Lab_10_Metagenomics_MBI3100A_2022_Tutorial

2022-11-12

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

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("tidyverse")) install.packages("tidyverse")
```

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

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

```
## Loading required package: DT
```

```
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(DT)
library(dendextend)
```

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 (2
## [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)"
```

Readable data table in Rmarkdown

To print a matrix/dataframe/vector in nice tabular format, use the command as.data.frme with pipe option (%>%) as shown in the example below. This provides a better view for matrix or vector type data without changing the underlying data structure.

```
phyloseq_datasets$results %>% DT::datatable()
```

Package LibPath Item Title

```
# 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 ]
```

Interpretation

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()
# sample_data()
# tax_table()
# phy_tree()
```

Question: How many samples are there in the 'GlobalPatterns' data set? (1 mark)

```
Answer: 26
```

Access the OTU table from a dataset

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

```
## 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
```

Read the table with as.data.frame( )

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

```
## 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) %>% DT::datatable()
```

X.SampleID Primer Final_Barcode Barcode_truncated_plus_T Barcode_full_length SampleType Description

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"
```

Read with as.data.frame

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

```
##                   .
## 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

```
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
```

```
GP_tutorial_1 = subset_taxa(GlobalPatterns, Species!="NA")
GP_tutorial_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 ]
```

Import data

We will combine a phyloseq object using otu_table, sample_data and taxonomy file. Will will read these three file 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
```

In order to read the OTU table as phyloseq object

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

```
## 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
```

Similarly to read sample data and taxonomy table as phyloseq objects

```
my_Sample_data = sample_data(GP_sp_tutorial_sample_data)
my_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
```

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

```
## 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 ]
```

Explore the catagories in a sample variable eg SampleType

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

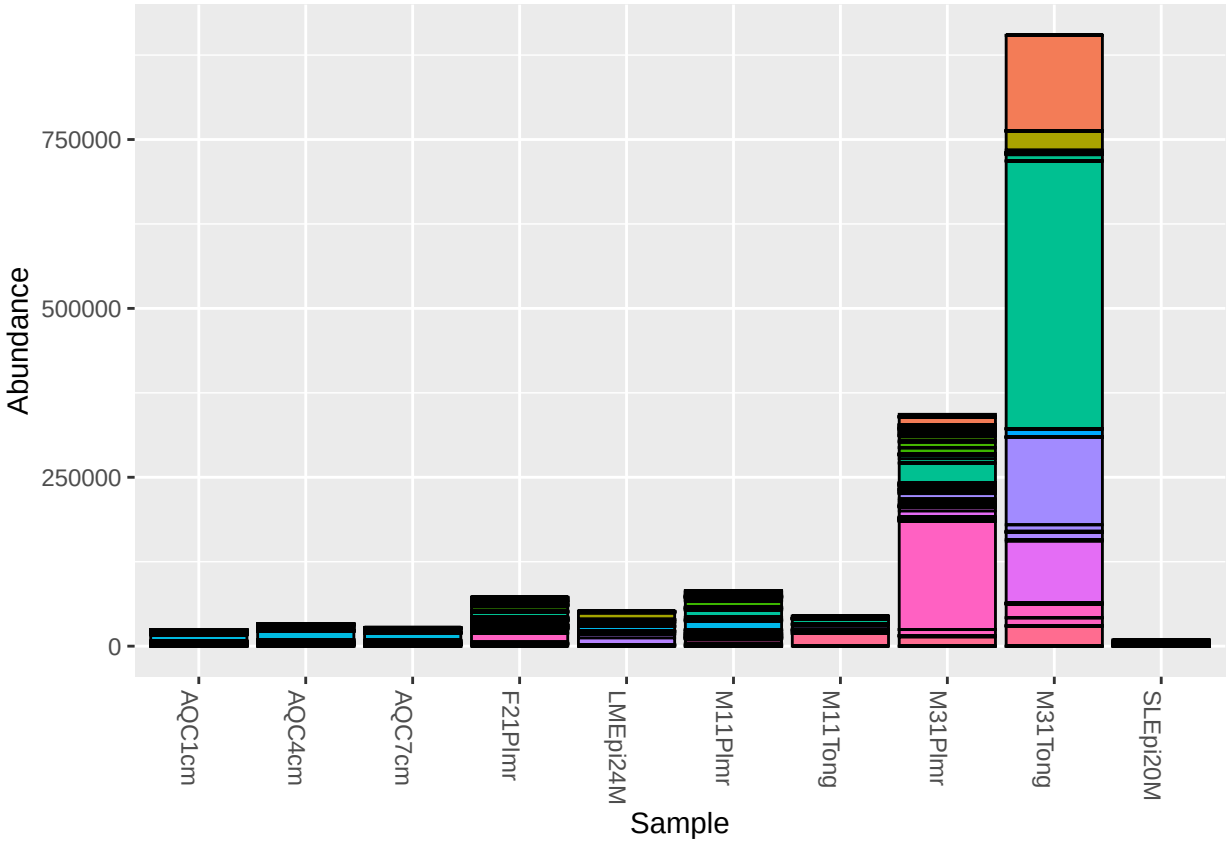
```
## [1] "Freshwater"         "Freshwater (creek)" "Skin"
## [4] "Tongue"
```

