

# Lab\_10\_Metagenomics\_MBI3100A\_2022\_Tutorial

2022-11-12

## R Markdown

**Acknowledgement:** The commands and datasets used in this tutorial are adapted from the phyloseq website and Yan Hui Microbiome tutorial.

## Install required libraries

Installing these libraries may take some time. Try to update all other dependencies when prompted (type “a” and enter).

```
if (!require("BiocManager")) install.packages("BiocManager")
```

```
## Loading required package: BiocManager
```

```
if (!require("phyloseq")) BiocManager::install("phyloseq")
```

```
## Loading required package: phyloseq
```

```
if (!require("microbiomeMarker")) BiocManager::install("microbiomeMarker")
```

```
## Loading required package: microbiomeMarker
```

```
## Registered S3 method overwritten by 'gplots':  
##   method      from  
##   reorder.factor DescTools
```

```
##  
## Attaching package: 'microbiomeMarker'
```

```
## The following object is masked from 'package:phyloseq':  
##  
##   plot_heatmap
```

```
if (!require("tidyverse")) install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## Found more than one class "atomicVector" in cache; using the first, from namespace 'Matrix'
```



```

if (!require("dendextend")) install.packages("dendextend")

## Loading required package: dendextend
## Registered S3 method overwritten by 'dendextend':
##   method      from
##   rev.hclust  vegan
##
## -----
## Welcome to dendextend version 1.16.0
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## You may ask questions at stackoverflow, use the r and dendextend tags:
##   https://stackoverflow.com/questions/tagged/dendextend
##
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))
## -----
##
##
## Attaching package: 'dendextend'
##
## The following object is masked from 'package:stats':
##
##   cutree

```

## Load Libraries

```

library(phyloseq)
library(ggplot2)
library(dplyr)
library(dendextend)
library(microbiomeMarker)

```

## List of packages available in phyloseq

Phyloseq come with preloaded datasets. The datasets can be explored using the following commands. The column named "Item" contains the list of all the datasets available in phyloseq package

```

phyloseq_datasets = data(package = "phyloseq")
phyloseq_datasets$results

```

```

##      Package   LibPath                                     Item
## [1,] "phyloseq" "C:/Users/HP/AppData/Local/R/win-library/4.2" "GlobalPatterns"
## [2,] "phyloseq" "C:/Users/HP/AppData/Local/R/win-library/4.2" "enterotype"
## [3,] "phyloseq" "C:/Users/HP/AppData/Local/R/win-library/4.2" "esophagus"

```

```
## [4,] "phyloseq" "C:/Users/HP/AppData/Local/R/win-library/4.2" "soilrep"
##      Title
## [1,] "(Data) Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample (2011)"
## [2,] "(Data) Enterotypes of the human gut microbiome (2011)"
## [3,] "(Data) Small example dataset from a human esophageal community (2004)"
## [4,] "(Data) Reproducibility of soil microbiome data (2011)"
```

```
# To load the GlobalPatterns dataset
data(GlobalPatterns)
```

```
GlobalPatterns
```

```
## phyloseq-class experiment-level object
## otu_table() OTU Table: [ 19216 taxa and 26 samples ]
## sample_data() Sample Data: [ 26 samples by 7 sample variables ]
## tax_table() Taxonomy Table: [ 19216 taxa by 7 taxonomic ranks ]
## phy_tree() Phylogenetic Tree: [ 19216 tips and 19215 internal nodes ]
```

## 1. Data exploration

GlobalPatterns is a phyloseq object which contains 1 out\_table, 1 sample\_data table, 1 taxonomy table and 1 phylogenetic tree.

We can access the different data type of data and table using following commands

```
# otu_table(GlobalPatterns) %>% head()
# sample_data(GlobalPatterns) %>% head()
# tax_table(GlobalPatterns) %>% head()
# phy_tree(GlobalPatterns) %>% head()
```

For example: How many samples are there in the ‘GlobalPatterns’ data set?

```
sample_data(GlobalPatterns)
```

```
##      X.SampleID Primer Final_Barcode Barcode_truncated_plus_T
## CL3          CL3 ILBC_01      AACGCA      TCGGTT
## CC1          CC1 ILBC_02      AACTCG      CGAGTT
## SV1          SV1 ILBC_03      AACTGT      ACAGTT
## M31Fcsw      M31Fcsw ILBC_04      AAGAGA      TCTCTT
## M11Fcsw      M11Fcsw ILBC_05      AAGCTG      CAGCTT
## M31Plmr      M31Plmr ILBC_07      AATCGT      ACGATT
## M11Plmr      M11Plmr ILBC_08      ACACAC      GTGTGT
## F21Plmr      F21Plmr ILBC_09      ACACAT      ATGTGT
## M31Tong      M31Tong ILBC_10      ACACGA      TCGTGT
## M11Tong      M11Tong ILBC_11      ACACGG      CCGTGT
## LMEpi24M     LMEpi24M ILBC_13      AACTGT      CAGTGT
## SLEpi20M     SLEpi20M ILBC_15      ACAGAG      CTCTGT
## AQC1cm       AQC1cm ILBC_16      ACAGCA      TGCTGT
## AQC4cm       AQC4cm ILBC_17      ACAGCT      AGCTGT
```

##	AQC7cm	AQC7cm ILBC_18	ACAGTG	CACTGT
##	NP2	NP2 ILBC_19	ACAGTT	AACTGT
##	NP3	NP3 ILBC_20	ACATCA	TGATGT
##	NP5	NP5 ILBC_21	ACATGA	TCATGT
##	TRRsed1	TRRsed1 ILBC_22	ACATGT	ACATGT
##	TRRsed2	TRRsed2 ILBC_23	ACATTC	GAATGT
##	TRRsed3	TRRsed3 ILBC_24	ACCACA	TGTGGT
##	TS28	TS28 ILBC_25	ACCAGA	TCTGGT
##	TS29	TS29 ILBC_26	ACCAGC	GCTGGT
##	Even1	Even1 ILBC_27	ACCGCA	TGCGGT
##	Even2	Even2 ILBC_28	ACCTCG	CGAGGT
##	Even3	Even3 ILBC_29	ACCTGT	ACAGGT
##		Barcode_full_length	SampleType	
##	CL3	CTAGCGTGCGT	Soil	
##	CC1	CATCGACGAGT	Soil	
##	SV1	GTACGCACAGT	Soil	
##	M31Fcsw	TCGACATCTCT	Feces	
##	M11Fcsw	CGACTGCAGCT	Feces	
##	M31Plmr	CGAGTCACGAT	Skin	
##	M11Plmr	GCCATAGTGTG	Skin	
##	F21Plmr	GTAGACATGTG	Skin	
##	M31Tong	TGTGGCTCGTG	Tongue	
##	M11Tong	TAGACACCGTG	Tongue	
##	LMEpi24M	CATGAACAGTG	Freshwater	
##	SLEpi20M	AGCCGACTCTG	Freshwater	
##	AQC1cm	GACCACTGCTG	Freshwater (creek)	
##	AQC4cm	CAAGCTAGCTG	Freshwater (creek)	
##	AQC7cm	ATGAAGCACTG	Freshwater (creek)	
##	NP2	TCGCGCAACTG	Ocean	
##	NP3	GCTAAGTGATG	Ocean	
##	NP5	GAACGATCATG	Ocean	
##	TRRsed1	CACGTGACATG	Sediment (estuary)	
##	TRRsed2	TGCGCTGAATG	Sediment (estuary)	
##	TRRsed3	GATGTATGTGG	Sediment (estuary)	
##	TS28	GCATCGTCTGG	Feces	
##	TS29	CTAGTCGCTGG	Feces	
##	Even1	TGACTCTGCGG	Mock	
##	Even2	TCTGATCGAGG	Mock	
##	Even3	AGAGAGACAGG	Mock	
##			Description	
##	CL3	Calhoun South Carolina Pine soil, pH 4.9		
##	CC1	Cedar Creek Minnesota, grassland, pH 6.1		
##	SV1	Sevilleta new Mexico, desert scrub, pH 8.3		
##	M31Fcsw	M3, Day 1, fecal swab, whole body study		
##	M11Fcsw	M1, Day 1, fecal swab, whole body study		
##	M31Plmr	M3, Day 1, right palm, whole body study		
##	M11Plmr	M1, Day 1, right palm, whole body study		
##	F21Plmr	F1, Day 1, right palm, whole body study		
##	M31Tong	M3, Day 1, tongue, whole body study		
##	M11Tong	M1, Day 1, tongue, whole body study		
##	LMEpi24M	Lake Mendota Minnesota, 24 meter epilimnion		
##	SLEpi20M	Sparkling Lake Wisconsin, 20 meter epilimnion		
##	AQC1cm	Allequash Creek, 0-1cm depth		
##	AQC4cm	Allequash Creek, 3-4 cm depth		

```
## AQC7cm           Allequash Creek, 6-7 cm depth
## NP2             Newport Pier, CA surface water, Time 1
## NP3             Newport Pier, CA surface water, Time 2
## NP5             Newport Pier, CA surface water, Time 3
## TRRsed1         Tijuana River Reserve, depth 1
## TRRsed2         Tijuana River Reserve, depth 2
## TRRsed3         Tijuana River Reserve, depth 2
## TS28            Twin #1
## TS29            Twin #2
## Even1           Even1
## Even2           Even2
## Even3           Even3
```

