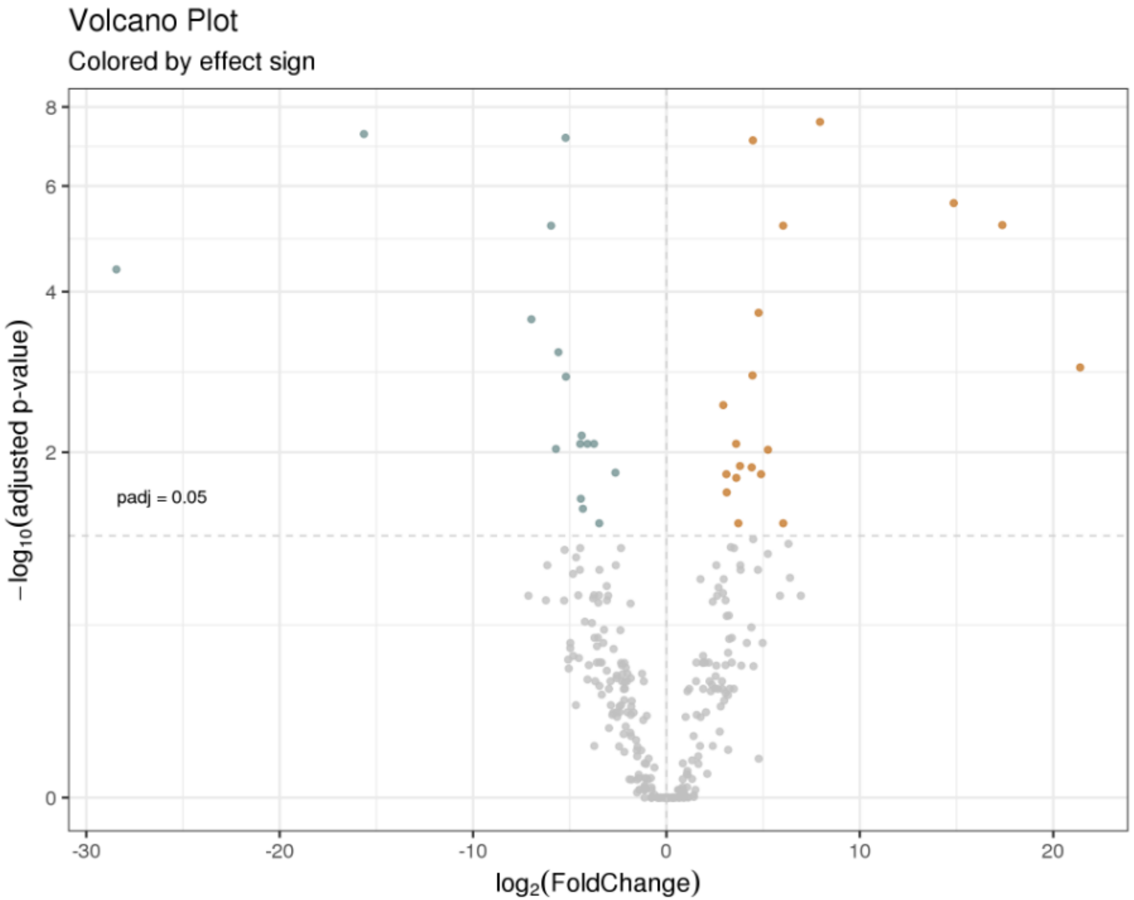


Pie chart to view OTUs number of Differential Abundance test

Most of the OTUs are not significantly affected between the conditions

35 OTUs are significantly affected between conditions

Differential abundance visualisation

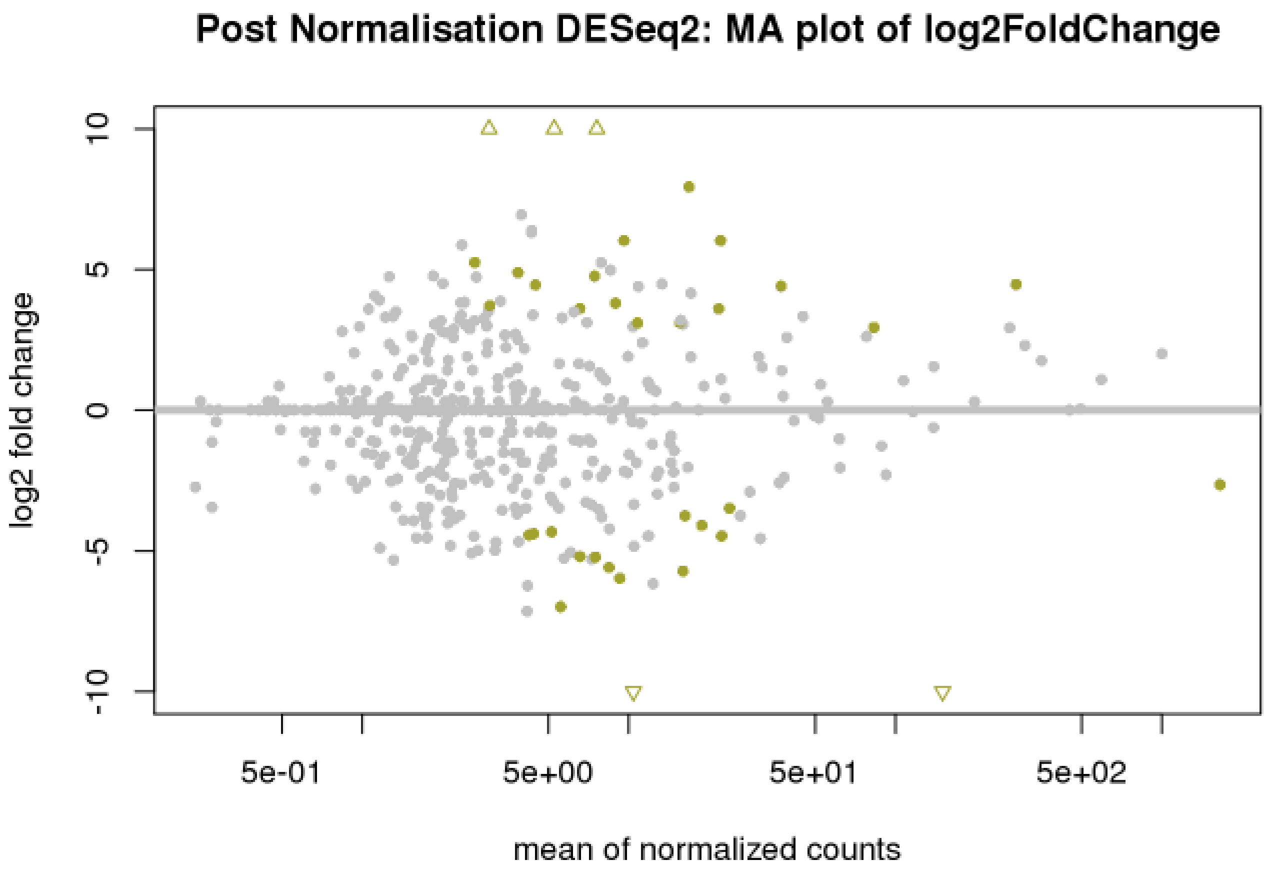


Visualisation of OTUs log2FoldChange and their associated adjusted p-values

Only OTUs with a significant adjusted p-value are colored

Differential abundance visualisation

MA plot



