

# 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 13, 2018*

## Abstract

This is the abstract. It consists of two paragraphs.

## Contents

<b>1</b>	<b>Abstract</b>	<b>1</b>
<b>2</b>	<b>Introduction</b>	<b>2</b>
<b>3</b>	<b>Materials and Experimental Configuration</b>	<b>3</b>
3.1	Experimental Protocols . . . . .	3
3.2	Methods . . . . .	3
3.2.1	Shannon Diversity Index (SDI) and Chao1 . . . . .	3
3.2.2	General Linear Model . . . . .	4
3.3	Data Preporocessing . . . . .	4
<b>4</b>	<b>Results</b>	<b>5</b>
4.1	Analysis of microbial community structure along with depth and oxygen concentration	5
4.2	Analysis of abundance information of Planctomyces along with depth and/or oxygen concentration . . . . .	8
4.3	Estimate richness (number of OTUs/ASVs) for Planctomyces . . . . .	11
4.4	Interpretation of abundance information of OTUs/ASVs of Planctomyces along with depth and/or oxygen concentration . . . . .	12
<b>5</b>	<b>Discussion</b>	<b>15</b>
<b>6</b>	<b>References</b>	<b>16</b>

---

## 1 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 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 changes 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.

## 2 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 hydrogen sulfide which is developed at the bottom of the water column (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 QIIME 2 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). Planctomycetes 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).

This report further addresses the planctomyces diversity correlation with oxygen and depth. We show that under the statistical framework there is little evidence to support the initial hypothesis. However, we claim that the linear model was not adequate to provide insightful knowledge regarding the existence of such correlations.

### 3 Materials and Experimental Configuration

#### 3.1 Experimental Protocols

To understand the correlation of microbial diversity and oxygen concentration across samples, we report four experimentally designed test protocols:

- P1.** Analysis of microbial community structure along with depth and oxygen concentration.
- P2.** Analysis of abundance information of Planctomyces along with depth and/or oxygen concentration.
- P3.** Estimate richness (number of OTUs/ASVs) for Planctomyces.
- P4.** Interpretation of abundance information of OTUs/ASVs of Planctomyces along with depth and/or oxygen concentration.

#### 3.2 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).

##### 3.2.1 Shannon Diversity Index (SDI) and Chao1

The Shannon diversity index (SDI) was used to estimate the microbial diversity of Saanich Inlet dataset with the following definition:  $SDI = -\sum_i^R p_i \log(p_i)$

Where  $p_i$  represents the distribution of individuals belonging to the  $i$ th species, and  $R$  represents the number of distinct species. 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 nor represents an accurate estimation of species diversity because the probability distribution of species is not knowable exactly; it is only an estimate from the 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,  $\alpha$  and  $\beta$  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.

### 3.2.2 General Linear Model

General linear model was employed to recover interactions between several factors that might be exhibited in Saanich Inlet dataset. A single regression model that relates a dependent variable  $y$  (abundance) to a single quantitative independent variable  $x_1$  (depth or oxygen) was used:

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

The parameter  $\theta_0$  is the  $y$ -intercept, which represents the expected value of  $y$  when  $x_1$  is zero. The parameter  $\theta_1$  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).  $\theta_1$  could be 0 indicating no effective change with  $x_1$ . And,  $\epsilon$  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  $\theta_1$ , 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  $\theta_1$  and the  $\gamma$  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.