**Answer: 26**

Access the OTU table from a dataset

```
otu_table(GlobalPatterns) %>% head()
```

Since the OTU tables are very big we use the command `head` to read the top 6 rows of a matrix/dataframe

```
## OTU Table:           [6 taxa and 26 samples]
##                      taxa are rows
##      CL3 CC1 SV1 M31Fcsw M11Fcsw M31Plmr M11Plmr F21Plmr M31Tong M11Tong
## 549322    0  0  0      0      0      0      0      0      0      0
## 522457    0  0  0      0      0      0      0      0      0      0
## 951       0  0  0      0      0      0      1      0      0      0
## 244423    0  0  0      0      0      0      0      0      0      0
## 586076    0  0  0      0      0      0      0      0      0      0
## 246140    0  0  0      0      0      0      0      0      0      0
##      LMEpi24M SLEpi20M AQC1cm AQC4cm AQC7cm NP2 NP3 NP5 TRRsed1 TRRsed2
## 549322      0      1      27      100      130  1  0  0      0      0
## 522457      0      0      0       2       6  0  0  0      0      0
## 951        0      0      0       0       0  0  0  0      0      0
## 244423      0      0      0      22      29  0  0  0      0      0
## 586076      0      0      0       2       1  0  0  0      0      0
## 246140      0      0      0       1       3  0  0  0      0      0
##      TRRsed3 TS28 TS29 Even1 Even2 Even3
## 549322      0  0  0      0      0      0
## 522457      0  0  0      0      0      0
## 951        0  0  0      0      0      0
## 244423      0  0  0      0      0      0
## 586076      0  0  0      0      0      0
## 246140      0  0  0      0      0      0
```

```
otu_table(GlobalPatterns)[1:5, 1:5]
```

Another way to see the content of a big table

```
## OTU Table:           [5 taxa and 5 samples]
##                      taxa are rows
##      CL3 CC1 SV1 M31Fcsw M11Fcsw
## 549322    0  0  0        0        0
## 522457    0  0  0        0        0
## 951       0  0  0        0        0
## 244423    0  0  0        0        0
## 586076    0  0  0        0        0
```

```
sample_data(GlobalPatterns) %>% head()
```

```
##      X.SampleID Primer Final_Barcode Barcode_truncated_plus_T
## CL3           CL3 ILBC_01      AACGCA                      TGC GTT
## CC1           CC1 ILBC_02      AACTCG                      CGA GTT
## SV1           SV1 ILBC_03      AACTGT                      ACA GTT
## M31Fcsw      M31Fcsw ILBC_04     AAGAGA                      TCTCTT
## M11Fcsw      M11Fcsw ILBC_05     AAGCTG                      CAGCTT
## M31Plmr      M31Plmr ILBC_07     AATCGT                      ACGATT
##      Barcode_full_length SampleType
## CL3           CTAGCGTGCGT      Soil
## CC1           CATCGACGAGT      Soil
## SV1           GTACGCACAGT      Soil
## M31Fcsw       TCGACATCTCT      Feces
## M11Fcsw       CGACTGCAGCT      Feces
## M31Plmr       CGAGTCACGAT      Skin
##
##      Description
## CL3      Calhoun South Carolina Pine soil, pH 4.9
## CC1      Cedar Creek Minnesota, grassland, pH 6.1
## SV1      Sevilleta new Mexico, desert scrub, pH 8.3
## M31Fcsw   M3, Day 1, fecal swab, whole body study
## M11Fcsw   M1, Day 1, fecal swab, whole body study
## M31Plmr   M3, Day 1, right palm, whole body study
```

Access the sample data table and the column content from a dataset

```
# To access the variables in the column 'SampleType'
# The column 'SampleType' is of class factor so get the levels using the command
sample_data(GlobalPatterns)$SampleType %>% levels()
```

```
## [1] "Feces"      "Freshwater"  "Freshwater (creek)"
## [4] "Mock"       "Ocean"       "Sediment (estuary)"
## [7] "Skin"       "Soil"        "Tongue"
```

```
sample_data(GlobalPatterns)$SampleType %>% levels() %>% as.data.frame()
```

Read with as.data.frame to see the results in a tabular format

```
##
## 1      Feces
## 2      Freshwater
## 3 Freshwater (creek)
## 4      Mock
## 5      Ocean
## 6 Sediment (estuary)
## 7      Skin
## 8      Soil
## 9      Tongue
```

How many sample types are available under the SampleType column?

*Answer: 9, "Feces", "Freshwater", "Freshwater (creek)", "Mock", "Ocean", "Sediment (estuary)", "Skin"*

Explore the taxonomy table

```
tax_table(GlobalPatterns) %>% head() %>% DT::datatable()
```

**Kingdom Phylum Class Order Family Genus Species**

```
tax_table(GlobalPatterns) %>% head()
```



```
## Taxonomy Table:      [6 taxa by 7 taxonomic ranks]:
##      Kingdom   Phylum   Class       Order       Family
## 549322 "Archaea" "Crenarchaeota" "Thermoprotei" NA       NA
## 522457 "Archaea" "Crenarchaeota" "Thermoprotei" NA       NA
## 951    "Archaea" "Crenarchaeota" "Thermoprotei" "Sulfolobales" "Sulfolobaceae"
## 244423 "Archaea" "Crenarchaeota" "Sd-NA"      NA       NA
## 586076 "Archaea" "Crenarchaeota" "Sd-NA"      NA       NA
## 246140 "Archaea" "Crenarchaeota" "Sd-NA"      NA       NA
##      Genus      Species
## 549322 NA        NA
## 522457 NA        NA
## 951    "Sulfolobus" "Sulfolobusacidocaldarius"
## 244423 NA        NA
## 586076 NA        NA
## 246140 NA        NA
```

## 2. Importing data in R as phyloseq object

We will combine a phyloseq object using `otu_table`, `sample_data` and taxonomy file. We will read these three files and then combine them to make a phyloseq object to work with them,

```
#data_dir =
```

### OTU table

```
# To import otus column as rownames, as required by phyloseq
GP_sp_tutorial_otu_table = read.table("./tutorial_files/GP_sp_tutorial_otu_table_df.csv",
  sep = "\t",
  header = T,
  row.names = "otus")
GP_sp_tutorial_otu_table[1:5, 1:5]
```

```
##      M31Plmr M11Plmr F21Plmr M31Tong M11Tong
## 951         0       1       0       0       0
## 155495      0       0       0       0       0
## 1029        0       0       0       0       0
## 341551      0       0       0       0       0
## 108964      0       2       6       1       0
```

### Sample data

```
# Import with sampleid column as rownames, as required by phyloseq
GP_sp_tutorial_sample_data = read.table("./tutorial_files/GP_sp_tutorial_sample_data_df.csv",
  sep = "\t", header = T,
  row.names = "sampleid")
GP_sp_tutorial_sample_data %>% head()
```

```
##      SampleOrigin X.SampleID  Primer Final_Barcode Barcode_truncated_plus_T
## M31Plmr      Human    M31Plmr ILBC_07      AATCGT          ACGATT
## M11Plmr      Human    M11Plmr ILBC_08      ACACAC          GTGTGT
## F21Plmr      Human    F21Plmr ILBC_09      ACACAT          ATGTGT
## M31Tong      Human    M31Tong ILBC_10      ACACGA          TCGTGT
## M11Tong      Human    M11Tong ILBC_11      ACACGG          CCGTGT
## LMEpi24M  Freshwater  LMEpi24M ILBC_13      ACACTG          CAGTGT
##      Barcode_full_length SampleType
## M31Plmr      CGAGTCACGAT      Skin
## M11Plmr      GCCATAGTGTG      Skin
## F21Plmr      GTAGACATGTG      Skin
## M31Tong      TGTGGCTCGTG      Tongue
## M11Tong      TAGACACCGTG      Tongue
## LMEpi24M      CATGAACAGTG Freshwater
##
##      Description
## M31Plmr      M3, Day 1, right palm, whole body study
## M11Plmr      M1, Day 1, right palm, whole body study
## F21Plmr      F1, Day 1, right palm, whole body study
## M31Tong      M3, Day 1, tongue, whole body study
## M11Tong      M1, Day 1, tongue, whole body study
## LMEpi24M Lake Mendota Minnesota, 24 meter epilimnion
```

