Spring 5-26-2021

Summer Marine Bacterial Community Composition of the Western Antarctic Peninsula

Codey Phoun

Summer Marine Bacterial Community Composition of the Western Antarctic Peninsula

A Project

Presented to the

Department of Computer Science

San José State University

In Partial Fulfillment of the

Requirements for the Degree

Master of Science

By

Codey Phoun

May 2021

ABSTRACT

The Western Antarctic Peninsula has experienced dramatic warming due to climate change over the last 50 years and the consequences to the marine microbial community are not fully clear. The marine bacterial community are fundamental contributors to biogeochemical cycling of nutrients and minerals in the ocean. Molecular data of bacteria from the surface waters of the Western Antarctic Peninsula are lacking and most existing studies do not capture the annual variation of bacterial community dynamics. In this study, 15 different 16S rRNA gene amplicon samples covering 3 austral summers were processed and analyzed to investigate the marine bacterial community composition and its changes over the summer season. Between the 3 summer seasons, a similar pattern of dominance in relative community composition by the classes of Alphaproteobacteria, Gammaproteobacteria, and Bacteroidetes was observed. Alphaproteobacteria were mainly composed of the order Rhodobacterales and increased in relative abundance as the summer progressed. Gammaproteobacteria were represented by a wide array of taxa at the order level. The class Bacteroidetes had the highest relative abundance in the early summer and decreased as the season progressed. Bacteroidetes were primarily represented by the order Flavobacteriales and genus *Polaribacter*. A high degree of interannual variability was observed for some taxa, like the order Sphingobacteriales, which exhibited a high relative abundance in only 1 season. Richness and evenness diversity measures were found to be at the lowest during phytoplankton blooms, and these diversity measures were observed to increase by the end of the summer. Code written for data processing and analysis are available at: https://github.com/codey-phoun/palmer_station_16S

Keywords: Marine bacteria, 16S rRNA gene amplicons, bacterial community composition, microbial oceanography, Western Antarctic Peninsula, Palmer LTER

TABLE OF CONTENTS

I. Introduction 1
A. Influence of Phytoplankton and Bacteria on Biogeochemical Cycles 1
B. Role of Metagenomics and Bioinformatics in Microbial Ecology 2
C. Past Research of the Bacterial Community in the WAP 3
D. Current Study 3
II. Materials and Methods 4
A. Environmental Data 4
B. 16S rRNA Gene Amplicon Samples 4
C. Sequence Processing 5
D. Microbiome Analysis in R 7
III. Results 10
A. Environmental Context of the Austral Summer Seasons 10
B. Quality Control of 16S Samples 17
C. Alpha Diversity Analysis 20
D. Relative Taxonomic Composition 23
E. Beta Diversity Analysis 26
F. Core Microbiome 29
G. Community Composition Dynamics 29
IV. Discussion 33
A. Summer Bacterial Community Composition 33
B. Limitations of this Study 35
C. Future Research 35
V. Conclusion 35
References 37
Appendices 43
Appendix A. Mid-point Rooted Phylogenetic Tree with Abundances 43
Appendix B. Taxa Names of Core Microbiome 44

LIST OF TABLES

TABLE I. 16S rRNA SAMPLE INFORMATION 5
TABLE II. 16S WATER SAMPLE ENVIRONMENTAL PROPERTIES 14
TABLE III. CUTADAPT QUALITY CONTROL RESULTS 17
TABLE IV. DADA2 QUALITY CONTROL RESULTS 18
TABLE V. ASV TAXONOMY CLASSIFICATION BY VSEARCH 19
TABLE VI. FINAL READ COUNTS OF 16S SAMPLE PROCESSING STEPS 19

LIST OF FIGURES

Figure 1. Bathymetry map of the Palmer Station LTER study area in the WAP 2
Figure 2. Bacterial properties in the austral summer 12
Figure 3. Phytoplankton proxy measurements 12
Figure 4. Chlorophyll *a* and bacterial production measurements 13
Figure 5. Water property measurements 15
Figure 6. Inorganic nutrient measurements 16
Figure 7. Rarefaction curves of 16S samples 20
Figure 8. Alpha diversity measurements 21
Figure 9. Kruskal-Wallis and Wilcoxon rank sum test results 22
Figure 10. Treemap of all 16S samples combined 24
Figure 11. Stacked bar plots of taxa relative frequency 25
Figure 12. Dendrogram of 16S samples 27
Figure 13. Ordination plots of 16S samples 28
Figure 14. Core microbiome of 16S samples 30
Figure 15. Relative abundance at the Order level over the summer season 31
Figure 16. Relative abundance at the Family level over the summer season 32

I. INTRODUCTION

A. Influence of Phytoplankton and Bacteria on Biogeochemical Cycles

Palmer Station is located on Anvers Island off the Western Antarctic Peninsula (WAP), as shown in Figure 1 [1]. The Palmer Station Long Term Ecological Research (PAL-LTER) was established in 1990 to study the marine ecosystem of the WAP [2]. During the austral winter, the coastal waters off of Palmer Station are covered in sea ice and low light levels, but in the austral summer, the water experiences increased solar irradiance, water temperatures, nutrient availability, water stratification, and reduced salinity and sea ice cover. Despite the extreme seasonal variations in the biogeochemical properties of the Southern Ocean, high levels of productivity by marine microorganisms occur in the spring and summer [3], making the WAP is one of the most productive regions in the Southern Ocean. The retreating sea ice and increased solar irradiance in the transition from the austral winter to summer elicits dramatic changes in the microbial ecosystem by inducing phytoplankton blooms in the water [4], [5].

These phytoplankton, bacteria, and other microbes are the primary form of biomass in the Southern Ocean and are crucial for supporting the Southern Ocean's food web [6], [7]. Increased production of phytoplankton-associated marine bacteria and changes in the bacterial community composition soon follow phytoplankton blooms due to the availability of dissolved organic carbon and other nutrients generated by the blooms [8]. These bacteria in the microbial community play a key role in the biogeochemical cycling of nutrients in the ocean. Heterotrophic bacteria are able to degrade and utilize the dissolved organic carbon derived from phytoplankton [9]. When these bacteria are consumed by bacterivores like zooplankton, organic carbon is introduced to higher trophic levels in the food web. Other nutrients like nitrogen and phosphorous are also processed through this microbial loop. Bacteria in turn can influence the phytoplankton community by competing with phytoplankton for the nutrients available in the water or by providing secondary metabolites to stimulate phytoplankton growth [10].

This microbial ecosystem in the Southern Ocean is a major sink for atmospheric CO_2. Overall, the world's oceans are estimated to sequester up to a third of the world's CO_2 from the atmosphere [11], with the Southern Ocean responsible for an estimated 40% of the CO_2 uptake by the world's oceans [12], [13]. Up to 50% of the organic carbon produced by phytoplankton is processed by heterotrophic bacteria [14]. Due to the effects of anthropogenic climate change, the WAP has experienced dramatic warming in the last 50 years [15]. The effects of global warming on this marine microbial ecosystem are not completely clear, but changes to primary production and phytoplankton community composition have been reported [16]. Heterotrophic bacteria are inextricably linked to phytoplankton, and changes in bacterial abundance and community composition can affect the rest of the microbial ecosystem and the higher trophic levels that depend on them. Given the importance of the bacterial community to global biogeochemical cycles, understanding these changes is essential. Temporal surveys of the marine bacterial community composition can be used as reference points to develop an understanding of the taxonomic structure and dynamics of the bacterial communities in marine ecosystems [17]. Information on how the marine bacterial community and functional diversity changes over time

can be used to gain insight into how they may respond and adapt to fill new ecological niches brought by climate change.

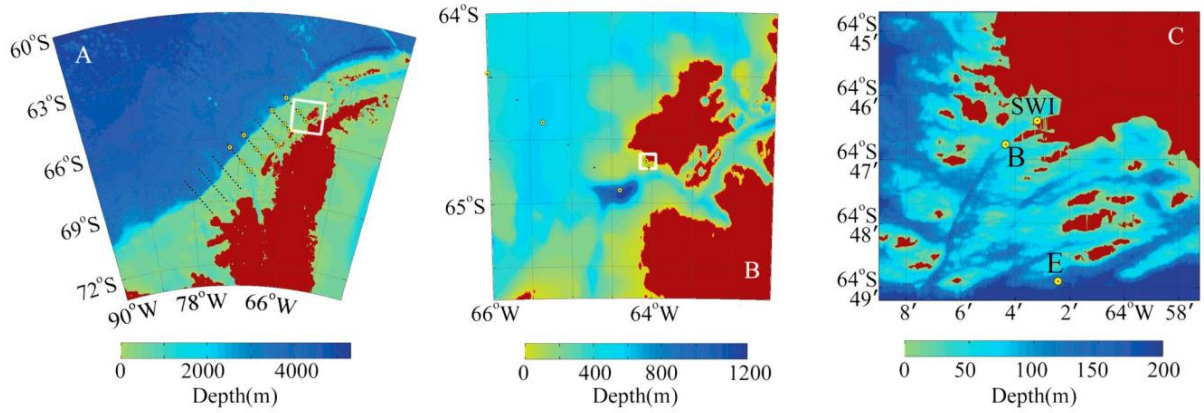


Figure 1. Bathymetry map of the Palmer Station LTER study area in the WAP
(A) Depicts the location of Anvers Island, located within the white box, in context to the WAP
(B) The white box shows the location of Palmer Station on Anvers Island
(C) Yellow dots show Palmer Station LTER Sampling Sites B, E, and seawater intake (SWI)
Figure is adapted from [1].

B. Role of Metagenomics and Bioinformatics in Microbial Ecology

Culture-dependent methods often underestimated marine microbial diversity because the majority of these microbes have not been cultivated [18].Through the use of metagenomic and bioinformatic techniques, a more complete microbiome of an environmental sample can be explored. Amplification and sequencing of the 16S rRNA gene is a commonly used technique in studying the microbial communities of prokaryotes (bacteria and archaea) [19], [20]. The 16S rRNA gene encodes the rRNA portion of the prokaryotic small ribosomal subunit, and this gene is highly conserved in prokaryotes but also has regions of high variability, making it suitable for taxonomic classification [21]. “Universal primers” are used in polymerase chain reaction (PCR) amplification on one or more hypervariable regions of the 16S rRNA gene. The abundance of these 16S gene amplicons is used as a proxy to determine the prokaryotic community composition of a sample. Though these primers attempt to capture all prokaryotic taxa present in the sample, certain groups can be missed depending on the set of primers used [22], [23].

Open-source bioinformatic tools like Cutadapt and QIIME 2 exist to process and analyze 16S gene amplicon data [24], [25]. QIIME 2 is an open-source bioinformatics platform with an array of both native and 3rd party bioinformatics tools available through different plugins. The DADA2 plugin in for QIIME 2 can be used to create amplicon sequence variants (ASVs) from the nucleotide sequencing read data [26]. ASVs have single nucleotide resolution, which are equivalent to 100% identical operational taxonomic units (OTUs) used in older methods. ASV methods have been recommended to be used to replace older OTU methods common in microbial ecology [27]. DADA2 is capable of creating these ASVs by incorporating the quality and abundance information from each sample to create a statistical error model to denoise the reads. In the full DADA2 pipeline, reads are quality filtered, dereplicated, denoised, chimera filtered, and merged. Taxonomic classification is accomplished by comparing ASVs to a

database of known 16S rRNA sequences. Analysis of these sequences can provide a taxonomic resolution down to the genus or even species level.

After the sample community composition is determined by taxonomic classification, this data can be integrated with environmental data to calculate various ecological community measures. For instance, the relative abundance of a different taxa of interest can be plotted over the course of the sampling period to observe the temporal dynamics of the bacterial community composition. Several different environmental measurements can be used to describe how the ecological context changes over time with bacterial community composition [28]. Bacterial abundance and bacterial production data are used to help measure the bacterial biomass and activity present in the water. Chlorophyll *a* concentration serves as a proxy for the amount of phytoplankton biomass. Primary production rates indicate the amount of uptake of inorganic carbon by the microbes through photosynthesis. Dissolved inorganic such as nutrients phosphate, silicate, and nitrite and nitrate are key sources of nutrition essential for phytoplankton growth. Temperature and salinity measurements help to describe the physical characteristics of the water. The incorporation, analysis, and visualization of the taxonomic classification results, community abundance data, and the environmental data can be completed with R and various metagenomic and ecological analysis packages.

C. Past Research of the Bacterial Community in the WAP

Previous studies on surface water bacterial community composition in or near the WAP have shown the classes of Bacteroidetes, Gammaproteobacteria, and Alphaproteobacteria to be dominate during the summer [29]–[38]. Bacteroidetes were primarily represented by the genus *Polaribacter* within the order Flavobacteriales. These bacteria are photoheterotrophs, which utilize sunlight and degrade organic matter available from phytoplankton blooms for energy [39]. Alphaproteobacteria were primarily represented by the orders of SAR11 (also known as Pelagibacterales) and Rhodobacterales. SAR11 has been reported to be the most abundant marine bacterium in the world has been observed at high abundances in both summer and winter populations [40]. Rhodobacterales have been identified to be primary colonizers of marine surface water and are mainly represented by members of the *Roseobacter* clade [41]. Gammaproteobacteria were represented by a mix of different orders dependent on the study. The most abundant orders of Gammaproteobacteria include Alteromondales, Cellvibrionales, Oceanospirillales, SAR86, and Vibrionales. Winter populations have been less studied, but several studies have found bacterial diversity to be at the highest in the winter season and lowest during periods of phytoplankton blooms in the spring and summer [29], [30], [32]. Research has shown winter communities to include chemolithoautotrophic bacteria and archaea [29].

D. Current Study

The PAL-LTER project has been collecting ecological data on the WAP since 1990, but detailed molecular data on the bacterial ecosystem is under sampled. Many of the previous studies surveyed the community dynamics of bacteria over a period of only one season or one year [29]–[35], [37]. While this may capture variability within a single season or between the summer and winter, it does not capture the interannual variability of bacterial community dynamics. Only a few studies have sampled the WAP bacterial composition over more than a single season or year [36], [38]. Long-term surveys of the bacterial community can help

elucidate how the bacterial community will respond and adapt to long-term changes brought by global warming.

In this study, the bacterial community composition in the surface waters of the WAP was explored over the 2012-2013, 2013-2014, and 2014-2015 austral summer seasons. Both seasonal and interannual variability of bacterial community composition was investigated. 16S rRNA amplicon samples were generated through high-throughput sequencing of the microbes in the water samples. The gene amplicon samples were then quality controlled and taxonomically classified with the QIIME 2 platform to determine the bacterial community composition. Data analysis in R included alpha and beta diversity analysis, ordination with the environmental data to investigate the effects of environmental factors on community composition, and relative abundance plots. A core microbiome was also determined to identify taxa which were present in all samples seasonally and annually. These results provide additional data and insight to understanding what bacterial taxa are present in the surface waters of the WAP during phytoplankton blooms in the summer, and how their relative abundances can change over time. Ultimately, this work contributes to the baseline knowledge of how the bacterial community composition and its functional diversity will respond to future ecological changes brought by global warming.

