Lab_10_Metagenomics_MBI3100A_2022_Assignment_Answers

2022-11-12

R Markdown

```
# clear the R environment
rm(list = ls())
```

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("microbiomeMarkera")
```

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

```
## -- 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("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)
```

Data import

Question 1: Import the three files named as 'GP_sp_assignment_otu_table_df.csv', 'GP_sp_assignment_sample_data_df.csv', and 'GP_sp_assignment_tax_table_df.csv' and make a phyloseq object named 'asgmt_physeq'? (2 points)

Please provide the correct file/folder path

```
GP_sp_assignment_otu_table = read.table("./assignment_files/GP_sp_assignment_otu_table_df.csv",
           sep = "\t",
           header =  T,
           row.names = "otus")
my_OTU_table = otu_table(GP_sp_assignment_otu_table, taxa_are_rows = TRUE)
GP_sp_assignment_sample_data = read.table("./assignment_files/GP_sp_assignment_sample_data_df.csv",
           sep = "\t", header =  T,
           row.names = "sampleid")
my_Sample_data = sample_data(GP_sp_assignment_sample_data)
GP_sp_assignment_tax_table = read.table("./assignment_files/GP_sp_assignment_tax_table_df.csv",
           sep = "\t",
           header =  T,
           row.names = "otus")
my_tax_table = tax_table(as.matrix(GP_sp_assignment_tax_table))

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

Print asgmt_physeq

```
print(asgmt_physeq)
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 1413 taxa and 9 samples ]
## sample_data() Sample Data:       [ 9 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 1413 taxa by 7 taxonomic ranks ]
```

Question 2: How many taxa, samples, and sample variables are there in asgmt_physeq? (1 point)

```

```

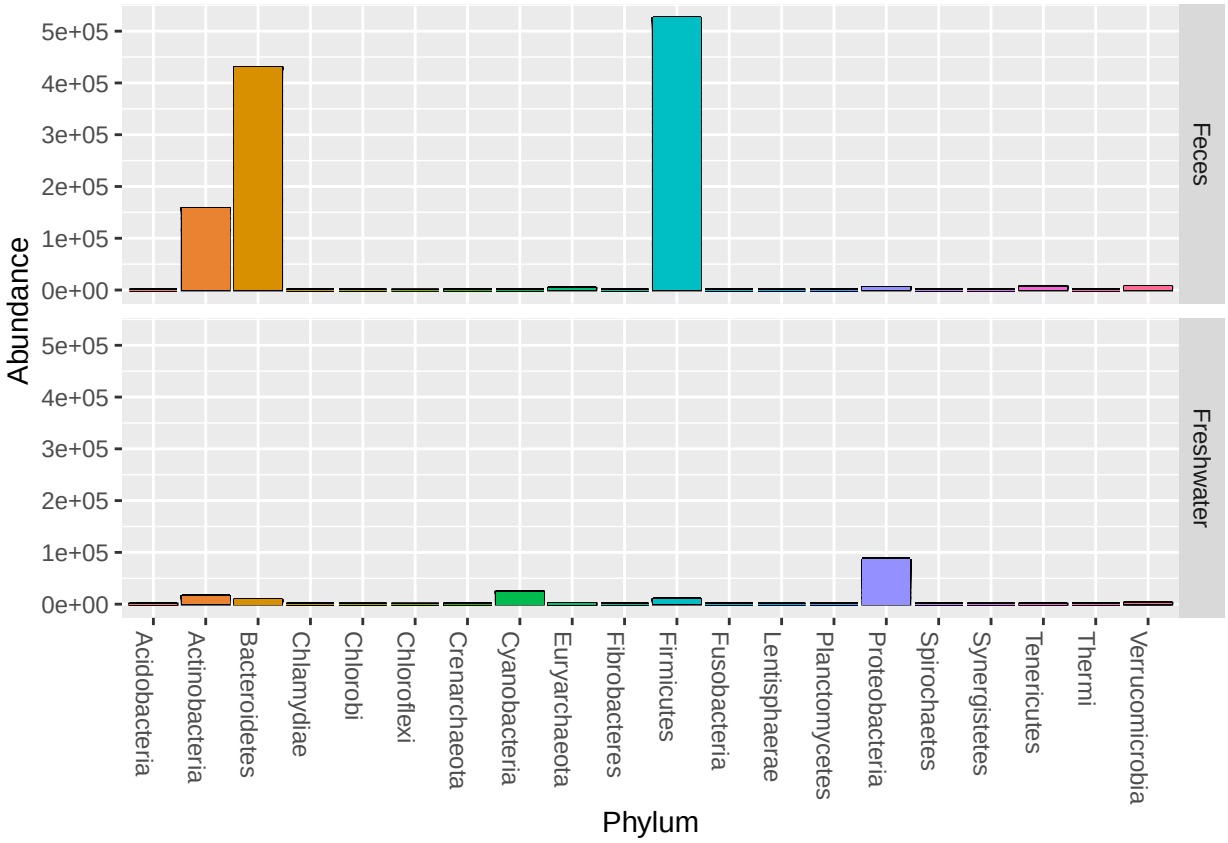
Question 3: List the categories present in the sample variable "SampleOrigin". (1 point)