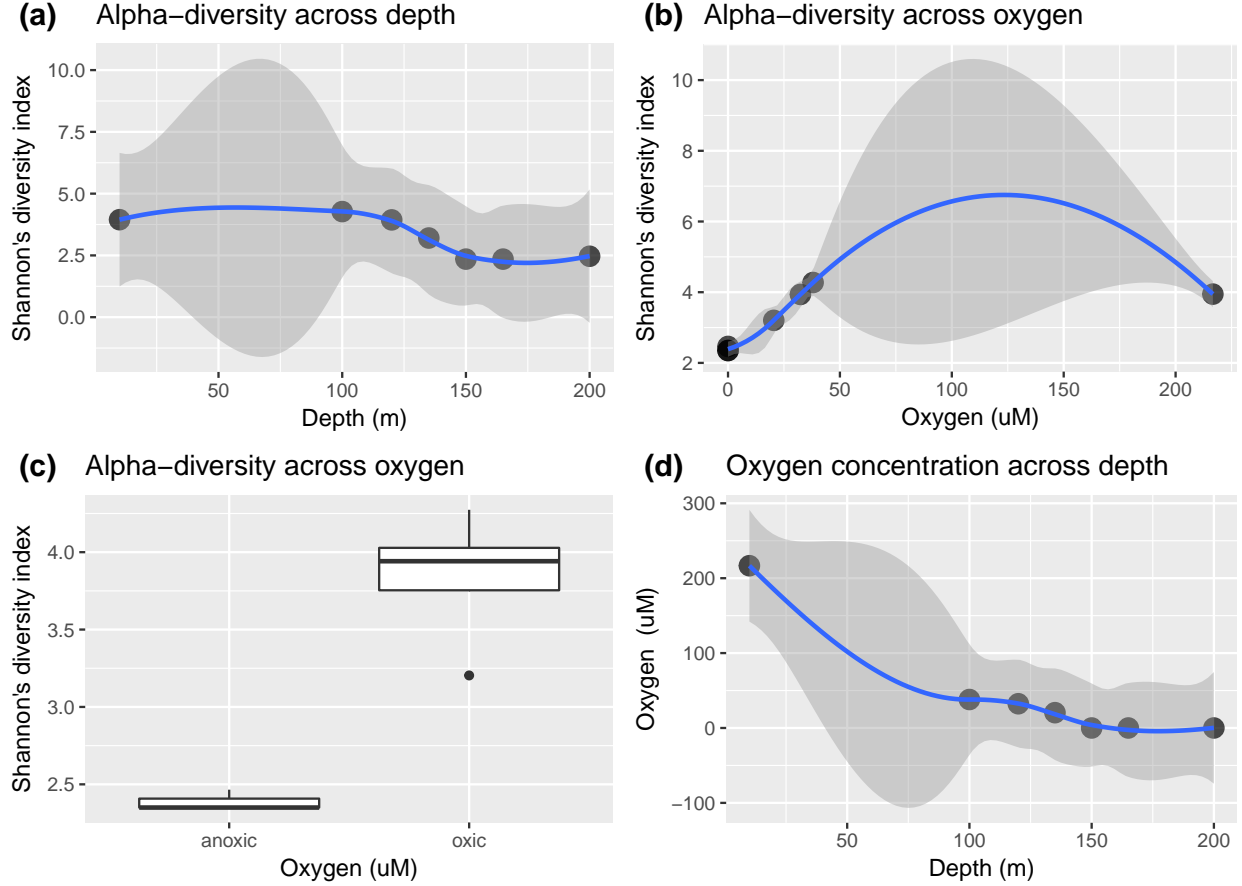
If the weight of  $\theta_1$ , referred to as the level of significance (or  $p$ -value) and defines as a probability, is below or equal to  $\gamma$  then  $H_1$  was accepted; otherwise, accept  $H_0$ .

## 3.3 Data Preprocessing

The samples were normalized to 100,000 sequences per sample to facilitate comparisons between samples. The rarefied counts were then converted to relative abundance percentages. Next, we perform a series of filterings according to three rules: i)- Exclude OTUs that are not observed for more than 4 samples, and ii)- Prune samples and OTUs with unknown values, such as `unclassified` value. This has resulted in 402 and 203 taxa from `mothur` and `QIIME2`, respectively. No other preprocessing were applied. The implementations are done entirely using R (v 3.4.3) and relied on some efficient third-party libraries, such as `phyloseq`, `tidyverse`, `gridExtra`, and `magrittr` (11-19).

