Training on Galaxy: Statistics to explore metagenomics

July 2019

Find, Rapidly, OTUs with Galaxy Solution

MARIA BERNARD, GÉRALDINE PASCAL, MAHENDRA MARIADASSOU, LAURENT CAUQUIL, STEPHANE CHAILLOU

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)?

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, picante

▪ Tree manipulation from ape

▪ Graphics from ggplot2

▪ (Differential analysis from DESeq2)

Overview

1. Part A: We play together on a first dataset

2. Part B: You play alone with our guideline on a 2nd dataset

PART A

Training Data1

A real analysis provided by Stéphane Chaillou *et al.*

Comparison of meat and seafood bacterial communities.

8 environment types (EnvType) :
- Meat → Ground Beef, Ground veal, Poultry sausage, Diced bacon
- Seafood → Cooked schrimps, Smoked salmon, Salmon filet, Cod filet

- 64 samples of 16S V1-V3
- Taxonomic affiliations was made with the Greengenes database

Chaillou, S. et al (2015). Origin and ecological selection of core and food-specific bacterial communities associated with meat and seafood spoilage. ISME J, 9(5):1105-1118.

Exercise A-1

1. Create a new history : « food »

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

➔ What supplementary data do we need to perform statistical analysis ?

Exercise A-1

1. Create a new history : « food »

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

➔ What supplementary data do we need to perform statistical analysis ?

2. Upload data
    1. chaillou/sample_metadata.tsv
    2. chaillou/chaillou.biom
    3. chaillou/tree.nwk
        (datatype nhx)

➔ Take a look at the data

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

Exercise A-1

➔ How many OTUs do we have here ?

➔ How many taxonomic levels do we have here?

Exercise A-1

➔ How many OTUs do we have here ?

➔ How many taxonomic levels do we have here?

15: FROGS Clusters stat: summary.html 👁 ✏ ✖

16: FROGS BIOM to TSV: abundance.tsv 👁 ✏ ✖

Clusters
508

Sequences
753,452

```
#taxonomy                                                                                                      observation_name
k__Bacteria;p__Tenericutes;c__Mollicutes;o__Mycoplasmatales;f__Mycoplasmataceae;g__Candidatus Lumbricincola;s__NA    otu_01778
k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella;s__NA                  otu_01838
k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Dyella;s__Ginsengisoli   otu_01386
```

Data import tool

PHYLOSEQ OBJECT CREATION

Phyloseq : Data import

The FROGS biom format contains:
- OTU count tables (required)
- OTU description : taxonomy

Others informations used in FROGSSTAT are:
- sample description in TSV file
- phylogenetic tree in Newick format (nwk or nhx)

FROGSSTAT Phyloseq Import Data from 3 files: biomfile, samplefile, treefile (Galaxy Version 1.0.0) ▼ Options

Biom file

2: chaillou.biom

The file contains the OTU informations (format: biom1).

Sample tsv file

3: sample_metadata.tsv

The file contains the samples informations (format: tabular).

Tree file

1: 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 analysis.

✔ Execute

Exercise A-2

1. Use FROGSSTAT Phyloseq Import Data, with and without samples normalization (rename datasets in consequence).
   ➜ What is the difference ?

2. Guess what is a Rdata file?

3. Explore the HTML results

FROGS Phyloseq: Import Data

Phyloseq 1.20.0

| Summary | Ranks Names | Sample metadata | Plot tree | R code |

```
phyloseq-class experiment-level object
otu_table()   OTU Table:          [ 508 taxa and 64 samples ]
sample_data() Sample Data:        [ 64 samples by 4 sample variables ]
tax_table()   Taxonomy Table:     [ 508 taxa by 7 taxonomic ranks ]
phy_tree()    Phylogenetic Tree:  [ 508 tips and 507 internal nodes ]
```

```
Number of sequences in each sample after normalization:  11718
```

Exercise A-2

3. Explore the HTML results

Phylogenetic tree colored by Phylum
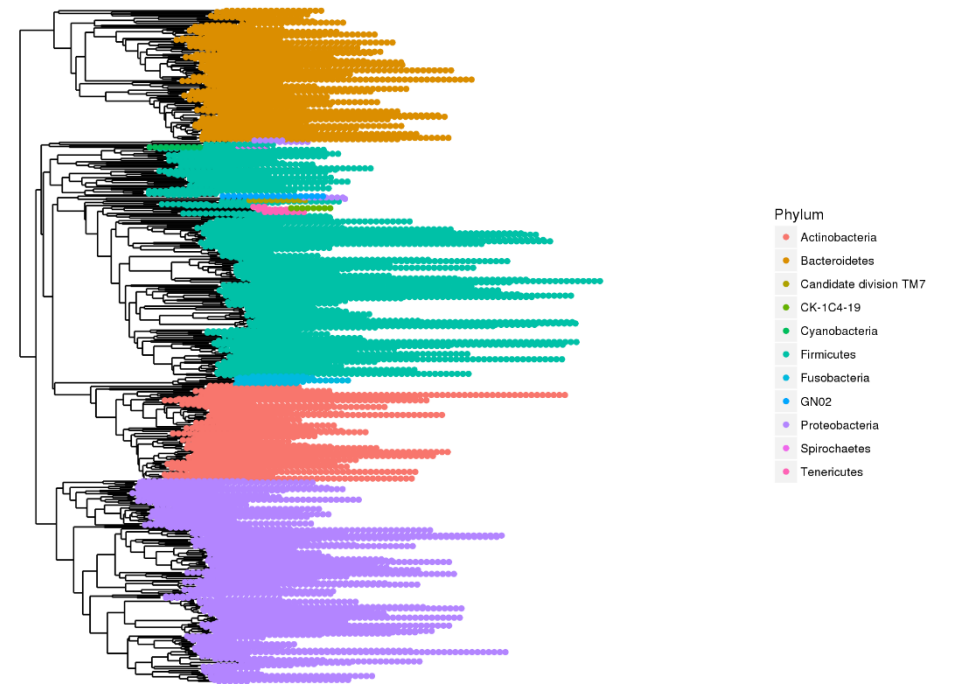
Exercise A-2

3. Explore the HTML results

FROGS Phyloseq: Import Data

Phyloseq 1.20.0

Summary   Ranks Names   **Sample metadata**   Plot tree   R code

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

```
EnvType :  BoeufHache, VeauHache, DesLardons, SaucisseVolaille,
Crevette, SaumonFume, FiletSaumon, FiletCabillaud

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

FoodType :  Meat, Seafood
```

FROGS Phyloseq: Import Data

Phyloseq 1.20.0

Summary   Ranks Names   Sample metadata   Plot tree   **R code**

Loading packages

```
library(phyloseq)
library(ape)
library(ggplot2)
```

Warning !

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

So take extra care when you construct your sample_metadata file

Biodiversity analysis

Biodiversity analysis

1. Exploring the 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 1.0.0)   ▼ Options

Phyloseq object (format rdata)

8: food.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'

✔ Execute

Explore the sample raw count

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

For the first usage, let the default parameters, but :
▪ Take care of your taxonomic level name
▪ Is the Taxon « Bacteria » in your data ?

19

Exercise A-3

➔ Interpretations ?

▪ Firmicutes and Proteobacteria are presents in all samples, but with a wide range of abundance

▪ Meat type share common Phylum composition with a majority of Firmicutes

▪ Seafoods seem to be much more variable

FROGS Phyloseq: Visualize Data Composition

Phyloseq 1.20.0

Bar plot   Composition plot   R code

Bar plot colored by Phylum
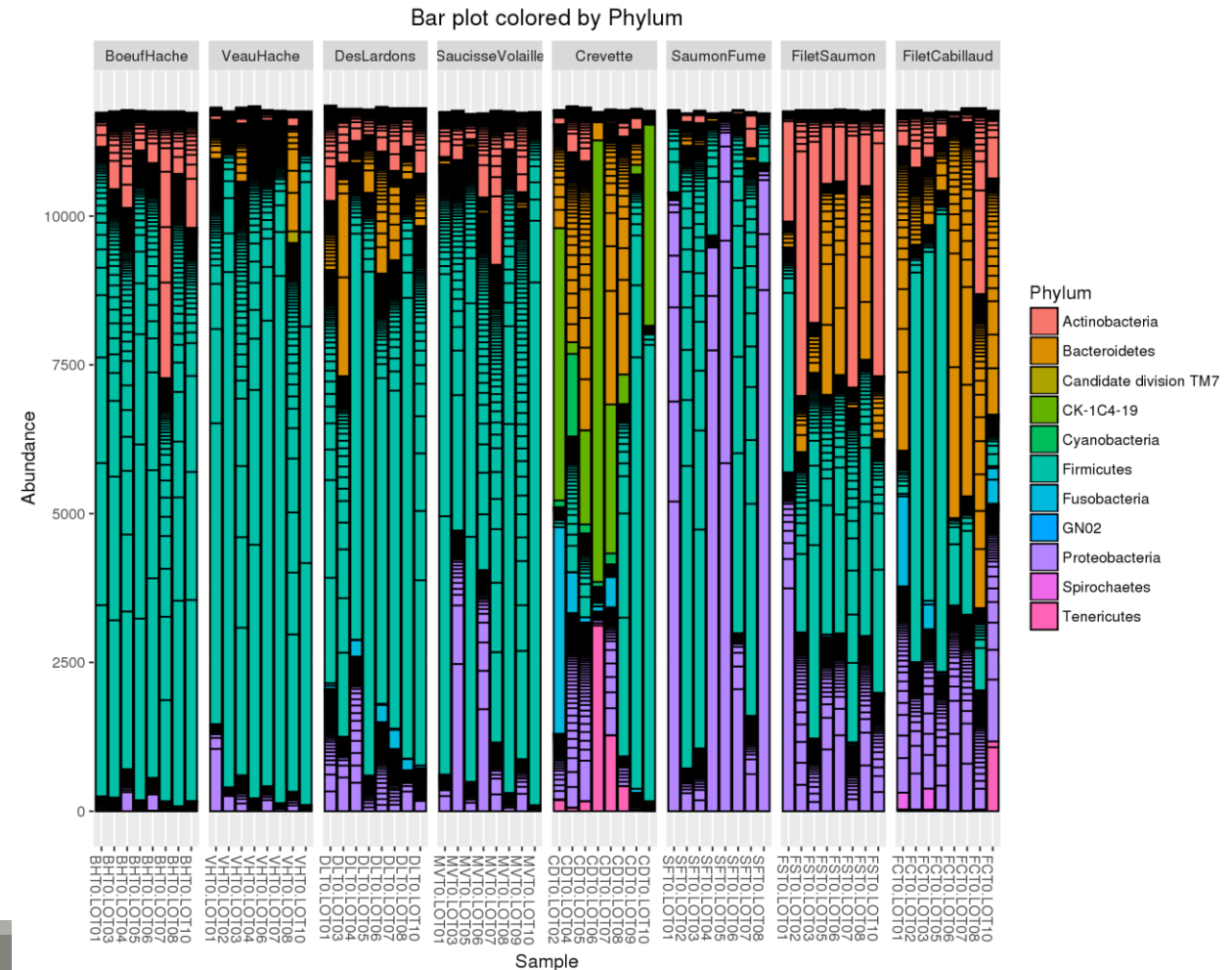
Exploring biodiversity : visualisation

➔ 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 normalized function)

Exploring biodiversity : visualisation

Customisation: plot_composition function :

- Works with relative abundances

- Subsets OTUs at a given taxonomic level

- Aggregates OTUs at another taxonomic level

- Shows only a given number of OTUs

Taxonomic level to filter your data

Kingdom

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

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

Bacteria

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

Taxonomic level used for aggregation

Phylum

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

Number of most abundant taxa to keep

9

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

Exercise A-4

Look at the « Composition plot » tab

Based on these results what would be interesting to look into ?

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

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

FROGS Phyloseq: Visualize Data Composition

Phyloseq 1.20.0

Bar plot | Composition plot | R code

Composition within Bacteria (9 top Phylum)

Phylum
- Actinobacteria
- Bacteroidetes
- CK-1C4-19
- Cyanobacteria
- Firmicutes
- Fusobacteria
- Proteobacteria
- Spirochaetes
- Tenericutes
- Other
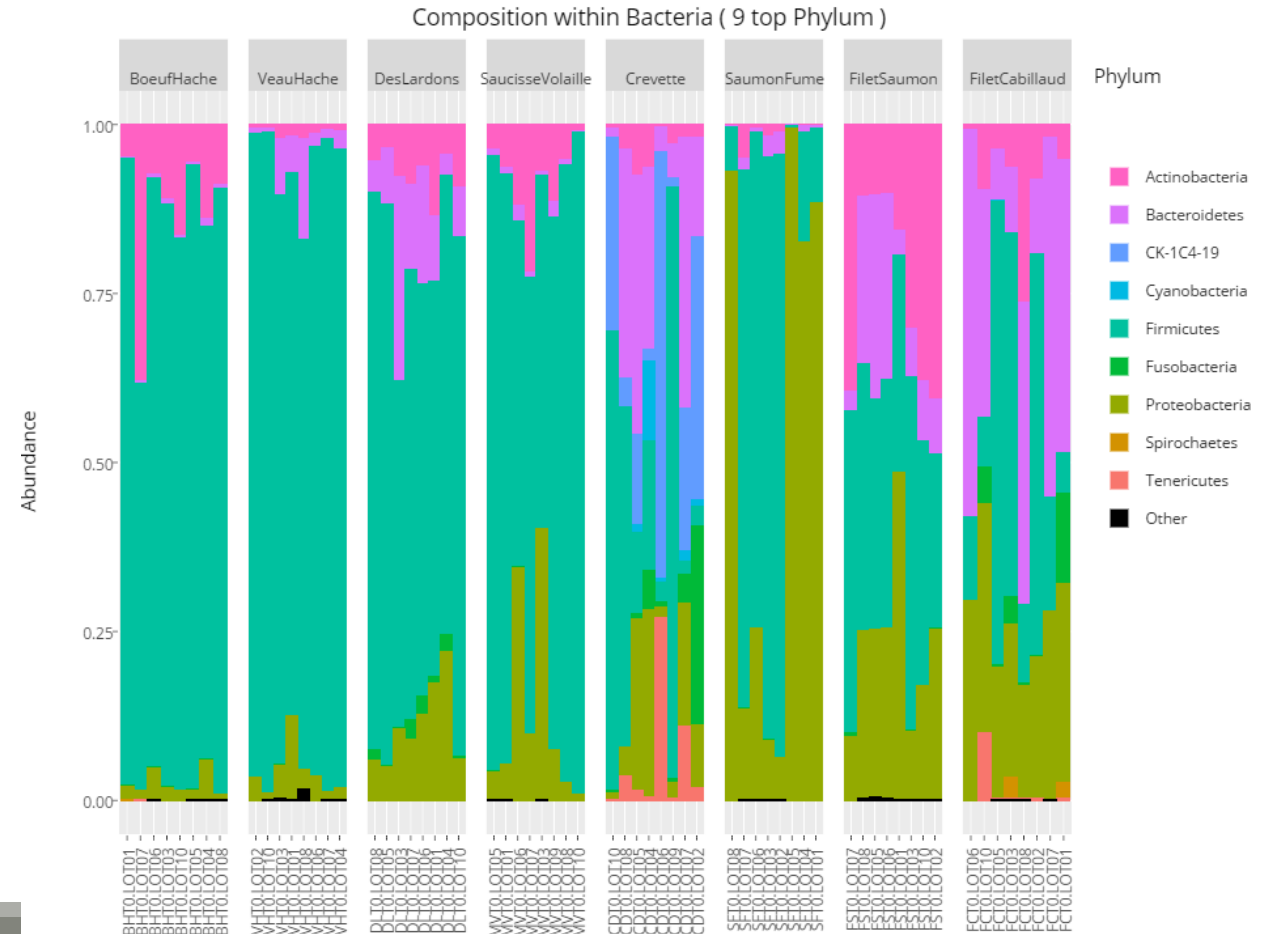
Exercise A-4

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)
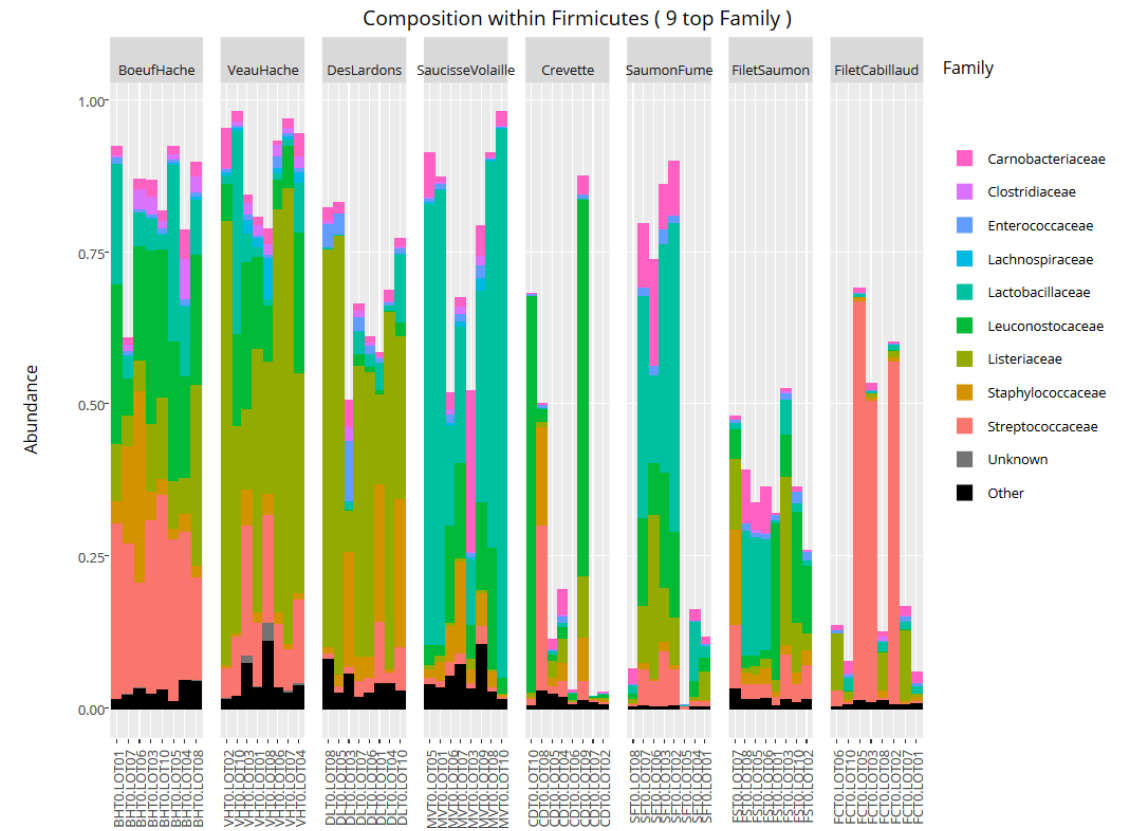
Exercise A-4

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'

Composition within Proteobacteria (9 top Family)
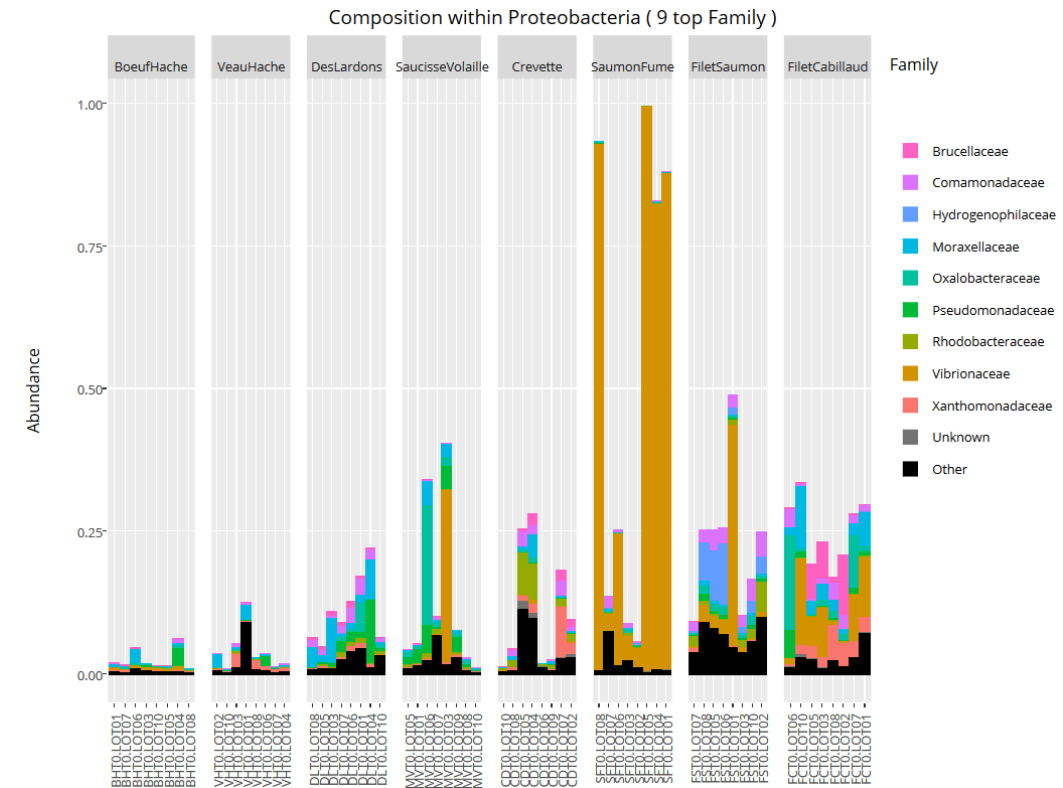
Exploring biodiversity : visualisation

