Phyloseq tutorial

Daniel Vaulot

7 juin 2017

Contents

1 Introduction

This document explains the use of the phyloseq R library to analyze metabarcoding data.

1.1 Phyloseq R library

- Phyloseq web site : https://joey711.github.io/phyloseq/index.html
- See in particular tutorials for
  - importing data: https://joey711.github.io/phyloseq/import-data.html
  - heat maps: https://joey711.github.io/phyloseq/plot_heatmap-examples.html

1.2 Data

This tutorial uses a reduced metabarcoding dataset obtained by C. Ribeiro and A. Lopes dos Santos. This dataset originates from the CARBOM cruise in 2013 off Brazil and corresponds to the 18S V4 region amplified on flow cytometry sorted samples (see pptx file for details) and sequenced on an Illumina run 2*250 bp analyzed with mothur.

1.3 References for data

- Gérikas Ribeiro, C., Lopes dos Santos, A., Marie, D., Helena Pellizari, V., Pereira Brandini, F., and Vaulot, D. (2016). Pico and nanoplankton abundance and carbon stocks along the Brazilian Bight. PeerJ 4, e2587. doi:10.7717/peerj.2587.

- Gérikas Ribeiro, C., Marie, D., Lopes dos Santos, A., Pereira Brandini, F., and Vaulot, D. (2016). Estimating microbial populations by flow cytometry: Comparison between instruments. Limnol. Oceanogr. Methods 14, 750–758. doi:10.1002/lom3.10135.
- Gérikas Ribeiro C, Lopes dos Santos A, Marie D, Brandini P, Vaulot D. (2018). Relationships between photosynthetic eukaryotes and nitrogen-fixing cyanobacteria off Brazil. ISME J in press.

2 Prerequisites to be installed

- R : https://pbil.univ-lyon1.fr/CRAN/
- R studio : https://www.rstudio.com/products/rstudio/download/#download
- Download and install the following libraries by running under R studio the following lines

```
install.packages("dplyr")     # To manipulate dataframes
install.packages("readxl")    # To read Excel files into R

install.packages("ggplot2")   # for high quality graphics

source("https://bioconductor.org/biocLite.R")
biocLite("phyloseq")
```

3 Script description

3.1 Load necessary libraries

```
library("phyloseq")
library("ggplot2")      # graphics
library("readxl")       # necessary to import the data from Excel file
library("dplyr")        # filter and reformat data frames
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

3.2 Read the data and create phyloseq objects

Three tables are needed :
* OTU
* Taxonomy
* Samples
They are read from a single Excel file where each sheet contains one of the tables

```r
otu_mat<- read_excel("CARBOM data.xlsx", sheet = "OTU matrix")
tax_mat<- read_excel("CARBOM data.xlsx", sheet = "Taxonomy table")
samples_df <- read_excel("CARBOM data.xlsx", sheet = "Samples")
```

Phyloseq objects need to have row.names

- define the row names from the otu column

```r
row.names(otu_mat) <- otu_mat$otu
```

```
## Warning: Setting row names on a tibble is deprecated.
```

- remove the column otu since it is now used as a row name

```r
otu_mat <- otu_mat %>% select (-otu)
```

- Idem for the two other matrixes

```r
row.names(tax_mat) <- tax_mat$otu
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```r
tax_mat <- tax_mat %>% select (-otu)

row.names(samples_df) <- samples_df$sample
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```r
samples_df <- samples_df %>% select (-sample)
```

Transform into matrixes otu and tax tables (sample table can be left as data frame)

```r
otu_mat <- as.matrix(otu_mat)
tax_mat <- as.matrix(tax_mat)
```

Transform to phyloseq objects

```r
OTU = otu_table(otu_mat, taxa_are_rows = TRUE)
TAX = tax_table(tax_mat)
samples = sample_data(samples_df)

carbom <- phyloseq(OTU, TAX, samples)
carbom
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 287 taxa and 55 samples ]
## sample_data() Sample Data:       [ 55 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 287 taxa by 7 taxonomic ranks ]
```

Visualize data

```r
sample_names(carbom)
```

```
##  [1] "X10n"    "X10p"    "X11n"    "X11p"    "X120n"   "X120p"   "X121n"
##  [8] "X121p"   "X122n"   "X122p"   "X125n"   "X125p"   "X126n"   "X126p"
## [15] "X127n"   "X13n"    "X13p"    "X140n"   "X140p"   "X141n"   "X141p"
## [22] "X142n"   "X142p"   "X155n"   "X155p"   "X156n"   "X156p"   "X157n"
## [29] "X157p"   "X15n"    "X15p"    "X165n"   "X165p"   "X166n"   "X166p"
## [36] "X167n"   "X167p"   "X1n"     "X1p"     "X2n"     "X2p"     "X3n"
## [43] "X3p"     "X5n"     "X5p"     "X7n"     "X7p"     "X9n"     "X9p"
## [50] "tri01n"  "tri01p"  "tri02n"  "tri02p"  "tri03n"  "tri03p"
```

```
rank_names(carbom)
```

```
## [1] "Domain"     "Supergroup" "Division"   "Class"      "Order"
## [6] "Family"     "Genus"
```

```
sample_variables(carbom)
```

```
##  [1] "fraction"           "Select_18S_nifH"   "total_18S"
##  [4] "total_16S"          "total_nifH"        "sample_number"
##  [7] "transect"           "station"           "depth"
## [10] "latitude"           "longitude"         "picoeuks"
## [13] "nanoeuks"           "bottom_depth"      "level"
## [16] "transect_distance" "date"              "time"
## [19] "phosphates"         "silicates"         "ammonia"
## [22] "nitrates"           "nitrites"          "temperature"
## [25] "fluorescence"       "salinity"          "sample_label"
```

Keep only samples to be analyzed

```
carbom <- subset_samples(carbom, Select_18S_nifH =="Yes")
carbom
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 287 taxa and 54 samples ]
## sample_data() Sample Data:       [ 54 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 287 taxa by 7 taxonomic ranks ]
```

Keep only photosynthetic taxa