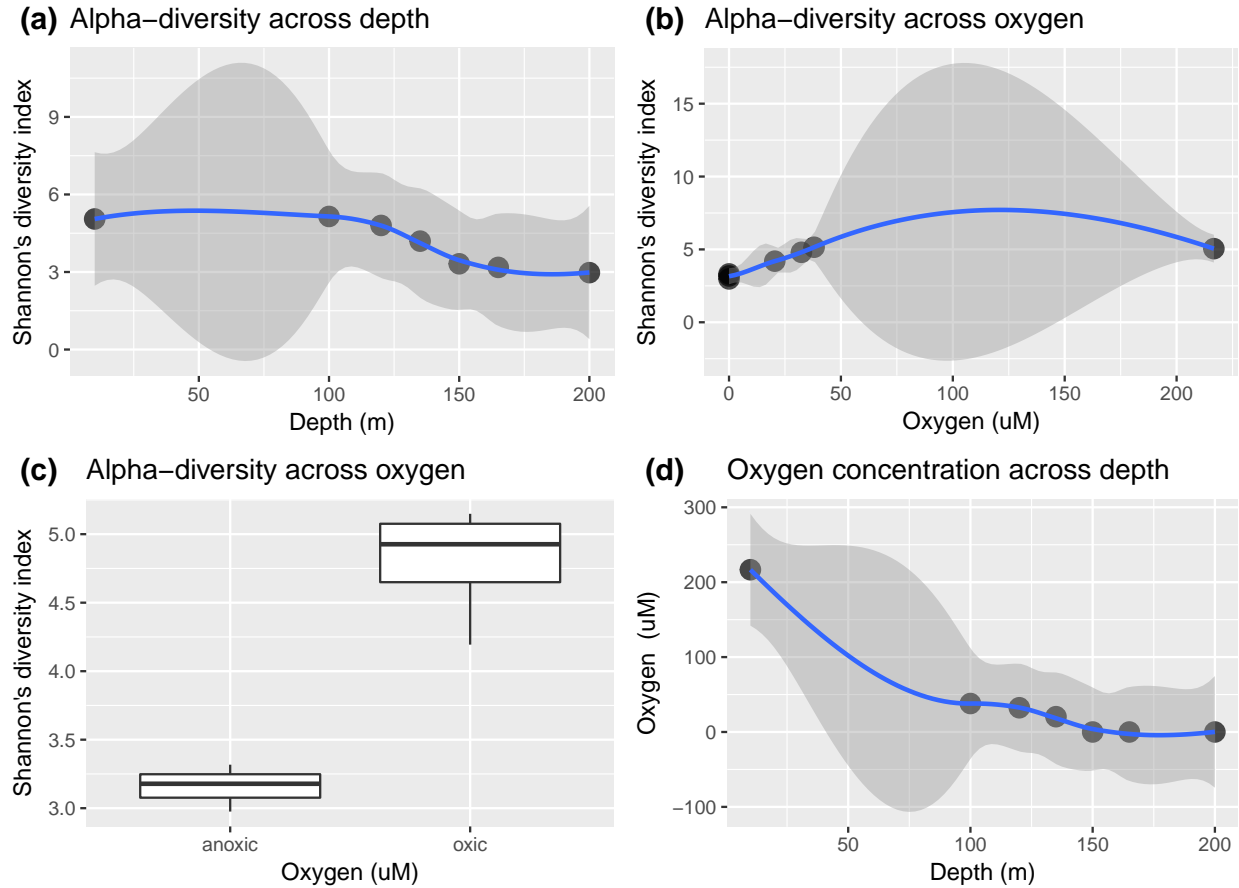
## 4 Results

### 4.1 Analysis of microbial community structure along with depth and oxygen concentration



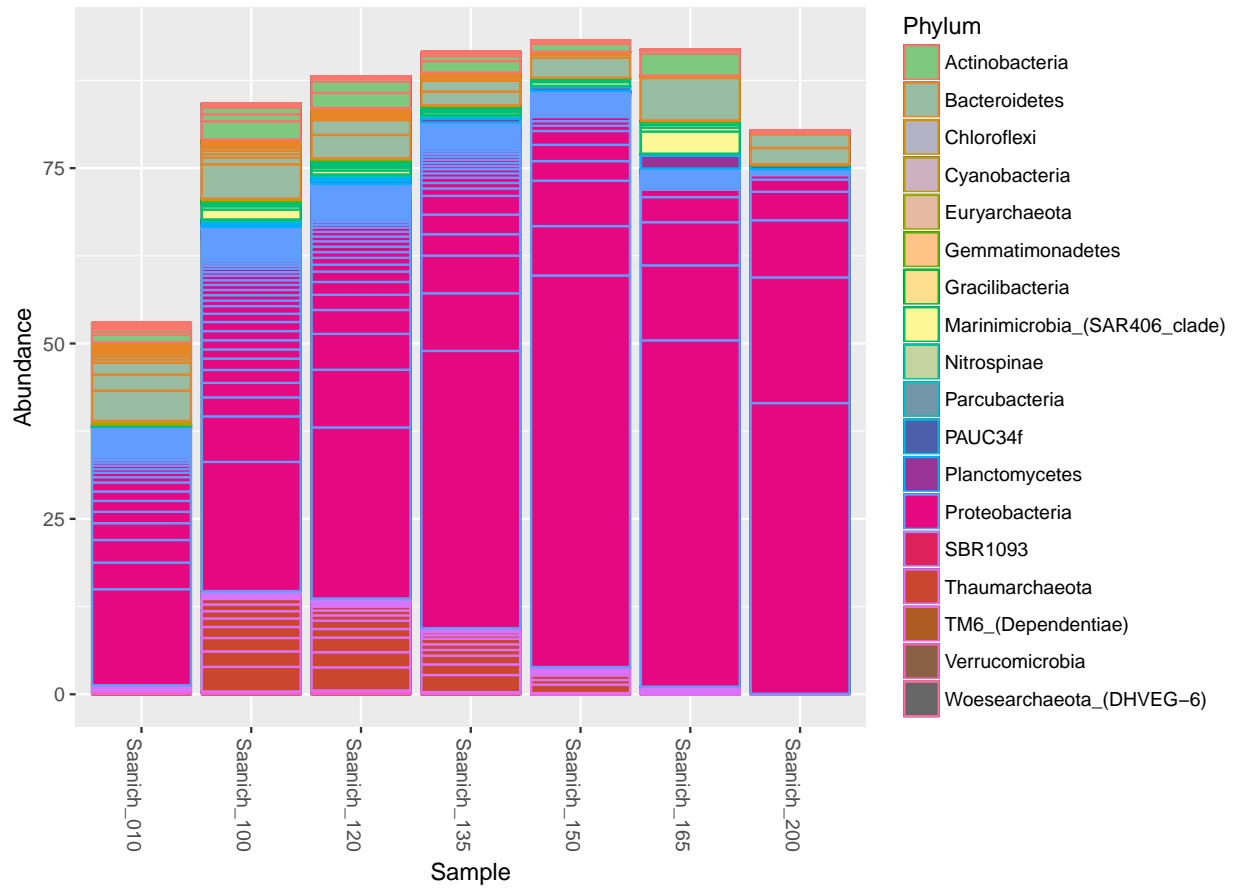
**Figure 1: Mothur**

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.