II. MATERIALS AND METHODS

A. Environmental Data

The environmental data of bacteria abundance, bacteria production, chlorophyll *a*, primary production, and dissolved inorganic nutrient concentrations of phosphate, silicate, and nitrite and nitrate were used to explore the biogeochemical context of the different austral summer seasons and the 16S samples (Table II). The date ranges of the environmental data gathered by the Palmer LTER research team at Palmer Sampling Station B ranged from 10/31/2012 to 03/21/2013 for the 2012-2013 summer season. In the 2013-2014 summer season sample dates were from 12/12/2013 to 03/24/2014, and from 11/13/2014 to 03/18/2015 in the 2014-2015 summer season. Environmental sampling of the 2013-2014 season did not begin at Palmer Sampling Station B until late December, due to the presence of sea ice. The environmental data are available at: http://pal.lternet.edu/data

Temperature and salinity data were extracted from raw conductivity, temperature, and depth (CTD) .cnv files with a custom R script called CTD.rmd which utilized the oce package [42]. The script extracts the CTD data closest to 10 meters in depth in the downcast measurements. Environmental data for each austral summer season was imported into R v.4.0.4 and visualized to show the environmental context of the summer seasons and each 16S sample.

B. 16S rRNA Gene Amplicon Samples

16S rRNA gene amplicon samples from the 2012-2013, 2013-2014, and 2014-2015 austral summers were created from sequencing the microbes in water samples collected from Palmer Sampling Station B. The water samples were collected at 10 meters depth by Dr. Shellie Bench as part of her project at Palmer Station. The prokaryotic component of the water samples was isolated by first filtering the water samples through a 3.0 µm filter to remove eukaryotic microbes. The water samples were then filtered through a 0.2 µm filter, and the microbes

retained on the 0.2 µm filter were PCR amplified and sequenced. The V4 hypervariable region of the 16S rRNA gene was amplified using the primers 515F (GTGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) [20]. The 16S rRNA gene amplicons were sequenced to produce forward and reverse reads of 250 base pairs in length. The FASTQ files containing the forward and reverse nucleotide sequence reads of each sample were in a mixed orientation. The "forward" labelled FASTQ sequence files also contained the reverse reads and "reverse" labelled FASTQ files also contained forward reads, which required additional considerations in the data processing pipeline. Of the three sample summer seasons, the 2012-2013 season had only 2 samples. 2013-2014 had 6 samples in total, and the 2014-2015 season had 7 samples in total. A total of 15 samples comprised of 12,071,180 total paired end reads were processed and analyzed in this study (Table I).

TABLE I. 16S RRNA SAMPLE INFORMATION

Water Sample Date	Sample Name	Sample Season	Summer Stage	Total Paired End Reads	Total Base Pairs
11/27/2012	S1L13	2012-2013	Early Summer	831,275	415,637,500
2/8/2013	S1L14	2012-2013	Mid-Summer	873,646	436,823,000
12/27/2013	S2L05	2013-2014	Early Summer	1,004,956	502,478,000
1/23/2014	S2L06	2013-2014	Mid-Summer	987,423	493,711,500
2/3/2014	S2L07	2013-2014	Mid-Summer	997,341	498,670,500
2/10/2014	S2L08	2013-2014	Mid-Summer	921,322	460,661,000
2/28/2014	S2L09	2013-2014	Late Summer	1,027,542	513,771,000
3/4/2014	S2L10	2013-2014	Late Summer	954,770	477,385,000
12/1/2014	S3L03	2014-2015	Early Summer	609,390	305,913,780
12/11/2014	S3L04	2014-2015	Early Summer	631,963	317,245,426
1/12/2015	S3L05	2014-2015	Mid-Summer	616,212	309,338,424
1/19/2015	S3L06	2014-2015	Mid-Summer	789,307	396,232,114
2/9/2015	S3L07	2014-2015	Mid-Summer	612,641	307,545,782
2/23/2015	S3L08	2014-2015	Late Summer	605,620	304,021,240
3/9/2015	S3L09	2014-2015	Late Summer	607,772	305,101,544

C. Sequence Processing

The pipeline 16s_full_pipeline.sh was created to process the 16S rRNA gene amplicon samples with Cutadapt and QIIME 2 on the San Jose State University College of Science High-Performance Computing Cluster (SJSU CoS HPC). Cutadapt v.3.10 was used to trim low quality bases from the 5' and 3' end of the forward and reverse reads (Table III). A minimum Phred quality score of 20, which equates to a 1% error rate, was set as the cutoff before a base would be trimmed from either end of the read. The 515F and 806R primers and Illumina adapters were removed from the forward and reverse reads with a minimum overlap of 10 base pairs and with a maximum mismatch of 1 base pair. Each primer and adapter's forward and reverse complement

sequence were checked for in both forward and reverse FASTQ files in each sample due to the mixed orientation of the reads.

The samples for each summer season were import separately into QIIME 2 v.2020.11 for further processing. The DADA2 plugin for QIIME 2 was used to create ASVs for each 16S rRNA gene amplicon sample. With DADA2 v.1.18.0, each summer season was processed separately to estimate the statistical error model unique to each sequencing run. All forward and reverse reads were first quality filtered. Any read with more than 4 expected errors was discarded by DADA2. Identical sequences were then dereplicated and the parameters for the error model were estimated based on the abundance and quality information. The DADA2 denoising algorithm then performs error correction on the nucleotide sequences based on the estimated error model to create ASVs. Forward and reverse reads are then merged together by DADA2 to create the final ASVs if the minimum criteria of an overlap of 20 bases is met. Each sample then underwent chimera detection and filtering to remove these contaminants (Table IV).

Chimeras are sequence artifacts formed in the PCR amplification process that are combination of two or more biological sequences. The min-fold-parent-over-abundance parameter sets the minimum fold threshold for determining if a sequence could be considered as a “parent” sequence to a potential chimera sequence. A default value of 1 indicates that a potential parent sequence must be more abundant that the potential chimera sequence. The parameter was changed from the default of 1 to 8 in order to decrease the number of false positive chimeras from being detected and filtered out. Initial processing runs with a default value of 1 led to over 30% of the reads to be flagged and removed as chimeric. The separate summer season samples were then combined to create a single QIIME 2 object for taxonomic classification.

ASVs were taxonomically classified with VSEARCH v.2.7.0 against the SILVA SSU Ref NR 99% v.138 database with the feature-classifier plugin for QIIME 2 [43]. The SILVA database contains a curated collection of taxonomically labeled and non-redundant 16S rRNA gene sequences that have been previously dereplicated by clustering at a 99% sequence identity threshold [44]. The QIIME 2 RESCRIPt v.2020.11.1 plugin was used to import the SILVA database and reverse transcribe the sequences from RNA to DNA [45]. The classify-consensus-vsearch command was used for the taxonomic classification of the ASVs.

VSEARCH conducts a global sequence alignment with both the forward and reverse complemented ASVs for taxonomic classification against the SILVA database. Potential matches of a query ASV against a potential reference sequence had to meet a minimum criteria of 80% sequence identity and 80% query coverage. A maximum of 1,000 matches were found for each ASV. Taxonomy was assigned for a query ASV at the lowest taxonomic level by finding the level where a minimum consensus of 51% of the top scoring matches agreed. For example, if a query ASV aligned to the sequences of Rhodobacteraceae;Yoonia-Loktanella, Rhodobacteraceae;Planktomarina, and Rhodobacteraceae;uncultured with equal alignment scores, the consensus taxonomy assigned to the ASV would be the family Rhodobacteraceae with no genus.

ASVs matching mitochondria, chloroplasts, eukaryote, or unassigned were filtered out from the final per-sample ASV abundance table and the ASV taxonomy classification file (Table V). ASVs that were unable to be taxonomically assigned by VSEARCH were extracted with the python script extract_unassigned.py. This script separated out the unassigned ASVs to a single

FASTA file. The unassigned ASVs in the FASTA file were then aligned to the NCBI BLAST 16S rRNA database using blastn with a max of 10 output target sequences. The blast_unassinged.sh script was written and used to perform this additional assignment step to check what taxonomy assignments VSEARCH may have missed.

A *de novo* phylogenetic tree was created for the ASVs with the align-to-tree-mafft-iqtree pipeline in the phylogeny plugin for QIIME 2. Default parameters were used for the pipeline of this plugin. Multiple sequence alignment (MSA) of the ASVs is first handled by MAFFT v7.475 [46]. The ambiguously aligned regions of the generated MSA are then masked by the pipeline to remove potentially misleading and noisy columns in the MSA. A maximum likelihood phylogenetic tree is constructed with IQ-TREE v.2.03 and midpoint rooted by the pipeline [47].

D. Microbiome Analysis in R

The final output files created by the 16s_full_pipeline.sh script for use in the downstream microbiome analysis in R consisted of the per-sample ASV abundance table, the ASV taxonomy classification file, and the phylogenetic tree. These files, along with an additional file containing each sample's metadata, were imported to a phyloseq object with the qiime2R v.0.99.5 package [48]. Phyloseq is an R package used for handling, analyzing, and visualizing microbiome data [49]. The phyloseq object is an object-oriented class that integrates the ASV abundance table, taxonomy information, phylogenetic tree, and sample metadata together as an experiment level object. Other R packages build off the phyloseq object and provide additional functions for processing, analyzing, and visualizing microbiome data. In this study, the packages microbiome v.1.12.0 and phylosmith v.1.0.5 were also used to conduct microbiome analyses [50], [51].

Several processing steps were required before microbiome analysis could be performed on the phyloseq object. First, samples were split into three different summer stages for comparisons that required categorical variables. The austral summer in Antarctica lasts from November through March. Samples between late November through mid-January were assigned as early summer. The mid-summer samples were from mid-January through mid-February, and late summer samples were between mid-February through March. The sample data in the phyloseq object was then agglomerated to the genus level. VSEARCH classified the taxonomy of the ASVs down to the species level if the samples met the minimum consensus criteria, but limitations 16S rRNA V4 region and SILVA database do not provide enough resolution to classify all ASVs to this level. Many of the ASVs are labelled as "uncultured" at the species level, which is not phylogenetically informative. Some analyses also agglomerated the data to even higher levels, such as the taxa levels of phylum, class, and order, to aid in the interpretation of different community composition analyses. Environmental data for each sample was also scaled and centered. For each set of environmental data, centering was performed by subtracting the mean value and scaling was achieved by dividing the centered values by the standard deviation. Finally, the 16S samples were normalized by rarefying the samples to the smallest sample size.

Rarefaction is a widely used method in microbial ecology for normalizing a set of samples for differences in the sample size due to sequencing depth [52]. In general, samples with a higher sequencing depth will capture more species and display a higher diversity measurement. Rarefaction attempts to address this issue by subsampling each sample without replacement to a specified equal size, like the smallest sample size. Rarefaction curves can be plotted to aid in

visualizing how the number of recovered taxa changes with sequencing depth for each 16S sample. These plots can show how, initially, the number of observed taxa rapidly increases with sequencing depth, but the curves will level out to an asymptote if only a few rare taxa remain to be sampled. If a sample curve does not converge to an asymptote, this indicates that the sample needs a higher sequencing depth to fully capture the diversity of the given environment. These results can be used to help guide how different samples may be underestimating the diversity of the environment after rarefaction.

The use of rarefaction is under debate in the literature [53], [54]. The main criticism of this method is that valid data is being discarded, subsampling creates additional uncertainty, and statistical power is reduced for certain analyses like differential abundance testing. Transforming the ASV abundance data to a relative proportion of the total sample size is an alternative normalization method, but this does not address how increased sampling depth tends to also increases the number of taxa in a sample. A sample with low sequencing depth may show 0 for a rare taxa, while a sample with a higher sequencing depth may show a fractional relative abundance value. Other methods proposed for normalizing 16S amplicon samples the log-ratio based centered log-ratio, additive log-ratio, and relative log expression transformations [55]. ASV abundance tables are sparse data sets containing a high proportion of zeros. Log-ratio transformations require the addition of a pseudo-count to the abundance data, as the log of zero is undefined, but the optimal pseudo-count value is also under debate in literature [56], [57]. Rarefaction was chosen as the normalization method for this study because most of the other Antarctic microbial composition profiling studies used rarefaction [29], [30], [33]–[35], [37], [38], with a few exceptions [36]. Normalization by rarefaction may help to facilitate comparisons with these studies. In addition, rarefaction-sensitive differential abundance testing was not conducted in this study.

Relative abundances of 16S rRNA gene amplicon data at different taxonomic levels were visualized with a treemap and stacked bar charts. The overall relative taxonomic composition of all samples combined was visualized at the class and order levels as a treemap. A treemap is a nested hierarchical plot of rectangles that are proportional in size to the relative abundance of each group. Stacked bar charts provided the visualization for each 16S sample's relative taxonomic composition at the phylum, class, order, family, and genus levels. Taxa that contributed less than 2% to the relative abundance of sample were labelled as "Other" to reduce the number of taxa shown in order to improve visualization.

Alpha diversity is the measure of diversity within a single sample and can be described in terms of the number of different observed taxa (richness) and the distribution of the abundance of different taxa (evenness). For this study, the number of observed taxa and the Chao1 index were both used to estimate the richness of each sample [58]. The Chao1 index estimates the expected number of taxa based on the number of rare taxa observed. For richness and evenness measures, Shannon diversity and the inverse Simpson index were used [59]. Shannon diversity gives more weight towards species richness than evenness and measures the uncertainty of predicting the identity of a randomly chosen taxa in the sample. The Simpson index gives more weight towards evenness than abundance and indicates the probability of two randomly sampled taxa of a sample are of a different classification. The inverse Simpson index is used to make this alpha diversity measure follow the same scale as the other measures used in this study, where a higher value indicates higher diversity.

Statistical testing of differences in alpha diversity measures between all summer groups was performed with the Kruskal-Wallis rank sum test. This nonparametric test is an alternative to ANOVA and checks whether all three summer stages come from populations with the same median alpha diversity measure. For pairwise comparisons between the summer stage groups, the Wilcoxon rank sum test was used. This nonparametric alternative to the t-test checks whether two specific summer stages come from populations with the same median alpha diversity metric. All p-values were adjusted for the false discovery rate by the Benjamini-Hochberg method [60].

Beta diversity is the measure of diversity between samples. A distance matrix between all samples was created with different beta diversity measures. Distance matrices are then used for hierarchical clustering and ordination through non-metric multidimensional scaling (NMDS). The Weighted UniFrac distance was used to represent the distances between samples for hierarchical clustering and NMDS [61]. This distance accounts for the relative abundance of taxa shared between samples and also incorporates the phylogenetic tree created earlier by IQ-TREE to determine the phylogenetically-weighted distances between samples. Unweighted Unifrac utilizes only taxa presence and absence information. This is a less appropriate representation of distance for this data set, due to how the relative taxonomic composition of the samples were largely dominated by only a few highly abundant taxa.