```
carbom <- subset_taxa(carbom, Division %in% c("Chlorophyta", "Dinophyta", "Cryptophyta",
                                              "Haptophyta", "Ochrophyta", "Cercozoa"))
carbom <- subset_taxa(carbom, !(Class %in% c("Syndiniales", "Sarcomonadea")))
carbom
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 205 taxa and 54 samples ]
## sample_data() Sample Data:       [ 54 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 205 taxa by 7 taxonomic ranks ]
```

Normalize number of reads in each sample using median sequencing depth.

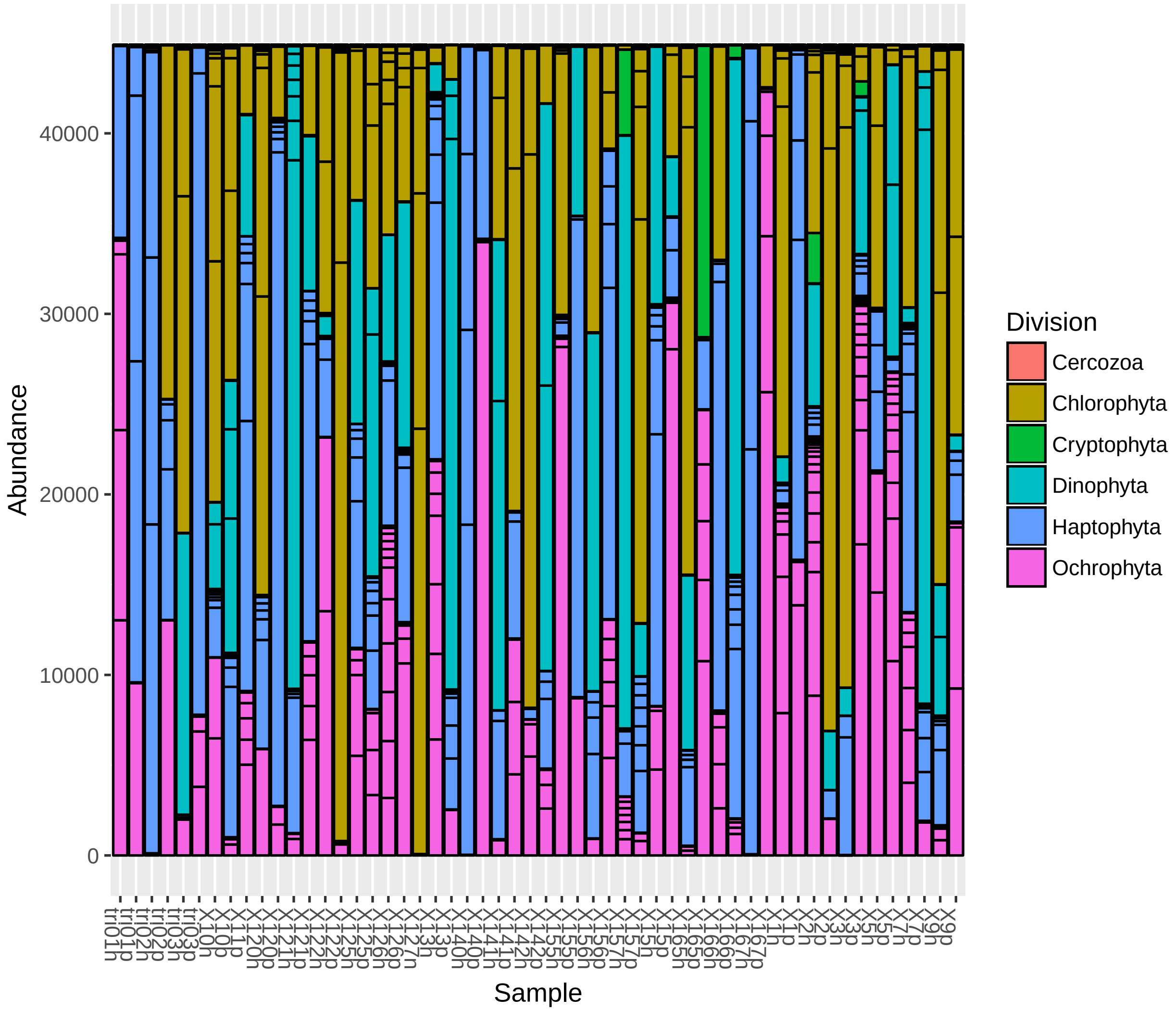
```
total = median(sample_sums(carbom))
standf = function(x, t=total) round(t * (x / sum(x)))
carbom = transform_sample_counts(carbom, standf)
```

The number of reads used for normalization is **44903**.

3.3 Bar graphs

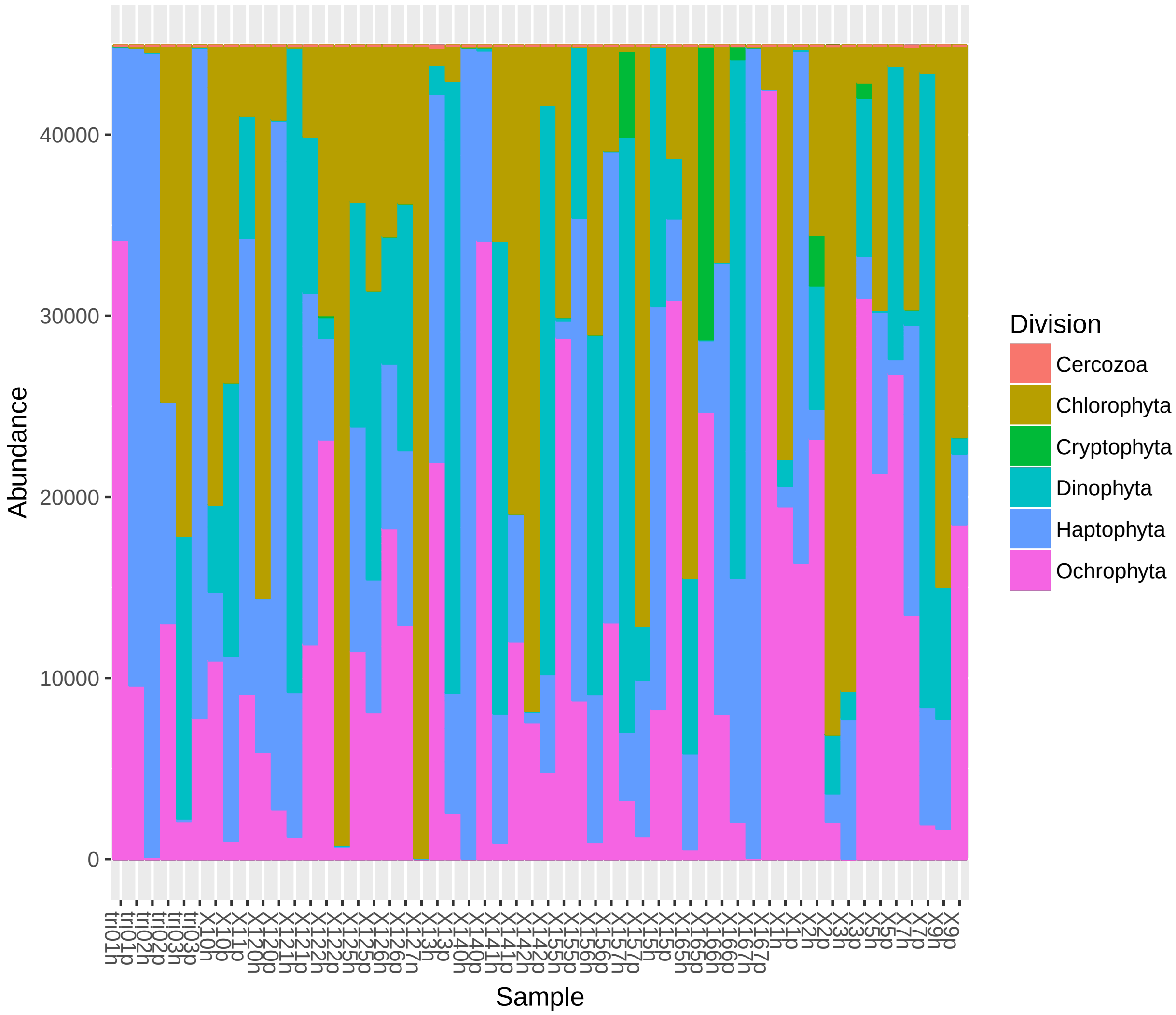
Basic bar graph based on Division

```
plot_bar(carbom, fill = "Division")
```



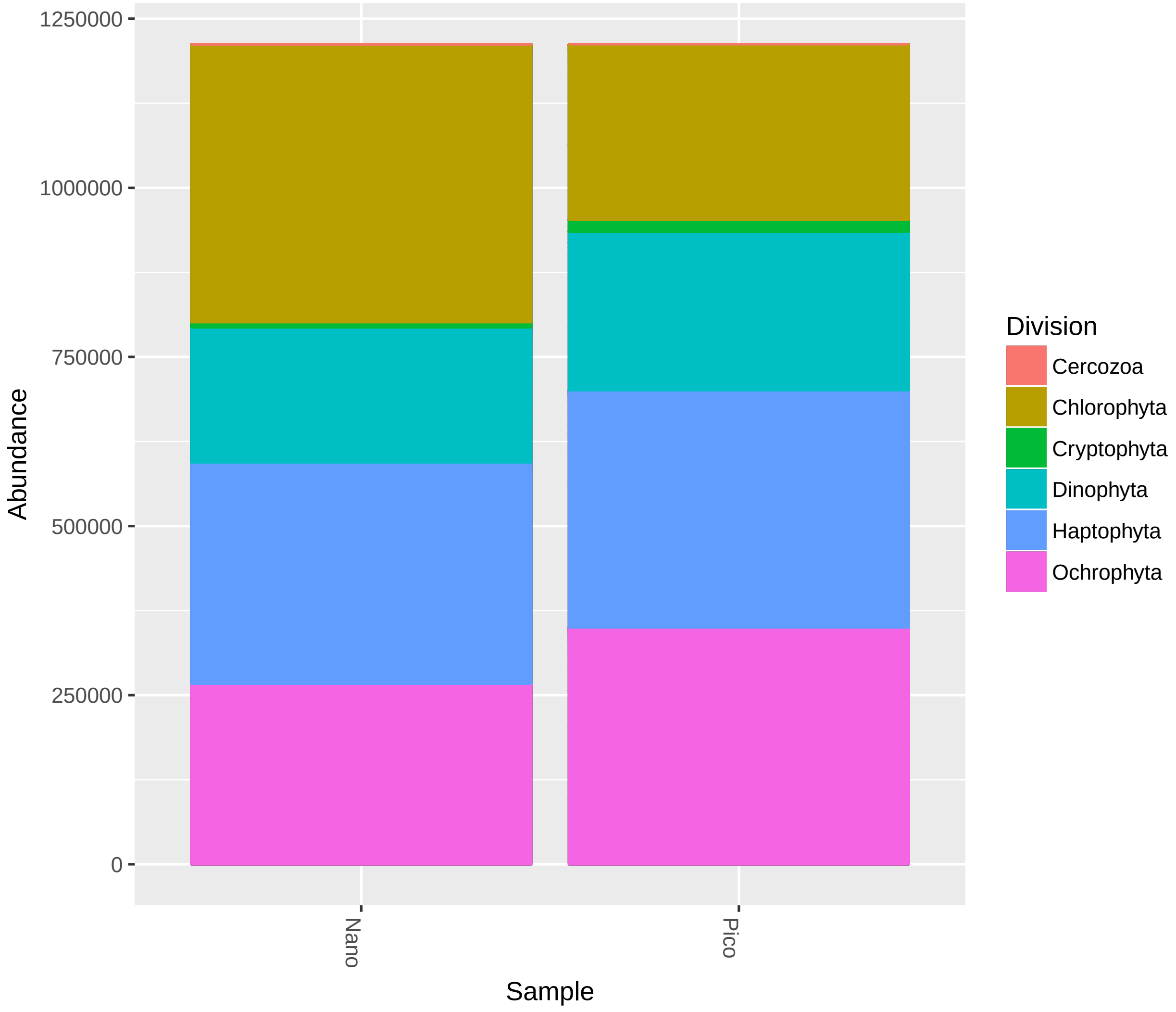
Make the bargraph nicer by removing OTUs boundaries. This is done by adding ggplot2 modifier.