## Taxonomy table

```
# To import otus column as rownames, as required by phyloseq
GP_sp_tutorial_tax_table = read.table("./tutorial_files/GP_sp_tutorial_tax_table_df.csv",
  sep = "\t",
  header = T,
  row.names = "otus")
GP_sp_tutorial_tax_table %>% head()
```

```
##      Kingdom      Phylum      Class      Order      Family
## 951      Archaea Crenarchaeota Thermoprotei Sulfolobales Sulfolobaceae
## 155495 Archaea Crenarchaeota Thaumarchaeota Cenarchaeales Cenarchaeaceae
## 1029      Archaea Crenarchaeota Thaumarchaeota Cenarchaeales Cenarchaeaceae
## 341551 Archaea Crenarchaeota Thaumarchaeota Cenarchaeales Cenarchaeaceae
## 108964 Archaea Crenarchaeota Thaumarchaeota Cenarchaeales Cenarchaeaceae
## 330416 Archaea Crenarchaeota Thaumarchaeota Cenarchaeales Cenarchaeaceae
##      Genus      Species
## 951      Sulfolobus Sulfolobusacidocaldarius
## 155495      Cenarchaeum      Cenarchaeumsymbiosum
## 1029      Cenarchaeum      Cenarchaeumsymbiosum
## 341551 Nitrosopumilus      pIVWA5
## 108964 Nitrosopumilus      pIVWA5
## 330416 Nitrosopumilus      pIVWA5
```

```
my_OTU_table = otu_table(GP_sp_tutorial_otu_table, taxa_are_rows = TRUE)
my_OTU_table %>% head()
```

In order to read the OTU table as phyloseq object we need to use the following command

```
## OTU Table:          [6 taxa and 10 samples]
##                      taxa are rows
##      M31Plmr M11Plmr F21Plmr M31Tong M11Tong LMEpi24M SLEpi20M AQC1cm AQC4cm
## 951          0      1      0      0      0      0      0      0      0
## 155495        0      0      0      0      0      0      0      0      0
## 1029          0      0      0      0      0      0      0      0      0
## 341551        0      0      0      0      0      0      0      0      0
## 108964        0      2      6      1      0      1      0      1      0
## 330416        0      0      0      0      0      0      0      0      0
##      AQC7cm
## 951          0
## 155495        0
## 1029          0
## 341551        0
## 108964        1
## 330416        0
```

```
my_Sample_data = sample_data(GP_sp_tutorial_sample_data)
my_Sample_data %>% head()
```

Similarly to read sample data and taxonomy table as phyloseq objects

```
##      SampleOrigin X.SampleID Primer Final_Barcode Barcode_truncated_plus_T
## M31Plmr          Human      M31Plmr ILBC_07      AATCGT      ACGATT
## M11Plmr          Human      M11Plmr ILBC_08      ACACAC      GTGTGT
## F21Plmr          Human      F21Plmr ILBC_09      ACACAT      ATGTGT
## M31Tong          Human      M31Tong ILBC_10      ACACGA      TCGTGT
## M11Tong          Human      M11Tong ILBC_11      ACACGG      CCGTGT
## LMEpi24M Freshwater LMEpi24M ILBC_13      AACTG      CAGTGT
##      Barcode_full_length SampleType
## M31Plmr          CGAGTCACGAT      Skin
## M11Plmr          GCCATAGTGT      Skin
## F21Plmr          GTAGACATGT      Skin
## M31Tong          TGTGGCTCGTG      Tongue
## M11Tong          TAGACACCGTG      Tongue
## LMEpi24M          CATGAACAGTG Freshwater
##
##      Description
## M31Plmr          M3, Day 1, right palm, whole body study
## M11Plmr          M1, Day 1, right palm, whole body study
## F21Plmr          F1, Day 1, right palm, whole body study
## M31Tong          M3, Day 1, tongue, whole body study
## M11Tong          M1, Day 1, tongue, whole body study
## LMEpi24M Lake Mendota Minnesota, 24 meter epilimnion
```

```
# the taxonomy table is required in matrix format
my_tax_table = tax_table(as.matrix(GP_sp_tutorial_tax_table))
my_tax_table %>% head()
```

## Reading taxonomy file as phyloseq object

```
## Taxonomy Table:      [6 taxa by 7 taxonomic ranks]:
##      Kingdom   Phylum   Class       Order
## 951    "Archaea" "Crenarchaeota" "Thermoprotei" "Sulfolobales"
## 155495 "Archaea" "Crenarchaeota" "Thaumarchaeota" "Cenarchaeales"
## 1029   "Archaea" "Crenarchaeota" "Thaumarchaeota" "Cenarchaeales"
## 341551 "Archaea" "Crenarchaeota" "Thaumarchaeota" "Cenarchaeales"
## 108964 "Archaea" "Crenarchaeota" "Thaumarchaeota" "Cenarchaeales"
## 330416 "Archaea" "Crenarchaeota" "Thaumarchaeota" "Cenarchaeales"
##      Family      Genus      Species
## 951    "Sulfolobaceae" "Sulfolobus" "Sulfolobusacidocaldarius"
## 155495 "Cenarchaeaceae" "Cenarchaeum" "Cenarchaeumsymbiosum"
## 1029   "Cenarchaeaceae" "Cenarchaeum" "Cenarchaeumsymbiosum"
## 341551 "Cenarchaeaceae" "Nitrosopumilus" "pIVWA5"
## 108964 "Cenarchaeaceae" "Nitrosopumilus" "pIVWA5"
## 330416 "Cenarchaeaceae" "Nitrosopumilus" "pIVWA5"
```

Combine to make a phyloseq object.

```
my_physeq = phyloseq(my_OTU_table, my_Sample_data, my_tax_table)
my_physeq
```

```
## phyloseq-class experiment-level object
## otu_table() OTU Table:      [ 1413 taxa and 10 samples ]
## sample_data() Sample Data:  [ 10 samples by 8 sample variables ]
## tax_table()  Taxonomy Table: [ 1413 taxa by 7 taxonomic ranks ]
```

'my\_physeq' is now a new phyloseq object which contains the data (OTU table, sample data, and taxonomy table) that we just imported.

```
sample_data(my_physeq)$SampleType %>% as.factor %>% levels()
```

Explore the categories in a sample variable eg SampleType

```
## [1] "Freshwater"      "Freshwater (creek)" "Skin"
## [4] "Tongue"
```

