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

March 27, 2018

Abstract

This is the abstract. It consists of two paragraphs.

Contents

---

1 Introduction

[Talk about microbial community deriving the most biogeochemical forces with citations]. The

wide diversity of microbes present in the different ecosystem implies that an almost infinite number of individuals needs to be identified to accurately describe such communities. Nevertheless, the advancements in analyzing high-throughput marker-gene sequencing data made it feasible to construct molecular operational taxonomic units (OTUs) through clustering the sequencing reads using a variety of dissimilarity distance methods (Chen et al. 2013). Another compelling approach to cluster group of reads is based on amplicon sequence variants (ASVs) (Callahan, McMurdie, and Holmes 2017).

We resort to studying microbial composition obtained from Saanich Inlet oxygen minimum zone (OMZ). Saanich Inlet is a seasonally anoxic fjord on the coast of Vancouver Island, British Columbia, Canada (Hawley et al. 2017; Torres-Beltrán et al. 2017). Most of the year, the fjord has an anoxic basin. At the end of summer, oxic waters flow into the basin, resulting in renewing the oxygen. From an OMZ research perspective, the Saanich Inlet is considered important as it provides an opportunity to study microbial ecology and various nutrient cycling in OMZs particularly under oxic-anoxic shifts (Hallam, Torres-Beltrán, and Hawley 2017) [You might add additional information]

Planctomyces is .... [Why did you choose this genus why is it important to study?]

To this end, we examined microbial diversity in Saanich Inlet dataset that was preprocess using mothur(Schloss et al. 2009) and QIIME2[I didn't find citation]. While mothur uses [Talk about OTU here], QIIME2 produces ASVs [give a brief description about ASVs]. [Talk about differences between these two methods].

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

The remaining of the paper is organized as follows. Section 2 describes the problem statements. Followed by [FILL] in Section 3. Finally, we summarize our contributions in Section 4.

2 Problem Formulation

[Talk in depth about Planctomyces]

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 Dataset

[Talk about Saanich Inlet dataset and various properties as documented in (Hawley et al. 2017; Torres-Beltrán et al. 2017)]

3.3 Methods

3.3.1 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 (Finotello, Mastrorilli, and Di Camillo 2016). 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 (Finotello, Mastrorilli, and Di Camillo 2016). However, in our experiments, the Chao1 method did not have any effects.

3.3.2 General Linear Model

General linear model (LM) (Hastie, Tibshirani, and Friedman 2009) 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 (OLS). However, to test the significance θ_1, we formulate hypothesis testing: specify the null and alternative hypotheses, specify an arbitrary cutoff probability value for γ from t-Student distribution, and determine the weight of evidence for rejecting the null hypothesis. This weight, given in terms of a probability, is referred to as the level of significance (or p-value) of the statistical test.

The usual null hypothesis for inference about θ_1 is $H_0 : \theta_1 > \gamma$ which asserts that no additional predictive value over and above, contributed by θ_1. While $H_1 : \theta_1 \leq \gamma$ measures whether x_j has additional predictive strengths. If the p-value for θ_1 is below or equal to γ then we accept H_1; otherwise, accept H_0.

3.4 Data Preporocessing

We used the Saanich Inlet dataset that was preprocessed using mothur and QIIME2. Afterward, samples were rarefied/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; ii)- prune samples and OTUs with unknown values, such as `unclassified` value; and iii)- any phylum fail to have more than 5 OTUs should be trimmed. This has resulted in 371 and 190 taxa from mothur and QIIME2, respectively. No other preprocessing were applied. The implementations are done entirely using R and relied on some efficient third-party libraries, such as `phyloseq` and `tidyverse`.

4 Results

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

Motivated by the recent report (Breitburg et al. 2018) regarding the oxygen depletion in the global in the global ocean, we analyze [the contribution of microbes and their existence along .... Fill]. Hereby, we try to understand the compositional complexity of a microbial community across Saanich Inlet samples. For this, we use Shannon's diversity index (SDI), which considers both the species abundances and the total number of distinct species in its diversity estimation. Figures 1(a) and 2(a) depicts Shannon's diversity index for mothur and QIIME2 datasets. Immediately, we observe that SDI values peak at depth 10 and 100 before monotonically decreasing at 200. The SDI values are maximal when the microbes are evenly distributed. Indeed, Figure 3 supports our claim, and we see an uneven distribution of Phylum at 200 more than at 10 or 100 depths. However, the Shannon index values do not capture the number of different species or richness varying across depths. Instead, Figures 1(a) and 2(a) shed some light in this regard.

