Module 3: Project 1 by Team 5

Mothur vs. QIIME2 Microbiome Data Analysis

Karl Abuan
May Ho
Jonah Lin
Leilynaz Malekafzali
Tiffany Yang
Abdur Rahman M. A. Basher

April 23, 2018

Abstract

Saanich Inlet is an intermittently anoxic fjord on the coast of Vancouver Island, and it is commonly used as a model system for analyzing microbial community response to ocean deoxygenation. A portion of 16S sequence data from Saanich Inlet were collected at various depths and were analyzed to estimate diversity and examine microbial abundance with respect to varying depths and oxygen levels. The analysis of pre-processed OTUs and ASVs - generated through Mothur and QIIME2 pipeline tools, respectively - showed that the community alpha-diversity values peaked at depths 10m and 100m and a minimum value was achieved at 200m. Further analysis of taxonomic levels showed that the phylum Proteobacteria was found to be the most abundant phylum among all samples. We used the genus *Planctomyces* to further examine how microbial communities change across various depths and the oxygen gradient in Saanich Inlet. Analysis revealed that abundance of *Planctomyces* had no relationship with varying depths nor oxygen concentration in datasets derived from both Mothur or QIIME2. The analysis of data using QIIME2 revealed four genera within the phylum Planctomycetes (*Candidatus_Scalindua*, *D_5_JL_ETNP_F27*, *FS140_16B_02_marine_group*, *D_5_Planctomyces*), while Mothur identified one extra genus in addition to those identified by QIIME2: *Pla3_lineage_ge*. The abundances of the OTUs and ASVs within the *Planctomyces* genus did not change significantly differ across depth and oxygen concentration.

Contents

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

---

1 Introduction

Saanich Inlet is a seasonally anoxic fjord located on the southeast coast of Vancouver Island, British Columbia, Canada [1], and it is commonly used as a model system for analyzing microbial community responses to ocean deoxygenation [2]. Established by the shallow sill opening to the Strait of Georgia, circulation of basin water in Saanich Inlet is reduced. Together with microbial respiration of organic matter from surface waters and increased stratification due to strong temperature gradients between the upper and bottom waters [3], oxygen levels down the inlet water column decrease. These oxygen-deprived regions, termed oxygen minimum zones [4], host a diverse microbial community that mediate important biogeochemical processes.

In Saanich Inlet, as the water column oxygen levels diminish, changes in nutrient and oxygen flow occur within its ecosystem. Microbes turn to use alternative terminal electron acceptors such as nitrate, followed by sulphate [4]. During the fall, tide and ocean current changes allow cold, nutrient-rich oxygenated water to creep inside and re-oxygenate the inlet [5]. This recurring phenomenon makes Saanich Inlet a great tractable model system to study biological processes that take place between oxic- anoxic-, suboxic- and sulfidic water layers.

The aim of this project is to investigate the changes in the microbial community structure found in Saanich Inlet with respect to various depths and decreasing oxygen concentrations. Through Mothur and QIIME2 pipeline tools, the data is processed to obtain operational taxonomic units (OTUs) and amplicon sequence variants (ASVs) derived from sequence reads. Using the 16S rRNA gene as marker, OTUs are constructed and filtered. OTUs are defined as the clusters of reads grouped together differing by less than a fixed, arbitrary sequence dissimilarity threshold, often 3% [6,7]. On the other hand, ASV is a new method developed to analyze a microbial community with finer resolution and it's independent from a reference database [6]. The ASV method can distinguish sequence variants that differ by one nucleotide and infer the biological sequence in the sample before introducing sequencing and amplification errors [6]. Using the OTUs and ASVs data sets, the microbial diversity at various depth and oxygen concentration of the Saanich Inlet samples is assessed with the main focus on the *Planctomyces* genus.

Planctomyces genus is part of the *Planctomycetes* phylum [8]. *Planctomyces* are known for their ubiquity, metabolic diversity, and unique features such as intracellular compartmentalization that are typical of eukaryotes [8]. They are abundant in oceans, freshwater, and soils. They also possess

diverse metabolism – such as aerobic chemoheterotrophs – and autotrophic anaerobes that oxidize ammonia to nitrogen (anammox process) [8]. These organisms facilitate the anammox process, contributing approximately 50% of the atmospheric nitrogen molecules in the global nitrogen cycle [8]. Additionally, the anammox process is important in the nitrogen-rich-wastewater remediation technology [8]. Thus, this report further addresses the change in the abundance of *Planctomyces* with depth and oxygen concentration, the richness (OTUs/ASVs) of *Planctomyces*, and the change in the abundance of OTUs/ASVs within *Planctomyces* genus with depth and oxygen concentration in Saanich Inlet.

2 Methods

For both Mothur and QIIME2 pipeline, sequences from Saanich Inlet were amplified using 515F and 808R primers. Sequences were generated on MiSeq with Phred33 quality scores [1]. For both the Mothur and QIIME2 pipelines, the data were cleaned up by filtering out various sequences (Low quality, chimeric sequences, etc.), aligned and classified with SILVA database, and trimmed by their start and end sites [9,10]. The end results for both pipelines were then included in a phyloseq object which contains the following: OTU/ASV Table, Taxonomy, and Sample Metadata [9,10].

The samples were normalized to 100,000 sequences per sample to facilitate comparisons between samples. The normalized counts were then converted to relative abundance percentages. Next, a series of filtering was applied according to two rules: i) Exclude OTUs that are not observed in more than 4 samples; ii) Prune samples and OTUs with unknown values, such as `unclassified` value. This has resulted in thirdMTaxa and thirdQTaxa taxa from Mothur and QIIME2, respectively. No other pre-processing was applied. The implementations are done entirely using R (v3.4.3) and relied on some efficient third-party libraries, such as `phyloseq`, `tidyverse`, `gridExtra`, and `magrittr` [11–16].

```
load("data/mothur_phyloseq.RData")
load("data/qiime2_phyloseq.RData")
set.seed(4832)
rarefiedM <- rarefy_even_depth(mothur, sample.size = 1e+05)
rarefiedQ <- rarefy_even_depth(qiime2, sample.size = 1e+05)
rarefiedMPer = transform_sample_counts(rarefiedM, function(x) 100 *
    x/sum(x))
rarefiedQPer = transform_sample_counts(rarefiedQ, function(x) 100 *
    x/sum(x))
```

```
# First rule
firstMTaxa <- filter_taxa(rarefiedMPer, function(x) sum(x ==
    0) <= 4, TRUE)
firstQTaxa <- filter_taxa(rarefiedQPer, function(x) sum(x ==
    0) <= 4, TRUE)

# Second rule
basedOnGenus <- as.data.frame(tax_table(firstMTaxa)) %>% filter(!str_detect(Genus,
    "uncultured|unclassified"))
secondMTaxa = subset_taxa(firstMTaxa, Genus %in% basedOnGenus$Genus)
```

```
basedOnGenus <- as.data.frame(tax_table(firstQTaxa)) %>% filter(!str_detect(Genus,
    "uncultured|unclassified|\\bD_5__\\b"))
secondQTaxa <- subset_taxa(firstQTaxa, Genus %in% basedOnGenus$Genus)
```

2.1 Shannon Diversity Index (SDI) and Chao1

We applied *Shannon diversity index (SDI)* to estimate the microbial diversity of Saanich Inlet dataset. It has the following definition:

$SDI = -\sum_i^R p_i \log(p_i)$

where p_i represents the distribution of individuals belonging to the ith species, and R represents the number of distinct species [17]. It can be noted that SDI takes both the richness and abundance information to measure the expected uncertainty about species contained in a sample. The high SDI value suggests that species are evenly distributed while low SDI value implies species are disproportionality situated. SDI value could be zero meaning the sample contains exactly one or no species at all. However, SDI does not directly model the expected richness of a sample and, neither, it represents an accurate estimation of species diversity because the probability distribution of species is not knowable exactly; it is only an estimate from a sample.

In contrast to SDI, *Chao1* could be used to recover approximate true richness:

$Chao1 = S_{obs} + \frac{\alpha}{2\beta}$

where S_{obs} represents the observed richness, α and β indicates the number of different species with exactly one or more than two counts, respectively. The Chao1 method is used to rectify the richness by including the distribution of the rarest species [17].

2.2 General Linear Model

General linear model (LM) [18] is employed to recover interactions between several factors that might be exhibited in Saanich Inlet dataset. In our experiments, we use a single regression model that relates a dependent variable y (abundance) to a single quantitative independent variable x_1 (depth or oxygen), and it has the following form:

$y = \theta_0 + \theta_1 x_1 + \epsilon$

The parameter θ_0 is the y-intercept, which represents the expected value of y when x_1 is zero. The parameter θ_1 is the slope of the regression line, and it represents the expected change (positive or negative) in y (abundance) for a unit increase in x_1 (depth or oxygen). θ_1 could be 0 indicating no effective change with x_1. And, ϵ is the error term and is usually set to 0.

All the parameters could be estimated using ordinary least squares. However, to test the significance θ_1, the following hypothesis testing was formulated: i) The null hypothesis $H_0 : \theta_1 > \gamma$, which asserts that no additional predictive value over and above, contributed by θ_1 and the γ is an arbitrary cutoff probability from t-student distribution ; ii) The alternative hypothesis $H_1 : \theta_1 \leq \gamma$ measures whether x_j has additional predictive strengths.