A dendrogram of the samples was created by performing hierarchical clustering with complete linkage on the Weighted UniFrac distance of the samples. In complete linkage clustering, also known as furthest neighbor linkage, a clusters are iteratively formed by determining the pair with the shortest distance and then by creating a new distance matrix. The distances between clusters of the new matrix are determined by the furthest pair of points between two clusters. Samples were colored by their summer stage to help depict the results of how this categorical variable clustered.

Ordination is a set of multivariate techniques that can be used to perform dimensionality reduction on a data set to visualize the ecological relationships and trends between samples. The R package vegan v.2.5-7 was used to perform the ordination [62]. Ordination techniques and code were adapted from examples in [63]. The unconstrained methods of NMDS and principal component analysis (PCA) were used in this study to perform exploratory analysis of the 16S samples.

A NMDS plot maximizes the rank correlation between the Weighted Unifrac distances and the plotted two-dimensional distances between all samples. The fit of the sample distances to the ordination plot is measured by a stress value, where plots with low stress values of <0.05 are considered as excellent representation of the data in reduced dimensions [64].

PCA was used to create two new axes which maximizes the overall variance of the abundance data between the 16S samples. Before PCA, the 16S abundance data was Hellinger transformed, as recommended by [65] to account for the sparse nature of the data set. Taxa with low counts and zero counts are given less weight after Hellinger transformation. The envfit function in vegan was used to fit test which environmental gradients best fit onto the ordination plots. Only environmental factors with a p-value ≤ 0.05 in the envfit results were plotted on the NMDS to improve visibility of the samples in the ordination space.

Based on these exploratory plots, constrained ordination by redundancy analysis (RDA) on the Hellinger transformed abundance data was performed to assess how much variation of the

abundance data could be explained by the different significant environmental variables [66]. Permutational multivariate analysis of variance (PERMANOVA) with the adonis function in vegan was then used to test for the difference in centroids and dispersions between the different groups of summer stages. PERMANOVA was also performed on the different austral summer seasons. The betadisper function was used to test for homogeneity of variances for each group.

A core microbiome was established by defining core sets of taxa for the different summer stages and sample seasons at taxa levels order, family, and genus. To be counted as a core taxa for a given group, the taxa must be detected in all samples of the group. Venn diagrams were created to help visualize the core microbiome of the summer stages at different taxa levels. The top ten taxa with the highest abundance proportion of the core microbiome at different taxa levels was also determined. Line plots were created to visualize the change in relative abundance over the austral summer for each sample season.

All code written to perform the data processing and analysis steps are available on the project GitHub repository located at: https://github.com/codey-phoun/palmer_station_16S

III. RESULTS

A. Environmental Context of the Austral Summer Seasons

Bacterial abundance (Figure 2A) ranged from 208,230,769 to 2,943,461,538 num/L for all austral summer seasons, where the maximum observed value occurred on 12/13/2012 in the 2012-2013 season. The peak bacterial abundance for the 2013-2014 summer season occurred on 01/27/2014 with a count of 1,613,384,615 num/L and peaked in the 2014-2015 season on 02/09/2015 at a count of 1,627,538,462 num/L. The average bacterial abundance for the combined three seasons was 671,243,811 num/L with a standard deviation of 386,703,648 num/L.

Bacterial production, measured by the leucine incorporation rate (Figure 2B), followed a similar pattern to the bacterial abundance data. The max value observed throughout the sampling period of this study was 133.63 pmol/L/hr on 12/13/2012 in the 2012-2013 season. The 2013-2014 summer season was nearly able to match this peak rate at 120.69 pmol/L/hr on 01/23/2014. The 2014-2015 summer season peaked at 45.98 pmol/L/hr on 02/03/2015. The average bacterial production of all three austral summer seasons was 31.89 pmol/L/hr with a standard deviation of 25.74 pmol/L/hr.

Phytoplankton bloom biomass was measured through the proxy of chlorophyll *a* concentration (Figure 3A). A large spike in chlorophyll *a* concentration during the early summer of the 2012-2013 signifies a large phytoplankton bloom. The peak of this spike occurred on 11/30/2012 at a concentration of 35.14 mg/m^3. After this spike, the chlorophyll *a* concentration of 2012-2013 had an average value of 2.09 mg/m^3 with a standard deviation of 1.13 mg/m^3. The 2013-2014 summer season had a max recorded chlorophyll *a* concentration of 5.77 mg/m^3 on 02/28/2014, while the 2014-2015 summer season had a recorded maximum concentration of 6.68 mg/m^3 on 01/19/2015. Chlorophyll *a* mean and standard deviation of the 2013-2014 summer season was at 2.81 ± 1.34 mg/m^3. In the 2014-2015 summer season, the mean and standard deviation for chlorophyll *a* was at 2.03 ± 1.57 mg/m^3.

Primary production rates (Figure 3B) showed a peak value of 627.06 mg/m³/day on 12/10/2012, during the large phytoplankton bloom of the early 2012-2013 summer season. A secondary, smaller spike in primary production occurs on 2/12/2013 at 347.17 mg/m³/day. For the total 2012-2013 summer season, the mean and standard deviation primary production was at 109.22 ± 121.92 mg/m³/day. The largest spike in primary production of the three austral summer seasons was on 2/6/2014 at 773.79 mg/m³/day in the 2013-2014 summer season. Overall, primary production in the 2013-2014 season had a mean and standard deviation of 141.26 ± 199.17 mg/m³/day. The 2014-2015 summer season did not have the large spikes in primary production rates observed in the previous seasons. Mean and standard deviation for this season was at 26.90 ± 28.08 mg/m³/day, with a max of 139.97 mg/m³/day on 11/25/2014.

For each austral summer season, water temperature gradually rose in the early and mid-summer before leveling off in the late summer (Figure 5A). Water temperature data ranged from -1.53 to 2.22 °C for the CTD data collected in the 2012-2013 summer season. The average temperature and standard deviation were 0.60 ± 1.10 °C. In the 2013-2014 summer season, the average and standard deviation of the temperature measurements was at 0.59 ± 0.56 °C. The minimum and maximum observed temperature was -0.75°C and 1.81°C, respectively. In the 2014-2015 season, the water temperature average and standard deviation was 0.14 ± 0.74 °C. A range of -1.38 °C to 1.10 °C was observed for this season.

In general, the water salinity data collected for the three austral summer seasons gradually decreased in the early summer to mid-summer periods (Figure 5B). Water salinity data showed a negative correlation to water temperature with a Pearson correlation coefficient R value of -0.47. The 2012-2013 summer season had an average and standard deviation of 33.61 ± 0.23, with a range from 33.09 to 33.95. For 2013-2014, the water salinity average and standard deviation was 33.21 ± 0.33 and had a range of 32.72 to 34.02. In the 2014-2015 summer season, average and standard deviation values were 33.45 ± 0.22 with an observed range of 32.98 to 33.77.

Inorganic nutrient concentrations of phosphate, silicate, and nitrite and nitrate for the three austral summer seasons are shown in Figure 6. These nutrients showed a large degree of variability, with multiple large dips in concentration in the early and mid-summer. The phosphate concentration range and average concentration with the standard deviation was 0.04 to 2.013 µmol/L and 1.47 ± 0.45 µmol/L for the 2012-2013 season; 0.81 to 2.21 µmol/L and 1.28 ± 0.34 µmol/L for the 2013-2014 season; and 0.61 to 2.10 µmol/L and 1.48 ± 0.29 µmol/L for the 2014-2015 season. Silicate concentration ranges and averages with standard deviation for the three austral summer seasons were 50.481 to 65.674 µmol/L and 59.61 ± 3.51 µmol/L for 2012-2013; 40.57 to 73.60 µmol/L and 52.46 ± 8.15 µmol/L for 2013-2014; and 20.98 to 62.01 µmol/L and 52.27 ± 11.46 µmol/L for 2014-2015. Finally, nitrite and nitrate concentration ranges and averages with standard deviation for the three seasons were 3.71 to 29.19 µmol/L and 21.69 ± 5.15 µmol/L for 2012-2013; 8.18 to 31.33 µmol/L and 17.49 ± 5.59 µmol/L for 2013-2014; and 3.56 to 29.34 µmol/L and 19.85 ± 6.11 µmol/L for 2014-2015.

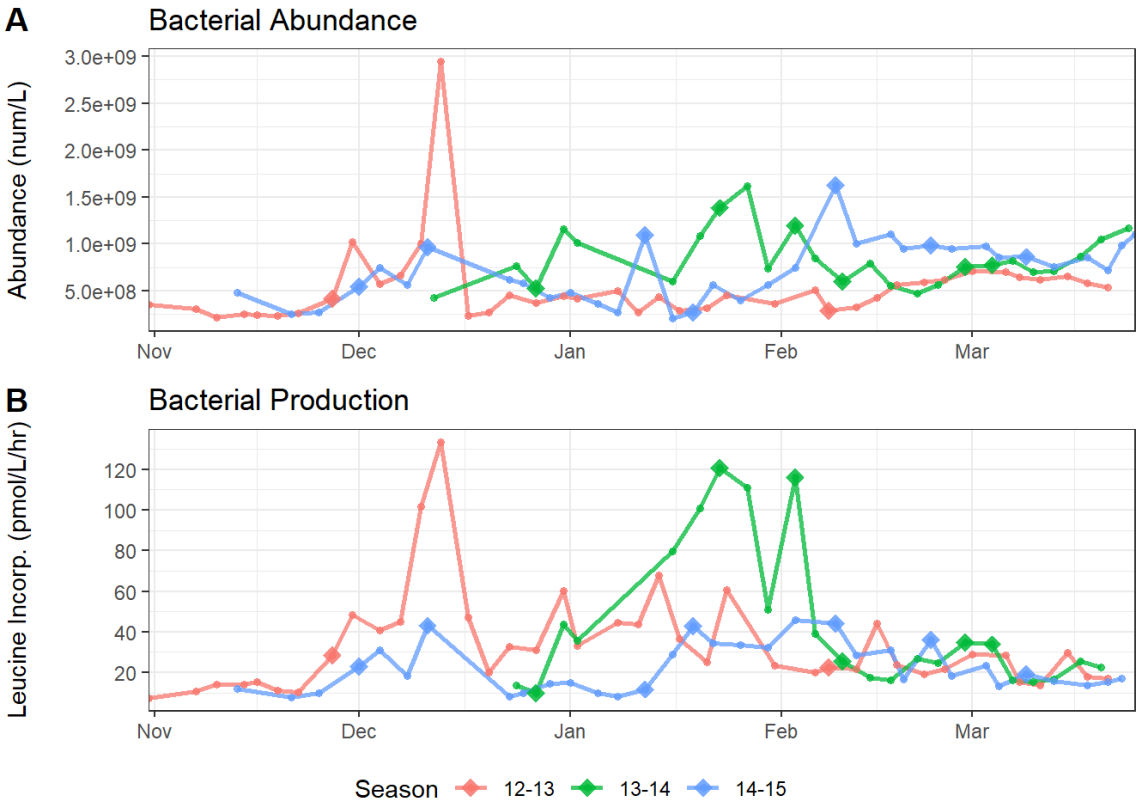


Figure 2. Bacterial properties in the austral summer
(A) Bacterial Abundance and (B) Bacterial Production

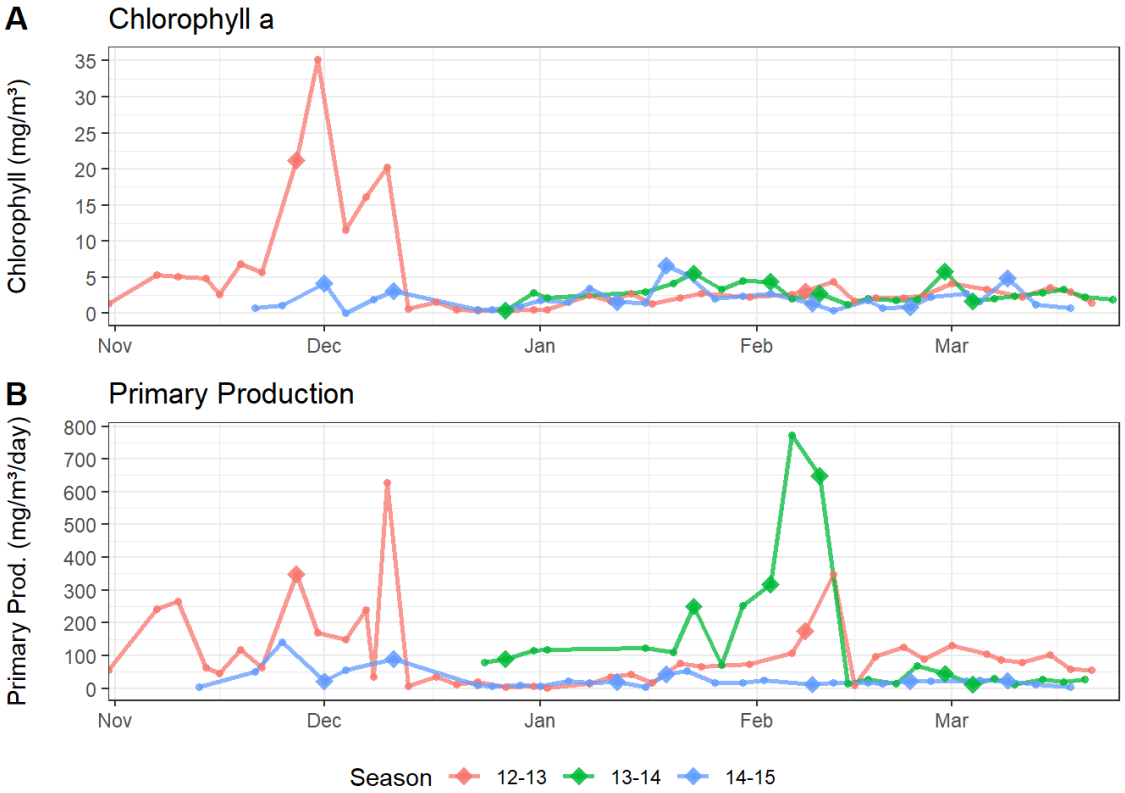


Figure 3. Phytoplankton proxy measurements
(A) Chlorophyll *a* (B) Primary Production

