Project2\_Report

Group 2: J. Atif, L. Chang, J. Gill, J. Huggins, L. Sonderegger, P. Suresh

April 17, 2018

Table of Contents

[Abstract 1](#_Toc512422756)

[Introduction 2](#_Toc512422757)

[Methods 4](#_Toc512422758)

[Sample collection, sequencing and processing 4](#_Toc512422759)

[Bioinformatic analysis and tree generation 4](#_Toc512422760)

[Data cleaning 5](#_Toc512422761)

[Results 6](#_Toc512422762)

[How do the DNA and RNA abundance of napA differ with depth? 6](#_Toc512422763)

[What taxa are responsible for napA gene expression and abundance? Are they the same with depth, and with DNA versus RNA? 8](#_Toc512422764)

[How does the abundance of napA relate to nitrogen species in Saanich? 13](#_Toc512422765)

[Discussion 16](#_Toc512422766)

[Future directions 18](#_Toc512422767)

[References 18](#_Toc512422768)

[Appendix 20](#_Toc512422769)

[Code to Generate Combined Dataset Containging Taxononmy and Abundance Information for both RNA and DNA at each Depth 20](#_Toc512422770)

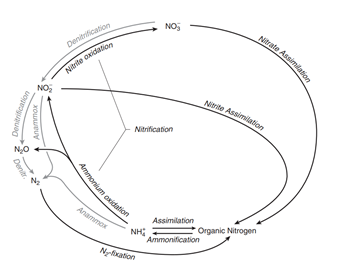
[Code to Generate Taxonomy and Relative Abundane Dataset for both RNA and DNA at each depth 22](#_Toc512422771)

# Abstract

Oxygen minimum zones are important for understanding the nitrogen cycle and the role that the ocean plays with respect to atmospheric greenhouse gases. The nitrogen cycle consists of three processes: N2 fixation, nitrification, and denitrification, each of which have specific groups of microorganisms contributing to the overall pathways. 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. The amount of nitrate increases (to a certain point) as oxygen decreases with depth in Saanich Inlet, which results in the production of the greenhouse gas N2O through denitrification. napA is the gene responsible for converting nitrate to nitrite, and is analyzed in this report. Using TreeSAPP to map the DNA and RNA reads to reference sequences, napA abundance data and phylogenetic relationships were inferred. The goal of this study is to understand the dynamics of the abunance of napA-bearing bacteria and napA expression by these bacteria down the water column in Sannich Inlet. To do this, we compared the DNA and RNA abundances and the taxonomic identities of these microbes at increasing depths. We found that increased napA expression was observed to be present at deeper depths, due to its known inhibition by oxygen. However, NapA expression was still observed at all depths, possibly due to its ability to catalyze the first step of assimilation of nitrate into organic compounds. Furthermore, we found that certain taxa closely mirrored their abundance to nitrate concentrations at different depths below the surface of the water. As expected, We also saw that RNA levels, at their peak, tend to be much higher than peak DNA levels. Different reaction steps of nutrient cycles are found distributed among different organisms due to early horizontal gene transfer followed by selection and amplification of traits that allow these species to adapt and thrive in their unique environments. Abundance of substrates and products played a role in influencing which reaction steps a species maintained. This specialization led to organisms taking up different specialized niches in their ecosystem, and interdependence between species developed.

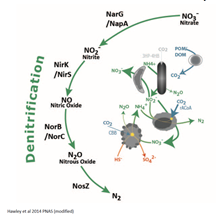
# Introduction

Oxygen minimum zones (OMZs) are areas in the ocean in which oxygen concentrations are lower than 20 µM. They are important for understanding the nitrogen cycle and the role that the ocean plays with respect to atmospheric greenhouse gases (1). 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 favors chemolithotrophs, which results in the production of several greenhouse gases (2). 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 (3). 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 (3). 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 (3). As oxygen levels decline in Saanich Inlet, the microorganisms present in the water column use alternative terminal electron acceptors, which allows denitrification to occur (4). The amount of nitrate increases (to a certain point) as oxygen decreases in Saanich Inlet, which results in the production of the greenhouse gas N2O through denitrification.  
Nitrogen is one of the most abundant substances in the atmosphere, but in its natural form, is inaccessible to most organisms who need it as a building block for survival (5). The nitrogen cycle consists of three processes: N2 fixation, nitrification, and denitrification, each of which have specific groups of microorganisms contributing to the overall pathways (6). A diagram of the distribution of the nitrogen cycle can be seen in intro Figure 1 (7).



Intro Figure 1: Nitrogen cycle in the marine environment, with different reactions for N2 fixation, nitrification, and denitrification (Gruber, 2008 (7))

Denitrification is the process of anaerobic respiration resulting in the conversion of nitrate, NO3-, to nitrogen gas, N2 (Bernard, 2010). The intermediates of this reaction are nitrite (NO2-), nitric oxide (NO) and nitrous oxide (N2O) (6). The microorganisms who perform these reactions are deemed denitrifiers, but most of these organisms only possess genes to perform a select few steps of the pathway (6). The complete denitrification pathway is therefore distributed and comes from the cooperation of these microorganisms. A model of the genes involved in denitrification can be seen in intro Figure 2, taken from Julia Ansett’s MICB 425 presentation (8).



Different genes involved in the denitrification pathway of the nitrogen cycle. Different organisms carry different genes, and their cooperation results in completion of the overall pathway.

The fact that microorganisms perform different steps of the pathway means that the intermediate gases are released into the environment, which leads to a greenhouse effect. Denitrification is an essential pathway that removes fixed nitrogen (nitrate) from the ecosystem and returns it to its biologically stable form on N2 so that the cycle can repeat, and harmful intermediate products and reactive nitrogen don’t accumulate (5).  
In this analysis, napA, the gene responsible for converting nitrate to nitrite, was analyzed at 7 different depths from Saanich Inlet. Using TreeSAPP to map the DNA and RNA reads to reference sequences, abundance data and phylogenetic relationships were inferred. This data was used to determine how the DNA and RNA abundance differ with depth, the taxa that are mainly responsible for napA and which depth(s) they are active at, and the relationship between the abundance of napA and nitrogen species analyzed in Project 1.

# Methods

## Sample collection, sequencing and processing

The sequence data was derived from Saanich Inlet Cruise 72. Specific target genes, specifically the V4-V5 region of small subunit (SSU) 16S rRNA, were amplified and sequenced in order to study microbial communities within a community. Samples for large volume SSU rRNA gene tags were collected from 6 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 and PCR amplification was used to generate SSU rRNA gene libraries (4). The samples were then sequenced using MiSeq system with Phred 33 quality scores. The Phred33 scores indicate the estimated probability of an sequencing error. The validity and quality of the SSU rRNA amplicons were determined at the Joint Genome Institute using the IMG/M pipeline (9). Gel electrophoresis was performed on SSU rRNA gene amplicons to check for molecular weight and possible degradation (4). Metagenomes were assembled and processed using MetaPathways 2.5 at UBC (10).

## Bioinformatic analysis and tree generation

In order to analyze the metagenomes, the TreeSAPP (11) (Tree-based Sensitive and Accurate Protein Profiler) pipeline was used to map reads to reference sequences, perform multiple sequence alignments, and create files that allowed for reconstruction of the nitrogen cycle along with determination of abundance of napA. TreeSAPP version Beta was used. A bash script using nano text editor was created to align the sequenced reads from all seven depths with metagenomic and metatranscriptomic assemblies. Example code is shown for the beginning of the bash script with 10m sequence data.