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

```
## [1] "Feces"      "Freshwater"
```

Question 4: Generate a bar plot for sample vs abundance and facet it for the catagories in SampleOrigin (1 point)

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



Question 5: Based on the plot generated in question 4, name all the phylum which have a big difference in abundance between "Feces" and "Freshwater" samples? (1 point)

Transform the absolute abundance into relative abundance and filter the taxa which have mean relative abundance less than 0.0001

```
# To convert to relative abundance
asgmt_physeq_r  = transform_sample_counts(asgmt_physeq, function(x) x / sum(x) )
# Keep the taxa which have a mean values at least 0.0001
asgmt_physeq_rf = filter_taxa(asgmt_physeq_r, function(x) mean(x) > 0.0001, TRUE)
asgmt_physeq_rf
```

```
## phyloseq-class experiment-level object
```

```
## otu_table()   OTU Table:         [ 238 taxa and 9 samples ]
## sample_data() Sample Data:       [ 9 samples by 8 sample variables ]
## tax_table()   Taxonomy Table:    [ 238 taxa by 7 taxonomic ranks ]
```

Question 6: How many taxa are left after the above filtering? (1 point)

For question 7 to 12, use dataset 'asgmt_physeq'.

Question 7: Generate a Hierarchical clustering plot using the distance "ward.D2". (2 points)

It will be a four step process

```
asgmt_physeq_otu_df = phyloseq::otu_table(asgmt_physeq) %>%  data.frame()
asgmt_physeq_otu_df[1:5, 1:5]
```

Step1: Extract OTU table as data frame

```
##        M31Fcsw M11Fcsw LMEpi24M SLEpi20M AQC1cm
## 951          0       0        0        0      0
## 155495       0       0        0        0      0
## 1029         0       0        0        0      0
## 341551       0       0        0        0      0
## 108964       0       0        1        0      1
```

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

Step2: Transpose the table (required by vegdist package)