Similarly, by analyzing the association between oxygen concentration within a microbial community, we observe, as in Figures 1(b) and 2(b), that SDI values increase when oxygen become more abundant. This is not surprising since the abundance of oxygen indicates..... [FILL]. The boxplots in Figures 1(c) and 2(c) supports this evidence too. Another compleing observation that follows the work in (Breitburg et al. 2018; Hawley et al. 2017) is the oxygen depletion along the water columns, as illustrated in Figures 1(d) and 2(d).

[ADD and EDIT]

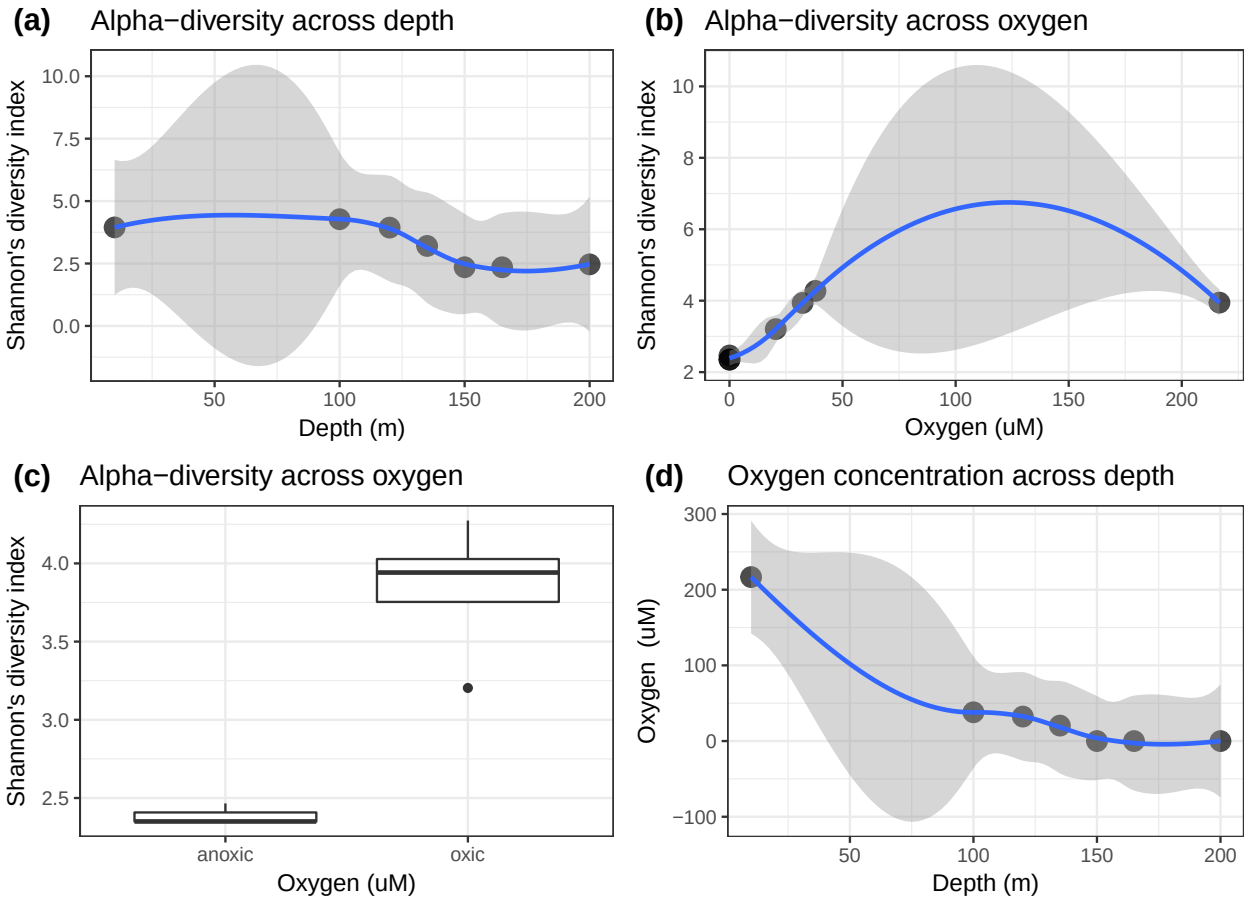


Figure 1: Mothur

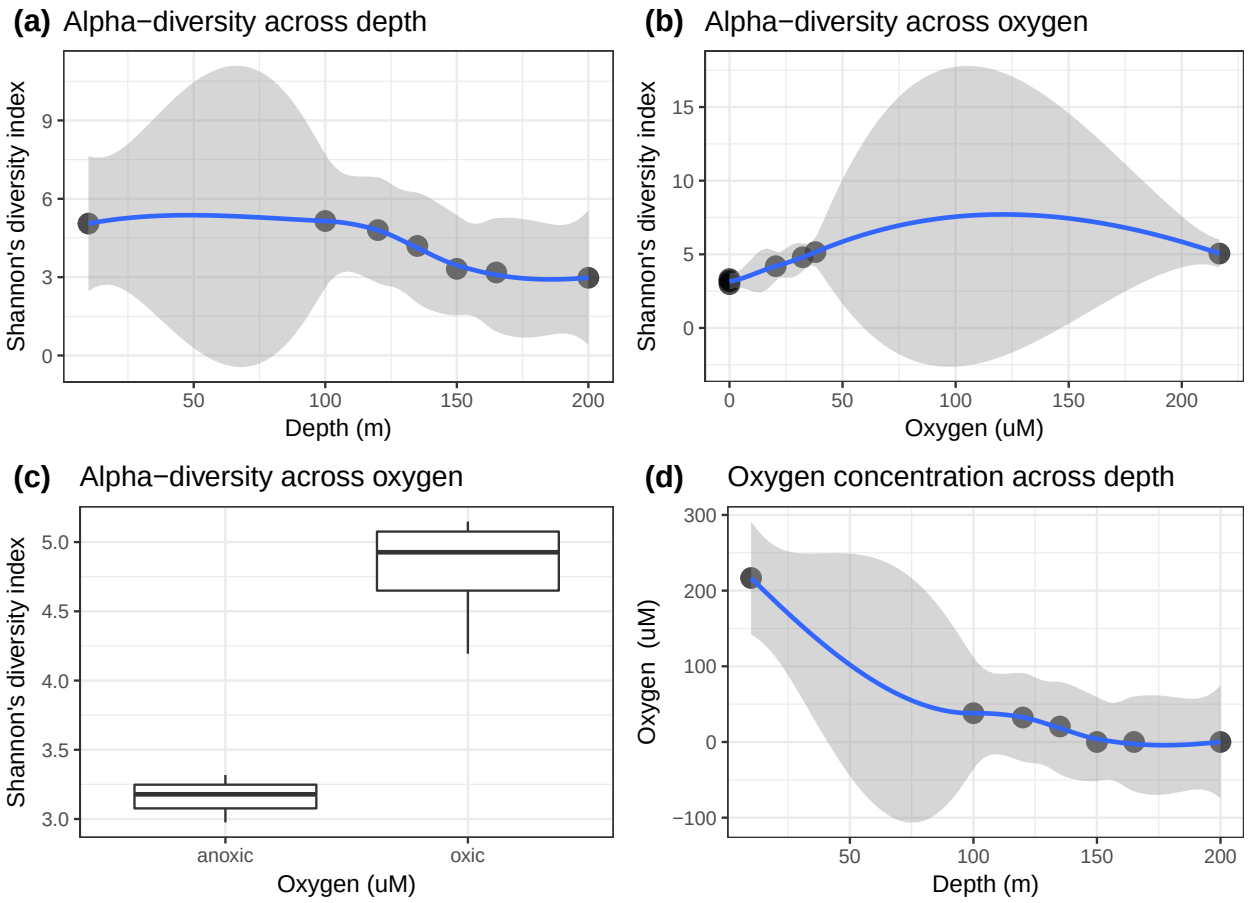


Figure 2: QIIME2

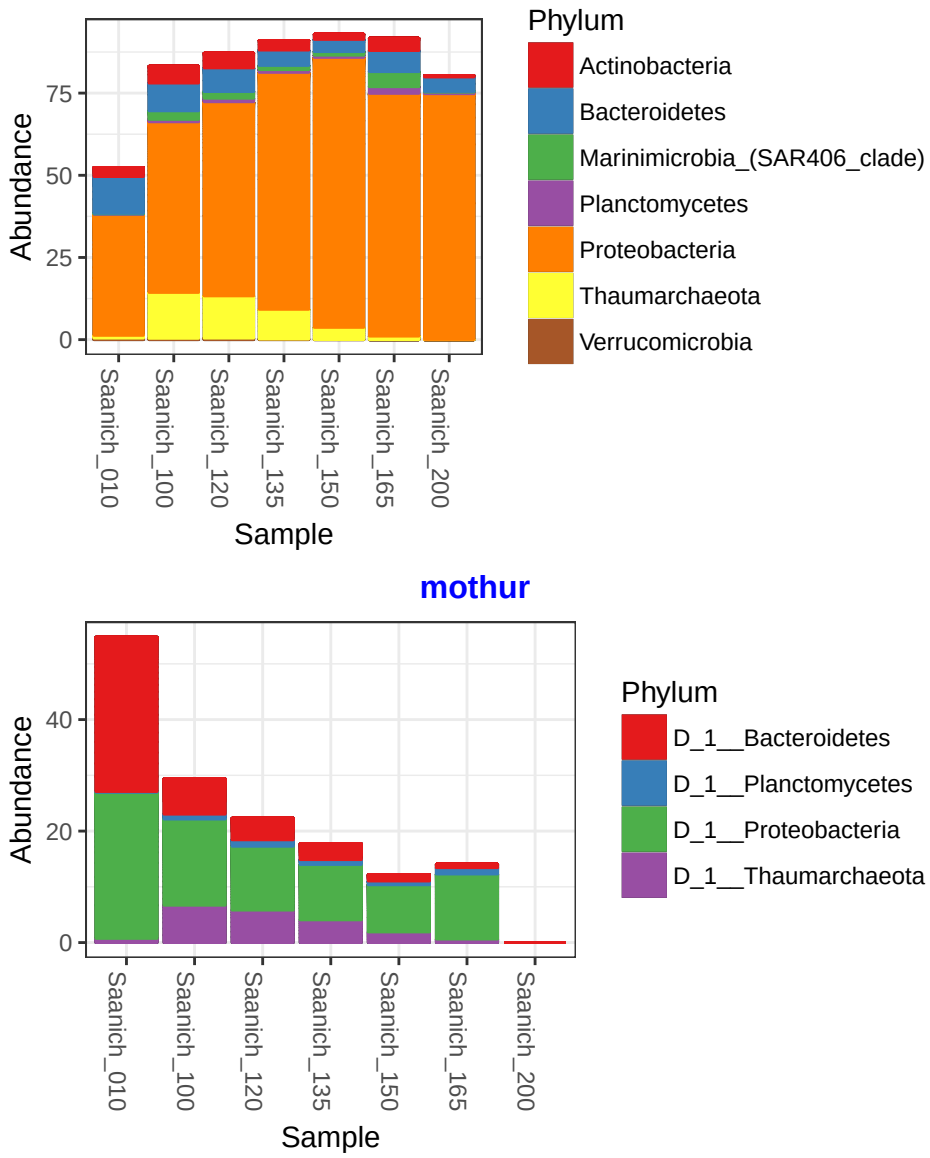


Figure 3: Phylum distribution across samples

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)	

Phylums from Mothur	Phylums from QIIME2
Planctomycetes	
Verrucomicrobia	

Based on our previous analysis and from Figure 3, it is clear that the microbial species and their distributions in Saanich inlet samples do indeed vary. Table 1 summarizes the phyla in both mothur and QIIME2 for Saanich Inlet dataset. We see that the number of hypothetical phyla present in mothur is estimated to be 7. Throught intesive investigation, we find that the *Planctomycetes* phylum is differentially presented across samples. And, after further exploring its genus distribution, as depicted in Figure 4, we exhibit that *Planctomyces* genus indeed are unevenly distributed across water columns. This is not coincidence because the [Fill based on some paper]. We intitally hypothesize that *Planctomyces* would be represented differently across depths, and for this, we performed regression tests.