Remark 1 : An example of what happens when sample_metadata file is not sorted in a meaning full way
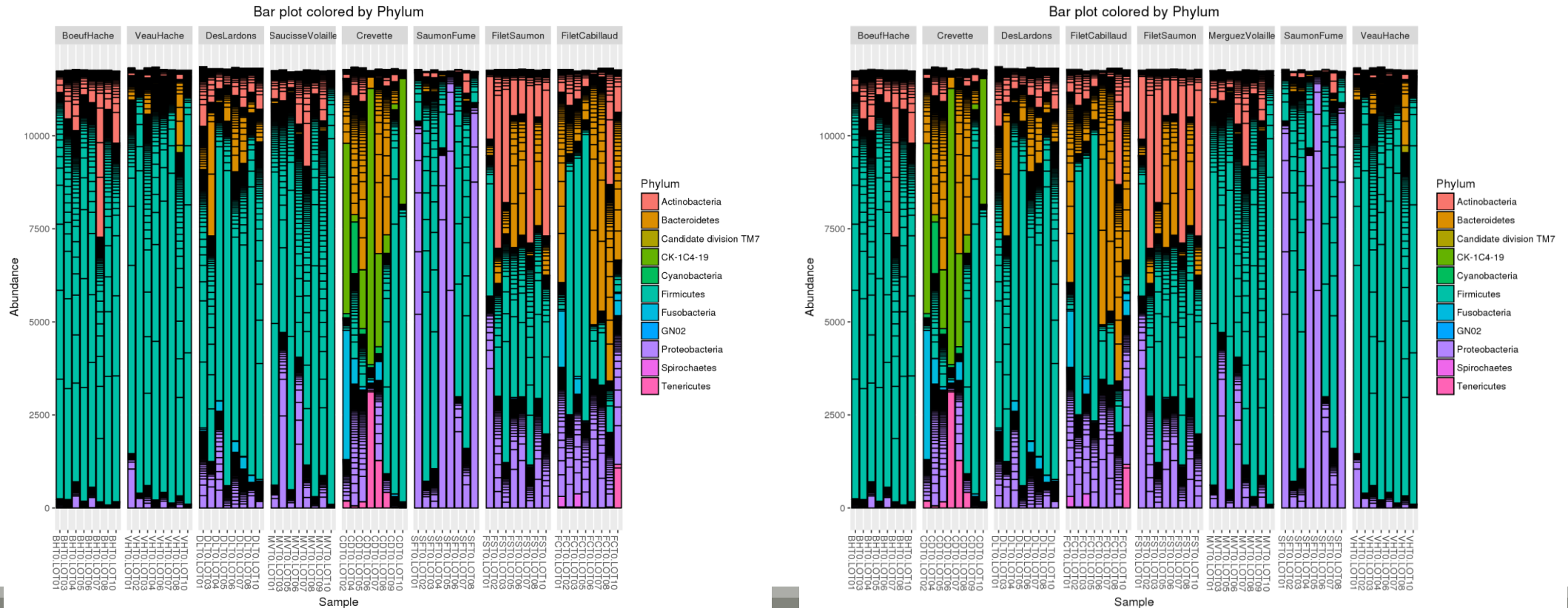
Exploring biodiversity : visualisation

Remark 2 : Keep in mind that human eye cannot distinguish more than 12 colours at the same time. Example of the 30 most abundant Families among Bacteria

Composition within Bacteria (30 top Family)
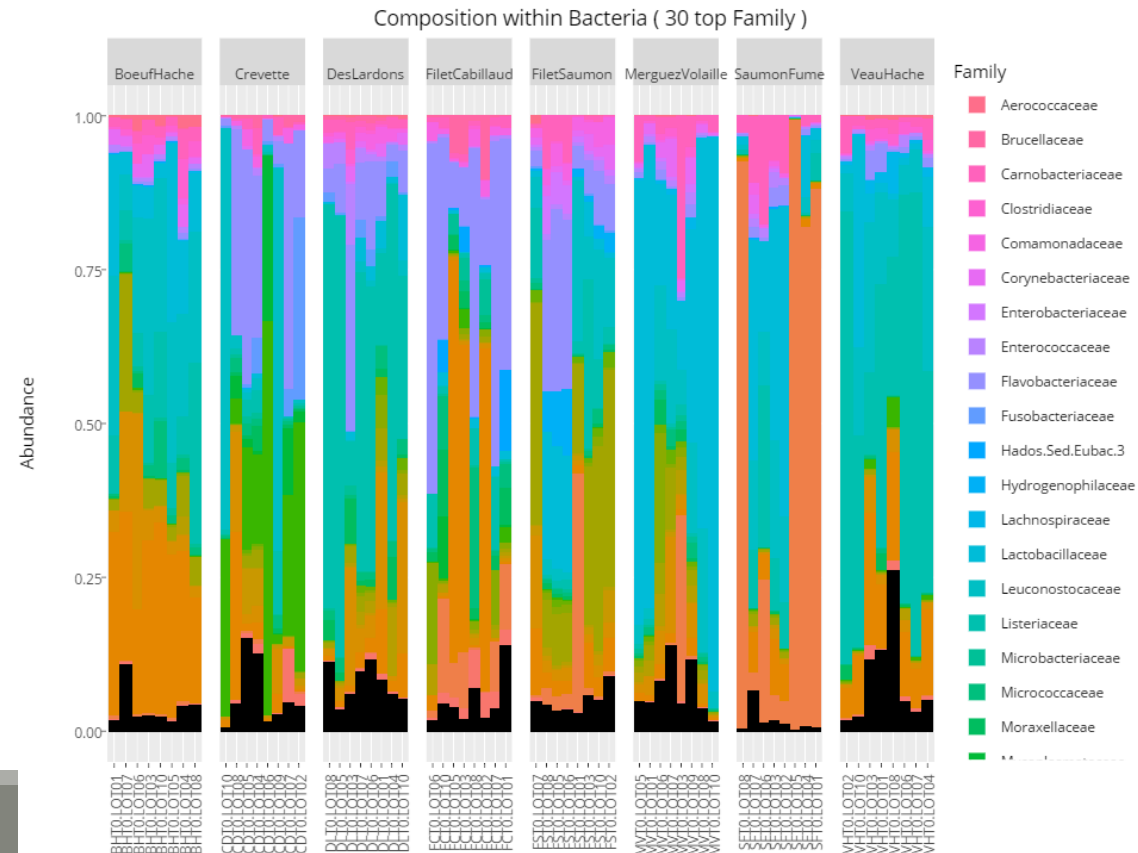
II. Biodiversity analysis

DIVERSITY INDICES

Exploring biodiversity : descriptors

- The **richness** corresponds to the number of OTUs or functional groups present in communities. It caracterises the **composition**.

- The **diversity** takes into account the relative abundancy of species. It caracterises the **structure**

Ecosystem 1

Ecosystem 2
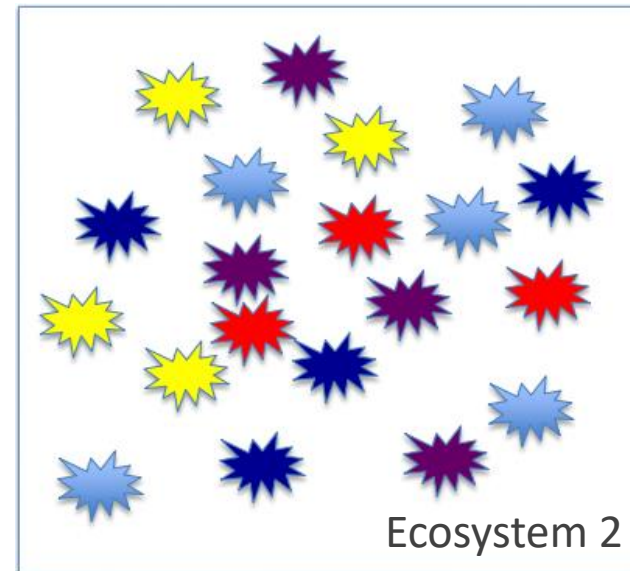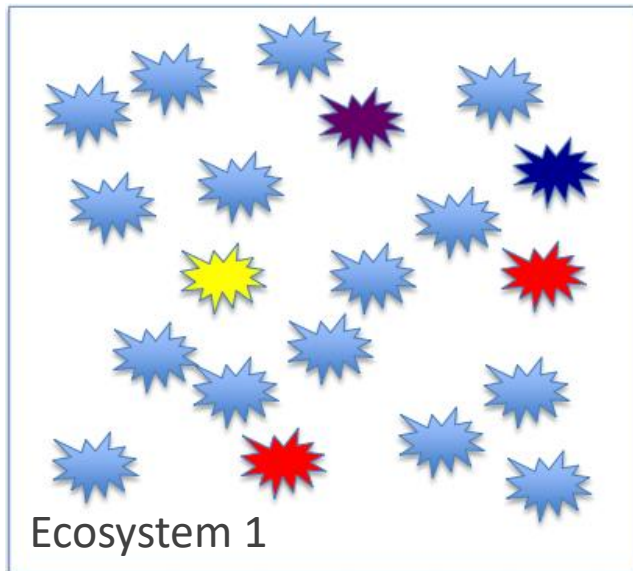
Richness : Eco1 = Eco2
Diversity: Eco2 > Eco1

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);
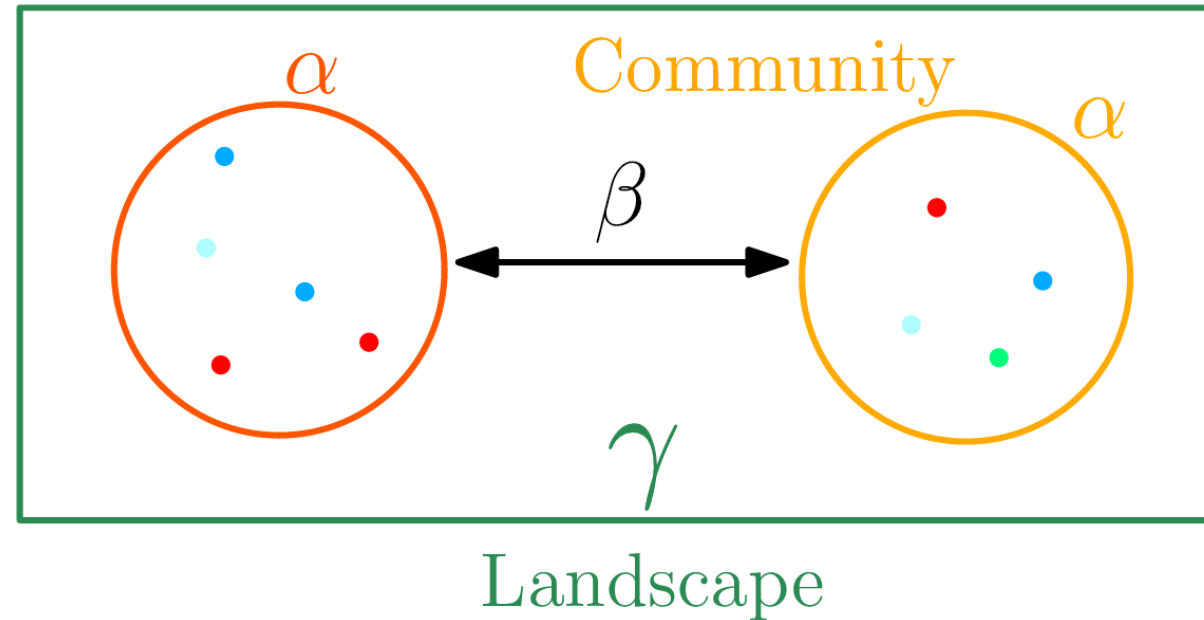
Exploring biodiversity : statistical indices

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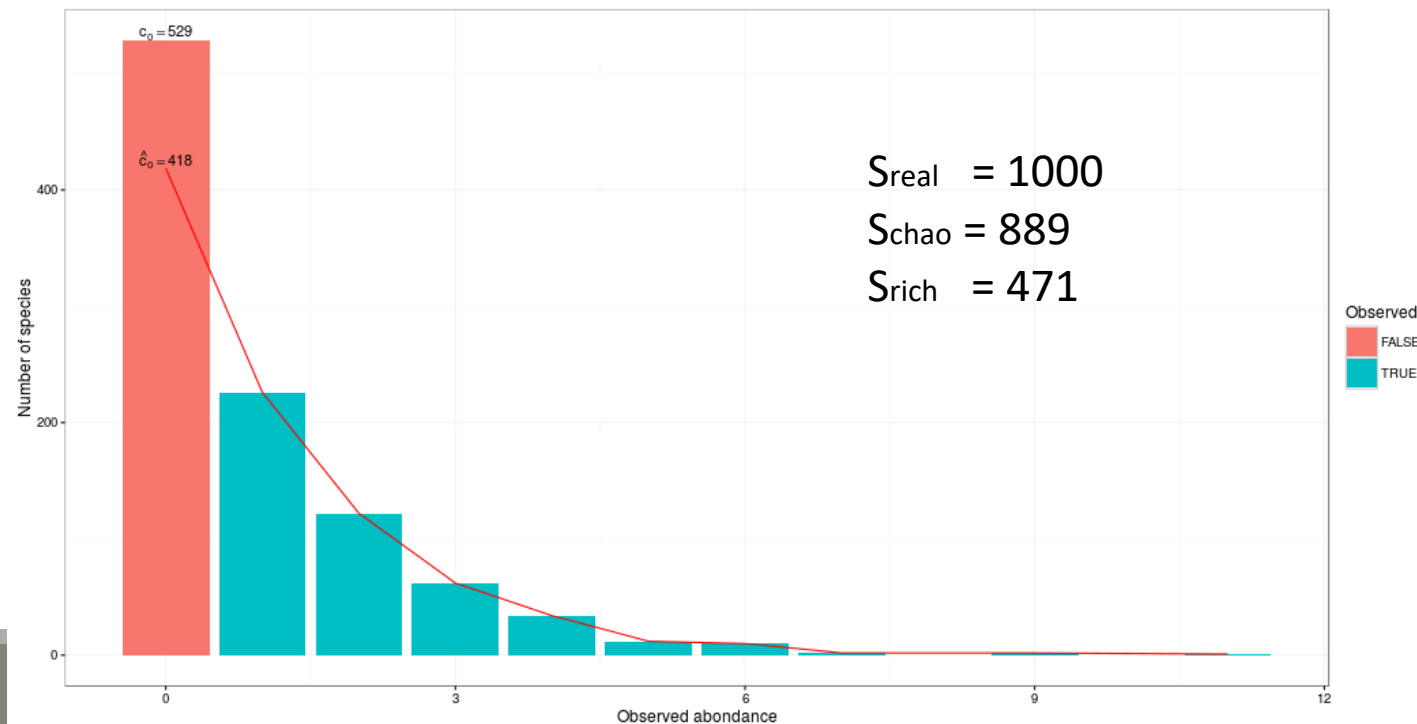
S_{real} = 1000
S_{chao} = 889
S_{rich} = 471

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

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

$S_{invsimp} = 15$
$S_{shan} = \log(15)$
$S_{rich} = 15$
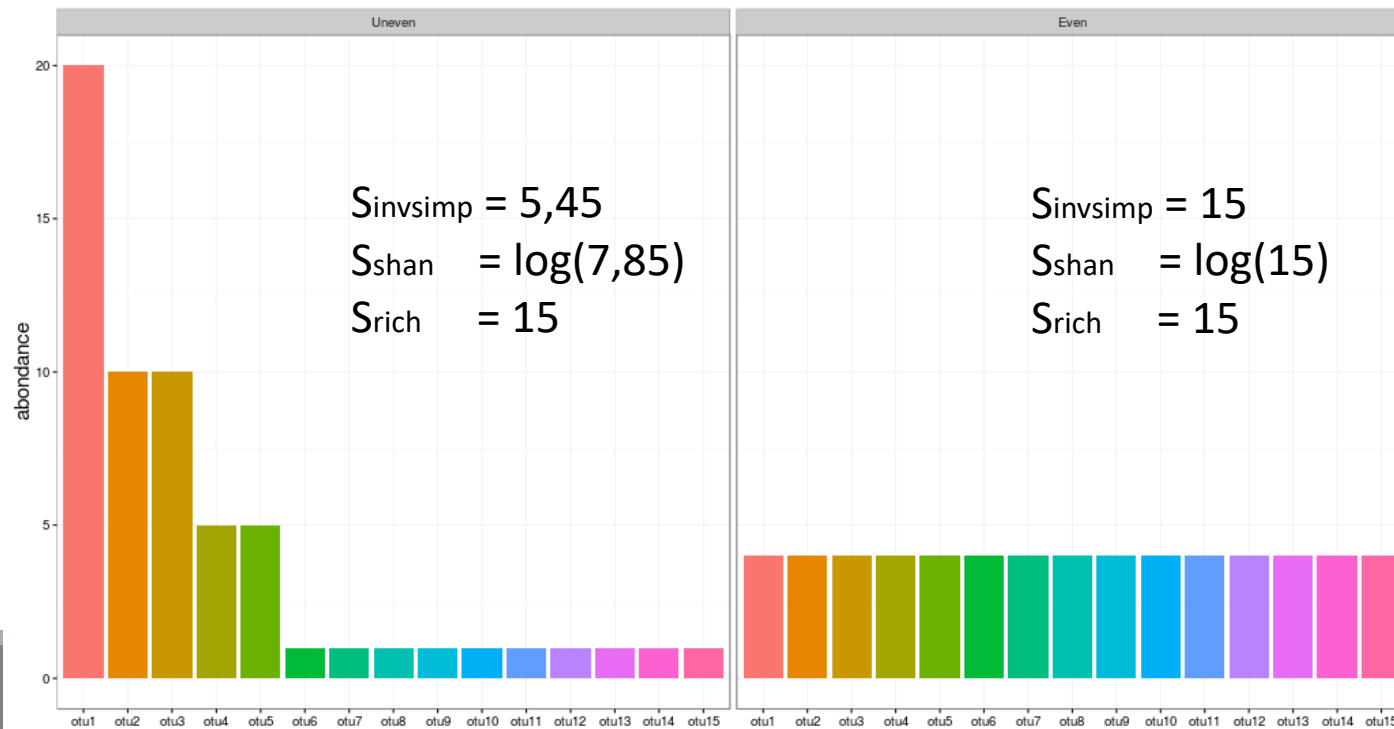
Interpretation :
15 observed species, but according to Shannon, the left example acts like there is 7.85 equally abundant species (5.45 for invSimp)
It is called effective diversities

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

Select R workspace including phyloseq object

Choose a sample variable to organise graphics

Choose which α-diversity indices you want to compute

Exercise A-5

Test it on EnvType

➔   What are the resulting datasets ?

➔   Which interpretation could you make on the boxplot results ?

➔   Have EnvType got an impact on α-diversity indice ?

Exercise A-5

➜ What are the resulting datasets ?

Report HTML file with graphical and statistical results

14: EnvType: alpha_diversity.html

Tabular file containing the detailed value of each indice in each sample

13: EnvType : alpha_diversity.tsv

1	2	3	4	5	6
	Observed	Chao1	se.chao1	Shannon	InvSimpson
DLT0.LOT08	205	215.344827586207	5.63455654866184	2.01591714100393	2.31393432719116
DLT0.LOT05	197	215.454545454545	9.04924368908291	1.76545015179311	1.90925718747888
DLT0.LOT03	219	226.916666666667	4.86003372343409	3.4340155003954	14.786255213252
DLT0.LOT07	220	224.714285714286	3.77924481885382	3.00227529842681	4.33279579199353

Exercise A-5

Boxplot interpretations

- Observed and Chao1 are very similar
  - ➔ All species have been detected

- Many taxa observed in Deslardons (high Chao1, high Observed)…

- …but low Shannon and Inverse-Simpson
  - ➔ communities are dominated by few abundant taxa

Alpha diversity distribution in function of EnvType
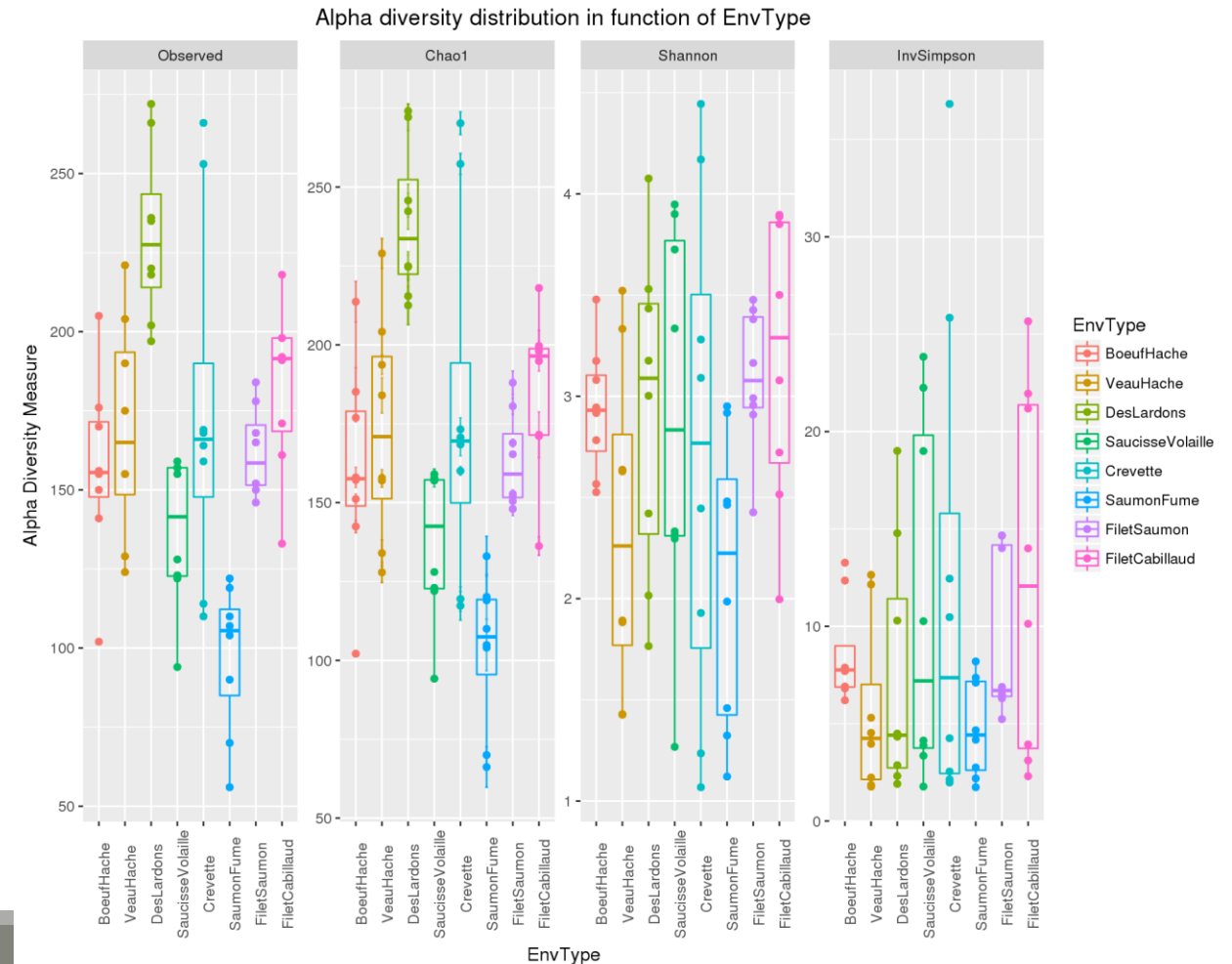
Exercise A-5

Anova interpretations

- Environments differ a lot in terms of richness...