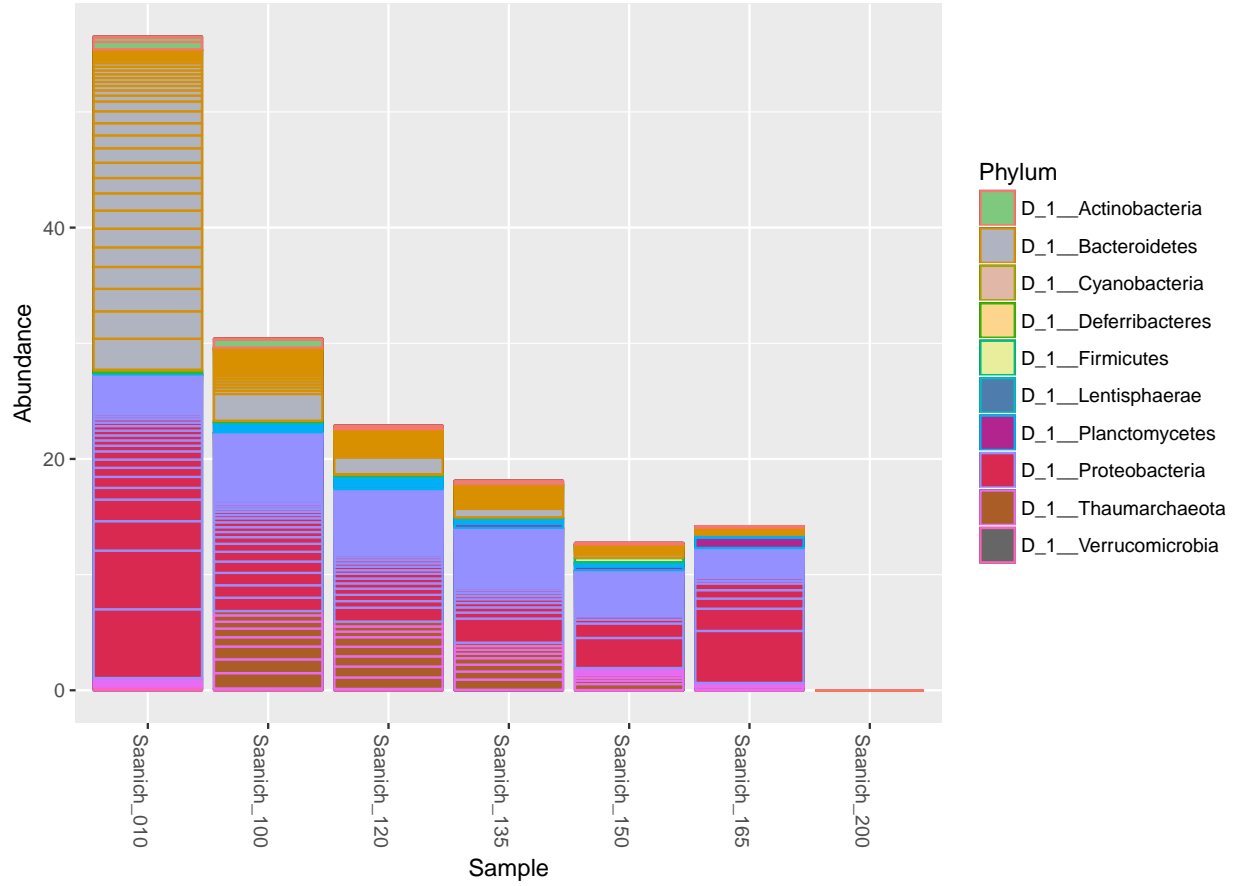


**Figure 2: QIIME2**

A linear model analysis (Table 2) was used to investigate the correlation of *Planctomyces*' abundance as a function of depth and/or oxygen concentration. From a statistical perspective, 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.



