Phyloseq tutorial

Daniel Vaulot

7 juin 2017

Introduction

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

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

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.

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.

Script description

Load necessary libraries

- phyloseq
- ggplot2
- readxl : necessary to import the data from Excel file
- dplyr : becessary to reformat dataframe

```
library("phyloseq")
library("ggplot2")
library("readxl")
library("dplyr")
```

```
##
## 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
```

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

```
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

```
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

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

- Idem for the two other matrixes

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

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

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

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

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

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

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

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

Transform to phyloseq objects

```
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:         [ 20 taxa and 10 samples ]
## sample_data() Sample Data:       [ 10 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 20 taxa by 7 taxonomic ranks ]
```

Visualize data

```
  sample_names(carbom)
```

```
##  [1] "X10n" "X10p" "X11n" "X11p" "X13n" "X13p" "X15n" "X15p" "X9n"  "X9p"
```

```
  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:         [ 20 taxa and 9 samples ]
## sample_data() Sample Data:       [ 9 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 20 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:         [ 15 taxa and 9 samples ]
## sample_data() Sample Data:       [ 9 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 15 taxa by 7 taxonomic ranks ]
```

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

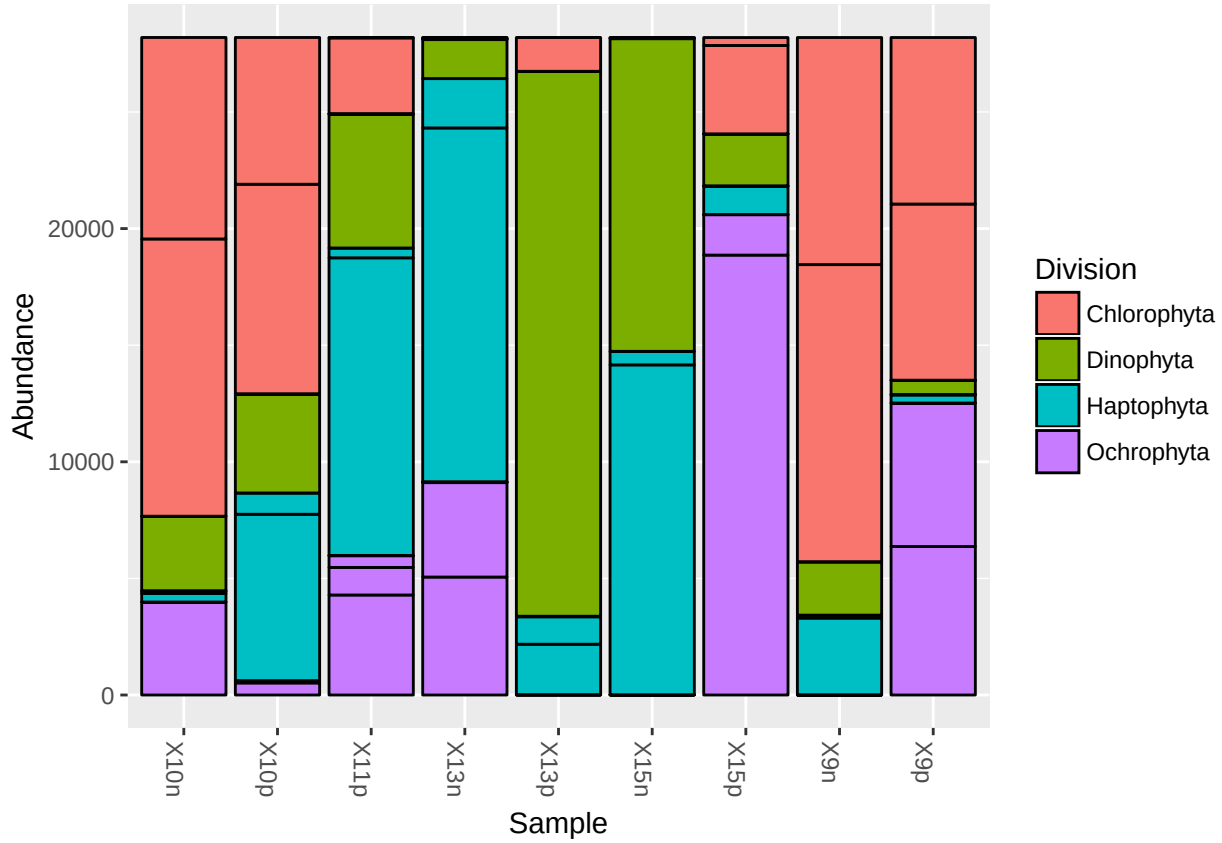
```
  total = median(sample_sums(carbom))
  standf = function(x, t=total) round(t * (x / sum(x)))
  carbom = transform_sample_counts(carbom, standf)
```

The number of reads used for normalization is **28193**.