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

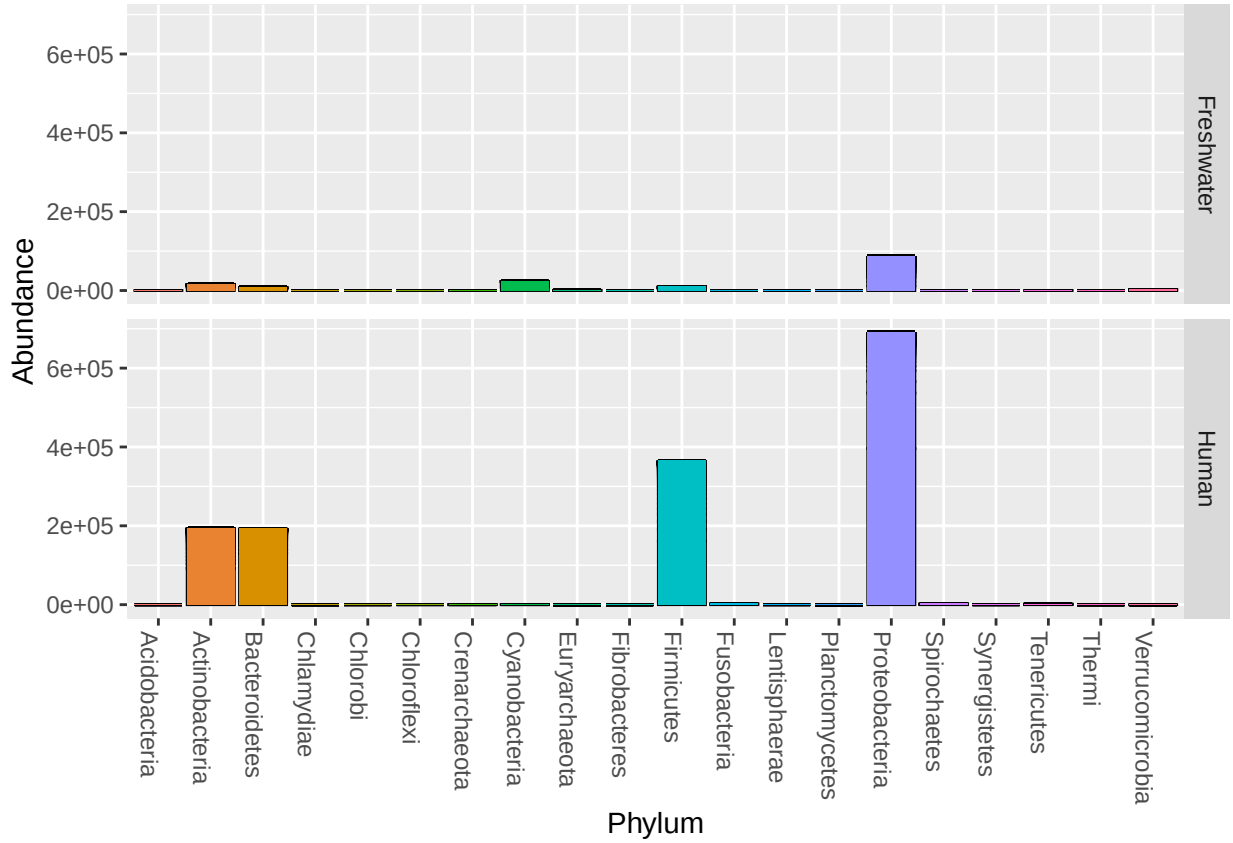
```
## [1] "Freshwater" "Human"
```

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

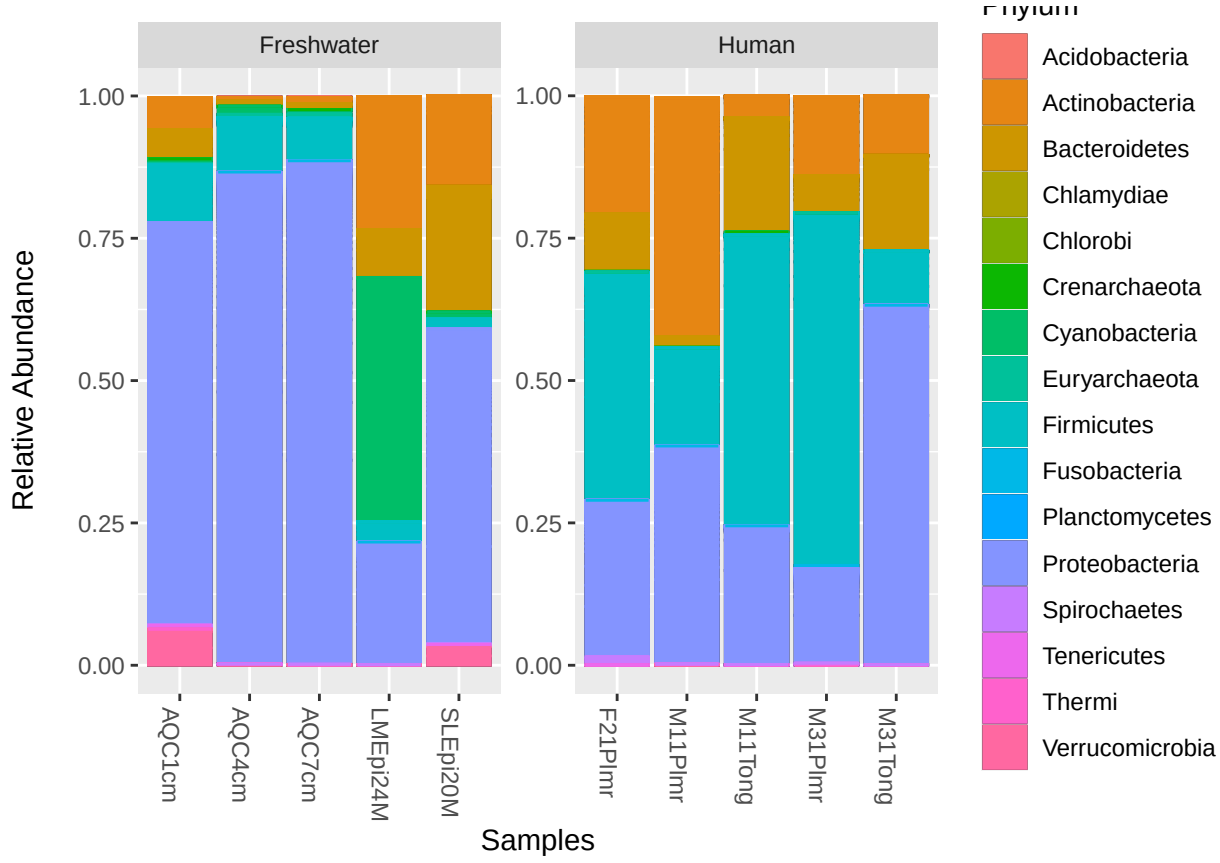
```
# To convert to relative abundance
my_physeq_r  = transform_sample_counts(my_physeq, function(x) x / sum(x) )
# Keep the taxa which have a mean values at least 1e-5
my_physeq_rf = filter_taxa(my_physeq_r, function(x) mean(x) > 1e-5, TRUE)
my_physeq_rf
```

```
## 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

