

# [Title]

Module 3: Project 1 by Team 5

*Karl Abuan*

*May Ho*

*Jonah Lin*

*Leilynaz Malekafzali*

*Tiffany Yang*

*Abdur Rahman M. A. Basher*

*March 11, 2018*

## Abstract

This is the abstract. It consists of two paragraphs.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Problem Formulation</b>	<b>2</b>
<b>3</b>	<b>Materials and Experimental Configuration</b>	<b>2</b>
3.1	Experimental Protocols . . . . .	2
3.2	Dataset . . . . .	2
3.3	Parameters Configuration . . . . .	2
3.4	Data Preprocessing . . . . .	2
3.4.1	Exploratory Data Analysis . . . . .	5
<b>4</b>	<b>Results</b>	<b>11</b>
4.1	Analysis of microbial community structure along with depth and oxygen concentration	11
4.2	Analysis of abundance information of [OTU****] along with depth and/or oxygen concentration . . . . .	11
4.3	Estimate richness (number of OTUs/ASVs) for [OTU****] . . . . .	11
4.4	Interpretation of abundance information of OTUs/ASVs of [OTU****] along with depth and/or oxygen concentration . . . . .	11
<b>5</b>	<b>Discussion</b>	<b>11</b>
	<b>References</b>	<b>12</b>

# 1 Introduction

## 2 Problem Formulation

## 3 Materials and Experimental Configuration

### 3.1 Experimental Protocols

Here . . . .:

**P1.** Analysis of microbial community structure along with depth and oxygen concentration (see Section 4.1).

**P2.** Analysis of abundance information of [OTU\*\*\*\*] along with depth and/or oxygen concentration (see Section 4.2).

**P3.** Estimate richness (number of OTUs/ASVs) for [OTU\*\*\*\*] (see Section 4.3).

**P4.** Interpretation of abundance information of OTUs/ASVs of [OTU\*\*\*\*] along with depth and/or oxygen concentration (see Section 4.4).

### 3.2 Dataset

### 3.3 Parameters Configuration

```
## try http:// if https:// URLs are not supported
source("https://bioconductor.org/biocLite.R")
biocLite("phyloseq")
biocLite("metagenomeSeq")
library("tidyverse")
library("ggplot2")
```

### 3.4 Data Preprocessing

Let us analyze the data in depth to summarize their main characteristics and plot some interesting patterns.

```
load("data/mothur_phyloseq.RData")
```

First, we need to understand the properties of the mothur object by peaking into this object.

```
mothur
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:             [ 4368 taxa and 7 samples ]
## sample_data() Sample Data:          [ 7 samples by 22 sample variables ]
## tax_table()   Taxonomy Table:        [ 4368 taxa by 7 taxonomic ranks ]
```

The phyloseq object can be accessed using special accessor functions, or **accessors**, which return specific information about phylogenetic sequencing data. For instance, we can get variety of information regarding the number of taxa, number of samples, covariates, taxa names, taxa rankings, OTU table structure representing OTU counts, with rows corresponding to samples and columns to OTUs, and phylogenetic tree for OTU.

```
ntaxa(mothur)
```

```
## [1] 4368
```

```
nsamples(mothur)
```

```
## [1] 7
```

```
sample_names(mothur)
```

```
## [1] "Saanich_010" "Saanich_100" "Saanich_120" "Saanich_135" "Saanich_150"
## [6] "Saanich_165" "Saanich_200"
```

```
sample_variables(mothur)
```

```
## [1] "Depth_m"          "O2_uM"            "PO4_uM"
## [4] "SiO2_uM"          "NO3_uM"            "NH4_uM"
## [7] "Std_NH4_uM"        "NO2_uM"            "Std_NO2_uM"
## [10] "H2S_uM"           "Std_H2S_uM"        "Cells.ml"
## [13] "N2O_nM"           "Std_N2O_nM"        "CH4_nM"
## [16] "Std_CH4_nM"        "Temperature_C"     "Conductivity_mScm_1"
## [19] "Fluorescence_mgm_3" "OxygenSBE_V"       "Salinity_PSU"
## [22] "Density_q"
```

```
head(taxa_names(mothur))
```

```
## [1] "Otu0001" "Otu0002" "Otu0003" "Otu0004" "Otu0005" "Otu0006"
```

```
rank_names(mothur)
```

```
## [1] "Domain" "Phylum" "Class" "Order" "Family" "Genus" "Species"
```

```
cat("The dimension of the otu_table of the mothur object: ", dim(otu_table(mothur)))
```

```
## The dimension of the otu_table of the mothur object: 7 4368
```