- ...but not so much in terms of Shannon 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  82142   11735   11.64 5.02e-09 ***
Residuals    56  56472    1008
---
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   7.91  1.1300   1.771  0.111
Residuals    56  35.72  0.6379
```

Exercise A-5

Rarefaction curves interpretation
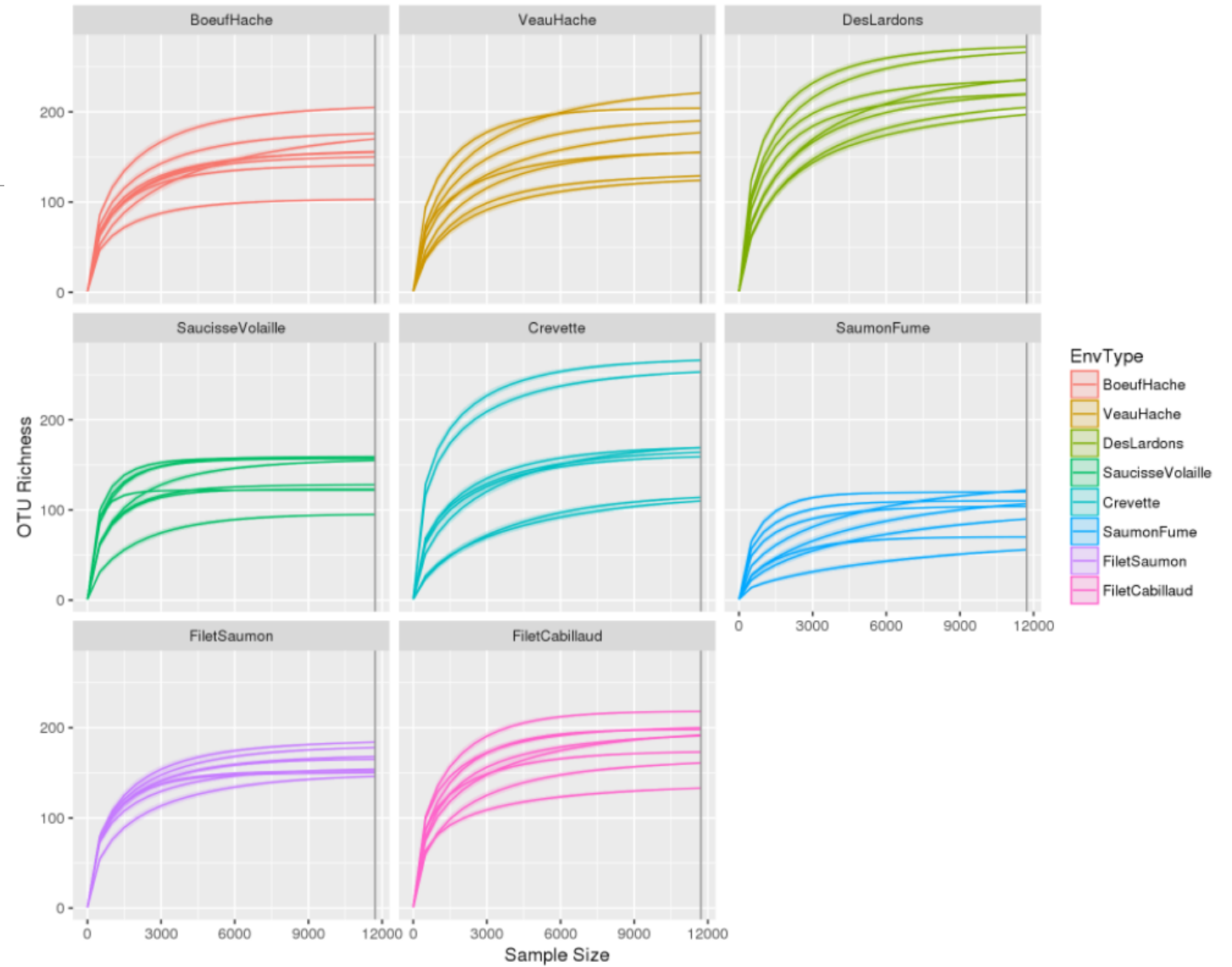
Exploring biodiversity : α-diversity

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)
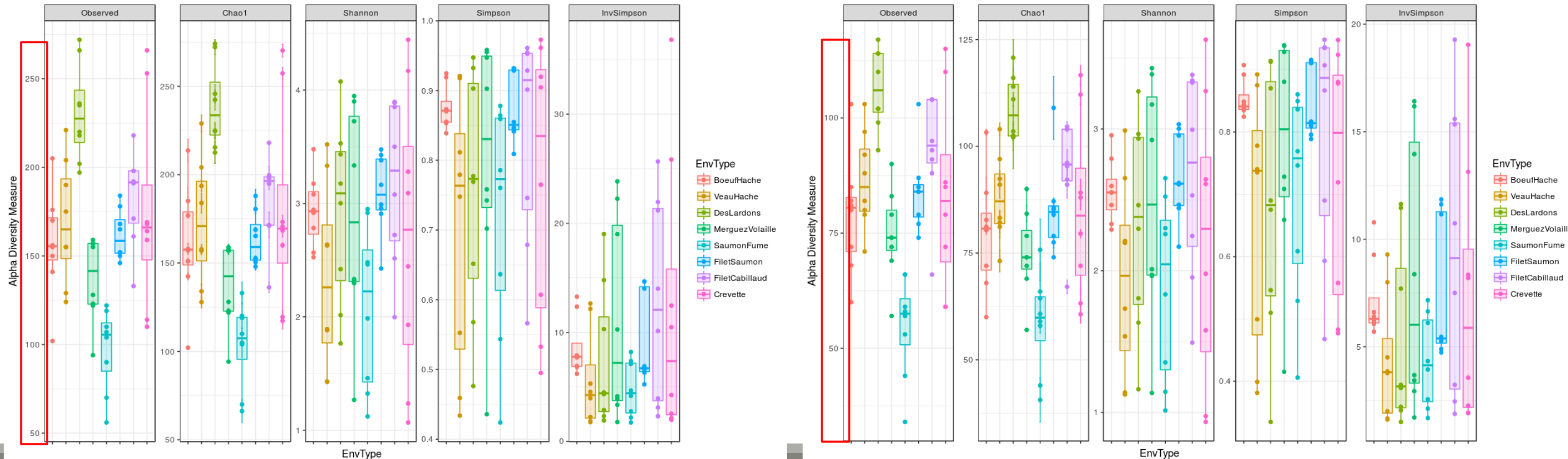
IV. Biodiversity analysis

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.

In this example :

- qualitatively, communities are very similar

- quantitatively, communities are very different

- phylogenetically, two communities seem to be closer than the third one.
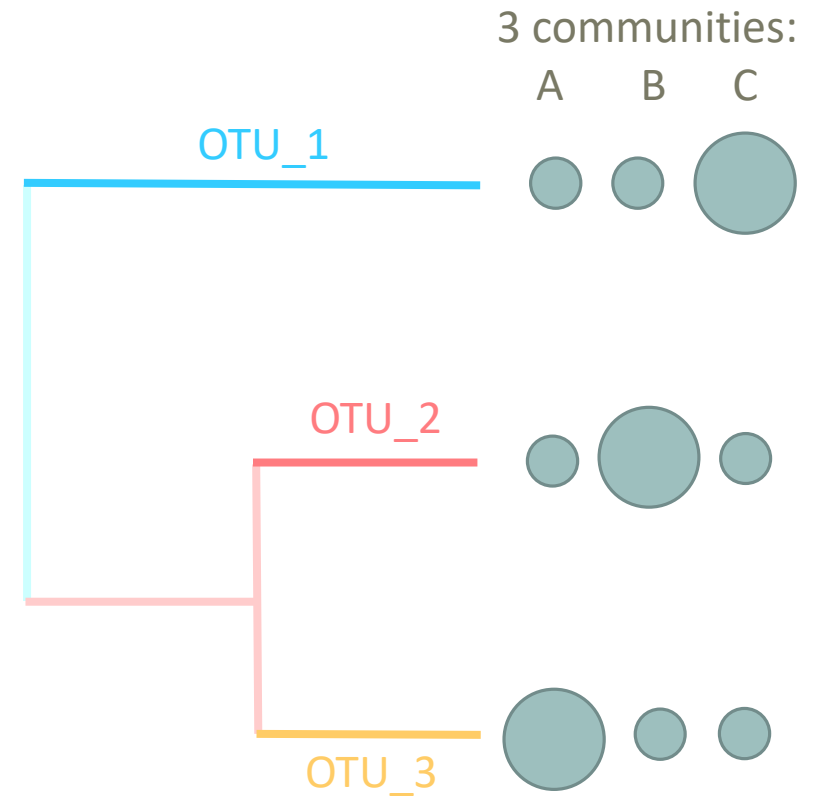
Exploring biodiversity : β-diversity

Jaccard:

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

Bray-Curtis:

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

Exploring biodiversity : β-diversity

- 2 communities

- 15 OTUs
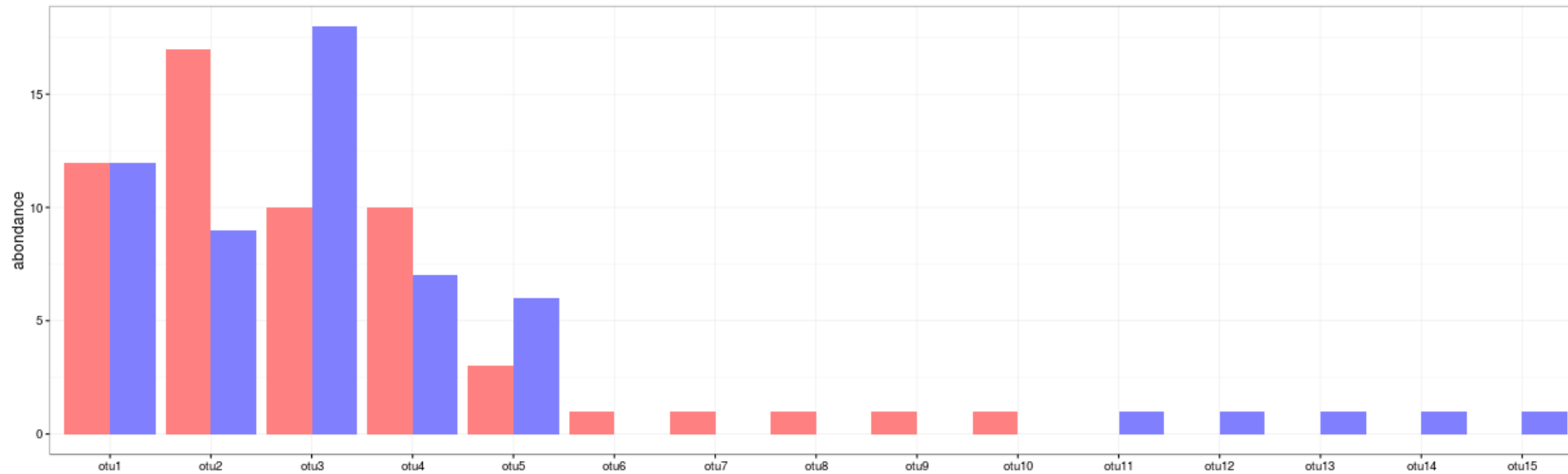
Exploring biodiversity : β-diversity

Jaccard:

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

D_{jac} = 10/15 = 0.667
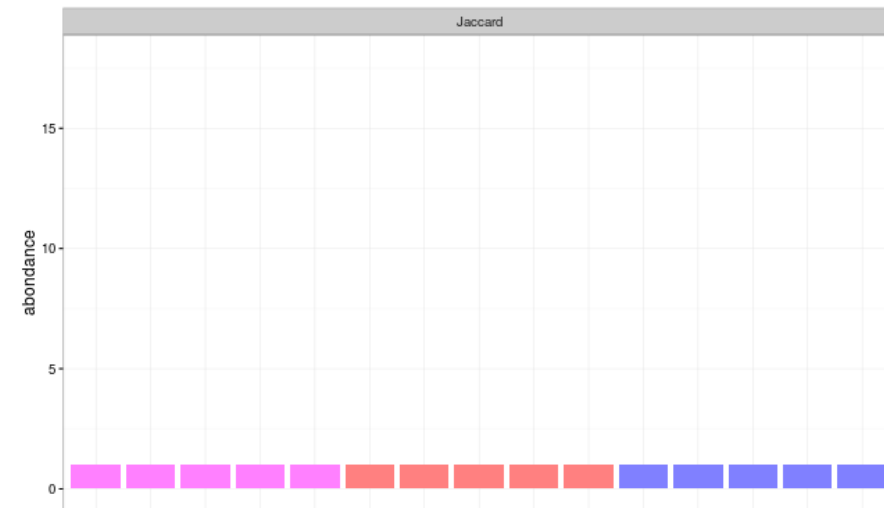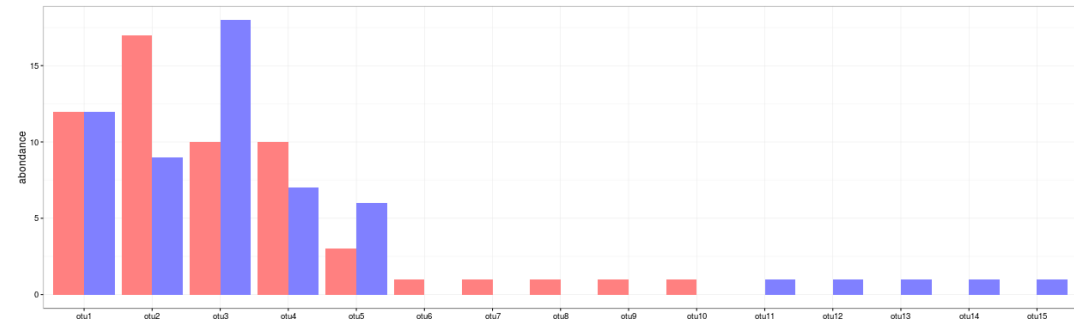
Exploring biodiversity : β-diversity

Bray-Curtis:

- Fraction of the community specific to either 1 or 2

$$D_{bc} = (8+8+3+3+10) / (24+26+28+17+9+10) = 0.281$$
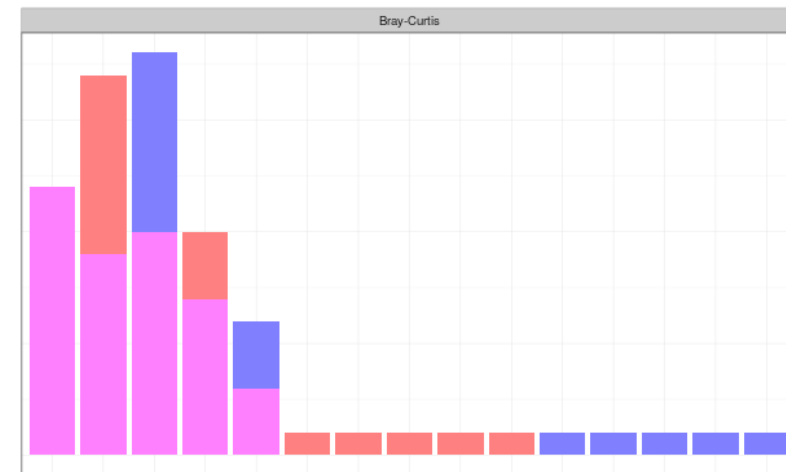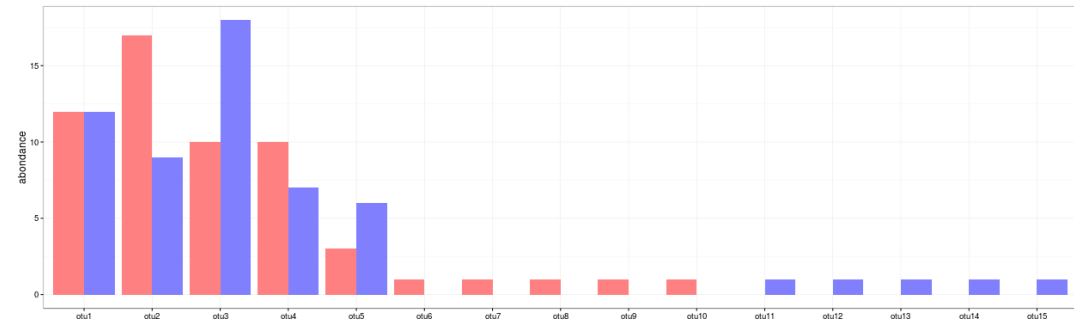
Exploring biodiversity : β-diversity

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

$D_{bc} = 0.909$
$D_{jac} = 0.667$
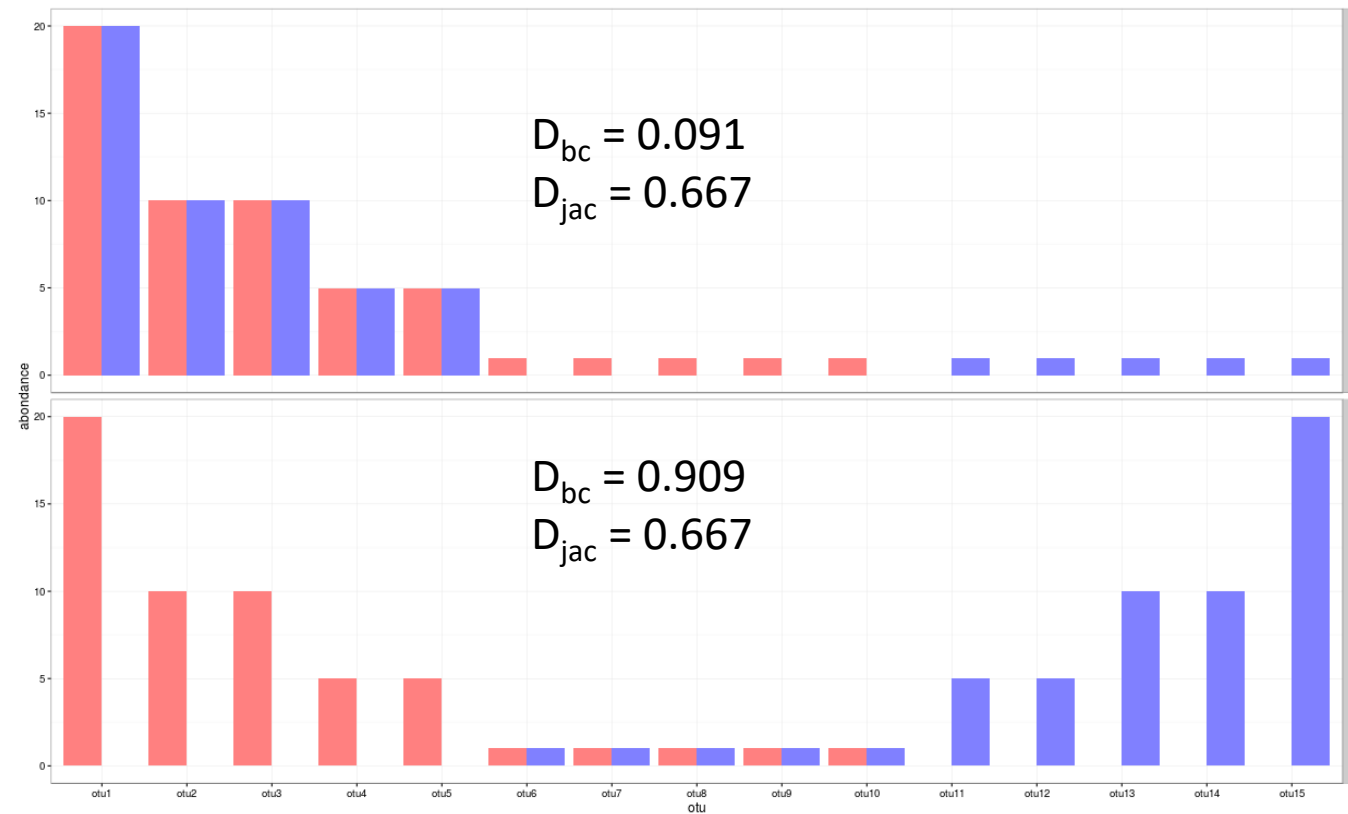
Exploring biodiversity : β-diversity

Unifrac:

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

Weigthed-Unifrac :

- Fraction of the <u>diversity</u> specific to either 1 or 2
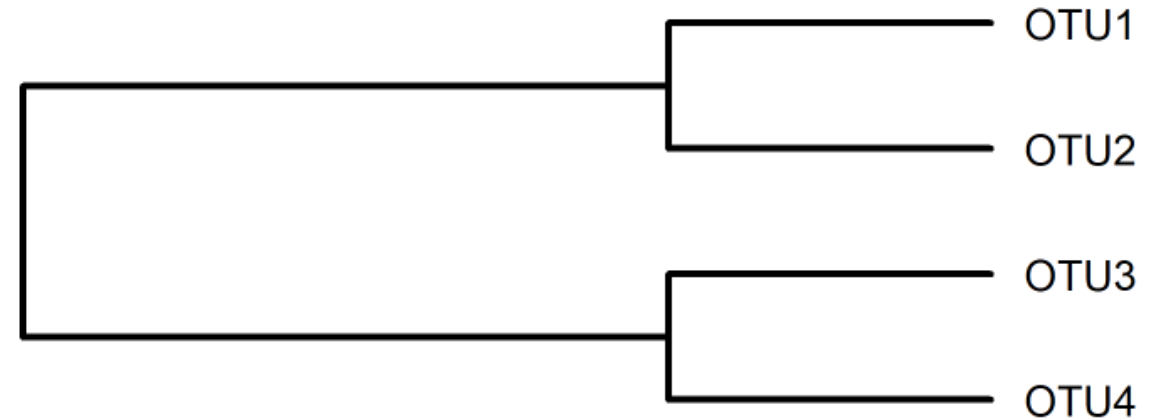
Exploring biodiversity : β-diversity

Unifrac:

- Fraction of the tree specific to either 1 or 2

$$Unifrac = \frac{\sum specific\_branch\_length}{\sum all\_branch\_length}$$
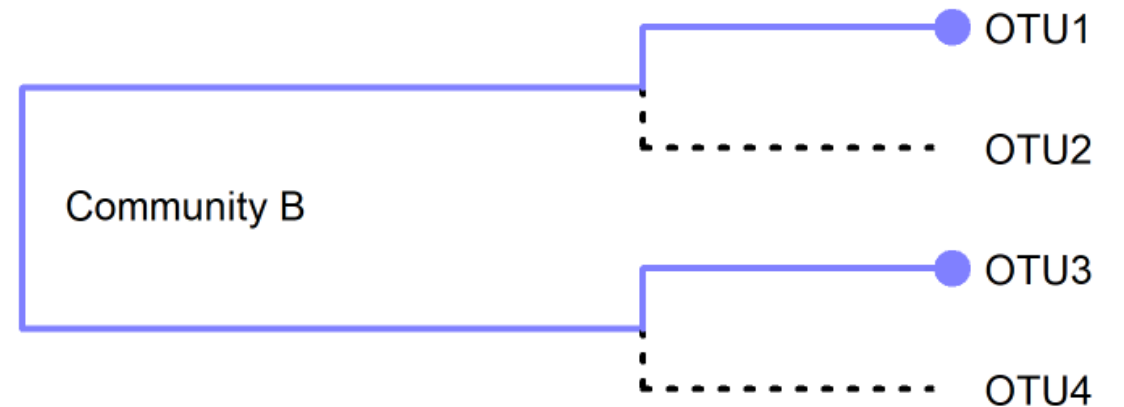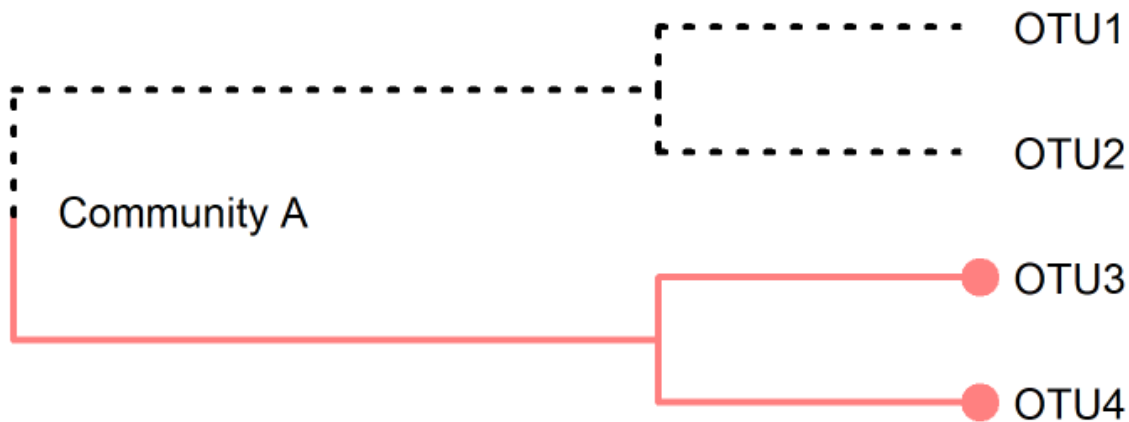
Exploring biodiversity : β-diversity

Unifrac:

- Fraction of <u>the tree </u>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 :

$$Unifrac= \frac{\sum specific\_branch\_length}{\sum all\_branch\_length} = 0.6$$
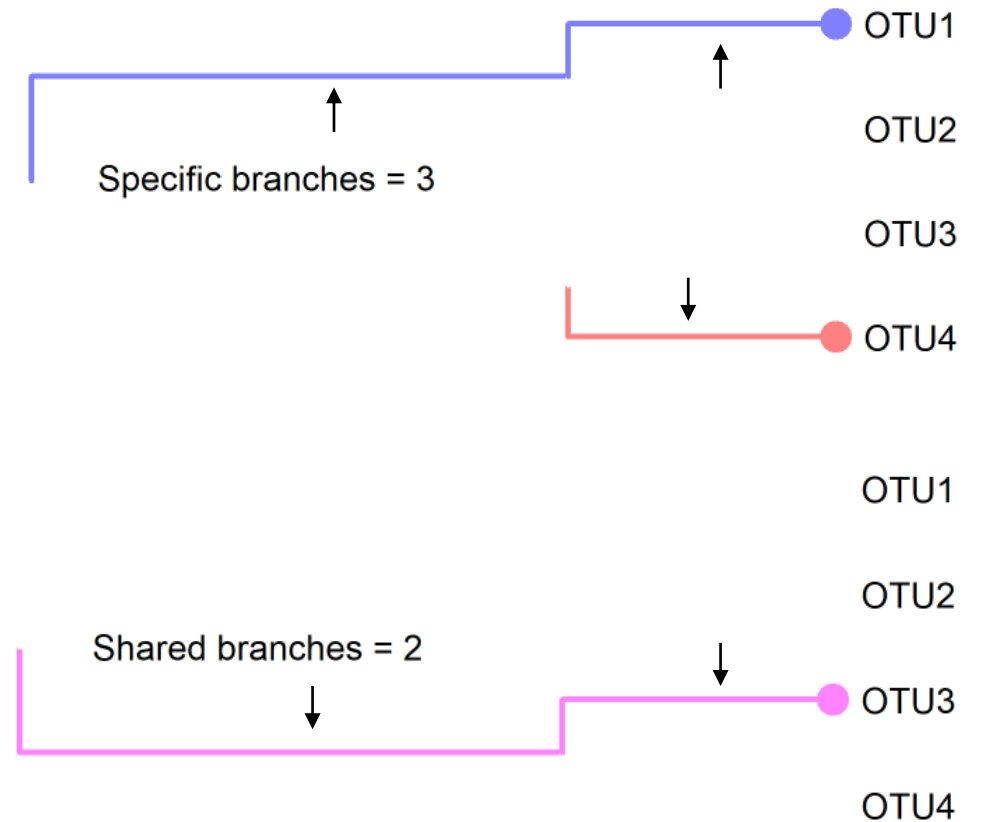
Specific branches = 3

Shared branches = 2

Exploring biodiversity : β-diversity

Weigthed-Unifrac :

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

$$WUnifrac = \frac{\sum reduced\_branch\_length}{\sum non\_reduced\_branch\_length}$$
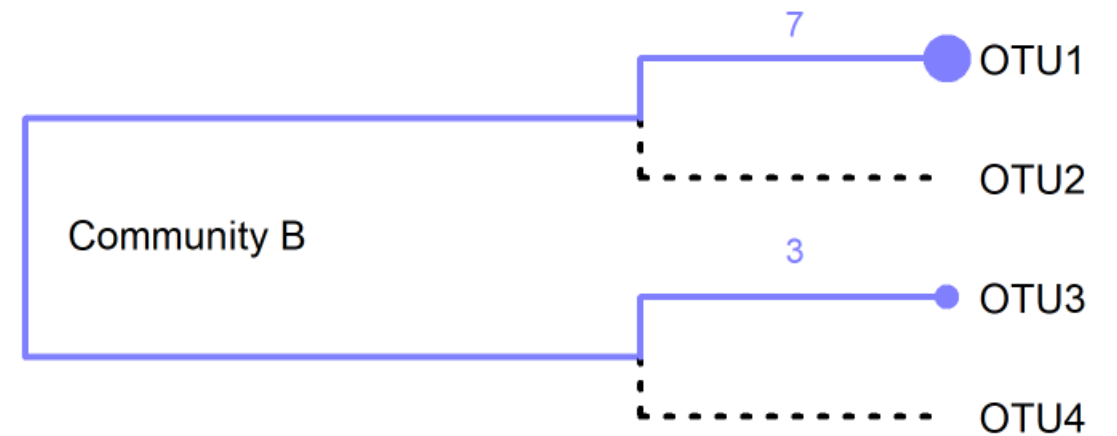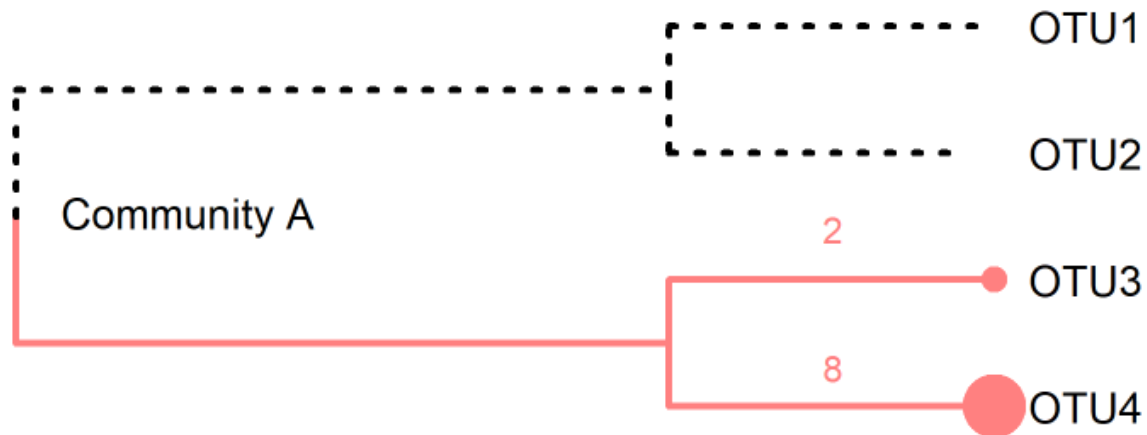
Exploring biodiversity : β-diversity

Weigthed-Unifrac :

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

$$WUnifrac = \frac{\sum reduced\_branch\_length}{\sum non\_reduced\_branch\_length}$$
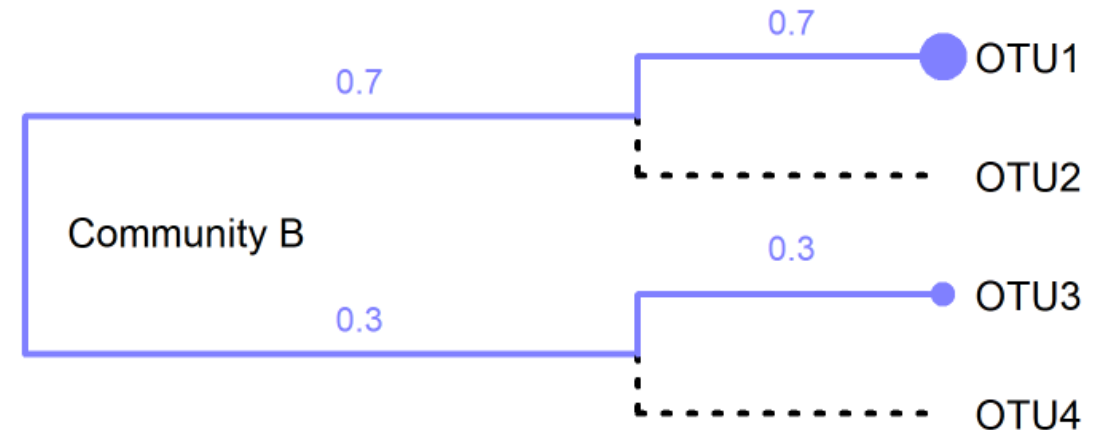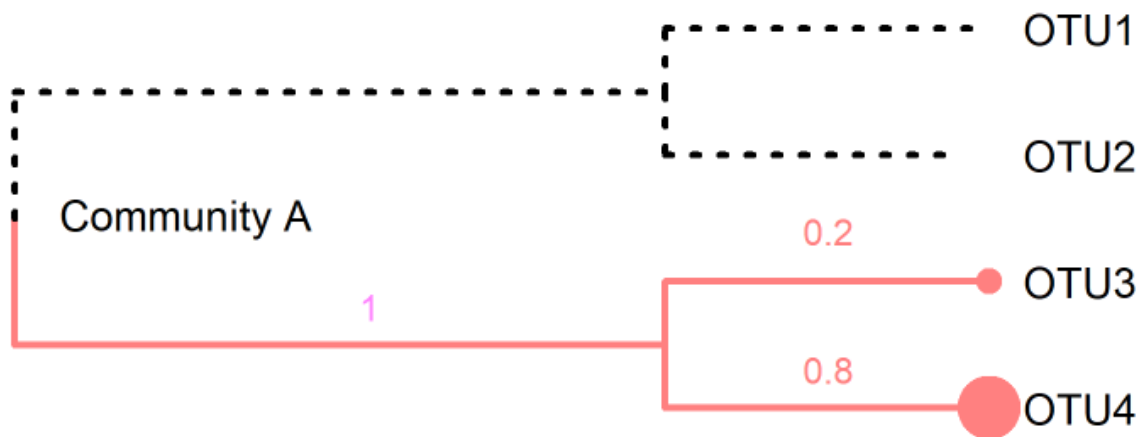
Exploring biodiversity : β-diversity

Weigthed-Unifrac :

- Fraction of the <u>diversity </u>specific to either 1 or 2
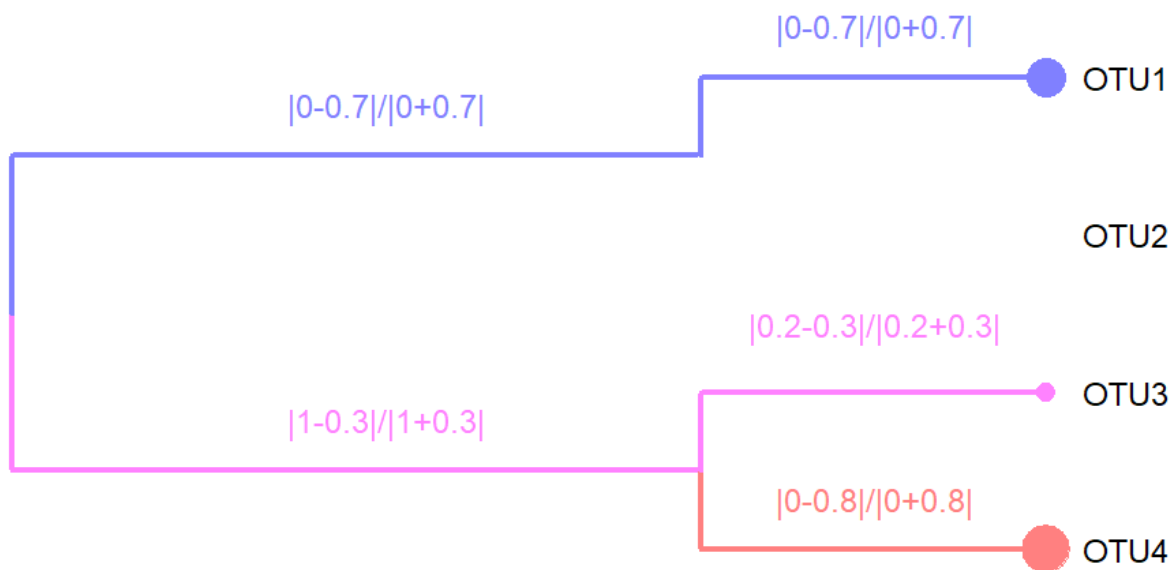
$$WUnifrac = \frac{\sum reduced\_branch\_length}{\sum non\_reduced\_branch\_length}$$

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

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

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

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

Tree branch labels:
- |0-0.7|/|0+0.7| (blue)
- |0-0.7|/|0+0.7| (blue) → OTU1
- OTU2
- |0.2-0.3|/|0.2+0.3| (pink) → OTU3
- |1-0.3|/|1+0.3| (pink)
- |0-0.8|/|0+0.8| (red) → OTU4

Exploring biodiversity : β-diversity

Weigthed-Unifrac :

- Fraction of the <u>diversity</u> specific to either 1 or 2
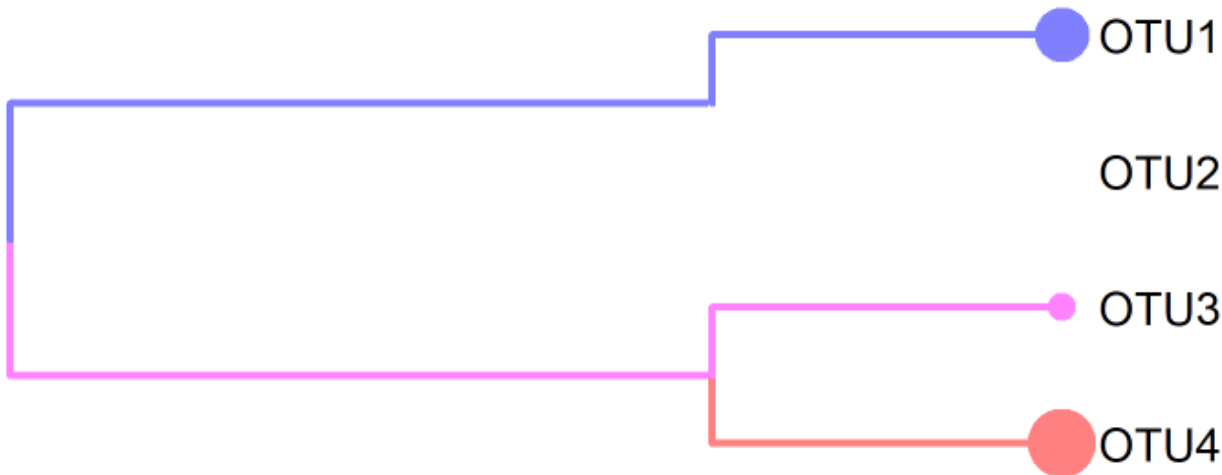
$$WUnifrac = \frac{\sum reduced\_branch\_length}{\sum non\_reduced\_branch\_length}$$

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

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

Exploring biodiversity : β-diversity

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

Low Unifrac / High Jaccard
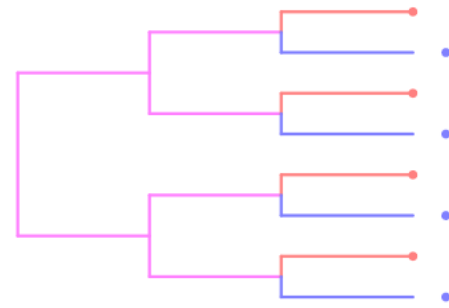
High Unifrac / High Jaccard
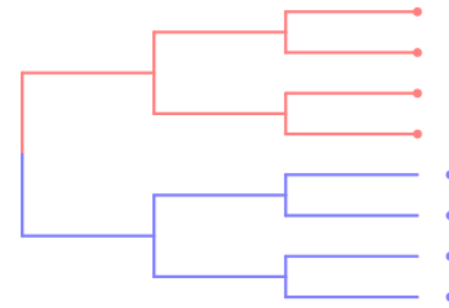
Low wUnifrac / High Bray Curtis
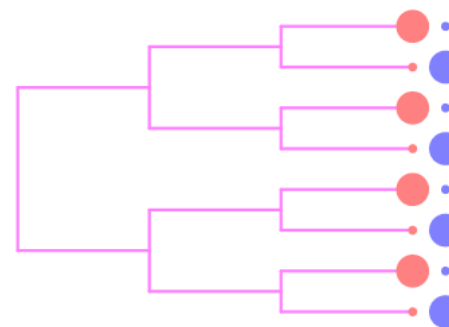
High wUnifrac / High Bray Curtis
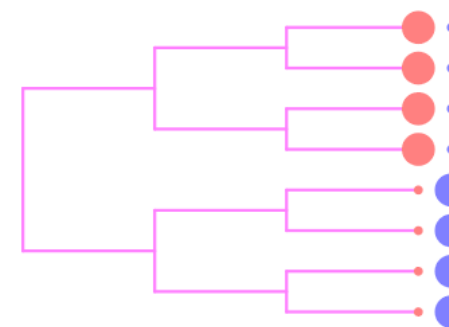
Exploring biodiversity : β-diversity

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"

"w" "-1" "c" "wb" "r" "I" "e" "t" "me" "j" "sor" …

Exploring biodiversity : β-diversity

Explore the sample normalised count

Choose a sample variable to organise graphics.

Choose which beta diversity distances you want to compute

Exercise A-6

Try it with the 4 most commonly used distance methods

➔ What are the output datasets ?