**Figure 3: Phylum distribution across samples from mothur**



**Figure 4: Phylum distribution across samples from QIIME2**

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

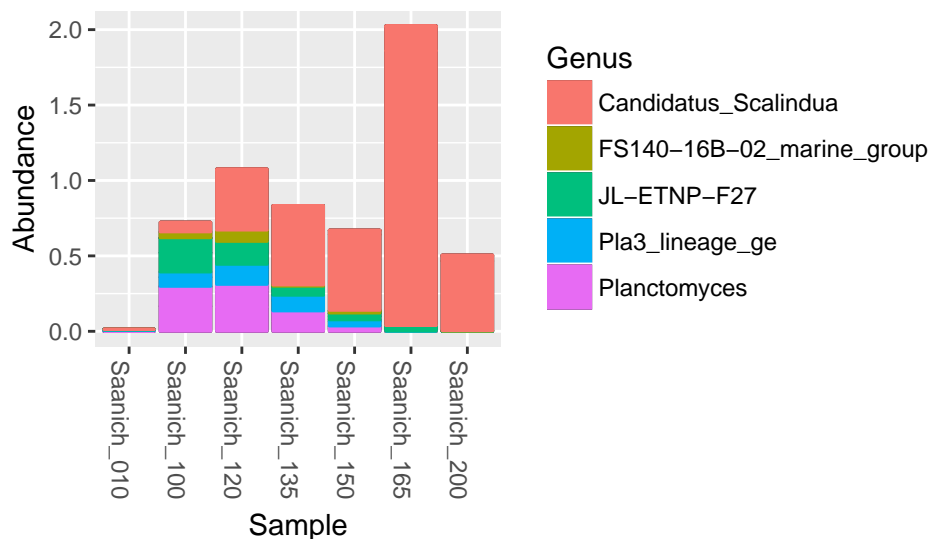
Table 1: Phylums from Mothur vs Phylums from QIIME2

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	

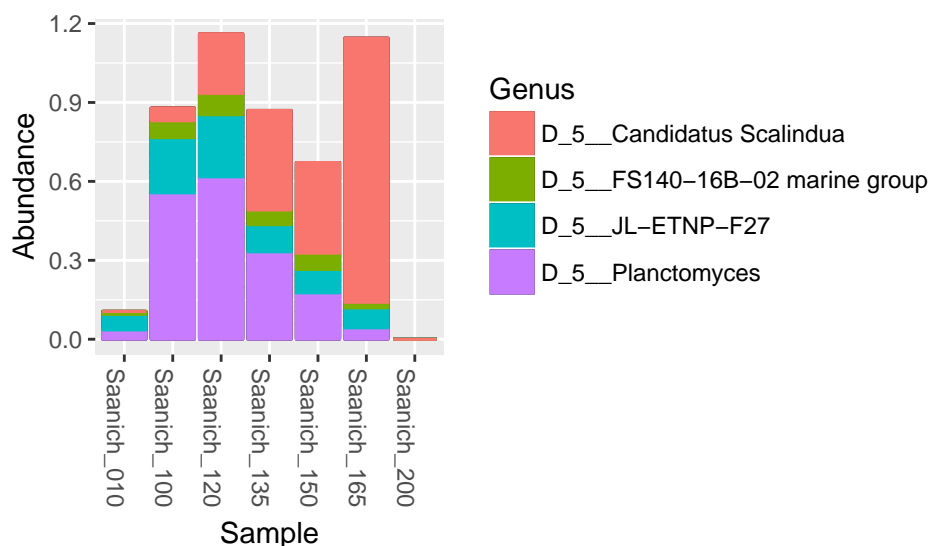


Phylums from Mothur	Phylums from QIIME2
Euryarchaeota	
PAUC34f	
Woesearchaeota_(DHVEG-6)	
Gracilibacteria	
Parcubacteria	

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, Marinimicrobia, Planctomycetes, and Verrucomicrobia were observed as depth-independent. In the QIIME2 data, there were much less phyla presented (4 types of phyla) in comparison with the Mothur data (7 types of phyla). The abundance of Bacteroidetes seemed to be decreasing from depth 10m to 200m. The abundance of Thaumarchaeota 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.



mothur



QIIME2

**Figure 5: Genus distribution of Planctomycetes across samples**

The well known general linear model is employed to recover relationships that might be exhibited between explanatory and target variables. We decompose our hypothesize in a series of experimental tests: i)- First, we investigate the correlation of Planctomyces abundance as a function of depth; and ii)- Then, we cross-examine the Planctomyces abundance as a function of oxygen concentration.

From the analyses, there were 4 OTUs within the Planctomyces genus as shown in Table 4. 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. The abundances of these OTUs also did not change significantly with oxygen concentration. 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.

