Analysis of Candidatus scalindua data from Saanich Inlet using mothur and QIIME2

J. Atif, L. Chang, J. Gill, J. Huggins, L. Sonderegger, P. Suresh

version March 28, 2018

Table of Contents

# **Introduction**

Oxygen minimum zones (OMZs) are areas in the ocean in which oxygen concentrations are lower than 20 µM. They are important for understand the nitrogen cycle and the role that the ocean plays with respect to atmospheric greenhouse gases (2). As a consequence of climate change and decreasing oxygen solubility, OMZs have become more widespread. Oxygen deficiency leads to decreased productivity of aerobic organisms and favours chemolithotrophs, which results in the production of several greenhouse gases (3). Saanich Inlet, located on the coast of Vancouver Island, is a seasonally anoxic fjord and is a model ecosystem for monitoring biogeochemical responses to changing oxygen levels (4). In spring and summer, a rise in primary productivity in surface waters combined with low levels of vertical mixing in the waters below results in anoxia and accumulation of methanol, ammonia, and hydrogen sulfide (47). During late summer and autumn, neap tidal flows bring in denser water from the Northeastern subarctic Pacific, which supplies the lower basin water with oxygen again. This pattern of anoxia followed by oxygen renewal is what makes Saanich Inlet a useful system in the study of biogeochemical shifts in response to changing oxygen (4).

Since the rise of next-generation sequencing technologies, amplicon sequencing is widely used to identify microorganisms as well as to infer relationships between them. In this analysis, the V4-5 regions of prokaryotic 16S rRNA (the small subunit of ribosomal RNA) was used. Once sequenced, the data can be classified into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs).

OTUs are defined as clusters of reads that differ by less than a fixed sequence dissimilarity, which is commonly set at 3% (1). This can be done either by de-novo clustering or closed-reference assignment. In closed-reference assignment, reads that fit within a closed range similar to a reference sequence are placed into one OTU (1). A drawback of this is that some reads get “clipped” into two OTUs even though they appear to be the same species. In de-novo clustering, reads are grouped into OTUs based on how similar they are to each other, not based on a reference sequence (1). Some drawbacks to this include the fact that different species have similar 16S sequences and situations where decisions have to be made about several clusters appearing to overlap.

ASVs have been recently introduced, and work by inferring the biological sequences in the sample prior to amplification and sequencing errors (1). In this case, everything that is not identical is unused in classification. A drawback of this method is that discarding data can lead to bias toward low-error sequences, which may not be the best representation of the cluster. Nevertheless, this method is gaining traction with its sensitivity and high resolution in distinguishing species (1).

This analysis is done to show how using different pipelines affects the end results. Using both mothur (OTU)- and QIIME2 (ASV)-processed data, the Saanich inlet data was studied to determine how changing depth and oxygen levels impacts community structure, abundance, and richness. After analysis was complete, the answers given by the separate pipelines were compared. The taxon of interest chosen after analyzing the processed mothur and QIIME2 was Candidatus Scalindua. This taxon was chosen after plotting the abundances of several taxa and seeing which ones varied by depth, which would likely be due to oxygen. The genus *Candidatus* Scalindua encompasses more than 5 OTUs and ASVs, and consistently increases with increasing depth. Interestingly, they are known to be annamox bacteria that thrive in oxygen-minimum zones (5). For these reasons, this taxon was chosen to pursue further.

# **Methods**

## Sample Collection and Processing

### Sample Processing and Sequencing

The sequence data was derived from Saanich Inlet Cruise 72. Specific target genes are amplified and sequenced in order to study microbial communities within a community. The small subunit ribosomal RNA (SSU rRNA) is the gene of interest as it is a universally conserved marker and therefore is useful for comparing microbial communities (4), and for this analysis the V4-5 region of the 16S rRNA was used. Samples for large volume SSU rRNA gene tags were collected from six major depths across the oxycline (4), while samples for high resolution SSU rRNA gene tag sequencing were taken from 16 depths across the oxycline. Genomic DNA was extracted from the collected samples (4). PCR amplification using 515F and 808R primers was then performed on the extracted genomic DNA to generate SSU rRNA gene libraries via (4). The samples were then sequenced using MiSeq system with Phred 33 quality scores. The Phred33 scores indicate the estimated probability of an error. The validity and quality of the SSU rRNA amplicons were determined Gel electrophoresis was performed on SSU rRNA gene amplicons to check for molecular weight and possible degradation (4). Sequences were then processed using either mother(6) or QIIME2(7).

### Data Analysis in mothur and QIIME2

The mothur software was used for further data processing . The data was cleaned up to identify potential errors and remove low quality sequences (6). Singletons were removed as sequences that only occur once throughout the entire data set are most likely due to an error. Sequences with less than 4 base pair differences are clustered together in order to reduce sequencing error and useless alignment data are removed (6). Then, operational taxonomic units (OTUs) are determined if clustered sequences display 97% similarity (6). The OTUs are then classified by first classifying all the sequences using the SILVA database. The taxonomies obtained from this database are then condensed for each OTU.

In the QIIME2 software the data is cleaned up and amplicon sequence variants (ASVs) are determined using the Dada2 protocol. The ASVS were then classified using the Silva version 119 database at a 99% similarity (7).