Bar graphs

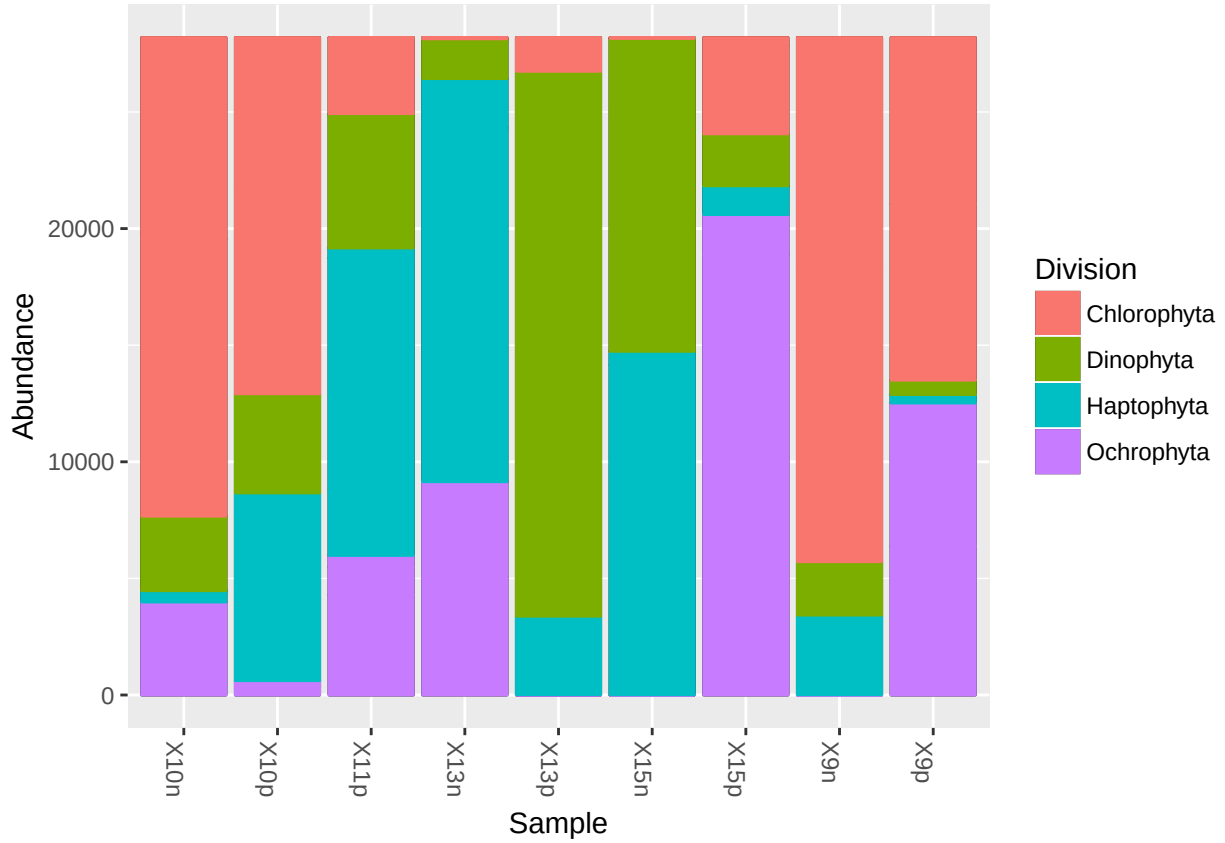
Basic bar graph based on Division

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



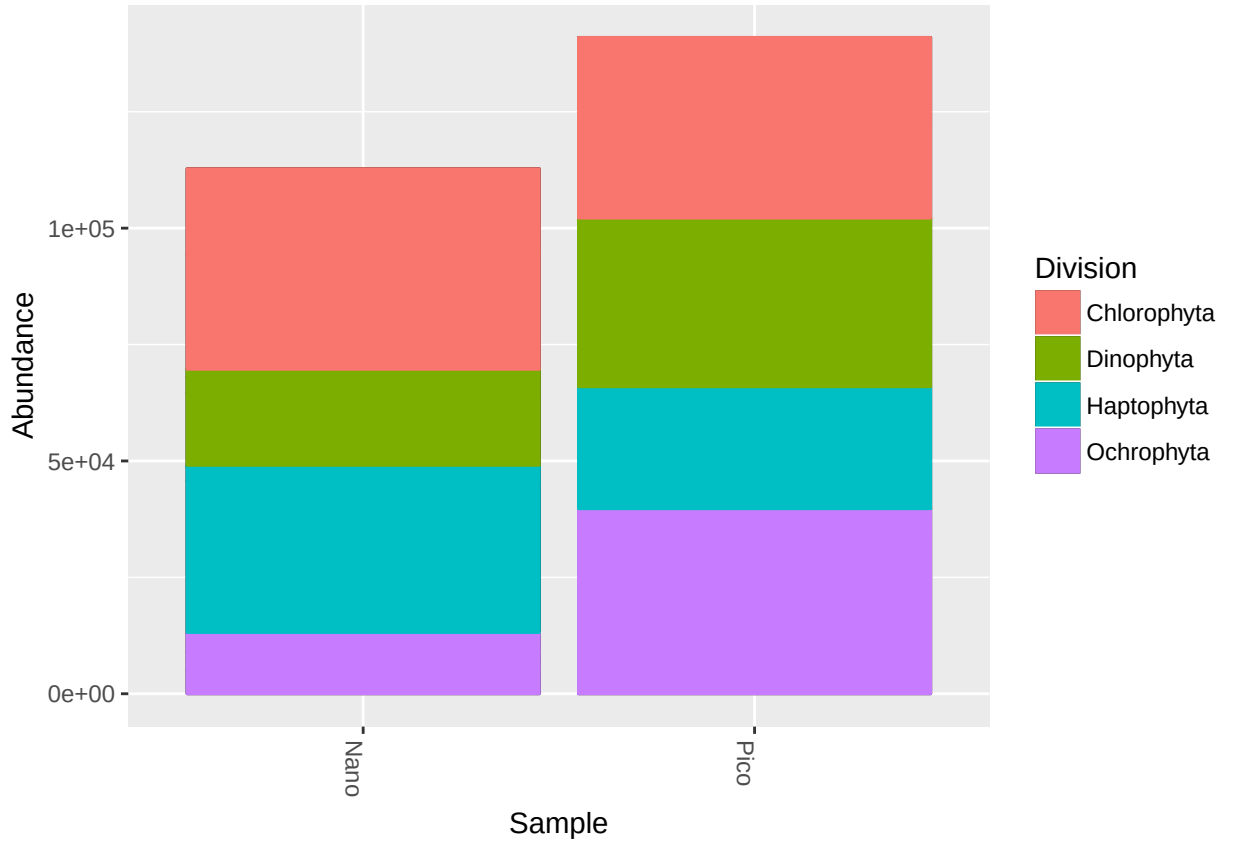
Make the bargraph nicer by removing OTUs boundaries. This is done by adding ggplot2 modifier.

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



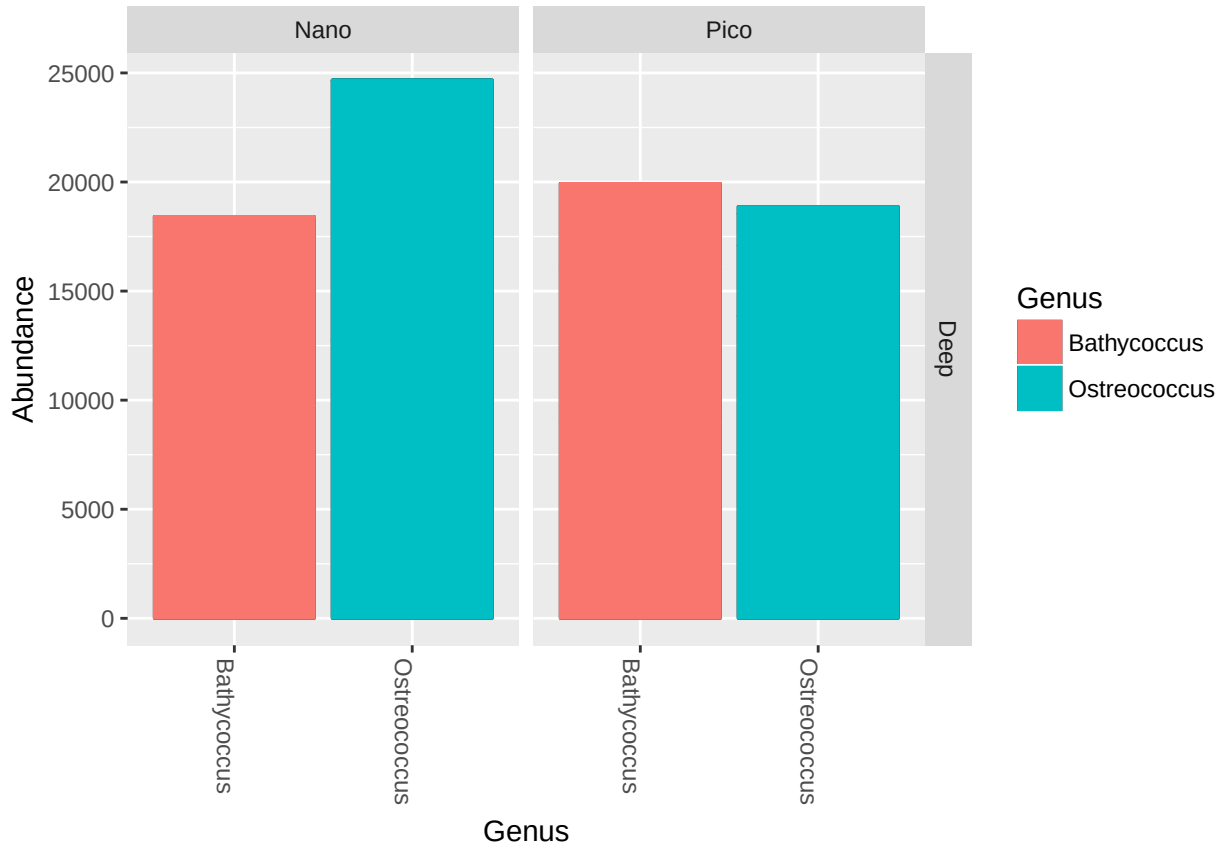
Regroup together Pico vs Nano samples

```
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")
```