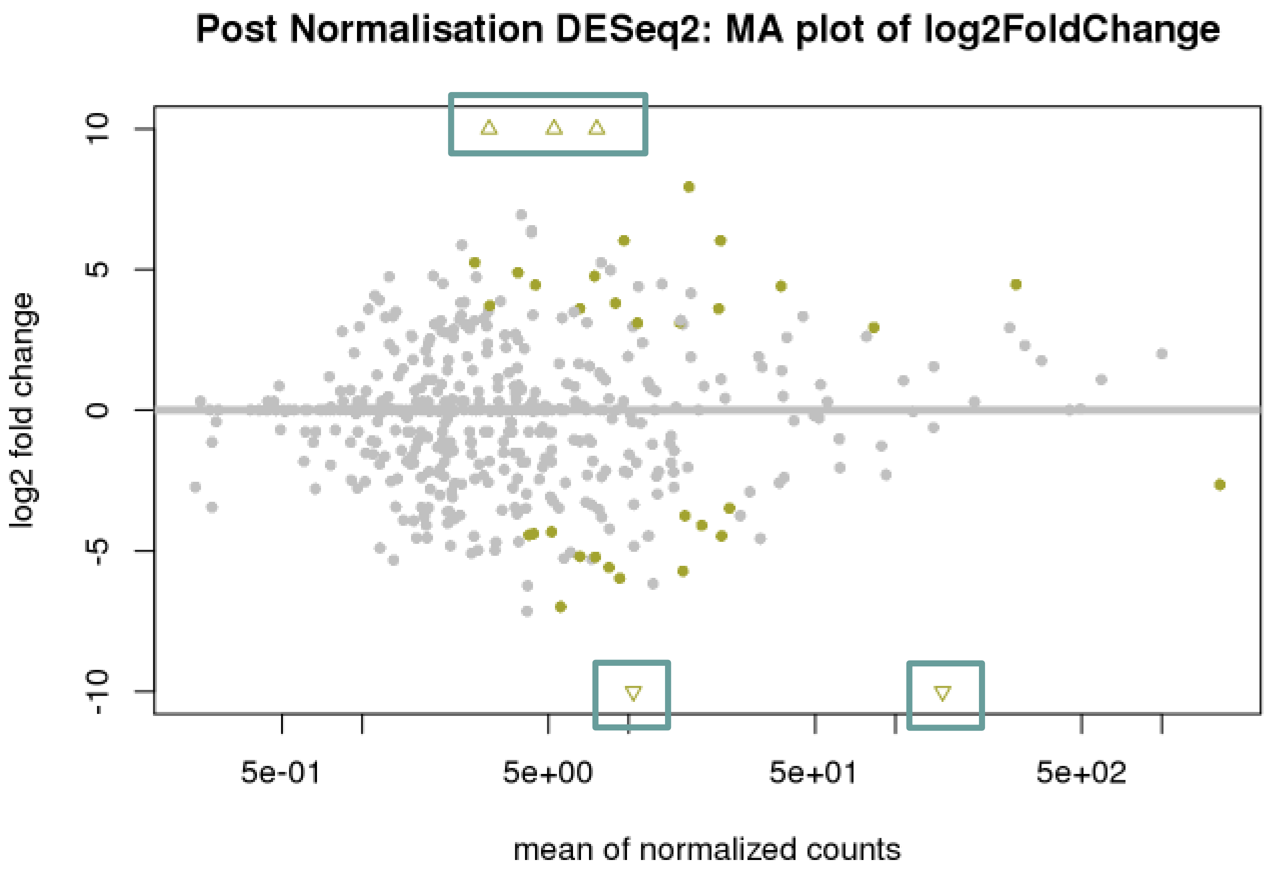
Visualisation of the relation between log2foldchange between conditions, and mean abundance of OTUs
(significantly affected OTUs are colored)

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

Colored OTUs on the left : affected rare OTUs

Differential abundance visualisation

MA plot

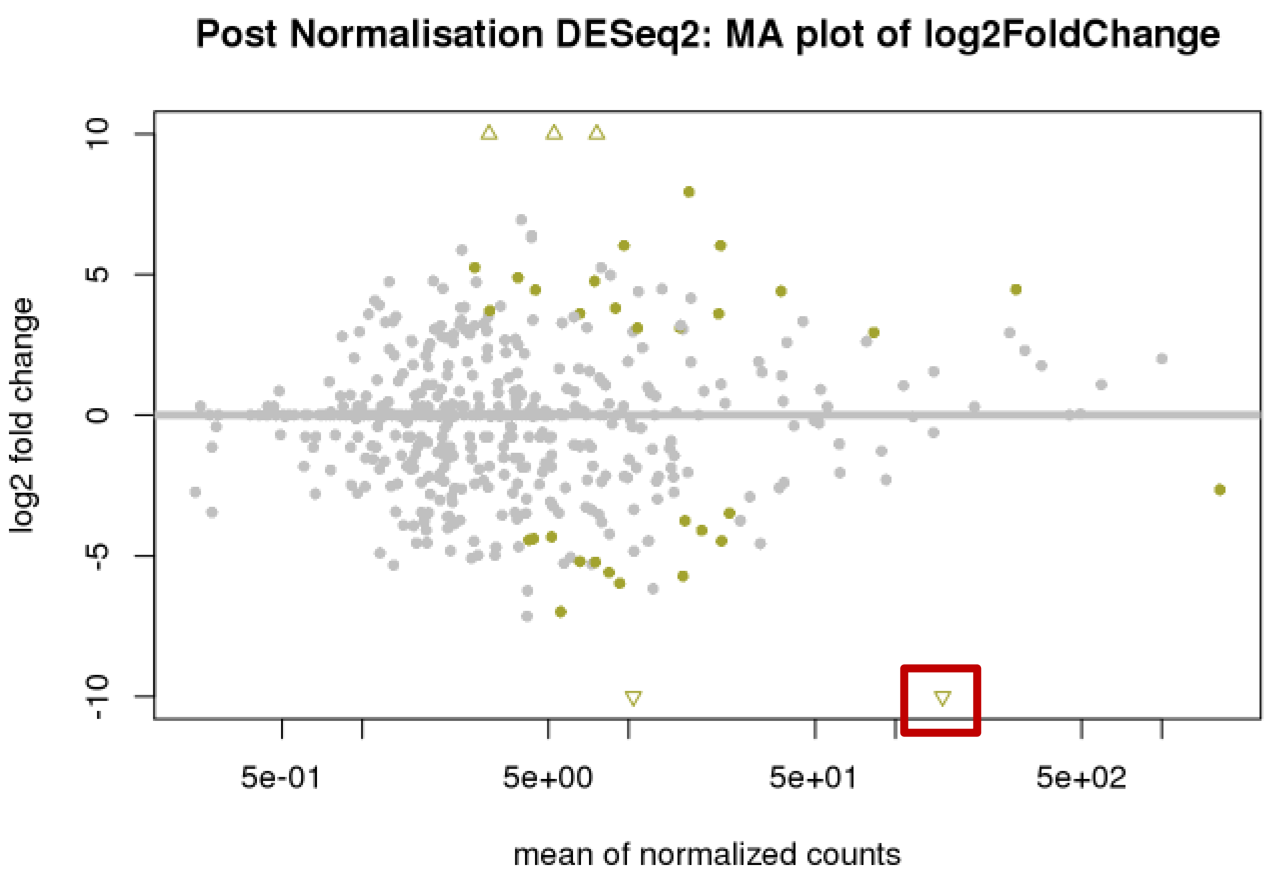


Visualisation of the relation between log2foldchange between conditions, and mean abundance of OTUs
(significantly affected OTUs are colored)