## Statistical Analysis of Relative Abundance Data

### Data cleaning

All statistical analysis of the data generated above was completed in R v3.4.3 (8) using the following packages:

library("tidyverse")

## -- Attaching packages -------------------------------------- tidyverse 1.2.1 --

## v ggplot2 2.2.1 v purrr 0.2.4  
## v tibble 1.4.2 v dplyr 0.7.4  
## v tidyr 0.7.2 v stringr 1.2.0  
## v readr 1.1.1 v forcats 0.2.0

## -- Conflicts ----------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library("magrittr")

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:purrr':  
##   
## set\_names

## The following object is masked from 'package:tidyr':  
##   
## extract

library("phyloseq")

After loading the data into R and before further analysis, a random seed was set to ensure reproducibility of our results. The Samples were then rarefied to 100,000 sequences per sample to allow the comparison of taxon relative abundance between samples.

load("~/MICB425\_portfolio/mothur\_phyloseq.RData")  
load("~/MICB425\_portfolio/qiime2\_phyloseq.RData")  
  
set.seed(9376)  
  
#variables starting with m: correspond to mothur data  
#variables starting with q; correspong to QIIME2 data  
  
m.norm = rarefy\_even\_depth(mothur, sample.size=100000)

## You set `rngseed` to FALSE. Make sure you've set & recorded  
## the random seed of your session for reproducibility.  
## See `?set.seed`

## ...

## 626OTUs were removed because they are no longer   
## present in any sample after random subsampling

## ...

q.norm = rarefy\_even\_depth(qiime2, sample.size=100000)

## You set `rngseed` to FALSE. Make sure you've set & recorded  
## the random seed of your session for reproducibility.  
## See `?set.seed`  
##   
## ...

## 11OTUs were removed because they are no longer   
## present in any sample after random subsampling

## ...

### Linear Model and F-test for testing significance of regression

Relative abundance percentages were calculated using the rarefied counts for each sample. This data was then piped into linear models and statistical testing using the magrittr package.

m.perc = transform\_sample\_counts(m.norm, function(x) 100 \* x/sum(x))  
q.perc = transform\_sample\_counts(q.norm, function(x) 100 \* x/sum(x))

For most of the analysis, the simple linear regression method was used. Briefly, a linear model was created to measure the strength of the relationship between relative taxon abundance and individual environmental variable within each sample. An F-test for testing significance of regression was then used to test the significance of the linear regression model in predicting the effect of the environmental variable on taxon relative abundance.

General example of computing the linear model:

#Summarizing the attributes of the linear model of Candidatus Scalindua relative abundance bundance against Depth   
m.norm %>%   
 subset\_taxa(Genus=="Candidatus\_Scalindua") %>%   
 tax\_glom(taxrank = 'Genus') %>%  
 psmelt() %>%  
   
 lm(Abundance ~ Depth\_m, .) %>%   
 summary()

##   
## Call:  
## lm(formula = Abundance ~ Depth\_m, data = .)  
##   
## Residuals:  
## 7 6 4 5 3 2 1   
## 911.5 550.3 -517.7 -733.2 -420.2 -498.9 708.2   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -849.823 705.010 -1.205 0.2820   
## Depth\_m 14.167 5.125 2.764 0.0396 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 757.1 on 5 degrees of freedom  
## Multiple R-squared: 0.6044, Adjusted R-squared: 0.5253   
## F-statistic: 7.64 on 1 and 5 DF, p-value: 0.03964

#Summarizing the attributes of the linear model of a specific OTU's relative abundance against Depth   
m.norm %>%   
 psmelt() %>%   
 filter(OTU=="Otu0242") %>%   
 lm(Abundance ~ Depth\_m, .) %>%   
 summary()

##   
## Call:  
## lm(formula = Abundance ~ Depth\_m, data = .)  
##   
## Residuals:  
## 1 2 3 4 5 6 7   
## 56.538 6.036 -2.715 -7.716 -19.213 -2.220 -30.711   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.7205 28.2965 0.025 0.981  
## Depth\_m 0.1500 0.2057 0.729 0.499  
##   
## Residual standard error: 30.39 on 5 degrees of freedom  
## Multiple R-squared: 0.09606, Adjusted R-squared: -0.08472   
## F-statistic: 0.5314 on 1 and 5 DF, p-value: 0.4987

# **Results**

## How does microbial community structure change with depth and oxygen concentration?

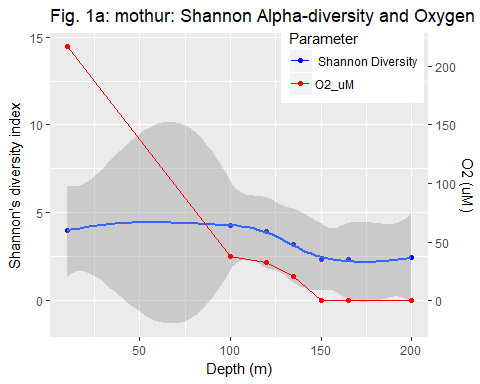
In R, the Alpha-diversity and richness were calculated for the total rarefied community in each sample for both mothur and QIIME2.

m.alpha = estimate\_richness(m.norm, measures = c("Chao1", "Shannon"))  
  
q.alpha = estimate\_richness(m.norm, measures = c("Chao1", "Shannon"))  
  
# Combine richness and alpha-diversity data with geochemical data:  
m.meta.alpha = full\_join(rownames\_to\_column(m.alpha), rownames\_to\_column(data.frame(m.perc@sam\_data)), by = "rowname")  
  
q.meta.alpha = full\_join(rownames\_to\_column(q.alpha), rownames\_to\_column(data.frame(m.perc@sam\_data)), by = "rowname")

By plotting the shannon alpha diversity and Oxygen concentration against depth for both mothur and QIIME2 data it is clear to see that the deiversity of taxons pesent decreases further down the water column (Figure 1a, 1b). The Shannon diversity predicted by using the QIIME2 data is higher at all depths than the diversity calculated by using the mothur data. Going down the watercolumn, a microbial diversity bottleneck is reached at a depth of 100m as the shannon alpha diversity fall suddenly.