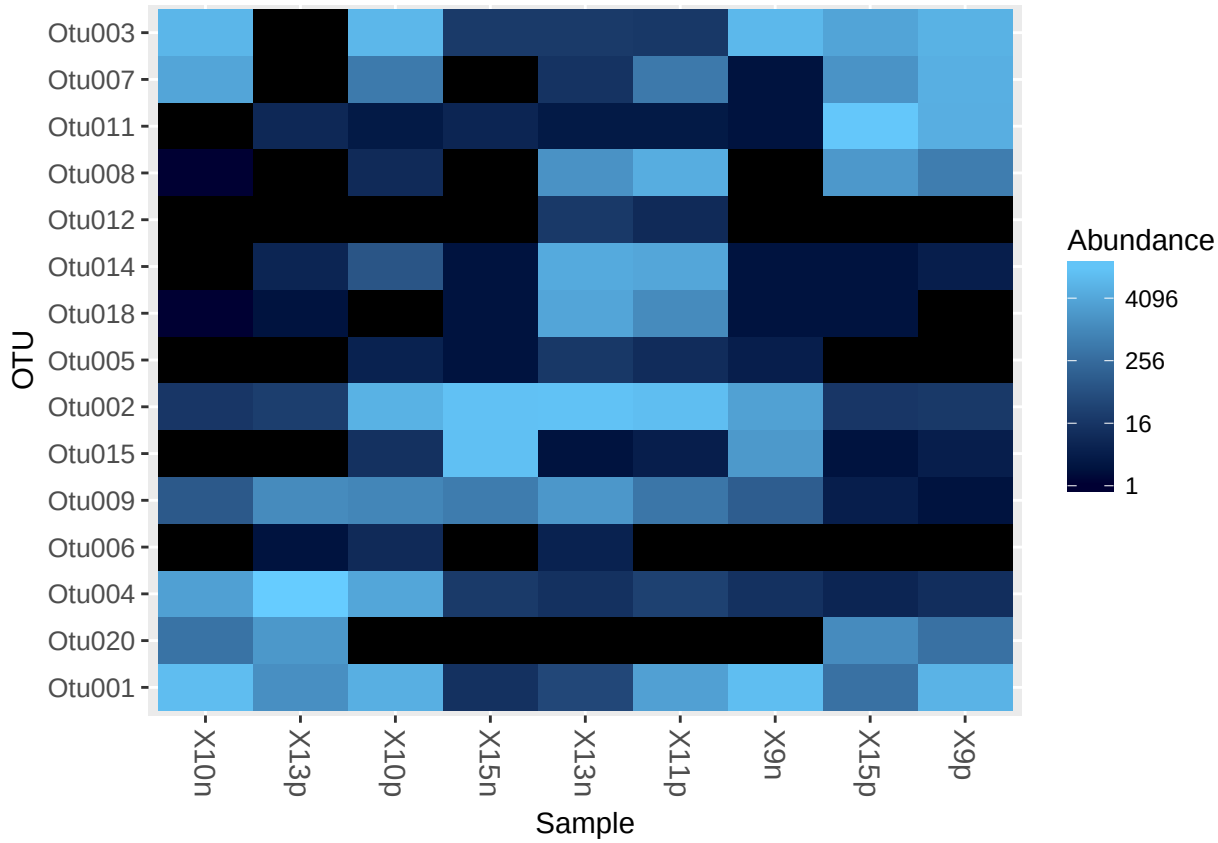


Heatmaps

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
```

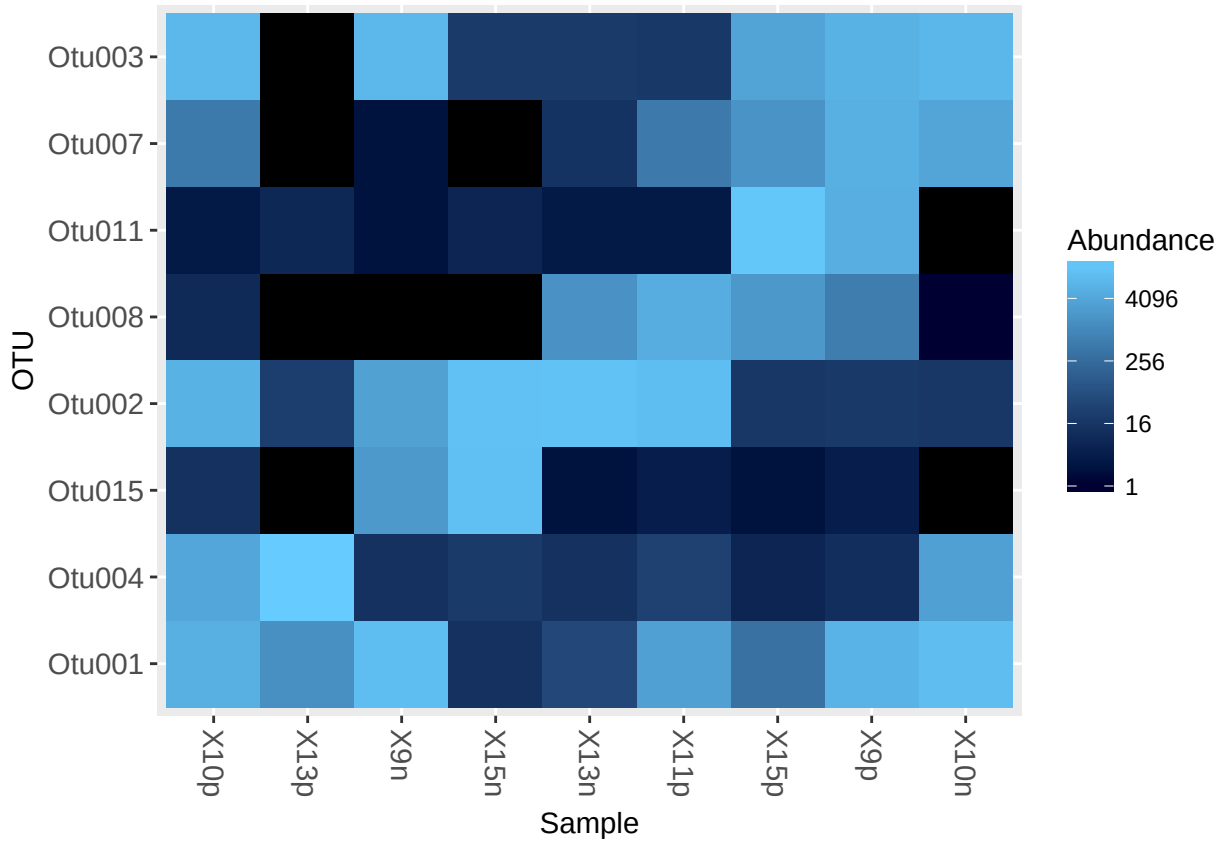
```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 8 taxa and 9 samples ]
## sample_data() Sample Data:       [ 9 samples by 27 sample variables ]
## tax_table()   Taxonomy Table:    [ 8 taxa by 7 taxonomic ranks ]
```

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