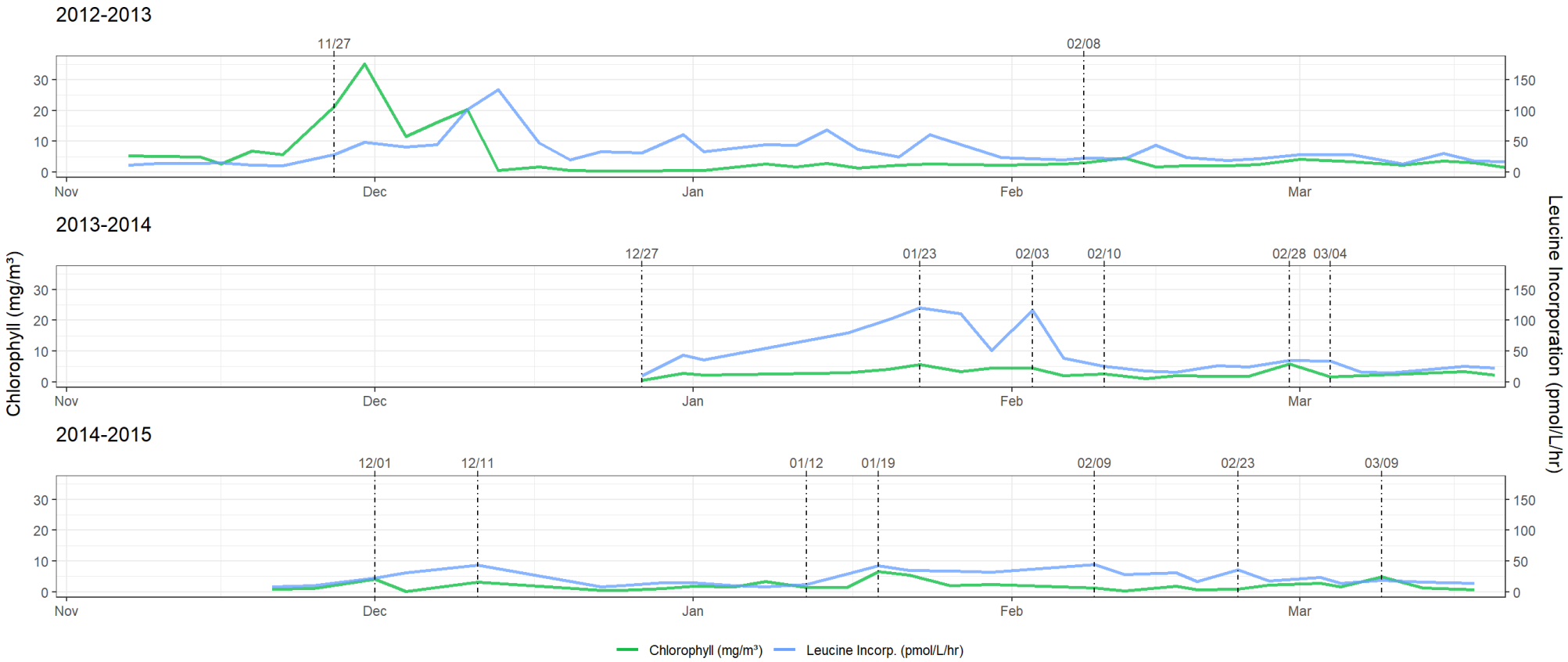


Figure 4. Chlorophyll *a* and bacterial production measurements
Vertical lines denote a 16S rRNA water sample date

TABLE II. 16S WATER SAMPLE ENVIRONMENTAL PROPERTIES

Water Sample Date	Abundance (num/L)	Leucine Incorp. (pmol/L/hr)	Chlorophyll (mg/m^3)	Primary Production (mg/m^3/day)	Phosphate (µmol/L)	Silicate (µmol/L)	Nitrite and Nitrate (µmol/L)	Temperature (°C)	Salinity
11/27/2012	419,846,154	28.3	21.2	349.2	0.6	56.3	7.6	-0.4	33.8
2/8/2013	288,000,000	22.6	3.0	175.6	1.5	58.8	20.9	1.6	33.4
12/27/2013	526,692,308	9.8	0.4	89.5	1.7	65.8	24.3	0.2	33.9
1/23/2014	1,384,357,143	120.7	5.6	251.1	1.0	53.6	11.7	0.8	33.0
2/3/2014	1,197,076,923	116.2	4.4	318.1	0.8	47.1	8.2	1.8	32.8
2/10/2014	597,923,077	25.6	2.7	647.7	1.4	53.6	17.2	0.9	33.2
2/28/2014	755,142,857	34.8	5.8	44.1	0.9	44.6	13.8	0.6	32.7
3/4/2014	775,846,154	33.8	1.7	12.5	1.3	48.9	17.2	0.3	32.9
12/1/2014	543,076,923	22.9	4.1	21.0	1.2	22.2	10.5	-0.7	33.7
12/11/2014	970,692,308	43.1	3.1	89.5	1.7	57.4	21.9	-0.8	33.4
1/12/2015	1,093,714,286	11.8	1.5	20.4	1.4	57.3	18.0	0.4	33.3
1/19/2015	274,000,000	42.8	6.7	42.7	1.2	47.8	12.1	1.0	33.2
2/9/2015	1,627,538,462	44.1	1.4	12.7	1.6	56.8	22.5	0.7	33.4
2/23/2015	984,769,231	36.1	0.9	21.1	1.5	54.2	22.4	0.8	33.3
3/9/2015	862,692,308	19.1	4.9	23.1	1.2	57.6	22.8	0.6	33.5

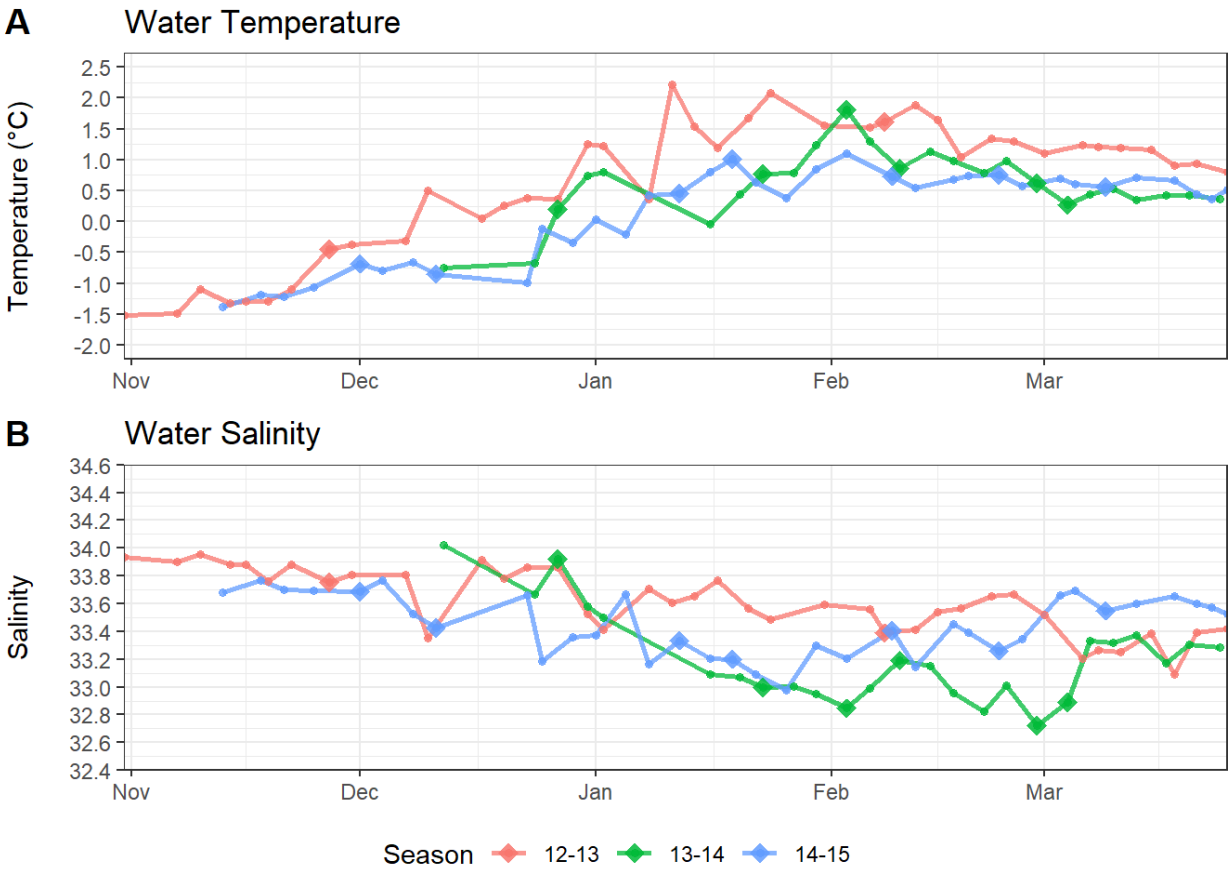


Figure 5. Water property measurements
(A) Water temperature (B) Water Salinity

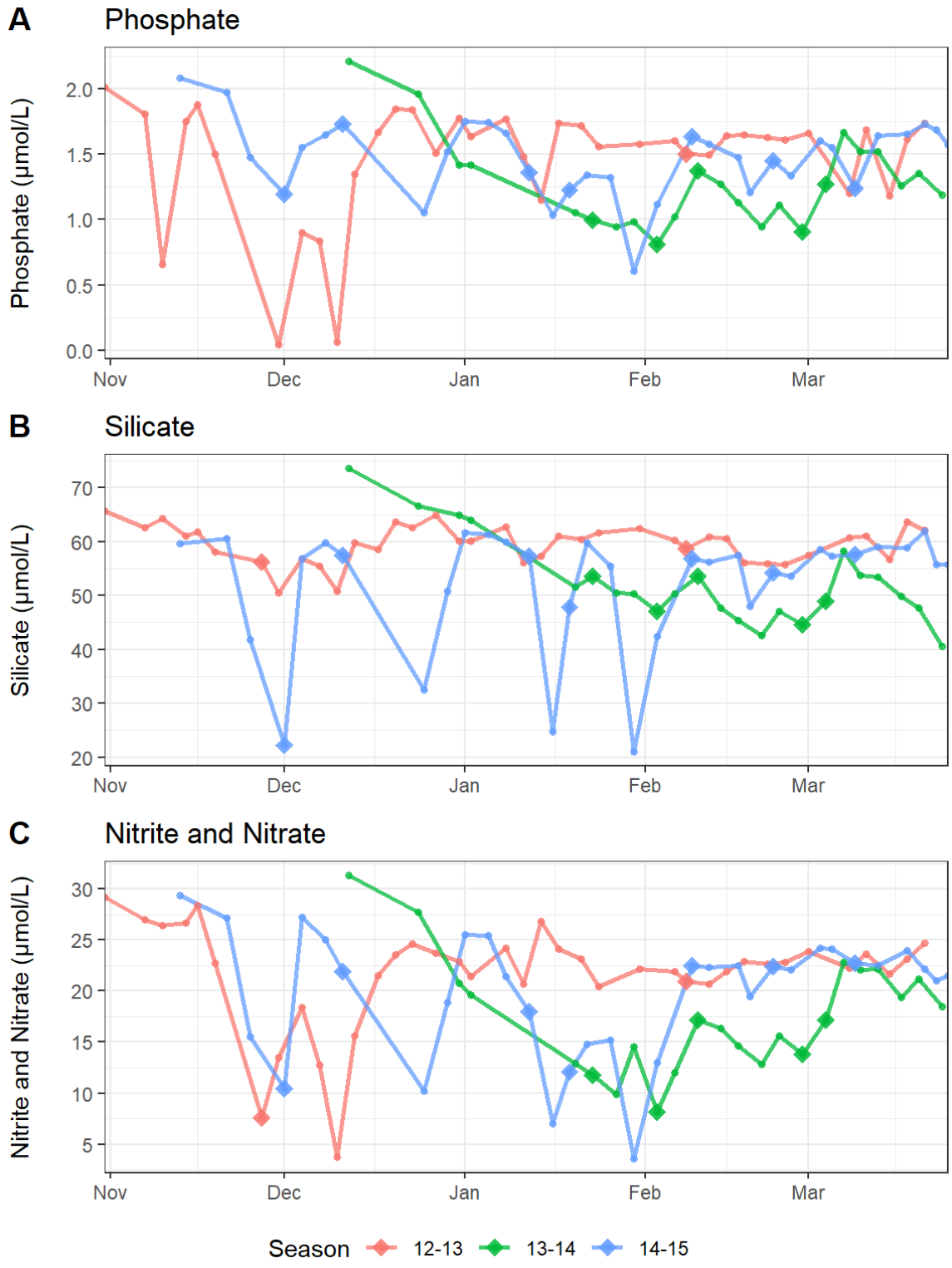


Figure 6. Inorganic nutrient measurements
(A) Phosphate (B) Silicate (C) Nitrite and Nitrate

B. Quality Control of 16S Samples

Table III shows the summarized results of quality trimming and the removal of adapters and primers by Cutadapt. About 8 to 10% of the base pairs of the initial FASTQ read files were removed by Cutadapt for each 16S sample. The 16S samples for the 2013-2014 (S2) were the largest samples and also had the largest amount of base pairs quality trimmed, with several samples reaching over 10 million total base pairs trimmed. Read pairs were not filtered out by Cutadapt.

TABLE III. CUTADAPT QUALITY CONTROL RESULTS

Sample	Paired Reads Processed	Read 1 With Adapters	Read 2 With Adapters	Base Pairs Quality Trimmed	Percent Trimmed
S1L13	831,275	818,871	813,325	3,835,646	9.11
S1L14	873,646	864,795	859,530	3,803,396	9.05
S2L05	1,004,956	997,551	964,192	10,287,476	9.93
S2L06	987,423	980,233	947,895	10,754,182	10.03
S2L07	997,341	989,834	966,295	9,899,680	9.86
S2L08	921,322	914,816	884,391	10,595,391	10.15
S2L09	1,027,542	1,021,658	989,301	11,590,963	10.16
S2L10	954,770	947,395	929,460	7,319,366	9.44
S3L03	609,390	566,585	563,482	1,838,728	8.53
S3L04	631,963	589,787	585,842	2,203,140	8.79
S3L05	616,212	573,532	568,085	2,382,860	8.63
S3L06	789,307	730,898	725,695	2,724,744	8.45
S3L07	612,641	568,066	564,139	2,083,920	8.53
S3L08	605,620	562,542	556,962	2,429,446	8.76
S3L09	607,772	563,111	559,599	2,075,538	8.63
Total	12,071,180	11,689,674	11,478,193	83,824,476	9.3%

Creation of the ASVs by DADA2 in QIIME 2 resulted in an average of 67.7% reads remaining after all quality control processing steps (Table IV). 16S rRNA samples for the 2014-2015 season (S3) lost about 10% of their reads to DADA2's quality filter, where only a maximum of 4 expected errors in each read was accepted. The denoising and merging process removed an additional 8.31 to 15.56% of reads from each sample. An average of 15.6% of reads were removed in the chimera checking step. Proportionally, sample S3L05 from 1/12/2015 lost the least number of reads after all processing steps, where 73.13% of the original reads were kept. In contrast, S2L09 from 2/28/2014 kept only 61.73% of the original reads. DADA2's output resulted in a total of 8,152,255 reads divided among 3,509 unique 16S ASVs. The length of the ASVs ranged from 73 to 487 base pairs and had a mean length of 249.77 base pairs with a

standard deviation of 24.64. The maximum frequency observed for an ASV was 378,067, while the median frequency of reads per ASV was 53.