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

```
## OTU Table:          [5 taxa and 7 samples]
##                      taxa are columns
##                      Otu0001 Otu0002 Otu0003 Otu0004 Otu0005
## Saanich_010         462         0    5317     41     169
## Saanich_100        11444         0   32026    1367    4884
## Saanich_120        40906         0   13932    5606    8947
## Saanich_135         52809         4    3764    7235   11042
## Saanich_150         83079        12     860   10392    9431
## Saanich_165         95560         9     342   20491    6933
## Saanich_200        15262       77958     92     530     141
```

```
print("First few rows of the tax_table of the mothur object: ")
```

```
## [1] "First few rows of the tax_table of the mothur object: "
```

```
head(tax_table(mothur))
```

```
## Taxonomy Table:      [6 taxa by 7 taxonomic ranks]:
##      Domain      Phylum      Class
## Otu0001 "Bacteria" "Proteobacteria" "Gammaproteobacteria"
## Otu0002 "Bacteria" "Proteobacteria" "Epsilonproteobacteria"
## Otu0003 "Bacteria" "Proteobacteria" "Gammaproteobacteria"
## Otu0004 "Bacteria" "Proteobacteria" "Gammaproteobacteria"
## Otu0005 "Bacteria" "Proteobacteria" "Deltaproteobacteria"
## Otu0006 "Bacteria" "Proteobacteria" "Gammaproteobacteria"
##      Order      Family
## Otu0001 "Oceanospirillales" "SUP05_cluster"
## Otu0002 "Campylobacteriales" "Campylobacteraceae"
## Otu0003 "Oceanospirillales" "JL-ETNP-Y6"
## Otu0004 "Chromatiales" "Ectothiorhodospiraceae"
## Otu0005 "SAR324_clade(Marine_group_B)" "SAR324_clade(Marine_group_B)_fa"
## Otu0006 "Oceanospirillales" "SUP05_cluster"
##      Genus      Species
## Otu0001 "SUP05_cluster_ge" ""
## Otu0002 "Arcobacter" ""
## Otu0003 "JL-ETNP-Y6_ge" ""
## Otu0004 "Thiorhodospira" ""
## Otu0005 "SAR324_clade(Marine_group_B)_ge" ""
## Otu0006 "SUP05_cluster_ge" ""
```

Next we perform a series of filterings according to three rules:

i)- exclude OTUs that are not observed for more than 4 samples:

```
abundantTaxa = filter_taxa(mothur, function(x) sum(x == 0) <= 4, TRUE)
cat("The number of taxa after applying filter 1: ", ntaxa(abundantTaxa))
```

```
## The number of taxa after applying filter 1: 873
```

ii)- prune samples and OTUs with unknown values, such as unclassified value:

```
basedOnGenus <- as.data.frame(tax_table(abundantTaxa)) %>%
  filter(!str_detect(Genus, 'uncultured'), !str_detect(Genus, 'unclassified')) %>%
  select(Genus)
knownTaxa = subset_taxa(abundantTaxa, Genus %in% basedOnGenus$Genus)
cat("The number of taxa after applying filter 2: ", ntaxa(knownTaxa))
```

```
## The number of taxa after applying filter 2: 485
```

iii)- any phylum fail to have more than 5 OTUs should be trimmed:

```
basedOnphylums <- as.data.frame(tax_table(knownTaxa)) %>%
  group_by(Phylum) %>%
```

```

count() %>%
filter( n > 5)

## In contrary we can run the following:
# workingTaxa <- prune_taxa(taxa_sums(knownTaxa) > 5, knownTaxa)

workingTaxa = subset_taxa(knownTaxa, Phylum %in% basedOnphylums$Phylum)
cat("The number of taxa after applying filter 3: ", ntaxa(workingTaxa))

## The number of taxa after applying filter 3: 457