If the weight of θ_1, referred to as the level of significance (or p-value) and defines as a probability, is below or equal to γ then H_1 was accepted; otherwise, accept H_0.

3 Results

3.1 Analysis of microbial community structure along with depth and oxygen concentration

Using Shannon's diversity index (SDI), which considered both the species abundances and the total number of distinct species in its diversity estimation, an attempt was made to understand the compositional complexity of a microbial community across samples from Saanich Inlet. Fig. 1(a) and 2(a) depicts the change in SDI across depth for Mothur and QIIME2 datasets. It was observed that SDI values peak at depths 10m and 100m before monotonically decreasing and reaching the minimum value at 200m. The SDI values were maximal when the microbes were evenly distributed. The results indicated in Fig. 3 supports this claim since an uneven distribution of phylum was observed at 200m more than at 10m or 100m depths. Analyses showed that oxygen level was slowly decreasing with increased depth of Saanich Inlet as shown in Fig. 1(d). The SDI was higher in the oxic part of the ocean and was much lower in the anoxic part of Saanich Inlet as presented in Fig. 1 (b) and 2(b). Therefore, alpha diversity was higher at high oxygen concentration levels at low depth and lower at low oxygen concentration at high depth.

```
rarefiedMRich <- estimate_richness(rarefiedM, measures = "Shannon")
rarefiedMRichAlpha <- full_join(rownames_to_column(rarefiedMRich),
    rownames_to_column(data.frame(sample_data(rarefiedMPer))),
    by = "rowname")

p1 <- rarefiedMRichAlpha %>% ggplot() + geom_point(aes(x = Depth_m,
    y = Shannon), size = 4, alpha = 0.7) + geom_smooth(method = "loess",
    aes(x = as.numeric(Depth_m), y = Shannon)) + labs(title = "Alpha-diversity across depth",
    y = "Shannon's diversity index", x = "Depth (m)")

p2 <- rarefiedMRichAlpha %>% ggplot() + geom_point(aes(x = O2_uM,
    y = Shannon), size = 4, alpha = 0.7) + geom_smooth(method = "loess",
    aes(x = as.numeric(O2_uM), y = Shannon)) + labs(title = "Alpha-diversity across oxygen",
    y = "Shannon's diversity index", x = "Oxygen (uM)")

p3 <- rarefiedMRichAlpha %>% mutate(O2_group = ifelse(O2_uM ==
    0, "anoxic", "oxic")) %>% ggplot() + geom_boxplot(aes(x = O2_group,
    y = Shannon)) + labs(title = "Alpha-diversity across oxygen",
    y = "Shannon's diversity index", x = "Oxygen (uM)")

p4 <- rarefiedMRichAlpha %>% ggplot() + geom_point(aes(x = Depth_m,
    y = O2_uM), size = 4, alpha = 0.7) + geom_smooth(method = "loess",
    aes(x = as.numeric(Depth_m), y = O2_uM)) + labs(title = "Oxygen concentration across depth
    y = "Oxygen  (uM)", x = "Depth (m)")

figureM <- ggarrange(p1, p2, p3, p4, ncol = 2, nrow = 2, labels = c("(a)",
    "(b)", "(c)", "(d)"))
annotate_figure(figureM, bottom = text_grob("Figure 1: Mothur",
    face = "bold", size = 12))
```

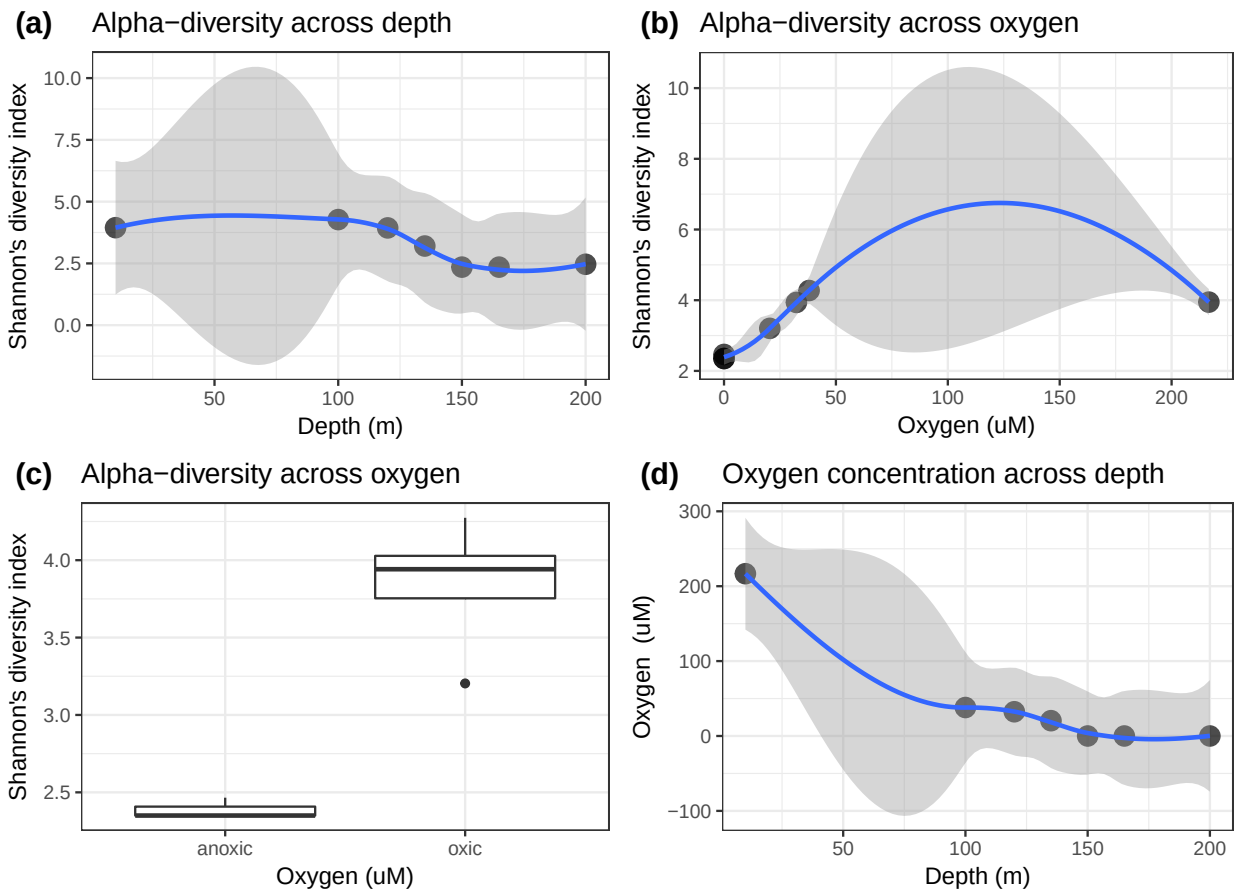

Figure 1: Mothur

```
rarefiedQRich <- estimate_richness(rarefiedQ, measures = "Shannon")
rarefiedQRichAlpha <- full_join(rownames_to_column(rarefiedQRich),
    rownames_to_column(data.frame(sample_data(rarefiedQPer))),
    by = "rowname")
p1 <- rarefiedQRichAlpha %>% ggplot() + geom_point(aes(x = Depth_m,
    y = Shannon), size = 4, alpha = 0.7) + geom_smooth(method = "loess",
    aes(x = as.numeric(Depth_m), y = Shannon)) + labs(title = "Alpha-diversity across depth",
    y = "Shannon's diversity index", x = "Depth (m)")
p2 <- rarefiedQRichAlpha %>% ggplot() + geom_point(aes(x = O2_uM,
    y = Shannon), size = 4, alpha = 0.7) + geom_smooth(method = "loess",
    aes(x = as.numeric(O2_uM), y = Shannon)) + labs(title = "Alpha-diversity across oxygen",
    y = "Shannon's diversity index", x = "Oxygen (uM)")
p3 <- rarefiedQRichAlpha %>% mutate(O2_group = ifelse(O2_uM ==
    0, "anoxic", "oxic")) %>% ggplot() + geom_boxplot(aes(x = O2_group,
    y = Shannon)) + labs(title = "Alpha-diversity across oxygen",
    y = "Shannon's diversity index", x = "Oxygen (uM)")
p4 <- rarefiedQRichAlpha %>% ggplot() + geom_point(aes(x = Depth_m,
    y = O2_uM), size = 4, alpha = 0.7) + geom_smooth(method = "loess",
    aes(x = as.numeric(Depth_m), y = O2_uM)) + labs(title = "Oxygen concentration across depth
    y = "Oxygen  (uM)", x = "Depth (m)")
```

```
figureQ <- ggarrange(p1, p2, p3, p4, ncol = 2, nrow = 2, labels = c("(a)",
    "(b)", "(c)", "(d)"))
annotate_figure(figureQ, bottom = text_grob("Figure 2: QIIME2",
    face = "bold", size = 12))
```