```
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")
```



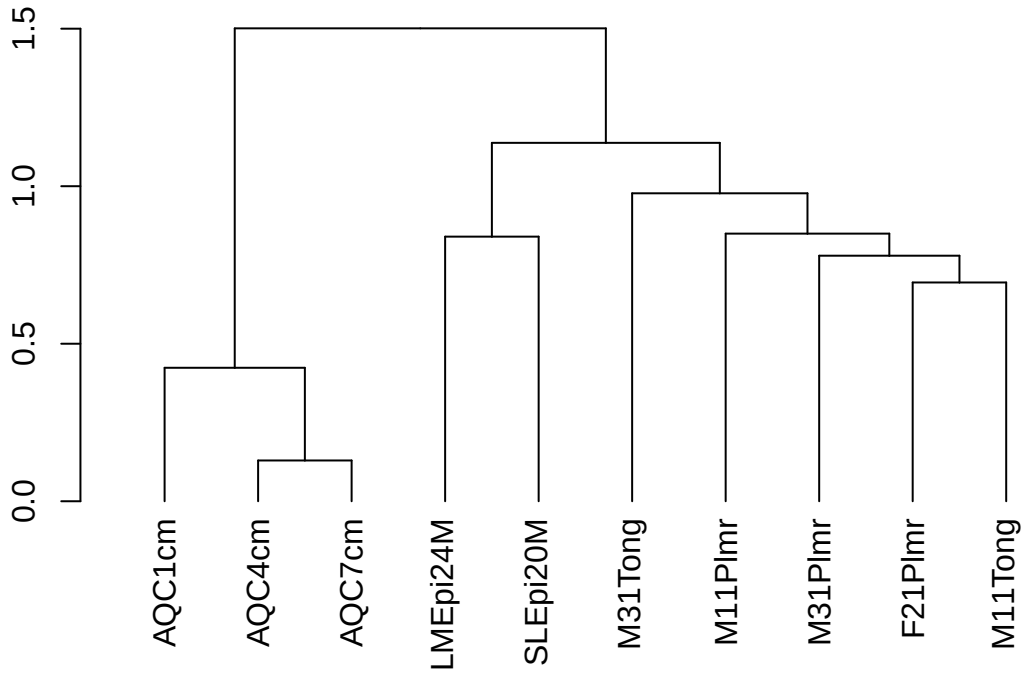
Hierarchical clustering

```
#Extract OTU table
ps_rel_otu = data.frame(phyloseq::otu_table(my_physeq))
ps_rel_otu = t(ps_rel_otu) # transpose the table (required by vegdist )
#compute Bray-curtis distance
bc_dist = vegan::vegdist(ps_rel_otu, method = "bray")
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
```

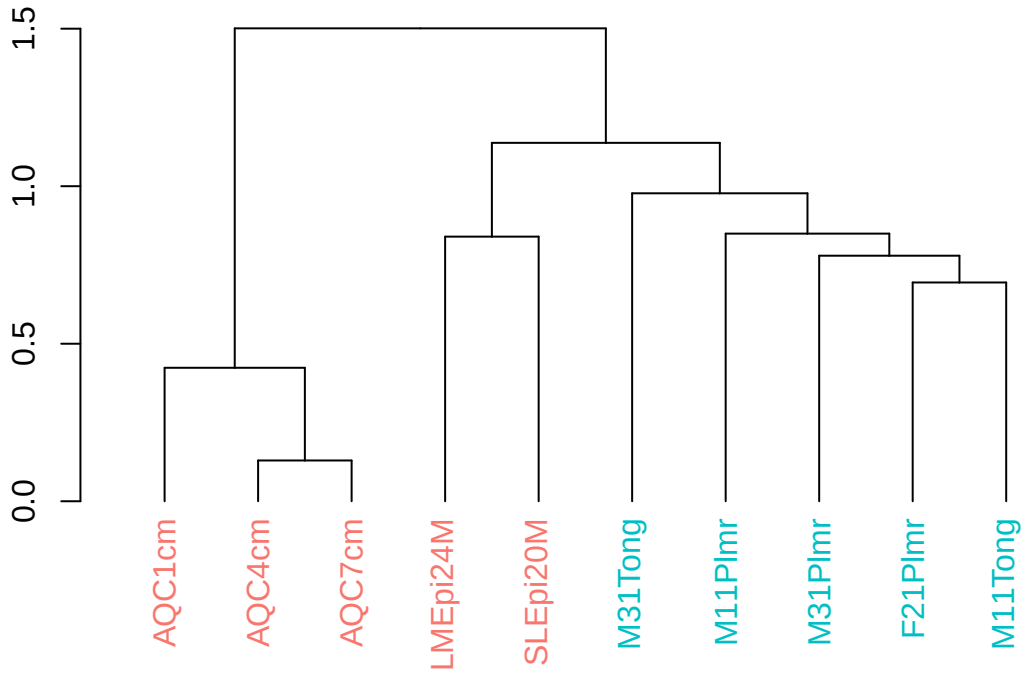
```
#Save as dendrogram
ward = as.dendrogram(hclust(bc_dist, method = "ward.D2"))