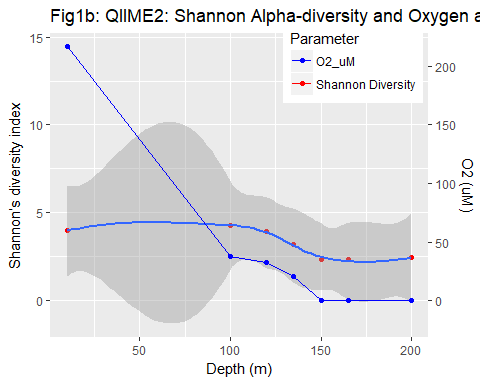
#Figure 1a  
m.meta.alpha %>%   
 ggplot() +  
 geom\_point(aes(x=Depth\_m, y=Shannon, colour= " Shannon Diversity")) +  
 geom\_smooth(method='auto', aes(x=as.numeric(Depth\_m), y=Shannon)) +  
 labs(title="Alpha-diversity across depth", y="Shannon's diversity index", x="Depth (m)") +  
 geom\_line(aes(x=Depth\_m, y=O2\_uM/15, colour="O2\_uM")) +   
 geom\_point(aes(x=Depth\_m, y=O2\_uM/15, colour="O2\_uM"))+  
 scale\_y\_continuous(sec.axis = sec\_axis(~.\*(15), name = "O2 (uM)")) +  
 scale\_colour\_manual(values = c("blue", "red"))+  
 labs(title="Fig. 1a: mothur: Shannon Alpha-diversity and Oxygen across depth", y = "Shannon's diversity index" , x = "Depth (m)" , colour = "Parameter") +  
 theme(legend.position = c(0.8, 0.9))

## `geom\_smooth()` using method = 'loess'



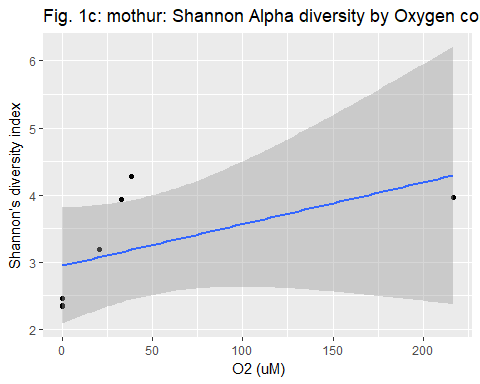
#Figure 1b:  
q.meta.alpha %>%   
 ggplot() +  
 geom\_point(aes(x=Depth\_m, y=Shannon, colour= "Shannon Diversity")) +  
 geom\_smooth(method='auto', aes(x=as.numeric(Depth\_m), y=Shannon)) +  
 labs(title="Alpha-diversity across depth", y="Shannon's diversity index", x="Depth (m)") +  
 geom\_line(aes(x=Depth\_m, y=O2\_uM/15, colour="O2\_uM")) +  
 geom\_point(aes(x=Depth\_m, y=O2\_uM/15, colour="O2\_uM"))+  
 scale\_y\_continuous(sec.axis = sec\_axis(~.\*(15), name = "O2 (uM)")) +  
 scale\_colour\_manual(values = c("blue", "red"))+  
 labs(title="Fig1b: QIIME2: Shannon Alpha-diversity and Oxygen across depth", y = "Shannon’s diversity index" , x = "Depth (m)" , colour = "Parameter") +  
 theme(legend.position = c(0.8, 0.9))

## `geom\_smooth()` using method = 'loess'

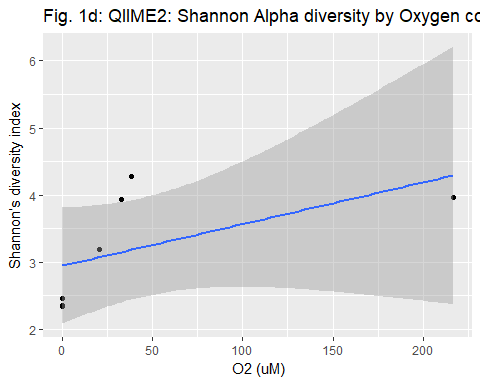


By plotting alpha diversity against the corresponding oxygen concentration of each sample and applying a simple linear regression to visualize the correlation between the two variables it is evident that despite microbial diversity decreasing with depth, the same strong trend is not seen for decreasing oxygen (Figure 1c, 1d).Regardless, there is a weak positive correlation between oxygen concentration and the corresponding alpha diversity. This pattern is similar for both the mothur and QIIME2 data.