```

### 3.4.1 Exploratory Data Analysis

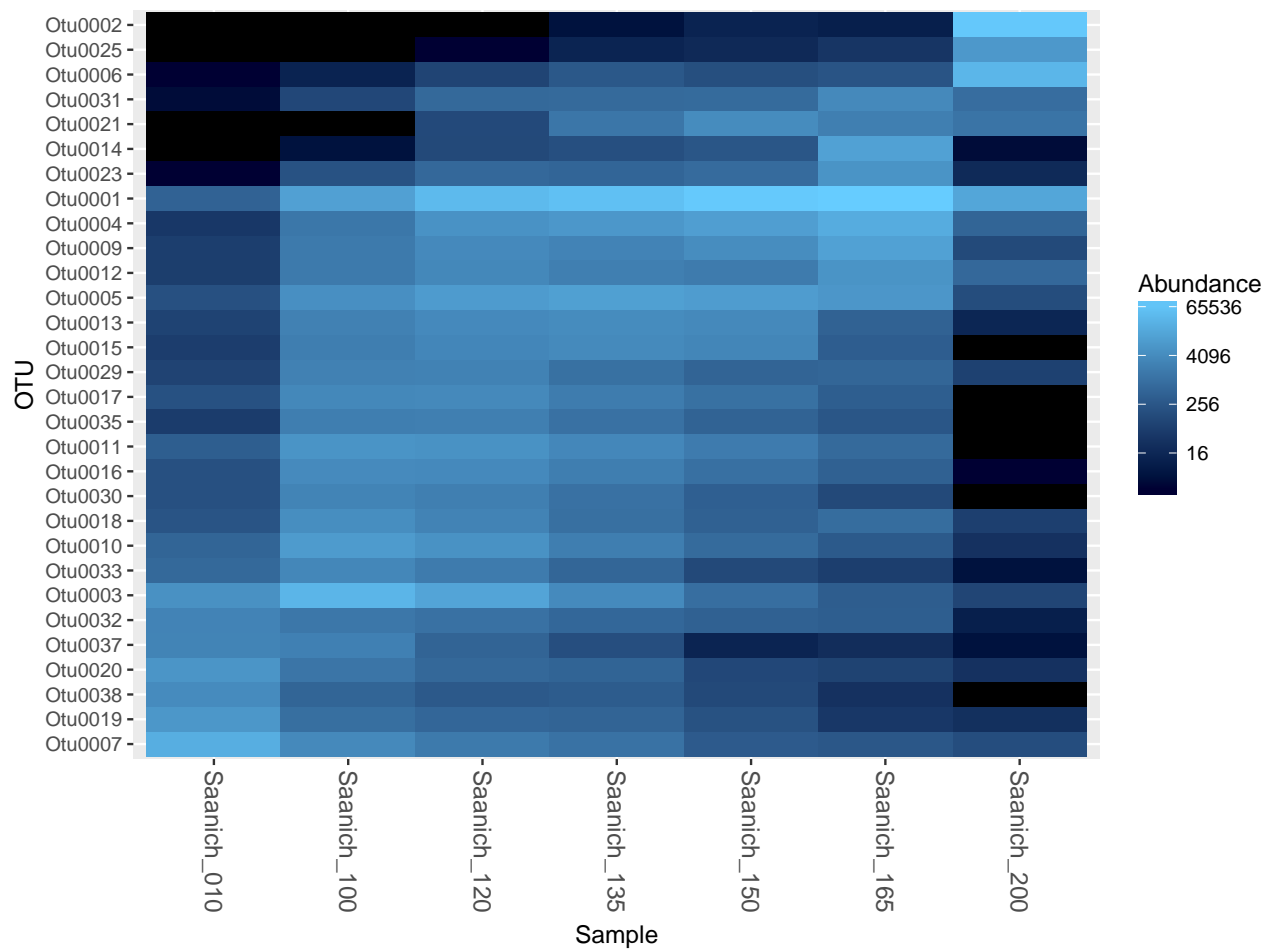
Before proceeding with the computations for differential expression across samples, it is possible to produce a plot showing the sample relations based on multidimensional scaling. The basic premise is that we make a plot so samples which are similar are near to each other in the plot while samples that are dissimilar are far from each other. Here is an example.