#!/bin/bash  
export WISECONFIGDIR=/home/connor/TreeSAPP//data//genewise\_support\_files//wisecfg  
  
time treesapp.py -T 8 --verbose --delete --pairing pe -t D0201 -i bucket/MetaG\_assemblies/SI072\_LV\_10m\_DNA.scaffolds.fasta --rpkm -r bucket/MetaG\_reads/SI072\_LV\_10m.anqdp.fastq.gz -o treesapp\_out\_dir\_depth10DNA  
  
rm treesapp\_out\_dir\_depth10DNA/RPKM\_outputs/\*.sam  
  
time treesapp.py -T 8 --verbose --delete --pairing pe -t D0201 -i bucket/MetaG\_assemblies/SI072\_LV\_10m\_DNA.scaffolds.fasta --rpkm -r bucket/MetaT\_reads/SI072\_LV\_10m.qtrim.3ptrim.artifact.rRNA.clean.fastq.gz -o treesapp\_out\_dir\_depth10RNA  
  
rm treesapp\_out\_dir\_depth10RNA/RPKM\_outputs/\*.sam

The script repeats, with the number 10 being substituted with the other 6 depths (100m, 120m, 135m, 150m, 165m, and 200m). SAM files are removed after every run due to their large size and limited space on the Google Cloud server. The bash script was run using screen and took about 2 days to complete. After the TreeSAPP processing was complete, the generated .jplace files were used to create phylogenetic trees on the integrative Tree of Life (iTOL) for both the DNA and RNA data (12). iTOL version 4.2 was used. the TreeSAPP processing also generated TSV files that had nucleotide type, depth, taxonomy and abundance information organized in a tab delimited spreadsheet.

### Data cleaning

Further manipulation of the TSV data generated above and plotting was completed in R v3.4.3 (13) using the following packages:

library("tidyverse")  
library("phyloseq")  
library("cowplot")

After loading the TSV data into R, the depth, taxonomy, abundance and query data for both RNA and DNA at all depths were manipulated into a single data frame. For the extended code used to generate this data table for either DNA or RNA, see the folder in the repo called “iTol Outputs”. Relative abundance for both RNA and DNA at each depth was also calculated and the code for this is also in the Appendix.

napA.all

## # A tibble: 15,918 x 11  
## Depth\_m Type Abundance Domain Phylum Class Order Family Genus Species  
## <dbl> <chr> <dbl> <chr> <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 10. DNA 0. Bacter~ <NA> <NA> <NA> <NA> <NA> <NA>   
## 2 10. DNA 0. Bacter~ Firmi~ Clos~ Clos~ Symbi~ <NA> <NA>   
## 3 10. DNA 0. Bacter~ Prote~ Delt~ Myxo~ Cysto~ <NA> <NA>   
## 4 10. DNA 0. Bacter~ Firmi~ Clos~ Clos~ Symbi~ <NA> <NA>   
## 5 10. DNA 0. Bacter~ Bacte~ Cyto~ Cyto~ Flamm~ <NA> <NA>   
## 6 10. DNA 0. Bacter~ Prote~ Delt~ uncl~ Dissu~ <NA> <NA>   
## 7 10. DNA 0. Bacter~ Acido~ uncl~ Acid~ <NA> <NA> <NA>   
## 8 10. DNA 0. Bacter~ <NA> <NA> <NA> <NA> <NA> <NA>   
## 9 10. DNA 0. Bacter~ Prote~ Gamm~ Ocea~ Sacch~ <NA> <NA>   
## 10 10. DNA 0. Bacter~ Firmi~ Clos~ Clos~ Symbi~ <NA> <NA>   
## # ... with 15,908 more rows, and 1 more variable: Query <chr>

The geochemical metadata associated with each depth was also used.

load("~/MICB425\_portfolio/mothur\_phyloseq.RData")  
  
metadata = data.frame(mothur@sam\_data)

## Depth\_m O2\_uM PO4\_uM SiO2\_uM NO3\_uM NH4\_uM Std\_NH4\_uM NO2\_uM  
## Saanich\_010 10 216.667 0.520 20.672 1.793 0.4080 0.0084 0.1275  
## Saanich\_100 100 38.012 3.672 69.062 26.400 0.1344 0.0092 0.0817  
## Saanich\_120 120 32.354 4.090 64.383 21.302 0.1782 0.0100 0.0978  
## Saanich\_135 135 20.446 4.342 70.321 15.917 0.1296 0.0166 0.0706  
## Saanich\_150 150 0.000 4.988 70.780 5.278 2.1754 0.0293 0.1127  
## Saanich\_165 165 0.000 5.599 62.580 0.000 4.7095 0.2112 0.0805  
## Saanich\_200 200 0.000 6.298 66.200 0.000 7.3582 0.2816 0.0000  
## Std\_NO2\_uM H2S\_uM Std\_H2S\_uM Cells.ml N2O\_nM Std\_N2O\_nM  
## Saanich\_010 0.0088 0.0000 0.0000 NAN 0.849 0.114  
## Saanich\_100 0.0105 0.0000 0.0000 NAN 18.087 1.275  
## Saanich\_120 0.0018 0.0000 0.0000 NAN 16.304 1.085  
## Saanich\_135 0.0018 0.0000 0.0000 NAN 12.909 2.577  
## Saanich\_150 0.0018 0.0000 0.0000 NAN 11.815 0.000  
## Saanich\_165 0.0053 3.5027 0.0423 NAN 6.310 0.732  
## Saanich\_200 0.0000 17.9867 0.0006 NAN 0.000 0.000  
## CH4\_nM Std\_CH4\_nM Temperature\_C Conductivity\_mScm\_1  
## Saanich\_010 1030.478 3.070 12.854 33.534  
## Saanich\_100 3.231 0.392 8.703 32.970  
## Saanich\_120 3.463 0.519 8.796 33.188  
## Saanich\_135 4.815 0.658 8.882 33.345  
## Saanich\_150 8.323 0.000 9.002 33.526  
## Saanich\_165 23.831 2.291 9.041 33.597  
## Saanich\_200 774.034 12.745 9.117 33.727  
## Fluorescence\_mgm\_3 OxygenSBE\_V Salinity\_PSU Density\_q  
## Saanich\_010 3.521 4.954 28.121 21.098  
## Saanich\_100 0.109 0.872 30.872 23.933  
## Saanich\_120 0.197 0.742 31.007 24.026  
## Saanich\_135 0.108 0.469 31.088 24.076  
## Saanich\_150 0.181 0.089 31.164 24.118  
## Saanich\_165 0.132 0.069 31.197 24.138  
## Saanich\_200 0.236 0.063 31.248 24.167

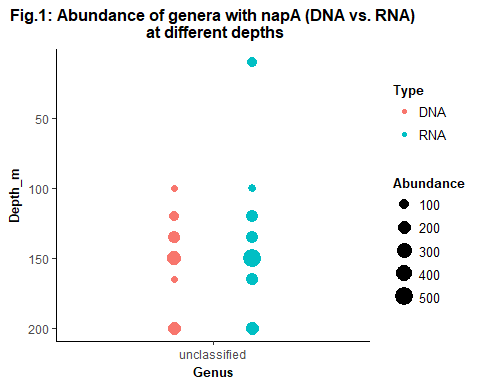
# Results

## How do the DNA and RNA abundance of napA differ with depth?

According to Figure 1, the abundance of the napA gene increases from 100m to 150m. The abundance then decreases dramatically at about 165m, and increases in abundance again at 200m. The presence of DNA is missing at the depth of 10m. DNA abundance is greatest at the depth of approximately 150m and is the least at 100m and 160m.