#Figure 1c:  
m.meta.alpha %>%  
 ggplot() +  
 geom\_point(aes(x=O2\_uM, y=Shannon)) +  
 geom\_smooth(method='lm', aes(x=as.numeric(O2\_uM), y=Shannon)) +  
 labs(title="Fig. 1c: mothur: Shannon Alpha diversity by Oxygen concentration", y = "Shannon’s diversity index" , x =" O2 (uM)" )



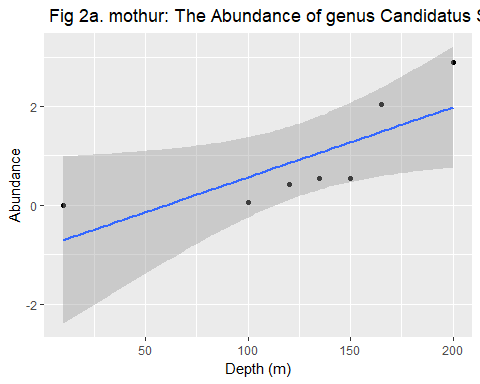
#Figure 1d:  
q.meta.alpha %>%  
 ggplot() +  
 geom\_point(aes(x=O2\_uM, y=Shannon)) +  
 geom\_smooth(method='lm', aes(x=as.numeric(O2\_uM), y=Shannon)) +  
 labs(title="Fig. 1d: QIIME2: Shannon Alpha diversity by Oxygen concentration", y = "Shannon’s diversity index" , x =" O2 (uM)" )



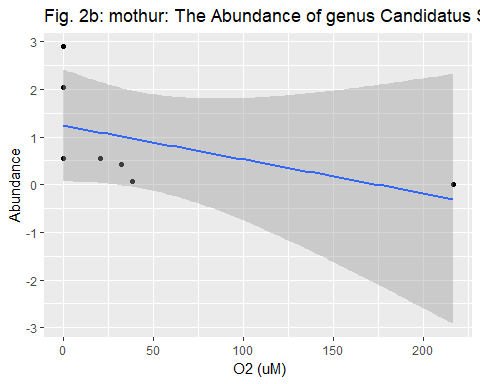
## Does the genus *Candidatus* Scalindua significantly differ in abundance with depth and oxygen concentration?

The genus *Candidatus* scalindua does differ significantly in abundance with depth uas the statistical test resulted in p values of 0.04 when using both mothur and QIIME2. On the other hand the this genus does not differ significantly in abundance with oxygen concentration using both mothur and QIIME2 as the p-values were greater than 0.05. The linear model for abundance with depth using mothur has a positive correlation with wide 95% confidence interval bands (Figure 2a). Whereas the linear model for abundance with oxygen concentration has a negative correlation with wide 95% confidence interval bands (Figure 2b). Similar to mothur, the linear model for abundance with depth using QIIME2 the 95% also has a positive correlation (Figure 2c). The linear model for the abundance of *Candidatus* scalindua across oxygen concentration using QIIME2 has a negative correlation (Figure 2d). It is evident that analysis of the data using both mother and QIIME 2 resulted in similar trends for the linear models. However, the 95% confidence intervals seem to be more narrower for the linear models that were generated using QIIME2 in comparison to the linear models that were generated using mothur.

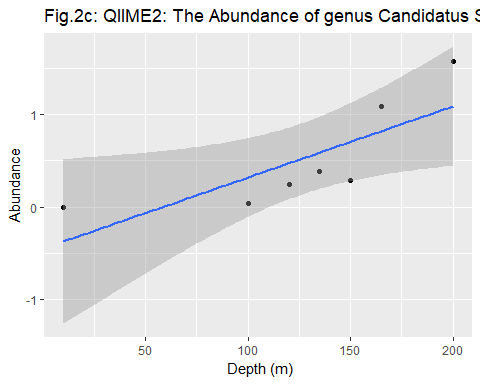
#Figure 2a  
m.perc %>%  
 subset\_taxa(Genus=="Candidatus\_Scalindua") %>%  
 psmelt() %>%  
 group\_by(Sample) %>%  
 summarize(Abundance\_sum=sum(Abundance), Depth\_m=mean(Depth\_m)) %>%  
   
 ggplot() +  
 geom\_point(aes(x=Depth\_m, y=Abundance\_sum)) +  
 geom\_smooth(method='lm', aes(x=as.numeric(Depth\_m), y=Abundance\_sum)) +  
 labs(title=" Fig 2a. mothur: The Abundance of genus Candidatus Scalindua across depth", y = "Abundance" , x = "Depth (m)" )



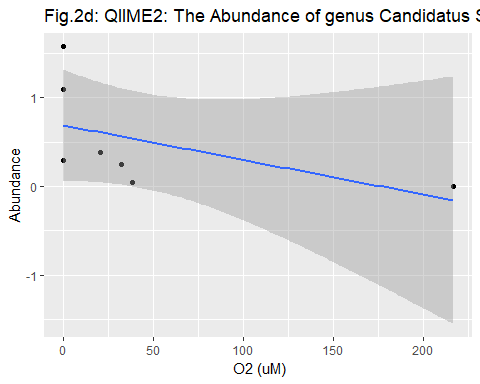
#Figure 2b  
m.perc %>%  
 subset\_taxa(Genus=="Candidatus\_Scalindua") %>%  
 psmelt() %>%  
 group\_by(Sample) %>%  
 summarize(Abundance\_sum=sum(Abundance), O2\_uM=mean(O2\_uM)) %>%  
   
 ggplot() +  
 geom\_point(aes(x=O2\_uM, y=Abundance\_sum)) +  
 geom\_smooth(method='lm', aes(x=as.numeric(O2\_uM), y=Abundance\_sum)) +  
 labs(title="Fig. 2b: mothur: The Abundance of genus Candidatus Scalindua across Oxygen conc.", y = "Abundance" , x =" O2 (uM)" )