Triangles represent OTU out of scale

Differential abundance visualisation

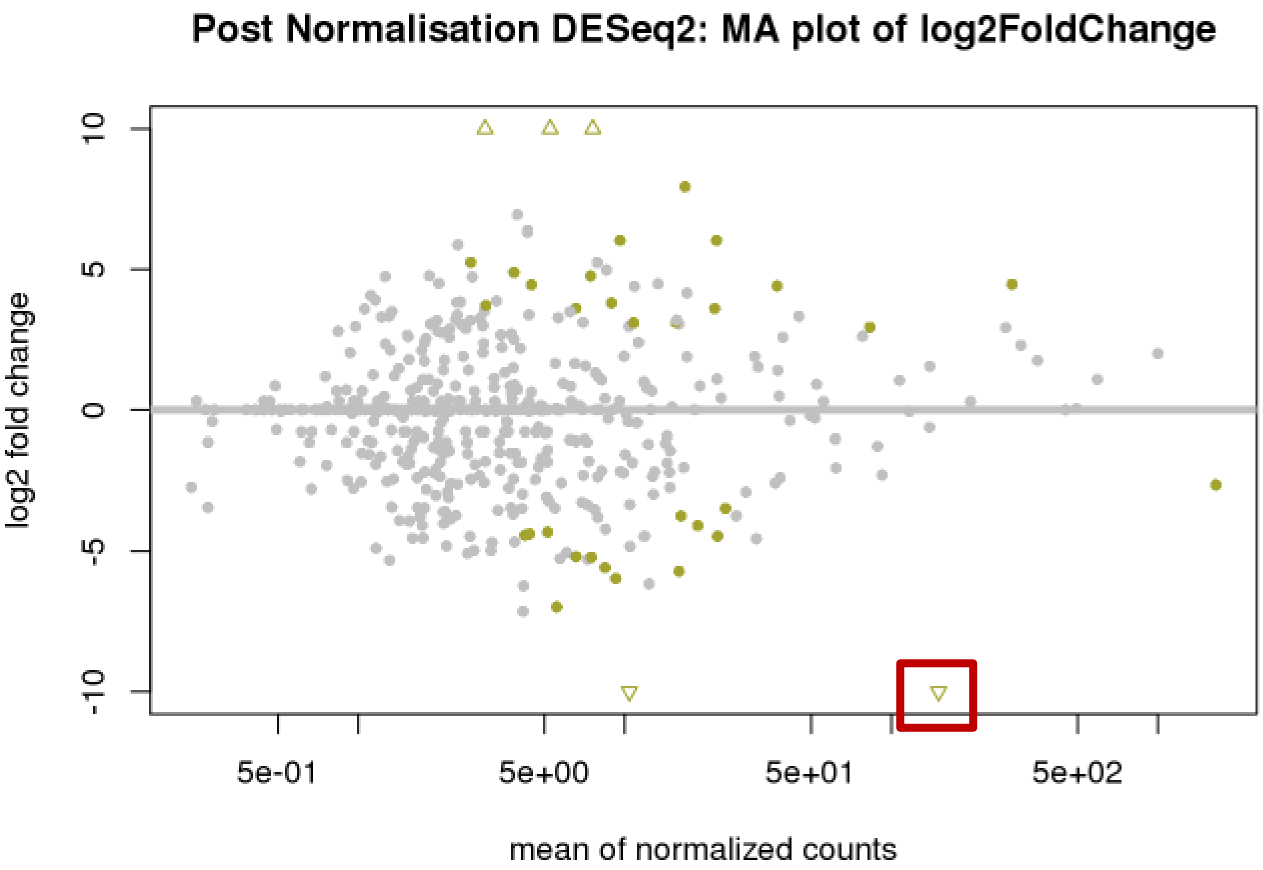
MA plot



➔ Which Cluster is the triangle spotted?

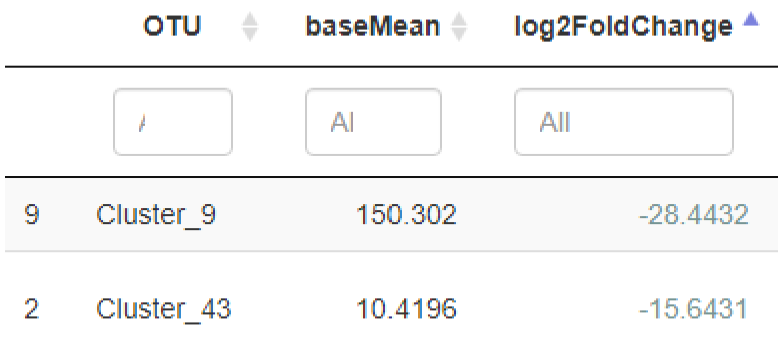
Differential abundance visualisation

MA plot



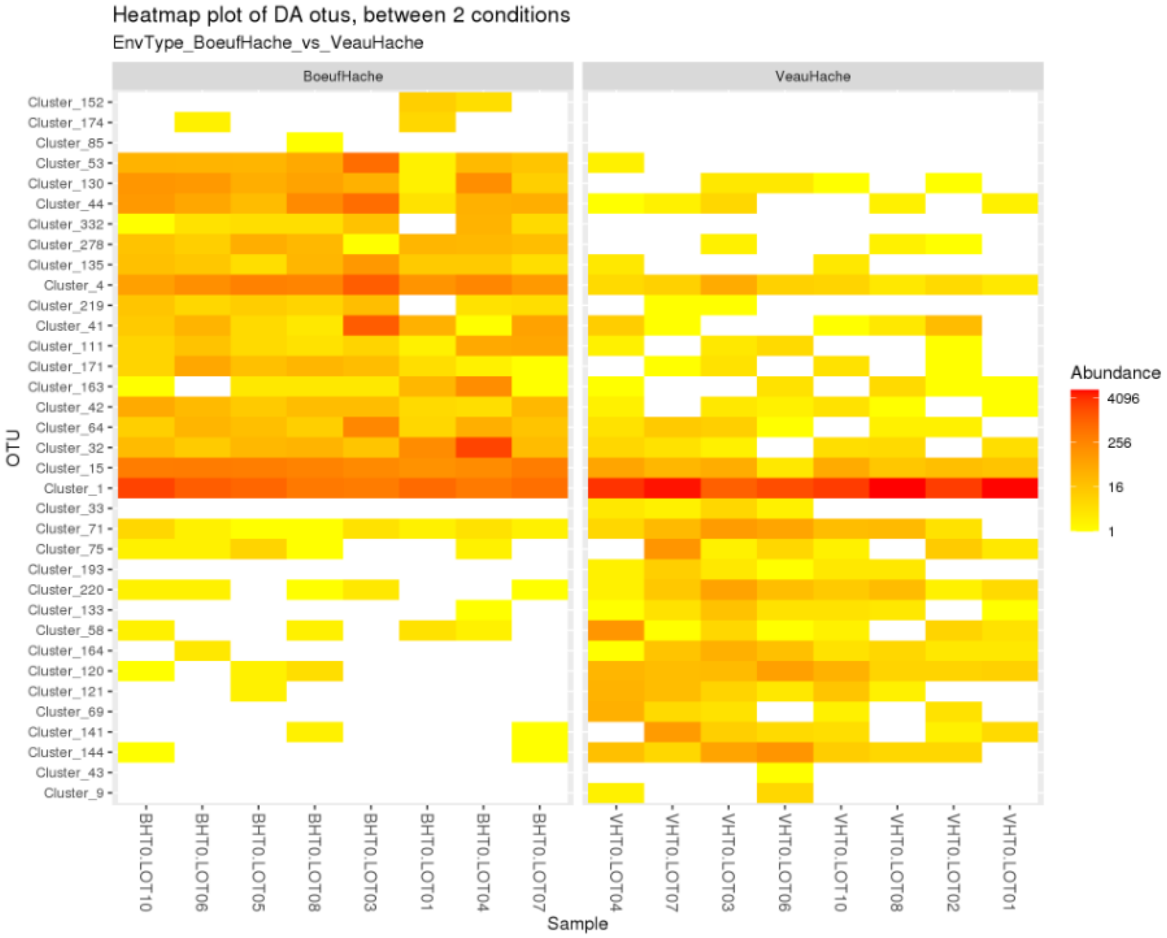
➔ Which Cluster is the triangle spotted?