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

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

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

Exercise A-6

→ What are the output datasets ?

Report HTML file with graphical and statistical results

17: FROGSSTAT Phyloseq Beta Diversity: beta diversity
947.7 KB
format: **html**, database: ?

One tabular file per distance method containing the all samples againt all beta diversity distance : a matrix

	DLT0.LOT08	DLT0.LOT05	DLT0.LOT03
DLT0.LOT08	0	0.239033964840416	0.724185014507595
DLT0.LOT05	0.239033964840416	0	0.817716333845366
DLT0.LOT03	0.724185014507595	0.817716333845366	0

21: FROGSSTAT Phyloseq Beta Diversity: beta diversity (wUnifrac.tsv)

20: FROGSSTAT Phyloseq Beta Diversity: beta diversity (Unifrac.tsv)

19: FROGSSTAT Phyloseq Beta Diversity: beta diversity (Jaccard.tsv)

18: FROGSSTAT Phyloseq Beta Diversity: beta diversity (Bray_Curtis.tsv)

Exercise A-6

- Jaccard lower than Bray-Curtis
➡ abundant taxa are not shared

- Jaccard higher than Unifrac
➡ communities' taxa are distinct but phylogenetically related

- Unifrac higher than weighted Unifrac
➡ abondant taxa in both communities are phylogenetically closed.
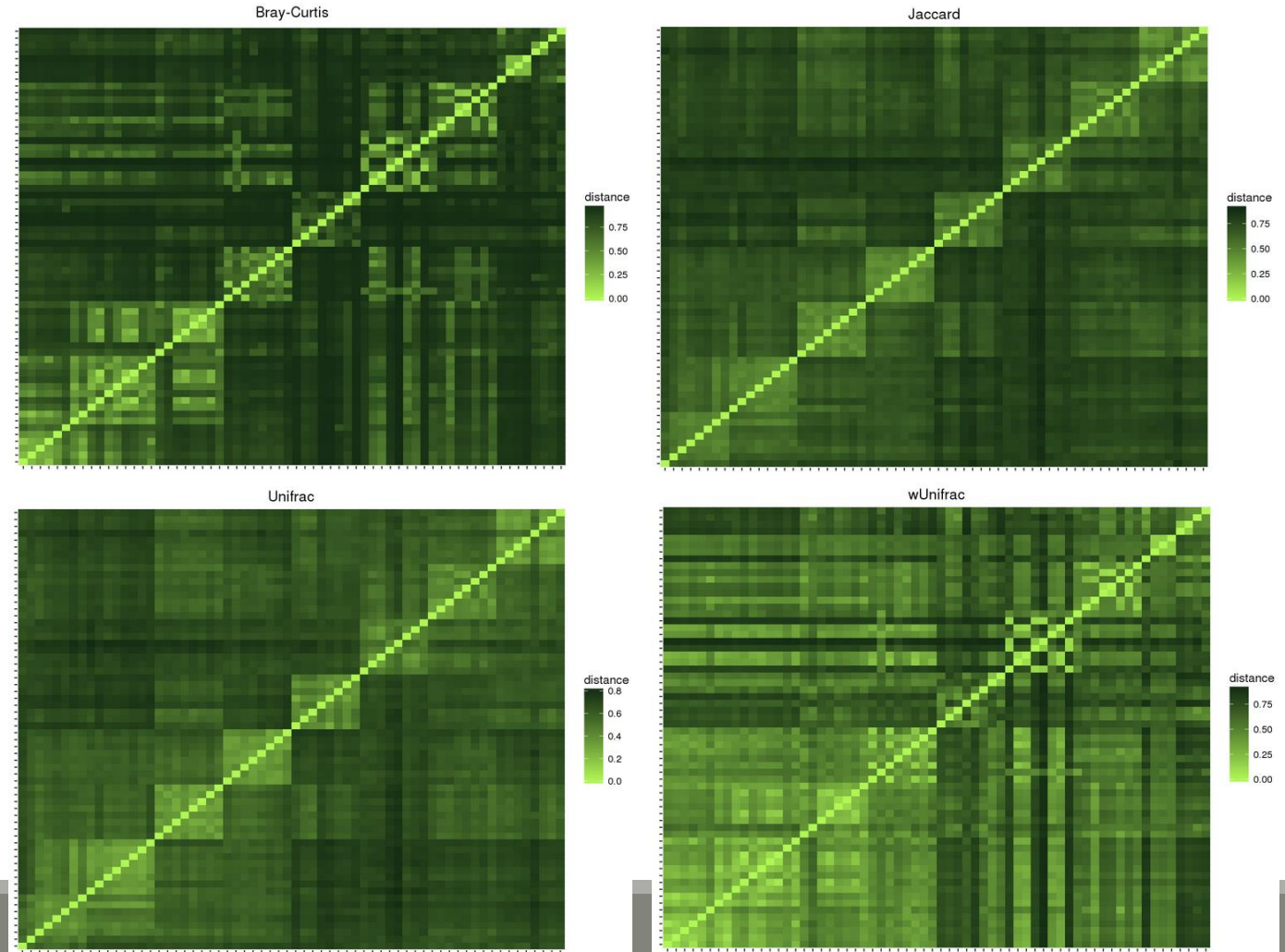
Exploring biodiversity : β-diversity

- 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
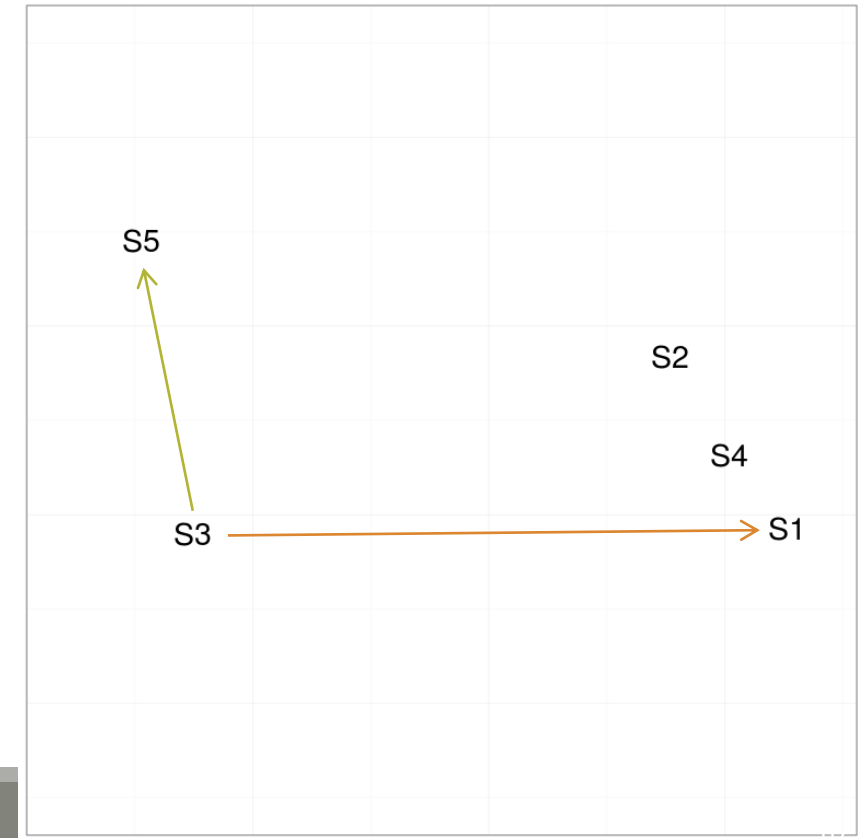
Exploring the structure : Heatmap

- 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 ▾ Options
(Galaxy Version 1.0.0)

Phyloseq object (format rdata)

| ▯ | ⊡ | ▭ | 8: food_normalized.Rdata | ▾ |

This is the result of FROGS Phyloseq Import Data Tool.

The beta diversity distance matrix file

| ▯ | ⊡ | ▭ | 21: FROGSSTAT Phyloseq Beta Diversity: beta_diversity (wUnifrac.tsv) | ▾ |

These file is the result of FROGS Phyloseq Beta Diversity tool.

Experiment variable

| EnvType |

The experiment variable that you want to analyse.

Ordination method

| MDS/PCoA | ▾ |

✔ Execute

Explore the sample normalised count

Choose the beta diversity distance matrix

Choose a sample variable to organise graphics.

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

Exercise A-7

Try it with one distance method matrix

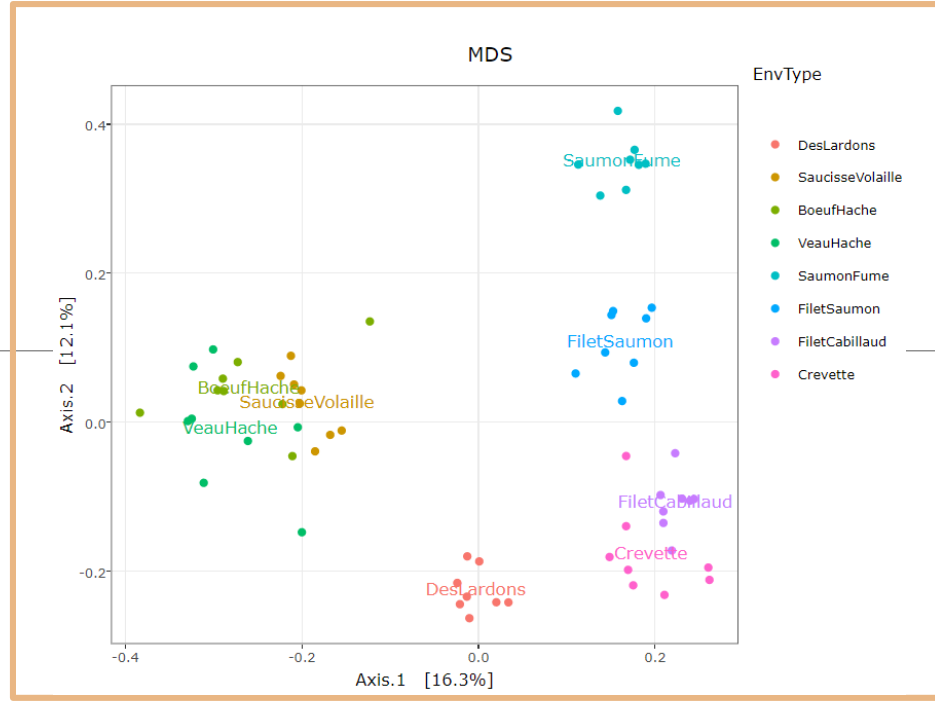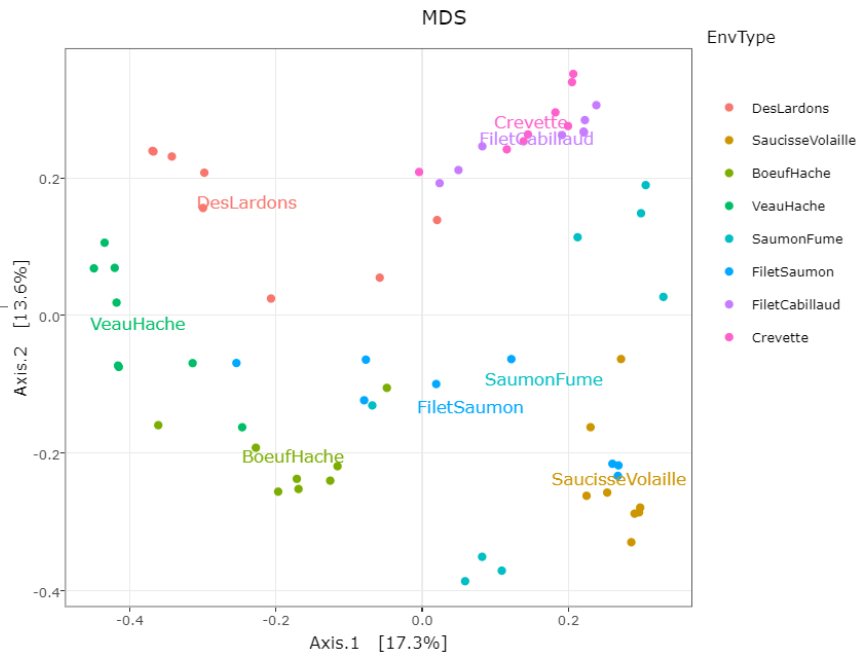➔ Are you satisfied of your ordination plot ?

Try with the other distance matrix

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

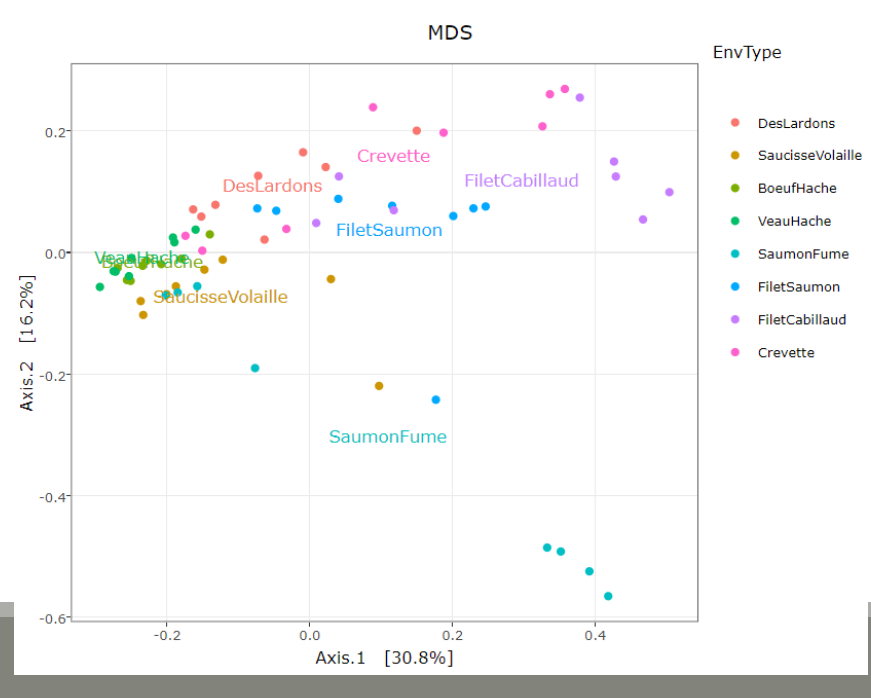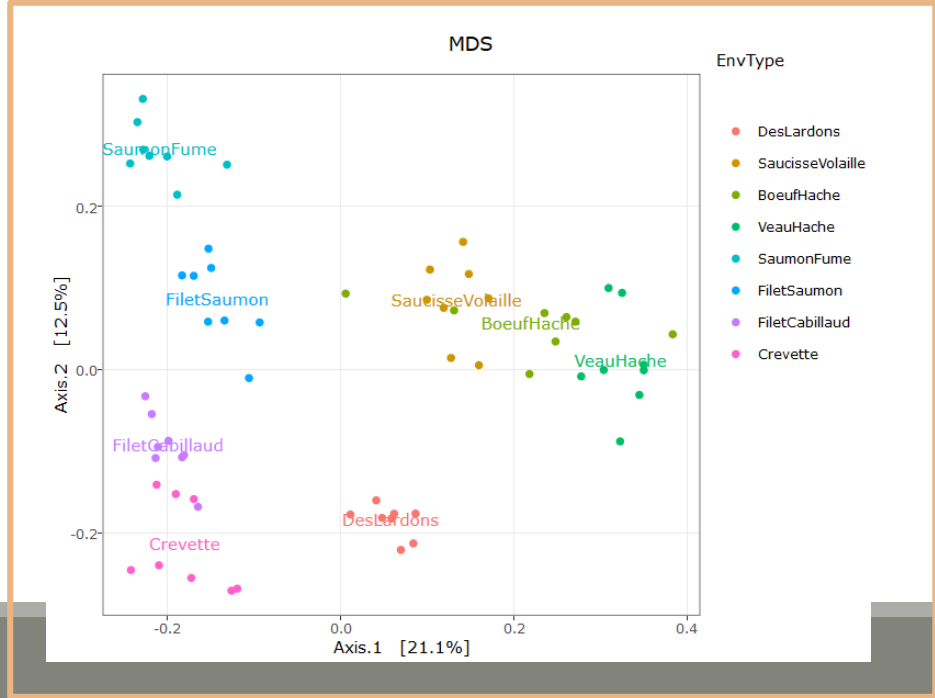➔ Guess why Lardon are somewhere between Meat and Seafood ?

➔ Based on your preferred distance matrix, what can you conclude on the heatmap ?

Bray Curtis

Jaccard

Unifrac

wUnifrac

Exercise A-7

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

➜ detected taxa segregate by origin.

- DesLardons is somewhere in between

➜ contamination induced by sea salt.

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

Exercise A-7

Heatmap plot with EnvType
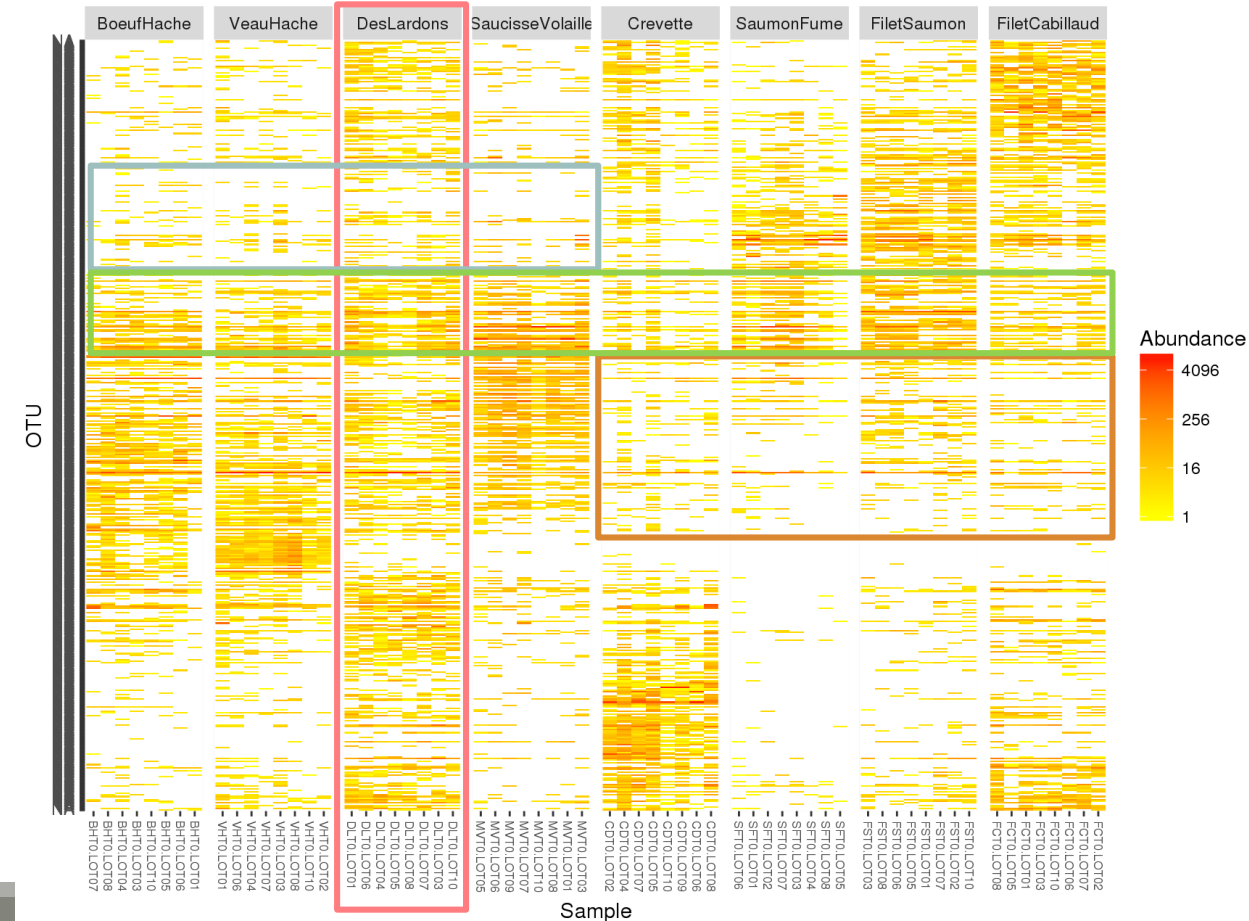
- Block-like structure of the abundance table

- Interaction between (groups of) taxa and (groups of) samples

- Core and condition-specific microbiota

➔ Classification of taxa and use of custom taxa order to highlight structure
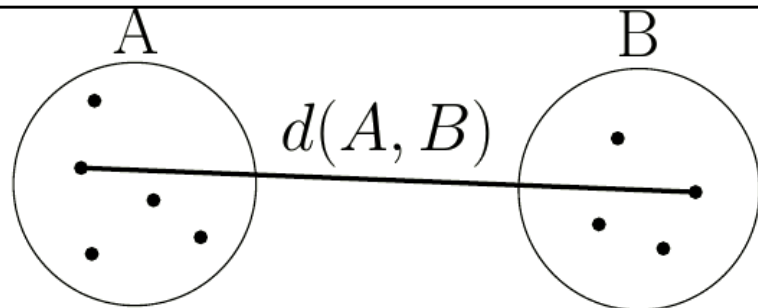
II. Exploring the structure

HIERARCHICAL CLUSTERING
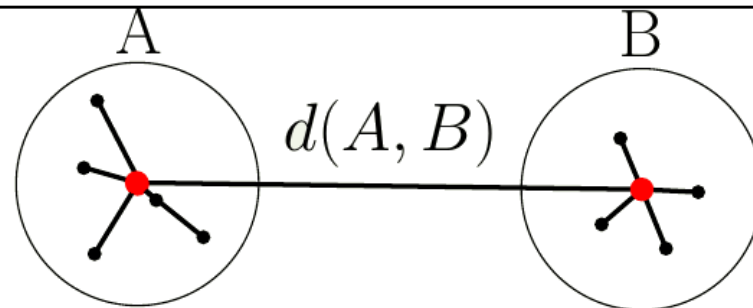
Exploring the structure : clustering

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.
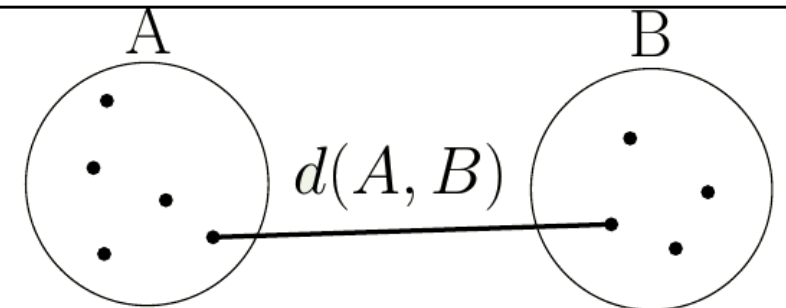
Exploring the structure : clustering

FROGSSTAT Phyloseq Sample Clustering of samples using different linkage methods ▼ Options
(Galaxy Version 1.0.0)

Phyloseq object (format rdata)

8: food_normalized.Rdata ▼

This is the result of FROGS Phyloseq Import Data tool.

The beta diversity distance matrix file

20: FROGSSTAT Phyloseq Beta Diversity: beta_diversity (Unifrac.tsv) ▼

This file is the result of FROGS Phyloseq Beta Diversity tool.

Experiment variable

EnvType

The experiment variable that you want to analyse.

✔ Execute

Explore the sample normalised count

Choose the beta diversity distance matrix

Choose a sample variable to organise graphics.

The tree different linkage functions will be used, generating three different trees

Exercise A-8

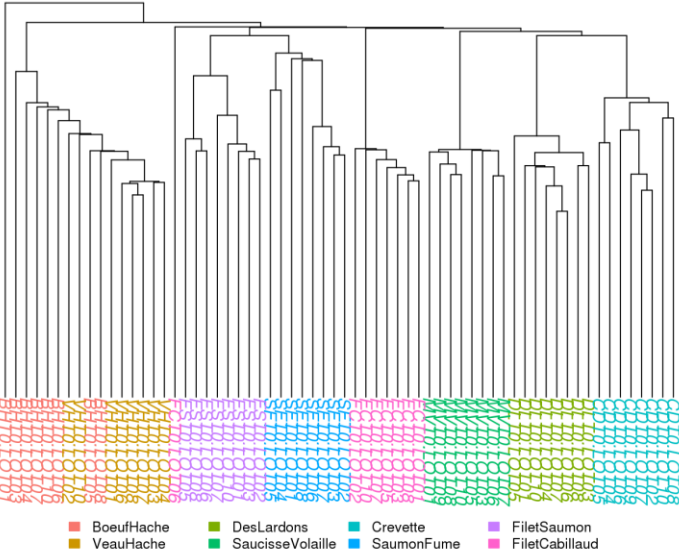Try it with « a good » distance method matrix on EnvType and on FoodType

➔ Which linkage method seems better to fit the data ?
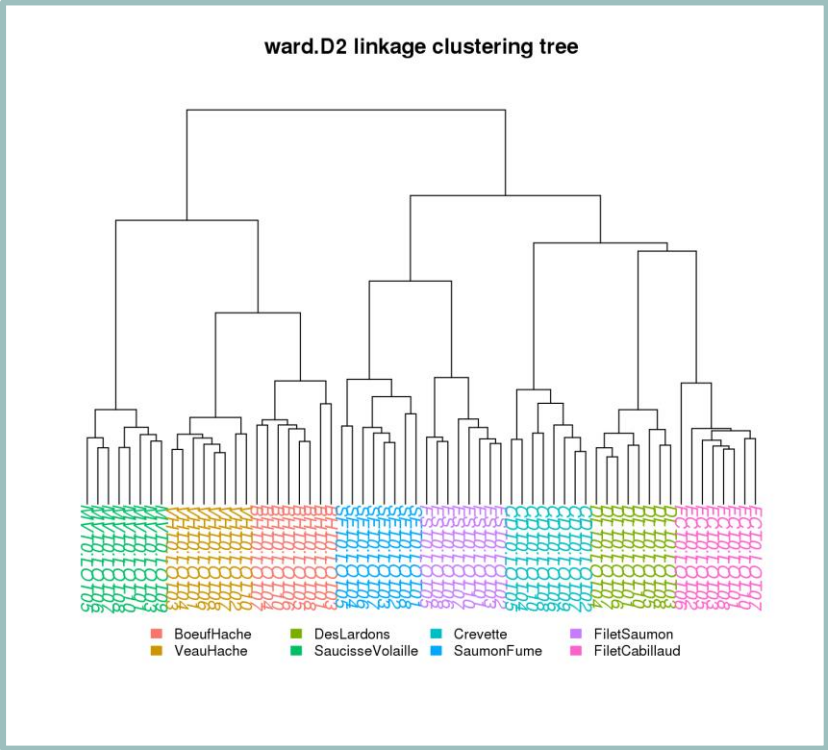
Try with « a bad » distance matrix
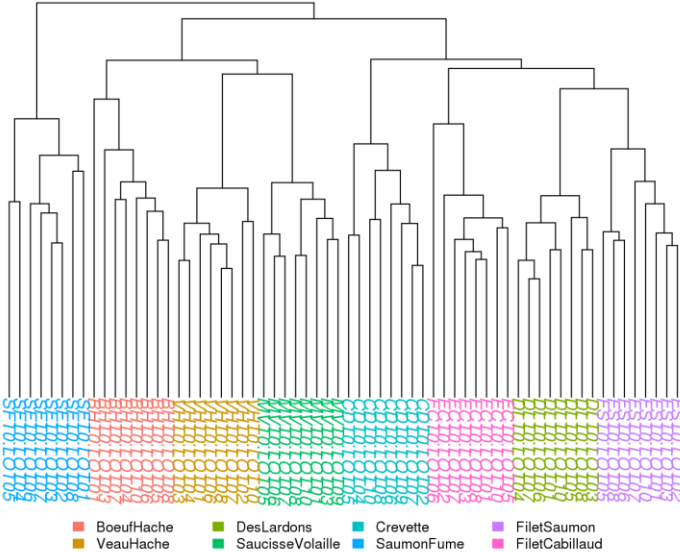
➔ Is there a big difference ?

Exercise A-8

single linkage clustering tree

ward.D2 linkage clustering tree

complete linkage clustering tree

BoeufHache   DesLardons   Crevette   FiletSaumon
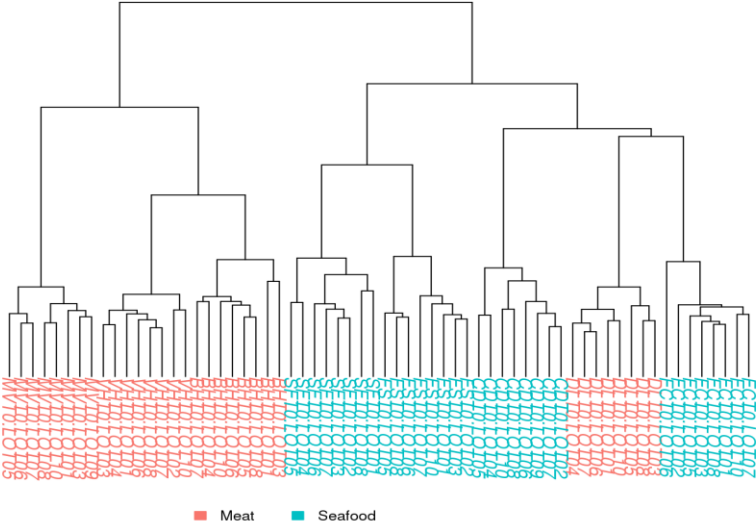VeauHache   SaucisseVolaille   SaumonFume   FiletCabillaud

Exercise A-8