#Figure 2c  
q.perc %>%  
 subset\_taxa(Genus=="D\_5\_\_Candidatus Scalindua") %>%  
 psmelt() %>%  
 group\_by(Sample) %>%  
 summarize(Abundance\_sum=sum(Abundance), Depth\_m=mean(Depth\_m)) %>%  
   
 ggplot() +  
 geom\_point(aes(x=Depth\_m, y=Abundance\_sum)) +  
 geom\_smooth(method='lm', aes(x=as.numeric(Depth\_m), y=Abundance\_sum)) +  
 labs(title="Fig.2c: QIIME2: The Abundance of genus Candidatus Scalindua across Depth", y = "Abundance" , x =" Depth (m)" )



#Figure 2d  
q.perc %>%  
 subset\_taxa(Genus=="D\_5\_\_Candidatus Scalindua") %>%  
 psmelt() %>%  
 group\_by(Sample) %>%  
 summarize(Abundance\_sum=sum(Abundance), O2\_uM=mean(O2\_uM)) %>%  
   
 ggplot() +  
 geom\_point(aes(x=O2\_uM, y=Abundance\_sum)) +  
 geom\_smooth(method='lm', aes(x=as.numeric(O2\_uM), y=Abundance\_sum)) +  
 labs(title="Fig.2d: QIIME2: The Abundance of genus Candidatus Scalindua across Oxygen conc.", y = "Abundance" , x =" O2 (uM)")



Interestingly, abundance of *Candidatus* scalindua across ammonium concentration using mothur resulted in a very low p-value of 0.000361 indicating that the difference is statistically significant.

## What is the richness of *Candidatus* Scalindua?

Across all samples, 24 OTUs were identified within *Candidatu*s scalindua using mother. Initially 30 OTUs were identified, however after rarefying the sample, only 24 OTUs remained. When analyzing the amount of OTUs from each sample, the richness of *Candidatus* scalindua seems to increase with depth. More OTUs are present in the sample obtained from depth 200m as opposed to 10m. A large increase in the number of OTUs is seen at depth of 165m and 200m. On the other hand, only 5 ASVs were identified within *Candidatus* scalindua across all samples using QIIME2. The richness of *Candidatus* scalindua does not differ greatly with depth in comparison to OTUs.

Detailed outputs of the richness outputs in the total dataset and across samples are shown below:

#Richness of Candidatus Scalindua in total dataset  
m.norm %>%  
 subset\_taxa(Genus == "Candidatus\_Scalindua")

## phyloseq-class experiment-level object  
## otu\_table() OTU Table: [ 24 taxa and 7 samples ]  
## sample\_data() Sample Data: [ 7 samples by 22 sample variables ]  
## tax\_table() Taxonomy Table: [ 24 taxa by 7 taxonomic ranks ]

q.norm %>%  
 subset\_taxa(Genus== "D\_5\_\_Candidatus Scalindua")

## phyloseq-class experiment-level object  
## otu\_table() OTU Table: [ 5 taxa and 7 samples ]  
## sample\_data() Sample Data: [ 7 samples by 22 sample variables ]  
## tax\_table() Taxonomy Table: [ 5 taxa by 7 taxonomic ranks ]

#Richness of Candidatus Scalindua across samples  
m.norm %>%   
 subset\_taxa(Genus== "Candidatus\_Scalindua") %>%  
 estimate\_richness(measures = c("Observed"))

## Observed  
## Saanich\_010 0  
## Saanich\_100 1  
## Saanich\_120 3  
## Saanich\_135 5  
## Saanich\_150 4  
## Saanich\_165 10  
## Saanich\_200 18

q.norm %>%  
 subset\_taxa(Genus=="D\_5\_\_Candidatus Scalindua") %>%  
 estimate\_richness(measures = c("Observed"))

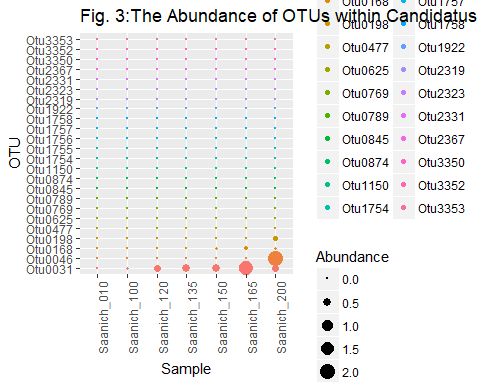
## Warning in estimate\_richness(., measures = c("Observed")): The data you have provided does not have  
## any singletons. This is highly suspicious. Results of richness  
## estimates (for example) are probably unreliable, or wrong, if you have already  
## trimmed low-abundance taxa from the data.  
##   
## We recommended that you find the un-trimmed data and retry.