TABLE IV. DADA2 QUALITY CONTROL RESULTS

Sample Name	Input Reads	Passed Filter	% Passed Filter	Denoised	Merged	% Merged	Non-chimeric	% Non-chimeric
S1L13	831,275	819,824	98.62	806,309	720,349	86.66	534,763	64.33
S1L14	873,646	863,127	98.8	848,249	769,075	88.03	542,836	62.13
S2L05	1,004,956	980,550	97.57	964,671	855,959	85.17	665,962	66.27
S2L06	987,423	963,481	97.58	953,833	848,383	85.92	681,908	69.06
S2L07	997,341	978,700	98.13	968,017	863,715	86.6	697,808	69.97
S2L08	921,322	898,868	97.56	886,183	780,854	84.75	614,562	66.7
S2L09	1,027,542	1,004,682	97.78	986,501	844,799	82.22	634,278	61.73
S2L10	954,770	937,580	98.2	921,660	826,758	86.59	684,382	71.68
S3L03	609,390	558,128	91.59	549,258	499,456	81.96	417,109	68.45
S3L04	631,963	578,114	91.48	565,566	503,698	79.7	415,529	65.75
S3L05	616,212	561,861	91.18	552,741	507,410	82.34	450,614	73.13
S3L06	789,307	713,928	90.45	706,476	648,330	82.14	561,892	71.19
S3L07	612,641	557,055	90.93	547,274	492,857	80.45	427,696	69.81
S3L08	605,620	548,708	90.6	538,005	479,385	79.16	413,606	68.29
S3L09	607,772	551,651	90.77	539,818	478,428	78.72	409,310	67.35
Total	12,071,180	11,516,257	95.40	11,334,561	10,119,456	83.83	8,152,255	67.53

Taxonomic classification of the 3,509 ASVs by VSEARCH in QIIME 2 resulted in 244,263 reads in 283 different ASVs were classified as “Unassigned” (Table V). 23 ASVs with a total of 8,758 reads were classified to the domain Archaea. 120 ASVs with 5,463 reads were classified to the domain Eukaryota. 3083 ASVs with a total of 7,893,771 reads were classified to the domain Bacteria. For the Bacteria classified reads, 475 ASVs with 2,136,552 reads were classified as a chloroplast. 201 ASVs with 233,840 total reads were classified as mitochondria. All ASVs classified as “Unassigned”, Eukaryota, chloroplast, and mitochondria were removed from the ASV results. 2,430 ASVs remained in the Archaea and Bacteria domains after this filtering step. Of the original 12,071,180 paired end reads from the 15 16S rRNA gene amplicon samples, 5,532,137 reads remained after all quality control and filtering steps. S3L05 from 1/12/2015 had the smallest sample size of 193,308 reads, leaving only 31% of the original reads. S2L09 from 2/28/2014 had the largest sample size of 594,964 reads, where 58% of the original reads remained (Table VI).

Samples were imported into R as a phyloseq object and agglomerated to the genus level. Agglomeration of the 2430 ASVs resulted in 384 unique genera. ASVs containing NAs in the taxonomy were also removed in this agglomeration step, resulting in a small loss of reads. S3L05 still had the smallest sample size at 192,952 reads, and all samples were rarefied to this level (Figure 7). 2,894,280 reads, spread evenly between the 15 samples, represented 375 different taxa at the genus level. 52.44% of the 5,518,789 reads are left after rarefaction to the smallest

library size. The number of observed taxa in the rarefaction curves for the 2012-2013 and 2014-2015 samples did not substantially increase after the rarefaction level of 192,952 reads. 2013-2014 samples showed a slight increase in observed taxa past the rarefaction level.

TABLE V. ASV TAXONOMY CLASSIFICATION BY VSEARCH

	ASVs	% of ASVs	Reads	% of Reads
Unassigned	283	8.06%	244,263	2.99%
Archaea	23	0.66%	8,758	0.11%
Eukaryota	120	3.42%	5,463	0.07%
Mitochondria	201	5.73%	233,840	2.87%
Chloroplast	475	13.54%	2,136,552	26.18%
Bacteria	2,407	68.60%	5,523,379	67.75%
Total	3,509		8,152,225	

TABLE VI. FINAL READ COUNTS OF 16S SAMPLE PROCESSING STEPS

Sample Name	Input Reads	DADA2 Output Reads	VSEARCH Output Reads	Phyloseq Output Reads	Final % Remaining
S1L13	831,275	534,763	326,102	324,962	39%
S1L14	873,646	542,836	363,707	362,535	41%
S2L05	1,004,956	665,962	331,698	329,636	33%
S2L06	987,423	681,908	501,364	501,253	51%
S2L07	997,341	697,808	548,281	548,000	55%
S2L08	921,322	614,562	533,557	532,506	58%
S2L09	1,027,542	634,278	594,964	593,411	58%
S2L10	954,770	684,382	496,252	494,396	52%
S3L03	609,390	417,109	221,852	221,245	36%
S3L04	631,963	415,529	275,339	274,018	43%
S3L05	616,212	450,614	193,308	192,952	31%
S3L06	789,307	561,892	377,484	377,212	48%
S3L07	612,641	427,696	290,269	289,685	47%
S3L08	605,620	413,606	242,134	241,639	40%
S3L09	607,772	409,310	235,826	235,339	39%
Total	12,071,180	8,152,255	5,532,137	5,518,789	46%

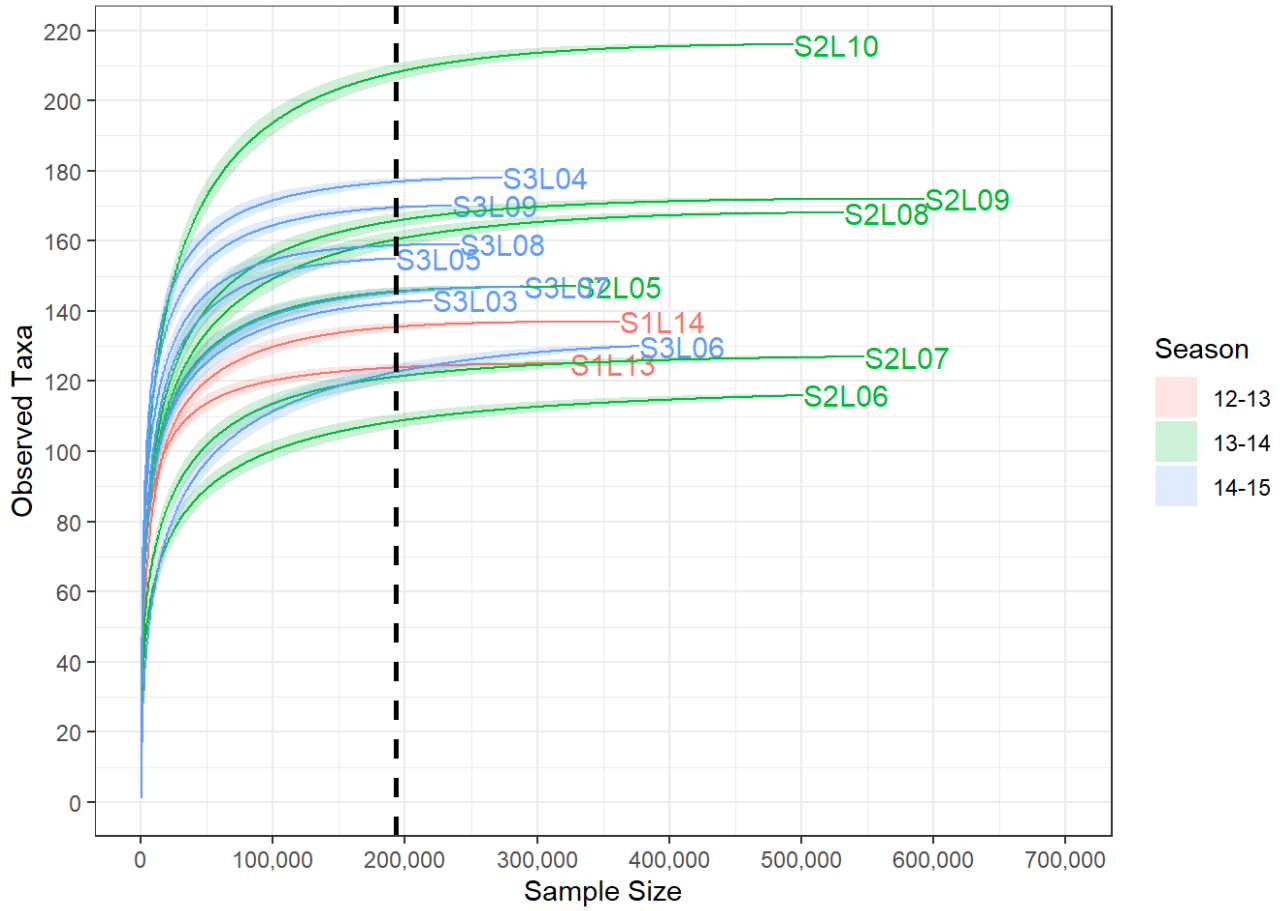


Figure 7. Rarefaction curves of 16S samples
Dashed vertical line shows the minimum library size of S3L05 at 192,952

C. Alpha Diversity Analysis

The alpha diversity metrics of observed taxa, Chao1 index, Shannon diversity, and Inverse Simpson index together showed a similar overall pattern (Figure 8). In general, the diversity measure increases from the early summer to late summer. The late summer water sample S2L10 from 3/4/2014 had the highest richness measures with the number of observed taxa at 209 and estimated taxa by the Chao1 index at 213.32. The mid-summer sample S2L06 from 1/23/2014 had the lowest richness measures with 108 observed taxa and an estimated 115.14 taxa by the Chao1 index. The richness and evenness measures showed mid-summer sample S3L06 from 1/19/2015 to have the lowest Shannon diversity at 1.43 and an Inverse Simpson index at 2.95. The highest values were observed in the late summer on 3/9/2015 with sample S3L09, where the Shannon diversity was 3.05 and the Inverse Simpson index was 13.84.

Results of statistical testing of these alpha diversity measures between the three different summer groups is shown in Figure 9. Kruskal-Wallis rank sum test on alpha diversity measures of the different summer stages showed a p-value under 0.05 for observed taxa, Shannon

diversity, and the Inverse Simpson index, while the test on the Chao1 index had a p-value of 0.052. Mid-summer and late summer comparisons with the Wilcoxon rank sum test had a p-value of 0.054 for observed taxa, 0.036 for Chao1 index, and 0.043 for the Shannon diversity. Between the early summer and late summer groups, only the Wilcoxon test with Shannon diversity showed a p-value under 0.05 at 0.043. The Wilcoxon test did not show any p-values under or near 0.05 between the early summer and mid-summer groups.

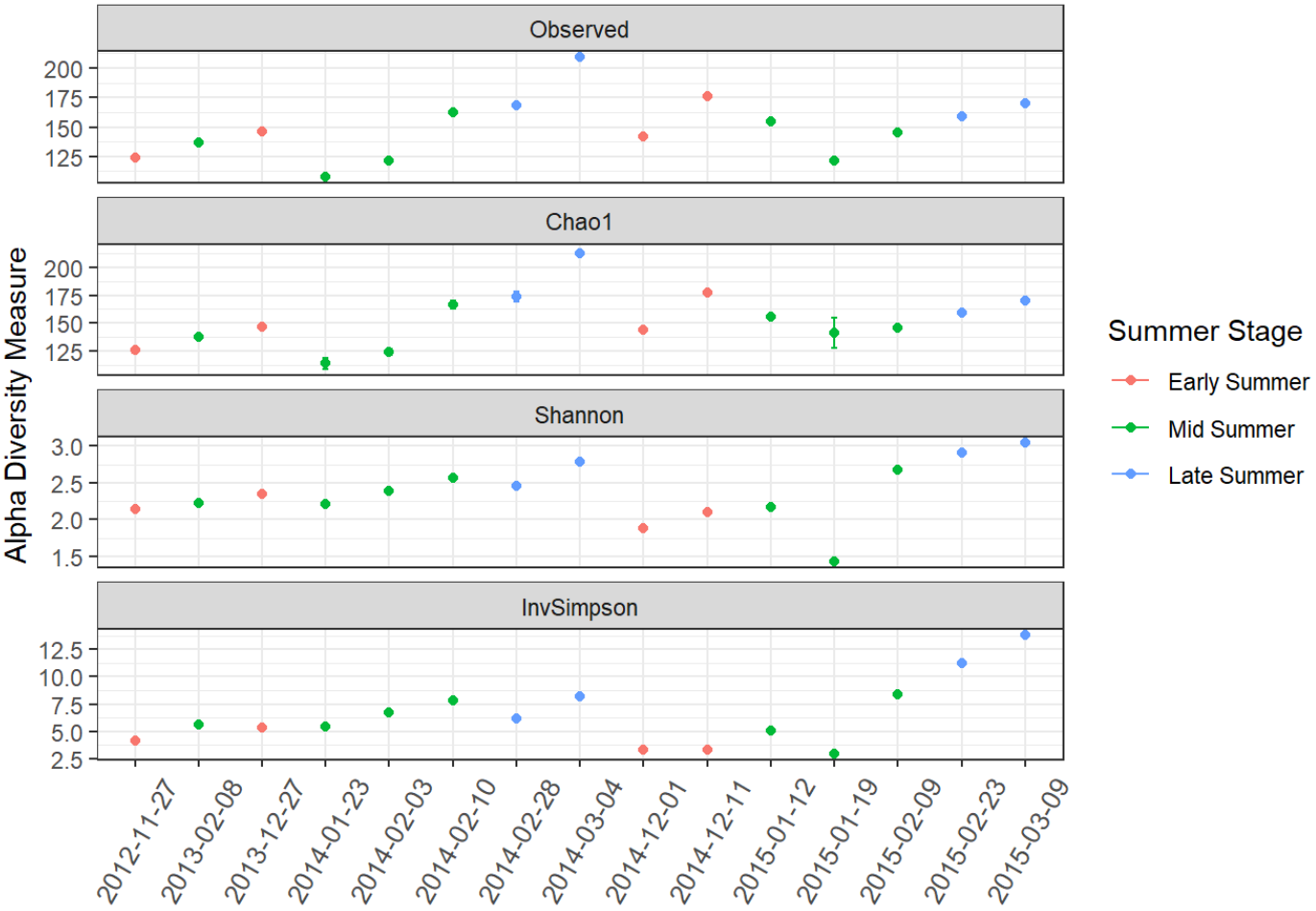


Figure 8. Alpha diversity measurements showing the number of observed taxa, Chao1 index, Shannon diversity, and the inverse Simpson index