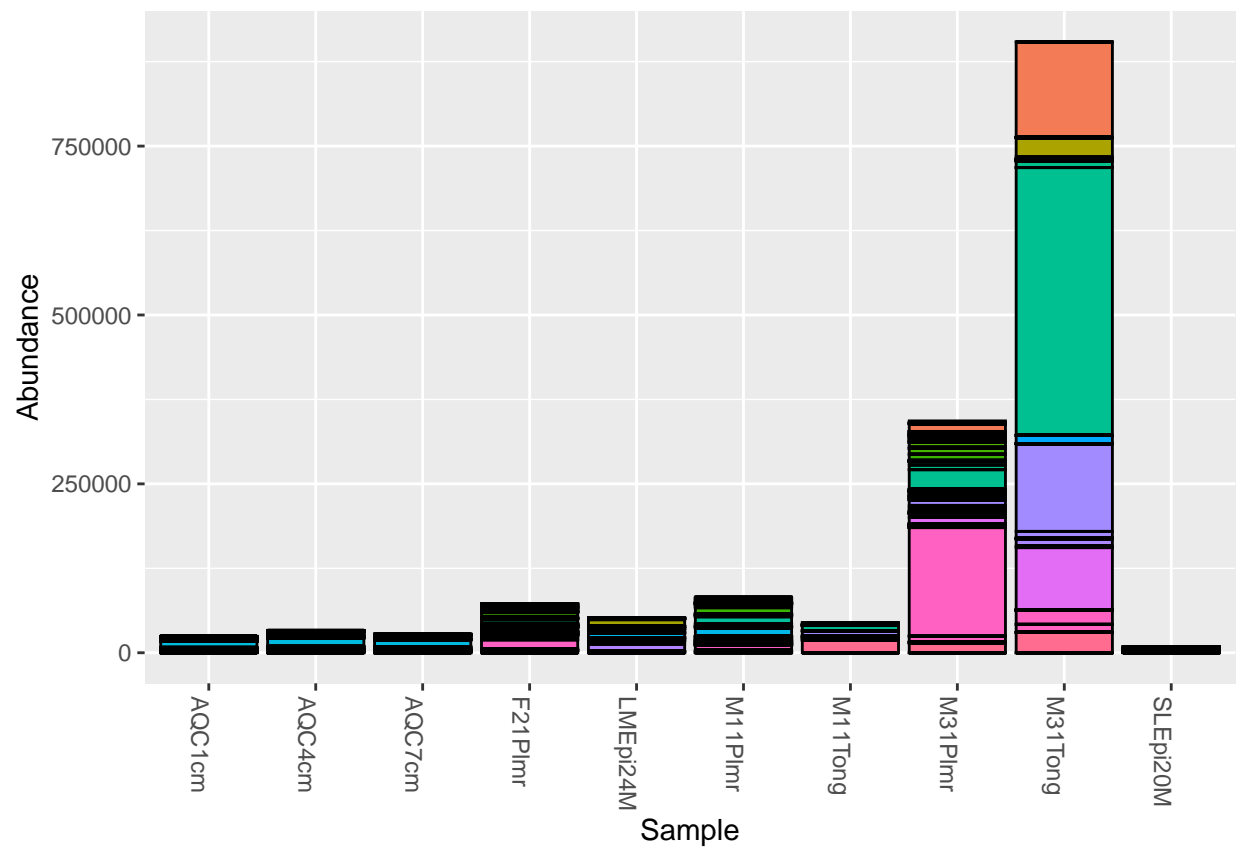
```
sample_data(my_physeq)$SampleOrigin %>% as.factor %>% levels()
```

```
## [1] "Freshwater" "Human"
```

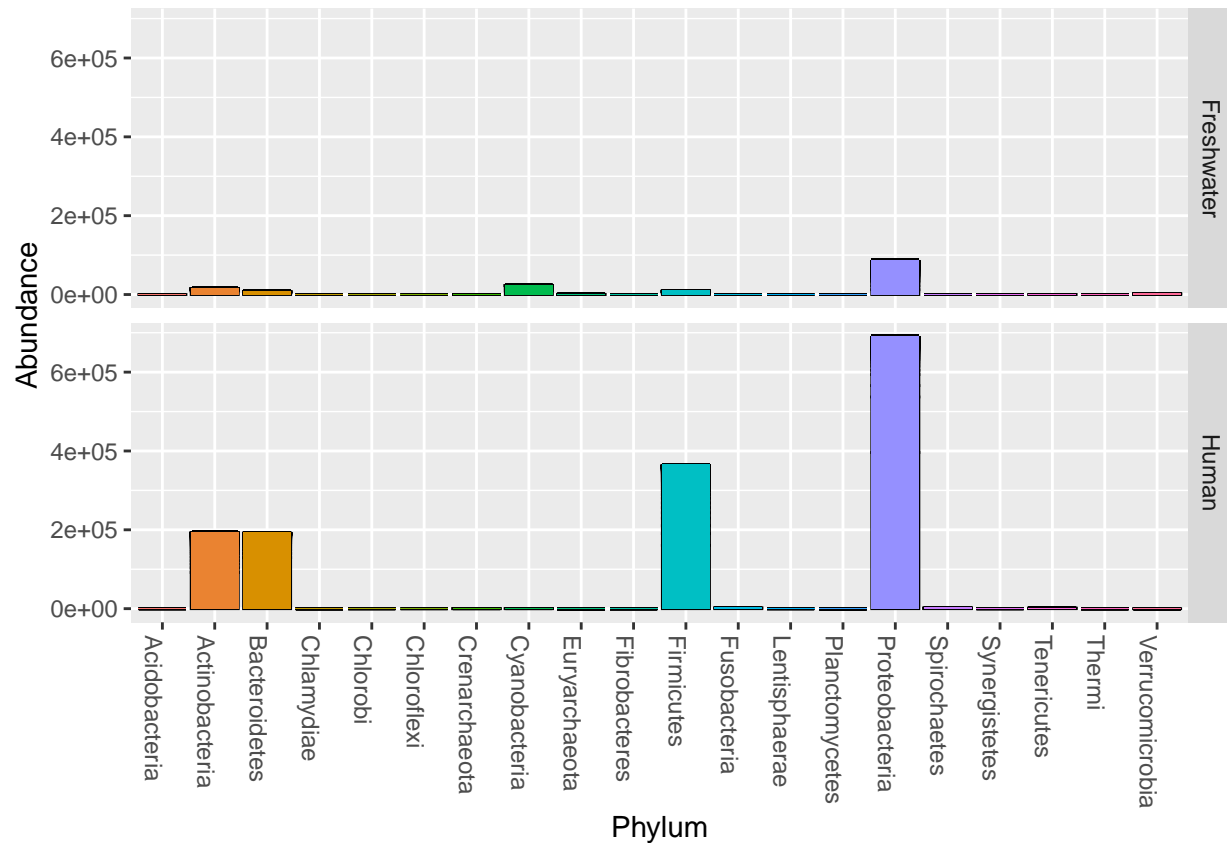
## 3. Data visualization and formatting

Basic plot

```
p = plot_bar(my_physeq, fill = "Species")
p + theme(legend.position="none")
```



```
p = plot_bar(my_physeq, x = "Phylum", fill = "Phylum", facet_grid=SampleOrigin~.)
p + theme(legend.position="none") + geom_bar(stat = "identity")
```



## Relative abundance and filtering

Some of the analysis provide better results when we work with relative abundance data (frequency table for OTU numbers). To avoid the spurious results, we can also filter some taxa which have very low abundance.

```
# To convert to relative abundance
my_physeq_r = transform_sample_counts(my_physeq, function(x) x / sum(x) )
# Keep the taxa which have a mean values at least 1e-5
my_physeq_rf = filter_taxa(my_physeq_r, function(x) mean(x) > 1e-5, TRUE)
my_physeq_rf
```

```
## phyloseq-class experiment-level object
## otu_table() OTU Table: [ 619 taxa and 10 samples ]
## sample_data() Sample Data: [ 10 samples by 8 sample variables ]
## tax_table() Taxonomy Table: [ 619 taxa by 7 taxonomic ranks ]
```

Now the number of remaining taxa after filtering low abundance taxa is 511 out of 1413 in the full dataset.

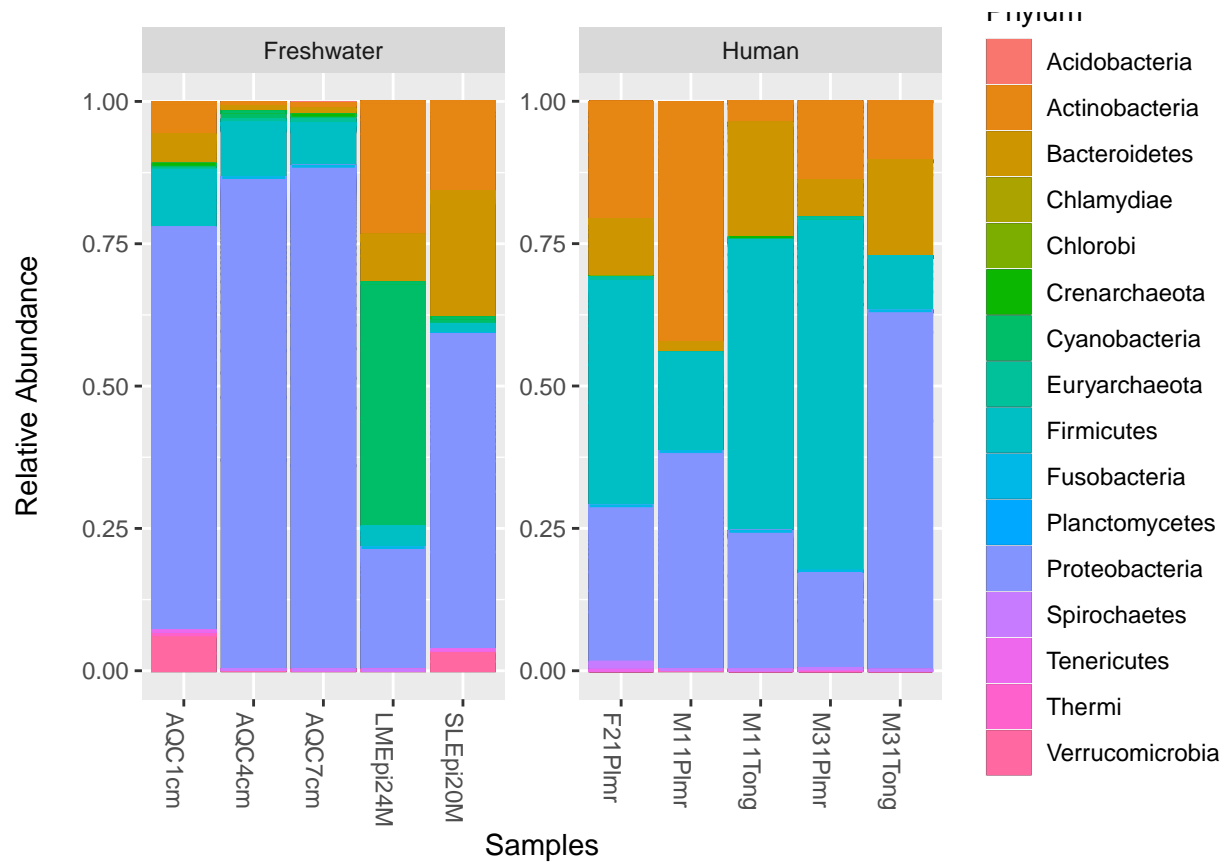
## Plot and compare the relative abundance