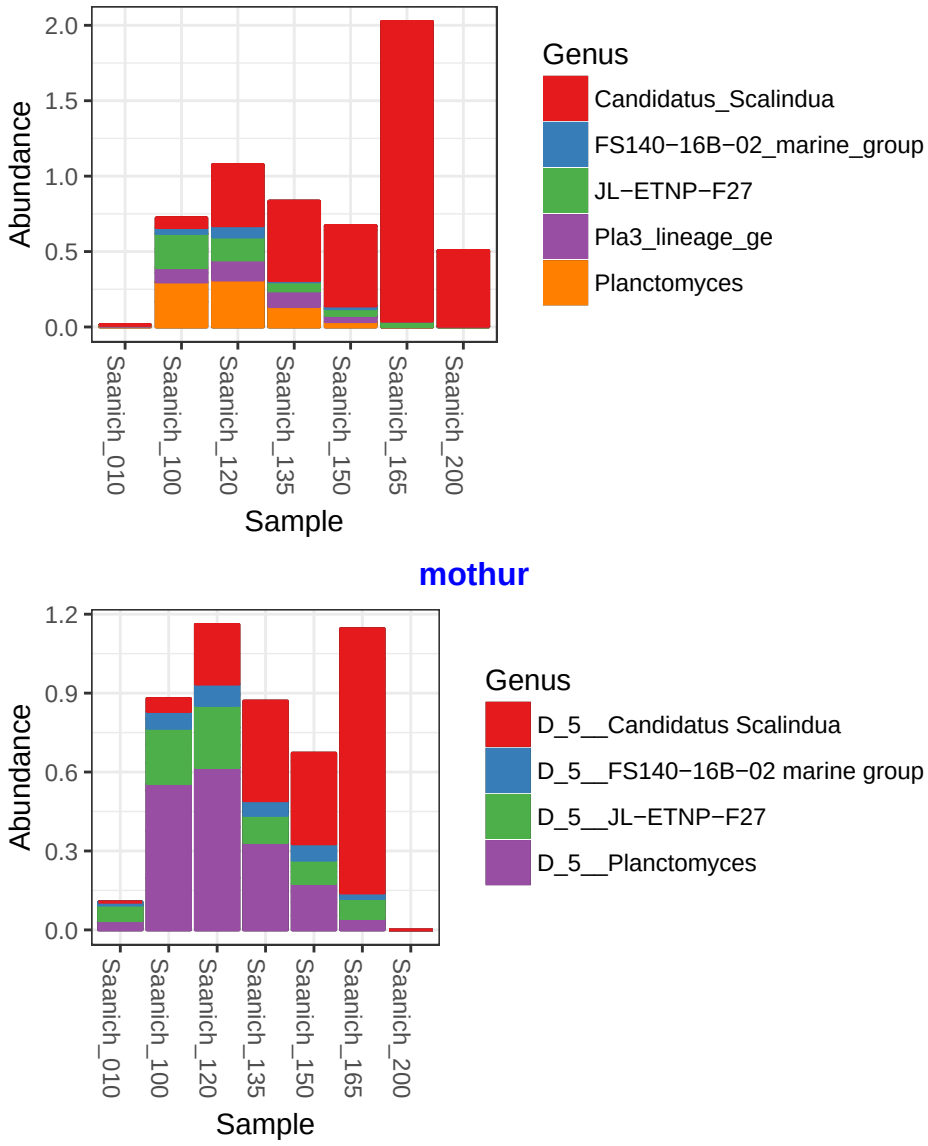


Figure 4: Genus distribution of Planctomycetes across samples

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.

Table 2. shows the results of our test analysis. From the statistical perspective, it can be inferred that there might be no relations with either the depth or the oxyzen. This is becasue the coefficients of depth and oxygen and their p-values were found to be (-0.0003609, p-value $= 0.7385598$) and (-0.0002544, p-value$= 0.7616253$), respectively, which are not statistically significant at 5% (an arbitrary cutoff). Hence, we might reason that there is a little statistical evidence to support our belief that Planctomyces indeed varies across depth and oxygen.

Such contradictory conclusion suggests to accept the null hyptothesis that states no intersting patterns exist for Planctomyces. However, the fitting problem associated with the general linear

model underestimate the existence of any kind of interesting relationships. Indeed, when we manually inspected the samples, we found that samples from depth $10, 150, 165$ and 200 do not or are less planctomyces abundant than at $100, 120$, and 135 depths, which imply that Planctomyces is quite differentially abundant in Saanich Inlet dataset. Perhaps, using more complex models might provide a better predictive analysis; but for now on, we stick with the statistical outputs.

[WRITE and EDIT]