```r
plot_bar(carbom, fill = "Division") +
geom_bar(aes(color=Division, fill=Division), stat="identity", position="stack")
```



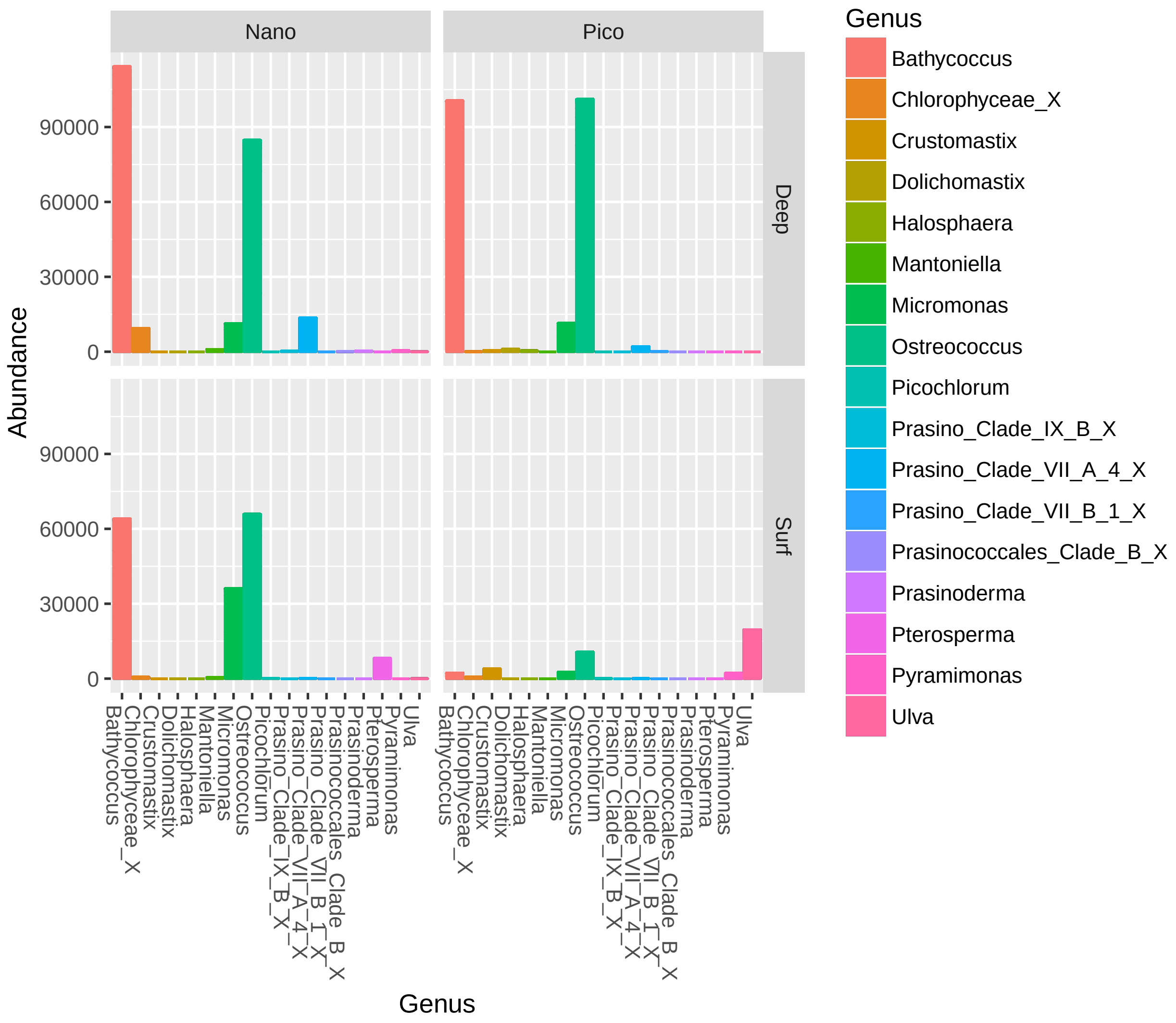
Regroup together Pico vs Nano samples

```
carbom_fraction <- merge_samples(carbom, "fraction")
plot_bar(carbom_fraction, fill = "Division") +
geom_bar(aes(color=Division, fill=Division), stat="identity", position="stack")
```



Keep only Chlorophyta and use color according to genus. Do separate panels Pico vs Nano and Surface vs Deep samples.

```
carbom_chloro <- subset_taxa(carbom, Division %in% c("Chlorophyta"))
plot_bar(carbom_chloro, x="Genus", fill = "Genus", facet_grid = level~fraction) +
geom_bar(aes(color=Genus, fill=Genus), stat="identity", position="stack")
```

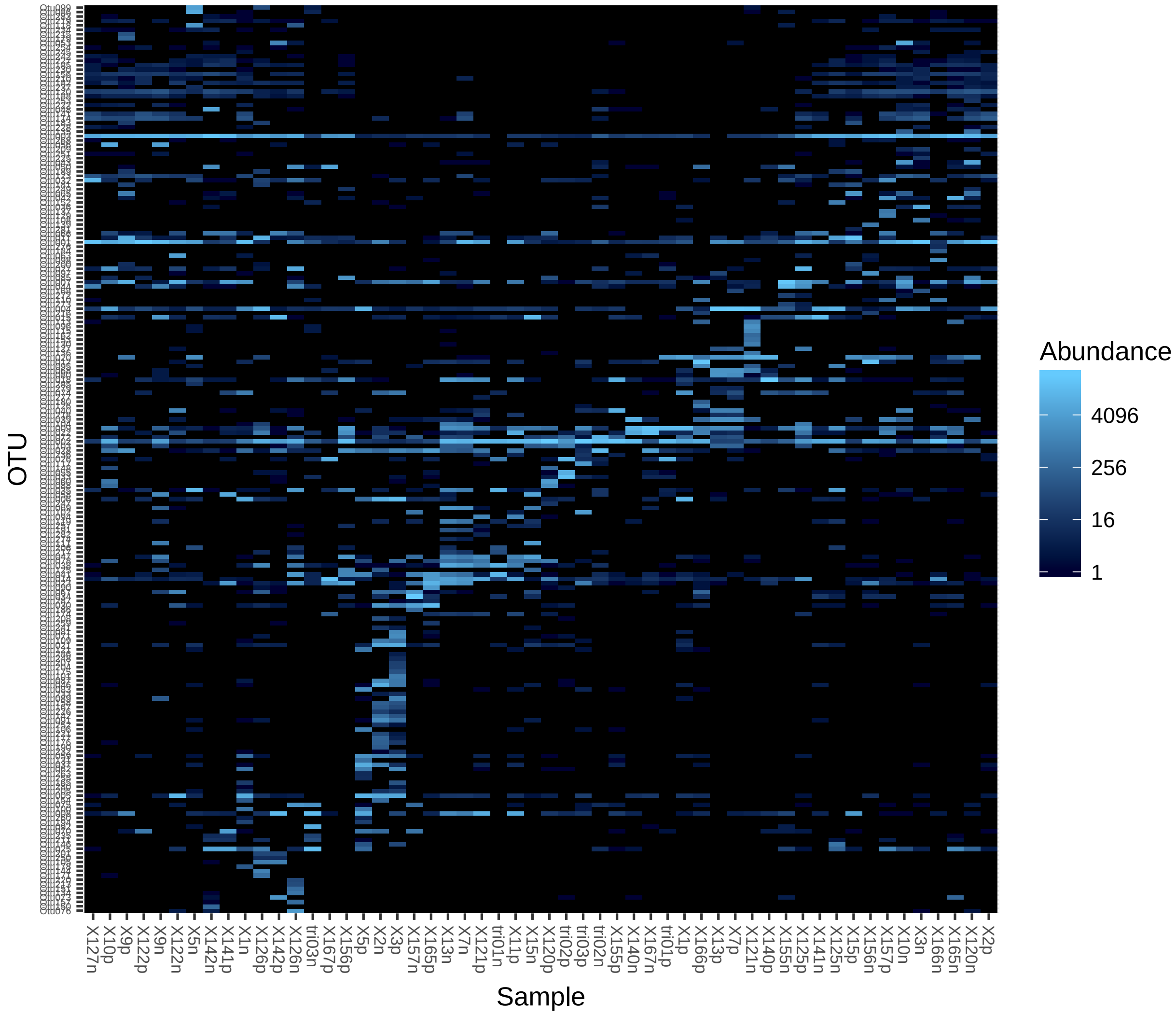


3.4 Heatmaps

A basic heatmap using the default parameters.

```
plot_heatmap(carbom, method = "NMDS", distance = "bray")
```

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



It is very very cluttered. It is better to only consider the most abundant OTUs for heatmaps. For example one can only take OTUs that represent at least 20% of reads in at least one sample. Remember we normalized all the sampples to median number of reads (total). We are left with only 33 OTUS which makes the reading much more easy.