Remarks

- Consistent with the ordination plots, clustering works quite well for the UniFrac distance for some linkage (Ward)

  → DesLardons seems to be much closer to Seafood than Meat.

- Clustering is based on the whole distance whereas ordination represents parts of the distance (the most it can with 2 dimensions)
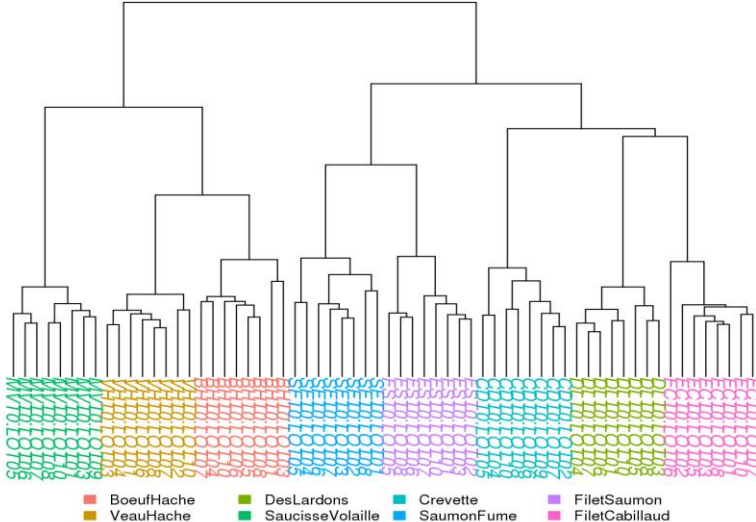
Exercise A-8

Ward linkage on Unifrac distance matrix

ward.D2 linkage clustering tree

■ Meat   ■ Seafood

ward.D2 linkage clustering tree

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

Diversity partitioning

Diversity partitioning

Are the structures seen linked to metadata ? Have the metadata got an effect on our communities composition ?

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

- tests composition differences of communities from different groups using a distance matrix
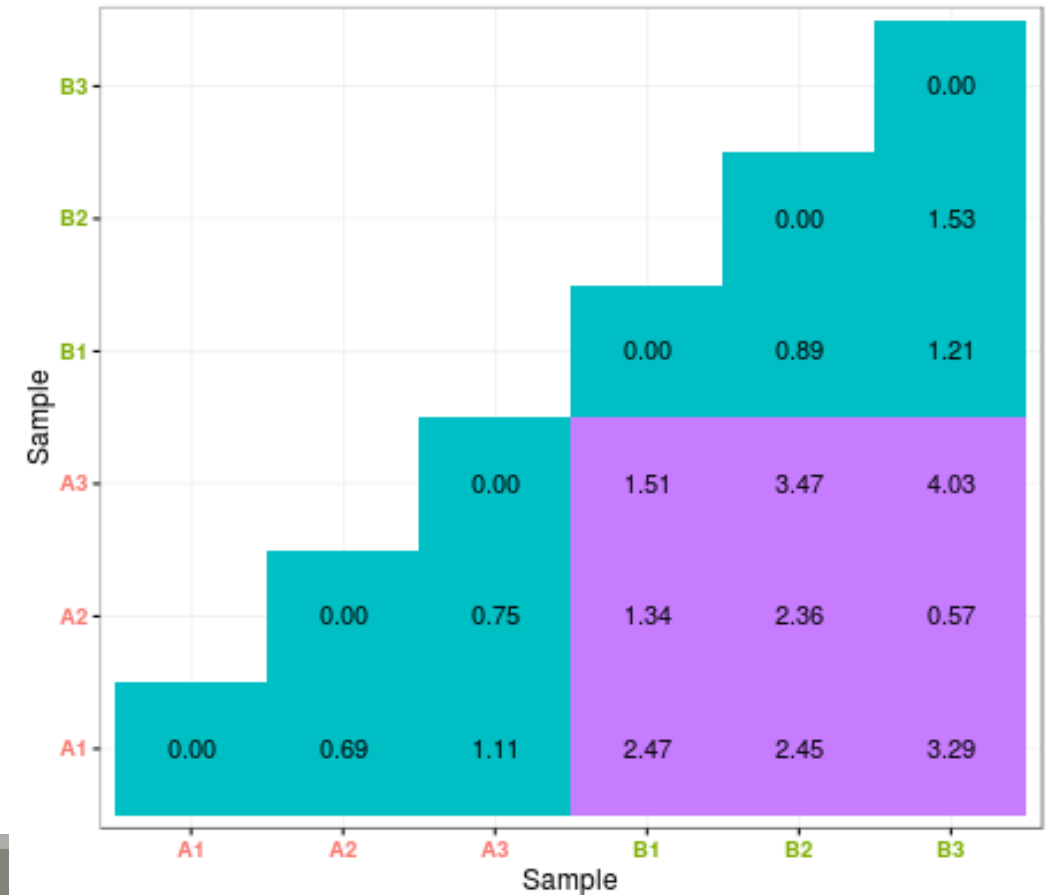- compares within group to between group distances
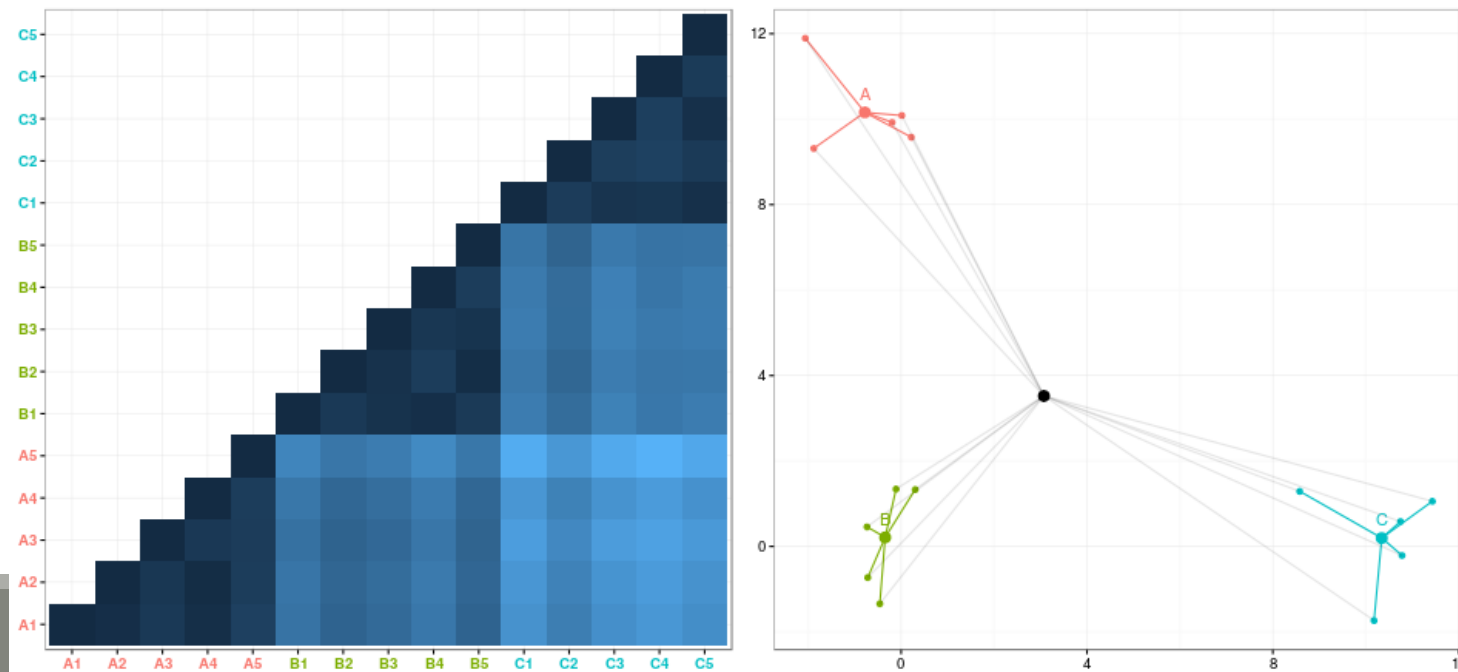
Diversity partitioning : Multivariate ANOVA

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

How it works ?

- Computes sum of square distance
- Variance analysis

Diversity partitioning : Multivariate ANOVA

FROGSSTAT Phyloseq Multivariate Analysis Of Variance (Galaxy Version 1.0.0) ▼ Options

Phyloseq object (format rdata)

8: food_normalized.Rdata ▼

This is the result of FROGS Phyloseq Import Data tool.

The beta diversity distance matrix file

20: FROGSSTAT Phyloseq Beta Diversity: beta_diversity (Unifrac.tsv) ▼

This file is the result of FROGS Phyloseq Beta Diversity tool.

Experiment variable

EnvType

The experiment variable that you want to analyse.

✔ Execute

Explore the sample normalised count

Choose the beta diversity distance matrix

Choose a sample variable to organise graphics.

Exercise A-9

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

➜ Does EnvType have an influence on the beta diversity variance ?

➜ What about FoodType ?

> Environment type explains roughly **62%** of the total variation

> Food type explains only **18 %** of the total variation

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    7.6445 1.09207  12.858 0.61645  1e-04 ***
Residuals 56    4.7564 0.08494         0.38355
Total     63   12.4009                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
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    2.2609 2.26092  13.824 0.18232  1e-04 ***
Residuals  62   10.1400 0.16355         0.81768
Total      63   12.4009                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
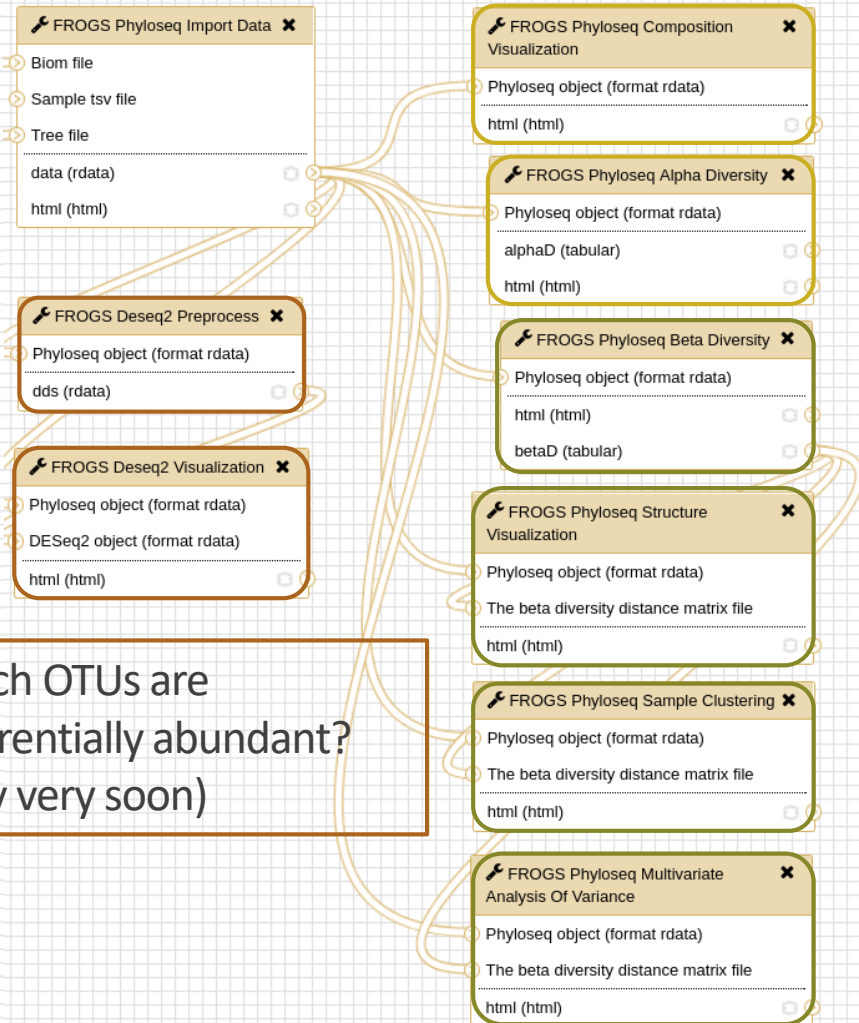
FROGSStat Summary

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? Possibly level name ?

- Well construct your sample_metadata TSV file, after import check that variable order is meaning full

- Keep in mind that :
  - Phyloseq composition and structure analysis need to be perform on normalised/rarefied counts
  - Different indices or distance methods will give different information
  - Test different distances or choose which one fits better our data
  - Richness indices depend lot on rare OTUs

PART B. Your turn !

Training Data2

A real analysis provided by Núria Mach *et al.*

16S survey of gut microbiomes from early life swines. Used (among others) to study the impact of weaning (Time and Weaned) on bacterial communities.

Along a kinetic of time 31 samples are analysed:

- Time : D14 (before weaning), D36, D48, D60, D70

- Weaned : TRUE, FALSE (Weaned is TRUE for TIME D14, else FALSE)

- sex : 1 (male), 2 (female)

155 samples of 16S V3-V4, and taxonomic affiliations was made with the Greengenes database

Exercise B-1

Upload this new dataset:
- kinetic.biom
- kinetic_sample_metadata.tsv
- tree.nwk

➔ How can you simply caracterise this dataset ?

➔ What is happening when you rarefy the counts ?

Exercise B-1

➜ How can you simply caracterise this dataset ?

- Number of OTUs and size / sample distribution with FROGS Clusters Stat

➜ More than 30% of OTUs are composed of just 1 sequence.

➜ But a small number of OTUs is specific to each sample.

- Number of taxonomic level, by converting biom to a tsv file with FROGS Biom to TSV

➜ Taxonomy are composed of 6 levels, from Kingdom to Genus

Root;Bacteroidetes;Bacteroidia;Bacteroidales;Prevotellaceae;Prevotella

Clusters size summary

Clusters size distribution

Clusters size distribution (decile)

Decile	Value
Min	1
1	1
2	1
3	1
4	2
Median	3
6	5
7	10
8	25
9	97
Max	13,867

Exercise B-1

➔ What is happening when you rarefy the counts ?

Import of raw counts

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 4031 taxa and 155 samples ]
sample_data() Sample Data:       [ 155 samples by 8 sample variables ]
tax_table()   Taxonomy Table:    [ 4031 taxa by 6 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 4031 tips and 4030 internal nodes ]
```