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

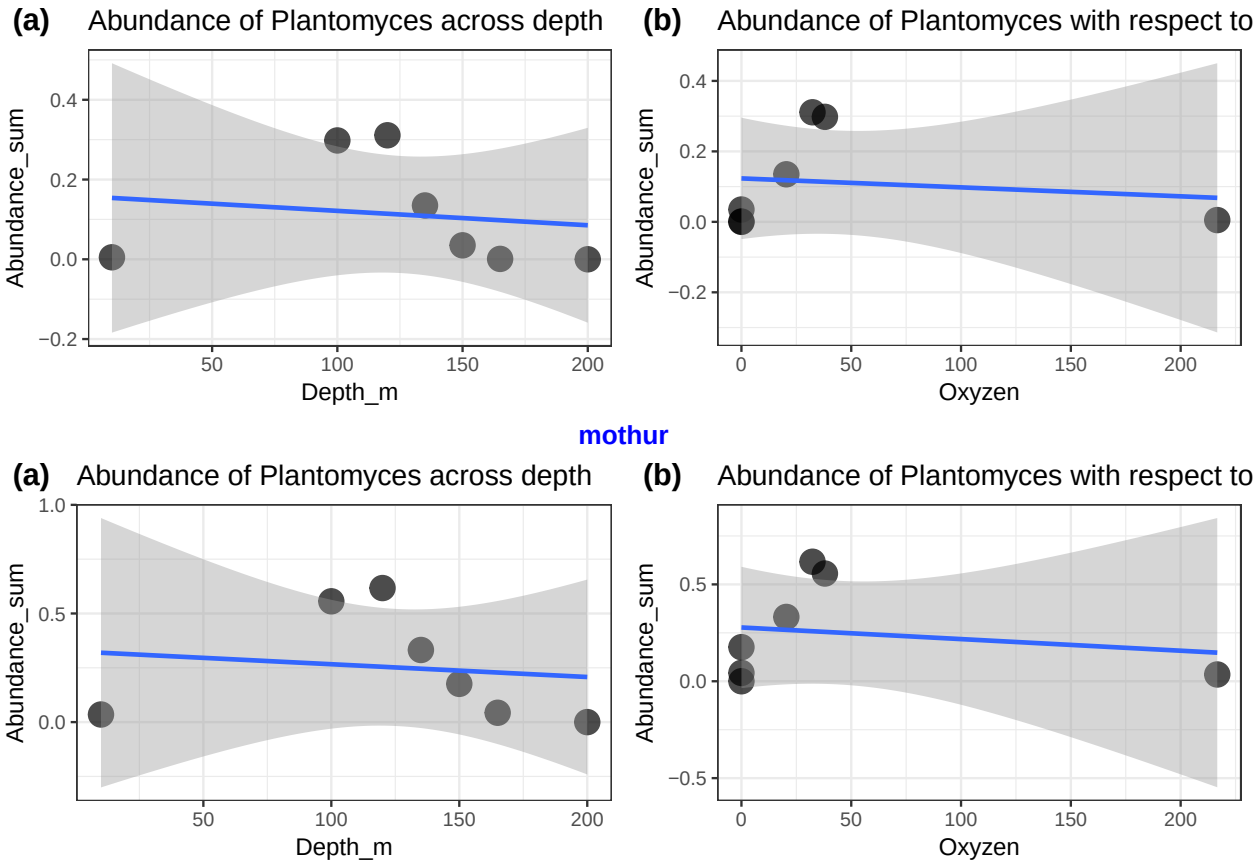


Figure 5: Regression analysis of Planctomyces across depth

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
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