As seen in Figure 1, the RNA abundance of the napA gene is very similar to the DNA abundance. However, unlike DNA, the RNA of the gene is present at the surface of the Saanich Inlet at 10m. Similar to the DNA levels, the abundance of RNA also increases from 100m to 150m before decreasing at 160m. The RNA increases in abundance again at the depth of 200m. The RNA abundance of napA is the greatest at 150m and is the least at 100m.

napA.all %>%  
 # Change NAs to "unclassified" at the level you want to plot  
 mutate(Genus = ifelse(is.na(Genus), "unclassified", Genus)) %>%  
   
 ggplot(aes(x = Genus, y = Depth\_m)) +  
 # Use an ifelse statement to make 0 values into NAs so that they don't show up on the plot  
 # Use position\_dodge to keep points from overlapping  
 geom\_point(aes(size = ifelse(Abundance == 0, NA, Abundance), color = Type), position = position\_dodge(0.5)) +  
 scale\_y\_reverse(lim=c(200,10)) +  
 labs(title = "Fig.1: Abundance of genera with napA (DNA vs. RNA)\n at different depths") +  
 theme\_classic() +  
 theme(plot.title=element\_text(size=12, face="bold", hjust = 0.5), legend.text=element\_text(size=10),legend.title=element\_text(size=10, face="bold"), axis.title = element\_text(face="bold", size=10))+  
 # Rename legend  
 scale\_size\_continuous(name = "Abundance")



## What taxa are responsible for napA gene expression and abundance? Are they the same with depth, and with DNA versus RNA?

At the genus level, as can be seen by Figure 1, None of the queries and their corresponding sequences can be confidently classified into taxonomic groups. However at a higher taxonomic level like family (Figure 2), the 16S rRNA data can be properly assigned taxonomic classification for both DNA and RNA. Figure 2 shows the relative RNA or DNA abundance for napA at each depth. It is clear from looking at the plot that the relative DNA and RNA abundance of most of the microorganisms responsible for napA changes with depth. However, the RNA abundanes seem to change more dramatically than the DNA abundances. For example, the members of the family Deferribacteraceae seem to comprise a good proportion of the bacteria (as seen by relative DNA abundance) at most depths. However, they also seem very sensitive to the geochemistry present at the different depths down the water column as the relative RNA abundance changes. Only members of the family Symbiobacteriaceae do not seem to change in their relative DNA abundance across depth, even though the relative RNA abundance does change.

RNA<-ggplot(data=R\_all, aes(x=Depth, y= Perc, fill=Family))+  
 geom\_bar(stat="identity")+  
 labs(title="napA Expression (RNA): Relative Abundance\n of Bacterial familiess across Depths", y="Relative Abundance")+  
theme(legend.position="none", plot.title = element\_text(size=12, face="bold", hjust=0.5), axis.title = element\_text(face="bold", size=10), axis.text= element\_text(size=9))  
  
  
DNA<-ggplot(data=D\_all, aes(x=Depth, y= Perc, fill=Family))+  
 geom\_bar(stat="identity")+  
 labs(title="napA in genome (DNA): Relative Abundance\n of Bacterial Families across Depths", y="Relative Abundance")+  
theme(legend.position="bottom",legend.text=element\_text(size=8),legend.title=element\_text(size=7, face="bold"), plot.title = element\_text(size=12, face="bold", hjust=0.5), axis.title = element\_text(face="bold", size=10), axis.text= element\_text(size=9))+  
 guides(fill=guide\_legend(ncol=3,bycol=TRUE))  
  
plot\_grid(RNA, DNA, labels=c("Fig.2A", "Fig.2B"), ncol=1, align = 'v', rel\_heights = c(1,1.8))

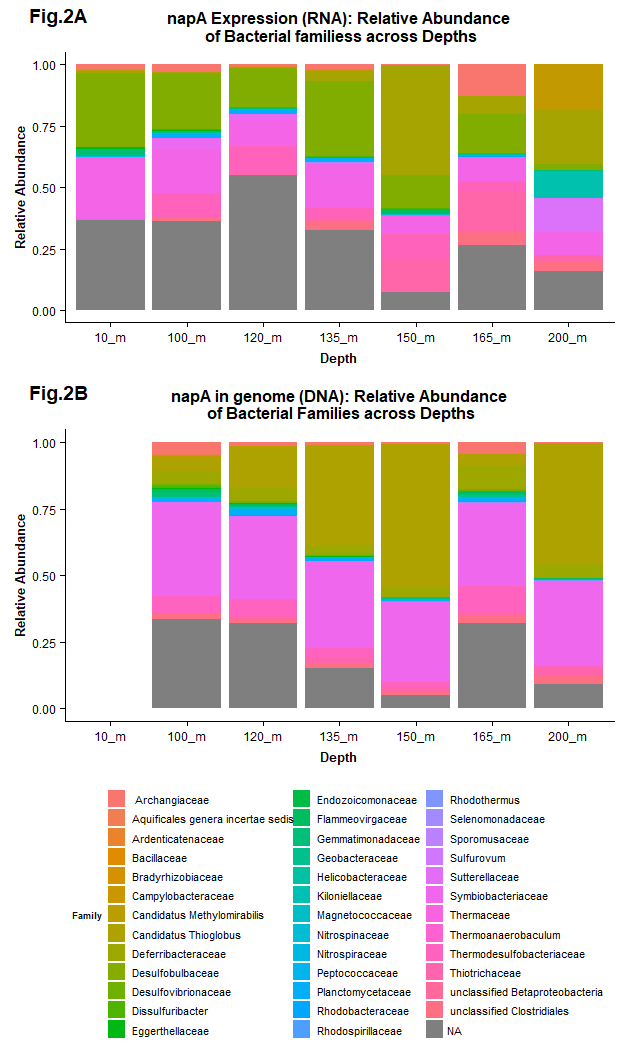
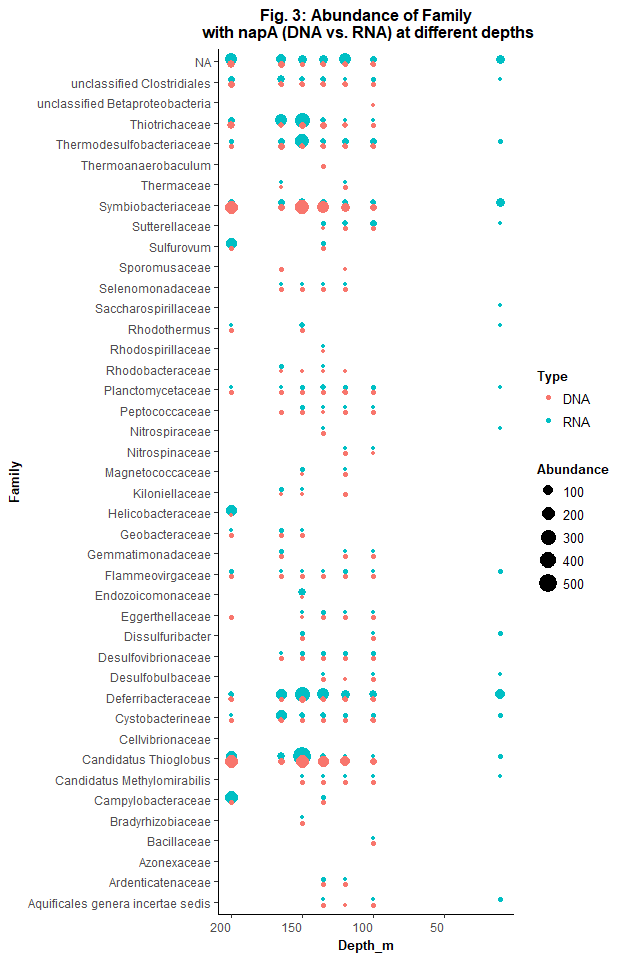