```
carbom_abund <- filter_taxa(carbom, function(x) sum(x > total*0.20) > 0, TRUE)
carbom_abund
```

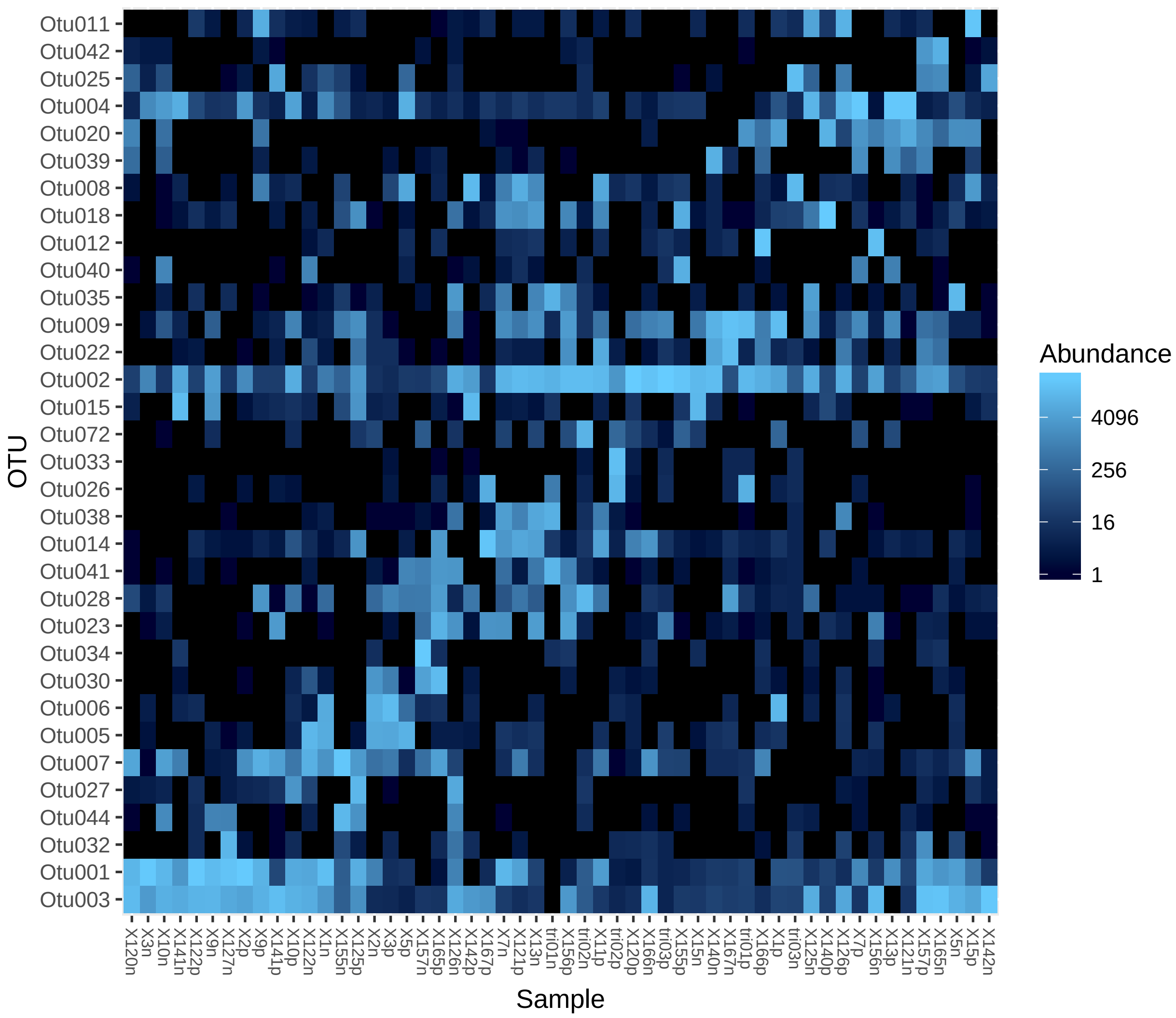
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 33 taxa and 54 samples ]
## sample_data() Sample Data:       [ 54 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 33 taxa by 7 taxonomic ranks ]
```

```
otu_table(carbom_abund)[1:8, 1:5]
```

```
## OTU Table:          [8 taxa and 5 samples]
##                      taxa are rows
##         X10n  X10p  X11p X120n X120p
## Otu001 13339  7346  3804 12662     3
## Otu002    18  8329 14958    30 36206
## Otu003  9692 10488    20 16537    11
## Otu004  3584  4943    33     7     9
## Otu005     0     6    11     0     5
## Otu006     0     9     0     0     5
## Otu007  4473   605   587  5894     3
## Otu008     1     9  6707     2    17
```

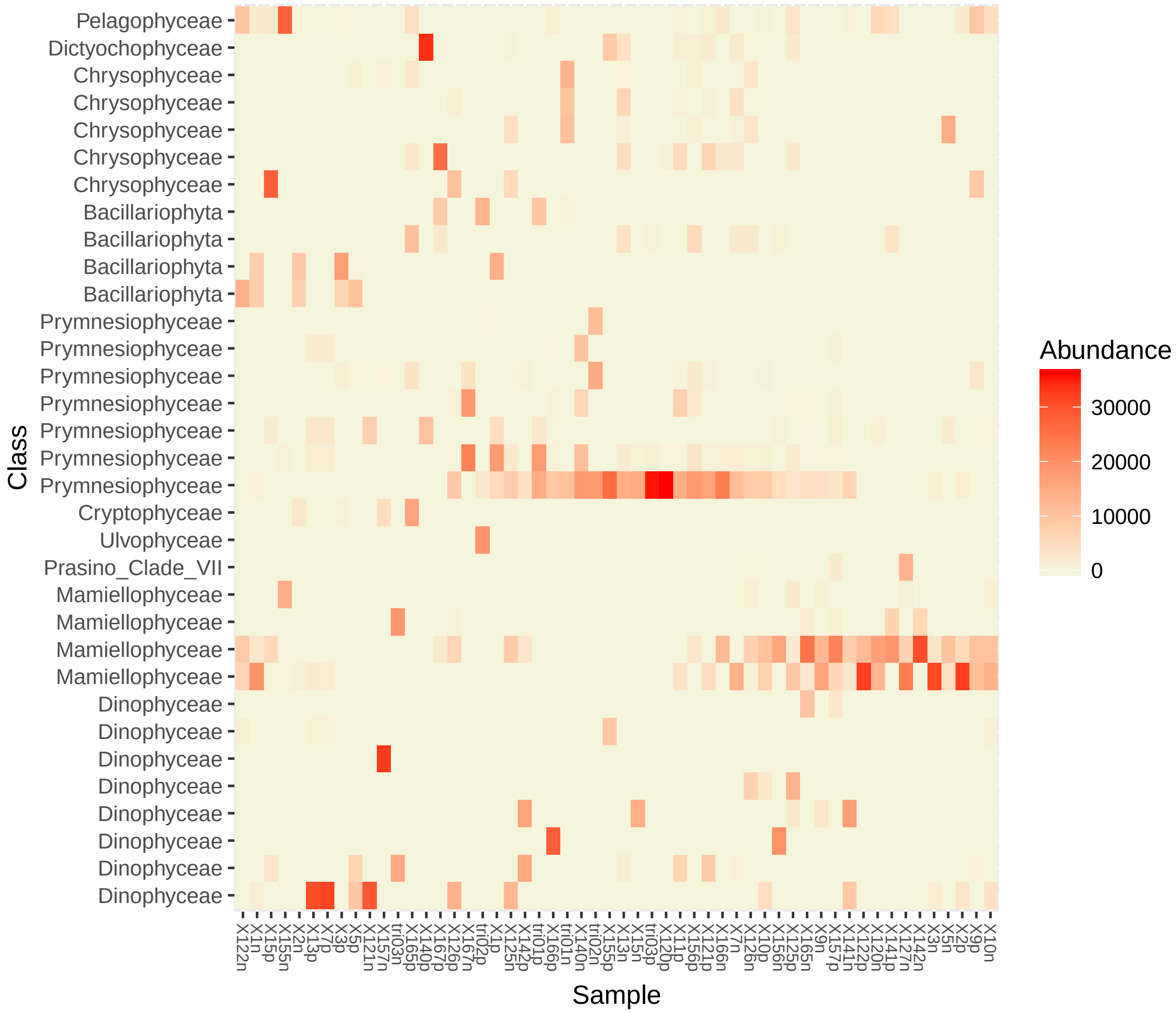
```
plot_heatmap(carbom_abund, method = "NMDS", distance = "bray")
```

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


Otu011
Otu042
Otu025
Otu004
Otu020
Otu039
Otu008
Otu018
Otu012
Otu040
Otu035
Otu009
Otu022
Otu002
Otu015
Otu072
Otu033
Otu026
Otu038
Otu014
Otu041
Otu028
Otu023
Otu034
Otu030
Otu006
Otu005
Otu007
Otu027
Otu044
Otu032
Otu001
Otu003
OTU
X120n
X3n
X10n
X141n
X122p
X9n
X127n
X2p
X9p
X141p
X10p
X122n
X1n
X155n
X125p
X2n
X3p
X5p
X157n
X165p
X126n
X142p
X167p
X7n
X121p
X13n
tri01n
X156p
tri02n
X11p
tri02p
X120p
X166n
tri03p
X155p
X15n
X140n
X167n
tri01p
X166p
X1p
tri03n
X125n
X140p
X126p
X7p
X156n
X13p
X121n
X157p
X165n
X5n
X15p
X142n
Sample
Abundance
4096
256
16
1

It is possible to use different distances and different multivaraite methods. For example Jaccard distance and MDS and label OTUs with Class, order by Class. We can also change the Palette (the default palette is a bit ugly…).