#Plot
plot(ward)
```

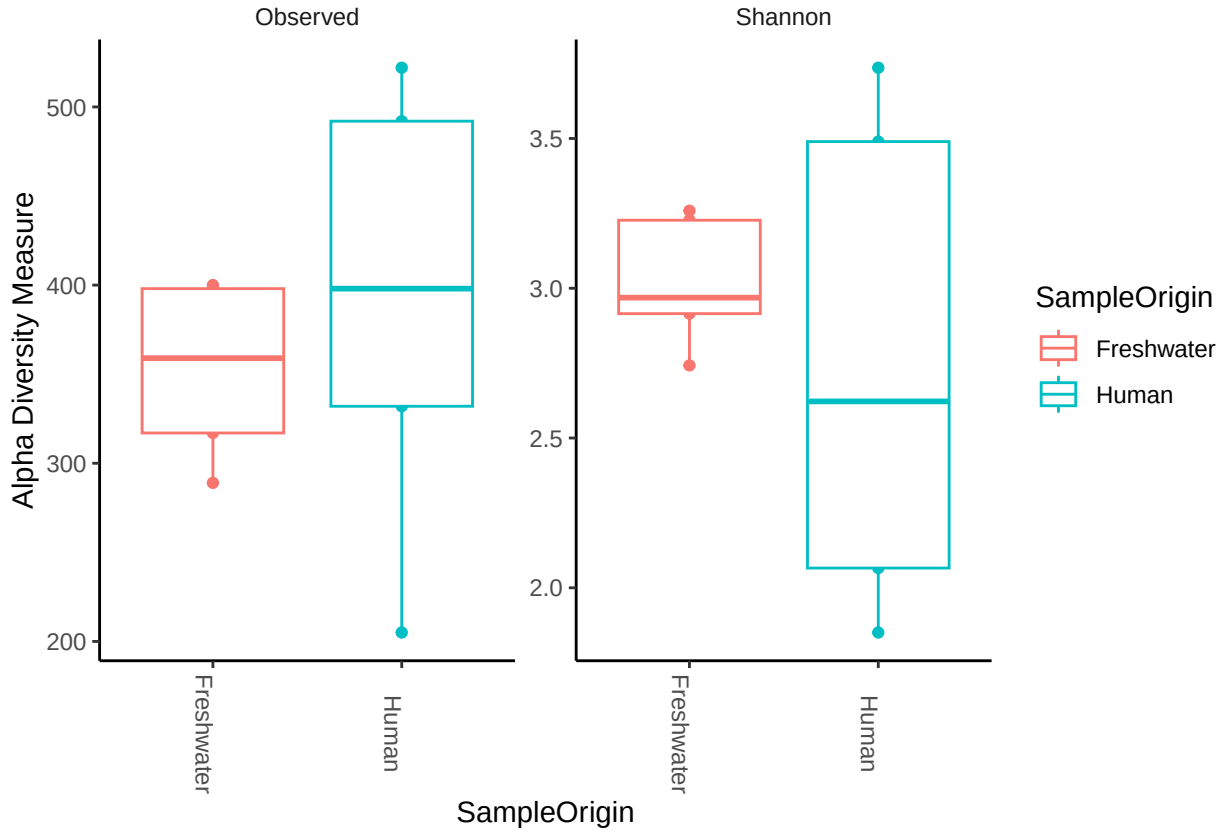


A nicer plot with color coding