Figure 3 displays the absolute DNA and RNA abundance of different families of microorganisms that are responsible for the napA gene. The families present, as seen by DNA abundance, do differ with depth, as some families are not present at certain depths while others are. For example, Ardenticatenaceae and Desulfobulbaceae are not present at depths below 150m while Cystobacterineae and Deferribacteraceae are present at depths of 165m and 200m. The abundance of DNA and RNA for the napA gene is generally similar within the different families. Figure 2 once again displays that only the RNA of the napA gene and not the DNA is present on the surface of the waters. The abundance of DNA for the gene is generally higher in the families of Candidatus Thioglobus and Symbiobacteriaceae. The abundance of RNA is high in the Candidatus Thioglobus, Deferribacteraceae, Helicobacteraceae, Therrodesulfobacteriaceae, and Thiotrichaceae families. For the most part, the abundance of DNA and RNA is similar at different depths within a family. However, in Deferribacteraceae, Therrodesulfobacteriaceae, and Thiotrichaceae families, the abundance of RNA is greater than DNA at different depths. On the other hand, for the Candidatus Thioglobus and Symbiobacteriaceae families, the abundance of DNA is greater than RNA at different depths.

It is also important to note that for most families the presence of DNA and RNA for the napA gene is absent 200m below the surface of the water. Only some families have the napA gene present at the depth of 200m. After the depth of 100m, a few families have only the DNA of the napA gene present at certain depths and not the RNA of the gene. Overall, the abundance of RNA and DNA of the napA gene is similar within the taxon, however there is some clear variance as well.

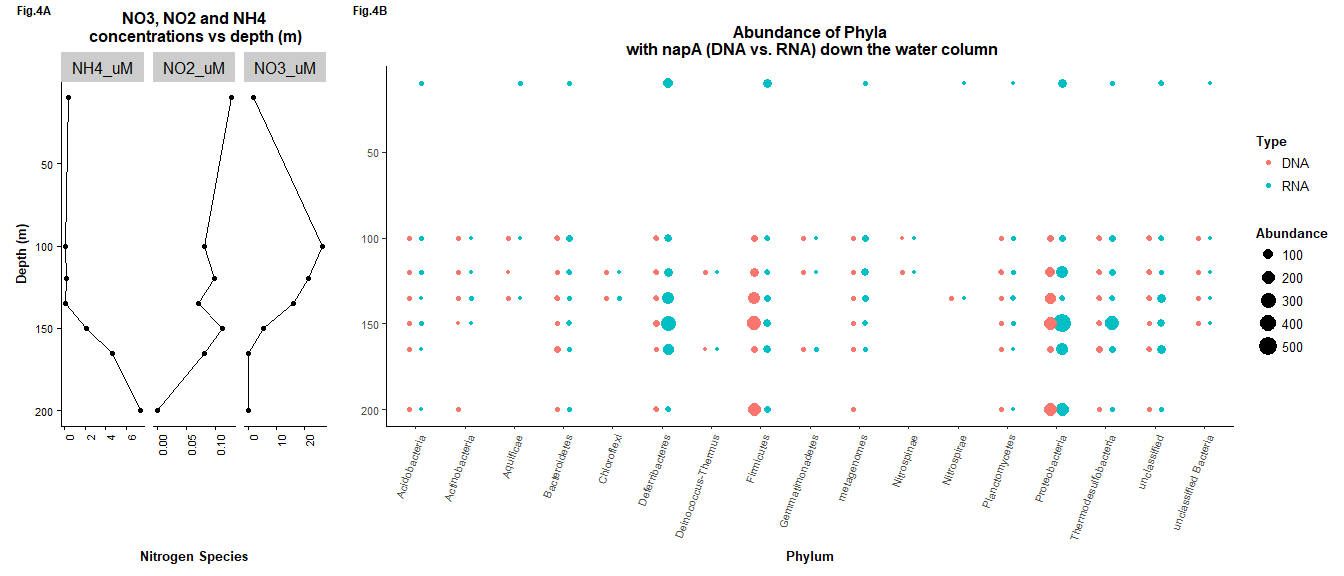
napA.all %>%  
 # Change NAs to "unclassified" at the level you want to plot  
 mutate(Genus = ifelse(is.na(Family), "unclassified", Family)) %>%  
   
 ggplot(aes(x = Family, y = Depth\_m)) +  
 # Use an ifelse statement to make 0 values into NAs so that they don't show up on the plot  
 # Use position\_dodge to keep points from overlapping  
 geom\_point(aes(x= Family, y= Depth\_m, size = ifelse(Abundance == 0, NA, Abundance), color = Type), position = position\_dodge(0.5)) +  
 scale\_y\_reverse(lim=c(200,10)) +  
 labs(title = "Fig. 3: Abundance of Family\n with napA (DNA vs. RNA) at different depths") +  
 theme\_classic()+  
 theme(axis.text.x = element\_text(hjust = 1), plot.title=element\_text(size=12, face="bold", hjust = 0.5), legend.text=element\_text(size=10),legend.title=element\_text(size=10, face="bold"), axis.title = element\_text(face="bold", size=10))+  
 # Rename legend  
 scale\_size\_continuous(name = "Abundance") + coord\_flip()