Plot a heatmap to peak the abundances of 30 taxa across samples:

```

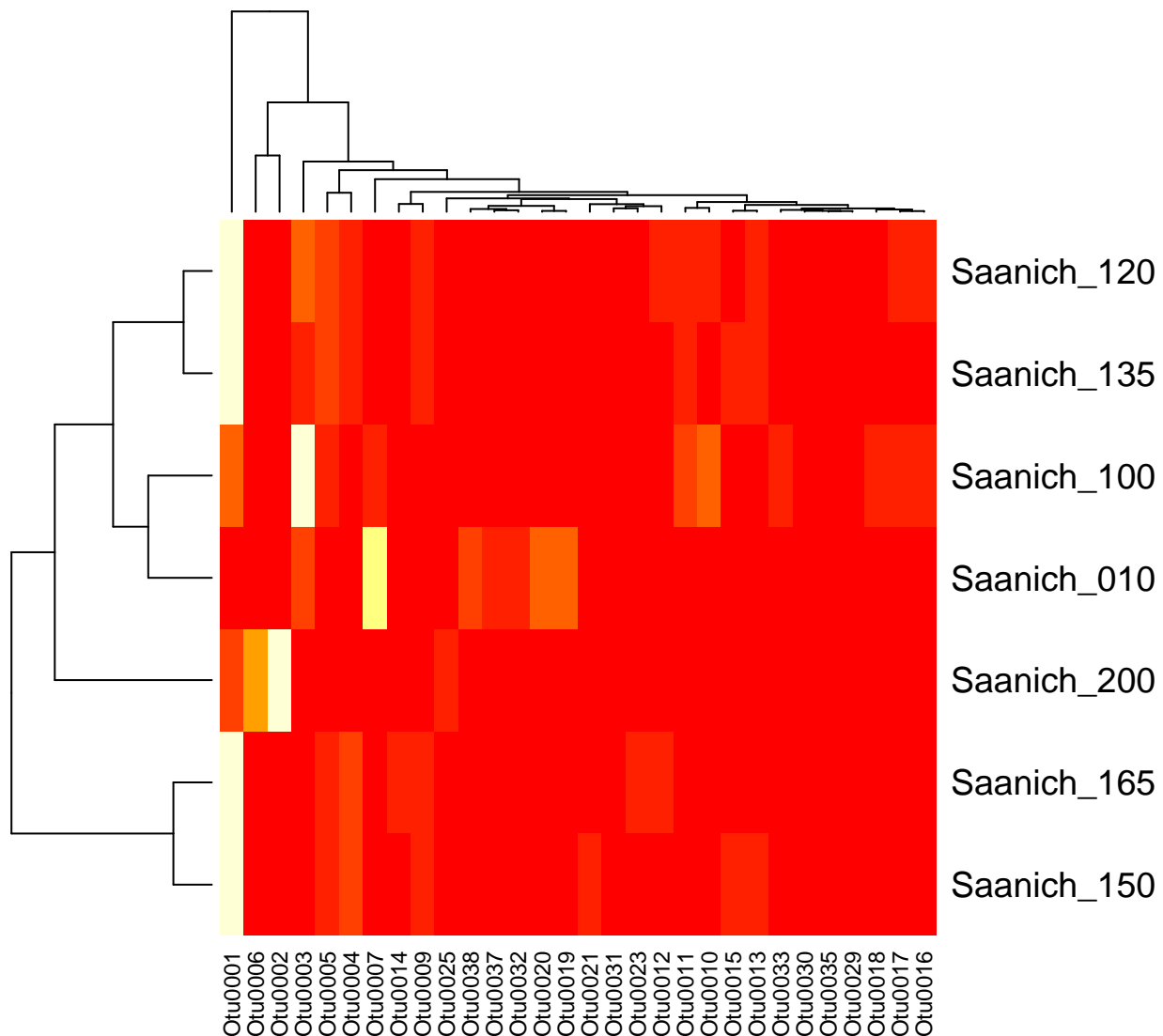
tmpTaxa = names(sort(taxa_sums(workingTaxa), decreasing = TRUE)[1:30])
prunedTaxa = prune_taxa(tmpTaxa, workingTaxa)
plot_heatmap(prunedTaxa, "PCoA", distance = "bray")

```



To observe similarities between samples and OTUs, we apply the following:

```
heatmap(otu_table(prunedTaxa))
```