```
## OTU Table:          [8 taxa and 5 samples]
##                      taxa are rows
##         X10n X10p  X11p  X13n  X13p
## Otu001 11888 6292  3245    40  1448
## Otu002    16 7134 12761 15170    24
## Otu003  8638 8983    17    19     0
## Otu004  3194 4234    29    12 23366
## Otu007  3986  518   501    13     0
## Otu008     1    8  5722  1672     0
## Otu011     0    3     3     3     7
## Otu015     0   12     4     2     0
```

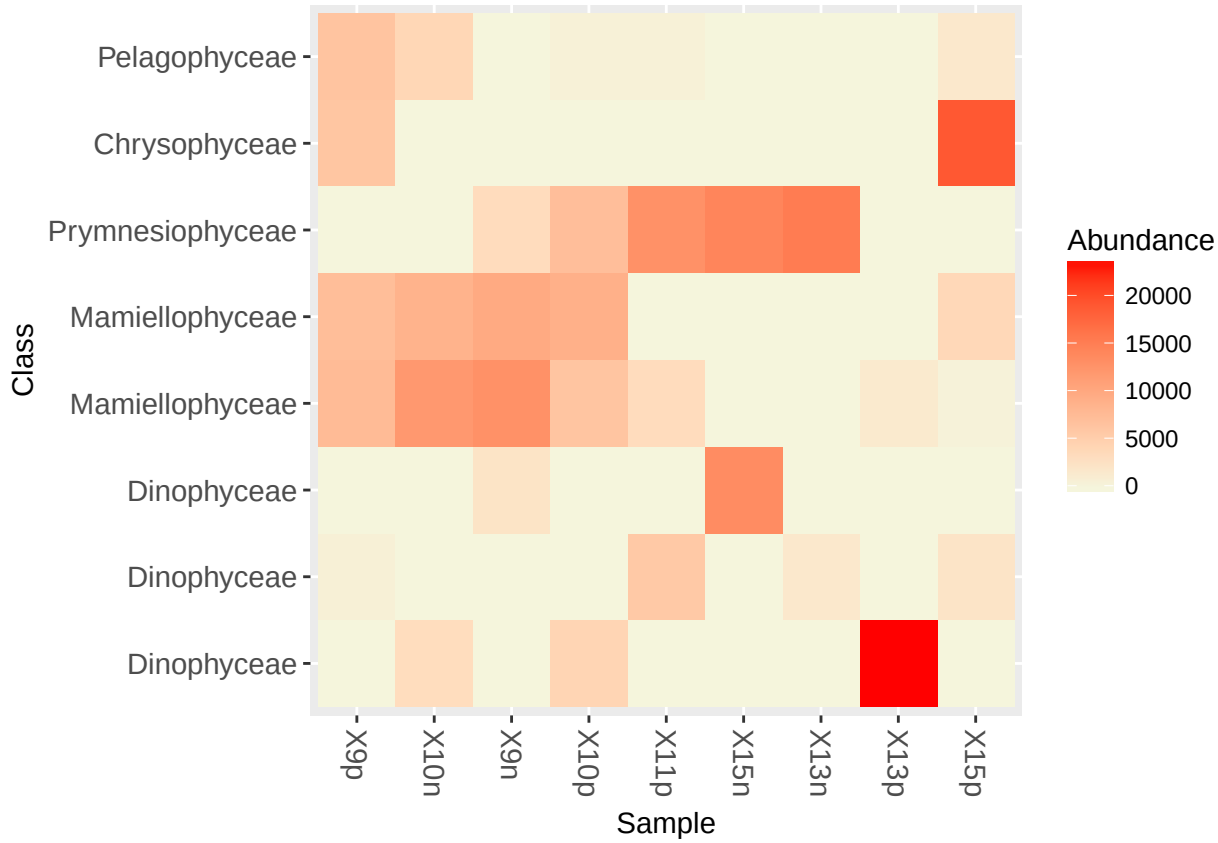
```
plot_heatmap(carbom_abund, method = "NMDS", distance = "bray")
```

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



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")
```

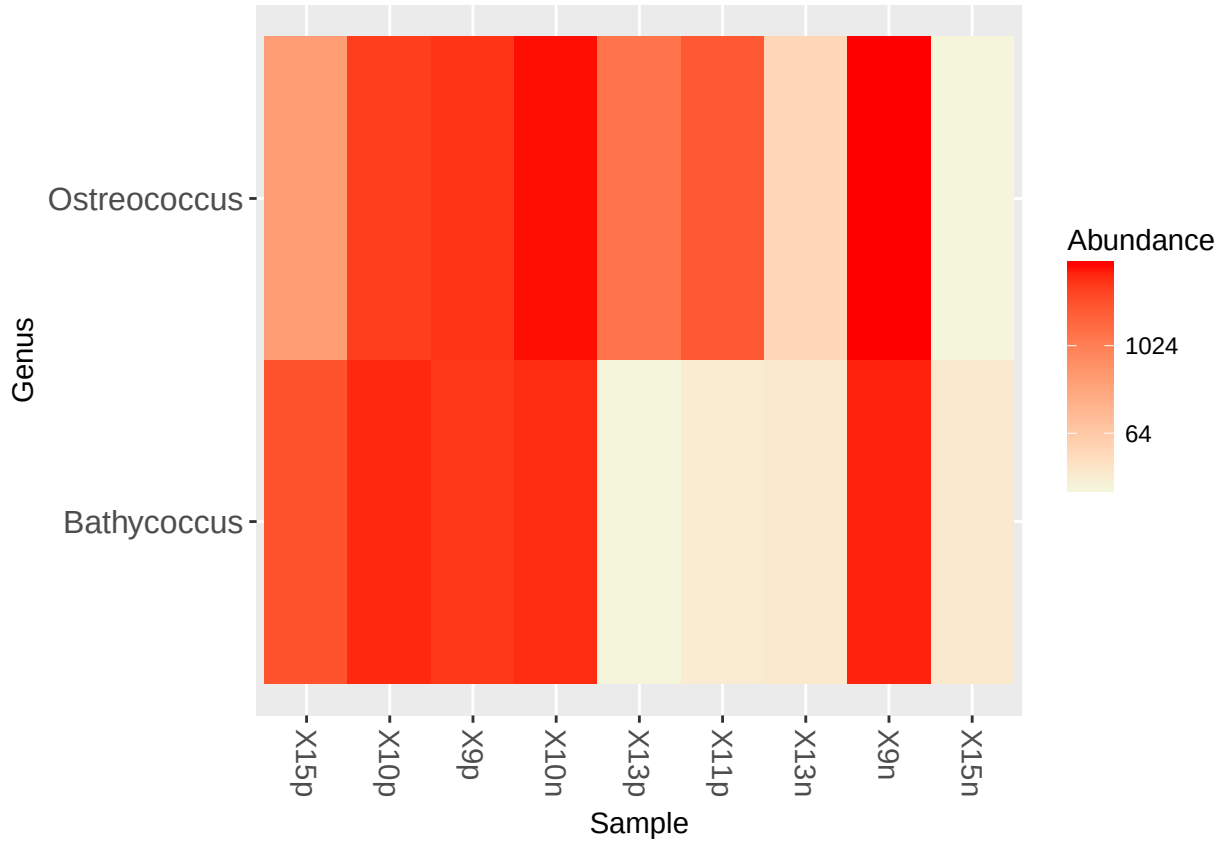


Another strategy is to do a heatmap for a specific taxonomy group. For example we can taget the Chlorophyta and then label the OTUs using the Genus.