## How does the abundance of napA relate to nitrogen species in Saanich?

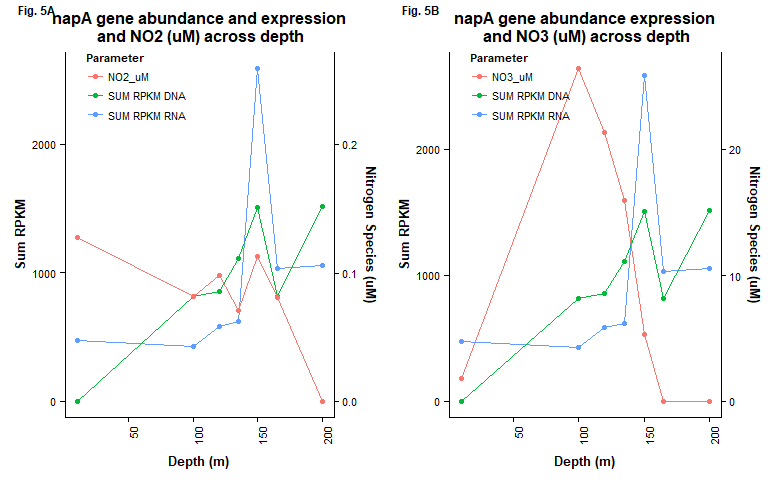
Figure 4 shows how different nitrogen species and the phylum of microorganisms differs with depth. As a general trend, it can be seen that as the abundance of NO2 and NO3 decreases with depth, the presence of DNA and RNA of the napA gene increases. The presence of DNA and RNA is primarily present below the depth of 100m, which is the point where the abundance of NO2 and NO3 are decreasing. On the contrary, the presence of the napA gene across different phyla is seen when NH4 increases. It is also important to note that the abundance of RNA with increasing depth in the phylum Deferribacteres specifically correlates with the abundance with NO3. The concentration of NO3 initially increases before it begins to gradually decrease to zero. Similarly, in the phylum Deferribacteres the abundance of RNA gradually increases and then decreases with depth.

plot1 = napA.all %>%  
 # Change NAs to "unclassified" at the level you want to plot  
 mutate(Phylum = ifelse(is.na(Phylum), "unclassified", Phylum)) %>%  
   
 ggplot(aes(x = Phylum, y = Depth\_m)) +  
 geom\_point(aes(size = ifelse(Abundance == 0, NA, Abundance), color = Type), position = position\_dodge(0.5)) +  
 scale\_y\_reverse(lim=c(200,10)) +  
 labs(y = "", title = "\nAbundance of Phyla\n with napA (DNA vs. RNA) down the water column") +  
 theme\_classic() +  
 scale\_size\_continuous(name = "Abundance")+  
 theme(axis.text.x = element\_text(angle = 70, hjust = 1, size=8), plot.title=element\_text(size=12, face="bold", hjust = 0.5), legend.text=element\_text(size=10),legend.title=element\_text(size=10, face="bold"), axis.title = element\_text(face="bold", size=10), axis.text.y = element\_text(size=8))  
  
plot2=metadata %>%  
 select (Depth\_m, NO2\_uM, NO3\_uM, NH4\_uM) %>%  
 gather("Nutrients","uM", NO2\_uM:NH4\_uM)%>%  
 arrange(Depth\_m) %>%  
 ggplot() + geom\_point(aes(y=Depth\_m, x=uM)) +  
 #geom\_line(aes(y=Depth\_m, x=uM)) +  
 geom\_path(aes(y=Depth\_m, x=uM,group = 1)) +  
 scale\_y\_reverse(lim=c(200,10)) +  
 facet\_wrap(~Nutrients, scales="free\_x", nrow=1) +  
 theme(legend.position="none")+  
 labs(y = "Depth (m)", x = "Nitrogen Species", title="NO3, NO2 and NH4\n concentrations vs depth (m)")+  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1, size=8), plot.title=element\_text(size=12, face="bold", hjust = 0.5), legend.text=element\_text(size=10),legend.title=element\_text(size=10, face="bold"), axis.title = element\_text(face="bold", size=10), axis.text.y = element\_text(size=8))  
   
  
plot\_grid(plot2, plot1, labels=c("Fig.4A", "Fig.4B"), align='h', rel\_widths = c(1.5/6, 4.5/6), axis="b", label\_size = 9, label\_fontface="bold")



Figures 5A and 5B show the abundance of napA in relation to NO2 and NO3 presence, respectively. In both figures, it can be seen that as the nitrogen species concentration is higher than 0 µM, the abundance of the DNA and RNA for napA increases. RNA abundance is lower than DNA at almost all depths except at the depth of 150m where a greater peak in the abundance of RNA in comparison to DNA is evident. Both RNA and DNA abundances peak at a depth of 10 m down the water column. The concentration of NO2 seems to be stable from 0m to 150m below the surface of the water before sharply decreasing. NO3, on the other hand, sharply increases in concentration before sharply declining at about 100m. Interestingly, the slope or rate of the decrease in RNA and DNA abundance after 150 m seems to mirror the slope of decrease that is seen in NO3 concentration.

Nit=metadata %>%  
 select (Depth\_m, NO2\_uM, NO3\_uM)  
  
Sum\_reads\_RNA=data.frame(  
 Depth\_m=c(10, 100, 120, 135, 150, 165, 200),  
 Sum\_RPKM=c(sum(R1$Abundance), sum(R2$Abundance), sum(R3$Abundance), sum(R4$Abundance),sum(R5$Abundance),sum(R6$Abundance), sum(R7$Abundance)))  
N\_RPKM\_RNA=cbind(Sum\_reads\_RNA, Nit)  
  
Sum\_reads\_DNA=data.frame(  
 Depth\_m=c(10, 100, 120, 135, 150, 165, 200),  
 Sum\_RPKM=c(sum(D1$Abundance), sum(D2$Abundance), sum(D3$Abundance), sum(D4$Abundance),sum(D5$Abundance),sum(D6$Abundance), sum(D7$Abundance)))  
N\_RPKM\_DNA=cbind(Sum\_reads\_DNA, Nit)  
  
Nit=metadata %>%  
 select (Depth\_m, NO2\_uM, NO3\_uM)  
  
# both RNA and DNA with NO2 uM  
plot7=ggplot() +  
 geom\_point(data= N\_RPKM\_RNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM RNA")) +  
 geom\_line(data= N\_RPKM\_RNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM RNA")) +  
 geom\_point(data=N\_RPKM\_DNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM DNA")) +  
 geom\_line(data=N\_RPKM\_DNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM DNA")) +  
 geom\_point(data=Nit, aes(x=Depth\_m, y=NO2\_uM\*10000, colour="NO2\_uM"))+  
 geom\_line(data=Nit, aes(x=Depth\_m, y=NO2\_uM\*10000, colour="NO2\_uM")) +  
 scale\_y\_continuous(sec.axis = sec\_axis(~.\*(1/10000), name = "Nitrogen Species (uM)")) +  
 labs(title="napA gene abundance and expression\n and NO2 (uM) across depth",y = "Sum RPKM" , x = "Depth (m)" , colour = "Parameter") +  
 theme(legend.position = c(0.05, 0.9), axis.text.x = element\_text(angle = 90, hjust = 1, size=8), plot.title=element\_text(size=12, face="bold", hjust = 0.5), legend.text=element\_text(size=8),legend.title=element\_text(size=8, face="bold"), axis.title = element\_text(face="bold", size=10), axis.text.y = element\_text(size=8))  
  