```
#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)
```




```
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 significane for the diversity between Feces and Freshwater

```
#
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

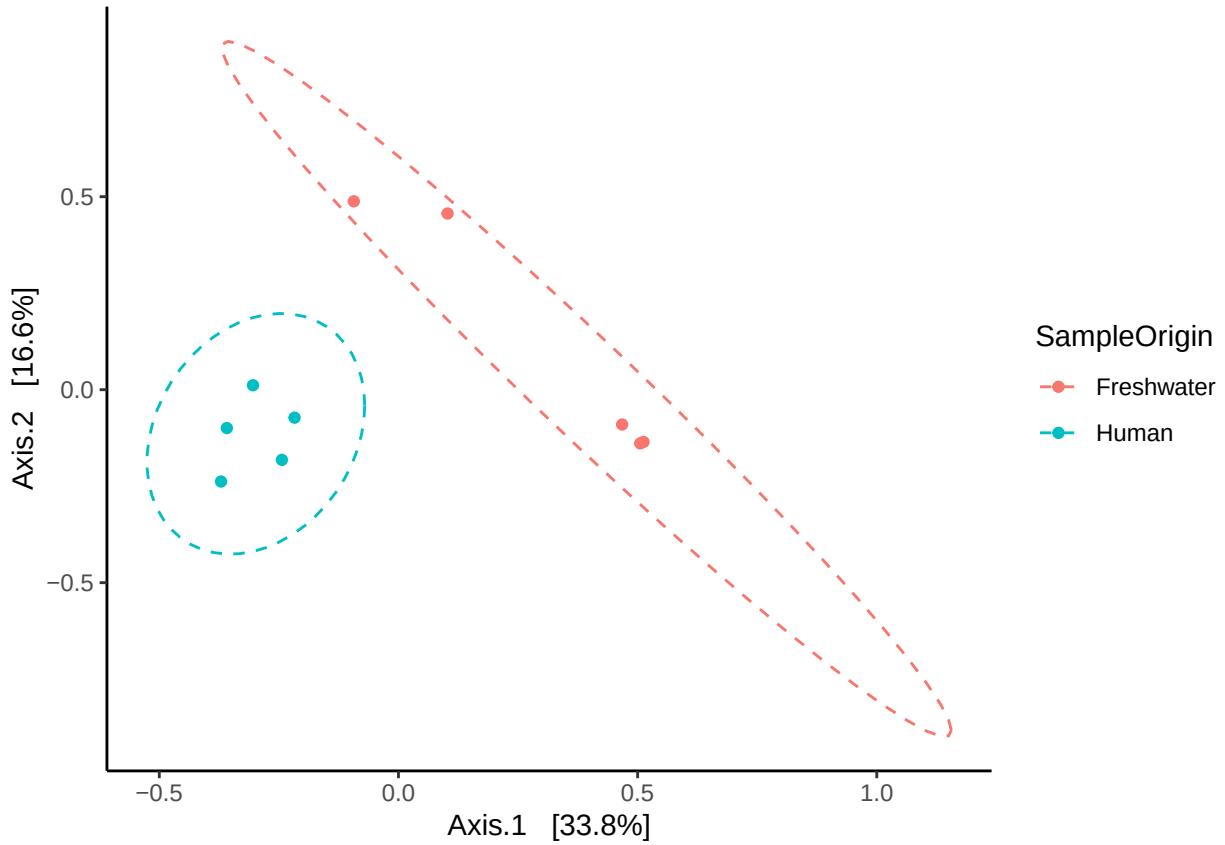
```
#Wilcoxon test for Shannon diversity for categories in SampleOrigin
wilcox.test(Shannon ~ SampleOrigin, data = my_alph_div, exact = FALSE, conf.int = TRUE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Shannon by SampleOrigin
## W = 15, p-value = 0.6761
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  -0.7676234  1.1929154
## sample estimates:
## difference in location
##              0.3463032
```

p-value is ~0.02 which is less than 0.05, Hence we can accept the alternative hypothesis which means that the Shannon diversity is significantly different between Freshwater and Feces samples.

Beta diversity

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



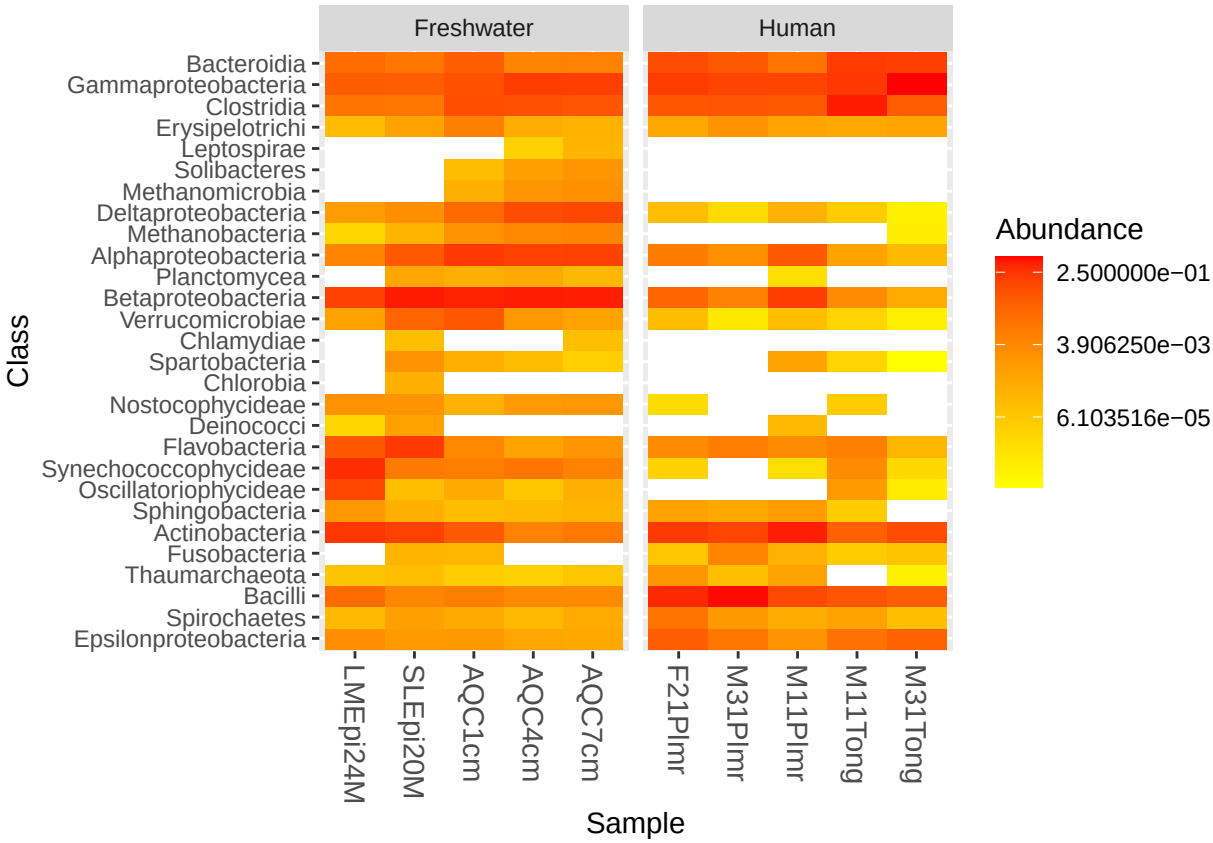