Import of rarefying counts

```
phyloseq-class experiment-level object
otu_table()   OTU Table:         [ 3002 taxa and 155 samples ]
sample_data() Sample Data:       [ 155 samples by 8 sample variables ]
tax_table()   Taxonomy Table:    [ 3002 taxa by 6 taxonomic ranks ]
phy_tree()    Phylogenetic Tree: [ 3002 tips and 3001 internal nodes ]
```

```
Number of sequences in each sample after normalization:  1056
```

➔ 4031 – 3002 = 1029 OTUs have been deleted, probably most of the singleton OTU

Exercise B-2

➔ What can you conclude with the composition plots ?

➔ What can you tell about alpha diversity indices ?
   Try it on raw counts and on rarefied counts.

Exercise B-2

➔ What can you conclude with the composition plots ?

Bar plot colored by Phylum
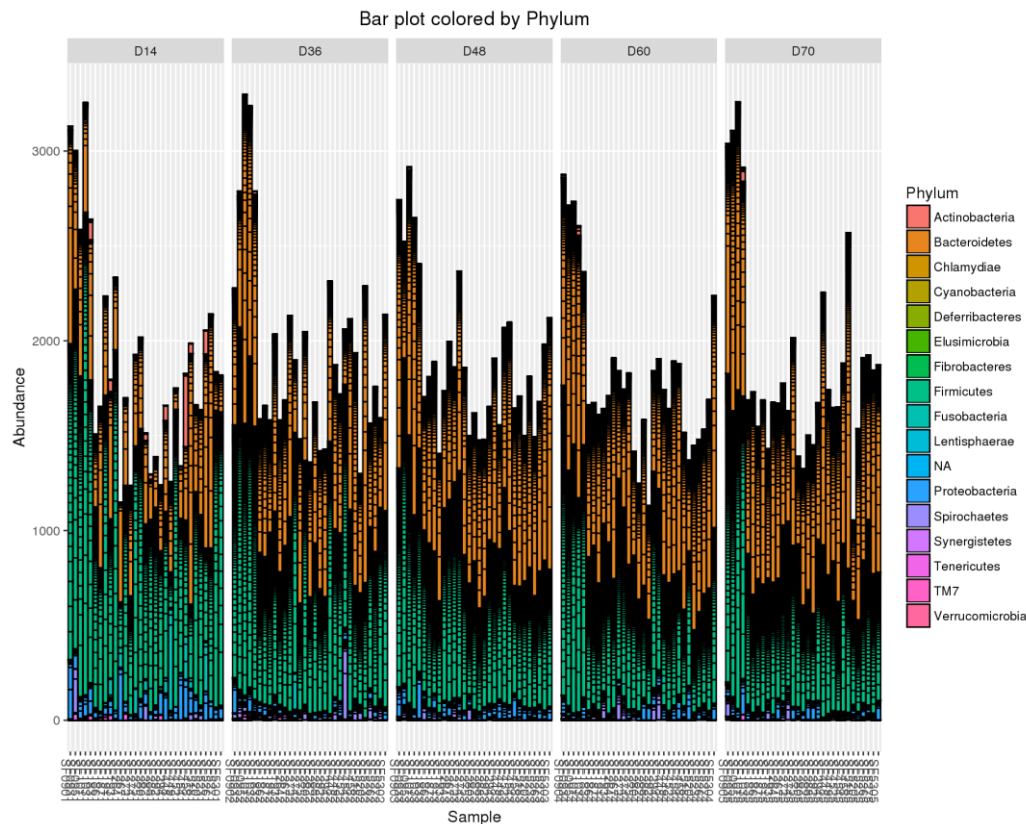
Plot bar coloured at the Phylum level on raw counts

➔ Clearly, samples are not sequenced at the same depth

➔ Data have to be rarefied

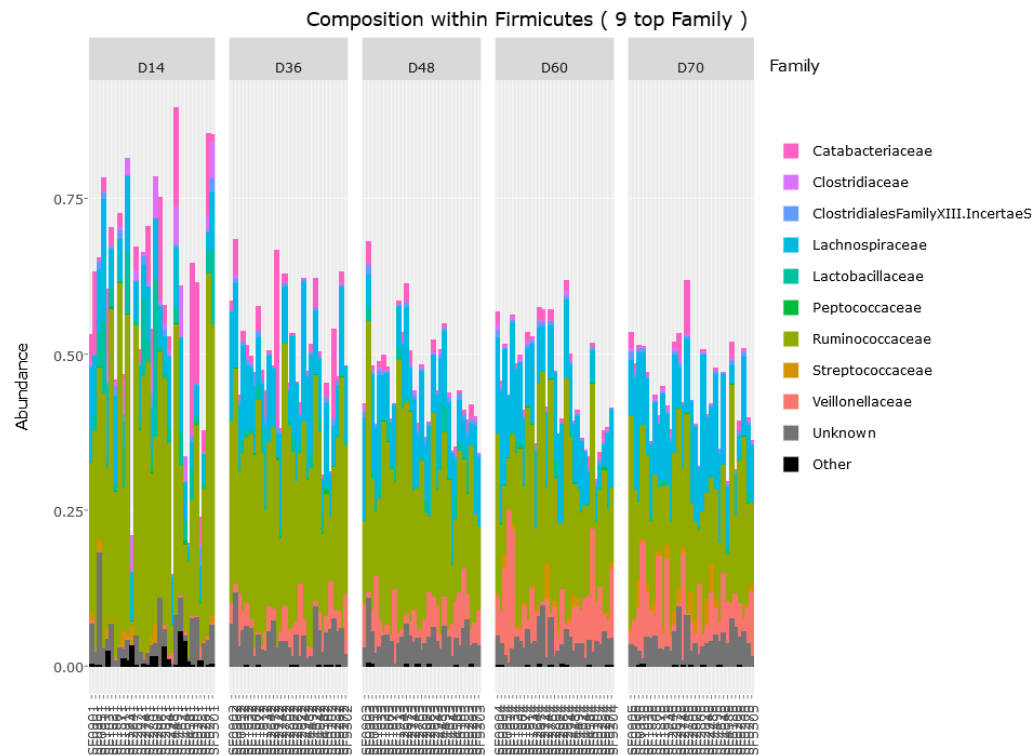# Exercise B-2

➔ What can you conclude with the composition plots ?