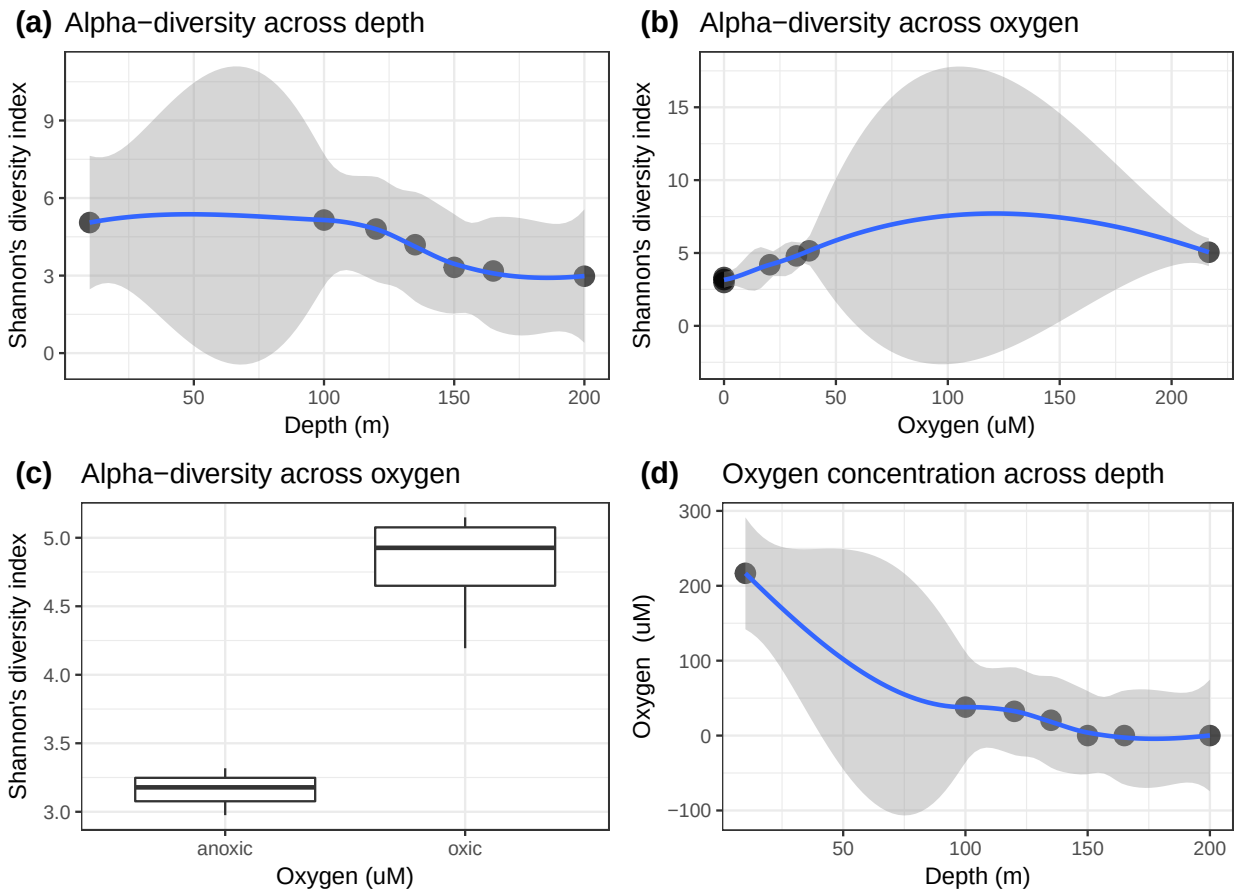

Figure 2: QIIME2

```
colorPlanctomycetes <- "black"

colourCount <- length(unique(tax_table(secondMTaxa)[, "Phylum"]))
myColors <- colorRampPalette(brewer.pal(8, "Accent"))(colourCount)
myColors[12] <- colorPlanctomycetes
p1 <- plot_bar(secondMTaxa, fill = "Phylum") + geom_bar(aes(color = Phylum,
    fill = Phylum), stat = "identity", position = "stack") +
    scale_fill_manual(values = myColors) + scale_colour_manual(values = myColors)
annotate_figure(ggarrange(p1, ncol = 1, nrow = 1), bottom = text_grob("Figure 3: Phylum distri
    face = "bold", size = 12))
```

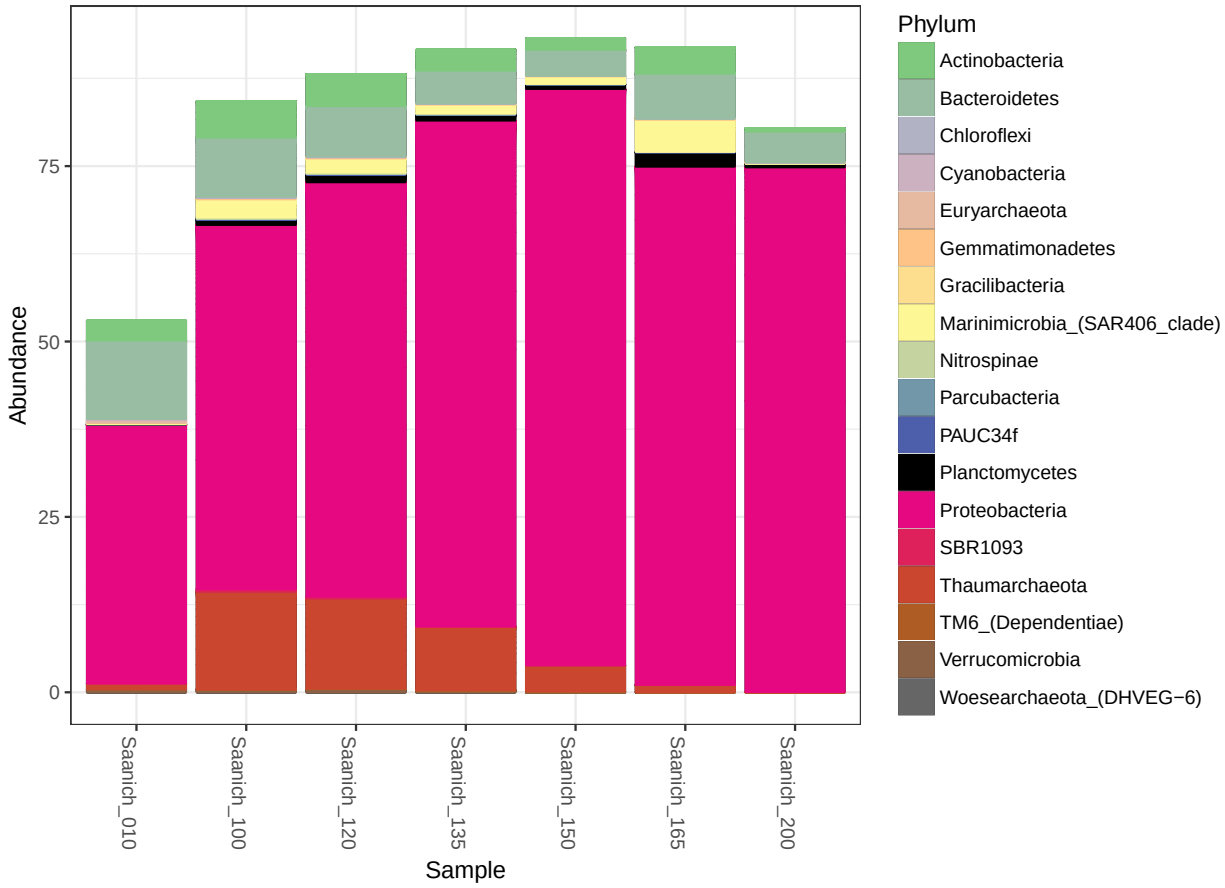

Figure 3: Phylum distribution across samples from Mothur

```
colourCount <- length(unique(tax_table(secondQTaxa)[, "Phylum"]))
myColors <- colorRampPalette(brewer.pal(8, "Accent"))(colourCount)
myColors[7] <- colorPlanctomycetes

p2 <- plot_bar(secondQTaxa, fill = "Phylum") + geom_bar(aes(color = Phylum,
    fill = Phylum), stat = "identity", position = "stack") +
    scale_fill_manual(values = myColors) + scale_colour_manual(values = myColors)
annotate_figure(ggarrange(p2, ncol = 1, nrow = 1), bottom = text_grob("Figure 4: Phylum distril
    face = "bold", size = 12))
```