Notice the taxa abundant in only one group.

```

phyloseq::plot_bar(my_physeq_rf , fill = "Phylum") +
  geom_bar(aes(color = Phylum, fill = Phylum), stat = "identity", position = "stack") +
  labs(x = "Samples", y = "Relative Abundance\n") +
  facet_wrap(~ SampleOrigin, scales = "free")

```



### An example of how to subset the data based on a condition

Here we will remove all the taxa which have “NA” in the Species column. i.e. Species is unknown for these OTUs

```
GP_subset_1 = subset_taxa(GlobalPatterns, Species!="NA")
```

```
## Found more than one class "phylo" in cache; using the first, from namespace 'phyloseq'
```

```
## Also defined by 'tidytree'
```

```
## Found more than one class "phylo" in cache; using the first, from namespace 'phyloseq'
```

```
## Also defined by 'tidytree'
```

```
## Found more than one class "phylo" in cache; using the first, from namespace 'phyloseq'
```

```
## Also defined by 'tidytree'
```

```
## Found more than one class "phylo" in cache; using the first, from namespace 'phyloseq'

## Also defined by 'tidytree'

## Found more than one class "phylo" in cache; using the first, from namespace 'phyloseq'

## Also defined by 'tidytree'

## Found more than one class "phylo" in cache; using the first, from namespace 'phyloseq'

## Also defined by 'tidytree'

## Found more than one class "phylo" in cache; using the first, from namespace 'phyloseq'

## Also defined by 'tidytree'
```

```
GP_subset_1
```

```
## phyloseq-class experiment-level object
## otu_table() OTU Table: [ 1413 taxa and 26 samples ]
## sample_data() Sample Data: [ 26 samples by 7 sample variables ]
## tax_table() Taxonomy Table: [ 1413 taxa by 7 taxonomic ranks ]
## phy_tree() Phylogenetic Tree: [ 1413 tips and 1412 internal nodes ]
```

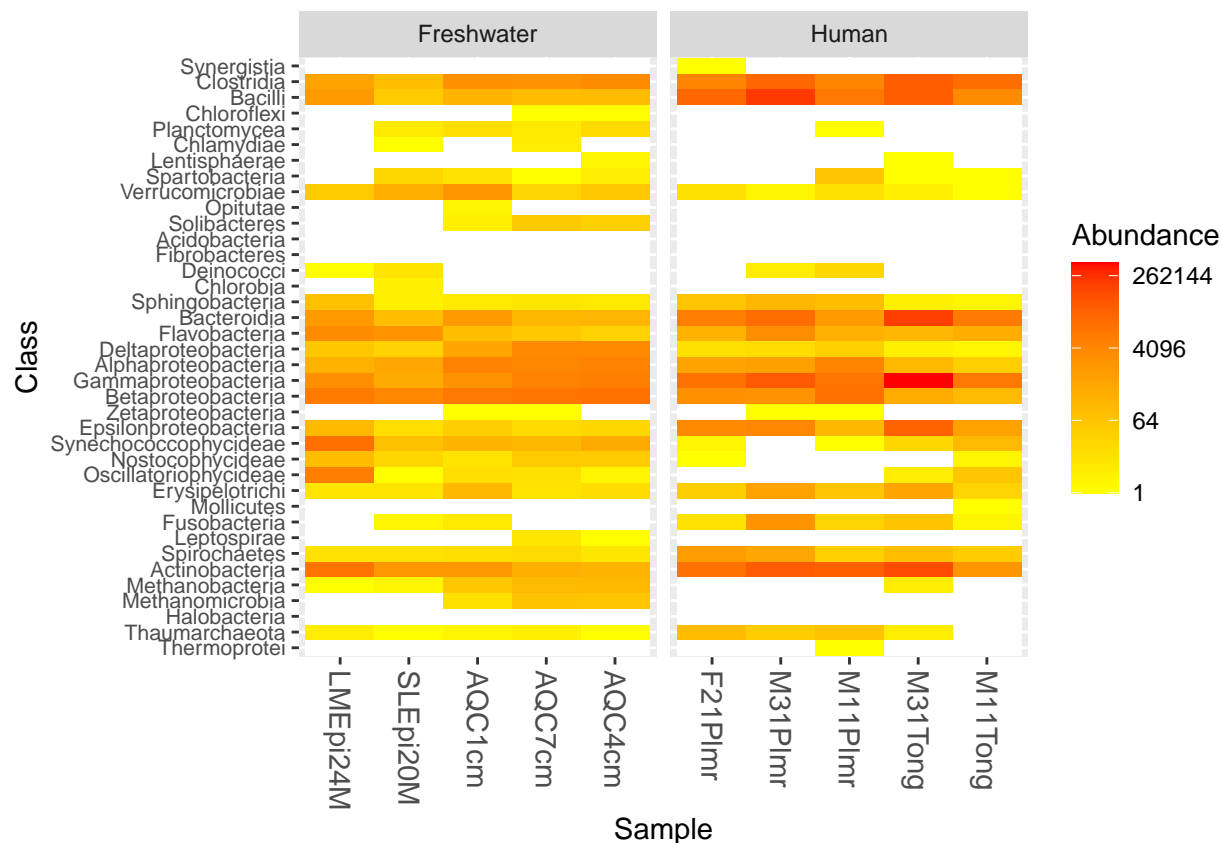
Agglomerate taxa at Class level (required by plot\_heatmap option)

```
my_physeq_glom = tax_glom(my_physeq, taxrank="Class")
```

```
phyloseq::plot_heatmap(my_physeq_glom, low = "yellow", high = "red", na.value = "white", taxa.label = "C",
  facet_grid(~SampleOrigin, scales = "free_x")
```

```
## Warning: Transformation introduced infinite values in discrete y-axis
```





## 4. Hierarchical clustering

```
#Extract OTU table as data frame
my_physeq_otu_df = phyloseq::otu_table(my_physeq) %>% data.frame()
my_physeq_otu_df[1:5, 1:5]
```

```
##           M31Plmr M11Plmr F21Plmr M31Tong M11Tong
## 951             0       1       0       0       0
## 155495          0       0       0       0       0
## 1029            0       0       0       0       0
## 341551          0       0       0       0       0
## 108964          0       2       6       1       0
```

```
# transpose the table (required by vegdist)
my_physeq_otu_df_t = t(my_physeq_otu_df)
my_physeq_otu_df_t[1:5, 1:5]
```

```
##           951 155495 1029 341551 108964
## M31Plmr    0     0    0     0     0
## M11Plmr    1     0    0     0     2
## F21Plmr    0     0    0     0     6
## M31Tong    0     0    0     0     1
## M11Tong    0     0    0     0     0
```

```
#compute Bray-Curtis dissimilarity
bc_dist = vegan::vegdist(my_physeq_otu_df_t, method = "bray")
bc_dist
```

```
##           M31Plmr  M11Plmr  F21Plmr  M31Tong  M11Tong  LMEpi24M  SEpi20M
## M11Plmr  0.8201904
## F21Plmr  0.7088825 0.7171987
## M31Tong  0.8100538 0.9747491 0.9275711
## M11Tong  0.8061918 0.8950000 0.6944554 0.9097781
## LMEpi24M 0.9651206 0.9228413 0.8950359 0.9863850 0.8614094
## SEpi20M  0.9903308 0.9559947 0.9673560 0.9984609 0.9663373 0.8399396
## AQC1cm   0.9885630 0.9196260 0.9576297 0.9951015 0.9537315 0.9318942 0.8449668
## AQC4cm   0.9922834 0.9462591 0.9733061 0.9976856 0.9738446 0.9451736 0.8869454
## AQC7cm   0.9919732 0.9472929 0.9722075 0.9978891 0.9716969 0.9485904 0.8737116
##           AQC1cm  AQC4cm
## M11Plmr
## F21Plmr
## M31Tong
## M11Tong
## LMEpi24M
## SEpi20M
## AQC1cm
## AQC4cm  0.3939152
## AQC7cm  0.3498352 0.1294176
```