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

```
## Warning in metaMDS(veganifyOTU(physeq), distance, ...): Stress is (nearly)
## zero - you may have insufficient data
```

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

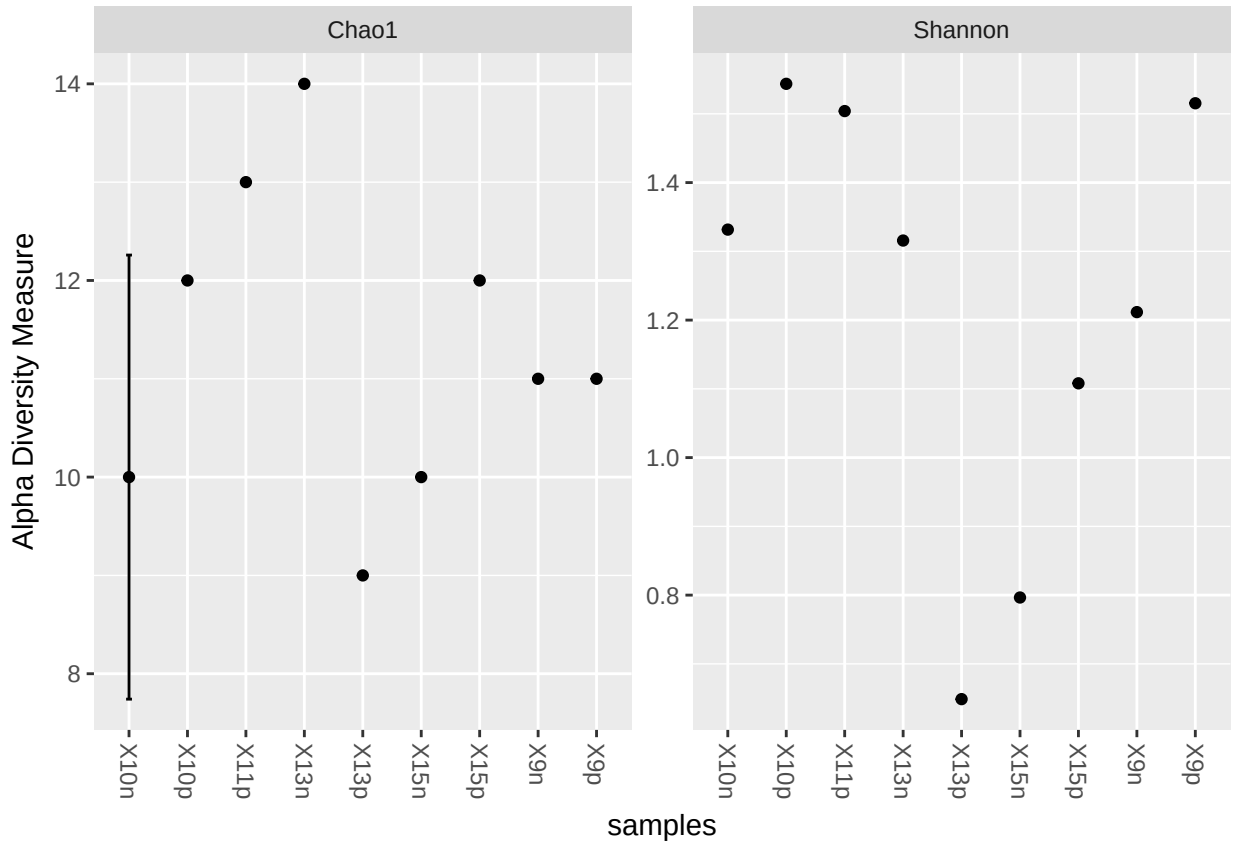


Alpha diversity

Plot Chao1 richness estimator and Shannon diversity estimator.

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

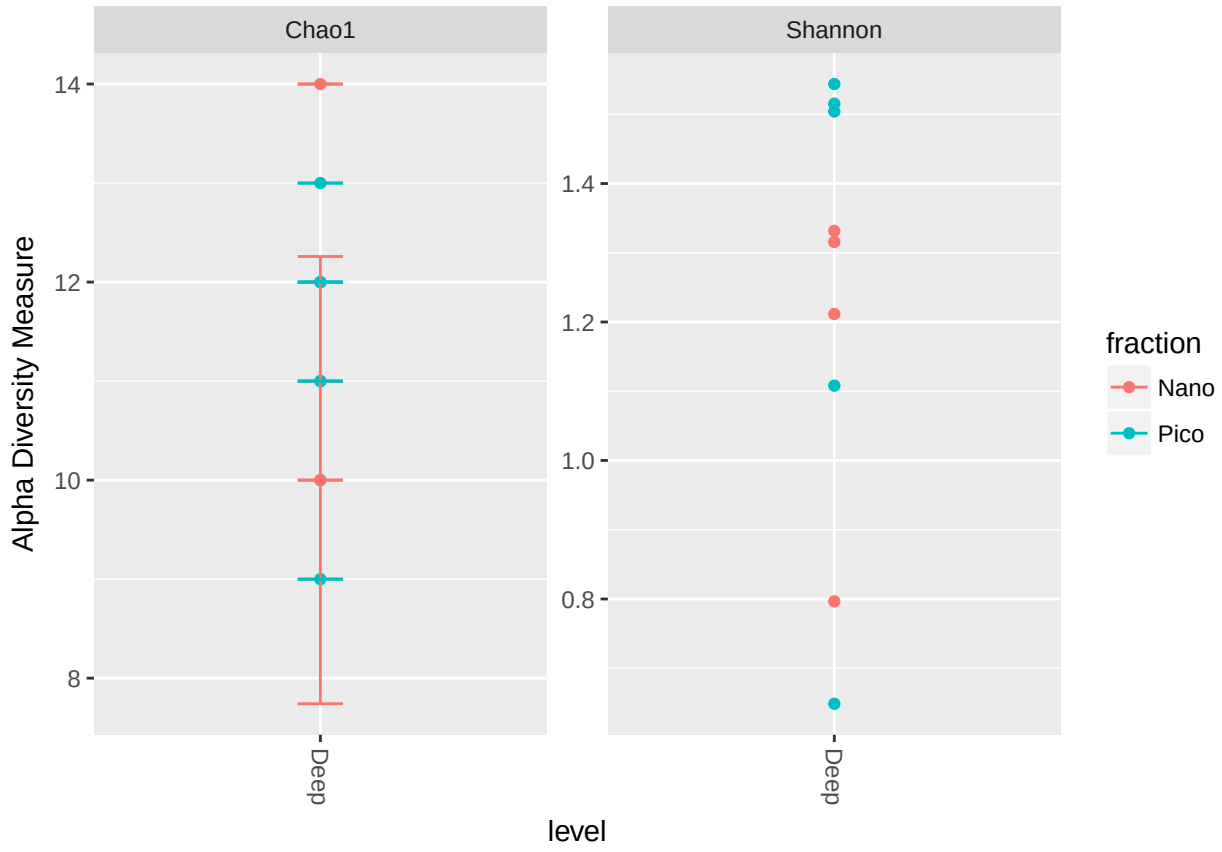
```
## Warning: Removed 9 rows containing missing values (geom_errorbar).
```



Regroup together samples from the same fraction.