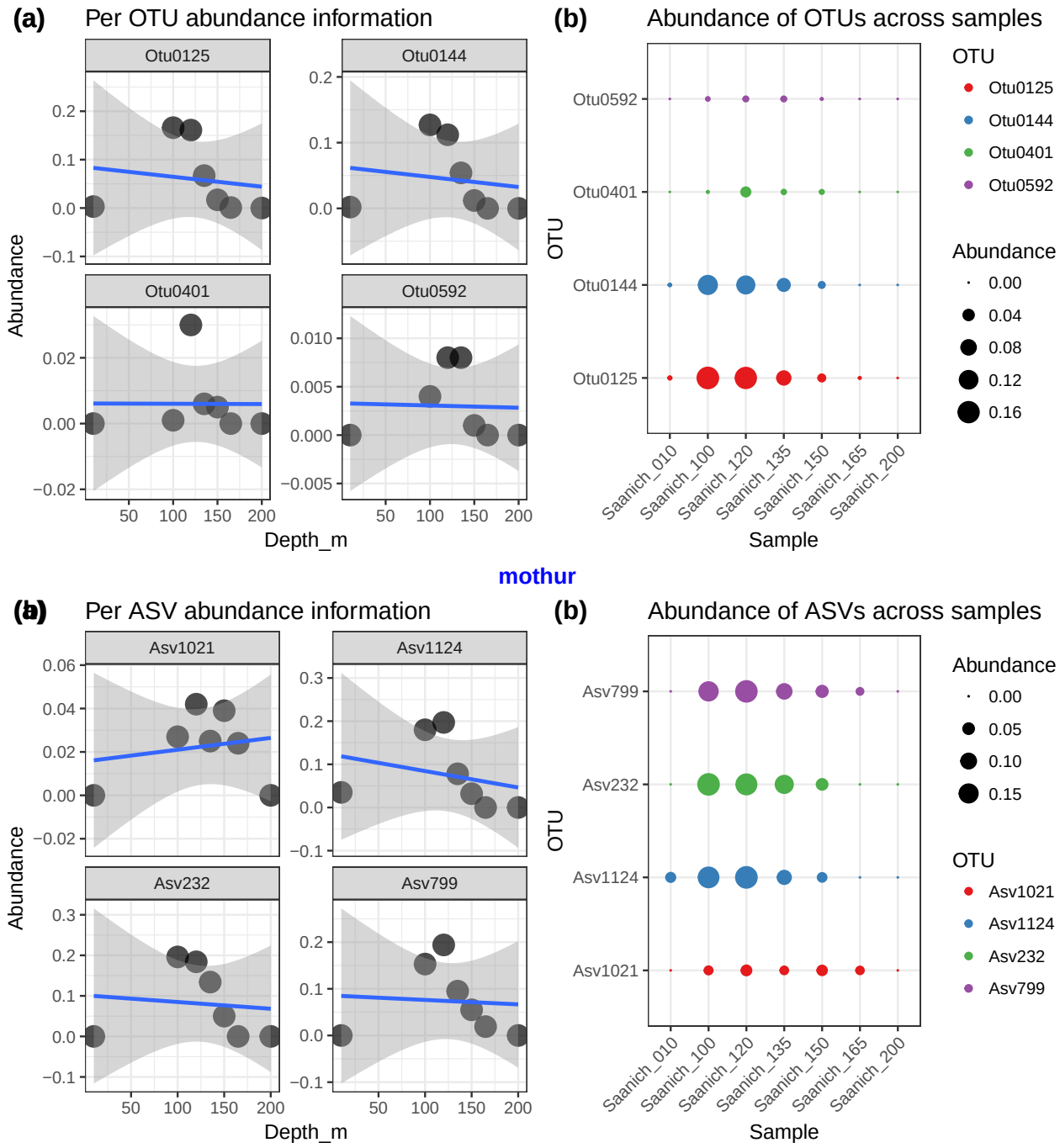
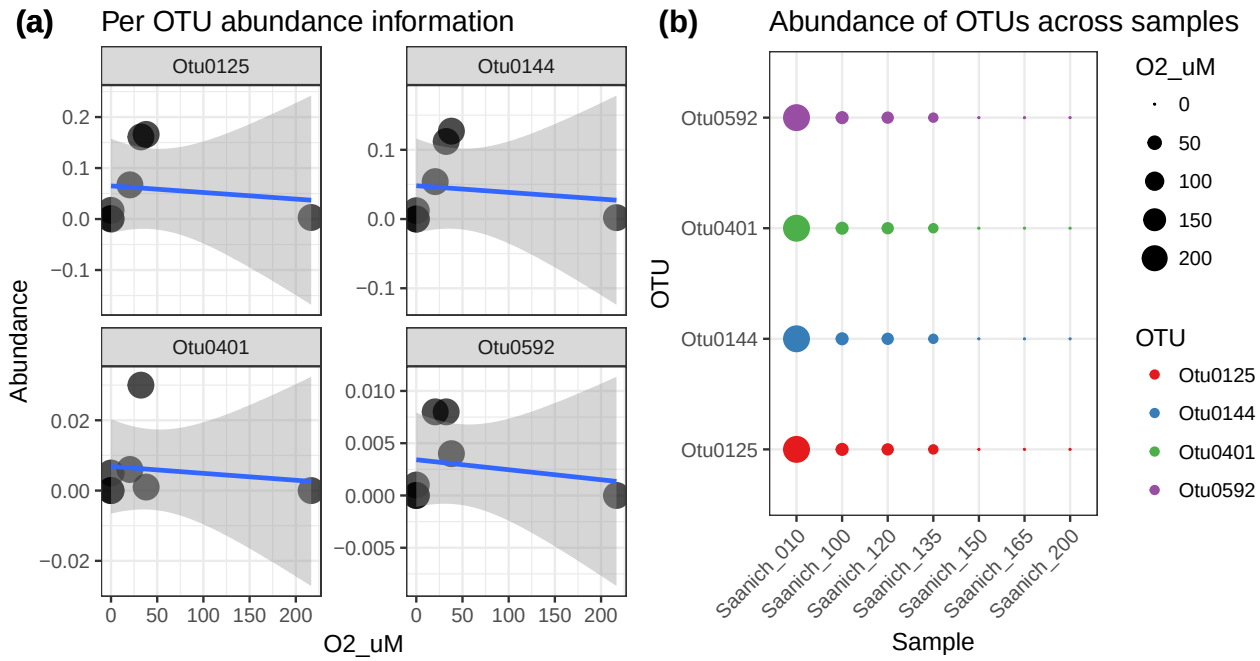