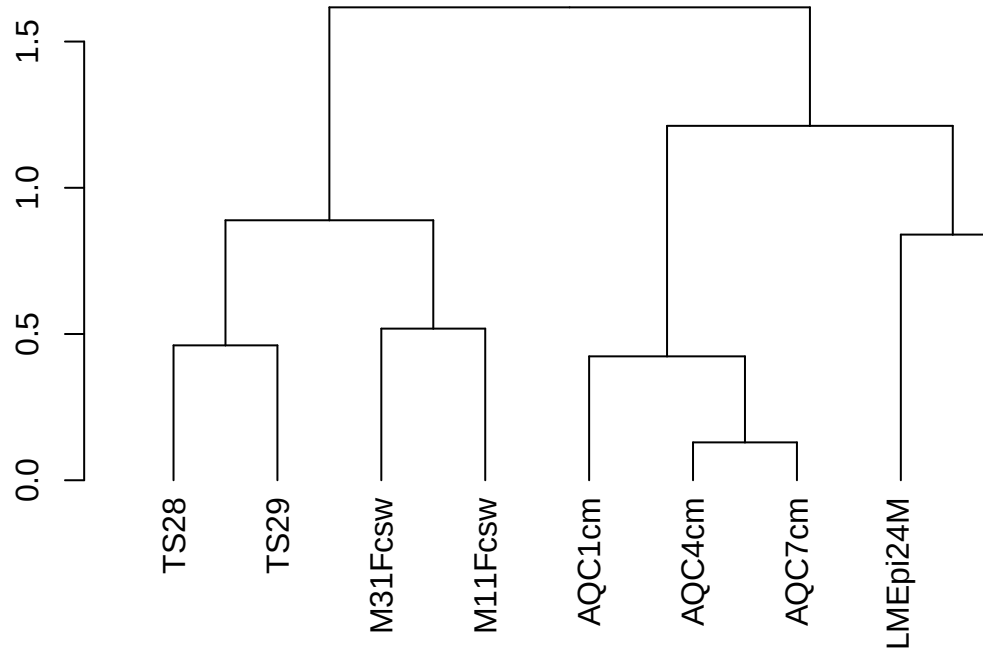
```
##          951 155495 1029 341551 108964
## M31Fcsw    0      0    0      0      0
## M11Fcsw    0      0    0      0      0
## LMEpi24M   0      0    0      0      1
## SLEpi20M   0      0    0      0      0
## AQC1cm     0      0    0      0      1
```

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

Step3: Compute Bray-Curtis dissimilarity

```
##            M31Fcsw   M11Fcsw  LMEpi24M  SLEpi20M    AQC1cm    AQC4cm    AQC7cm
## M11Fcsw  0.5184034
## LMEpi24M 0.9904053 0.9938943
## SLEpi20M 0.9970146 0.9972846 0.8399396
## AQC1cm   0.9805934 0.9828535 0.9318942 0.8449668
## AQC4cm   0.9957373 0.9961299 0.9451736 0.8869454 0.3939152
## AQC7cm   0.9960703 0.9965983 0.9485904 0.8737116 0.3498352 0.1294176
## TS28     0.6222115 0.7872989 0.9934247 0.9967061 0.9720616 0.9954971 0.9960830
## TS29     0.6348875 0.8073339 0.9918941 0.9974866 0.9784274 0.9965166 0.9969603
##               TS28
## M11Fcsw
## LMEpi24M
## SLEpi20M
## AQC1cm
## AQC4cm
## AQC7cm
## TS28
## TS29     0.4612095
```