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

```
## Warning: Removed 9 rows containing missing values (geom_errorbar).
```



Ordination

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

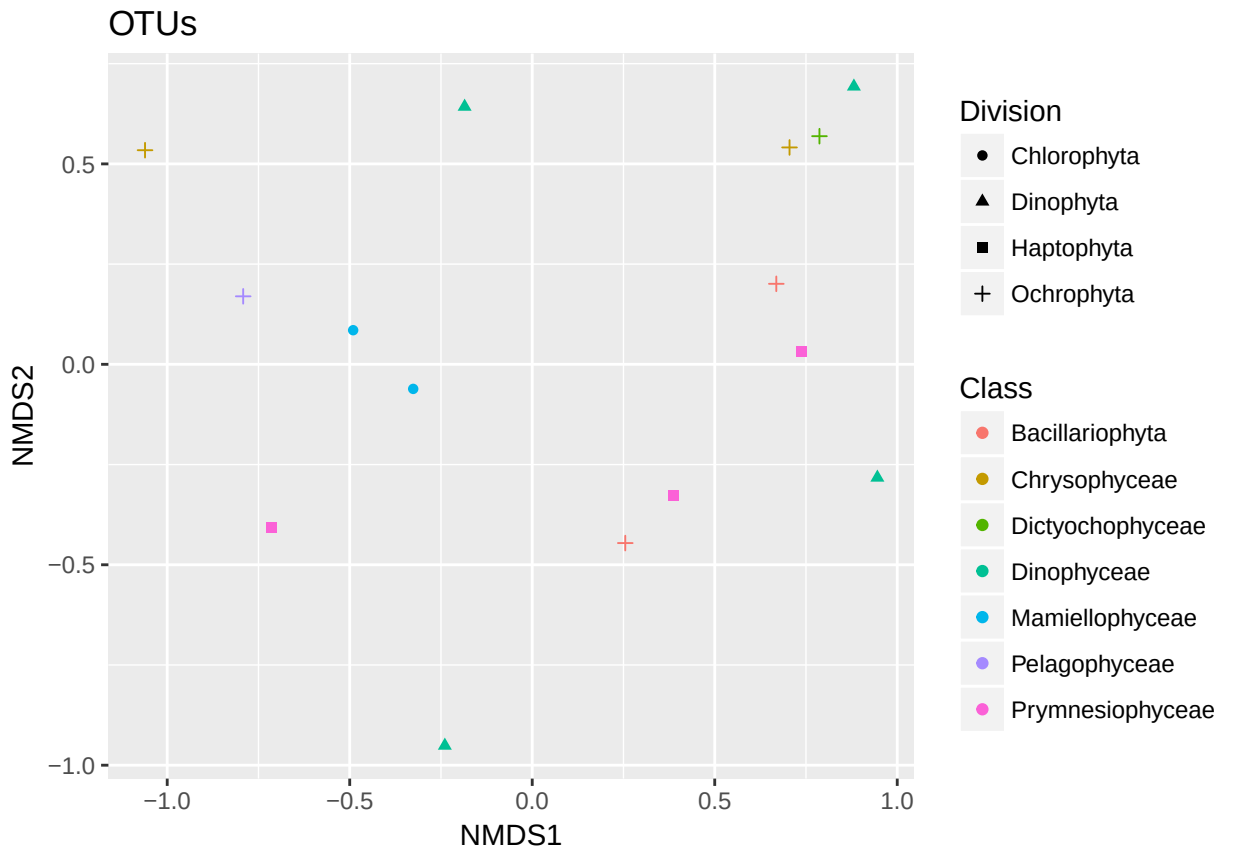
```
carbom.ord <- ordinate(carbom, "NMDS", "bray")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.1069238
## Run 1 stress 0.1024627
## ... New best solution
## ... Procrustes: rmse 0.2095831  max resid 0.2941203
## Run 2 stress 0.1069239
## Run 3 stress 0.1024627
## ... Procrustes: rmse 7.66764e-05  max resid 0.0001460526
## ... Similar to previous best
## Run 4 stress 0.1860346
## Run 5 stress 0.1024627
## ... Procrustes: rmse 7.337751e-05  max resid 0.0001551218
## ... Similar to previous best
## Run 6 stress 0.1480373
## Run 7 stress 0.1175085
## Run 8 stress 0.1024627
## ... New best solution
## ... Procrustes: rmse 1.574708e-05  max resid 2.874615e-05
## ... Similar to previous best
```

```
## Run 9 stress 0.1069238
## Run 10 stress 0.1069239
## Run 11 stress 0.1480373
## Run 12 stress 0.1024627
## ... Procrustes: rmse 7.820401e-05  max resid 0.0001552184
## ... Similar to previous best
## Run 13 stress 0.1024627
## ... Procrustes: rmse 2.478219e-05  max resid 4.497876e-05
## ... Similar to previous best
## Run 14 stress 0.1069238
## Run 15 stress 0.1024627
## ... New best solution
## ... Procrustes: rmse 1.172297e-05  max resid 2.149502e-05
## ... Similar to previous best
## Run 16 stress 0.1024627
## ... Procrustes: rmse 7.333179e-05  max resid 0.0001506098
## ... Similar to previous best
## Run 17 stress 0.1626382
## Run 18 stress 0.2652792
## Run 19 stress 0.1069238
## Run 20 stress 0.1069238
## *** Solution reached
```

PLot **OTUs**

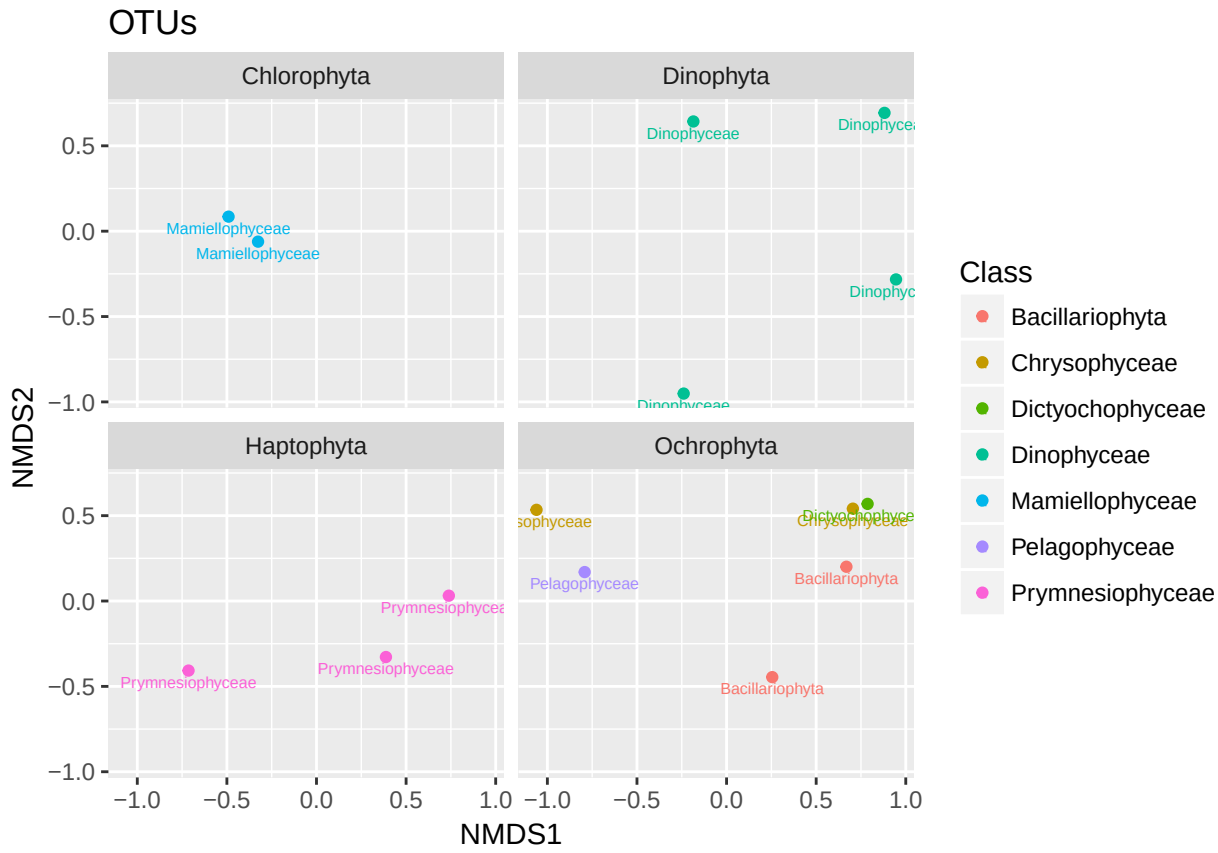
```
plot_ordination(carbom, carbom.ord, type="taxa", color="Class", shape= "Division",
                title="OTUs")
```