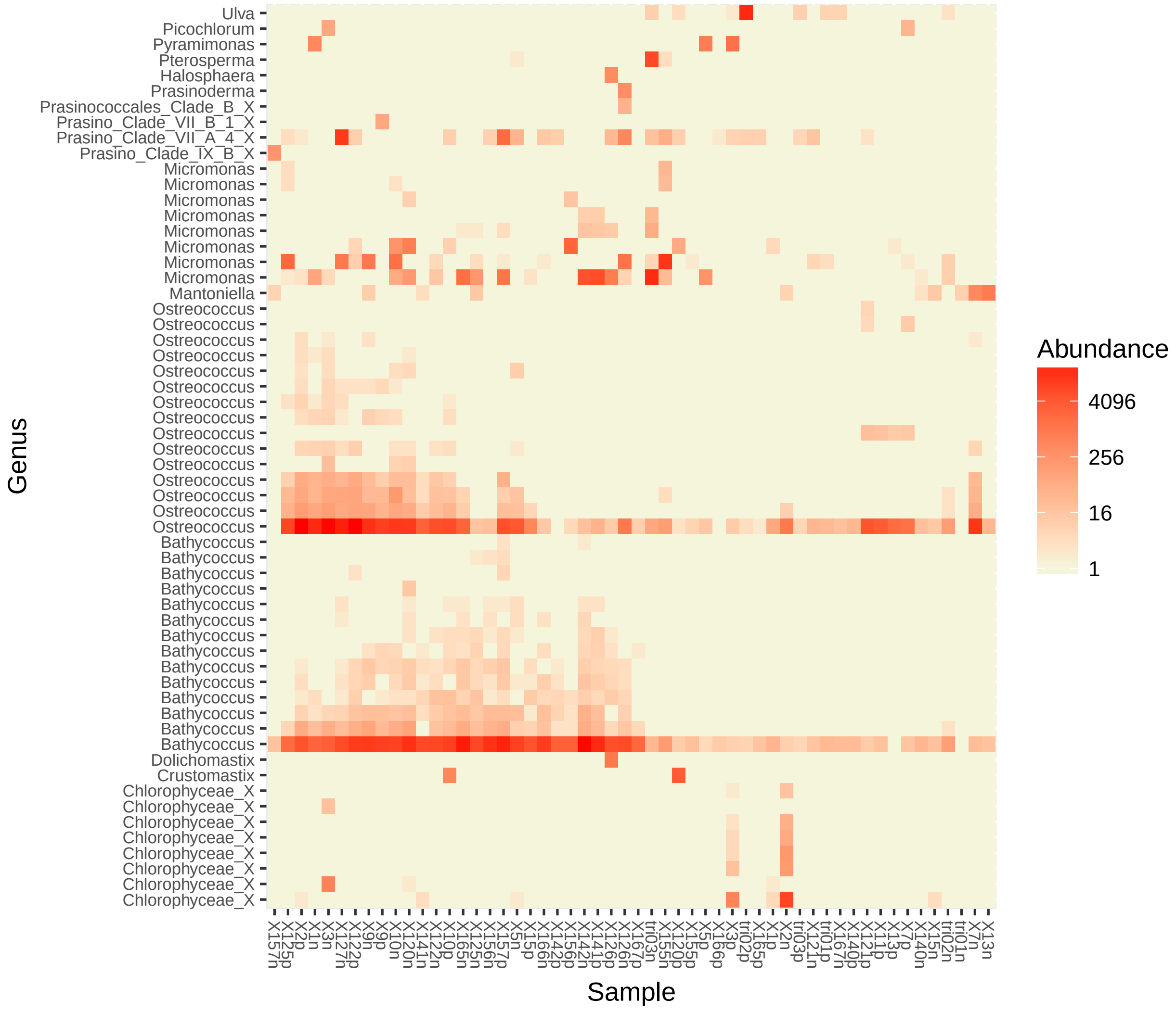
```
plot_heatmap(carbom_abund, method = "MDS", distance = "jaccard",
             taxa.label = "Class", taxa.order = "Class",
             trans=NULL, low="beige", high="red", na.value="beige")
```



Another strategy is to do a heatmap for a specific taxonomy group. For example we can taget the Chlorophyta and then label the OTUs using the Genus.

```r
plot_heatmap(carbom_chloro, method = "NMDS", distance = "bray",
             taxa.label = "Genus", taxa.order = "Genus",
             low="beige", high="red", na.value="beige")
```

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

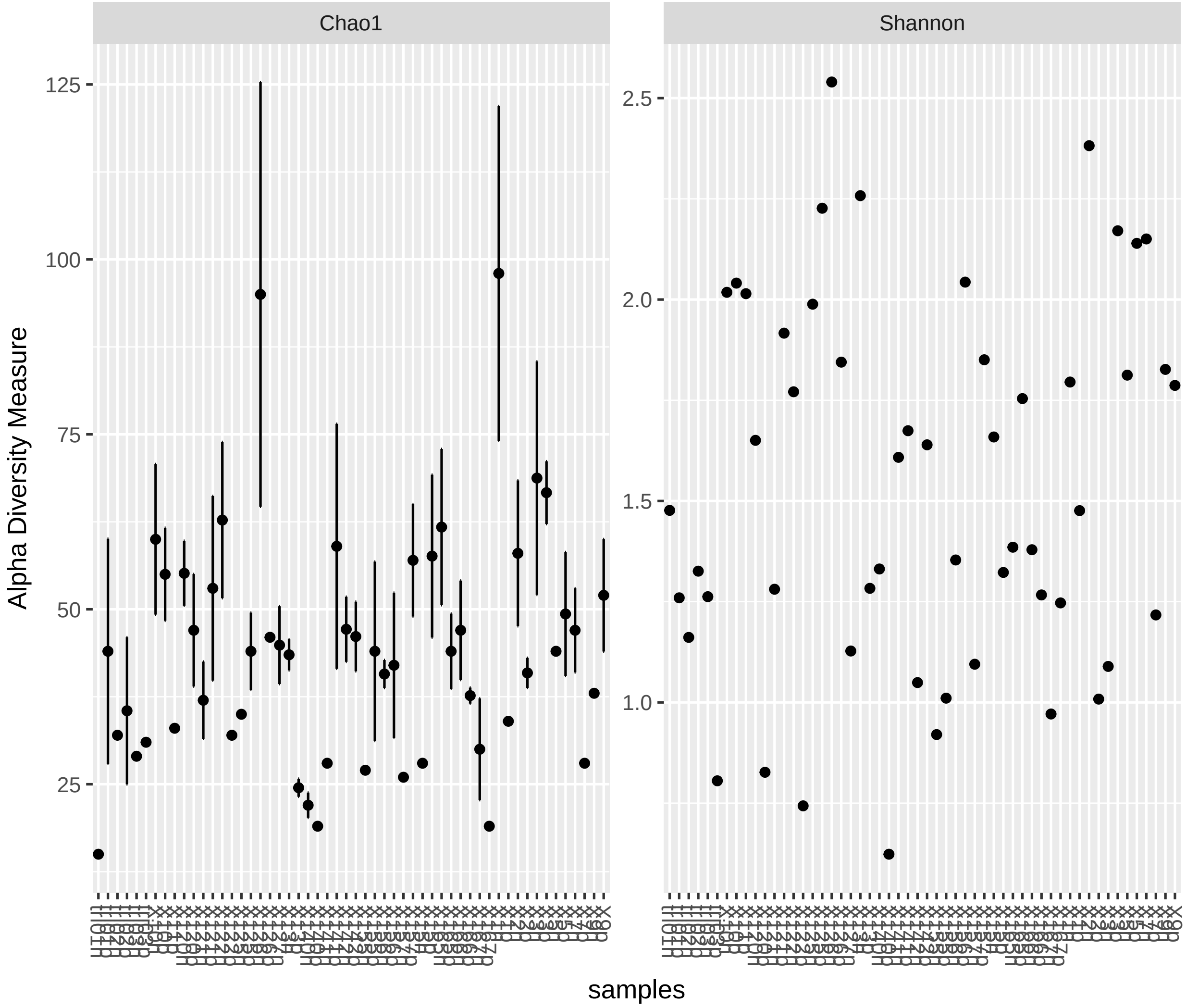


3.5 Alpha diversity

Plot Chao1 richness estimator and Shannon diversity estimator.

```r
plot_richness(carbom, measures=c("Chao1", "Shannon"))
```