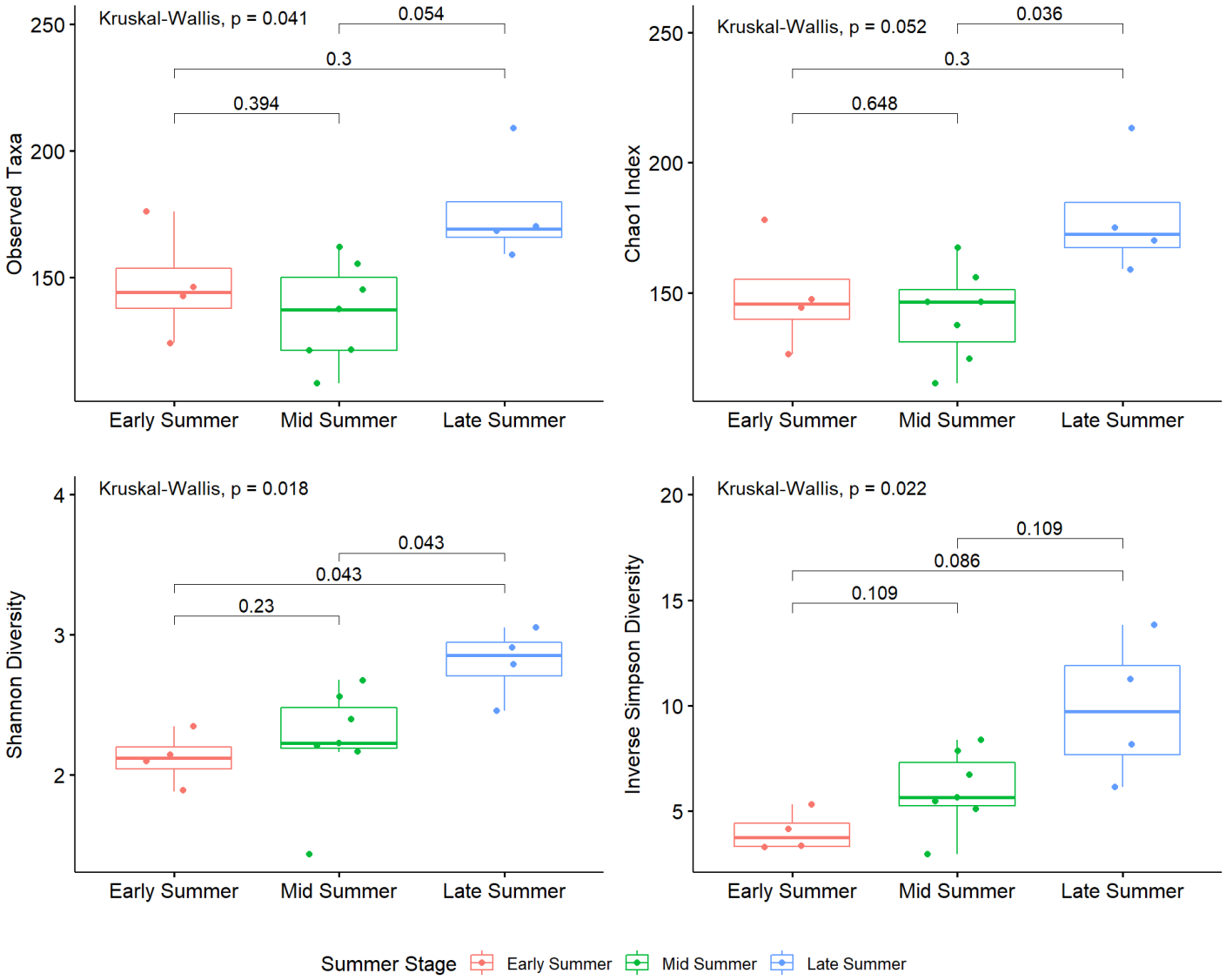


Figure 9. Kruskal-Wallis and Wilcoxon rank sum test results on alpha diversity measurements for each summer stage. Horizontal brackets indicate pairwise group comparison and the resulting p-value of the Wilcoxon rank sum test.

D. Relative Taxonomic Composition

A treemap of all 16S samples in the phyloseq object showed the relative taxonomic composition of all reads in the combined samples (Figure 10). At the class and order levels, the treemap was primarily dominated by three classes. 42.03% of the reads belonged to the class Bacteroidia. 29.77% were of the class Gammaproteobacteria and 27.16% Alphaproteobacteria. The class Bacteroidia mainly consisted of the order Flavobacteriales, which contributed 40.37% of the total 16S reads. The next largest order in Bacteroidia were the Sphingobacteriales which contributed 1.14% to the total reads. Gammaproteobacteria were more diverse in terms of the number of different taxa at the order level that contributed relatively large numbers of reads. For these taxa, Oceanospirillales had an overall relative abundance of 18.13%, Cellvibrionales contributed 3.29%, Nitrosococcales contributed 2.10%, Thiomicrospirales contributed 1.88%, Burkholderiales contributed 1.62%, and Alteromonadales contributed 1.44%. Similar to the Bacteroidia class, the Alphaproteobacteria class was also mainly composed of just a single order, where the order Rhodobacterales had a relative abundance of 25.19%. The SAR11 clade contributed 0.77% to the total reads. The class Actinobacteria contributed 0.45% to the relative abundance of 16S reads.

Stacked bar charts of sample-wise relative frequency for phylum to genus level taxa are shown in Figure 11. The taxonomic composition per sample is not static and changes in the relative abundance were observed throughout the summer seasons. At the phylum and class level (Figure 11A), the relative abundance per sample was dominated by Proteobacteria and Bacteroidia. On 1/19/2015, Alphaproteobacteria had a relative abundance of 76.24%. Gammaproteobacteria contributed 3.75 % while Bacteroidia contributed 19.7% on this date. The dominance in the relative abundance of Alphaproteobacteria on this date was attributed mainly to the order and family Rhodobacterales *Rhodobacteraceae*. At the genus level, *Rhodobacteraceae* were split between *Yoonia-Loktanella* at 49.75% and *Sulfitobacter* at 24.96% on sampling date 1/19/2015. Oceanospiralles were present in all samples except for this date. Their relative abundance dropped to below 5% on 1/12/2015 and then disappears from the stacked bar chart on 1/19/2015. Similar to the results of the treemap, the class Bacteroidia was mainly composed of the order Flavobacteriales in all samples. The genus *Polaribacter* are the main contributors to the abundance of Flavobacteriales. The family *Cryomorphaceae*, in the order Flavobacteriales, had a high relative abundance of 17.72% on 1/12/2014. The order Sphingobacteriales had a relatively high abundance of 9.78% on the sampling date 1/23/2014 but did not show a large abundance in any other sample. The taxa of the order Nitrosococococcales and family *Methylophagaceae* was only noticeable in the 2013-2014 summer season and appears in all but the first sample of this season. At the genus level, nearly half of the relative frequency is either labelled as Uncultured or contribute less than 2% to the relative abundance.

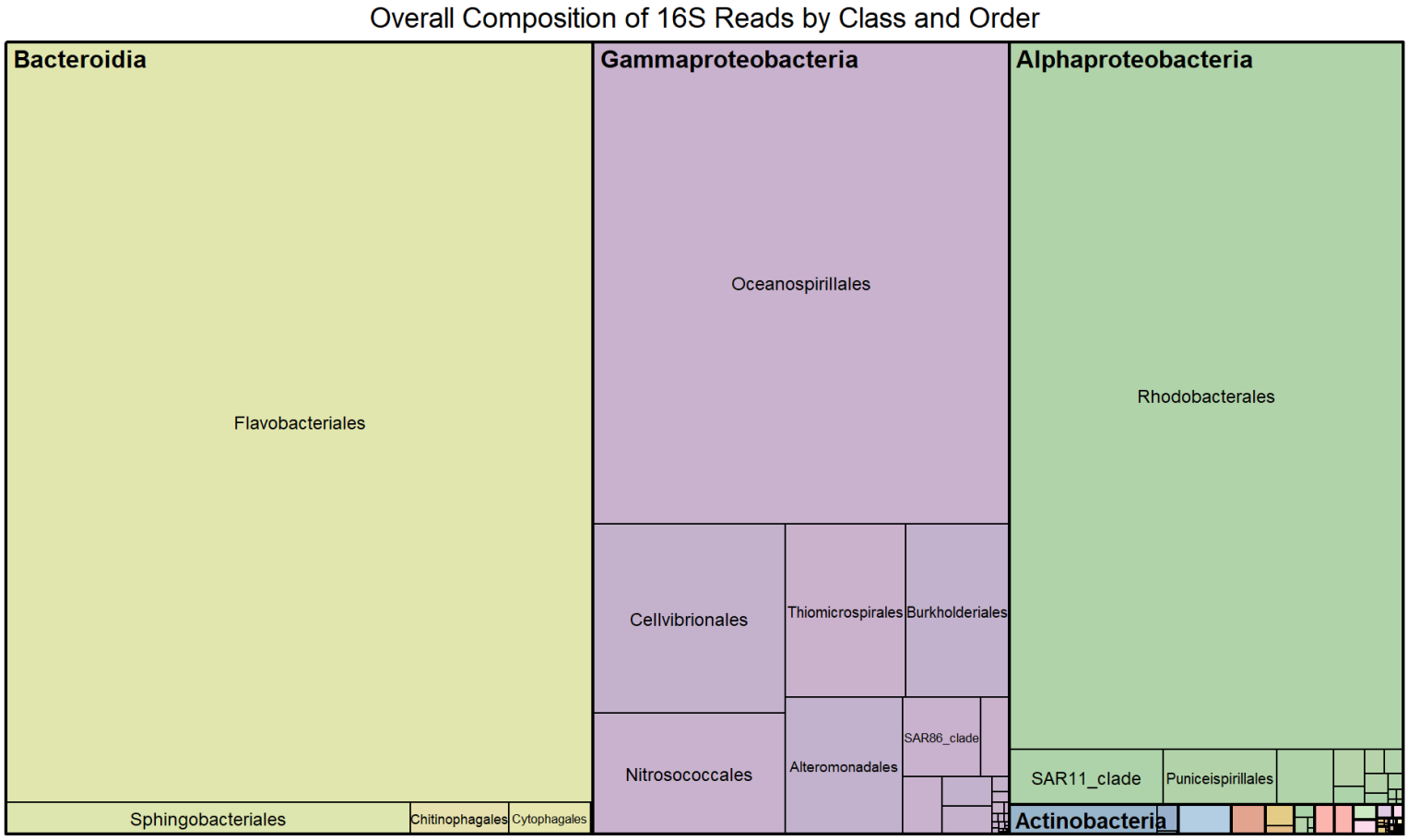


Figure 10. Treemap of all 16S samples combined
to show the relative abundance of the reads by class and order taxonomy levels

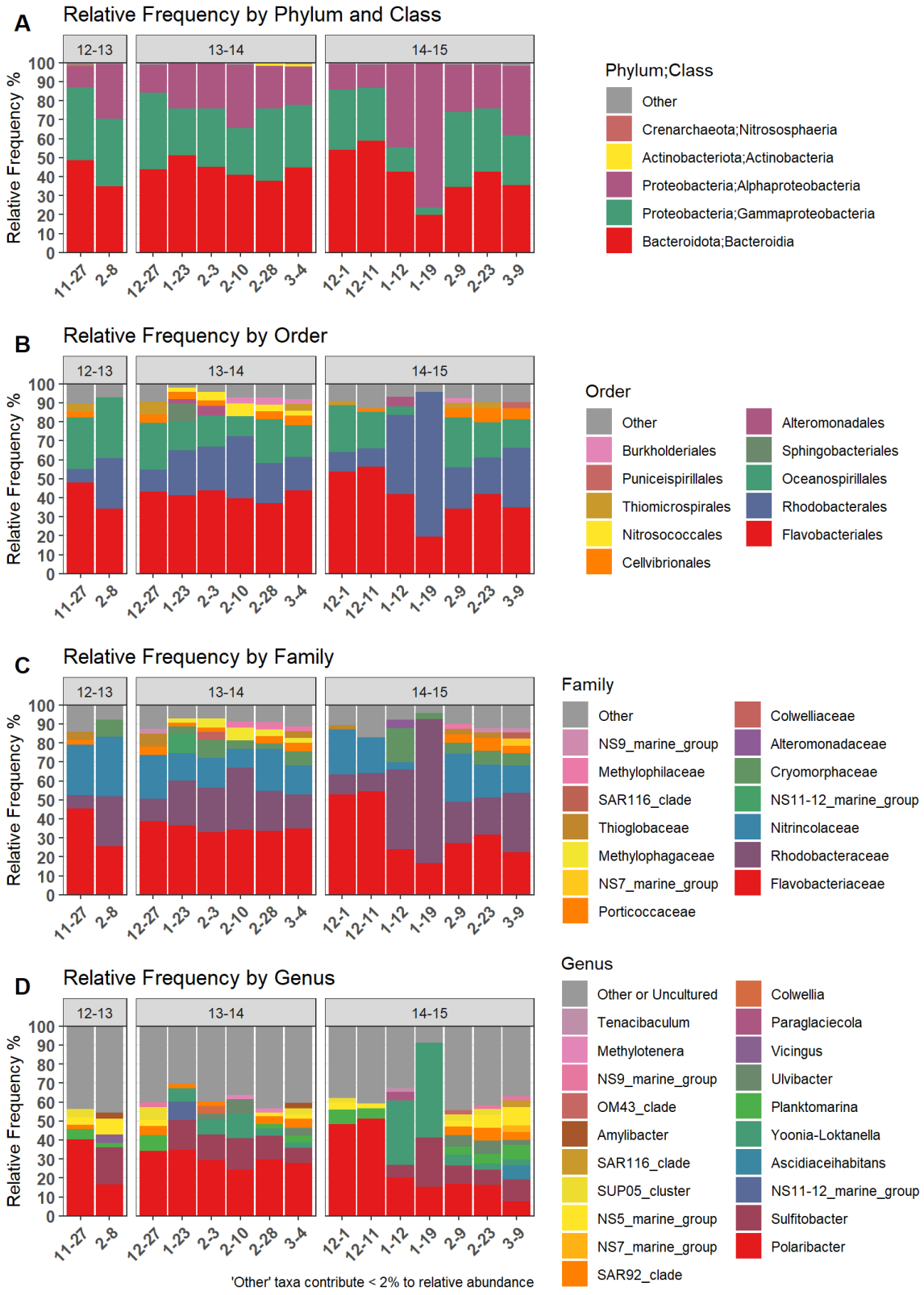


Figure 11. Stacked bar plots of taxa relative frequency
(A) Phylum and Class level (B) Order level (C) Family level (D) Genus level

E. Beta Diversity Analysis

Hierarchical clustering on the Weighted Unifrac distance by complete linkage showed three clusters at a distance of 0.08 (Figure 12). One branch of four samples consisted solely of mid-summer groups from both the 2013-2014 and 2014-2015 summer seasons. The late summer sample from 3/9/2015 is a single branch by itself until a distance at about 0.11. The third cluster is a mix of early summer, mid-summer, and late summer groups. At distances under 0.04, early summer samples do not cluster with any late summer samples.