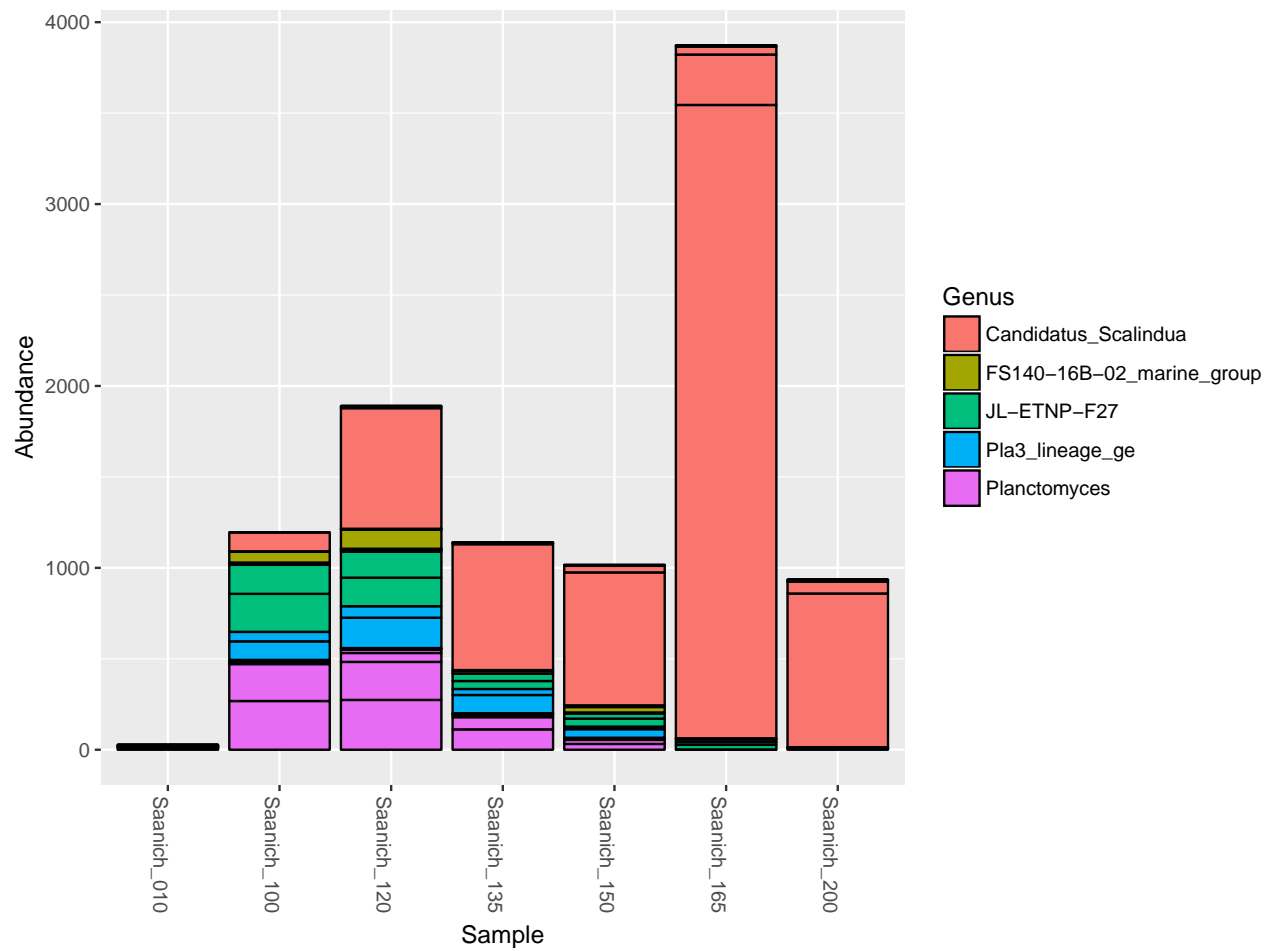
The following function will assist us to understand the unique Phylum rank:

```
get_taxa_unique(physeq = workingTaxa, taxonomic.rank = "Phylum")
```

```
## [1] "Proteobacteria"          "Bacteroidetes"
## [3] "Thaumarchaeota"         "Actinobacteria"
## [5] "Marinimicrobia_(SAR406_clade)" "Planctomycetes"
## [7] "Gemmatimonadetes"       "Verrucomicrobia"
## [9] "Chloroflexi"
```

We choose the *Planctomycetes* phylum and we explore the distribution of genera of this phylum.