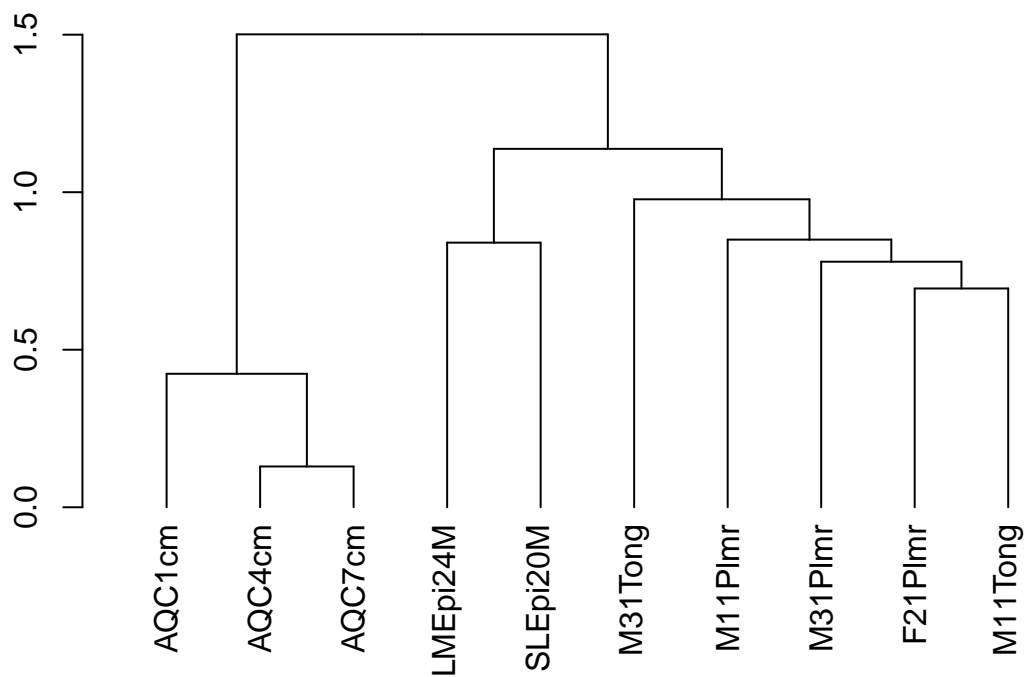
```
# View the distance matrix as matrix
as.matrix(bc_dist)[1:5, 1:5]
```

```
##           M31Plmr  M11Plmr  F21Plmr  M31Tong  M11Tong
## M31Plmr  0.0000000 0.8201904 0.7088825 0.8100538 0.8061918
## M11Plmr  0.8201904 0.0000000 0.7171987 0.9747491 0.8950000
## F21Plmr  0.7088825 0.7171987 0.0000000 0.9275711 0.6944554
## M31Tong  0.8100538 0.9747491 0.9275711 0.0000000 0.9097781
## M11Tong  0.8061918 0.8950000 0.6944554 0.9097781 0.0000000
```

The distance table records the Bray\_Curtis distance between all samples. For example, the distance between M31Plmr and M11Plmr is 0.8201904. Next this matrix will be used for plotting the dendrogram. 'hclust' is the command for hierarchical clustering on distance matrix

```
#Save as dendrogram
ward = as.dendrogram(hclust(bc_dist, method = "ward.D2"))

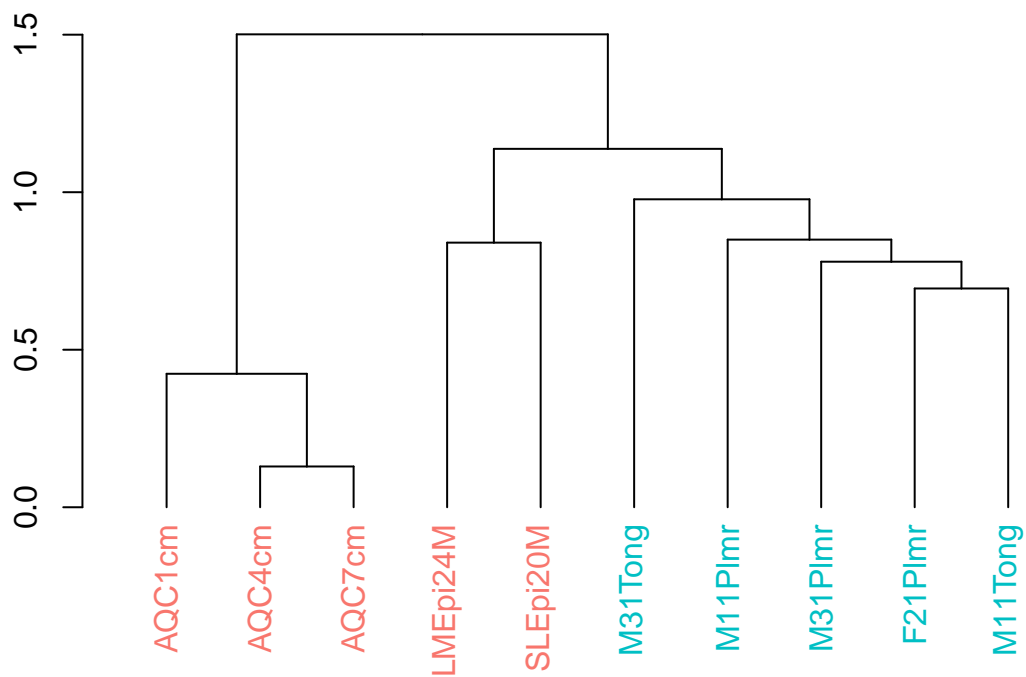
#Plot
plot(ward)
```



### A nicer dendrogram plot with color coding

Below is the code to color the samples based on their category in SampleOrigin.

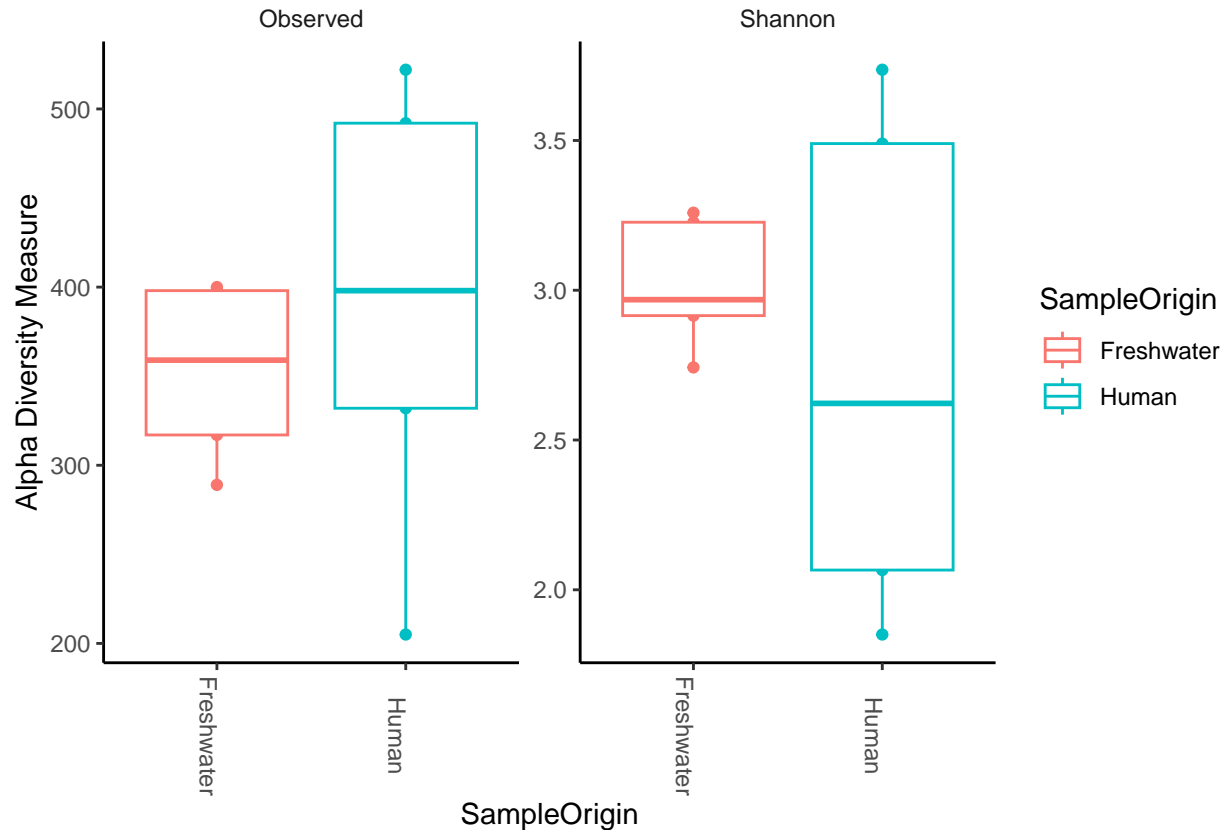
```
#Provide color codes
meta = data.frame(phyloseq::sample_data(my_physeq))
colorCode = c(`Freshwater` = "#F8766D", Human = "#00BFC4" )
labels_colors(ward) = colorCode[meta$SampleOrigin][order.dendrogram(ward)]
#Plot
plot(ward)
```