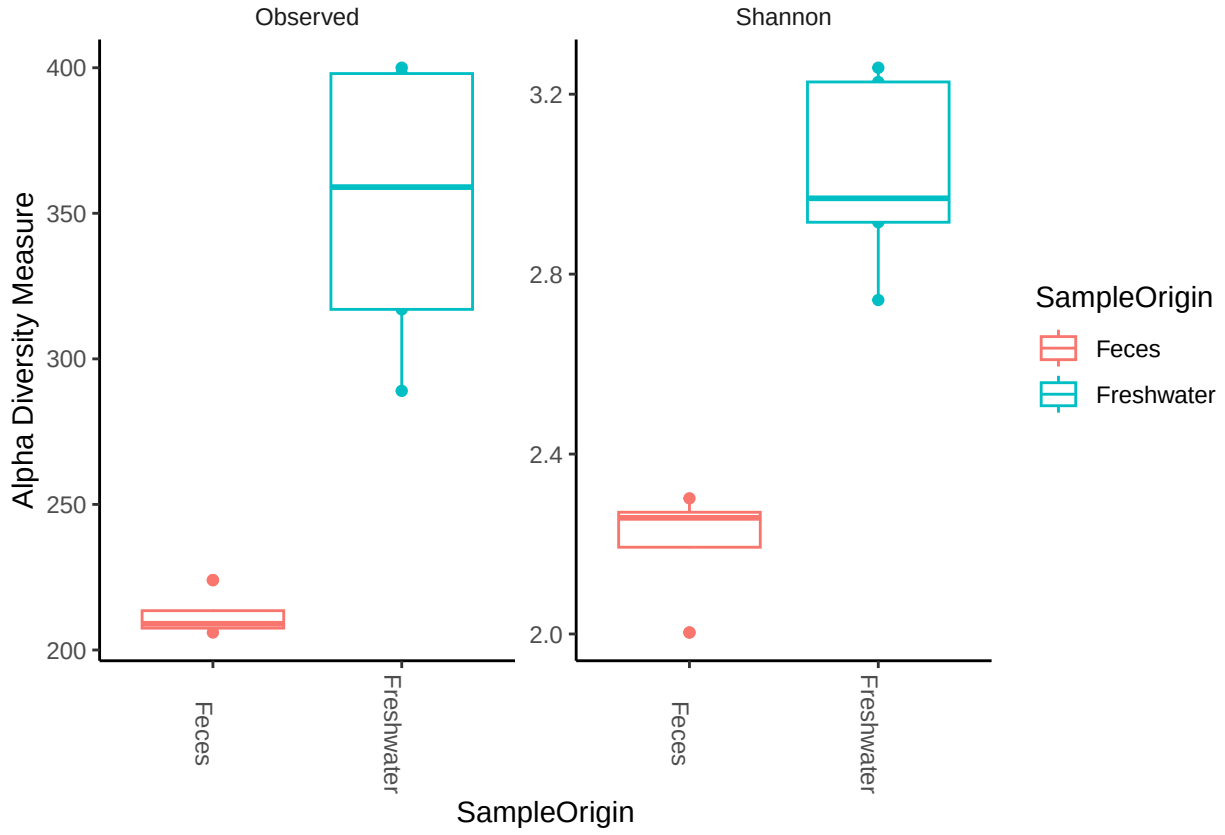
```
#Save as dendrogram
ward = as.dendrogram(hclust(bc_dist, method = "ward.D2"))

#Plot
plot(ward)
```



Step4: Save as dendrogram

Question 8: Plot for alpha diversity using two measures, "Observed" and "Shannon". (1 points)

```
plot_richness(asgmt_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))
```



Question 9: Apply wilcox.test to see if the Observed diversity is significantly different for SampleOrigin. (2 points)

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

Step 1: Make a dataframe to combine the outputs of Observed and SampleOrigin.

```
##          Observed  Shannon SampleOrigin
## M31Fcsw       210 2.256019        Feces
## M11Fcsw       206 2.003266        Feces
## LMEpi24M      317 2.742431   Freshwater
## SLEpi20M      289 3.227190   Freshwater
## AQC1cm        400 3.258820   Freshwater
## AQC4cm        398 2.915304   Freshwater
```

```
#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)
```