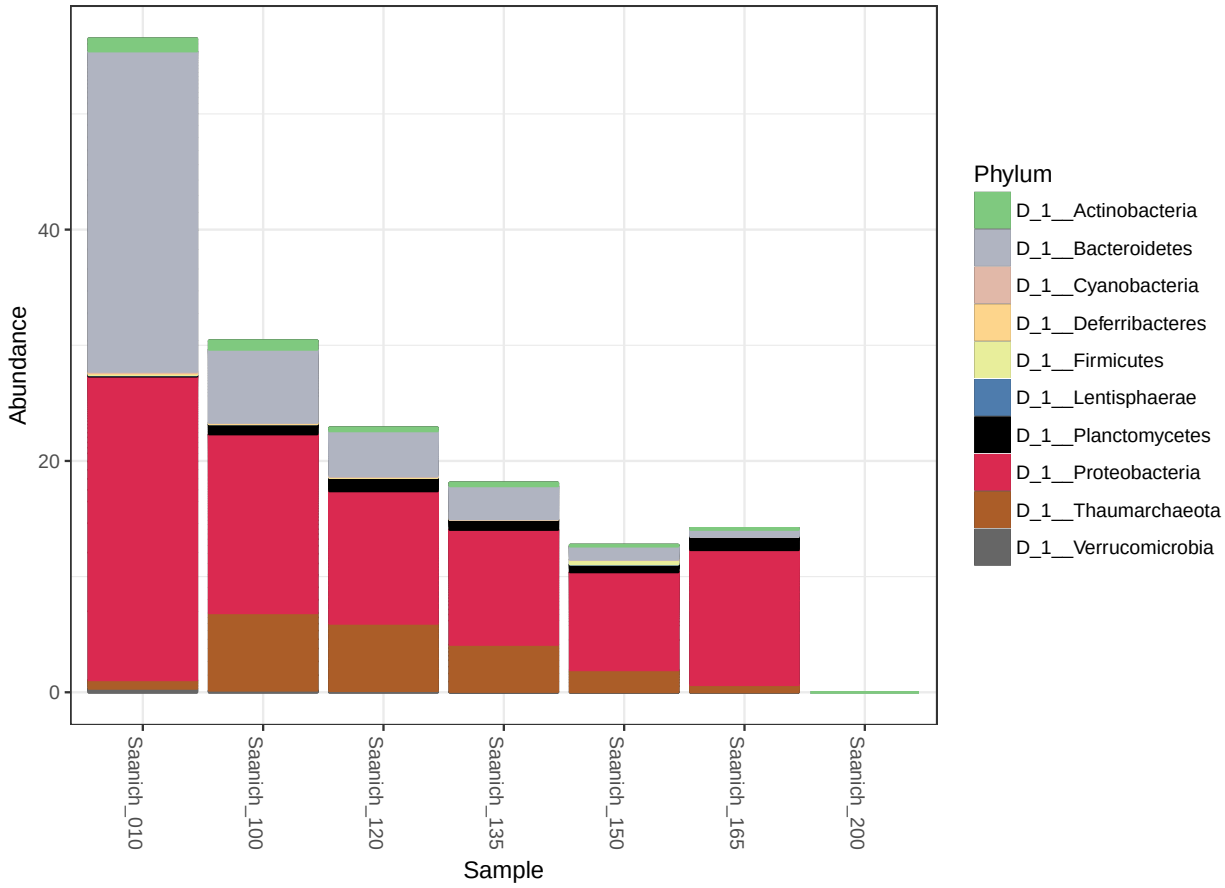

Figure 4: Phylum distribution across samples from QIIME2

3.2 Analysis of abundance information of Planctomyces along with depth and/or oxygen concentration

```
workingMTaxa = subset_taxa(secondMTaxa, Genus == "Planctomyces")
workingQTaxa = subset_taxa(secondQTaxa, Genus == "D_5__Planctomyces")
```

```
phylumM <- get_taxa_unique(physeq = secondMTaxa, taxonomic.rank = "Phylum")
phylumQ <- get_taxa_unique(physeq = secondQTaxa, taxonomic.rank = "Phylum")
phylumMQ <- list(phylumM, phylumQ)
phylumMQ <- data.frame(sapply(phylumMQ, "[", seq(max(sapply(phylumMQ,
    length)))), stringsAsFactors = FALSE)
phylumMQ[is.na(phylumMQ)] <- " "
colnames(phylumMQ) <- c("Phylums from Mothur", "Phylums from QIIME2")
kable(phylumMQ, caption = "Table 1: Phylums from Mothur vs Phylums from QIIME2")
```

Phylums from Mothur	Phylums from QIIME2

Table 1: Table 1: Phylums from Mothur vs Phylums from QIIME2

Phylums from Mothur	Phylums from QIIME2
Proteobacteria	D_1__Proteobacteria
Bacteroidetes	D_1__Bacteroidetes
Thaumarchaeota	D_1__Planctomycetes
Actinobacteria	D_1__Thaumarchaeota
Marinimicrobia_(SAR406_clade)	D_1__Actinobacteria
Planctomycetes	D_1__Deferribacteres
Gemmatimonadetes	D_1__Verrucomicrobia
Verrucomicrobia	D_1__Firmicutes
Nitrospinae	D_1__Lentisphaerae
SBR1093	D_1__Cyanobacteria
TM6_(Dependentiae)	
Chloroflexi	
Cyanobacteria	
Euryarchaeota	
PAUC34f	
Woesearchaeota_(DHVEG-6)	
Gracilibacteria	
Parcubacteria	

Based on Fig. 3, the phylum Proteobacteria was found to be the most abundant from depth of 10m to 200m compared to the other phyla. The abundance of phylum Thaumarchaeota was found to be decreasing from 100m to 200m. The abundance of the other phyla Actinobacteria, Bacteroidetes, Chloroflexi, Cyanobacteria, Euryarchaeota, Gemmatimonadetes, Gracilibacteria, Marinimicrobia_(SAR406_clade), Nitrospinae, Parcubacteria, PAUC34f, Planctomycetes, Proteobacteria, SBR1093, Thaumarchaeota, TM6_(Dependentiae), Verrucomicrobia, and Woesearchaeota_(DHVEG-6) were observed as depth-independent. In the QIIME2 data, there were much less phyla presented (10 types of phyla) in comparison with the Mothur data (18 types of phyla), see Table 1. The abundance of Bacteriodetes seemed to be decreasing from depth 10m to 200m. The abundance of Thaumarcheota was found to be decreasing from depth of 100m to 165m. There was no significant relationship found between the abundance of Proteobacteria and Planctomycetes in regards to depth. Overall, it was observed that the distribution of phyla was changing with depth in the Mothur and QIIME2 data.

```
subMTaxa <- subset_taxa(secondMTaxa, Phylum == "Planctomycetes")
colourCount <- length(unique(tax_table(subMTaxa)[, "Genus"]))
myColors <- colorRampPalette(brewer.pal(8, "Accent"))(colourCount)
myColors[5] <- colorPlanctomycetes
p1 <- plot_bar(subMTaxa, fill = "Genus") + geom_bar(aes(color = Genus,
    fill = Genus), stat = "identity", position = "stack") + scale_fill_manual(values = myColors
    scale_colour_manual(values = myColors)
```

```
figureM <- annotate_figure(ggarrange(p1, ncol = 1, nrow = 1,
    widths = 0.7, heights = 0.2), bottom = text_grob("Mothur",
    color = "blue", face = "bold", size = 12))

subQTaxa = subset_taxa(secondQTaxa, Phylum == "D_1__Planctomycetes")
colourCount <- length(unique(tax_table(subMTaxa)[, "Genus"]))
myColors <- colorRampPalette(brewer.pal(8, "Accent"))(colourCount)
myColors[4] <- colorPlanctomycetes
p2 <- plot_bar(subQTaxa, fill = "Genus") + geom_bar(aes(color = Genus,
    fill = Genus), stat = "identity", position = "stack") + scale_fill_manual(values = myColors
    scale_colour_manual(values = myColors)
figureQ <- annotate_figure(ggarrange(p2, ncol = 1, nrow = 1,
    widths = 0.7, heights = 0.2), bottom = text_grob("QIIME2",
    color = "blue", face = "bold", size = 12))
annotate_figure(ggarrange(figureM, figureQ, ncol = 1, nrow = 2),
    bottom = text_grob("Figure 5: Genus distribution of Planctomycetes across samples",
        face = "bold", size = 12))
```