Unconstrained ordination by PCA and NMDS is shown in Figure 13A and 13B. The first two principal components of the PCA plot explain 65.5% of the variation in the Hellinger transformed abundance data [65]. Environmental factors projected onto the PCA with the envfit function in vegan showed temperature, nitrite and nitrate, and salinity to have the strongest correlations to the first two principal components. Salinity and temperature gradients showed a strong correlation to the first principal component, while nitrite and nitrate appear to correlate with the second principal component. Temperature correlated with both axes. Statistical significance of the fitting of these variables to the first two principal components showed p-values of 0.020 for temperature, 0.068 for nitrite and nitrate, and 0.163 for salinity. NMDS on the Weighted Unifrac distances shows a similar pattern in the relative positioning of the samples. The NMDS plot has a stress value of 0.034. Only environmental factors with a p-value under 0.05 with the envfit results were shown on the NMDS ordination. Constrained ordination by RDA on temperature, nitrite and nitrate, and salinity is shown in Figure 13C. 24.57% of the variation in the abundance table could be explained by temperature, 8.98% by nitrite and nitrate, and 3.33% by salinity. The total proportion of variation explained by these three variables was 36.88%.

A PERMANOVA test with the adonis function in vegan tested for differences in the centroids and dispersions in the different summer stages with the Hellinger transformed taxa abundance data. With the summer stages, the PERMANOVA p-value was 0.0003, indicating support for differences in the centroid and/or dispersions of the summer groups. The test was repeated for the different sample seasons, and the p-value was 0.2481. The summer stages failed the assumption of homogeneity of variance when tested with the betadisper function in vegan. This test reported a p-value of 0.026. For the category of different sample seasons, this test reported a p-value of 0.2702.

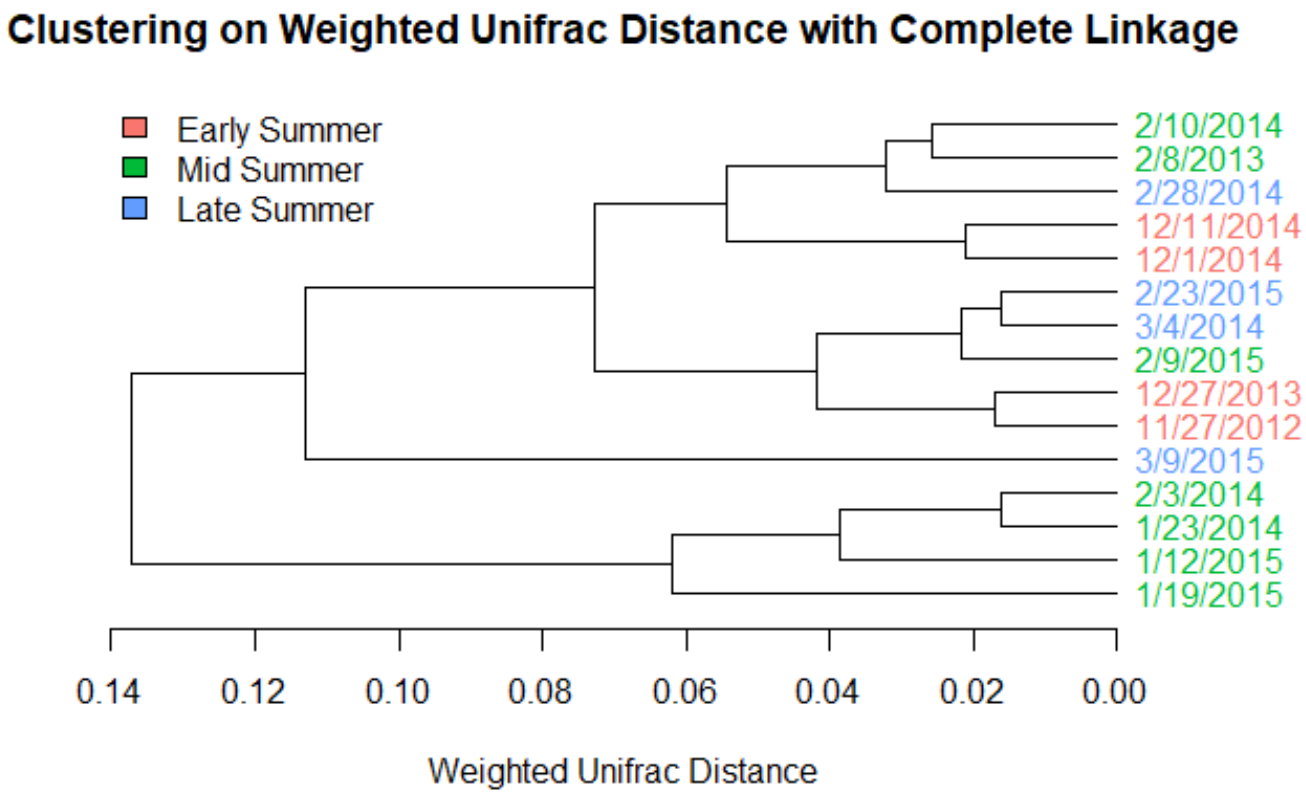


Figure 12. Dendrogram of 16S samples clustered on the Weighted Unifrac distance by complete linkage and colored by summer stages

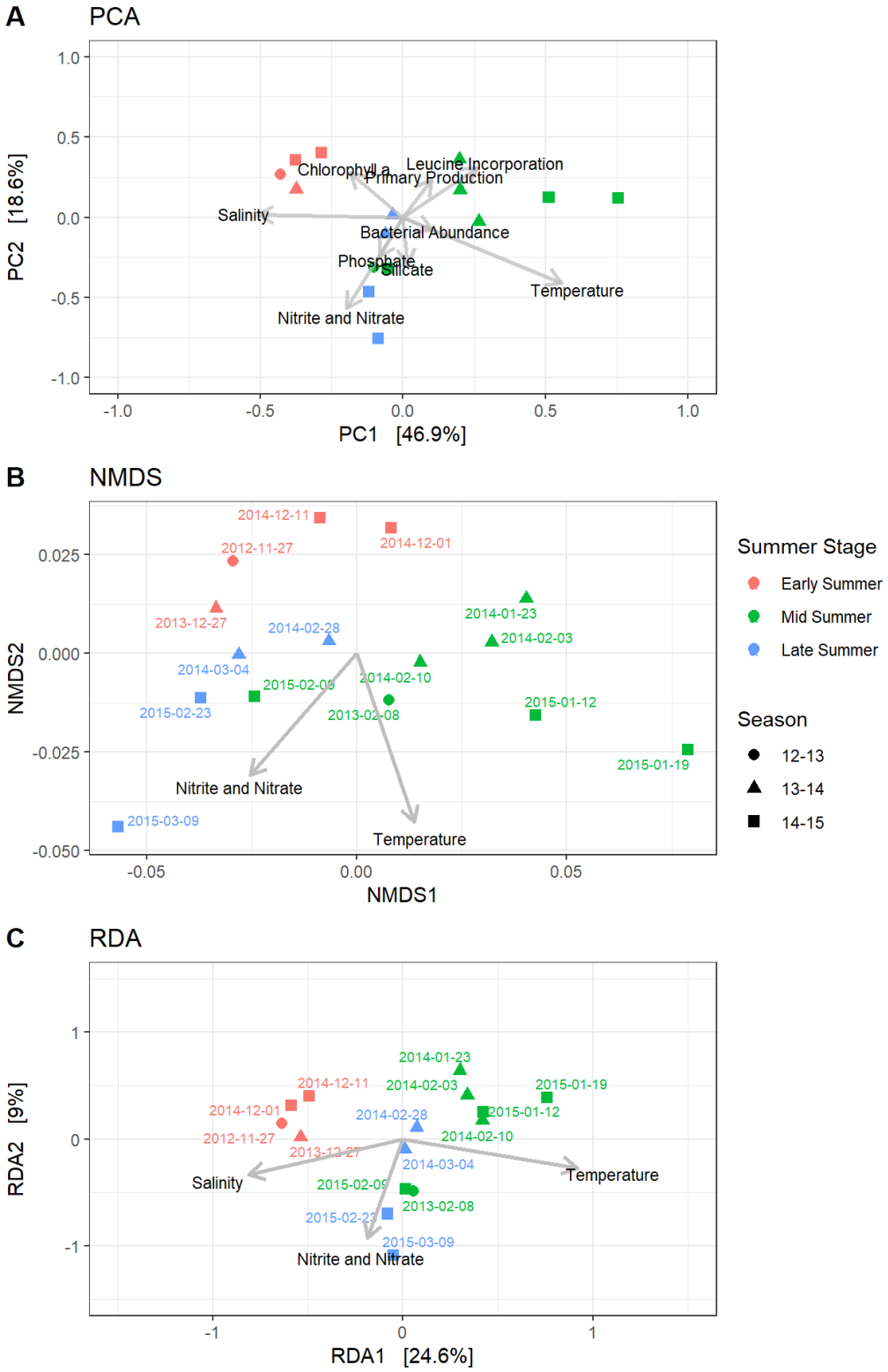


Figure 13. Ordination plots of 16S samples
(A) PCA on Hellinger distance and fit with environmental vectors
(B) NMDS by Weighted Unifrac distance and fit with significant environmental vectors (stress = 0.034)
(C) RDA on Hellinger distance and constrained on Temperature, Salinity, and Nitrite and Nitrate

F. Core Microbiome

At the order level of the core microbiome (Figure 14A), 33 different taxa were shared between all summer stages. The early summer stage had 6 unique taxa at the order level and the late summer stage had 11 unique taxa. The mid-summer stage did not contain any unique taxa. This pattern was consistent with the family and genus level, where only the early summer and late summer had taxa which were unique to those summer stages only. At the family level (Figure 14B), there was a share of 45 taxa between all summer stages, and 62 taxa at the genus level (Figure 14C). The core microbiome of the sample seasons shares the same core of 62 genus taxa as the summer stages. Each sample season has unique taxa at the genus level that are not present in the other sample seasons (14D). Season 2012-2013 has the most unique taxa, with 13 total unique taxa at the genus level.

G. Community Composition Dynamics

The relative abundance values of the top 10 most abundant core taxa at the order level was plotted for each sample in Figure 15. The plot shows how the relative abundances of different order taxa change over the season and annually. Flavobacteriales began the summer season with a relatively high abundance but tended to decrease over the season, while Rhodobacterales showed an increase in the relative abundance over the season for all sampled summers. Several taxa like Nitrosococcales and Sphingobacteriales were detected with a high relative abundance in the 2013-2014 summer season only. Oceanospirillales, Alteromondales, and SAR11 displayed high variability in the dynamics of their relative abundances over the different years. Alteromondales showed a spike in relative abundance, up to 5%, at separate times in the 2013-2014 and 2014-2015 summer seasons. The relative abundance of Oceanospirillales was noticeably higher in the two sample points from the 2012-2013 season than any samples in the 2013-2014 and 2014-2015 season. For the 2013-2014 and 2014-2015 summer seasons, the Gammaproteobacteria orders of Burkholderiales and Cellvibrionales showed a relatively low relative abundance in early summer before increasing in the mid and late summer. Relative abundance values of the top 10 most abundant core taxa at the family level is shown in Figure 16. Results of the family level largely reflected the order level, except the family *Cryomorphaceae* of the order Flavobacteriales replaces the order of Alteromondales in the top 10 taxa. The *Cryomorphaceae* appear to increase over the summer and exhibits a peak just before the *Rhodobacteraceae* peak in the 2014-2015 season.

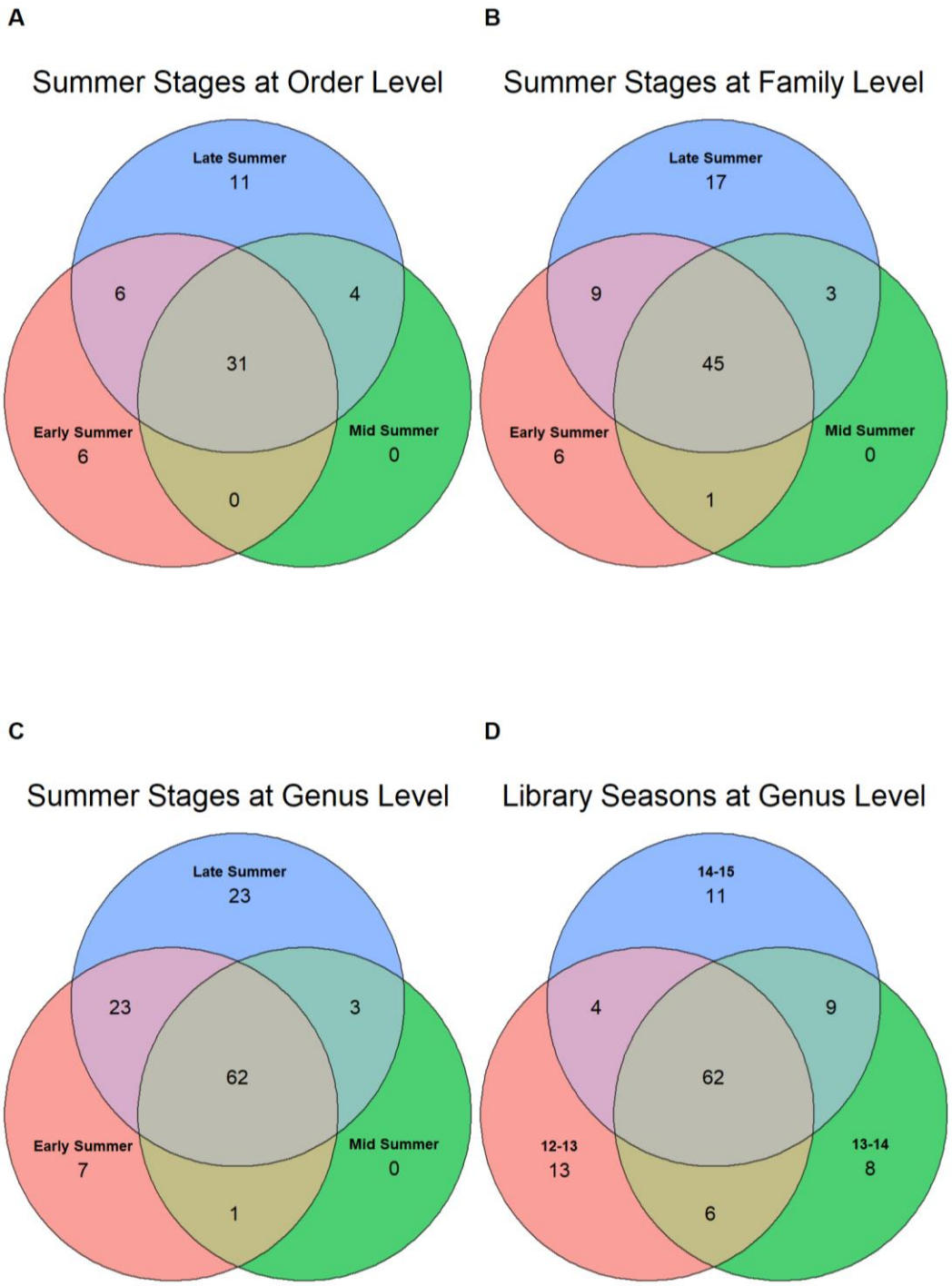


Figure 14. Core microbiome of 16S samples
(A) Core taxa in the summer stages at Order level
(B) Core taxa in the summer stages at Family level
(C) Core taxa in the summer stages at Genus level
(D) Core taxa in the library seasons at Order level