## 5. Alpha diversity

Now we will plot for alpha diversity using two measures, c("Observed", "Shannon"),

```
plot_richness(my_physeq, x="SampleOrigin", measures=c("Observed", "Shannon"), color = "SampleOrigin") +
  geom_boxplot() +
  theme_classic() +
  theme(strip.background = element_blank(), axis.text.x.bottom = element_text(angle = -90))
```



Identifying the level of significance for the diversity between Feces and Freshwater

```
# Make a dataframe to combine the outputs of Observed, Shannon and SampleOrigin
my_alph_div = data.frame(
  "Observed" = phyloseq::estimate_richness(my_physeq, measures = "Observed"),
  "Shannon" = phyloseq::estimate_richness(my_physeq, measures = "Shannon"),
  "SampleOrigin" = phyloseq::sample_data(my_physeq)$SampleOrigin)
head(my_alph_div)
```

```
##      Observed  Shannon SampleOrigin
## M31Plmr      492  2.622235         Human
## M11Plmr      522  3.736234         Human
## F21Plmr      398  3.489691         Human
## M31Tong      332  1.850676         Human
## M11Tong      205  2.065849         Human
## LMEpi24M      317  2.742431    Freshwater
```

Check the level of significance

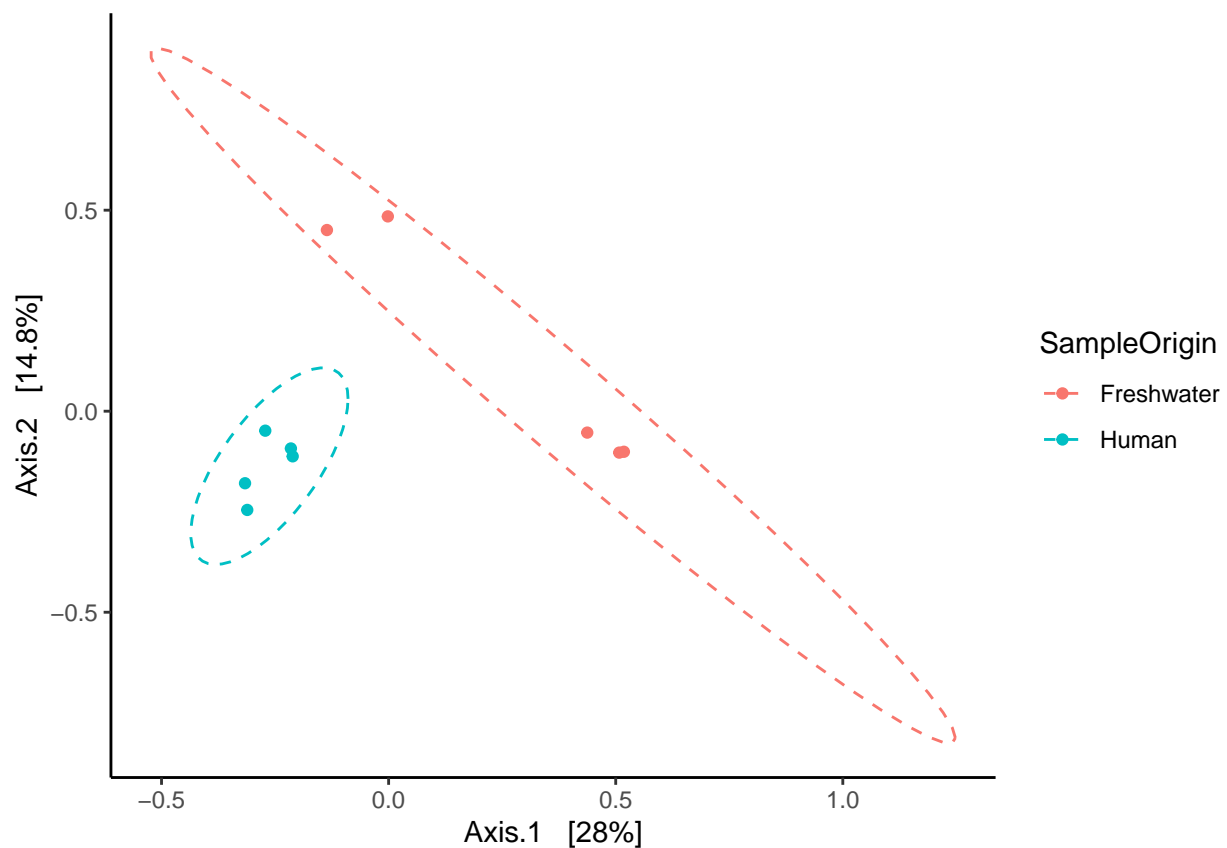
```
#Wilcoxon test for Shannon diversity for categories in SampleOrigin
my_alph_div_wt = wilcox.test(Shannon ~ SampleOrigin, data = my_alph_div, exact = FALSE, conf.int = TRUE)
print(my_alph_div_wt$p.value)
```

```
## [1] 0.6761033
```

p-value = 0.6761. Here the p-value is more than 0.05. If the p-value in the output is less than 0.05, it means the difference is significant. In this case we can **NOT** accept the alternative hypothesis which means that the diversity is **NOT** significantly different between Freshwater and Human samples.

## 6. Beta diversity

```
ordination = ordinate(my_physeq, method="PCoA", distance="jaccard")
plot_ordination(my_physeq, ordination, color="SampleOrigin") +
  theme_classic() +
  theme(strip.background = element_blank()) +
  stat_ellipse(linetype = 2)
```



It shows that the between sample diversity is very high between Freshwater and very low between Feces samples

## 7. Differential abundance (DA) analysis using deseq2

```

set.seed(2345)
# run_deseq2 command run the program deseq2 to identify DA taxa
# Running this command takes a few seconds
my_physeq_deseq2 = run_deseq2(my_physeq,
                             group = "SampleOrigin",
                             transform = "log10p", # log transformation
                             norm = "rarefy", # common method for normalization
                             p_adjust = "BH", # adjusted p-value methods
                             )

## You set 'rngseed' to FALSE. Make sure you've set & recorded
## the random seed of your session for reproducibility.
## See '?set.seed'

## ...

## 4610TUs were removed because they are no longer
## present in any sample after random subsampling

## ...

## converting counts to integer mode

## -- note: fitType='parametric', but the dispersion trend was not well captured by the
## function: y = a/x + b, and a local regression fit was automatically substituted.
## specify fitType='local' or 'mean' to avoid this message next time.