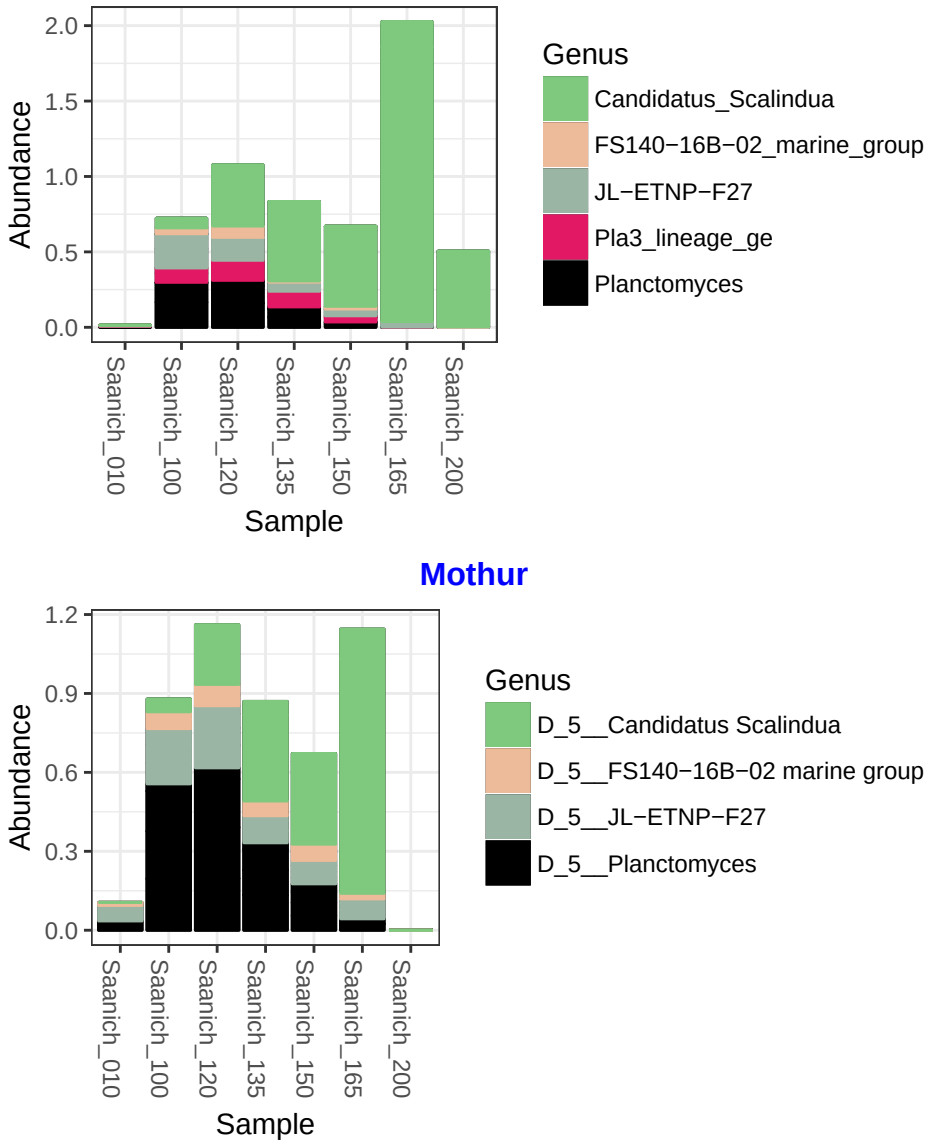

Figure 5: Genus distribution of Planctomycetes across samples

A general linear model was used to investigate the correlation of *Planctomyces*' abundance as a function of depth and/or oxygen concentration. From Table 2, it can be inferred that the abundance of *Planctomyces* had no relationship with varying depths nor oxygen concentration. The p-values, which were above the 5% arbitrary cut-off, also suggested that differences in abundance of *Planctomyces* with respect to depth and oxygen concentration did not differ significantly. These results were consistent with either data sets derived from Mothur and QIIME2.

```
otu_stats <- data.frame()
lmMTaxa1 <- workingMTaxa %>% tax_glom(taxrank = "Genus") %>%
    psmelt() %>% lm(Abundance ~ Depth_m, .) %>% summary()
lmMTaxa2 <- workingMTaxa %>% tax_glom(taxrank = "Genus") %>%
    psmelt() %>% lm(Abundance ~ O2_uM, .) %>% summary()
lmQTaxa1 <- workingQTaxa %>% tax_glom(taxrank = "Genus") %>%
```

```
    psmelt() %>% lm(Abundance ~ Depth_m, .) %>% summary()
lmQTaxa2 <- workingQTaxa %>% tax_glom(taxrank = "Genus") %>%
    psmelt() %>% lm(Abundance ~ O2_uM, .) %>% summary()

otu_stats <- rbind(otu_stats, lmMTaxa1$coefficients["Depth_m",
    ])
otu_stats <- rbind(otu_stats, lmMTaxa2$coefficients["O2_uM",
    ])
otu_stats <- rbind(otu_stats, lmQTaxa1$coefficients["Depth_m",
    ])
otu_stats <- rbind(otu_stats, lmQTaxa2$coefficients["O2_uM",
    ])
tmpNames <- c("Depth (mothur)", "O2_uM (mothur)", "Depth (QIIME2)",
    "O2_uM (QIIME2)")
otu_stats <- cbind(tmpNames, otu_stats)
colnames(otu_stats) <- c("Covariates", "Estimate", "Std. Error",
    "t-value", "Pr(>|t|)")
kable(otu_stats, caption = "Table 2: Correlation data of OTUs within *Planctomyces* genus acros
```

Table 2: Table 2: Correlation data of OTUs within *Planctomyces* genus across depth and oxyzen concentration from Mothur and QIIME2

Covariates	Estimate	Std. Error	t-value	Pr(>|t|)
Depth (mothur)	-0.0003609	0.0010227	-0.3528908	0.7385598
O2_uM (mothur)	-0.0002544	0.0007941	-0.3203956	0.7616253
Depth (QIIME2)	-0.0005878	0.0018774	-0.3130933	0.7668485
O2_uM (QIIME2)	-0.0005997	0.0014441	-0.4152436	0.6951812

```
dfworkingMTaxa <- workingMTaxa %>% psmelt() %>% group_by(Sample) %>%
    summarize(Abundance_sum = sum(Abundance), Depth_m = mean(Depth_m))
p1 <- ggplot(dfworkingMTaxa) + geom_point(aes(x = Depth_m, y = Abundance_sum),
    size = 5, alpha = 0.7) + geom_smooth(method = "lm", aes(x = as.numeric(Depth_m),
    y = Abundance_sum)) + labs(title = "Abundance of Plantomyces across depth")

dfworkingMTaxa <- workingMTaxa %>% psmelt() %>% group_by(Sample) %>%
    summarize(Abundance_sum = sum(Abundance), Oxyzen = mean(O2_uM))
p2 <- ggplot(dfworkingMTaxa) + geom_point(aes(x = Oxyzen, y = Abundance_sum),
    size = 5, alpha = 0.7) + geom_smooth(method = "lm", aes(x = as.numeric(Oxyzen),
    y = Abundance_sum)) + labs(title = "Abundance of Plantomyces with respect to oxyzen concen

dfworkingQTaxa <- workingQTaxa %>% psmelt() %>% group_by(Sample) %>%
    summarize(Abundance_sum = sum(Abundance), Depth_m = mean(Depth_m))
p3 <- ggplot(dfworkingQTaxa) + geom_point(aes(x = Depth_m, y = Abundance_sum),
    size = 5, alpha = 0.7) + geom_smooth(method = "lm", aes(x = as.numeric(Depth_m),
    y = Abundance_sum)) + labs(title = "Abundance of Plantomyces across depth")
```

```
dfworkingQTaxa <- workingQTaxa %>% psmelt() %>% group_by(Sample) %>%
    summarize(Abundance_sum = sum(Abundance), Oxyzen = mean(O2_uM))
p4 <- ggplot(dfworkingQTaxa) + geom_point(aes(x = Oxyzen, y = Abundance_sum),
    size = 5, alpha = 0.7) + geom_smooth(method = "lm", aes(x = as.numeric(Oxyzen),
    y = Abundance_sum)) + labs(title = "Abundance of Plantomyces with respect to oxyzen concen

figureM <- annotate_figure(ggarrange(p1, p2, ncol = 2, nrow = 1,
    labels = c("(a)", "(b)")), bottom = text_grob("Mothur", color = "blue",
    face = "bold", size = 12))
figureQ <- annotate_figure(ggarrange(p3, p4, ncol = 2, nrow = 1,
    labels = c("(a)", "(b)")), bottom = text_grob("QIIME2", color = "blue",
    face = "bold", size = 12))
annotate_figure(ggarrange(figureM, figureQ, ncol = 1, nrow = 2),
    bottom = text_grob("Figure 6: Regression analysis of *Planctomyces* across depth",
        face = "bold", size = 12))
```