It shows that the between sample diversity is very high between Freshwater and very low between Feces samples

Agglomerate taxa at Class level (required by plot_heatmap option)

```
my_physeq_rf_glom = tax_glom(my_physeq_rf, taxrank="Class")
```

```
plot_heatmap(my_physeq_rf_glom, low = "yellow", high = "red", na.value = "white",taxa.label = "Class") +
  facet_grid(~SampleOrigin, scales = "free_x")
```

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


Freshwater
Human
Bacteroidia
Gammaproteobacteria
Clostridia
Erysipelotrichi
Leptospirae
Solibacteres
Methanomicrobia
Deltaproteobacteria
Methanobacteria
Alphaproteobacteria
Planctomycea
Betaproteobacteria
Verrucomicrobiae
Chlamydiae
Spartobacteria
Chlorobia
Nostocophycideae
Deinococci
Flavobacteria
Synechococcophycideae
Oscillatoriophycideae
Sphingobacteria
Actinobacteria
Fusobacteria
Thaumarchaeota
Bacilli
Spirochaetes
Epsilonproteobacteria
Class
LMEpi24M
SLEpi20M
AQC1cm
AQC4cm
AQC7cm
F21PImr
M31PImr
M11PImr
M11Tong
M31Tong
Sample
Abundance
2.500000e−01
3.906250e−03
6.103516e−05