```

```
my_physeq_deseq2
```

```

## microbiomeMarker-class inherited from phyloseq-class
## normalization method:          [ RLE ]
## microbiome marker identity method: [ DESeq2: Wald ]
## marker_table() Marker Table:    [ 16 microbiome markers with 5 variables ]
## otu_table() OTU Table:          [ 959 taxa and 10 samples ]
## sample_data() Sample Data:      [ 10 samples by 8 sample variables ]
## tax_table() Taxonomy Table:     [ 959 taxa by 1 taxonomic ranks ]

```

Plot the differentially abundant taxa identified by deseq2 method

```
marker_table(my_physeq_deseq2) %>% head()
```

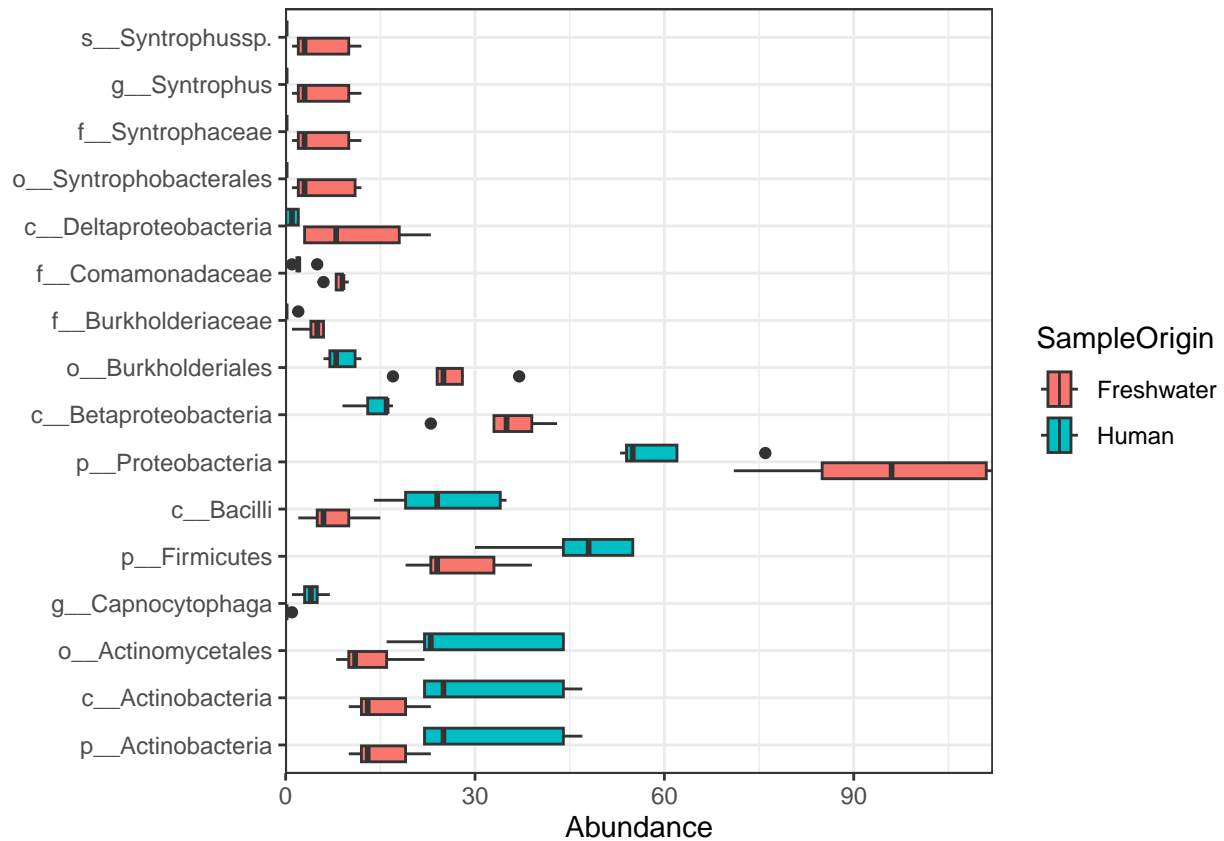
```

##
## marker1                                     k__Bacteria|p__l
## marker2                                     k__Bacteria|p__Proteobacteria|c__Betap
## marker3                                     k__Bacteria|p__Proteobacteria|c__Betaproteobacteria|o__B
## marker4                                     k__Bacteria|p__Proteobacteria|c__Deltaproteobacteria|o__Syntro
## marker5                                     k__Bacteria|p__Proteobacteria|c__Deltaproteobacteria|o__Syntrophobacteriales|f__
## marker6 k__Bacteria|p__Proteobacteria|c__Deltaproteobacteria|o__Syntrophobacteriales|f__Syntrophaceae

```

```
##      enrich_group  ef_logFC      pvalue      padj
## marker1   Freshwater -0.6629649 4.009616e-05 0.001116965
## marker2   Freshwater -1.2848808 3.611545e-05 0.001116965
## marker3   Freshwater -1.5739909 3.313312e-05 0.001116965
## marker4   Freshwater -4.9787402 2.470348e-05 0.001116965
## marker5   Freshwater -4.9281142 3.137184e-05 0.001116965
## marker6   Freshwater -4.9281142 3.137184e-05 0.001116965
```

```
plot_DA = microbiomeMarker::plot_abundance(my_physeq_deseq2, group = "SampleOrigin")
plot_DA
```

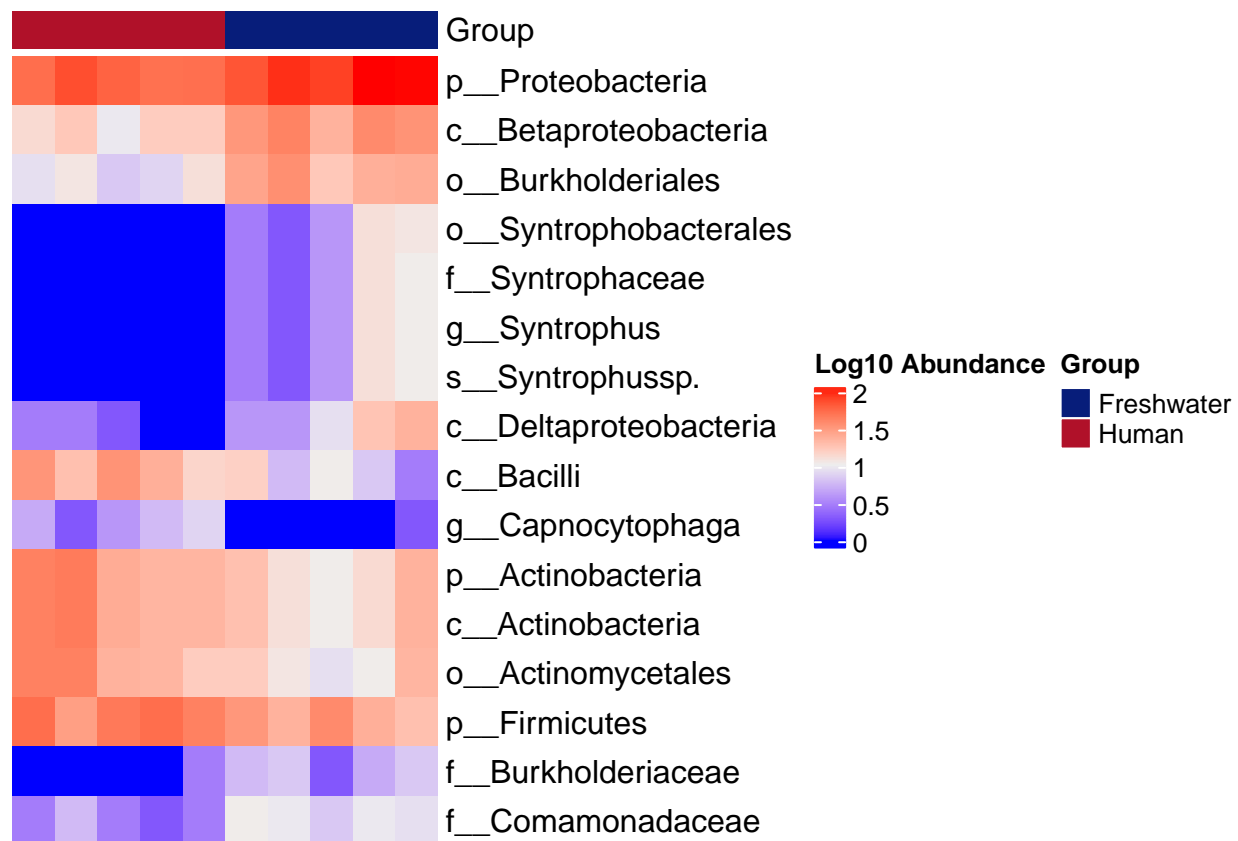


```
plot_DA_hmap = microbiomeMarker::plot_heatmap(my_physeq_deseq2, group = "SampleOrigin")
```

```
## Warning in transform_log10(otu): OTU table contains zeroes. Using log10(1 + x)
## instead.
```

```
plot_DA_hmap
```





```
#otu_table(my_physeq_deseq2) %>% head()
```