## Observed  
## Saanich\_010 0  
## Saanich\_100 1  
## Saanich\_120 1  
## Saanich\_135 2  
## Saanich\_150 2  
## Saanich\_165 3  
## Saanich\_200 2

## Do the abundances of OTUs/ASVs within *Candidatus* Scalindua change significantly with depth and oxygen concentration?

The abundance of OTUs within *Candidatus* scalindua does not differ significantly with depth when using mothur. This was determined by performing statistical tests of each OTU at all the depths which resulted in p-values greater than 0.05. As shown in Figure 3, it is evident that the abundance of the OTU’s remains constant for almost all of the identified 24 OTUs. Only OTU 0031 seems to display a change in abundance however the statistical test indicated that this difference is still not significant. The abundance of OTUs with oxygen concentration also does not change significantly with depth when using mothur. Similar to mothur, the abundance of ASVs does not differ significantly with depth or oxygen concentration when using QIIME2. The statistical tests performed on each of the five ASVs resulted in a p-value greater than 0.05 indicating that the difference is not significant.

#Figure 3  
m.perc %>%  
 subset\_taxa(Genus=="Candidatus\_Scalindua") %>%  
 psmelt() %>%  
   
 ggplot() +  
 geom\_point(aes(x=Sample, y=OTU, size=Abundance, color=OTU)) +  
 scale\_size\_continuous(range = c(0,5)) +  
 labs(title="Fig. 3:The Abundance of OTUs within Candidatus Scalindua across depth")+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1))



# **Discussion**

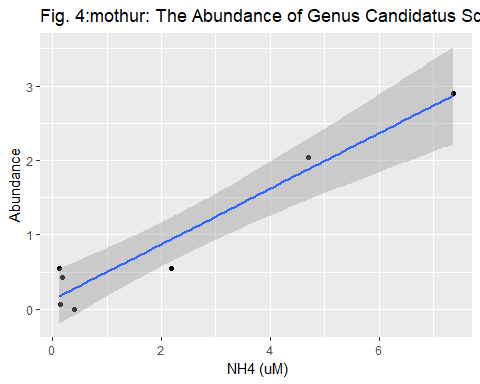
At first sight, alpha diversity seems to significantly increase with higher oxygen concentrations. However, this significant effect can only be observed if the samples are sorted in only two groups, oxic vs. anoxic. When the diversity is plotted against the actual oxygen concentrations in a linear model, no significant change can be observed anymore. This means, we cannot make any conclusions about changes of alpha diversity across oxygen concentrations.

The abundance of our chosen taxon *Candidatus* scalindua does not significantly differ with decreasing oxygen concentration for either QIIME2 or mothur analyzed datsets. The abundance across depth shows a significant correlation for both mothur and QIIME2, with a p-value of 0.03964 and 0.03787 respectively. However, this p-value is not very low and would not be significant in a 99% cutoff approach. Therefore, gathering more sampling data and repeating the statistical analysis might be worth considering to further investigate the correlation between abundance and depth. In addition, QIIME2 does not show a significant increase of our taxon with decreasing oxygen concentration. Also, none of the observed single OTUs or ASVs differs significantly across the two variables oxygen and depth. These results are somewhat surprising, as *Candidatus* scalindua have been shown to be anaerobic ammonium oxidizing bacteria (9). We expected the abundances of the bacteria to increase with greater depth and decreasing oxygen concentrations as the anammox process should be favored in the absence of oxygen. One possible explanation for the lack of a significant change with oxygen concentrations might be, that *Candidatus* scalindua could be aerotolerant and actually does not directly rely on an anaerobic environment. Therefore, we investigated the correlation of abundance with the actual metabolites of the anammox process. Interestingly, the abundances indeed change significantly when observed across concentrations of NH4 or NO2, which are the two educts of the anammox process. The abundance of *Candidatus\_Scalindua* genus is tightly correlated with NH4 concentration (p= 0.0003607) as can be seen in Figure 4.This result aligns well with the annotation of *Candidatus* scalindua as anammox process performing bacteria.

#Linear model: Abundance Against NH4  
m.perc %>%  
 subset\_taxa(Genus=="Candidatus\_Scalindua") %>%  
 psmelt() %>%  
 group\_by(Sample) %>%  
 summarize(Abundance\_sum=sum(Abundance), NH4\_uM=mean(NH4\_uM)) %>%  
 lm(Abundance\_sum ~ NH4\_uM, .) %>%  
 summary()

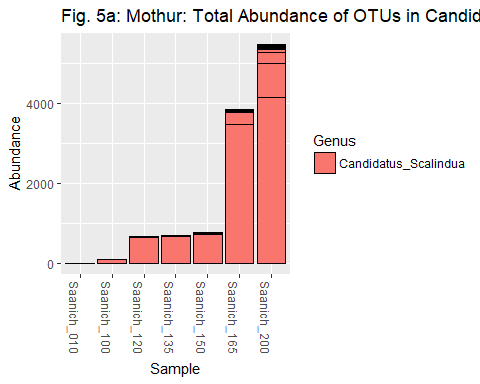
##   
## Call:  
## lm(formula = Abundance\_sum ~ NH4\_uM, data = .)  
##   
## Residuals:  
## 1 2 3 4 5 6 7   
## -0.27907 -0.10902 0.23664 0.36977 -0.39631 0.15448 0.02352   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.12689 0.14868 0.853 0.432398   
## NH4\_uM 0.37300 0.04363 8.549 0.000361 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3046 on 5 degrees of freedom  
## Multiple R-squared: 0.936, Adjusted R-squared: 0.9232   
## F-statistic: 73.08 on 1 and 5 DF, p-value: 0.0003607