Table 2: Correlation data of OTUs within Planctomyces genus across depth and oxygen 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

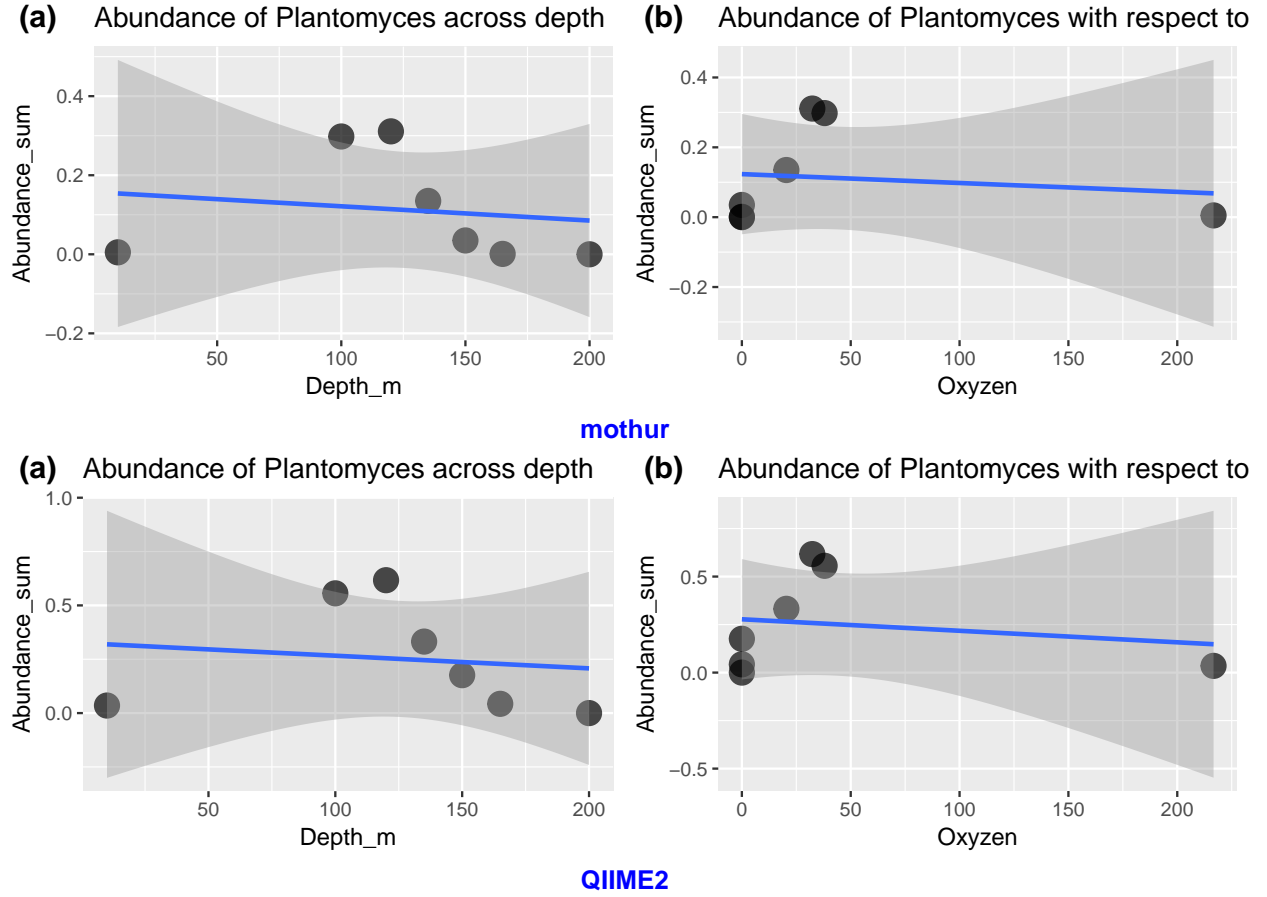


Figure 6: Regression analysis of Planctomyces across depth

#### 4.3 Estimate richness (number of OTUs/ASVs) for Planctomyces

Table 3: OTUs from Mothur vs ASVs from QIIME2

OTUs from Mothur	ASVs from QIIME2
Otu0125	Asv232
Otu0144	Asv799

OTUs from Mothur	ASVs from QIIME2
Otu0401	Asv1021
Otu0592	Asv1124

We explore the diversity of *Planctomyces* across depth.