It's Cluster_9 !



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

Differential abundance visualisation

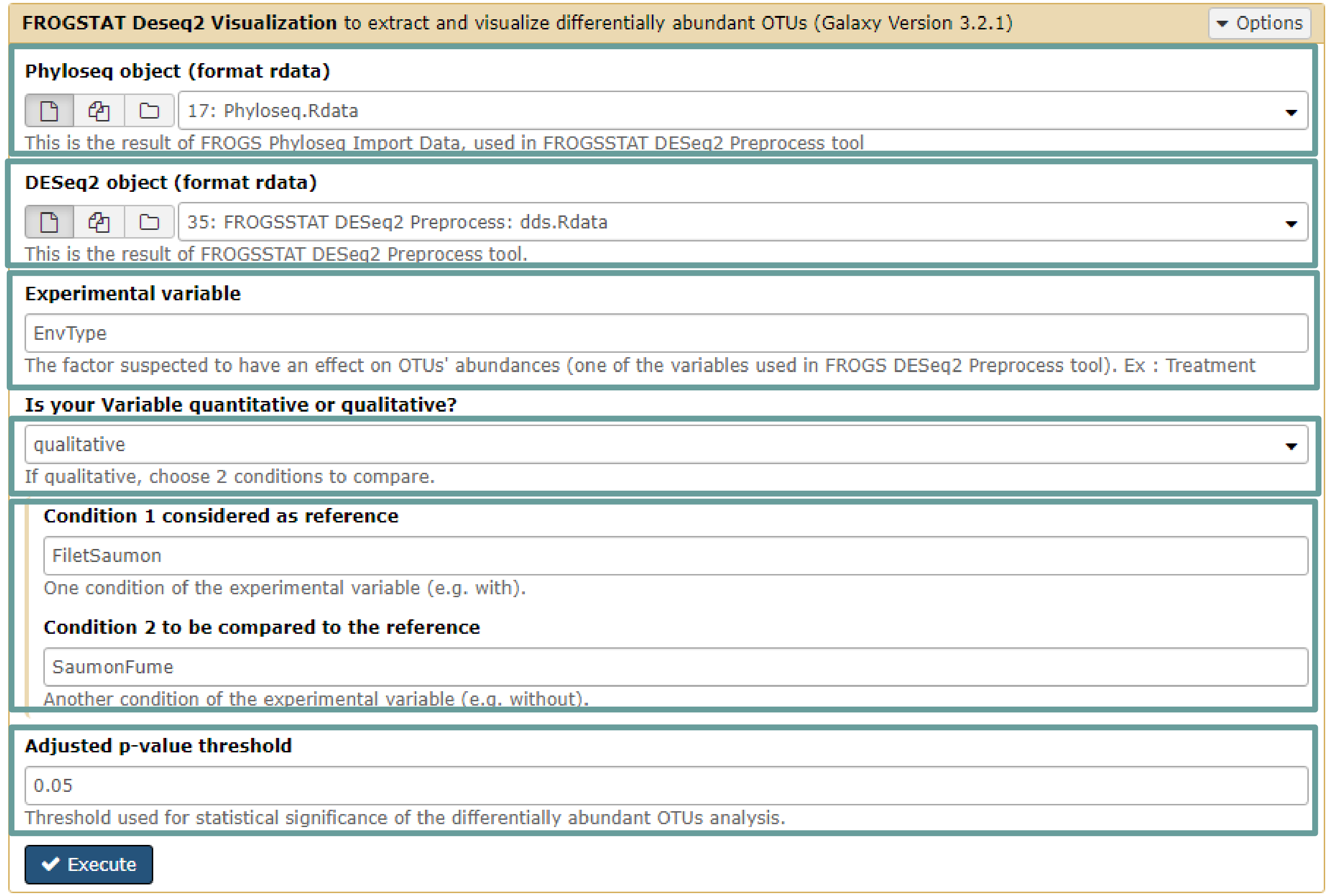


Heatmap plot

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

OTUs are ordered from top to bottom in descending order

Differential abundance visualisation



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

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).

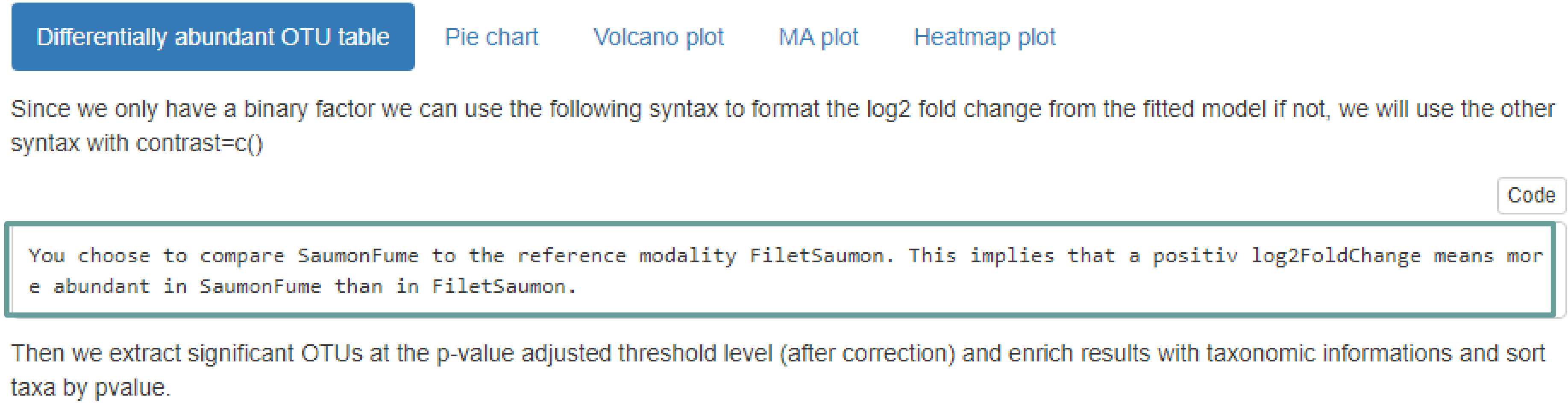
Adjusted p-value threshold

0.05

Threshold used for statistical significance of the differentially abundant OTUs analysis.