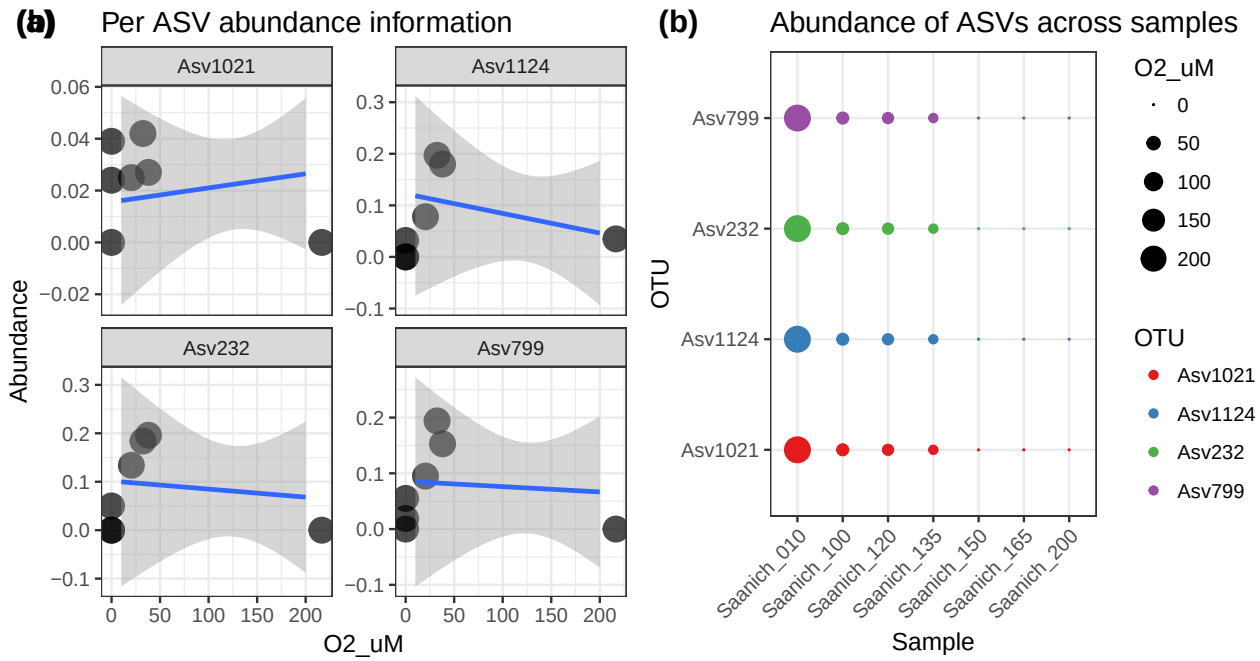