[SIMILAR to BOTH PARTS WRITE and EDIT; Consider the abundance information in describing the correlations and shannons diveristy index]

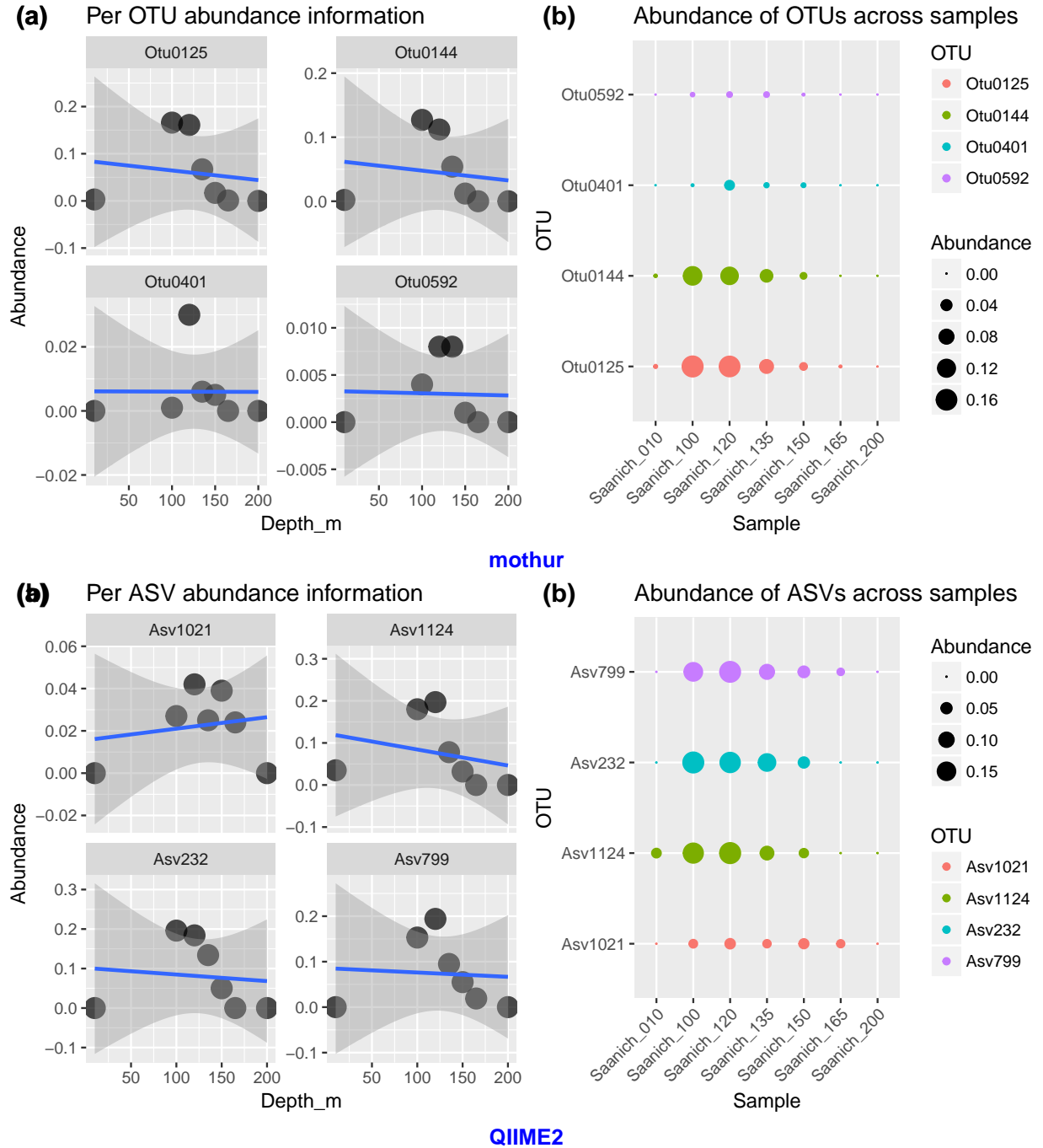
#### 4.4 Interpretation of abundance information of OTUs/ASVs of *Planctomyces* along with depth and/or oxygen concentration