Compare FiletSaumon vs SaumonFume

Differential abundance visualisation



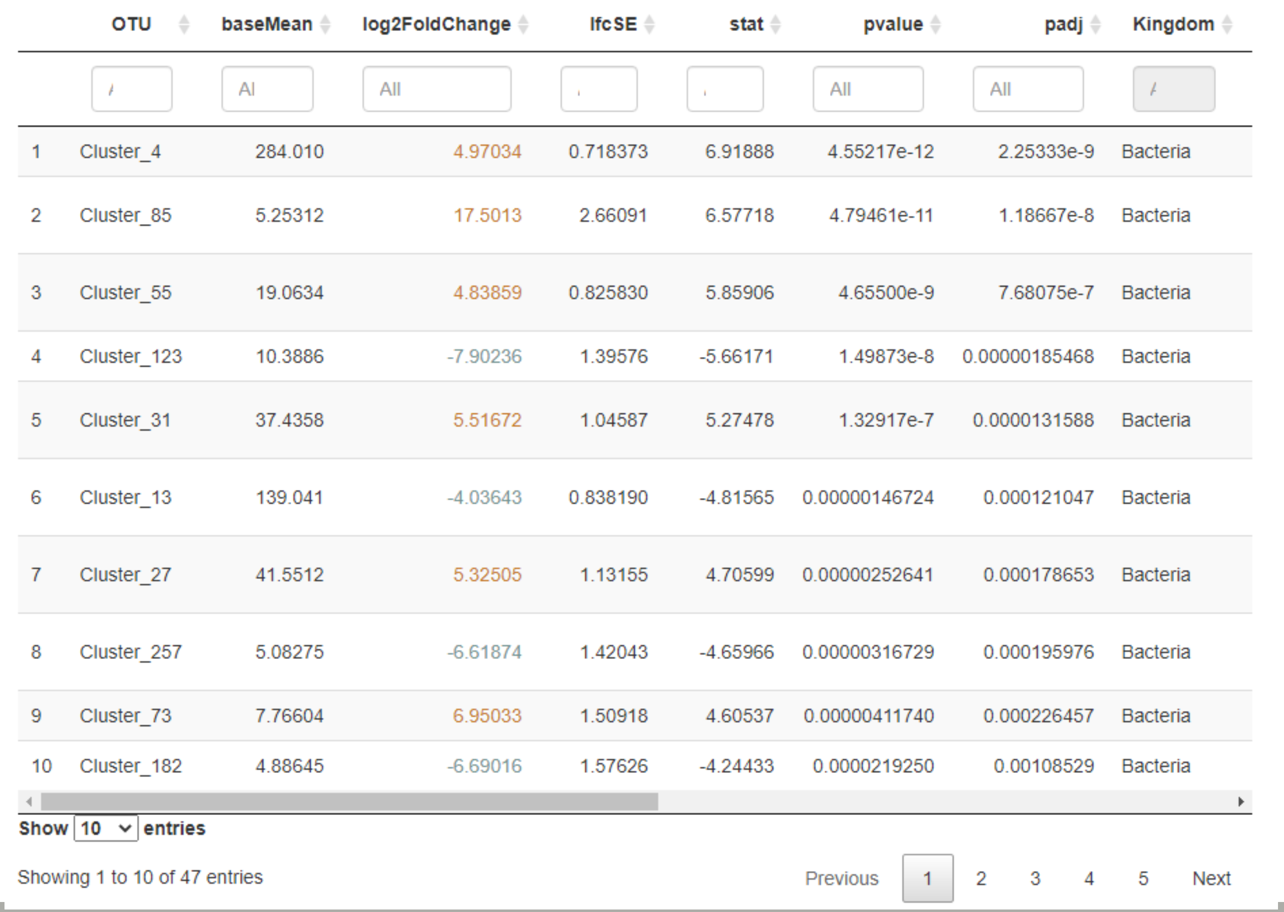
Since we only have a binary factor we can use the following syntax to format the log2 fold change from the fitted model if not, we will use the other syntax with contrast=c()

```
You choose to compare SaumonFume to the reference modality FiletSaumon. This implies that a positiv log2FoldChange means mor
e abundant in SaumonFume than in FiletSaumon.
```

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 visualisation

Diferentially abundant OTU table



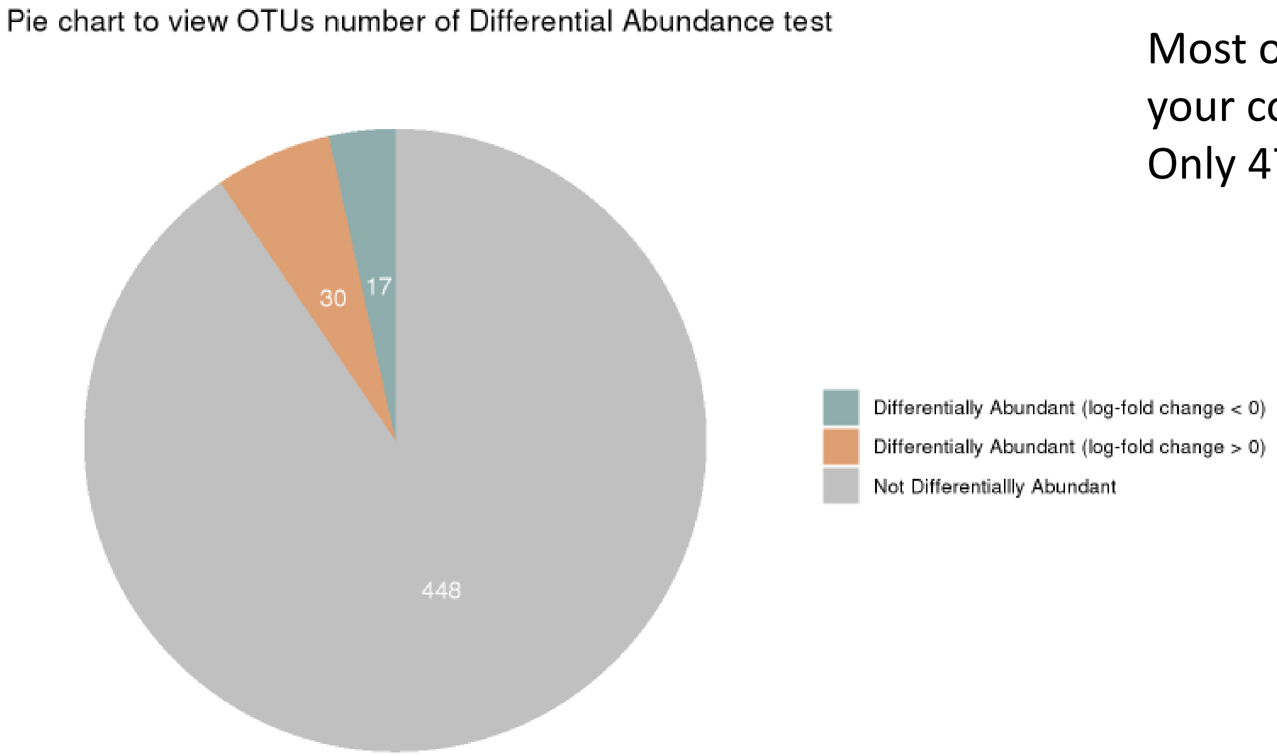
	OTU	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Kingdom
	/	Al	All	,	,	All	All	/
1	Cluster_4	284.010	4.97034	0.718373	6.91888	4.55217e-12	2.25333e-9	Bacteria
2	Cluster_85	5.25312	17.5013	2.66091	6.57718	4.79461e-11	1.18667e-8	Bacteria
3	Cluster_55	19.0634	4.83859	0.825830	5.85906	4.65500e-9	7.68075e-7	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
7	Cluster_27	41.5512	5.32505	1.13155	4.70599	0.00000252641	0.000178653	Bacteria
8	Cluster_257	5.08275	-6.61874	1.42043	-4.65966	0.00000316729	0.000195976	Bacteria
9	Cluster_73	7.76604	6.95033	1.50918	4.60537	0.00000411740	0.000226457	Bacteria
10	Cluster_182	4.88645	-6.69016	1.57626	-4.24433	0.0000219250	0.00108529	Bacteria