# both RNA and DNA with NO3 uM  
plot8=ggplot() +  
 geom\_point(data= N\_RPKM\_RNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM RNA")) +  
 geom\_line(data= N\_RPKM\_RNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM RNA")) +  
 geom\_point(data=N\_RPKM\_DNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM DNA")) +  
 geom\_line(data=N\_RPKM\_DNA, aes(x=Depth\_m, y=Sum\_RPKM, colour= "SUM RPKM DNA")) +  
 geom\_point(data=Nit, aes(x=Depth\_m, y=NO3\_uM\*100, colour="NO3\_uM"))+  
 geom\_line(data=Nit, aes(x=Depth\_m, y=NO3\_uM\*100, colour="NO3\_uM")) +  
 scale\_y\_continuous(sec.axis = sec\_axis(~.\*(1/100), name = "Nitrogen Species (uM)")) +  
 labs(title="napA gene abundance expression\n and NO3 (uM) across depth", y = "Sum RPKM" , x = "Depth (m)" , colour = "Parameter") +  
 theme(legend.position = c(0.05, 0.9),axis.text.x = element\_text(angle = 90, hjust = 1, size=8), plot.title=element\_text(size=12, face="bold", hjust = 0.5), legend.text=element\_text(size=8),legend.title=element\_text(size=8, face="bold"), axis.title = element\_text(face="bold", size=10), axis.text.y = element\_text(size=8))  
  
plot\_grid(plot7, plot8, labels=c("Fig. 5A", "Fig. 5B"), rel\_widths=c(1/2, 1/2), label\_size=9, label\_fontface="bold")



# Discussion

NapA is a periplasmic nitrate reductase that catalyzes the reduction of nitrate to nitrite (4). This reduction of nitrate is the first step in the denitrification process which eventually leads to the production of nitrogen (4). As the denitrification pathway is an anaerobic process (4), the expression of napA and the other denitrification enzymes could be expected to be restricted to depths where no oxygen is present. Our results indeed show a strong increase in napA expression at the depth of 150m, which is a zone of minimal oxygen concentration. Furthermore, the expression of napA could be expected to correlate with the concentration of nitrate, the substrate of the enzyme. However, the highest concentration of nitrate was observed at 100m, 50 meters above the peak of napA expression. napA expression not matching the peak of nitrate concentration could be explained by the notably higher oxygen concentration at 100m. Therefore, at this depth, oxygen could still be used as electron acceptor, which yields much more energy because of the higher reduction potential of oxygen compared to nitrate. Another reason for the peak of napA not matching the highest nitrate concentration could be the distribution of the denitrification pathway amongst different bacterial species. The comparison of bacterial species expressing genes of the denitrification pathway showed that this process is shared amongst several distinct taxa. Therefore, for highest efficiency of the steps catalyzed by one enzyme (or one bacterial species), it is advantageous to perform this reaction at a location where also the other components of the pathway are abundant. napA can only work efficiently when the generated nitrite is not accumulated, but further processed by the downstream enzymes of the denitrification process. These downstream enzymes could be expressed in bacteria that are restricted to depths of at least 150m and therefore force the napA expression also take place in this region.

A question that still remains remains to be addressed is: why is napA expressed at all measured depths? Expression of napA even at locations of high oxygen concentration does not seem to be advantageous, as oxygen could be used as more efficient electron acceptor. The explanation for this observation could be that napA also can be used for assimilation of nitrogen instead of the dissimilatory denitrification pathway. In oxic zones, nitrate is produced by chemolithotrophic bacteria that use nitrite as their electron donor and oxygen as their electron acceptor (5). The accumulated nitrate could then serve as substrate for napA (5). In a later step, nitrite can be turned into ammonia, which can be integrated into organic compounds like amino acids (5). To summarize, napA can catalyze the first step of assimilation of nitrate into organic compounds and can thus be useful at all depths of the ocean.

Another interesting observation is that at 10m only RNA was observed, but no DNA. This can be explained by low abundance of the napA-containing bacteria at this depth. If only few bacteria are present in a sample, their presence might be missed on the level of DNA as each bacteria contains only one copy of the gene. DNA fragments of low abundance easily get lost during processing and sequencing. However, if these bacteria are strongly expressing the gene, many copies of the mRNA are present and therefore are still present after processing of the sample.

There are many different environments on Earth, and for organisms to grow in these locations, each require a mechanism to convert nutrients from an inaccessible form, to a usable substance for life. Such an example is N2 gas, which requires transformation into ammonia (NH3) to be used by plants. Because different environments exist, we would expect environment-specific organisms to take up the niche of transforming (or recycling) nutrients, thus explaining the distribution of metabolic pathways across different species. Microbes have also been known to be “guardians of metabolism”, allowing for the protection of metabolic pathways throughout Earth’s history. Through horizontal gene transfer (HGT), different genes or even whole pathways may be given to different microbial species, some of which have innate characteristics to survive the harsh events on Earth (14). HGT helps explain why metabolic pathways are distributed among different species, and in combination with selective pressures, why some organisms may have genes that process only parts of a nutrient cycle.

In general, reactions in a nutrient cycle are split into steps performed by different enzymes, and each reaction step is distributed among different species (5). This creates an interdependence between species to provide substrates, or remove products to allow for the continuation of each reaction step. But how did distribution of these genes come to be? Although HGT helped spread metabolic pathways among microbial species, selection for only some of these genes may occur as it allows an organism to adapt to the environment.

Studies have shown that ancestral genes of known enzymes may have actually played different roles on early Earth than their current counterparts do (15). Thus, these genes may have spread and have been selected for reasons other than metabolic capability. For example, nitrogenase (which converts N2 to NH4) also has the ability to reduce other molecules such as acetylene, azide, and cyanide. It contains subunits encoded by different genes such as the nifD gene. Early forms of nifD are thought to have roles in coenzyme and metallocluster biosynthesis in the active sites of enzymes (15). Thus, ancestral nitrogenase may have originally provided adaptive traits via coenzyme synthesis, rather than its nitrogen fixing ability. This may explain why we currently see only a few reaction steps of a metabolic cycle in some species, rather than the whole metabolic cycle.

As well, it is possible that loss of a gene may occur in a community if the substrate for that reaction-step in the nutrient cycle is no longer present (16). This may explain why species have only some genes of a metabolic cycle. Different enzymes in a metabolic cycle may be regulated by various environmental factors, thus creating situations where expression/function of an enzyme may be constantly inhibited, making the gene irrelevant. This gene will no longer be selected for in the community and may eventually disappear in abundance. Such an example is nitrogenase which is O2-sensitive. Because of this sensitivity, oxygen-abundant environments may contain organisms that have genes for the rest of the nitrogen cycle, but lack nitrogenase genes. To prevent N2 build-up, another organism would have to take up the niche for nitrogen fixation. This organism may obtain nitrogenase genes via HGT, and may happen to have unique traits to protect the nitrogenase from oxygen, such as through the use of heterocysts (5).

Due to environmental regulation of different genes in a metabolic cycle, it is favorable to have different organisms specialize in different reaction steps. Specialization of a reaction step may be more efficient, as an organism does not need to spend excess energy with all steps of the metabolic cycle. Specialization also prevents different species from competing with each other for the same substrates, but facilitates cooperation and interdependency instead.

# Future directions

Our data is based on single samples of one liter at each depth, which means these small samples might not represent the actual microbiome very well. Therefore, to confirm our results, it would be good to obtain data from more and larger samples. To confirm the hypothesis of a shared denitrification pathway, it would be interesting to follow the conversion of metabolites in vivo, for example by providing labeled substrates to a water sample. The distribution of the labeled molecules amongst the bacterial taxa could then be tracked during the process of denitrification.