#Figure 4: plot corresponding to the Abundance vs NH4 linear model  
m.perc %>%  
 subset\_taxa(Genus=="Candidatus\_Scalindua") %>%  
 psmelt() %>%  
 group\_by(Sample) %>%  
 summarize(Abundance\_sum=sum(Abundance), NH4\_uM=mean(NH4\_uM)) %>%  
   
 ggplot() +  
 geom\_point(aes(x=NH4\_uM, y=Abundance\_sum)) +  
 geom\_smooth(method='lm', aes(x=as.numeric(NH4\_uM), y=Abundance\_sum)) +  
 labs(title="Fig. 4:mothur: The Abundance of Genus Candidatus Scalindua across ammonium conc.", y = "Abundance" , x =" NH4 (uM)")

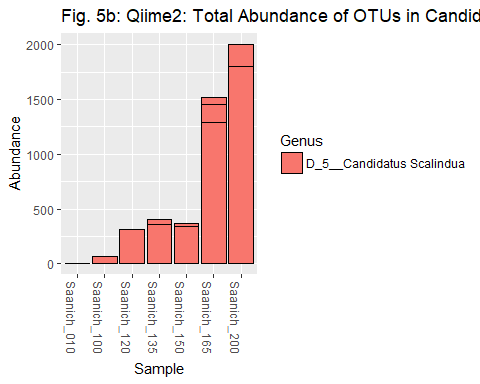


The general trend in abundance for *Candidatus* scalindua across depth determined by mothur and QIIME2 showed similar results. However these two methods differed in the numerical quantity observed at each depth, with mothur obtaining values approximately double the abundance of those determined by QIIME2 (Figure 5a). This is expected as QIIME2 produces ASVs through denoising of the data, thus producing fewer hits compared to the mothur pipeline (1). QIIME2 discards more data, potentially biasing the relative abundances towards a low error-prone sequence. Mothur on the other hand, keeps more of the data, although not all of it will represent *Candidatus* scalindua in the community sampled.

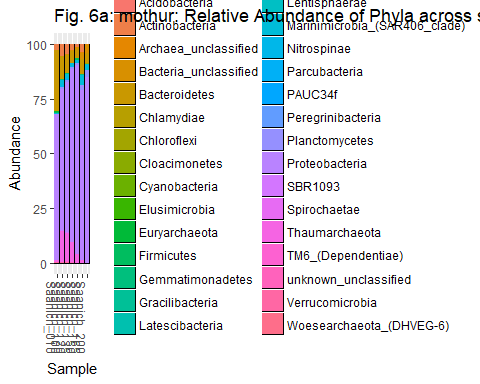
#Figure 5A  
gp = subset\_taxa(mothur, Genus== "Candidatus\_Scalindua")  
plot\_bar(gp, fill="Genus") +  
 labs(title="Fig. 5a: Mothur: Total Abundance of OTUs in Candidatus Scalindua")



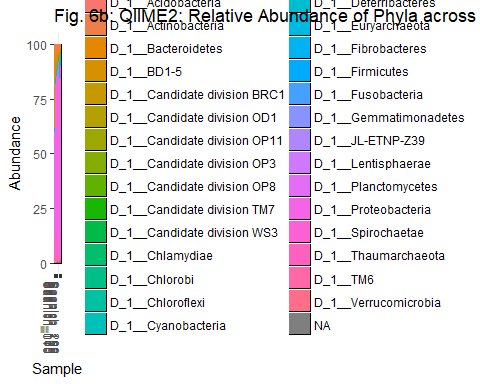
#Figure 5B  
gp2 = subset\_taxa(qiime2, Genus== "D\_5\_\_Candidatus Scalindua")  
plot\_bar(gp2, fill="Genus")+  
 labs(title="Fig. 5b: Qiime2: Total Abundance of OTUs in Candidatus Scalindua")

 Although there were a few differences, mothur and QIIME in general were both able to detect mostly the same phyla and classes in similar abundances (Figure 6a, 6b, 7a, 7b). Likewise, the general trend in Shannon’s diversity index with depth and O2 concentration was similar between mothur and QIIME2.

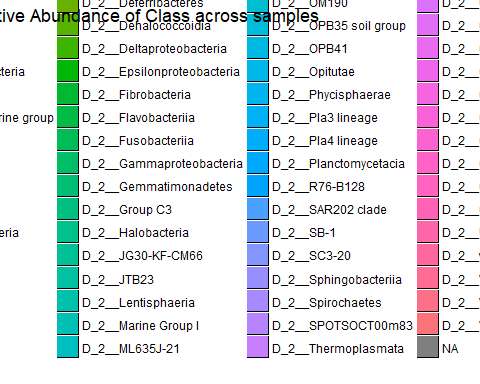
#FIgure 6A  
m.perc %>%  
 plot\_bar(fill="Phylum") +  
 geom\_bar(aes(fill=Phylum), stat="identity") +  
 labs(title="Fig. 6a: mothur: Relative Abundance of Phyla across samples")