Show 10 entries

Showing 1 to 10 of 47 entries

Differential abundance visualisation

Pie chart

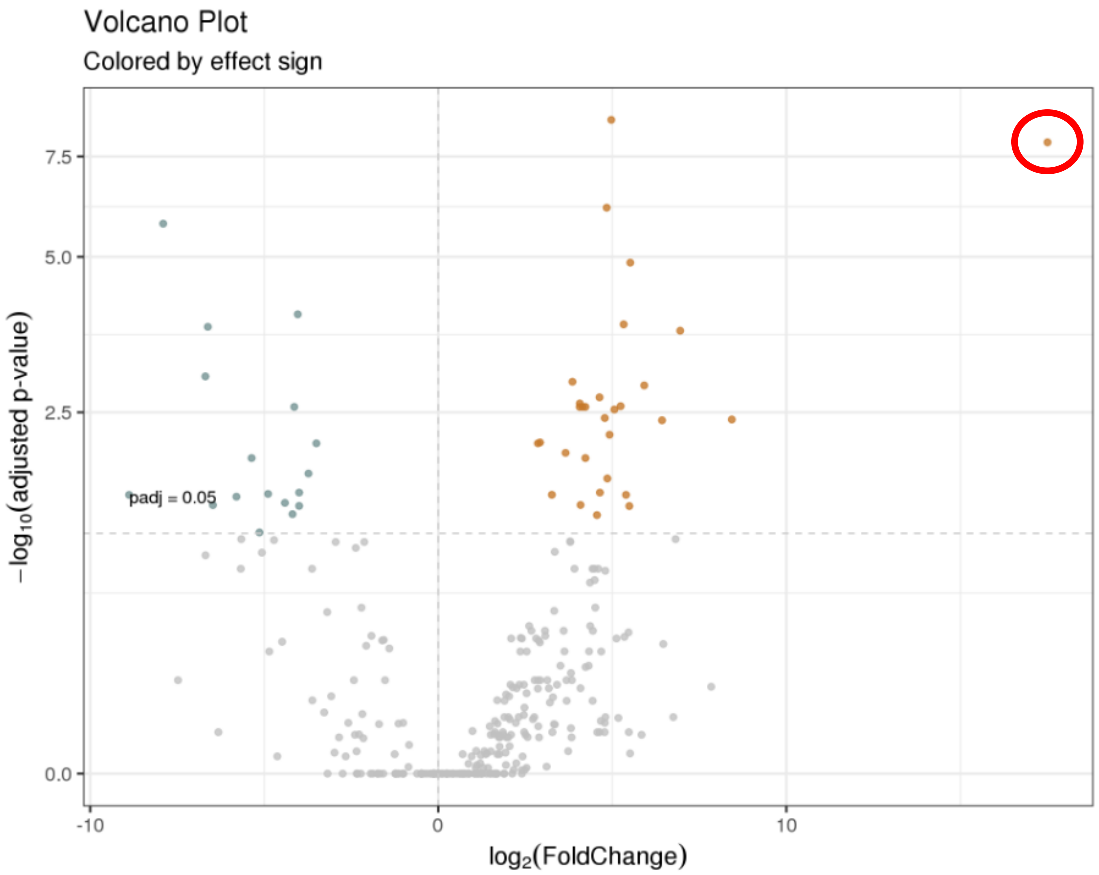


Most of the OTU are not significantly affected between your conditions

Only 47 OTUs are significantly affected between conditions

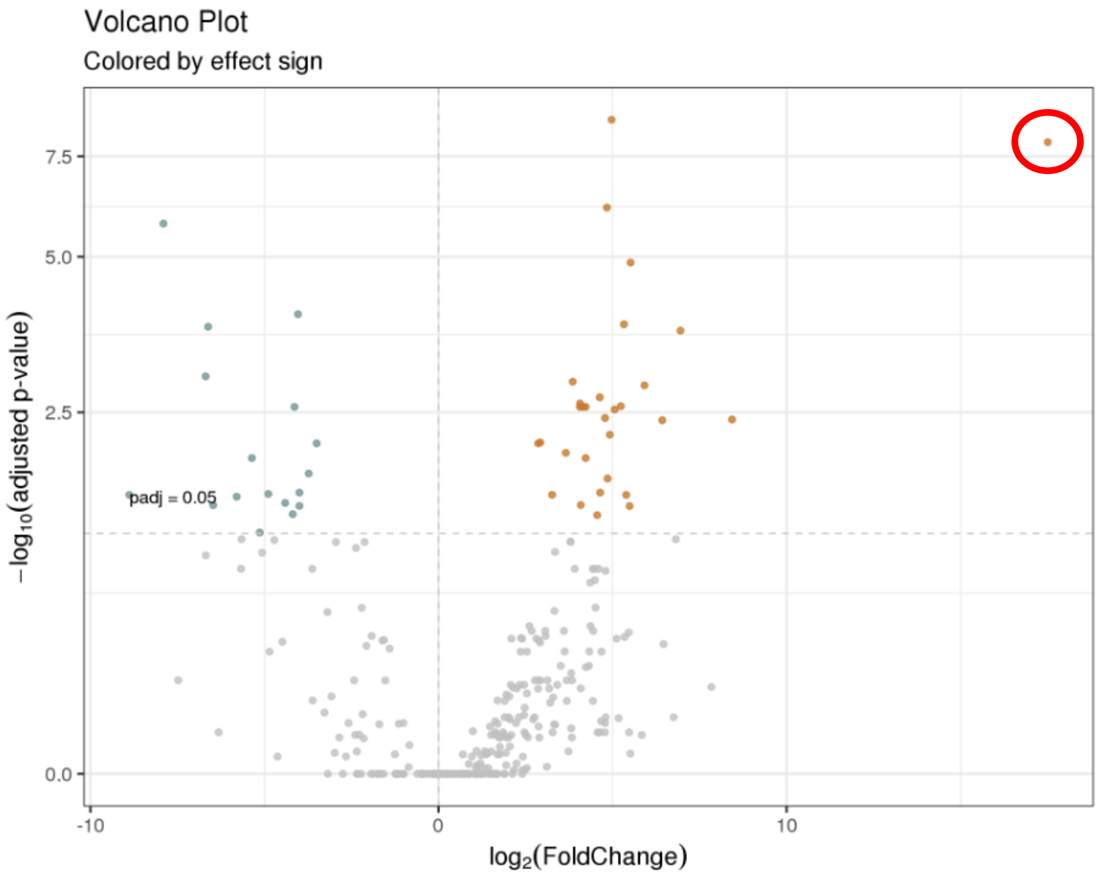
Differential abundance visualisation

Volcano plot



➔ Which Cluster is it ?