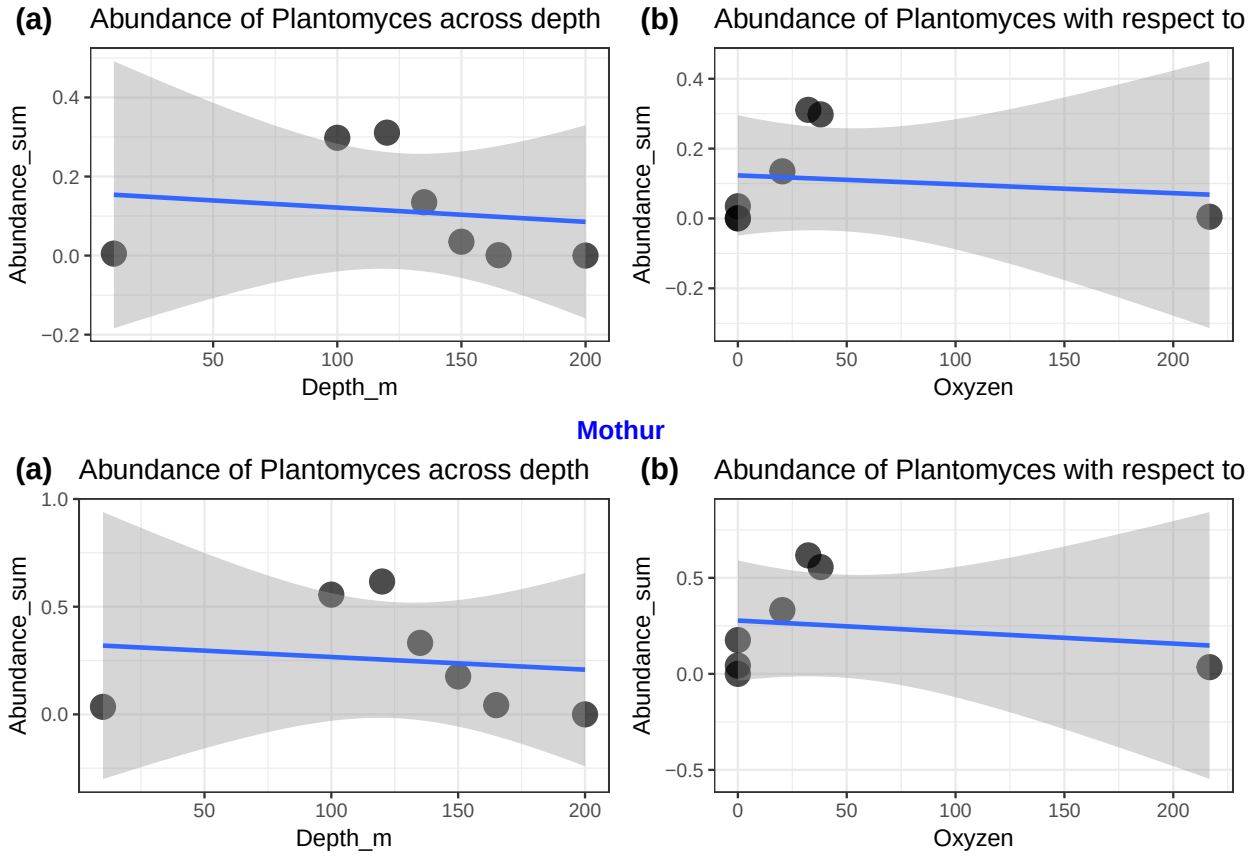

Figure 6: Regression analysis of *Planctomyces* across depth

3.3 Estimate richness (number of OTUs/ASVs) for *Planctomyces*

```
suggestedMOTUs <- colnames(otu_table(workingMTaxa))
suggestedQOTUs <- rownames(otu_table(workingQTaxa))
suggestedMQOTUs <- list(suggestedMOTUs, suggestedQOTUs)
suggestedMQOTUs <- data.frame(sapply(suggestedMQOTUs, "[", seq(max(sapply(suggestedMQOTUs,
    length)))))
colnames(suggestedMQOTUs) <- c("OTUs from Mothur", "ASVs from QIIME2")
kable(suggestedMQOTUs, caption = "Table 3: OTUs from Mothur vs ASVs from QIIME2")
```

Table 3: Table 3: OTUs from Mothur vs ASVs from QIIME2

OTUs from Mothur	ASVs from QIIME2
Otu0125	Asv232
Otu0144	Asv799
Otu0401	Asv1021
Otu0592	Asv1124

Table 3 shows that in our analysis we observe only 4 OTUs/ASVs within the *Planctomyces* genus.

3.4 Interpretation of abundance information of OTUs/ASVs of *Planctomyces* along with depth and/or oxygen concentration

```
otu_stats <- data.frame()
for (otu in suggestedMOTUs) {
    linear_fit <- workingMTaxa %>% psmelt() %>% filter(OTU ==
        otu) %>% lm(Abundance ~ Depth_m, .) %>% summary()
    otu_stats <- rbind(otu_stats, linear_fit$coefficients["Depth_m",
        ])
}
tmpNames <- paste(suggestedMOTUs, " (Mothur) ")
for (otu in suggestedQOTUs) {
    linear_fit <- workingQTaxa %>% psmelt() %>% filter(OTU ==
        otu) %>% lm(Abundance ~ Depth_m, .) %>% summary()
    otu_stats <- rbind(otu_stats, linear_fit$coefficients["Depth_m",
        ])
}
tmpNames <- c(tmpNames, paste(suggestedQOTUs, " (QIIME2) "))
otu_stats <- cbind(tmpNames, otu_stats)
colnames(otu_stats) <- c("Covariates", "Estimate", "Std. Error",
    "t-value", "Pr(>|t|)")
kable(otu_stats, caption = "Table 4: Correlation data of OTUs within *Planctomyces* genus acros
```

Table 4: Table 4: Correlation data of OTUs within *Planctomyces* genus across depth

Covariates	Estimate	Std. Error	t-value	Pr(>|t|)
Otu0125 (Mothur)	-0.0002045	0.0005479	-0.3731784	0.7243139
Otu0144 (Mothur)	-0.0001533	0.0004035	-0.3798581	0.7196506
Otu0401 (Mothur)	-0.0000009	0.0000807	-0.0113619	0.9913741
Otu0592 (Mothur)	-0.0000023	0.0000274	-0.0836392	0.9365887
Asv232 (QIIME2)	-0.0001665	0.0006541	-0.2545201	0.8092302
Asv799 (QIIME2)	-0.0000953	0.0005682	-0.1676505	0.8734282
Asv1021 (QIIME2)	0.0000544	0.0001222	0.4454921	0.6745908
Asv1124 (QIIME2)	-0.0003805	0.0005859	-0.6493922	0.5447317

```
otu_stats <- data.frame()
for (otu in suggestedMOTUs) {
    linear_fit <- workingMTaxa %>% psmelt() %>% filter(OTU ==
        otu) %>% lm(Abundance ~ O2_uM, .) %>% summary()
    otu_stats <- rbind(otu_stats, linear_fit$coefficients["O2_uM",
        ])
}
tmpNames <- paste(suggestedMOTUs, " (Mothur) ")
for (otu in suggestedQOTUs) {
    linear_fit <- workingQTaxa %>% psmelt() %>% filter(OTU ==
        otu) %>% lm(Abundance ~ O2_uM, .) %>% summary()
    otu_stats <- rbind(otu_stats, linear_fit$coefficients["O2_uM",
        ])
}
tmpNames <- c(tmpNames, paste(suggestedQOTUs, " (QIIME2) "))
otu_stats <- cbind(tmpNames, otu_stats)
colnames(otu_stats) <- c("Covariates", "Estimate", "Std. Error",
    "t-value", "Pr(>|t|)")
kable(otu_stats, caption = "Table 5: Correlation data of OTUs within *Planctomyces* genus acros
```

Table 5: Table 5: Correlation data of OTUs within *Planctomyces* genus across oxygen concentration

Covariates	Estimate	Std. Error	t-value	Pr(>|t|)
Otu0125 (Mothur)	-0.0001290	0.0004265	-0.3025609	0.7744070
Otu0144 (Mothur)	-0.0000962	0.0003142	-0.3062077	0.7717866
Otu0401 (Mothur)	-0.0000196	0.0000619	-0.3172313	0.7638869
Otu0592 (Mothur)	-0.0000096	0.0000208	-0.4601604	0.6647195
Asv232 (QIIME2)	-0.0002374	0.0004989	-0.4759372	0.6541875
Asv799 (QIIME2)	-0.0002374	0.0004285	-0.5541303	0.6033590
Asv1021 (QIIME2)	-0.0001100	0.0000831	-1.3250528	0.2424753
Asv1124 (QIIME2)	-0.0000147	0.0004727	-0.0311450	0.9763589

```
p1 <- workingMTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = Depth_m,
    y = Abundance), size = 5, alpha = 0.7) + geom_smooth(method = "lm",
    aes(x = Depth_m, y = Abundance)) + facet_wrap(~OTU, scales = "free_y") +
    labs(title = "Per OTU abundance information")
p2 <- workingMTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = Sample,
    y = OTU, size = Abundance, color = OTU)) + scale_size_continuous(range = c(0,
    5)) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Abundance of OTUs across samples")
p3 <- workingQTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = Depth_m,
    y = Abundance), size = 5, alpha = 0.7) + geom_smooth(method = "lm",
    aes(x = Depth_m, y = Abundance)) + facet_wrap(~OTU, scales = "free_y") +
    labs(title = "Per ASV abundance information")
p4 <- workingQTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = Sample,
    y = OTU, size = Abundance, color = OTU)) + scale_size_continuous(range = c(0,
    5)) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Abundance of ASVs across samples")

figureM <- annotate_figure(ggarrange(p1, p2, ncol = 2, nrow = 1,
    labels = c("(a)", "(b)")), bottom = text_grob("Mothur", color = "blue",
    face = "bold", size = 12))
figureQ <- annotate_figure(ggarrange(p3, p4, ncol = 2, nrow = 1,
    labels = c("(a)", "(b)")), bottom = text_grob("QIIME2", color = "blue",
    face = "bold", size = 12))

annotate_figure(ggarrange(figureM, figureQ, ncol = 1, nrow = 2,
    labels = c("(a)", "(b)")), bottom = text_grob("Figure 7: Abundance of OTUs/ASVs within *Pla
    face = "bold", size = 12))
```