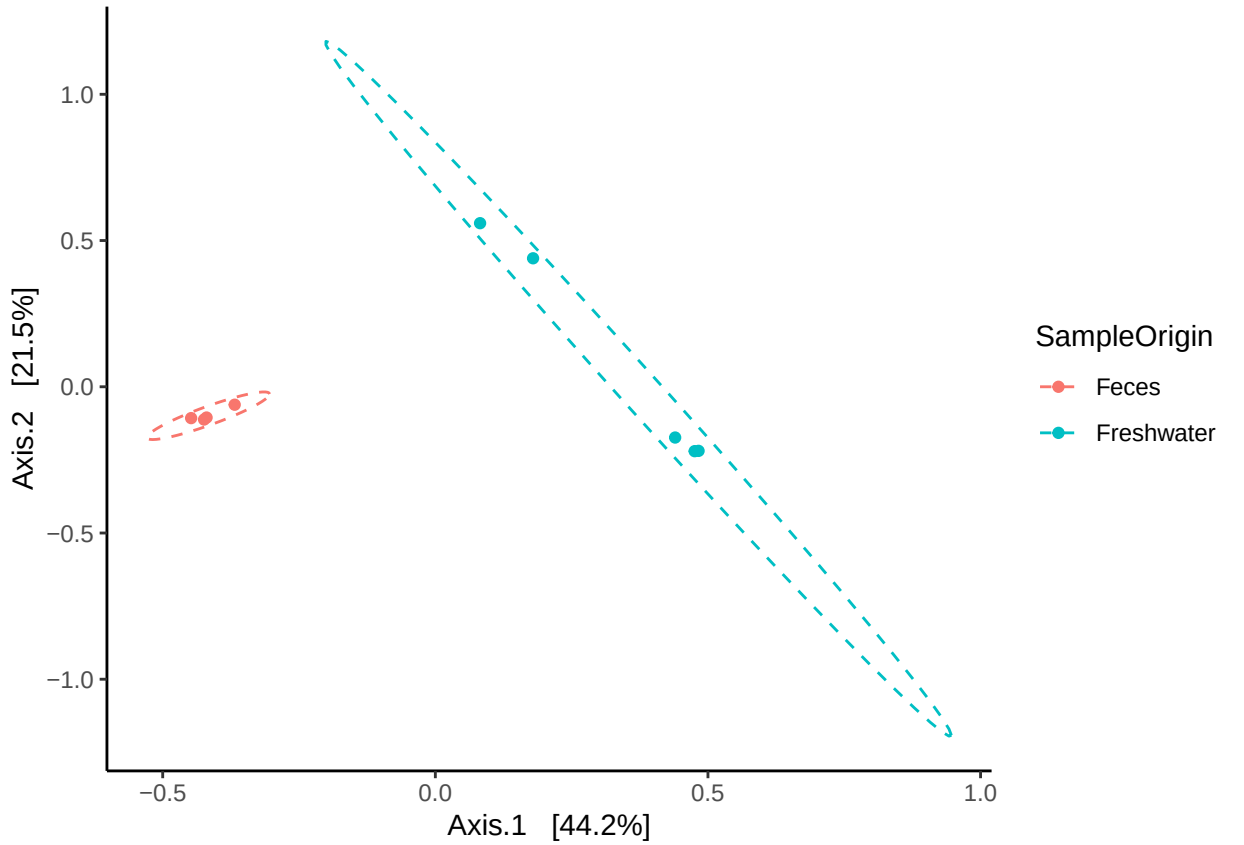
Step 2: Check the significance level for wilcox.test

```
## [1] 0.01996445
```

Note down the p-value? is the difference significant i.e is p-value less than 0.05?

Question 10: Make a PCoA plot using the "bray" method as distance for beta diversity. (1 point)

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



Question 11: Apply DESeq2 method to identify the differentially abundant taxa based on SampleOrigin column. (1 point)

```
set.seed(2345)
# run_deseq2 command run the program deseq2 to identify DA taxa
# Running this command takes a few seconds
asgmt_physeq_deseq2 = run_deseq2(asgmt_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‘
```

```
## ...
```

```
## 313OTUs were removed because they are no longer
## present in any sample after random subsampling
```

```
## ...
```

```
## converting counts to integer mode
```