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

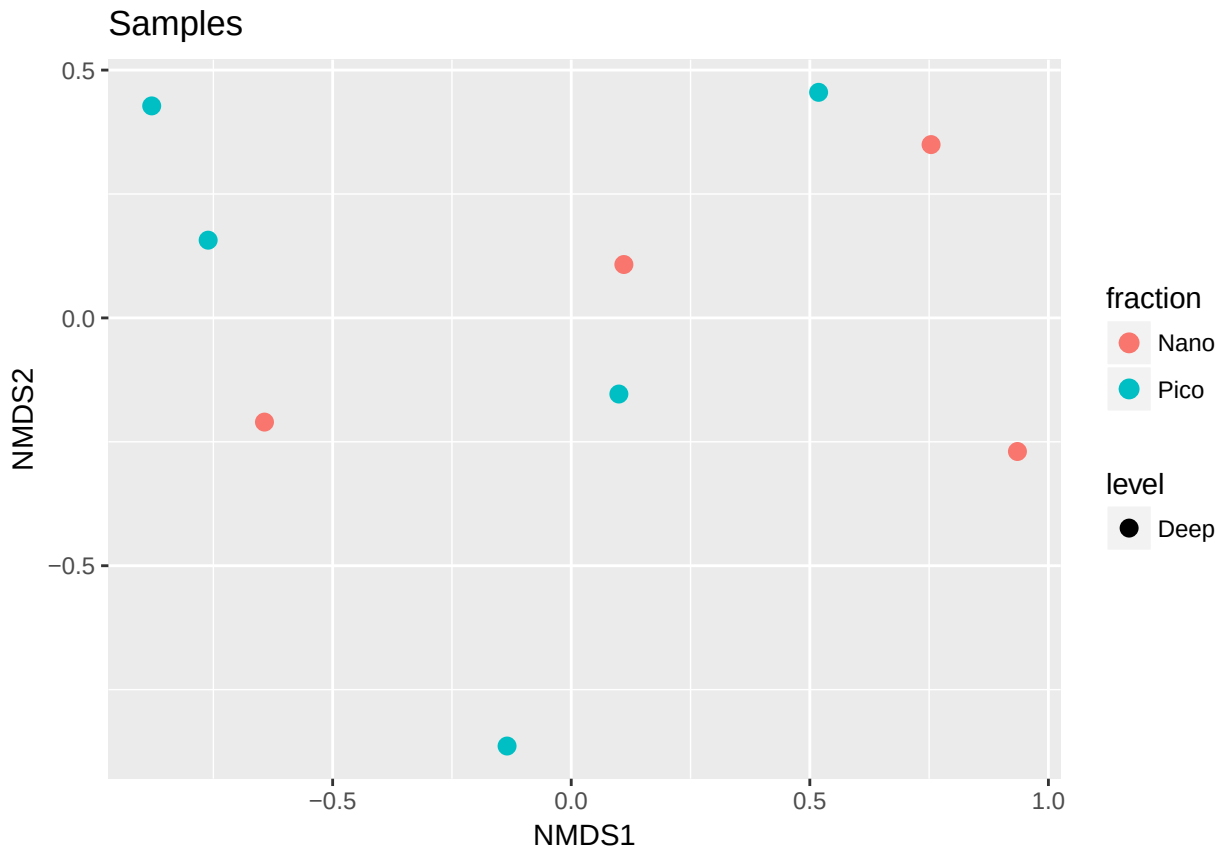
```
plot_ordination(carbom, carbom.ord, type="taxa", color="Class",
                title="OTUs", label="Class") +
facet_wrap(~Division, 3)
```

```
## Warning: Ignoring unknown aesthetics: na.rm
```



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

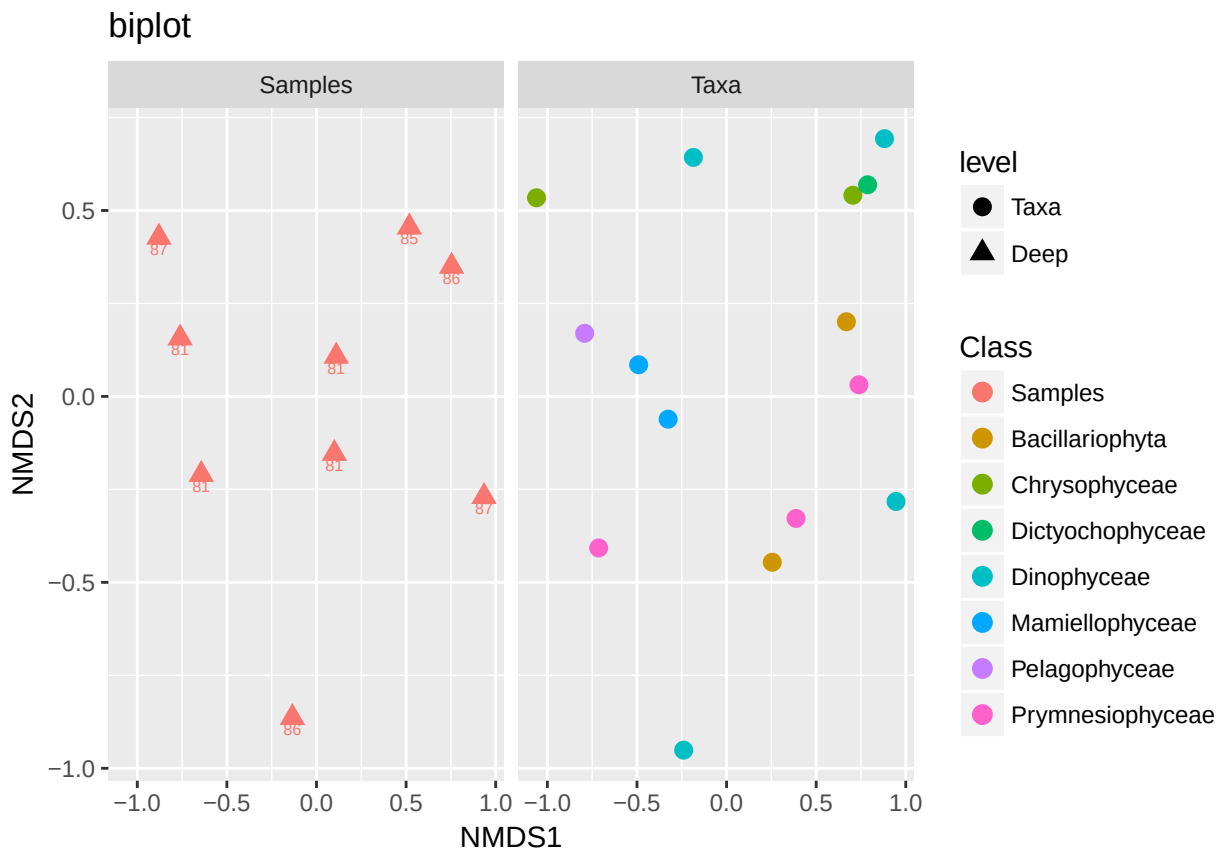
```
plot_ordination(carbom, carbom.ord, type="samples", color="fraction",
                shape="level", title="Samples") + geom_point(size=3)
```



Diplay both samples and OTUs but in 2 different panels.