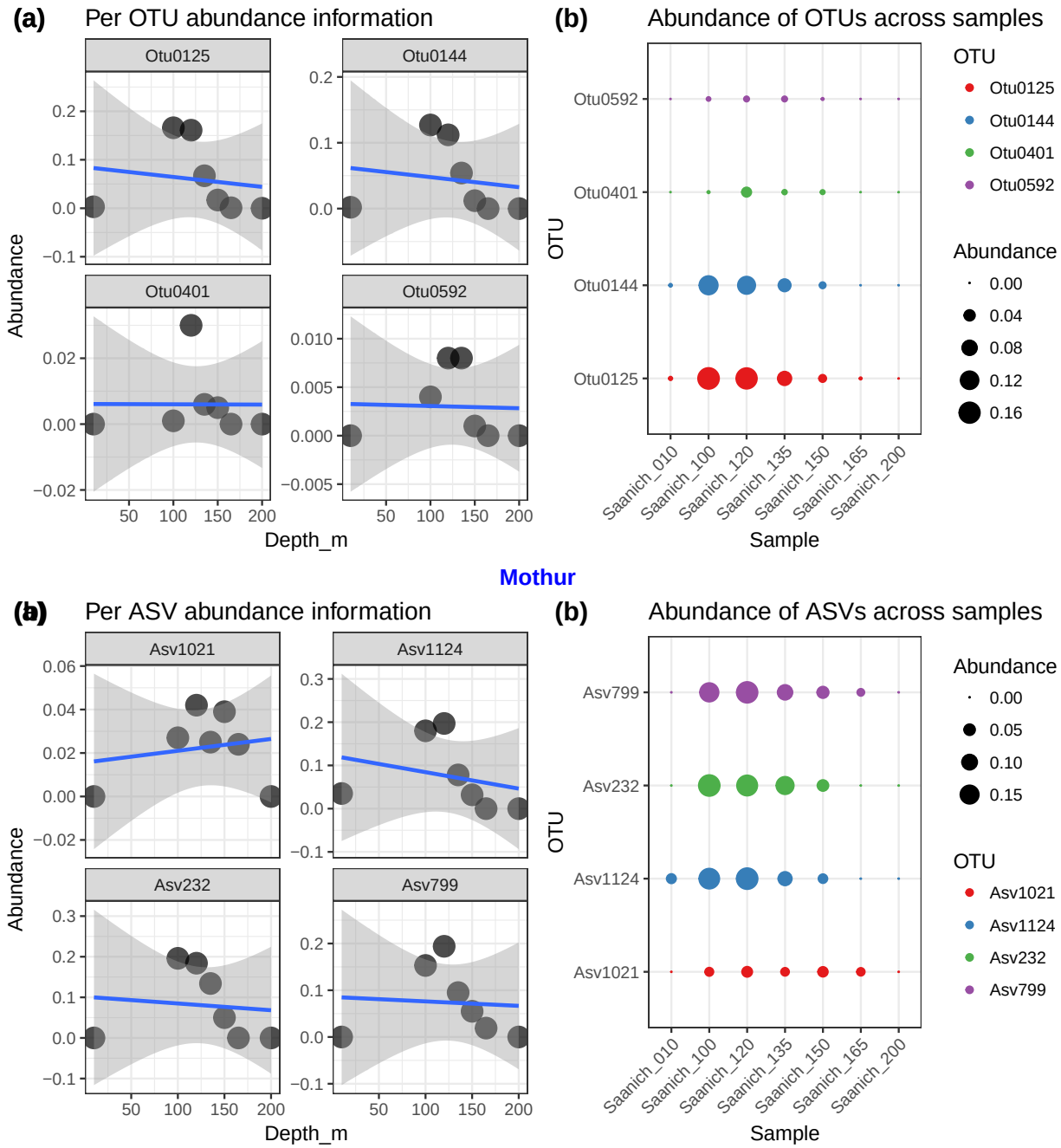

Figure 7: Abundance of OTUs/ASVs within *Planctomyces* genus across depth

```
p1 <- workingMTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = O2_uM,
    y = Abundance), size = 5, alpha = 0.7) + geom_smooth(method = "lm",
    aes(x = O2_uM, y = Abundance)) + facet_wrap(~OTU, scales = "free_y") +
    labs(title = "Per OTU abundance information")
p2 <- workingMTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = Sample,
    y = OTU, size = O2_uM, color = OTU)) + scale_size_continuous(range = c(0,
    5)) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Abundance of OTUs across samples")
```

```
p3 <- workingQTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = O2_uM,
    y = Abundance), size = 5, alpha = 0.7) + geom_smooth(method = "lm",
    aes(x = Depth_m, y = Abundance)) + facet_wrap(~OTU, scales = "free_y") +
    labs(title = "Per ASV abundance information")
p4 <- workingQTaxa %>% psmelt() %>% ggplot() + geom_point(aes(x = Sample,
    y = OTU, size = O2_uM, color = OTU)) + scale_size_continuous(range = c(0,
    5)) + theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Abundance of ASVs across samples")

figureM <- annotate_figure(ggarrange(p1, p2, ncol = 2, nrow = 1,
    labels = c("(a)", "(b)")), bottom = text_grob("Mothur", color = "blue",
    face = "bold", size = 12))

figureQ <- annotate_figure(ggarrange(p3, p4, ncol = 2, nrow = 1,
    labels = c("(a)", "(b)")), bottom = text_grob("QIIME2", color = "blue",
    face = "bold", size = 12))

annotate_figure(ggarrange(figureM, figureQ, ncol = 1, nrow = 2,
    labels = c("(a)", "(b)")), bottom = text_grob("Figure 8: Abundance of OTUs/ASVs within *Pla
    face = "bold", size = 12))
```

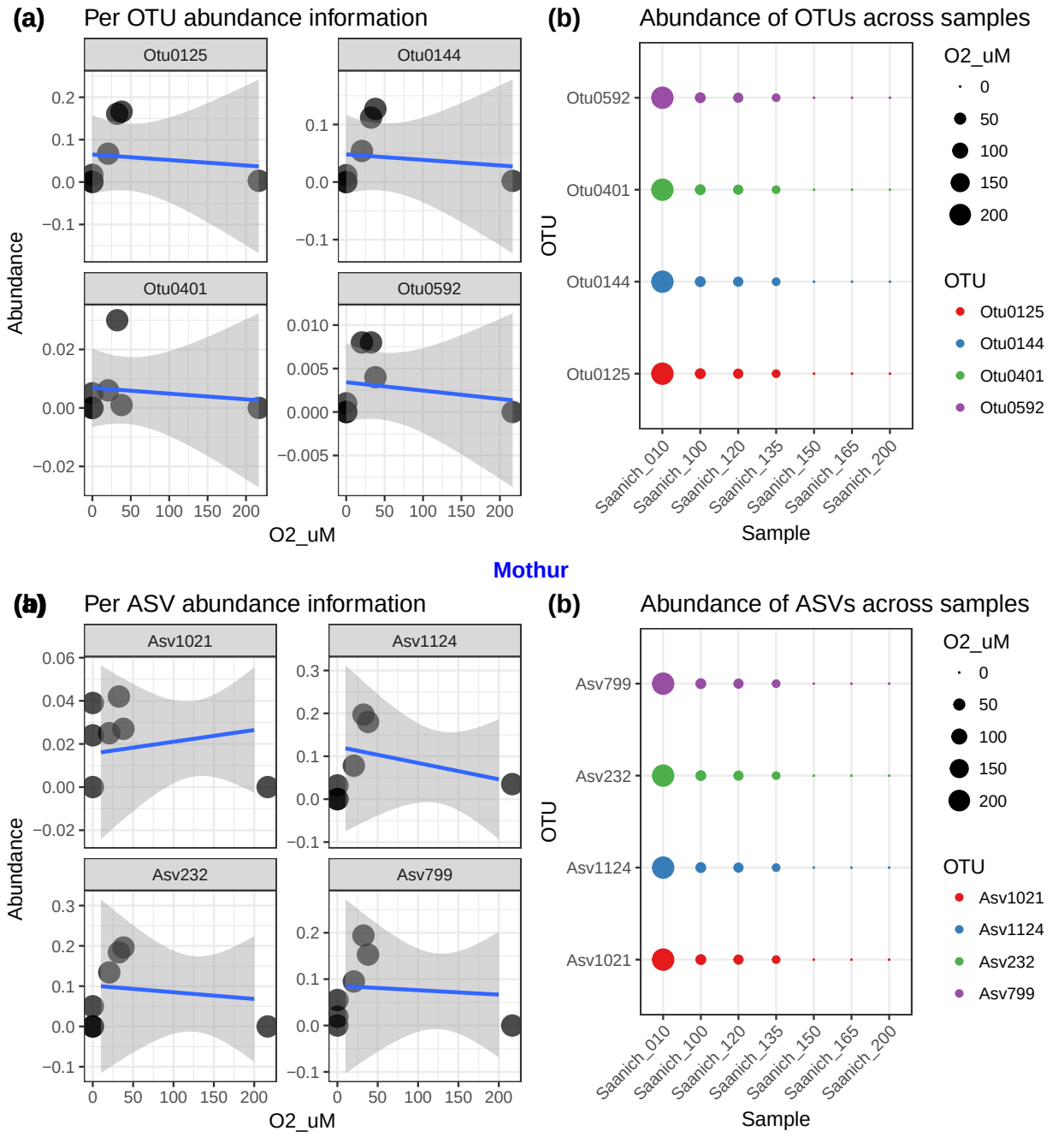

Figure 8: Abundance of OTUs/ASVs within *Planctomyces* genus across oxygen concentration

As we highlighted in Table 3, there were 4 OTUs within the *Planctomyces* genus. The abundances of the OTUs within the *Planctomyces* genus did not significantly differ across depth. None of the linear models generated for these OTUs have a p-value below 0.05, hence a relationship cannot be established between abundance and depth, as summerized in Table 4. The abundances of these OTUs also did not change significantly with oxygen concentration either in Table 5. The linear models generated to compare the relationship between abundance and oxygen concentration also had p-values that were above 0.05. This indicated that the abundances of these OTUs were unaffected by oxygen concentration. Similar results were observed from the ASV results generated from QIIME2,

where none of the p-values generated from the linear models were below 0.05.