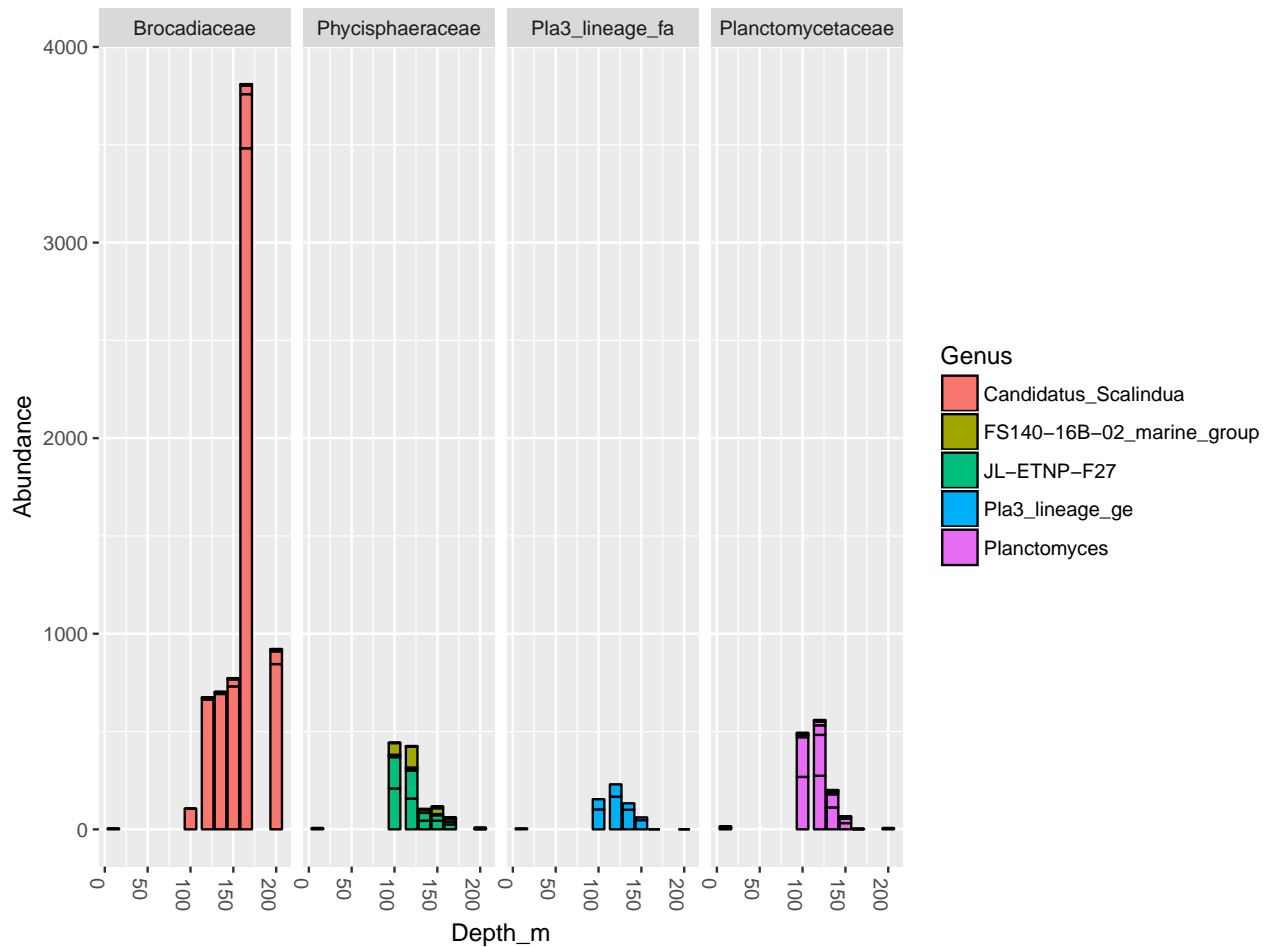
```
subTaxa = subset_taxa(workingTaxa, Phylum == "Planctomycetes")
plot_bar(subTaxa, fill="Genus")
```



We explore how the genus of this phylum varies across samples (representing depth of the ocean) based on genus abundances.

```
plot_bar(subTaxa, "Depth_m", fill="Genus", facet_grid=~Family)
```





The genus was chosen to be *Planctomyces* and we identify the following OTUs.