Differential abundance visualisation

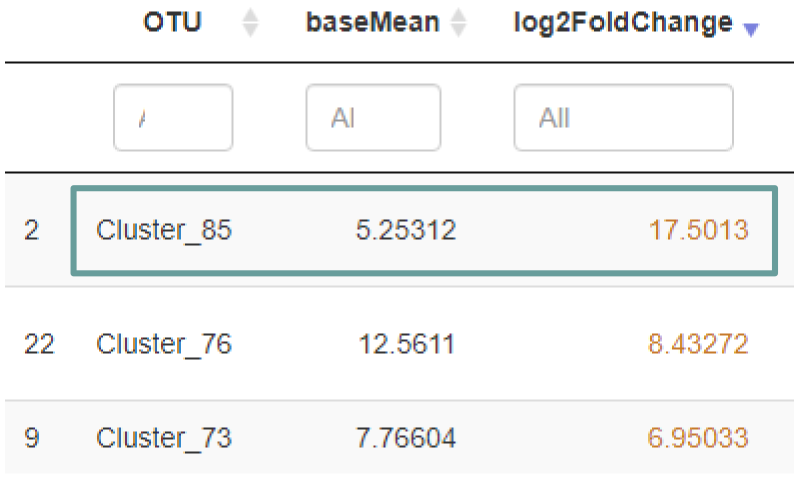


Volcano Plot

Colored by effect sign

Volcano plot

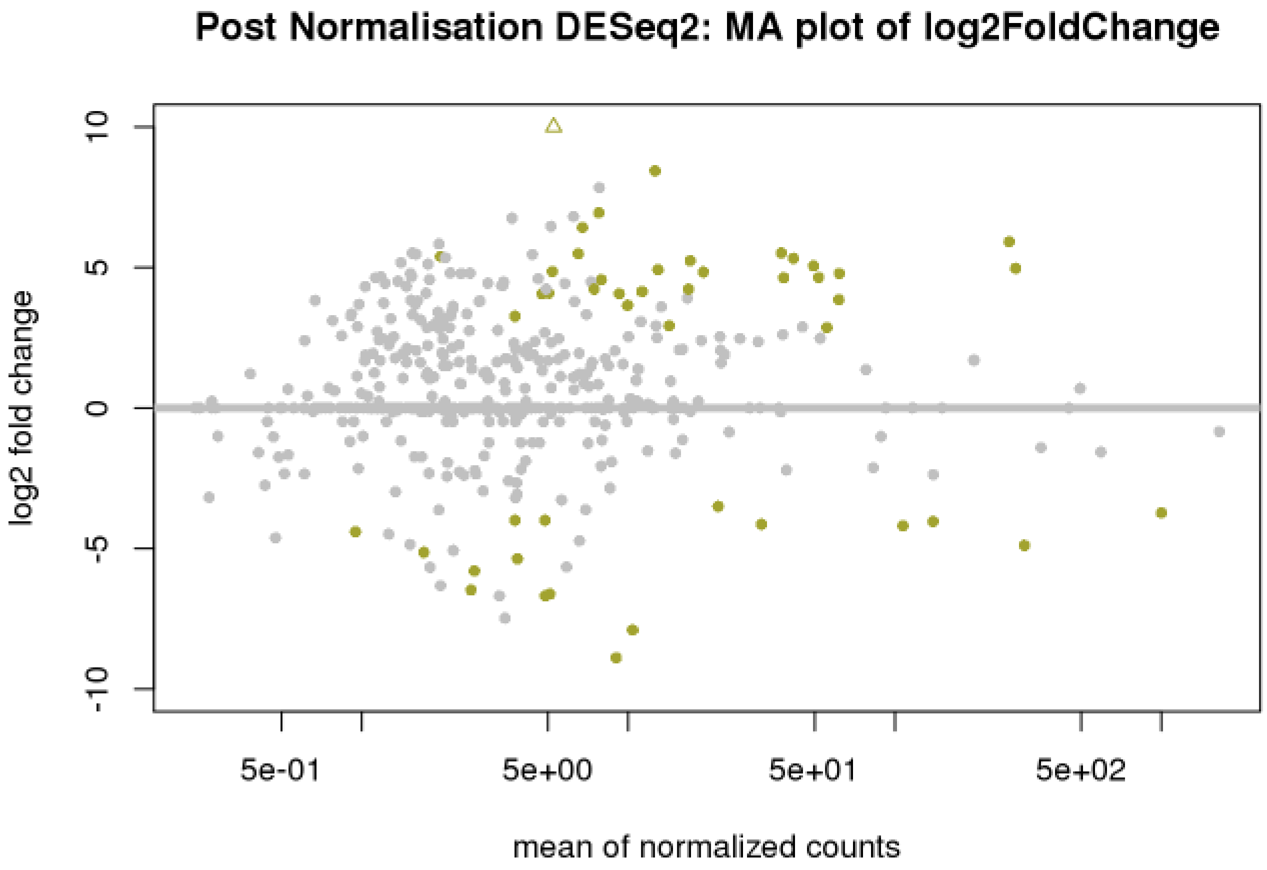
➔ Which Cluster is it ?