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

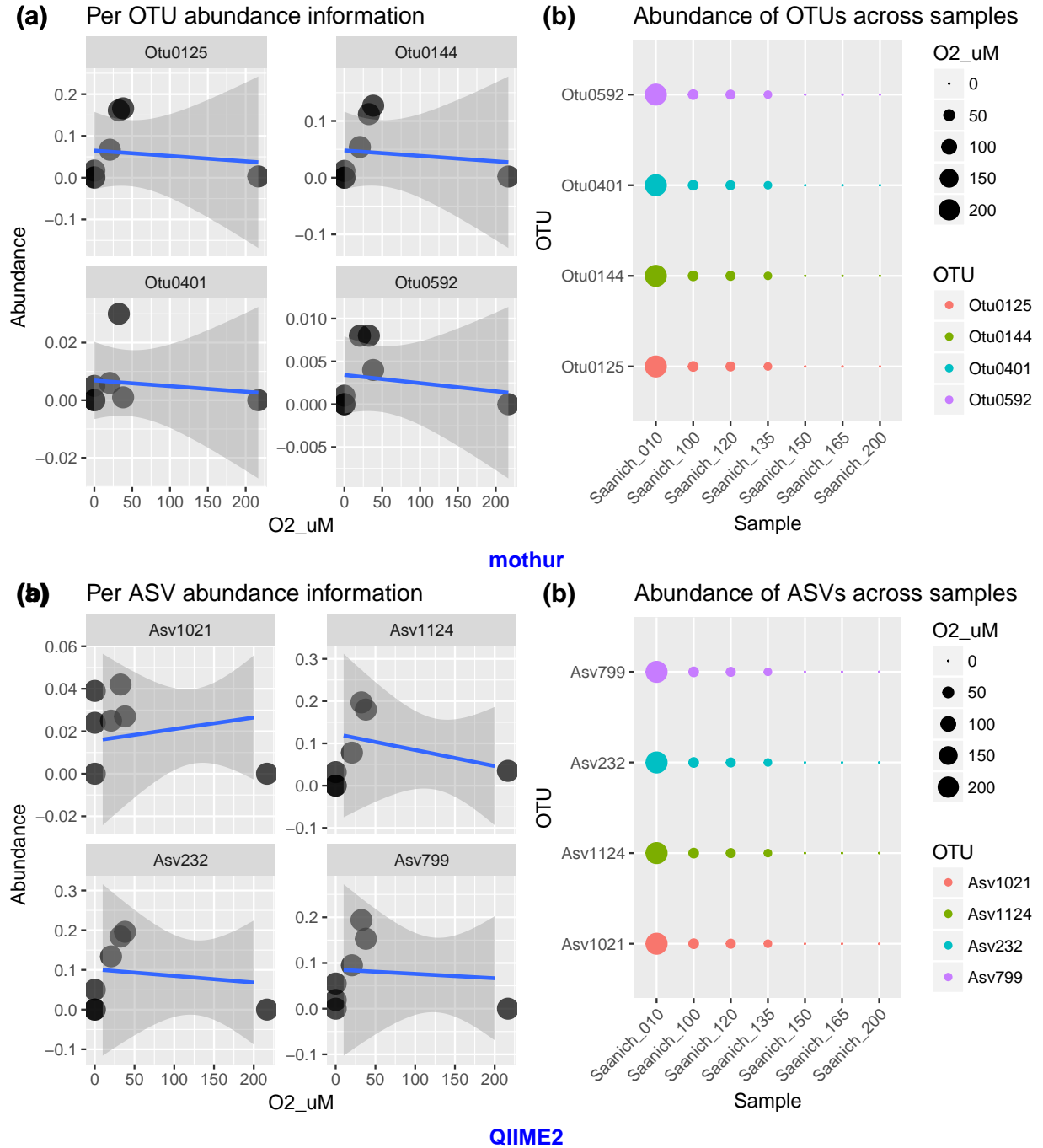
Covariates	Estimate	Std. Error	t-value	Pr(> t )
Otu0125 (mothur)	-0.0002045	0.0005479	-0.3731784	0.7243139
Otu0144 (mothur)	-0.0001533	0.0004035	-0.3798581	0.7196506
Otu0401 (mothur)	-0.0000009	0.0000807	-0.0113619	0.9913741
Otu0592 (mothur)	-0.0000023	0.0000274	-0.0836392	0.9365887
Asv232 (QIIME2)	-0.0001665	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

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



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



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

To summarize, there were 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)** An extra genus was present in the genus distribution of the Planctomycetes phylum across samples from Mothur: Pla3\_lineage\_ge (Fig. 4). **3)** In terms of abundance, the genus distribution of Planctomyces in QIIME2 data was higher than Mothur (Figure 4). Specifically, in Fig. 4 the QIIME2 data on Planctomyces (purple) almost doubled the number when compared to the Mothur data on Planctomyces (orange). **4)**

The scaling of abundance (Fig. 4) 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).

## 5 Discussion

Planctomyces, also known as anammox bacteria, are known to perform an important process in the nitrogen cycle: Anammox ( $\text{NO}_2^- + \text{NH}_4^+ \rightarrow \text{N}_2 + 2 \text{H}_2\text{O}$ ). 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 (21).

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 (21); it was reported that anammox bacteria were very sensitive to oxygen and that oxygen concentrations as little as 1  $\mu\text{M}$  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 (21). 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 (22). The respective cell numbers were in the range  $1.1\text{-}6.7 \times 10^7$  cells  $\text{g}^{-1}$  of wet peat, comprising 2-14% of total bacterial cells, and displaying linear correlation to the peat water pH (22). This result contradicted with the research study done by Koo et al, which observed the decline of Planctomycetes population at the oxic part of the ocean (21). 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 (22). 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 (24). 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 (23). 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 (24).

**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 (24).

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 extracted from this dataset for more in-depth exploration in 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 between the other genera that are present in both the Mothur and QIIME2 data.

---

## 6 References