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

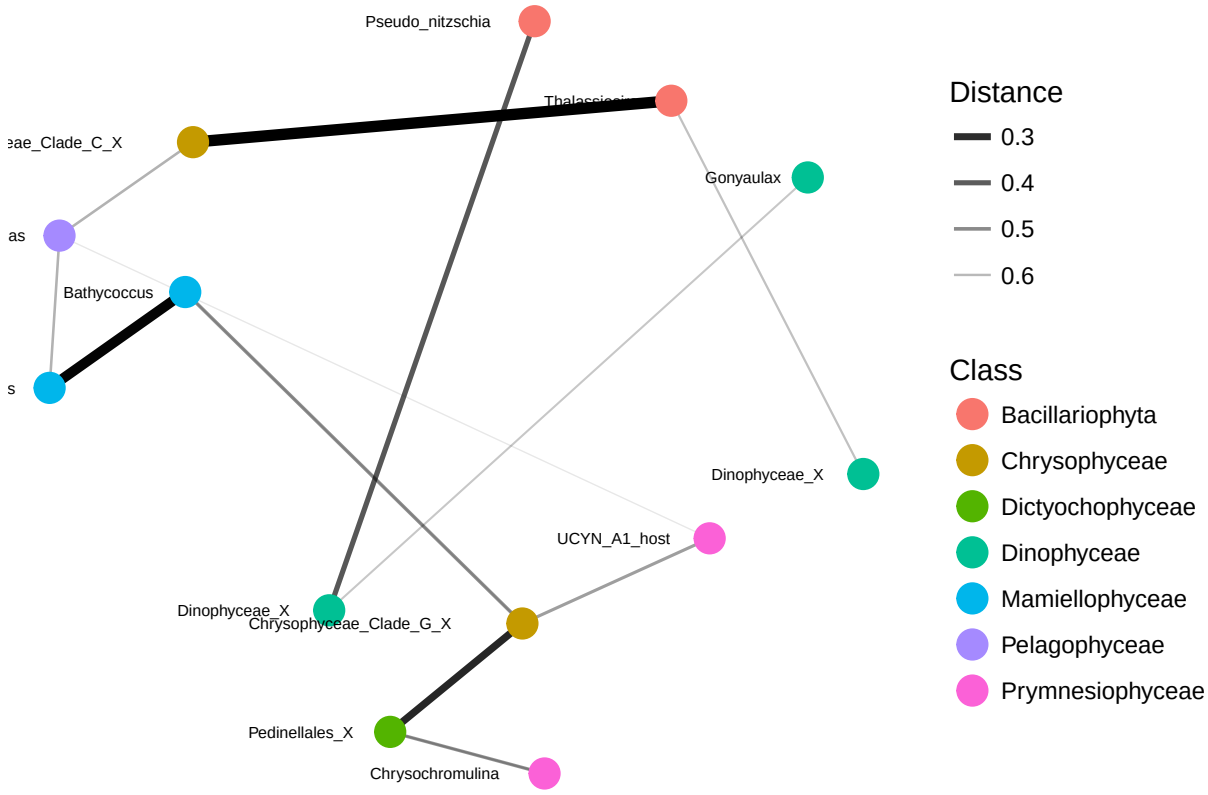
```
## Warning: Ignoring unknown aesthetics: na.rm
```



Network analysis

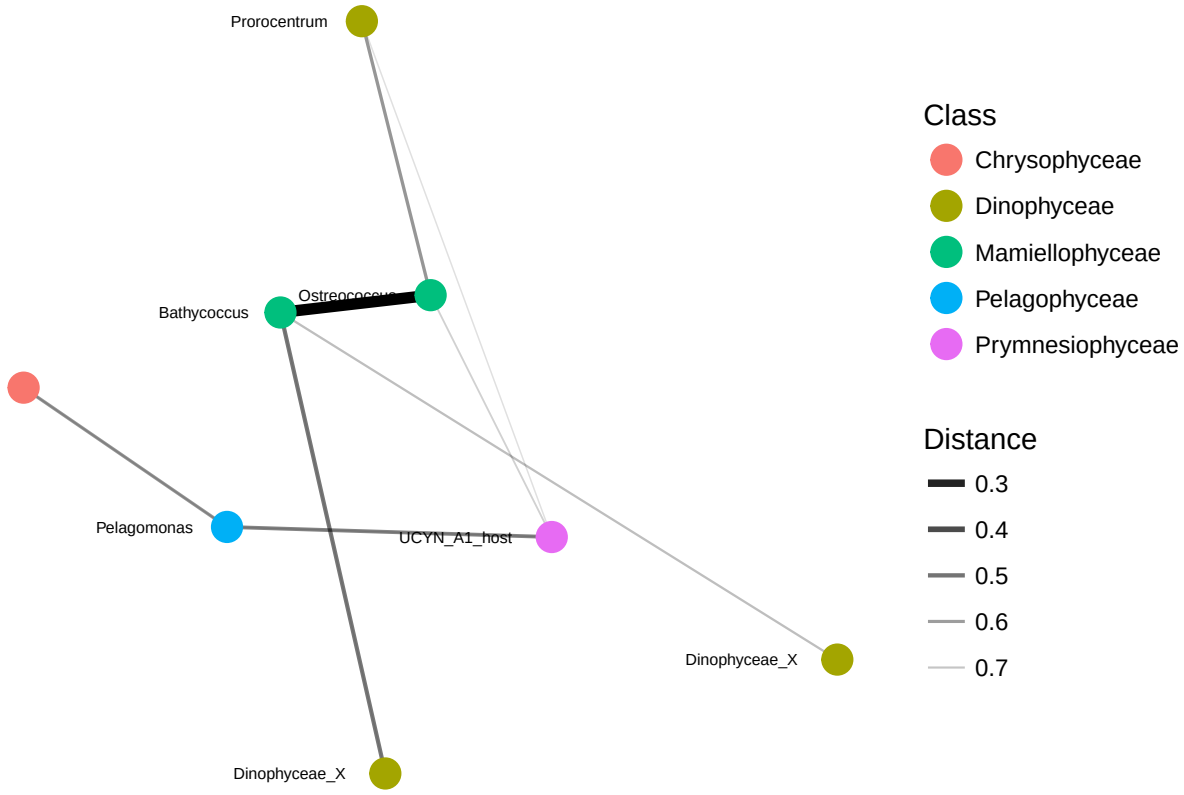
Simple network analysis

```
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")
```


Prorocentrum
Bathycoccus
Ostreococcus
Pelagomonas
UCYN_A1_host
Dinophyceae_X
Dinophyceae_X
Class
Chrysophyceae
Dinophyceae
Mamiellophyceae
Pelagophyceae
Prymnesiophyceae
Distance
0.3
0.4
0.5
0.6
0.7