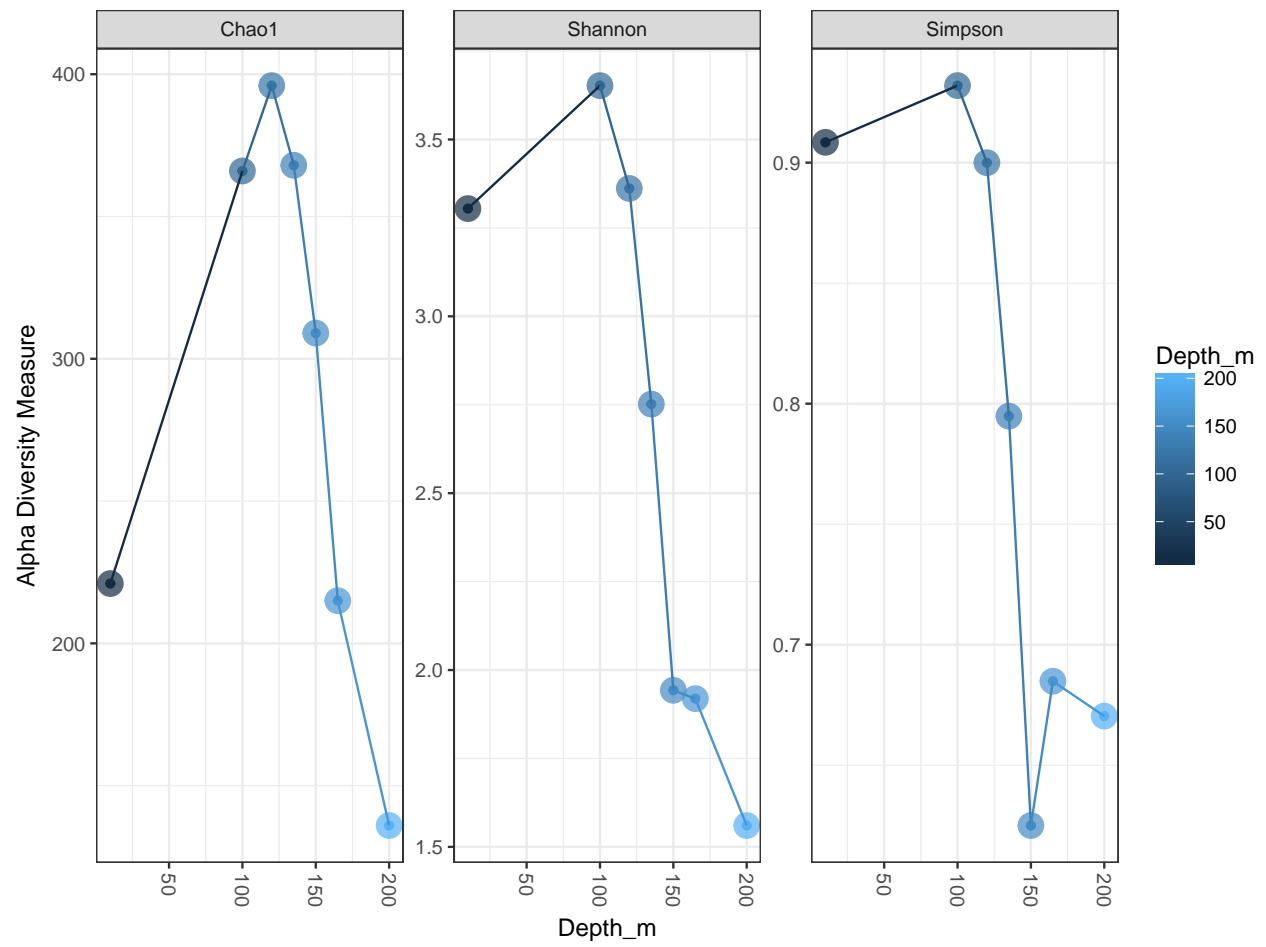
```
subGenTaxa = subset_taxa(workingTaxa, Genus == "Planctomyces")
colnames(otu_table(subGenTaxa))
```

```
## [1] "0tu0125" "0tu0144" "0tu0401" "0tu0592" "0tu1038" "0tu1262"
```