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

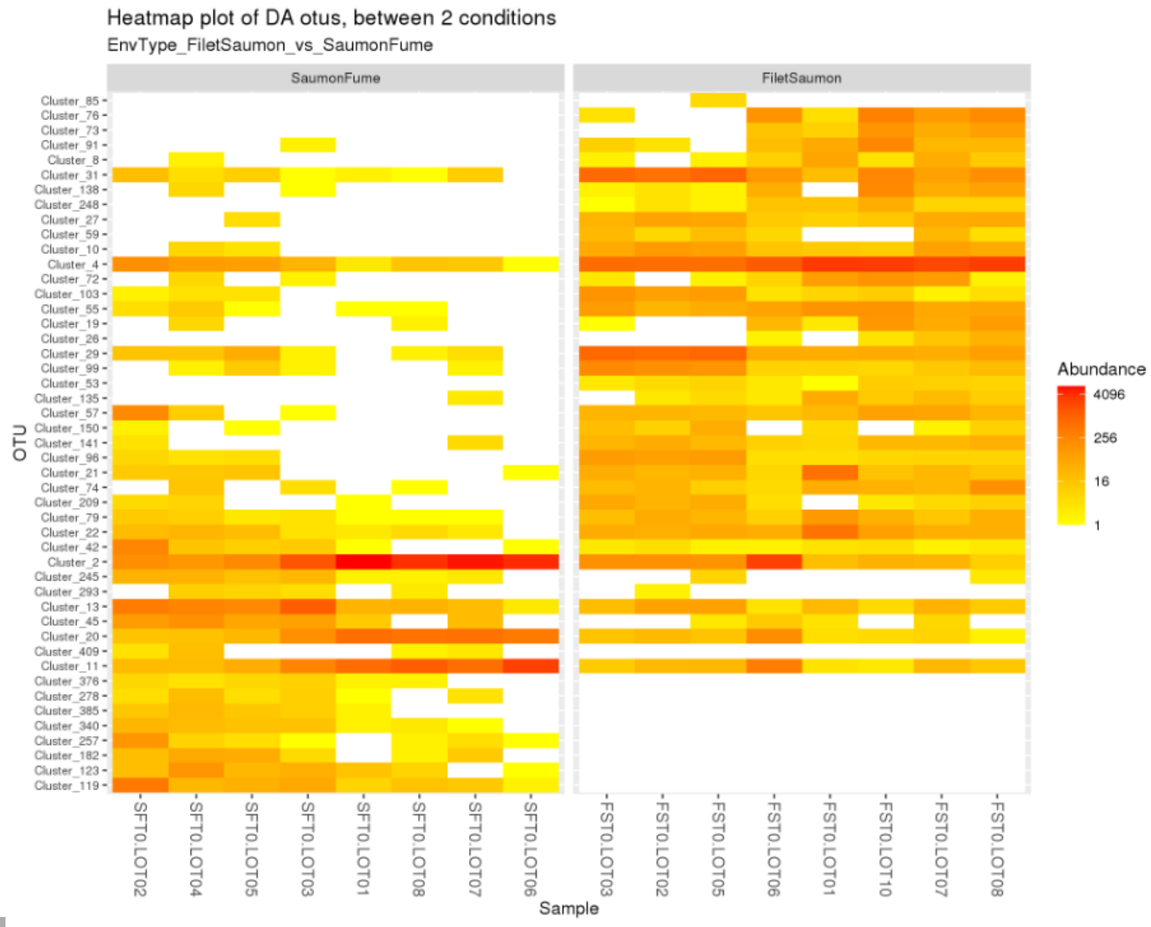
Differential abundance visualisation

MA plot

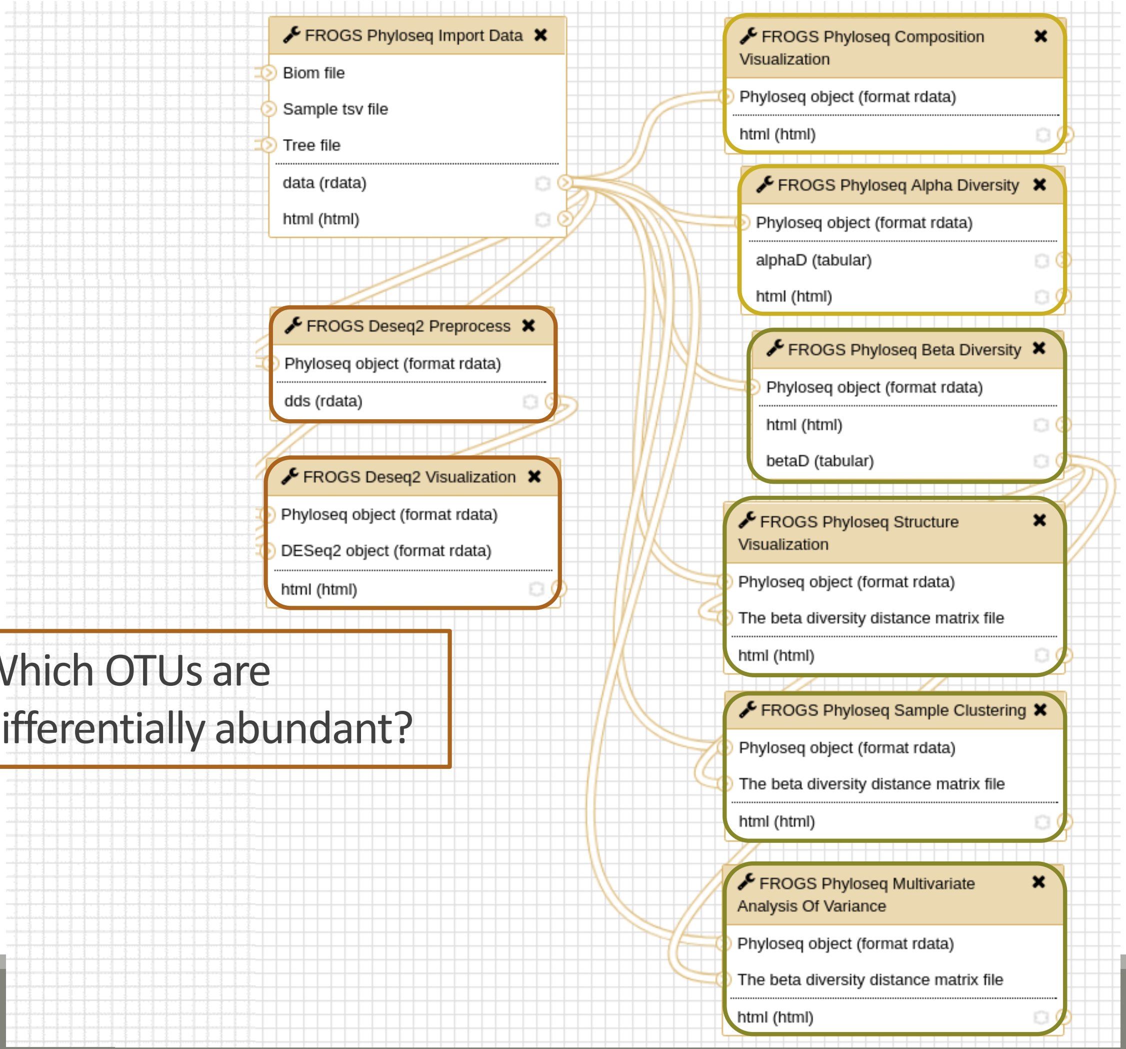


Differential abundance visualisation

Heatmap plot



FROGSStat Summary


FROGS Phyloseq Import Data
Biom file
Sample tsv file
Tree file
data (rdata)
html (html)
FROGS Deseq2 Preprocess
Phyloseq object (format rdata)
dds (rdata)
FROGS Deseq2 Visualization
Phyloseq object (format rdata)
DESeq2 object (format rdata)
html (html)
FROGS Phyloseq Composition Visualization
Phyloseq object (format rdata)
html (html)
FROGS Phyloseq Alpha Diversity
Phyloseq object (format rdata)
alphaD (tabular)
html (html)
FROGS Phyloseq Beta Diversity
Phyloseq object (format rdata)
html (html)
betaD (tabular)
FROGS Phyloseq Structure Visualization
Phyloseq object (format rdata)
The beta diversity distance matrix file
html (html)
FROGS Phyloseq Sample Clustering
Phyloseq object (format rdata)
The beta diversity distance matrix file
html (html)
FROGS Phyloseq Multivariate Analysis Of Variance
Phyloseq object (format rdata)
The beta diversity distance matrix file
html (html)

Which OTUs are differentially abundant?

What is the sample composition ?

What are the sample diversities ?

Composition analysis

What is the samples dissimilarity ?

Is there any relation between species or communities?

how do the communities cluster?

Which variable influence the diversity ?

Structure analysis

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 meaningful
- Keep in mind that :
  - Phyloseq composition and structure analyses need to be perform on normalised (=rarefied) counts
  - Different indices or distance methods will give different but complementary information
  - Test different distances and choose which one fits better your data
  - Richness indices are highly dependent on rare OTUs
  - DESeq analysis need to be performed on counts without normalisation

Annexes

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.