3.5 Analysis of results from Mothur and QIIME2 processed data

We observe no significant differences between the Mothur and QIIME2 data in terms of depths, oxygen concentration, and richness. However, there were some differences that should be noted:

1) Higher Shannon's Diversity Index was noted in QIIME2 vs. Mothur in terms of oxygen concentrations and anoxic versus oxic levels (Fig. 1 and 2).

2) More phyla observed from Mothur (18) than QIIME2 (10) (Table 1).

3) An extra genus was present in the genus distribution of the Planctomycetes phylum across samples from Mothur: *Pla3_lineage_ge* (Fig. 5).

4) In terms of abundance, the genus distribution of *Planctomyces* in QIIME2 data was higher than Mothur (Fig. 5).

5) The scaling of abundance (Fig. 5) in the genus distribution of Planctomycetes across samples from Mothur were also noted to be higher than the scaling of QIIME2 data (Mothur has the Saanich depth of 165m to be approx. 2.0 for abundance whereas QIIME2 has the Saanich depth of 165m to be approx. 1.2 for abundance).

4 Discussion

Planctomyces, also known as anammox bacteria, are known to perform an important process in the nitrogen cycle: Anammox ($NO_2^- + NH_4^+$ -> $N_2 + 2H_2O$). Anammox can be carried out under anaerobic conditions and is an important process for recycling nitrogen back into the environment and for waste-water treatment [19].

One study have suggested that approximately $30-50\%$ of the total nitrogen loss is currently estimated to take place in OMZs (Oxygen Minimum Zone), and very low concentrations of ammonium in suboxic waters indicate that the anammox process could play an important role in these ecosystems [19]; it was reported that anammox bacteria were very sensitive to oxygen and that oxygen concentrations as little as 1μM can reversibly inhibited the anammox reaction. It has also been suggested that the ammonium concentrations were below the detection limit in OMZs and anammox, rather than heterotrophic denitrification, was responsible for the nitrogen loss [19]. Since anammox bacteria abundance was correlated with nitrogen loss, it was expected that there will be higher levels of anammox bacteria such as *Planctomyces* in OMZs. Therefore, based on the information provided by the study, the fact that no relationship is observed between the abundance of *Planctomyces* with oxygen level and ocean depth is not expected.

However, another study had pointed out that *Planctomyces* were most abundant in the oxic part of the wetland profiles [20]. The respective cell numbers were in the range $1.1 - 6.7 \times 10^7$ cells g^{-1} of wet peat, comprising $2 - 14\%$ of total bacterial cells, and displaying linear correlation to the peat water pH [20]. This result contradicted with the research study done by Koo et al. [19], which observed the decline of *Planctomycetes* population at the oxic part of the ocean [19]. Also, Dedysh and Ivanova suggested that different species of Planctomycetes colonized different parts of the ocean: Oxic peat layers were dominated by representatives of the Isosphaera-Singulisphaera

group, while anoxic peat was inhabited mostly by Zavarzinella- and Pirellula-like *Planctomyces* [20]. Therefore, the fact that no was no significance between the *Planctomyces* population versus the oxygen concentration and depth were justifiable, since different stains of Planctomycetes colonized different parts of the ocean. The specific strain of anammox Planctomycetes was found to be more abundant in anoxic part of the ocean, but different strains of Planctomycetes were found to be distributed across parts of the ocean.

In terms of the bioinformatic pipelines commonly used in metagenomics, some differences were observed between the Mothur and QIIME2 data:

1) Higher Shannon's diversity indexes were observed in the QIIME2 data when compared to the Mothur data in terms of oxygen concentrations and oxic versus anoxic levels. One possible explanation for the higher Shannon's diversity indexes in the QIIME2 data could be due to the difference in treatment of OTU and ASV. QIIME2 treats each ASV as an individual species, whereas Mothur uses the representative sequence of each OTU to determine the taxonomy [21]. This could potentially lead to higher Shannon's diversity indexes in the QIIME2 when compared to the Mothur data.

2) The absence of *Pla3_lineage_ge* at the genus level in the QIIME2 data were also noted when compared to the Mothur data. One possibility for this observation is that QIIME2 discards more data through stricter filtering [22]. Another possibility for this observation is that Mothur keeps more of these types of data even if they might not represent the "real" taxa in the community [21].

3) The relative abundance of *Planctomyces* at the genus level were indicated to be higher in the QIIME2 data in comparison to the Mothur data. The removal of *Pla3_lineage_ge* by QIIME2 described above would affect the relative abundance calculation, which is the percent composition of an organism of a particular kind relative to the total number of organisms in the area. By not considering *Pla3_lineage_ge* in the total number of organisms, this would lead to a higher relative abundance of *Planctomyces* across all depths in the QIIME2 data.

4) The relative abundance in the genus distribution of planctomycetes across samples from Mothur data were also noted to be higher than the relative abundance of QIIME2 data. This could be due to Mothur keeping more data than QIIME2, which would affect the scaling of relative abundance. Similar to the previous point, these extra data kept by Mothur might not represent the "real" taxa in the community [21].

In addition to the comparisons of Mothur and QIIME2 in our analyses, another study demonstrated that between QIIME, Mothur, and MG-RAST, differences were mostly observed at the genus level due to Mothur's tendency to have unclassified reads [2]. These inconsistencies highlight the limitations of bioinformatics pipelines and their ability to distinguish between some 16S rRNA sequences at a genus and species level because of their near identical 16s rRNA sequences.

Additional research questions can be addressed with this dataset for more in-depth exploration of future directions. In terms of the unclassified OTUs/ASVs, future research can be done to determine how significant the unclassified taxonomies are in impacting the geochemical gradients in Saanich Inlet in order to better understand their roles in the nutrient cycles. Another alternative question that can be addressed would be whether there are any significant differences with the classification of the other genera (including the ones that were unclassified) by both Mothur and QIIME2.

References

1. Hawley AK, Torres-Beltrán M, Zaikova E, Walsh DA, Mueller A, Scofield M, et al. A compendium of multi-omic sequence information from the saanich inlet water column. Scientific data. Nature Publishing Group; 2017;4:170160.

2. Hallam SJ, Torres-Beltrán M, Hawley AK. Monitoring microbial responses to ocean deoxygenation in a model oxygen minimum zone. Scientific data. Nature Publishing Group; 2017;4.

3. Ulloa O, Canfield DE, DeLong EF, Letelier RM, Stewart FJ. Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences. National Acad Sciences; 2012;109:15996–6003.

4. Wright JJ, Konwar KM, Hallam SJ. Microbial ecology of expanding oxygen minimum zones. Nature Reviews Microbiology. Nature Publishing Group; 2012;10:381.

5. Anderson JJ, Devol AH. Deep water renewal in saanich inlet, an intermittently anoxic basin. Estuarine and Coastal Marine Science. Elsevier; 1973;1:1–10.

6. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal. Nature Publishing Group; 2017;11:2639.

7. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Applied and environmental microbiology. Am Soc Microbiol; 2007;73:278–88.

8. Fuerst JA, Sagulenko E. Beyond the bacterium: Planctomycetes challenge our concepts of microbial structure and function. Nature Reviews Microbiology. Nature Publishing Group; 2011;9:403.

9. McFarland KD. mothur Pipeline. https://github.com/EDUCE-UBC/MICB425/blob/master/Module_03/Project1/data/mothur_pipeline.html; 2018.

10. Beni J. QIIME2 Pipeline. https://github.com/EDUCE-UBC/MICB425/blob/master/Module_03/Project1/data/qiime2_pipeline.html; 2018.

11. McMurdie PJ, Holmes S. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. PloS one. Public Library of Science; 2013;8:e61217.

12. Wickham H. Tidyverse: Easily install and load'tidyverse'packages. R package version. 2017;1.

13. Auguie B. GridExtra: Functions in grid graphics. R package version 09. 2012;1.

14. Bache SM, Wickham H. Magrittr: A forward-pipe operator for r. R package version. 2014;1.

15. Kassambara A. Ggpubr: 'Ggplot2' based publication ready plots, r packag. Version 01. 2017;2.

16. Xie Y. Dynamic documents with r and knitr. CRC Press; 2015.

17. Finotello F, Mastrorilli E, Di Camillo B. Measuring the diversity of the human microbiota with targeted next-generation sequencing. Briefings in bioinformatics. Oxford University Press; 2016;bbw119.

18. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: Data mining, inference

and prediction. 2nd ed. Springer; 2009.

19. Koo H, Hakim JA, Morrow CD, Eipers PG, Davila A, Andersen DT, et al. Comparison of two bioinformatics tools used to characterize the microbial diversity and predictive functional attributes of microbial mats from lake obersee, antarctica. Journal of microbiological methods. Elsevier; 2017;140:15–22.

20. Dedysh SN, Ivanova AO. Abundance, diversity, and depth distribution of planctomycetes in acidic northern wetlands. Frontiers in microbiology. Frontiers; 2012;3:5.

21. McFarland KD. OTU vs. ASV. https://github.com/EDUCE-UBC/MICB425/blob/master/Module_03/Project1/20180307_intro_OTU_ASV.pdf; 2018.

22. McFarland KD. Major differences in alph diversity between q1 and q2. https://forum.qiime2.org/t/major-differences-in-alph-diversity-between-q1-and-q2/2419/3; 2018.