Composition within Root (5 top Phylum)
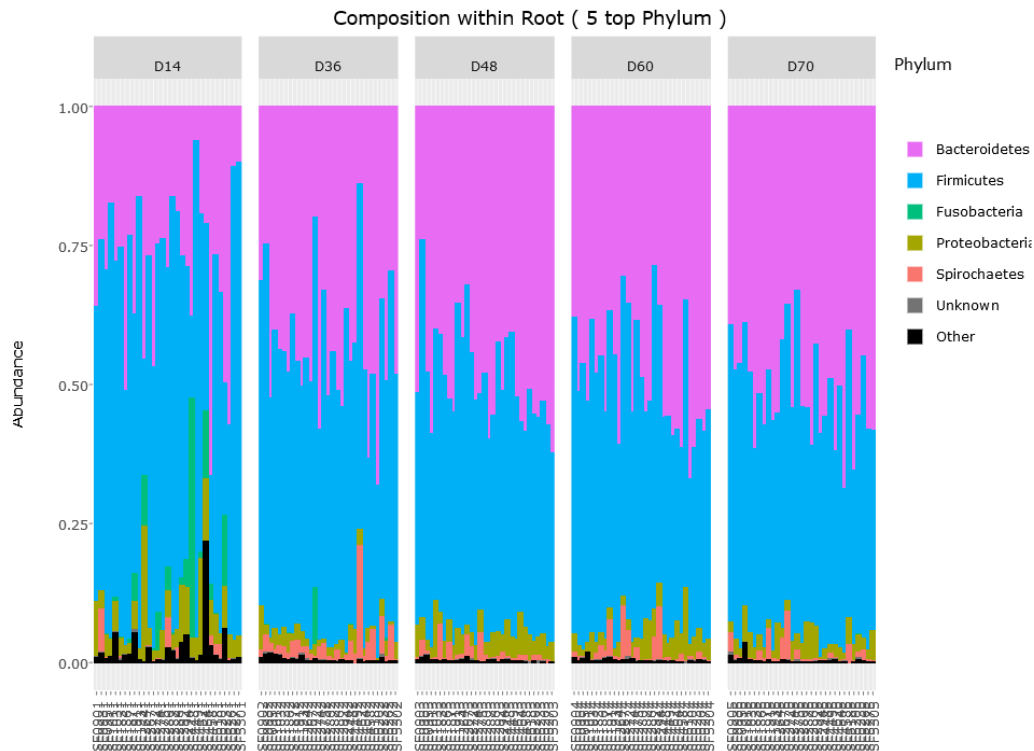
Composition plot of the 5 top Phylum coloured at the Phylum level on rarefied counts

➔ The 2 most abundant Phylum are the Firmicutes and the Bacteroidetes

Exercise B-2

➔ What can you conclude with the composition plots ?

Composition within Firmicutes (9 top Family)

Composition plot of the 9 top Firmicutes families coloured at the Family level on rarefied counts

➔ Veillonellaceae seems to rise after weaning, but the Firmicutes are not drastically change

Exercise B-2

➔ What can you conclude with the composition plots ?

Composition within Bacteroidetes (9 top Family)
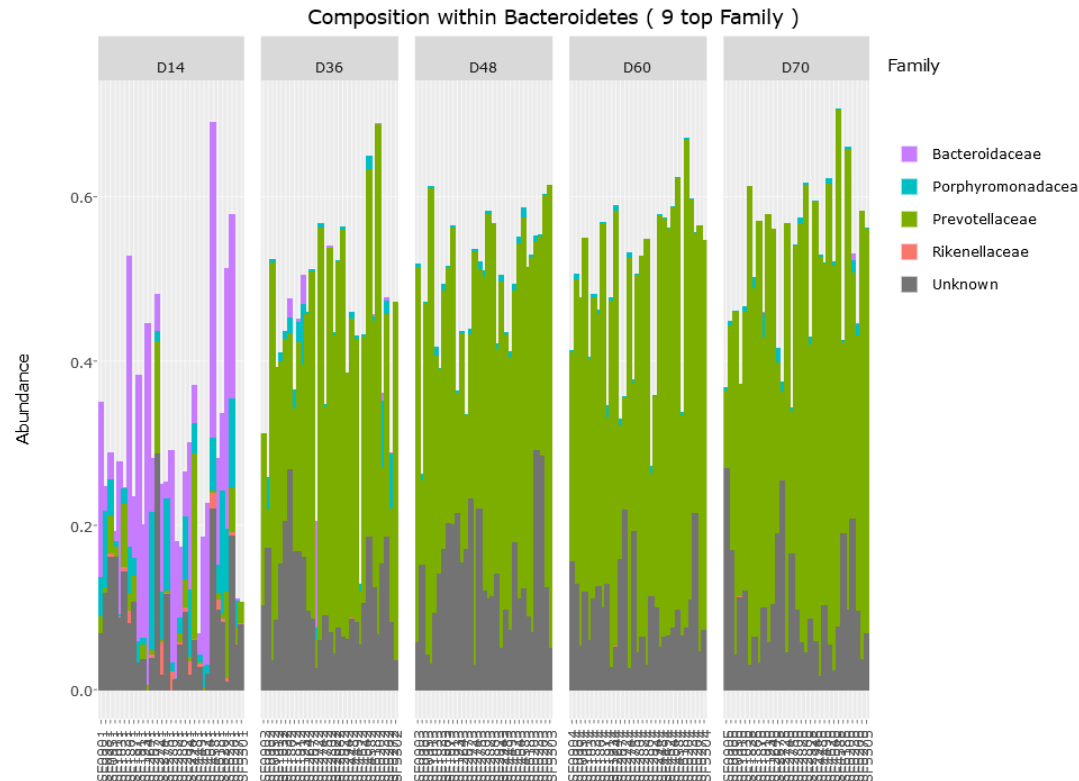
Composition plot of the 9 top Bacteroidetes families coloured at the Family level on rarefied counts

➔ After weaning Bacteroidetes composition has clearly changed.

Exercise B-2

➔ What about alpha diversity indices ?

Interpretation

Diversity increases with time (with strong housing effect)

Low shannon/InvSimpson diversities compared to Observed, Chao1

➔ communities are dominated by a moderate number of abundant taxa

Alpha diversity distribution in function of Time
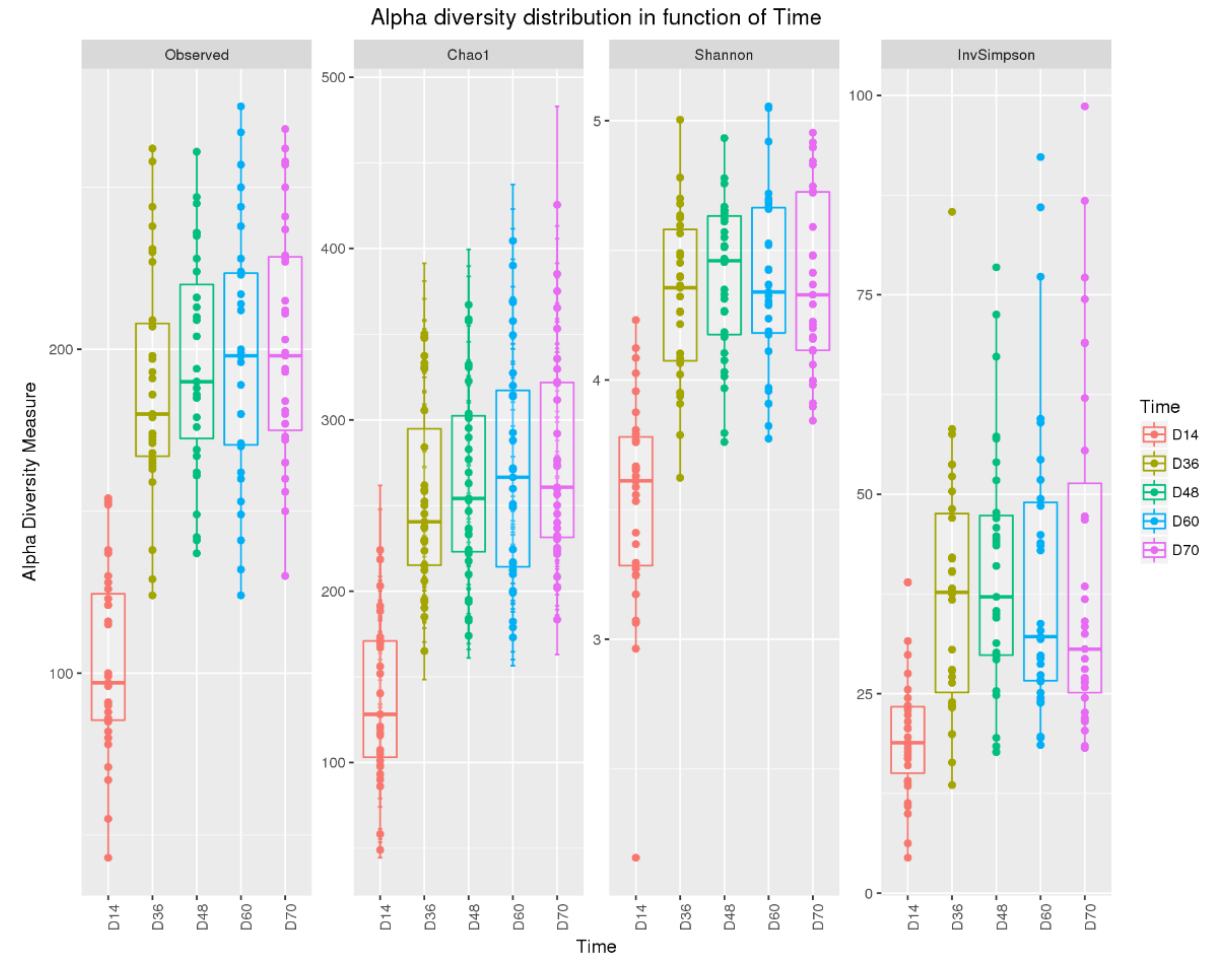
Exercise B-2

Effective diversities are more robust to depth bias

➔ Either correct for depth or perform rarefaction before comparing diversities
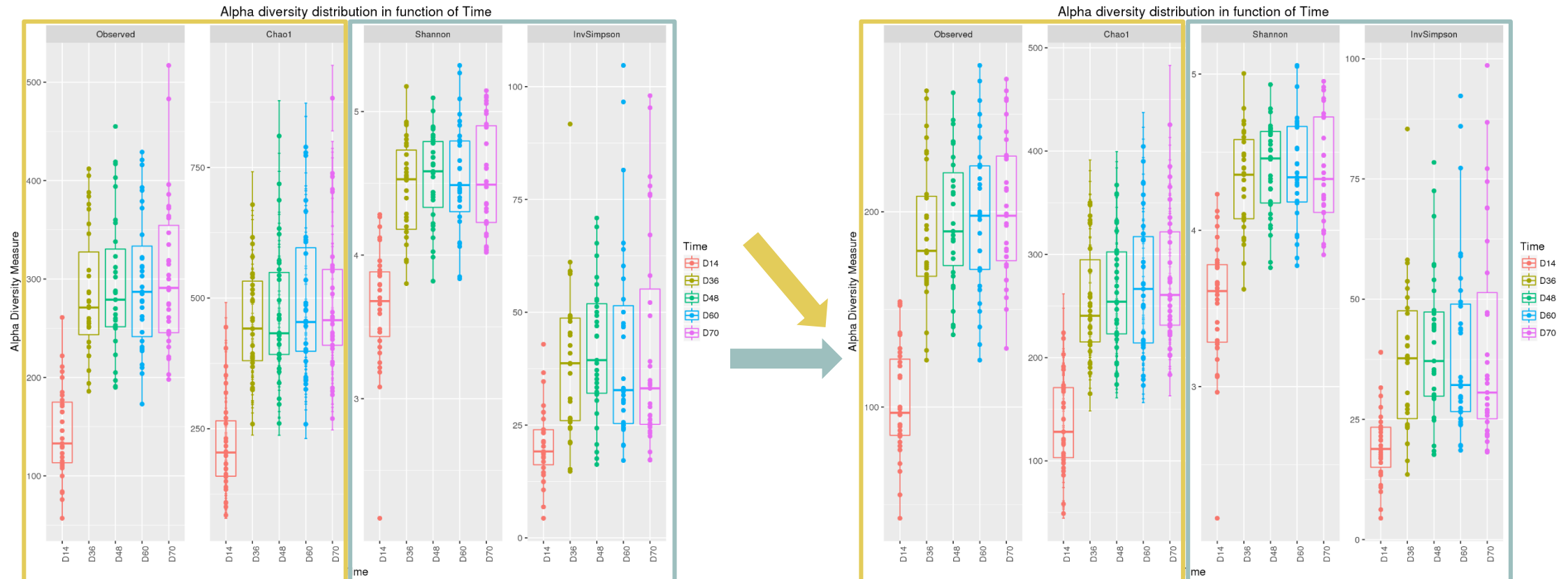
Alpha diversity indices on raw counts

Alpha diversity indices on rarefied counts

Exercise B-3

➔ Now, how to analyse the OTU/sample structure?

  ➔ First step is to compute distance matrix : beta diversities also called dissimilarities

  ➔ Then use it to :

  ▪ represent samples in a 2D graphic that best respect this distance matrix.

  ▪ test that clustering samples based on dissimilarities looks like expected.

  ▪ construct heatmap to discover if samples/OTUs are connected.

Exercise B-4

Test the 4 most common distances.

→ Can you conclude something based on distance matrix comparison

→ Can you conclude something based on the graphical representations of samples/OTUs, which type of distance fit the most our data ?
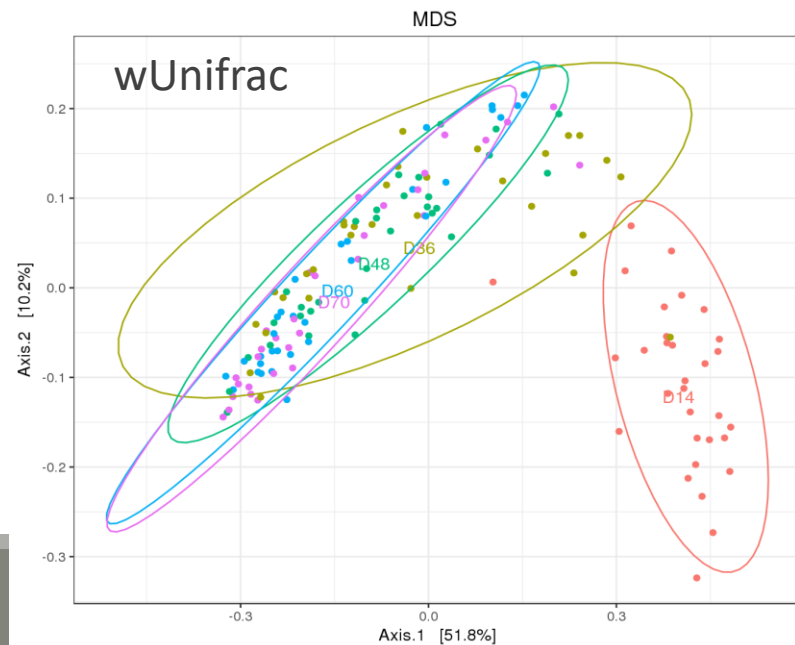
Exercise B-4

➔Can you conclude something based on distance matrix comparison
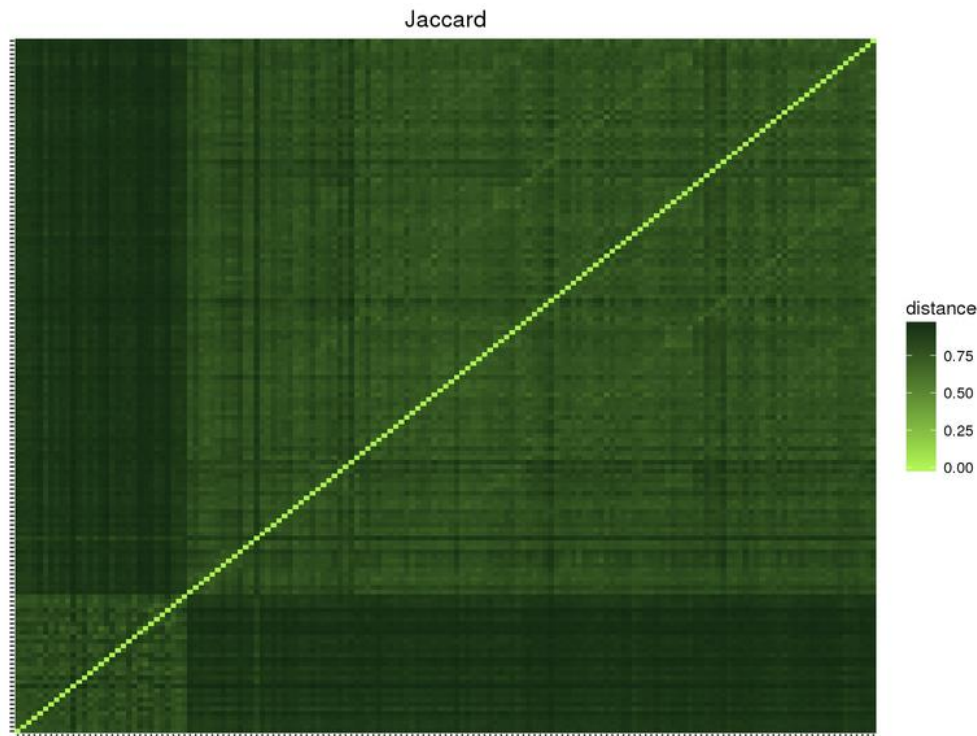
Jaccard higher than Bray-Curtis ➔ abundant taxa are shared

Exercise B-4

➔ Can you conclude something based on distance matrix comparison
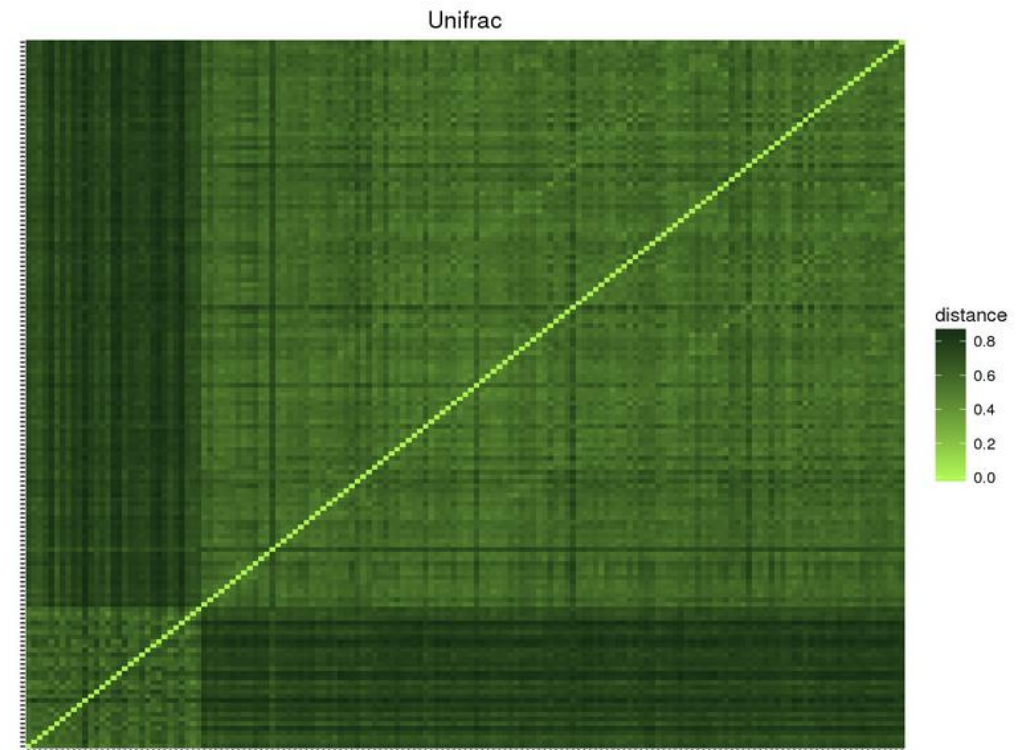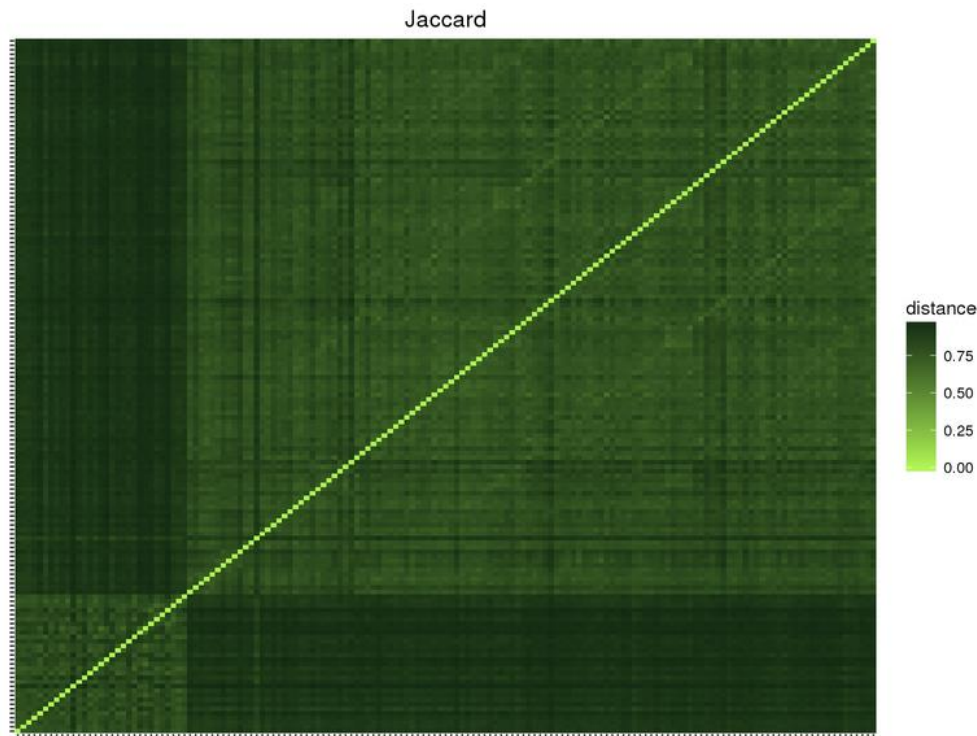
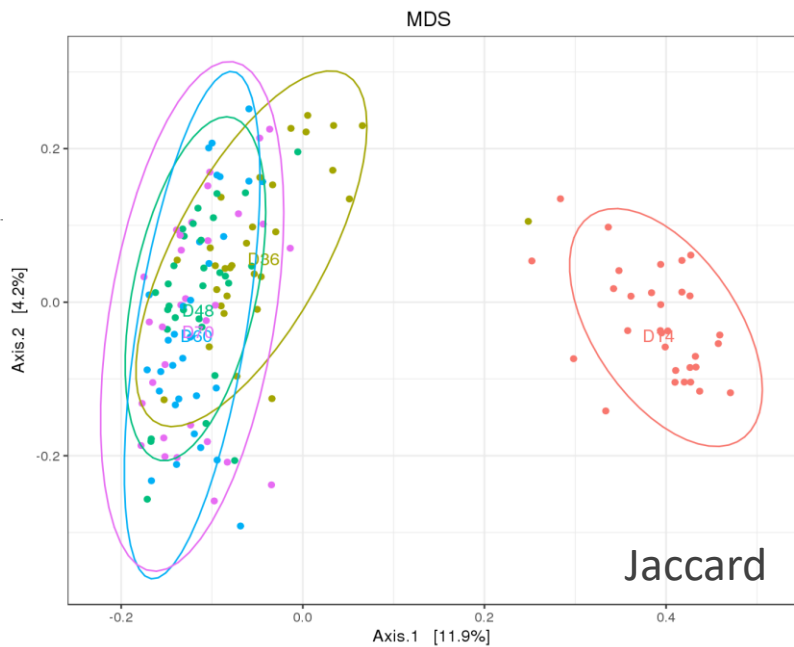Jaccard higher than Unifrac   ➔ community taxa are distinct but phylogenetically related