# References

1. **Paulmier, A, Ruiz-Pino, D.** 2009. Oxygen minimum zones (OMZs) in the modern ocean. Prog. Oceanogr. **80**:113-128.
2. **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**:578-582.
3. **Torres-Beltrán, M, Hawley, AK, Capelle, D, Zaikova, E, Walsh, DA, Mueller, A, Scofield, M, Payne, C, Pakhomova, L, Kheirandish, S.** 2017. A compendium of geochemical information from the Saanich Inlet water column. Scientific Data. **4**:170159.
4. **Hawley, AK, Torres-Beltrán, M, Zaikova, E, Walsh, DA, Mueller, A, Scofield, M, Kheirandish, S, Payne, C, Pakhomova, L, Bhatia, M.** 2017. A compendium of multi-omic sequence information from the Saanich Inlet water column. Scientific Data. **4**:170160.
5. **Bernhard, A.** 2010. The nitrogen cycle: processes. Players, and Human.
6. **Gruber, N.** 2008. The marine nitrogen cycle: overview and challenges. Nitrogen in the Marine Environment. **2**:1-50.
7. **Stein, LY, Klotz, MG.** 2016. The nitrogen cycle. Current Biology. **26**:R98.
8. **Ansett, Julia** 2018. MICB425: Microbial Ecological Genomics. Automated reconstruction of the nitrogen cycle along defined redox gradients in Saanich Inlet using the TreeSAPP pipeline [PowerPoint slides].
9. **Chen, IA, Markowitz, VM, Chu, K, Palaniappan, K, Szeto, E, Pillay, M, Ratner, A, Huang, J, Andersen, E, Huntemann, M.** 2016. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. gkw929.
10. **Konwar, KM, Hanson, NW, Bhatia, MP, Kim, D, Wu, S, Hahn, AS, Morgan-Lang, C, Cheung, HK, Hallam, SJ.** 2015. MetaPathways v2. 5: quantitative functional, taxonomic and usability improvements. Bioinformatics. **31**:3345-3347.
11. Treesapp version: Beta <https://github.com/hallamlab/treesapp>
12. **R Core Team**. 2017. R: A language and environment for statistical computing. 3.4.3. R Foundation for Statistical Computing, Vienna, Austria.
13. **Letunic, I, Bork, P.** 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. **44**:W245.
14. **Falkowski, PG, Fenchel, T, Delong, EF.** 2008. The microbial engines that drive Earth’s biogeochemical cycles. Science. **320**:1034-1039.
15. **Dos Santos, PC, Fang, Z, Mason, SW, Setubal, JC, Dixon, R.** 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics. **13**:162.
16. **Maslov, S, Krishna, S, Pang, TY, Sneppen, K.** 2009. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proceedings of the National Academy of Sciences. **106**:9743-9748.

# Appendix

## Code to Generate Combined Dataset Containging Taxononmy and Abundance Information for both RNA and DNA at each Depth

RNA\_10<-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/10m/10m\_RNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.RNA.10 = Confident\_Taxonomy, Abund.RNA.10 = Abundance, Query)  
RNA\_100<-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/100m/100m\_RNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.RNA.100 = Confident\_Taxonomy, Abund.RNA.100 = Abundance, Query)  
RNA\_120 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/120m/120m\_RNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.RNA.120 = Confident\_Taxonomy, Abund.RNA.120 = Abundance, Query)  
RNA\_135 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/135m/135m\_RNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.RNA.135= Confident\_Taxonomy, Abund.RNA.135= Abundance, Query)  
RNA\_150 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/150m/150m\_RNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.RNA.150= Confident\_Taxonomy, Abund.RNA.150= Abundance, Query)  
RNA\_165 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/165m/165m\_RNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.RNA.165= Confident\_Taxonomy, Abund.RNA.165= Abundance, Query)  
RNA\_200 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/200m/200m\_RNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.RNA.200= Confident\_Taxonomy, Abund.RNA.200= Abundance, Query)  
  
DNA\_10<-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/10m/10m\_DNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.DNA.10 = Confident\_Taxonomy, Abund.DNA.10 = Abundance, Query)  
DNA\_100<-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/100m/100m\_DNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.DNA.100 = Confident\_Taxonomy, Abund.DNA.100 = Abundance, Query)  
DNA\_120 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/120m/120m\_DNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.DNA.120 = Confident\_Taxonomy, Abund.DNA.120 = Abundance, Query)  
DNA\_135 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/135m/135m\_DNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.DNA.135= Confident\_Taxonomy, Abund.DNA.135= Abundance, Query)  
DNA\_150 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/150m/150m\_DNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.DNA.150= Confident\_Taxonomy, Abund.DNA.150= Abundance, Query)  
DNA\_165 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/165m/165m\_DNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.DNA.165= Confident\_Taxonomy, Abund.DNA.165= Abundance, Query)  
DNA\_200 <-read\_tsv("~/MICB425\_portfolio/gcloud raw outputs/200m/200m\_DNA/final\_outputs/marker\_contig\_map.tsv")%>%  
 select(Tax.DNA.200= Confident\_Taxonomy, Abund.DNA.200= Abundance, Query)  
  
  
napA.all= DNA\_10 %>%  
 full\_join(RNA\_10, by="Query") %>%  
 full\_join(DNA\_100, by="Query") %>%  
 full\_join(RNA\_100, by="Query") %>%  
 full\_join(DNA\_120, by="Query") %>%  
 full\_join(RNA\_120, by="Query") %>%  
 full\_join(DNA\_135, by="Query") %>%  
 full\_join(RNA\_135, by="Query") %>%  
 full\_join(DNA\_150, by="Query") %>%  
 full\_join(RNA\_150, by="Query") %>%  
 full\_join(DNA\_165, by="Query") %>%  
 full\_join(RNA\_165, by="Query") %>%  
 full\_join(DNA\_200, by="Query") %>%  
 full\_join(RNA\_200, by="Query")%>%  
  
 mutate(Taxonomy = ifelse(!is.na(Tax.RNA.10), Tax.RNA.10,  
 ifelse(!is.na(Tax.DNA.10), Tax.DNA.10,  
 ifelse(!is.na(Tax.RNA.100), Tax.RNA.100,  
 ifelse(!is.na(Tax.DNA.100), Tax.DNA.100,  
 ifelse(!is.na(Tax.RNA.120), Tax.RNA.120,  
 ifelse(!is.na(Tax.DNA.120), Tax.DNA.120,  
 ifelse(!is.na(Tax.RNA.135), Tax.RNA.135,  
 ifelse(!is.na(Tax.DNA.135), Tax.DNA.135,  
 ifelse(!is.na(Tax.RNA.150), Tax.RNA.150,  
 ifelse(!is.na(Tax.DNA.150), Tax.DNA.150,  
 ifelse(!is.na(Tax.RNA.165), Tax.RNA.165,  
 ifelse(!is.na(Tax.DNA.165), Tax.DNA.165,  
 ifelse(!is.na(Tax.RNA.200), Tax.RNA.200,  
 ifelse(!is.na(Tax.DNA.200), Tax.DNA.200,  
 "unclassified"))))))))))))))) %>%  
 select(-starts\_with("Tax.")) %>%  
 gather("Key", "Abundance", starts\_with("Abund")) %>%  
 separate(Key, c("Key","Type","Depth\_m"), by = ".") %>%  
 select(Depth\_m, Type, Abundance, Taxonomy, Query) %>%  
 mutate(Depth\_m = as.numeric(Depth\_m)) %>%  
 separate(Taxonomy, into = c("Domain", "Phylum", "Class", "Order", "Family", "Genus", "Species"), sep="; ")