Figure 6: Abundance of OTUs/ASVs within Planctomyces genus across depth



mothur



QIIME2

Figure 7: Abundance of OTUs/ASVs within Planctomyces genus across oxygen concentration

5 Discussion

References

Breitburg, Denise, Lisa A Levin, Andreas Oschlies, Marilaure Grégoire, Francisco P Chavez, Daniel J Conley, Véronique Garçon, et al. 2018. "Declining Oxygen in the Global Ocean and Coastal Waters." *Science* 359 (6371). American Association for the Advancement of Science: eaam7240.

Callahan, Benjamin J, Paul J McMurdie, and Susan P Holmes. 2017. "Exact Sequence Variants Should Replace Operational Taxonomic Units in Marker-Gene Data Analysis." *The ISME Journal* 11 (12). Nature Publishing Group: 2639.

Chen, Wei, Clarence K Zhang, Yongmei Cheng, Shaowu Zhang, and Hongyu Zhao. 2013. "A Comparison of Methods for Clustering 16s rRNA Sequences into Otus." *PloS One* 8 (8). Public Library of Science: e70837.

Finotello, Francesca, Eleonora Mastrorilli, and Barbara Di Camillo. 2016. "Measuring the Diversity of the Human Microbiota with Targeted Next-Generation Sequencing." *Briefings in Bioinformatics*. Oxford University Press, bbw119.

Hallam, Steven J, Mónica Torres-Beltrán, and Alyse K Hawley. 2017. "Monitoring Microbial Responses to Ocean Deoxygenation in a Model Oxygen Minimum Zone." *Scientific Data* 4. Nature Publishing Group.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. Springer.

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.

Schloss, Patrick D, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23). Am Soc Microbiol: 7537–41.

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.