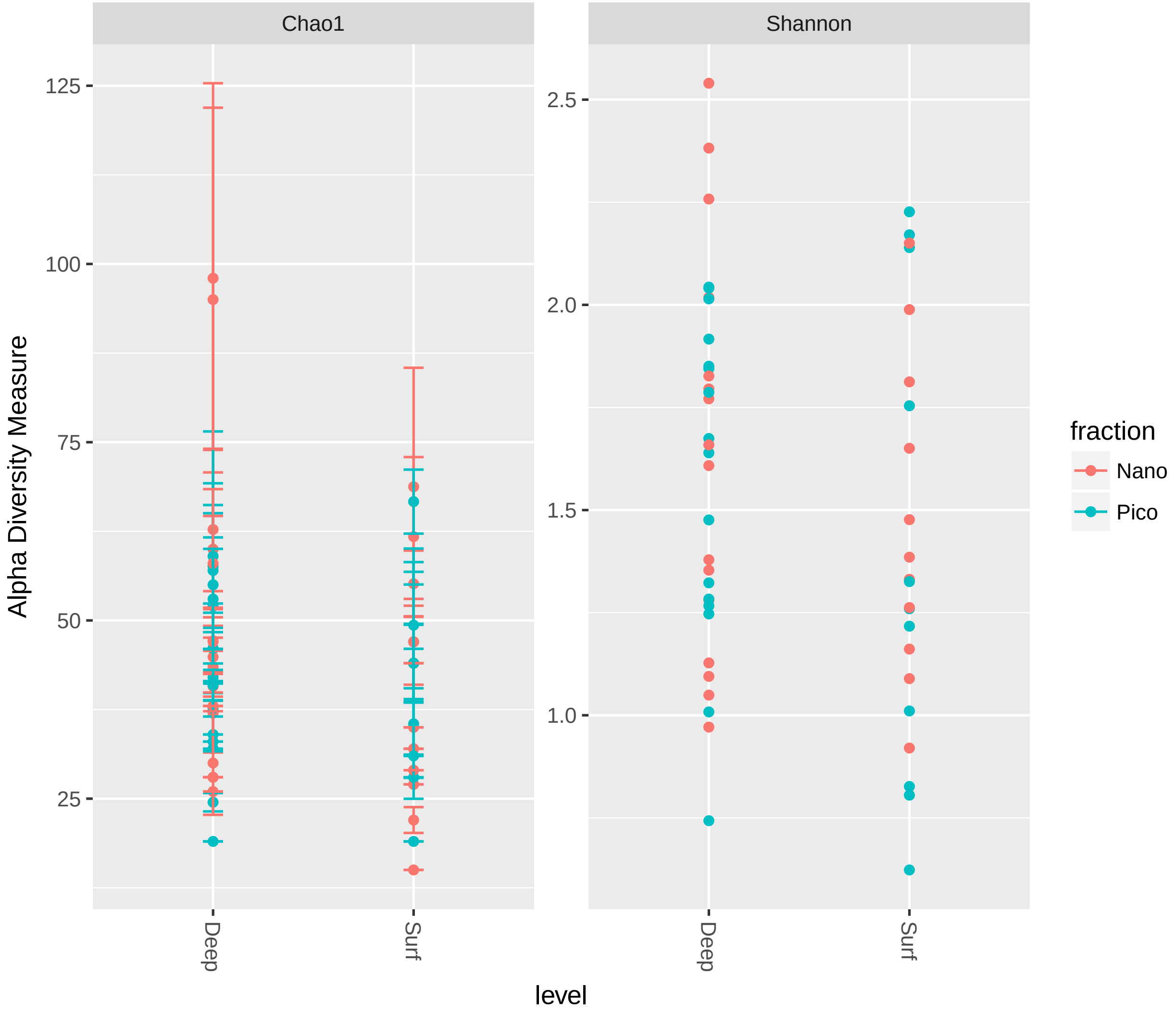
```
## Warning: Removed 54 rows containing missing values (geom_errorbar).
```



Regroup together samples from the same fraction.

```r
plot_richness(carbom, measures=c("Chao1", "Shannon"), x="level", color="fraction")
```

```
## Warning: Removed 54 rows containing missing values (geom_errorbar).
```



3.6 Ordination

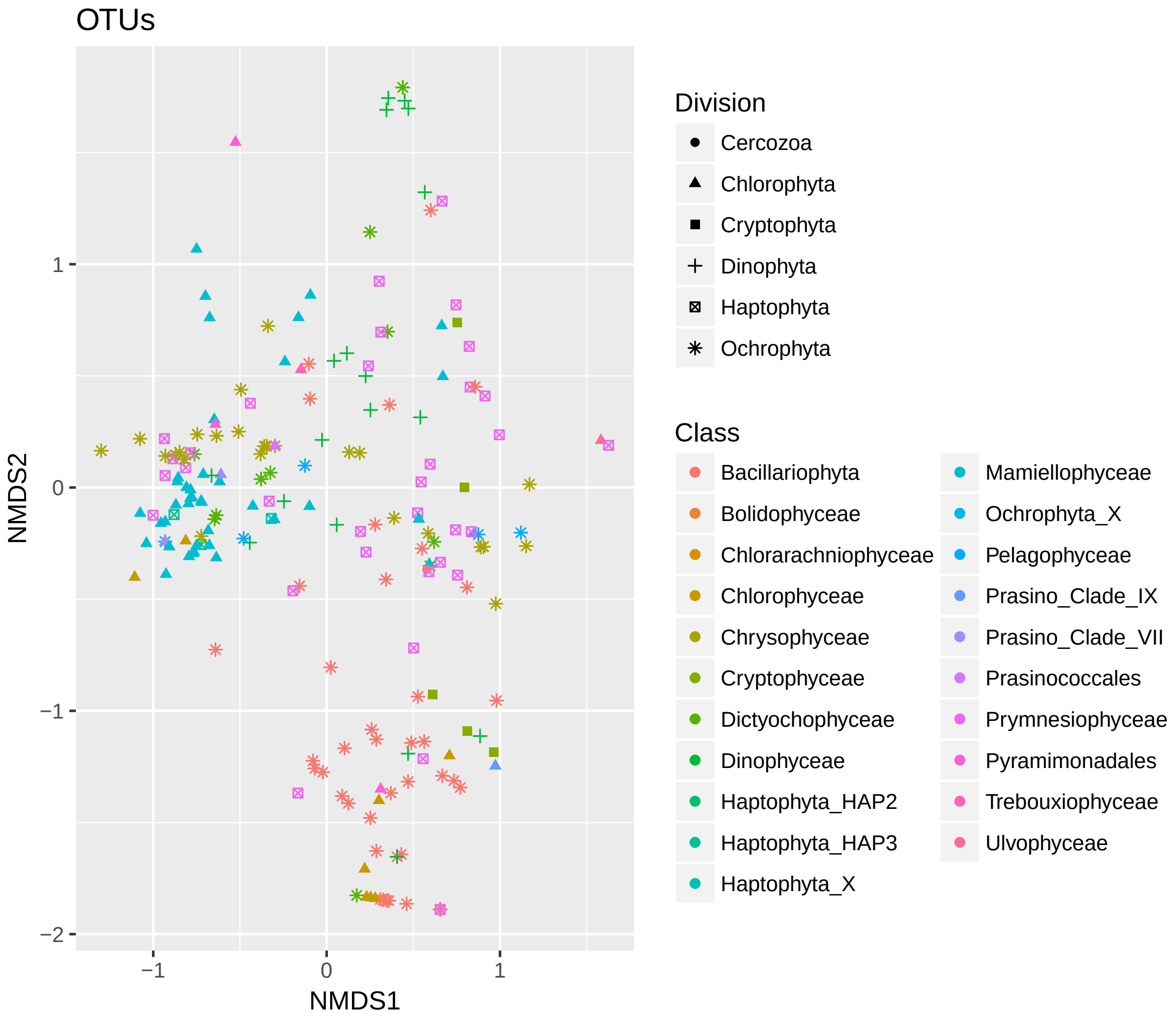
Do multivariate analysis based on Bray-Curtis distance and NMDS ordination.

```
carbom.ord <- ordinate(carbom, "NMDS", "bray")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.2317058
## Run 1 stress 0.2322769
## Run 2 stress 0.2576847
## Run 3 stress 0.2561343
## Run 4 stress 0.2504777
## Run 5 stress 0.2524615
## Run 6 stress 0.2494219
## Run 7 stress 0.2335095
## Run 8 stress 0.2405252
## Run 9 stress 0.2542259
## Run 10 stress 0.2511304
## Run 11 stress 0.2537237
## Run 12 stress 0.2585122
## Run 13 stress 0.2441821
## Run 14 stress 0.2488392
## Run 15 stress 0.2486157
## Run 16 stress 0.2613244
## Run 17 stress 0.2570676
## Run 18 stress 0.2482644
## Run 19 stress 0.2316124
## ... New best solution
## ... Procrustes: rmse 0.1112088  max resid 0.2958914
## Run 20 stress 0.2512223
## *** No convergence -- monoMDS stopping criteria:
##     20: stress ratio > sratmax
```

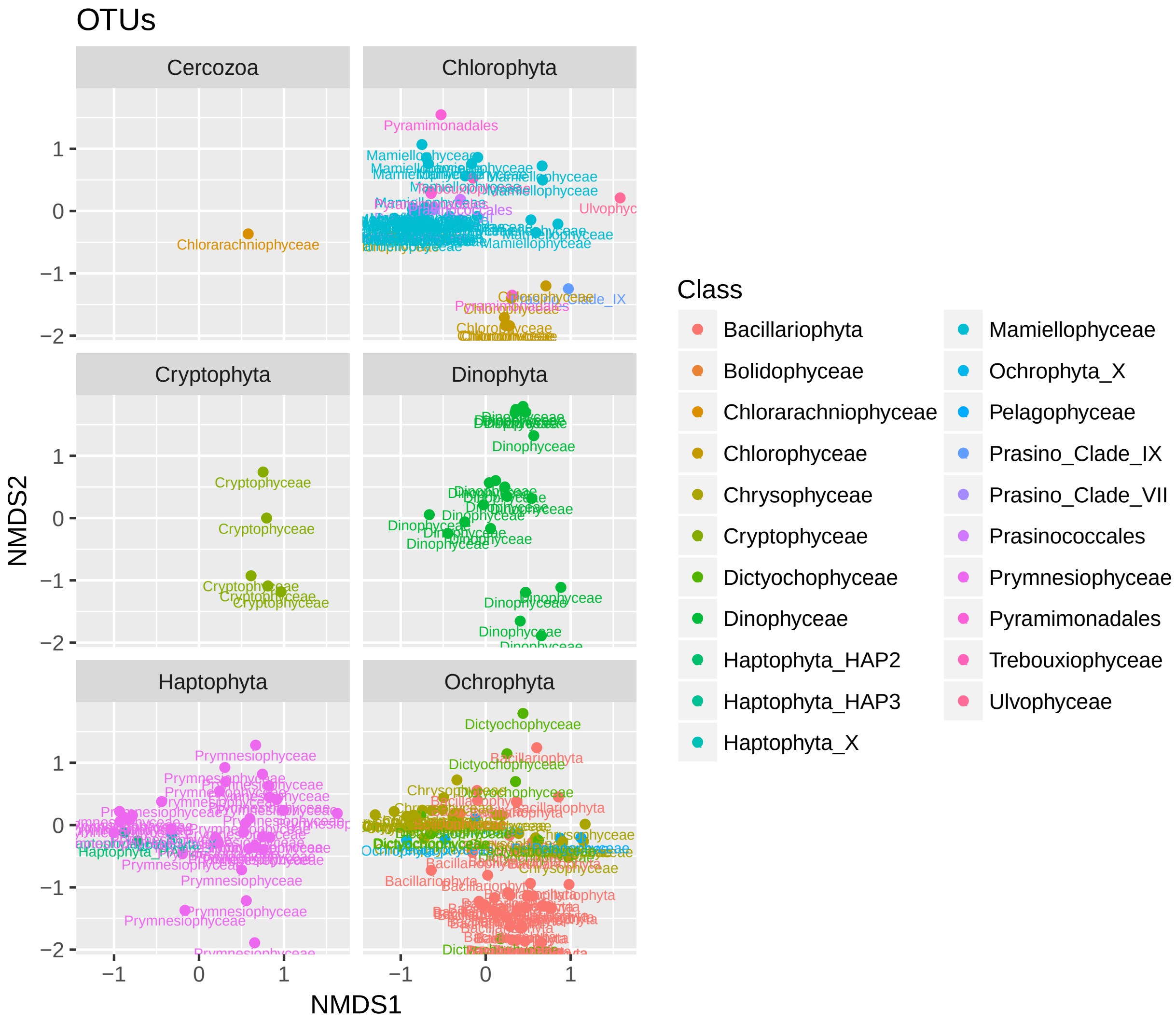
Plot **OTUs**

```
plot_ordination(carbom, carbom.ord, type="taxa", color="Class", shape= "Division",
                title="OTUs")
```