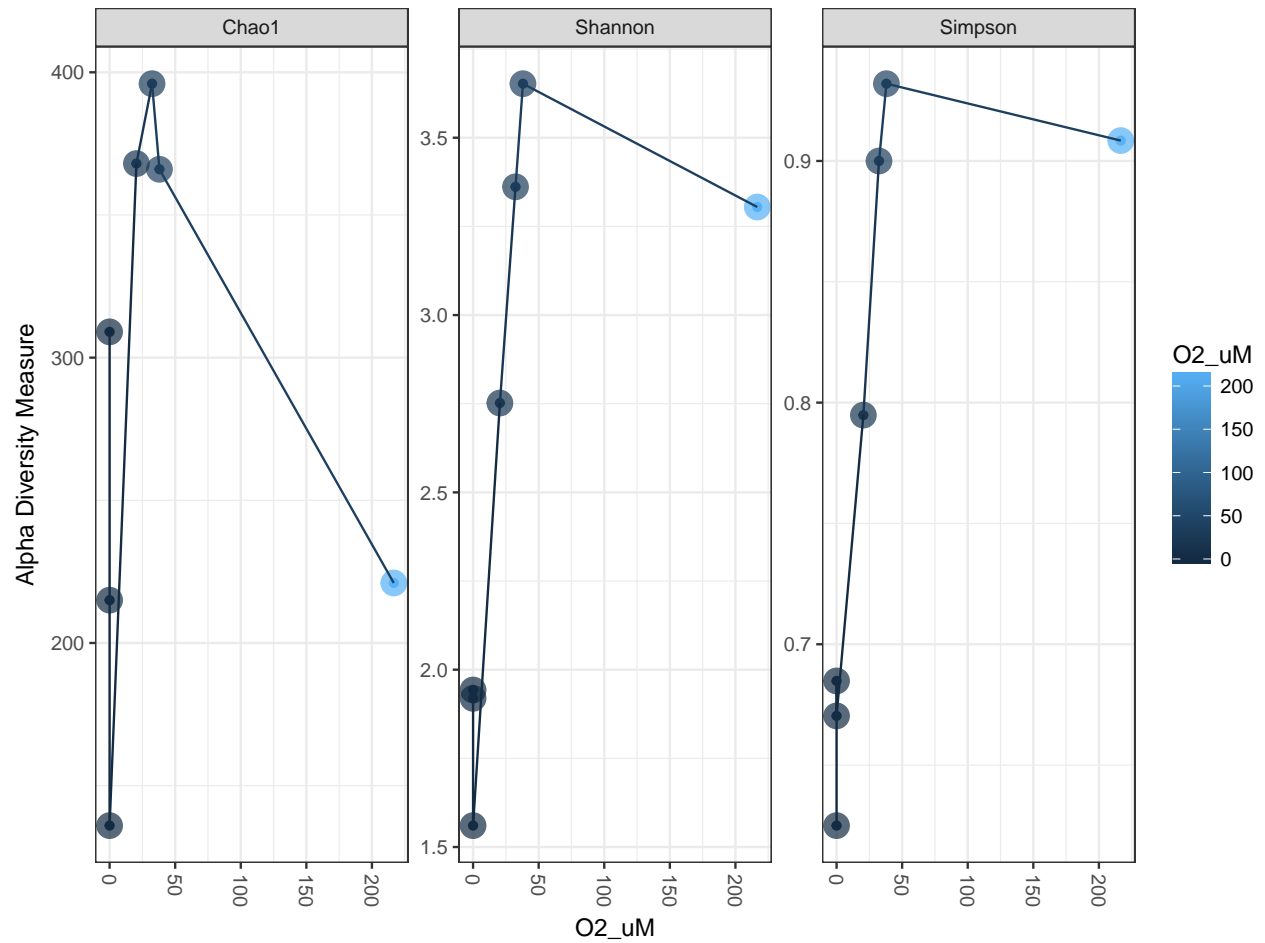
Estimate diversity of *Planctomyces* across depth and oxygen concentration.

```
theme_set(theme_bw())
pal = "Set1"
scale_colour_discrete <- function(palname=pal, ...){
  scale_colour_brewer(palette=palname, ...)
}
scale_fill_discrete <- function(palname=pal, ...){
  scale_fill_brewer(palette=palname, ...)
}
```

```
plot_richness(workingTaxa, x="Depth_m", color="Depth_m", measures = c("Chao1", "Shannon", "Simpson"))
```



```
plot_richness(workingTaxa, x="O2_uM", color="O2_uM", measures = c("Chao1", "Shannon", "Simpson"))
```



## 4 Results

- 4.1 Analysis of microbial community structure along with depth and oxygen concentration
- 4.2 Analysis of abundance information of [OTU\*\*\*\*] along with depth and/or oxygen concentration
- 4.3 Estimate richness (number of OTUs/ASVs) for [OTU\*\*\*\*]
- 4.4 Interpretation of abundance information of OTUs/ASVs of [OTU\*\*\*\*] along with depth and/or oxygen concentration

## 5 Discussion

(Hawley et al. 2017; Torres-Beltrán et al. 2017)

## References

Hawley, Alyse K, Mónica Torres-Beltrán, Elena Zaikova, David A Walsh, Andreas Mueller, Melanie Scofield, Sam Kheirandish, et al. 2017. “A Compendium of Multi-Omic Sequence Information from the Saanich Inlet Water Column.” *Scientific Data* 4. Nature Publishing Group: 170160.

Torres-Beltrán, Mónica, Alyse K Hawley, David Capelle, Elena Zaikova, David A Walsh, Andreas Mueller, Melanie Scofield, et al. 2017. “A Compendium of Geochemical Information from the Saanich Inlet Water Column.” *Scientific Data* 4. Nature Publishing Group: 170159.