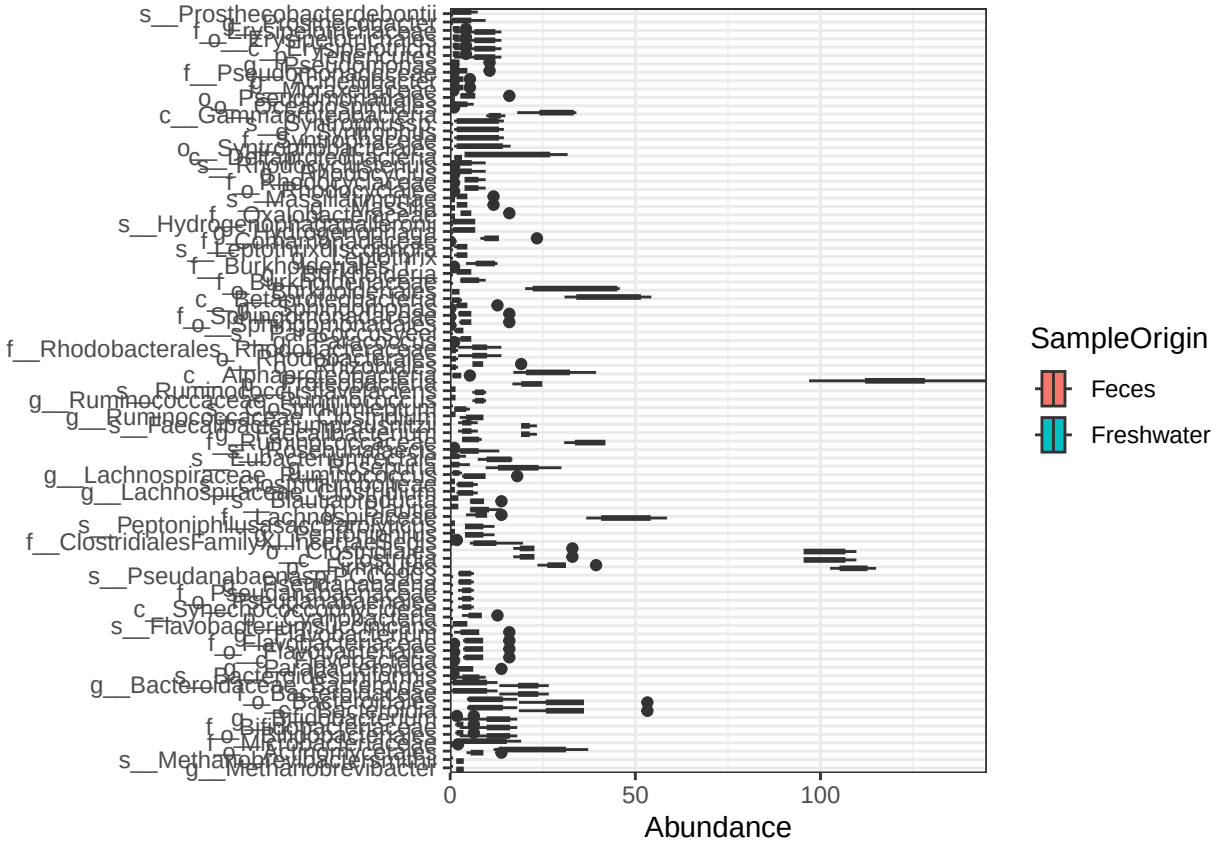
```
## -- note: fitType=’parametric’, but the dispersion trend was not well captured by the
##    function: y = a/x + b, and a local regression fit was automatically substituted.
##    specify fitType=’local’ or ’mean’ to avoid this message next time.
```

```
asgmt_physeq_deseq2
```

```
## microbiomeMarker-class inherited from phyloseq-class
## normalization method:              [ RLE ]
## microbiome marker identity method: [ DESeq2: Wald ]
## marker_table() Marker Table:       [ 91 microbiome markers with 5 variables ]
## otu_table()    OTU Table:          [ 837 taxa and  9 samples ]
## sample_data()  Sample Data:        [ 9 samples by  8 sample variables ]
## tax_table()    Taxonomy Table:     [ 837 taxa by 1 taxonomic ranks ]
```

Question 12: Plot the differentially abundant taxa identified by deseq2 method . (1 point)