Exercise B-4

➔ Can you conclude something based on distance matrix comparison
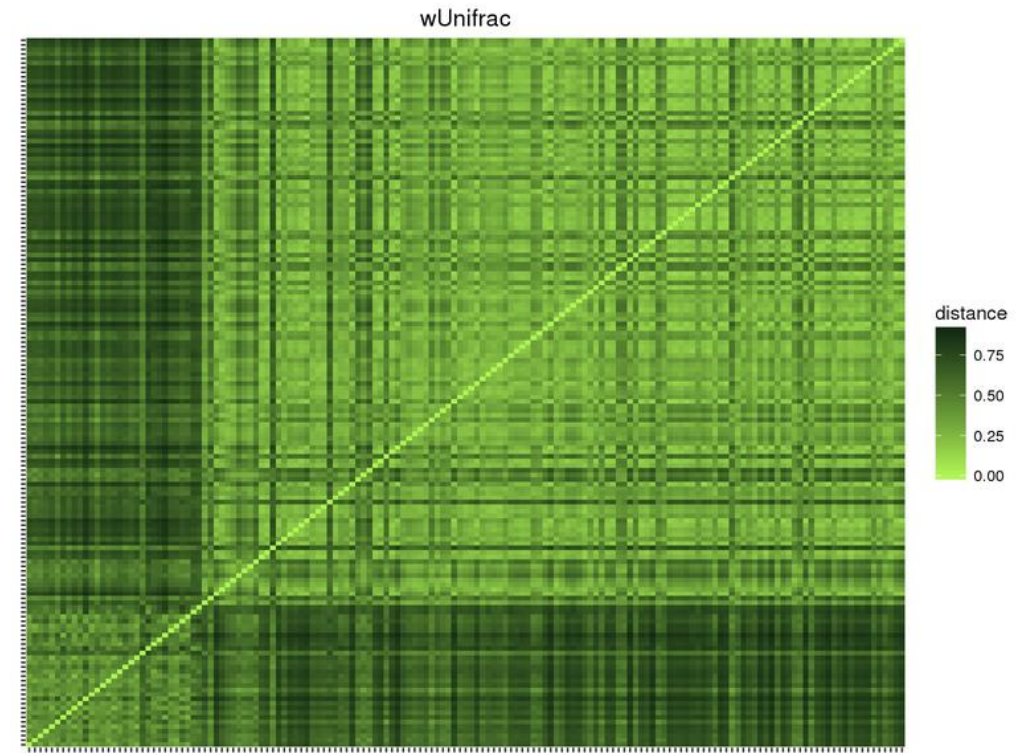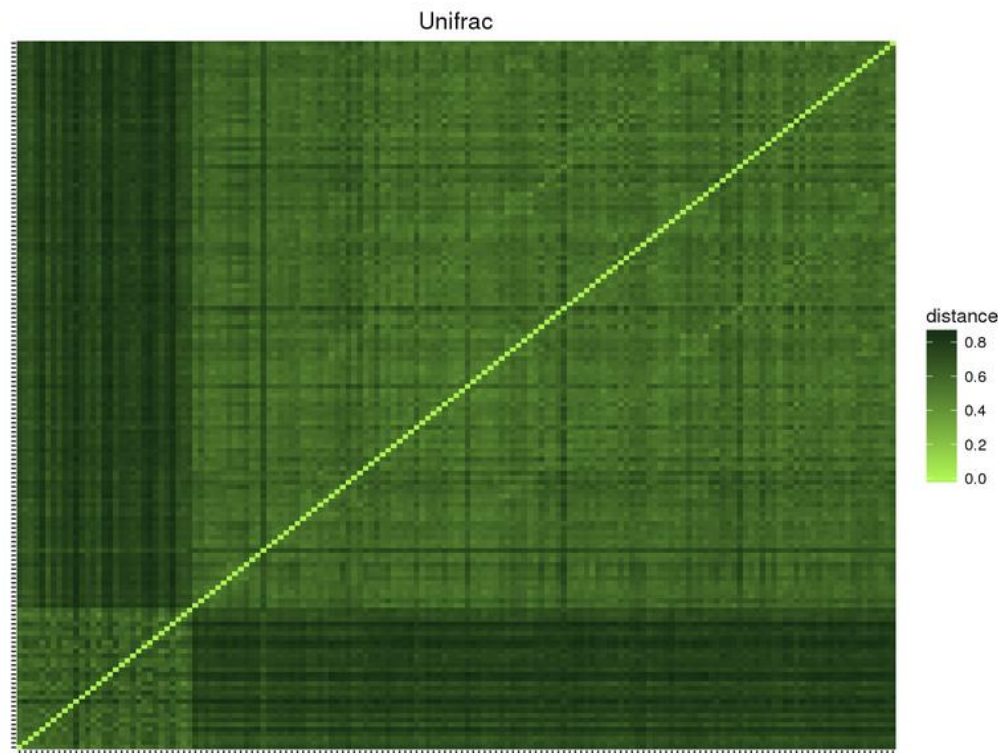
Unifrac higher than weighted Unifrac ➔ abundant taxa in communities are phylogenetically close
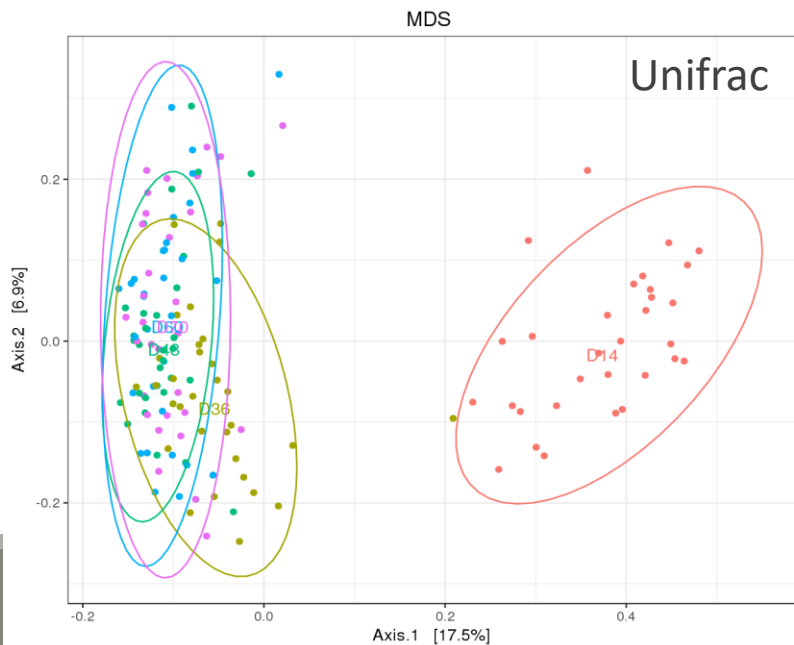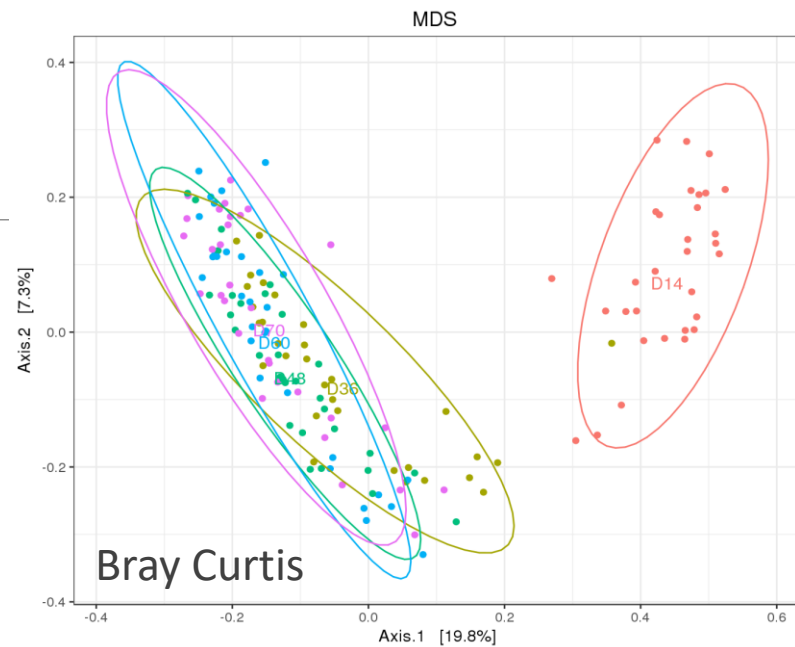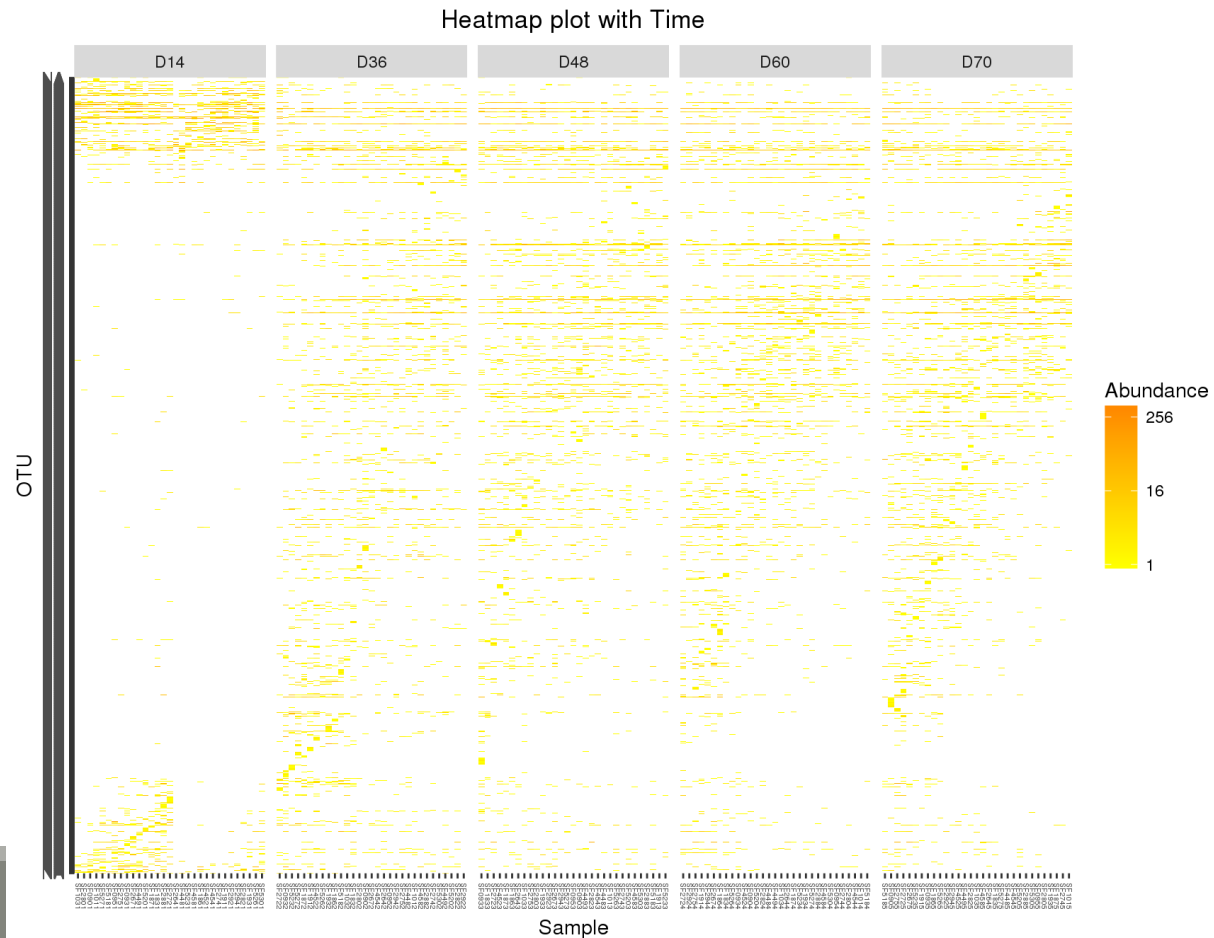
Exercise B-4

➔ Based on the graphical representations of samples/OTUs, which type of distance fits the most our data ?

- Qualitative distances (Unifrac, Jaccard) separate D14 and the rest.

- weighted Unifrac mixes up some samples: the taxa separating D14 from the rest may be replaced by (phylogenetically) close siblings.

- All distances (weighted Unifrac) exhibit a high gradient corresponding to high heterogeneity of samples on axis 2.

- Distance between groups seems to be smaller with qualitative distances (Jaccard/Unifrac) than quantitative distance ➔ specific species before or after weaning must be pretty rare.

- Warning: The 2-D representation capture only part of the original distances.
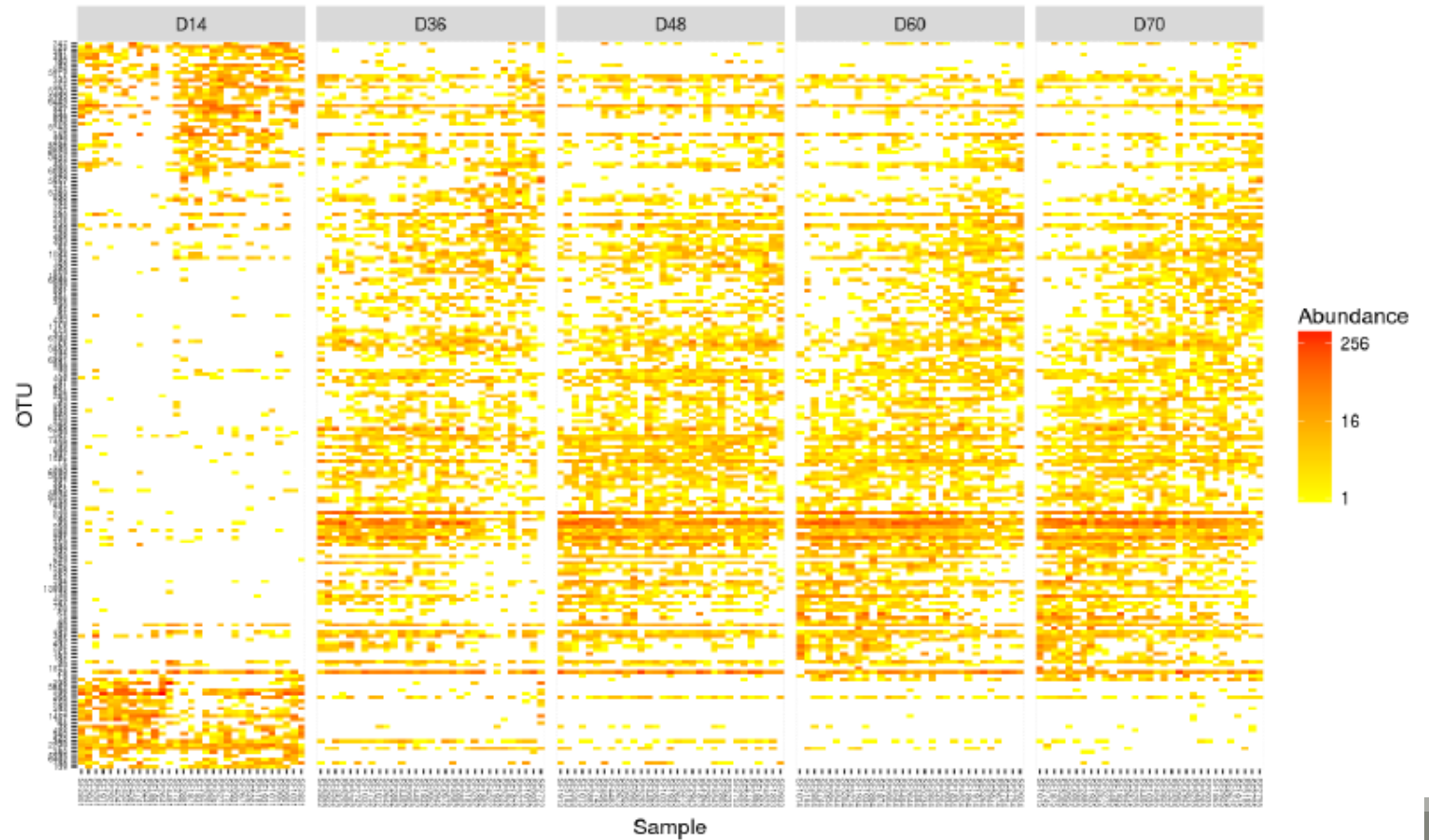
Exercise B-4

➔ Based on the heatmap representation are samples/OTUs connected?

Heatmap plot with Time

Exercise B-4

➔ Based on the heatmap representation are samples/OTUs connected?

Heatmap on 200 most abundant OTU

Exercise B-4

➔ Based on the graphical representations of samples/OTUs, which type of distance fit the most our data ?

Hierarchical clustering plots :

▪ Consistent with the ordination plots, clustering shows a good structure (D14 vs. rest or Weaned FALSE vs TRUE) for the Bray-Curtis distance for the Ward linkage

▪ Different distances would result (in this case) in similar results.

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

Sample Clustering with Ward.D2 linkage
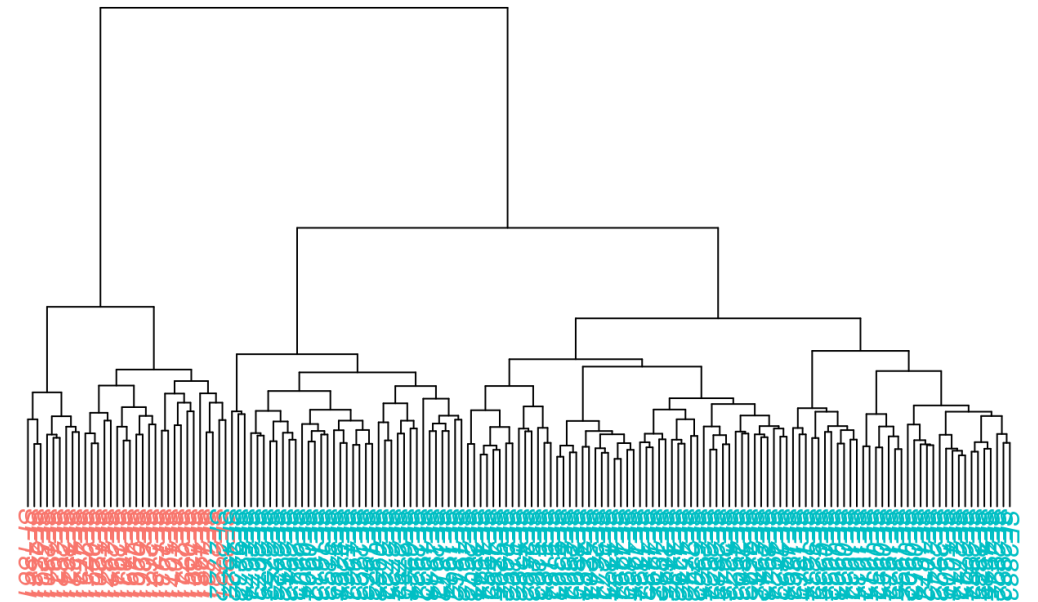
Exercise B-5

We found that Time or Weaned seems to have an effect on sample diversities.

➔ How can we measure this effect ?

➔ by performing a multivariate analysis of the variance

FROGSSTAT Phyloseq Multivariate Analysis Of Variance (Galaxy Version 1.0.0)   ▼ Options

Phyloseq object (format rdata)

8: kinetic_normalized.Rdata
This is the result of FROGS Phyloseq Import Data tool.

The beta diversity distance matrix file

23: FROGSSTAT Phyloseq Beta Diversity: beta_diversity (Bray_Curtis.tsv)
This file is the result of FROGS Phyloseq Beta Diversity tool.

Experiment variable

Time
The experiment variable that you want to analyse.

✔ Execute

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

Permutation: free
Number of permutations: 9999

Terms added sequentially (first to last)

              Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
Time           4     9.560  2.3899  9.6484 0.20464  1e-04 ***
Residuals    150    37.155  0.2477         0.79536
Total        154    46.714                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Time explains significantly around
20% of the beta diversity variance

Exercise B-5

Comment:

You can use more complexe formula:

- to analyse multiple variable at the same time

FROGSSTAT Phyloseq Multivariate Analysis Of Variance (Galaxy Version 1.0.0) ▼ Options

Phyloseq object (format rdata)

8: kinetic_normalized.Rdata

This is the result of FROGS Phyloseq Import Data tool.

The beta diversity distance matrix file

23: FROGSSTAT Phyloseq Beta Diversity: beta_diversity (Bray_Curtis.tsv)

This file is the result of FROGS Phyloseq Beta Diversity tool.

Experiment variable

Weaned + sex

The experiment variable that you want to analyse.

✔ Execute

```
Call:
adonis(formula = dist ~ Weaned + sex, data = metadata, permutations = 9999)

Permutation: free
Number of permutations: 9999

Terms added sequentially (first to last)

              Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
Weaned         1     7.840  7.8397 30.9042 0.16782 0.0001 ***
sex            1     0.315  0.3155  1.2437 0.00675 0.1583
Residuals    152    38.559  0.2537         0.82542
Total        154    46.714                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only Weaned has an effect and it explains significantly around 17% of the beta diversity variance

Exercise B-5

Comment:

You can use more complexe formula:

- to analyse multiple variable at the same time
- to analyse variable interaction


```
Call:
adonis(formula = dist ~ Time * Bande + sex, data = metadata,        permutations = 9999)

Permutation: free
Number of permutations: 9999

Terms added sequentially (first to last)

             Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
Time          4     9.560 2.38988 10.3916 0.20464 0.0001 ***
Bande         5     2.804 0.56076  2.4383 0.06002 0.0001 ***
sex           1     0.302 0.30170  1.3118 0.00646 0.1233
Time:Bande   20     5.531 0.27656  1.2025 0.11841 0.0116 *
Residuals   124    28.518 0.22998         0.61048
Total       154    46.714                 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Time and Bande have independantly an effect as well as their combination which explains significantly around 37% of the beta diversity variance

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.

▪ Núria Mach, Mustapha Berri, Jordi Estellé, Florence Levenez, Gaëtan Lemonnier, Catherine Denis, Jean-Jacques Leplat, Claire Chevaleyre, Yvon Billon, Joël Doré, Claire Rogel-Gaillard and Patricia Lepage(2015). Early-life establishment of the swine gut microbiome and impact on host phenotypes. Environmental Microbiology Reports (2015) 7(3), 554–569.

▪ Jacques Ravela, Pawel Gajera, Zaid Abdob, G. Maria Schneiderc, Sara S. K. Koeniga, Stacey L. McCullea, Shara Karlebachd, Reshma Gorlee, Jennifer Russellf, Carol O. Tacketf, Rebecca M. Brotmana, Catherine C. Davisg, Kevin Aultd, Ligia Peraltae, and Larry J. Forneyc (2011). Vaginal microbiome of reproductive-age women. PNAS Vol.108

▪ 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.