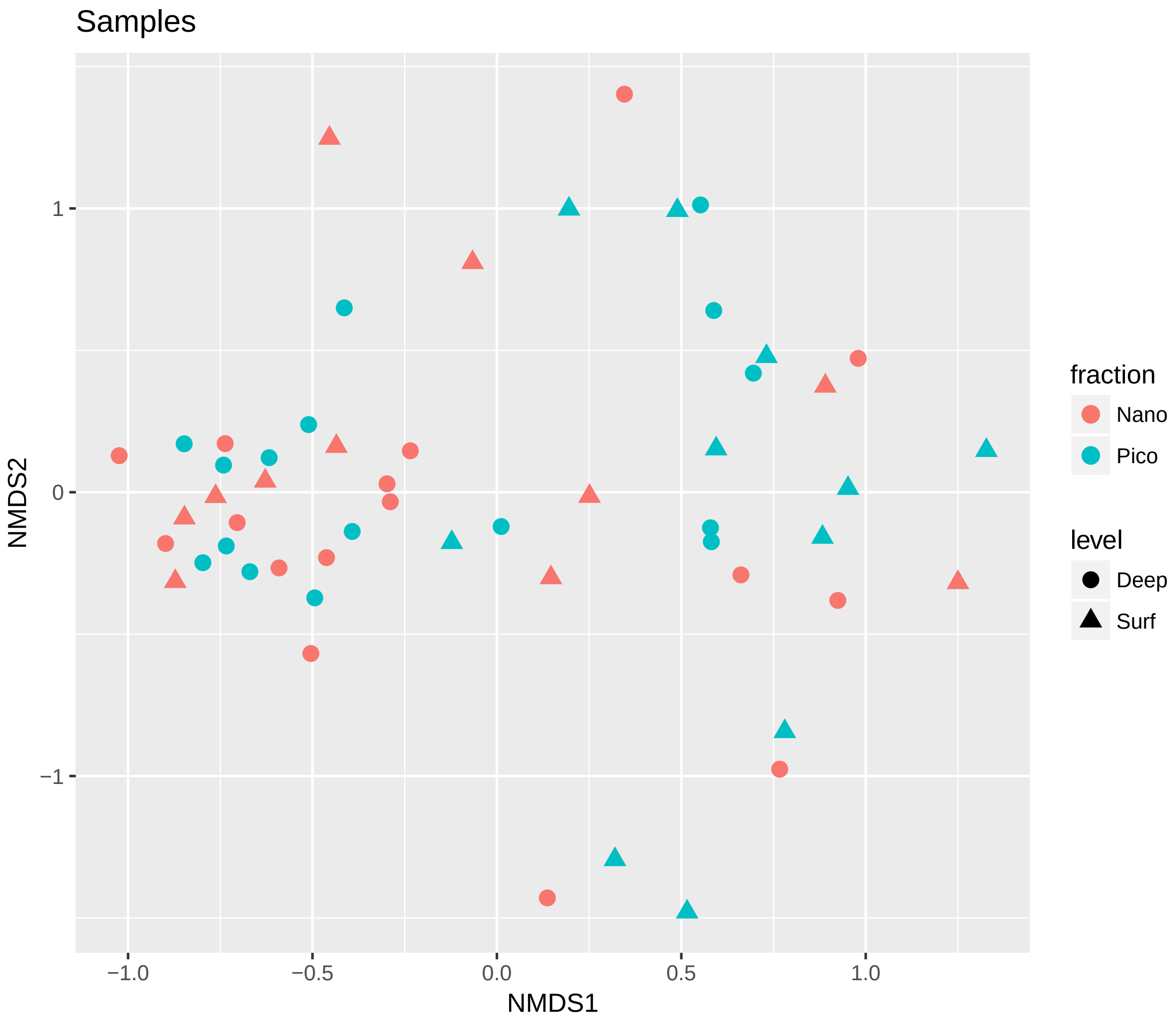
A bit confusing, so make it more easy to visualize by breaking according to taxonomic division.

```
plot_ordination(carbom, carbom.ord, type="taxa", color="Class",
                title="OTUs", label="Class") +
facet_wrap(~Division, 3)
```



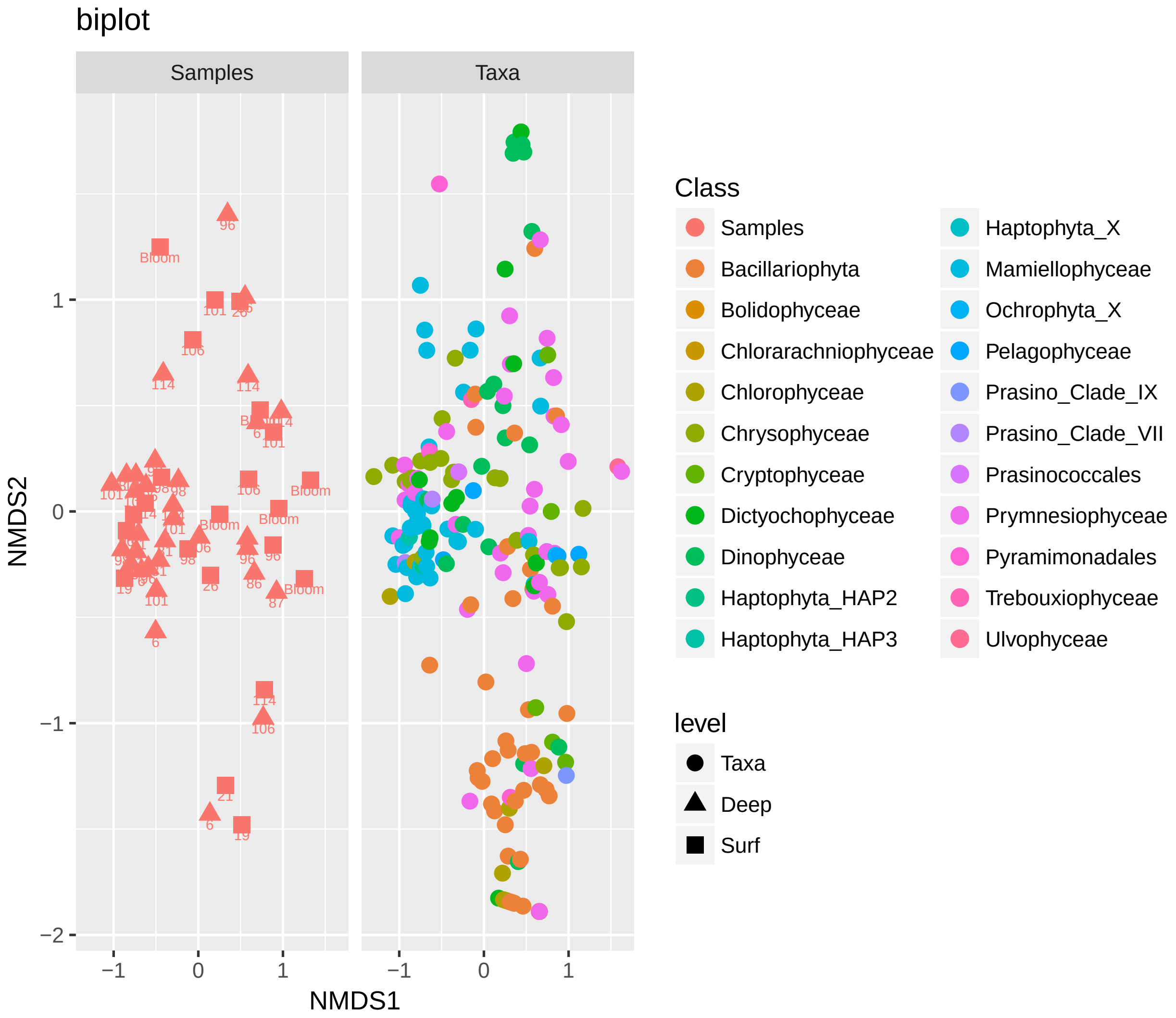
Now display **samples** and enlarge the points to make it more easy to read.

```
plot_ordination(carbom, carbom.ord, type="samples", color="fraction",
                shape="level", title="Samples") + geom_point(size=3)
```



Display both samples and OTUs but in 2 different panels.