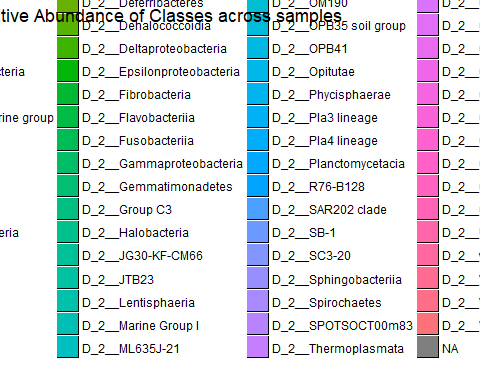
#Figure 6b  
q.perc %>%  
 plot\_bar(fill="Phylum") +  
 geom\_bar(aes(fill=Phylum), stat="identity") +  
 labs(title="Fig. 6b: QIIME2: Relative Abundance of Phyla across samples")



#Figure 7A  
q.perc %>%  
 plot\_bar(fill="Class") +  
 geom\_bar(aes(fill=Class), stat="identity") +  
 labs(title="Fig. 7a: mothur: Relative Abundance of Class across samples")



#Figure 7B  
q.perc %>%  
 plot\_bar(fill="Class") +  
 geom\_bar(aes(fill=Class), stat="identity") +  
 labs(title="Fig. 7b: QIIME2: Relative Abundance of Classes across samples")



Both OTU and ASV data revealed similar trends between each other when analyzing abundance of Candidatus scalindua across depth and oxygen concentration. Abundance vs depth shows a slight positive correlation, however the 95% confidence interval is still relatively large in both the OTU and ASV data, implying that the correlation may not as strong as the trend line predicts (Figures 2a, 2c). Abundance vs oxygen concentration shows a slight negative correlation, however again, both the OTU and ASV data show a large 95% confidence interval, implying that O2 levels may not be a significant influential factor (Figures 2a, 2c).

The relatively small quantity of identified OTUs and ASVs for Candidatus scalindua may explain why the analysis from mothur and QIIME2 do not seem to differ significantly. However it is argued that by using ASVs over OTUs, gene sequence analysis will become more comprehensive, reproducible, reusable, and precise. The connection between OTUs and species are unsupported, and output numbers of OTUs are usually much larger than the actual number of strains in a community (10). Unlike OTUs, ASVs are reproducible and comprehensive fundamental units and have a biological meaning as a DNA sequence (11).

Future directions for this project may include analysis of other taxonomic groups in the samples obtained from Saanich Inlet. Specifically, the analysis of taxonomic groups with a larger amount of OTUs and ASVs. With more data, we may then see a more significant difference between the mothur and QIIME2 analysis results compared to the minimal contrast observed with the *Candidatus* scalindua data.

# **References**

1. **Callahan, BJ, McMurdie, PJ, Holmes, SP.** 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal. 11:2639.
2. **Paulmier A, Ruiz-Pino D.** 2009. Oxygen minimum zones (OMZs) in the modern ocean. Progress in Oceanography. 80: 3-4.
3. **Walsh DA, Zaikova E, Howes CG, Song YC, Wright JJ, Tringe SG, Tortell PD, Hallam, SJ.** 2009. Metagenome of a versatile chemolithoautotroph from Expanding Oceanic Dead Zones. Science. 326:5952.
4. **Hawley AK, Torres-Beltrán M, Zaikova E, Walsh DA, Mueller A, Scofield M, Kheirandish S, Payne C, Pakhomova L, Bhatia M, Shevchuk O, Gies EA, Fairley D, Malfatti SA, Norbeck AD, Brewer HM, Pasa-Tolic L, del Rio TG, Suttle CA, Tringe S, Hallam SJ.** 2017. A compendium of multi-omic sequence information from the Saanich Inlet water column. Sci Data 4:170160.
5. **Wang J and Gu JD.** 2013. Dominance of Candidatus Scalindua species in annamox community revealed in soils with different duration of rice paddy cultivation in Northeast China. Appl Microbial Biotechnol. 97:4.
6. **Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF.** 2009. Introducing mothur: open-source, platform-Independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537-7541.
7. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R.** 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335-336.
8. **R Core Team.** 2017. R: A language and environment for statistical computing. 3.4.3. R Foundation for Statistical Computing, Vienna, Austria.
9. **Schmid, M., Walsh, K. Webb, R. Rijpstra WI, van de Pas-Schoonen K., Verbruggen MJ, Hill T, Moffett B, Fuerst J, Schouten S, Sinninghe Damste JS, Harris J, Shaw P, Jetten M, Strous M.** Candidatus “Scalindua brodae”, sp. nov., Candidatus “Scalindua wagneri”, sp. nov., two new species of anaerobic ammonium oxidizing bacteria. Syst Appl Microbiol, 2003. 26(4): p. 529-38.
10. **Stackebrandt E, Ebers J.** 2006. Taxonomic parameters revisited: tarnished gold standards. Microbiology Today 33: 152.
11. **Kopylova, E, Navas-Molina, JA, Mercier, C, Xu, ZZ, Mahé, F, He, Y, Zhou, H, Rognes, T, Caporaso, JG, Knight, R.** 2016. Open-source sequence clustering methods improve the state of the art. Msystems. 1:3.