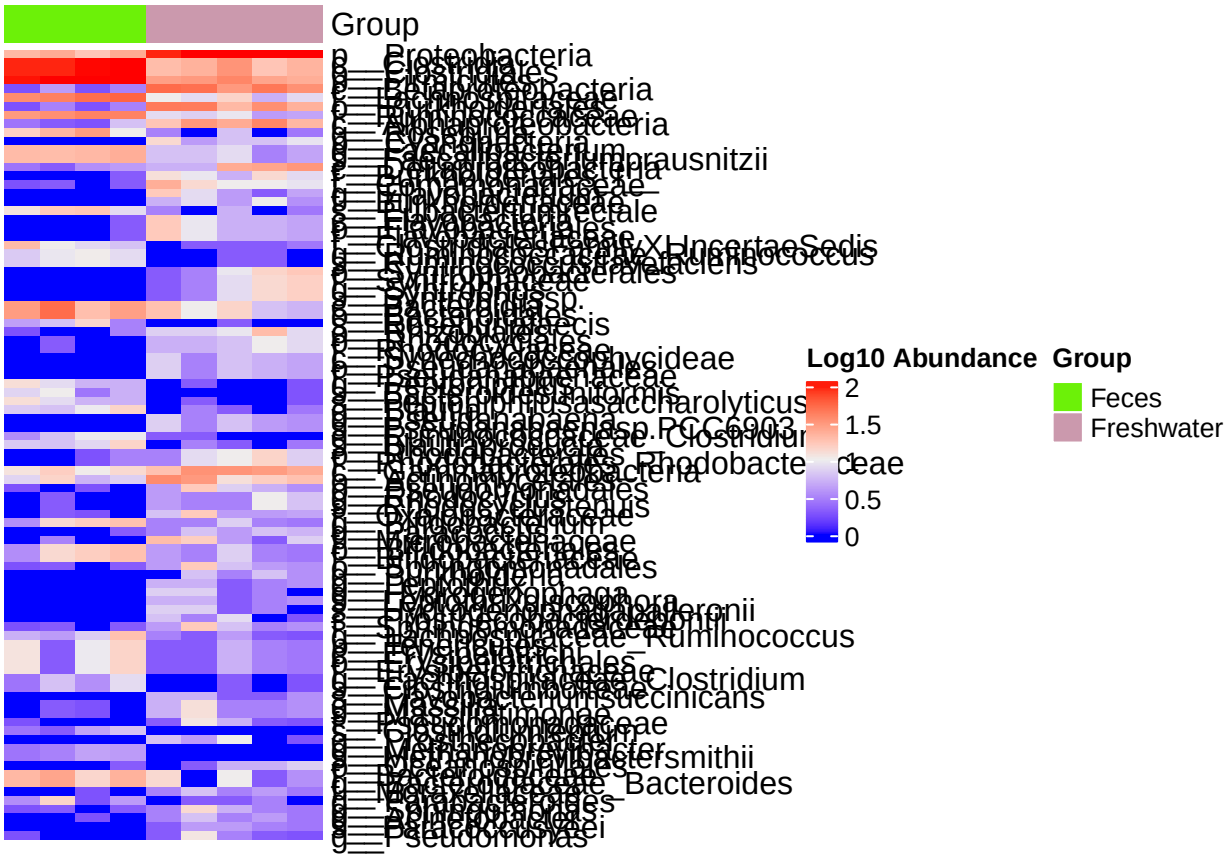
```
plot_DA =  microbiomeMarker::plot_abundance(asgmt_physeq_deseq2, group = "SampleOrigin")
plot_DA
```




```
plot_DA_hmap = microbiomeMarker::plot_heatmap(asgmt_physeq_deseq2, group = "SampleOrigin")
```

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

```
plot_DA_hmap
```


Group
Log10 Abundance
2
1.5
1
0.5
0
Group
Feces
Freshwater