## Code to Generate Taxonomy and Relative Abundane Dataset for both RNA and DNA at each depth

RNA\_10m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/10m/10m\_RNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
RNA\_10m["Domain"]<-RNA\_10m$Taxonomy  
RNA\_10m[8] <- lapply(RNA\_10m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
RNA\_10m\_sep = separate(data=RNA\_10m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
RNA\_10m\_sep["Depth"]<- "10\_m"  
R1<-select(RNA\_10m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
R1["Perc"]<-R1[2]\*(1/sum(R1$Abundance))  
  
  
RNA\_100m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/100m/100m\_RNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
RNA\_100m["Domain"]<-RNA\_100m$Taxonomy  
RNA\_100m[8] <- lapply(RNA\_100m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
RNA\_100m\_sep = separate(data=RNA\_100m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
RNA\_100m\_sep["Depth"]<- "100\_m"  
R2<-select(RNA\_100m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
R2["Perc"]<-R2[2]\*(1/sum(R2$Abundance))  
  
RNA\_120m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/120m/120m\_RNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
RNA\_120m["Domain"]<-RNA\_120m$Taxonomy  
RNA\_120m[8] <- lapply(RNA\_120m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
RNA\_120m\_sep = separate(data=RNA\_120m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
RNA\_120m\_sep["Depth"]<- "120\_m"  
R3<-select(RNA\_120m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
R3["Perc"]<-R3[2]\*(1/sum(R3$Abundance))  
  
  
RNA\_135m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/135m/135m\_RNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
RNA\_135m["Domain"]<-RNA\_135m$Taxonomy  
RNA\_135m[8] <- lapply(RNA\_135m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
RNA\_135m\_sep = separate(data=RNA\_135m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
RNA\_135m\_sep["Depth"]<- "135\_m"  
R4<-select(RNA\_135m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
R4["Perc"]<-R4[2]\*(1/sum(R4$Abundance))  
  
RNA\_150m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/150m/150m\_RNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
RNA\_150m["Domain"]<-RNA\_150m$Taxonomy  
RNA\_150m[8] <- lapply(RNA\_150m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
RNA\_150m\_sep = separate(data=RNA\_150m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
RNA\_150m\_sep["Depth"]<- "150\_m"  
R5<-select(RNA\_150m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
R5["Perc"]<-R5[2]\*(1/sum(R5$Abundance))  
  
RNA\_165m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/165m/165m\_RNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
RNA\_165m["Domain"]<-RNA\_165m$Taxonomy  
RNA\_165m[8] <- lapply(RNA\_165m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
RNA\_165m\_sep = separate(data=RNA\_165m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
RNA\_165m\_sep["Depth"]<- "165\_m"  
R6<-select(RNA\_165m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
R6["Perc"]<-R6[2]\*(1/sum(R6$Abundance))  
  
RNA\_200m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/200m/200m\_RNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
RNA\_200m["Domain"]<-RNA\_200m$Taxonomy  
RNA\_200m[8] <- lapply(RNA\_200m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
RNA\_200m\_sep = separate(data=RNA\_200m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
RNA\_200m\_sep["Depth"]<- "200\_m"  
R7<-select(RNA\_200m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
R7["Perc"]<-R7[2]\*(1/sum(R7$Abundance))  
  
  
R\_all=rbind(R1, R2, R3, R4, R5, R6, R7)  
  
#DNA from here on out  
  
DNA\_10m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/10m/10m\_DNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
DNA\_10m["Domain"]<-DNA\_10m$Taxonomy  
DNA\_10m[8] <- lapply(DNA\_10m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
DNA\_10m\_sep = separate(data=DNA\_10m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
DNA\_10m\_sep["Depth"]<- "10\_m"  
D1<-select(DNA\_10m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
D1["Perc"]<-D1[2]\*(1/sum(D1$Abundance))  
  
DNA\_100m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/100m/100m\_DNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
DNA\_100m["Domain"]<-DNA\_100m$Taxonomy  
DNA\_100m[8] <- lapply(DNA\_100m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
DNA\_100m\_sep = separate(data=DNA\_100m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
DNA\_100m\_sep["Depth"]<- "100\_m"  
D2<-select(DNA\_100m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
D2["Perc"]<-D2[2]\*(1/sum(D2$Abundance))  
  
DNA\_120m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/120m/120m\_DNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
DNA\_120m["Domain"]<-DNA\_120m$Taxonomy  
DNA\_120m[8] <- lapply(DNA\_120m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
DNA\_120m\_sep = separate(data=DNA\_120m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
DNA\_120m\_sep["Depth"]<- "120\_m"  
D3<-select(DNA\_120m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
D3["Perc"]<-D3[2]\*(1/sum(D3$Abundance))  
  
  
DNA\_135m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/135m/135m\_DNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
DNA\_135m["Domain"]<-DNA\_135m$Taxonomy  
DNA\_135m[8] <- lapply(DNA\_135m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
DNA\_135m\_sep = separate(data=DNA\_135m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
DNA\_135m\_sep["Depth"]<- "135\_m"  
D4<-select(DNA\_135m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
D4["Perc"]<-D4[2]\*(1/sum(D4$Abundance))  
  
DNA\_150m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/150m/150m\_DNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
DNA\_150m["Domain"]<-DNA\_150m$Taxonomy  
DNA\_150m[8] <- lapply(DNA\_150m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
DNA\_150m\_sep = separate(data=DNA\_150m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
DNA\_150m\_sep["Depth"]<- "150\_m"  
D5<-select(DNA\_150m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
D5["Perc"]<-D5[2]\*(1/sum(D5$Abundance))  
  
DNA\_165m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/165m/165m\_DNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
DNA\_165m["Domain"]<-DNA\_165m$Taxonomy  
DNA\_165m[8] <- lapply(DNA\_165m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
DNA\_165m\_sep = separate(data=DNA\_165m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
DNA\_165m\_sep["Depth"]<- "165\_m"  
D6<-select(DNA\_165m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
D6["Perc"]<-D6[2]\*(1/sum(D6$Abundance))  
  
DNA\_200m<-read.delim("~/MICB425\_portfolio/gcloud raw outputs/200m/200m\_DNA/final\_outputs/marker\_contig\_map.tsv", na.strings=c(""))  
DNA\_200m["Domain"]<-DNA\_200m$Taxonomy  
DNA\_200m[8] <- lapply(DNA\_200m[8], gsub, pattern = "Cystobacterineae;", replacement = "", fixed = TRUE)  
DNA\_200m\_sep = separate(data=DNA\_200m, col=Domain, into=c("Domain", "Phylum", "Class","Order","Family", "Genus", "Species", "Subsp."), sep="\\;", na.strings=c(""))  
DNA\_200m\_sep["Depth"]<- "200\_m"  
D7<-select(DNA\_200m\_sep, Depth, Abundance, Domain, Phylum, Class, Order, Family, Genus, Species, Subsp.)  
D7["Perc"]<-D7[2]\*(1/sum(D7$Abundance))  
  
  
D\_all=rbind(D1, D2, D3, D4, D5, D6, D7)