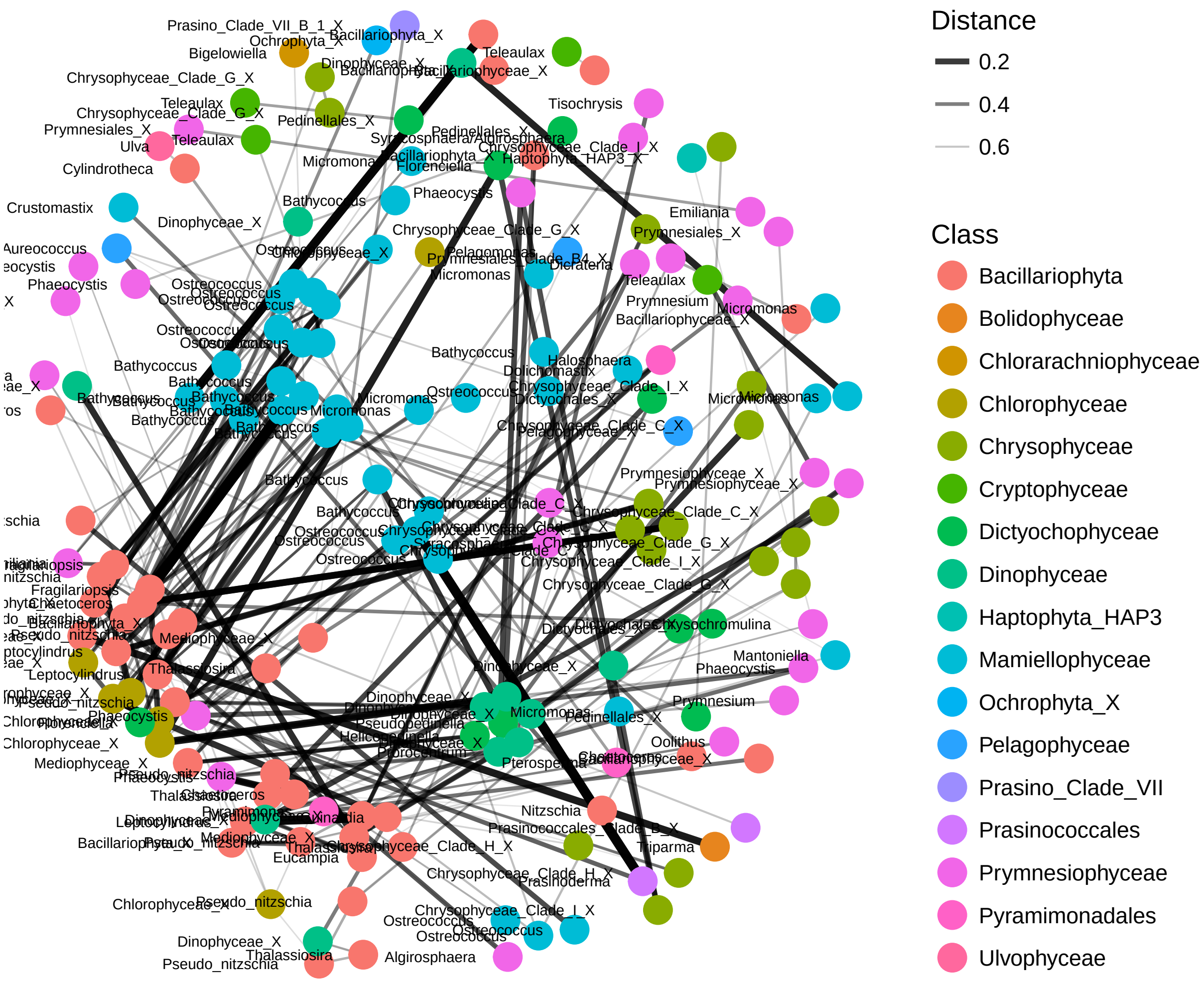
```r
plot_ordination(carbom, carbom.ord, type="split", color="Class",
                shape="level", title="biplot", label = "station") +
geom_point(size=3)
```



3.7 Network analysis

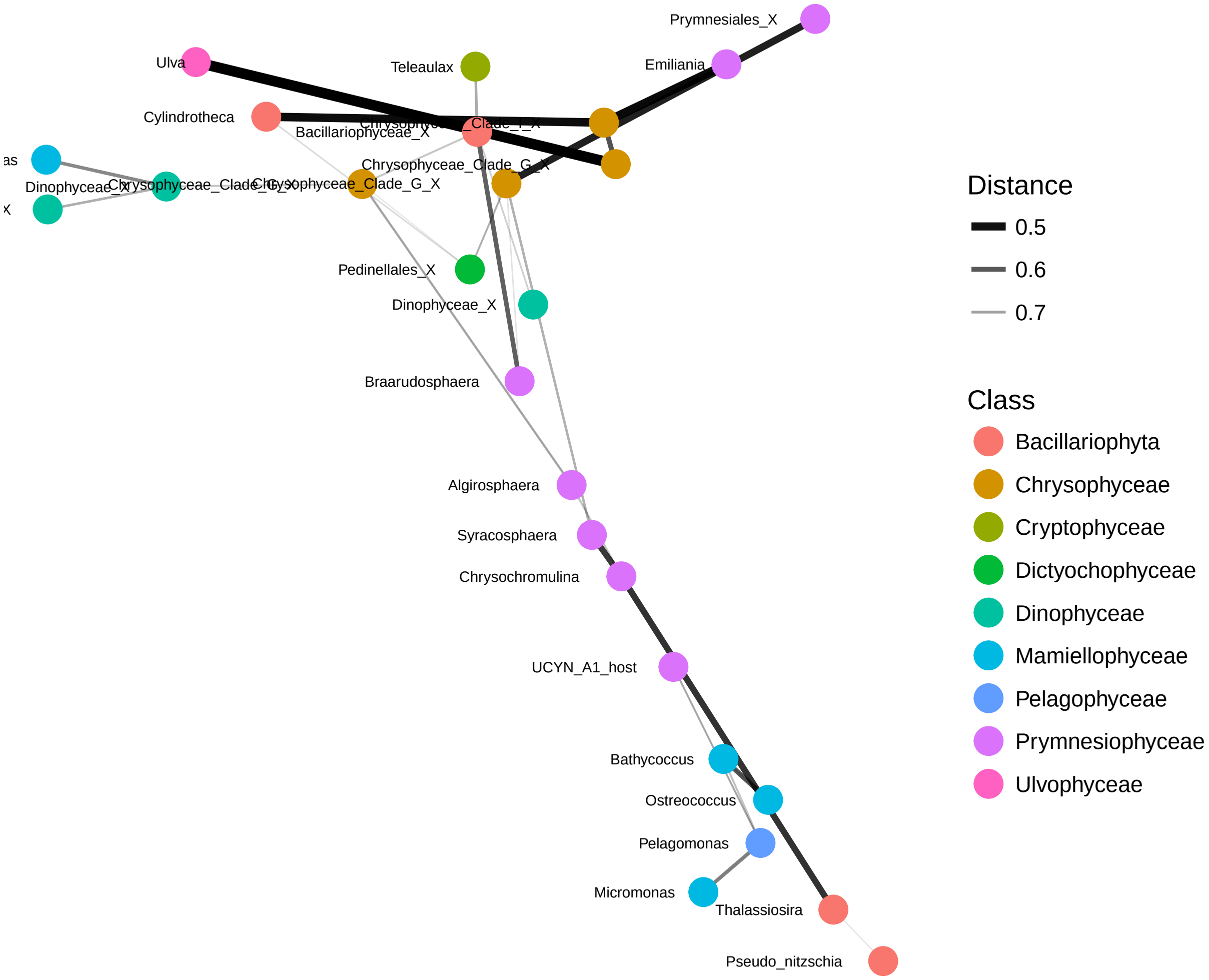
Simple network analysis

```r
plot_net(carbom, distance = "bray", type = "taxa",
         maxdist = 0.7, color="Class", point_label="Genus")
```



This is quite confusing. Let us make it more simple by using only major OTUs

```
plot_net(carbom_abund, distance = "bray", type = "taxa",
         maxdist = 0.8, color="Class", point_label="Genus")
```


Prymnesiales_X
Ulva
Teleaulax
Emiliania
Cylindrotheca
Chrysophyceae_Clade_I_X
Bacillariophyceae_X
as
Chrysophyceae_Clade_G_X
Dinophyceae_X
Chrysophyceae_Clade_G_X
Chrysophyceae_Clade_G_X
X
Pedinellales_X
Dinophyceae_X
Braarudosphaera
Algirosphaera
Syracosphaera
Chrysochromulina
UCYN_A1_host
Bathycoccus
Ostreococcus
Pelagomonas
Micromonas
Thalassiosira
Pseudo_nitzschia
Distance
Class
Bacillariophyta
Chrysophyceae
Cryptophyceae
Dictyochophyceae
Dinophyceae
Mamiellophyceae
Pelagophyceae
Prymnesiophyceae
Ulvophyceae