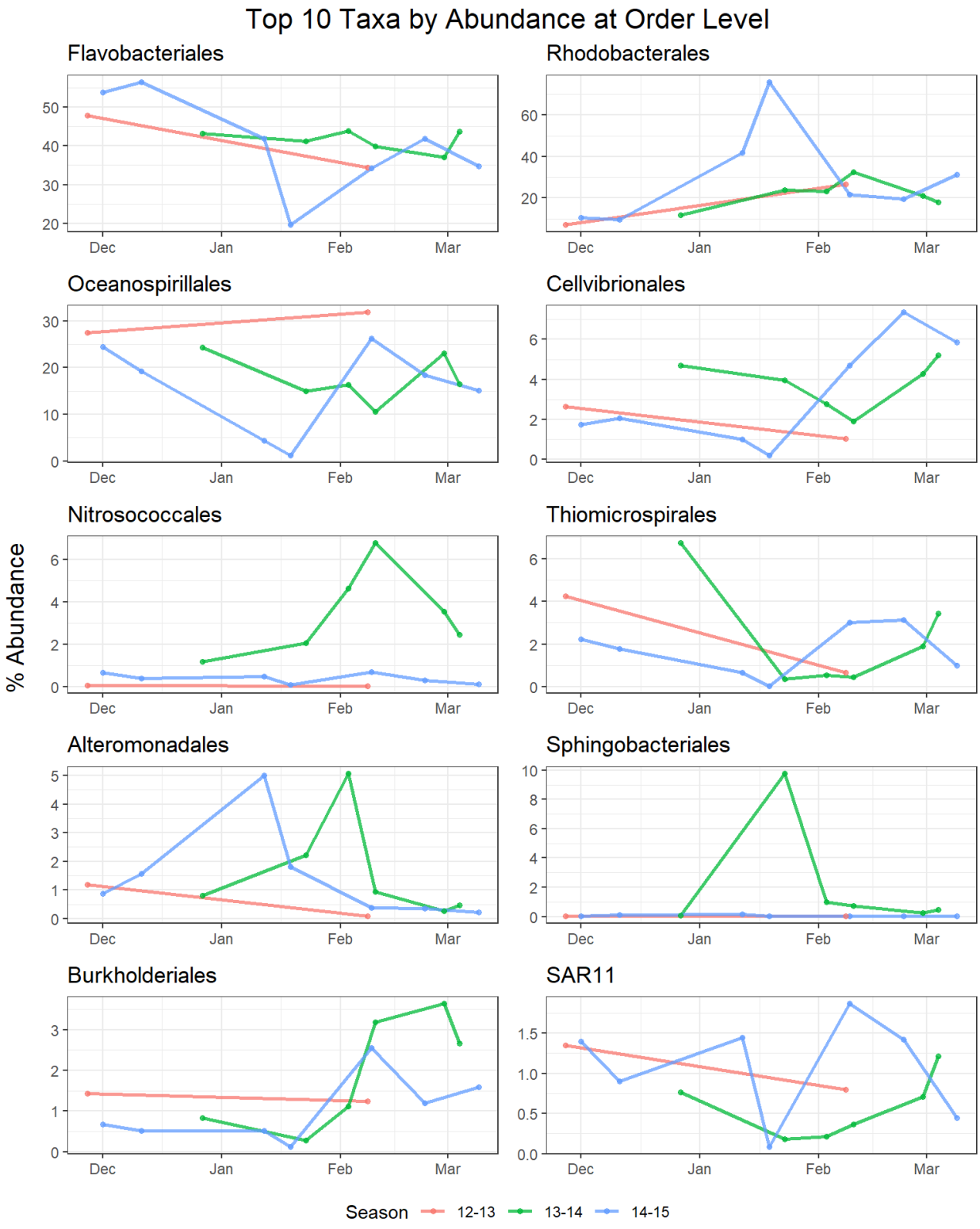


Figure 15. Relative abundance at the Order level over the summer season of the top 10 most abundant taxa in the core microbiome

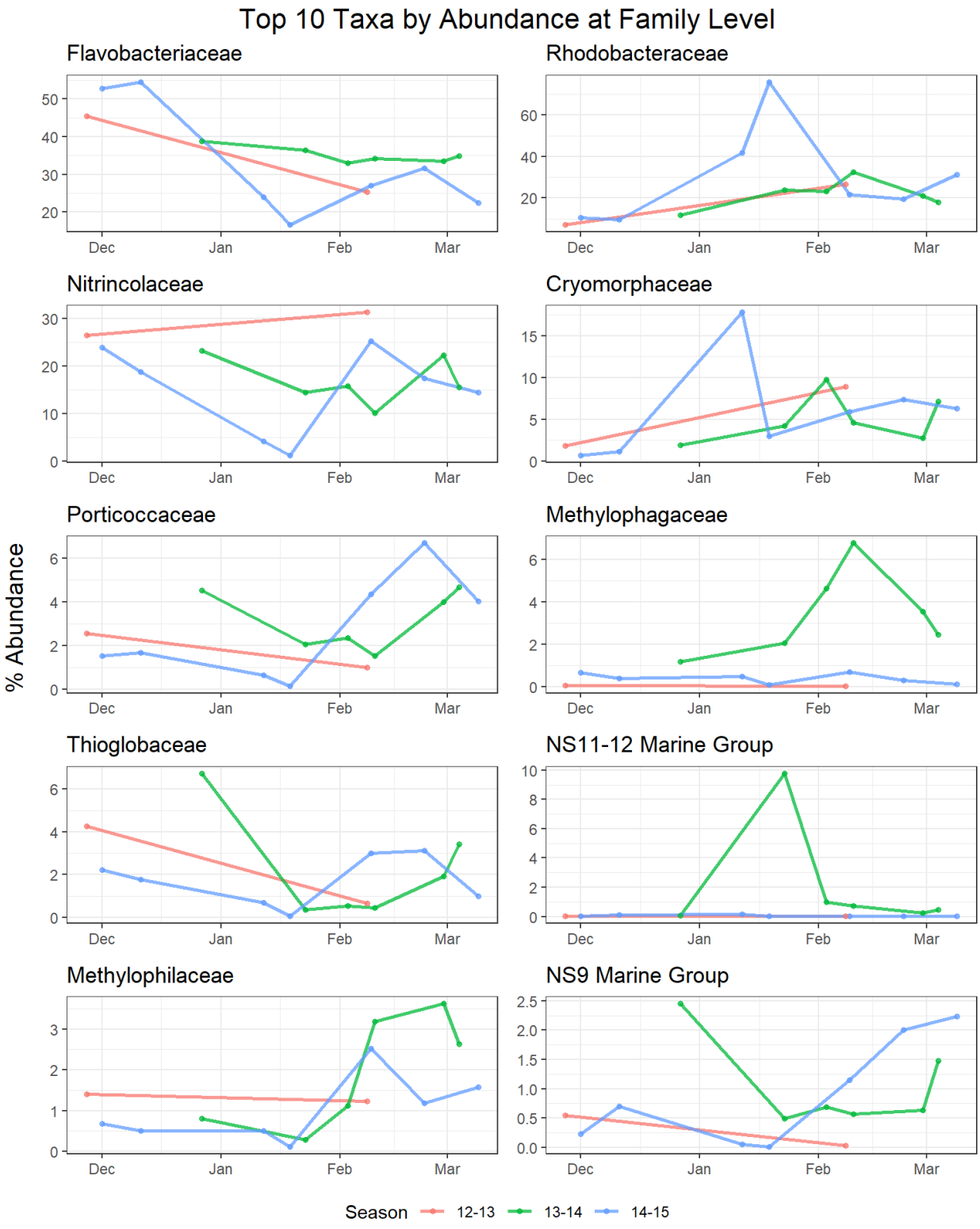


Figure 16. Relative abundance at the Family level over the summer season of the top 10 most abundant taxa in the core microbiome

IV. DISCUSSION

A. Summer Bacterial Community Composition

The results from this study demonstrated how surface waters of the WAP, during an austral summer, were dominated by the classes of Bacteroidia, Gammaproteobacteria, and Alphaproteobacteria (Figure 10). Bacteroidia 16S rRNA gene amplicons represented 42.03% of the final reads used in this study. Gammaproteobacteria accounted for 29.77% and Alphaproteobacteria accounted for 27.16% of the reads. Bacteroidia were primarily represented by the order of Flavobacteriales (96.05% of all Bacteroidia) and the genus *Polaribacter* (66.15 % of all Bacteroidia). Alphaproteobacteria reads largely consisted of the order Rhodobacterales, which accounted for 92.75% of all reads for this class. The Rhodobacterales class was split mainly between the genus of *Sulfitobacter* (44.79%), *Yoonia-Loktanella* (34.44%), and *Planktomarina* (12.70%). The order of SAR 11 only accounted for 3% of the total Alphaproteobacteria reads. Gammaproteobacteria composition was more diverse than Bacteroidia and Alphaproteobacteria. This class was represented by the orders of Oceanospirillales (60.89%), Cellvibrionales (11.04 %), Nitrosococcales (7.05%), Thiomicrospirales (6.32%), Burkholderiales (5.45%), and Alteromonadales (4.85%). Of the reads in Oceanospirillales, 95.76% belonged to the family *Nitrincolaceae* but were unclassified down at the genus level.

These three classes have also been reported to dominate the surface waters of the WAP in other studies [29]–[38]. Visualization of the relative abundances of specific taxa throughout each austral summer displayed the within-season variation and annual variation of the taxa (Figure 11). Particularly, the orders of Flavobacteriales and Rhodobacterales showed a temporal pattern that repeated in each sampling year (Figure 15). The relative abundance of Flavobacteriales generally decreased as the austral summer progressed, while the Rhodobacterales' relative abundance increased. This pattern was also observed at the family level with *Flavobacteriaceae* and *Rhodobacteraceae*.

A core microbiome of 31 unique taxa at the order level and 62 unique taxa at the genus level were found to be present in all 15 different 16S rRNA gene amplicon samples (Figure 14). Of the 31 taxa at the order level, the top 10 taxa by relative abundance accounted for 96.9% of all core taxa reads. The mid-summer samples did not have any taxa which were unique to this stage only, unlike the early and late summer stages. Each sample summer season had taxa which were unique to their respective seasons only. These results indicate that some rare taxa may be below detectable levels for certain summer stages and years and highlights the temporal variability of the marine bacterial community.

SAR11 relative abundance was noticeably low at 3% of the total reads when compared to the expected relative abundance of 9-42% reported in several other studies of the WAP [29], [32], [35]–[37]. Relative abundance of SAR11 has been reported to highest in the winter season and lowest in the summer season in the Southern Ocean in some studies [29], [35]. The low relative abundance of SAR11 in this study is likely due to primer bias against this taxa in the PCR amplification step. The original 806R primer is reported to have a nucleotide mismatch to

the 16S rRNA V4 region of the SAR11 clade [23]. This reduces the number of SAR11 amplicons from being created and leads to the underrepresentation observed in the data. Other studies which used utilized a different reverse primer for the V4 region or amplified other hypervariable regions of the 16S gene detected the SAR11 clade at a higher relative abundance.

Archaea was detected with a relative abundance of only 0.16% at the genus level. The original 515F and 806R primers have also been shown to be biased against the archaea Thaumarchaea [22]. Other studies have shown a similar result for the austral summer surface water and found archaea to have a higher relative abundance during the winter season and in deeper waters [29], [32]. Archaea metabolism in the WAP has been described as chemolithoautotrophic, which likely cannot compete in the nutrient conditions present during phytoplankton blooms. The archaea may not be able to fully utilize the high levels of available DOM during the austral summer.

Sample S3L06 from water sample date 1/19/2015 stood as an outlier among the other 14 16S rRNA samples (Figure 11). On this date, the order Rhodobacterales relative abundance reached a high of 76%, while Flavobacteriales and Oceanospirillales had a low of 19.64% and 1.22% respectively. Chlorophyll *a* concentration reached a season high of 6.68 mg/m³ on this date, while water temperature was recorded to be 1.02 °C, near the season high of 1.10°C (Figure 3A and Figure 5A). The environmental conditions of the peak phytoplankton bloom in the summer and warm water temperature may have helped to induce the high relative abundance observed for Rhodobacterales. For the 2014-2015 summer season, alpha diversity measures were at the lowest on 1/19/2015 (Figure 8). Sample S2L06 from water sample date 1/23/2014 had the lowest alpha diversity measures of the 2013-2014 season (Figure 8). This was also during a peak phytoplankton bloom, where the chlorophyll *a* concentration reached 5.62 6.68 mg/m³ (Figure 3A). This sample was not dominated by Rhodobacterales but did show a spike in the order of Sphingobacteriales of the class Bacteroidota instead at a relative abundance of 9.78% (Figure 11B).

Analyses with alpha and beta diversity measures showed significant differences between the three summer stages (Figure 9). The Kruskal-Wallis test showed strong evidence for difference in median richness among the three summer stages. The pairwise comparisons of the summer stages with the Wilcoxon Rank Sum Test on the Shannon diversity values showed a significant difference between early summer against late summer and mid-summer against late summer. The results showed alpha diversity metrics of bacterial community composition to increase by the late austral summer season. Other studies have shown alpha diversity metrics to decrease during phytoplankton blooms and to increase by the end of the summer [29], [30], [32], [35]. Samples within these stages also tend to cluster together in PCA, NMDS, and RDA ordination plots (Figure 13). Early summer samples are grouped together in the PCA plot, but the mid-summer and late summer samples were more intermixed. The temperature, salinity, and the dissolved inorganic nutrients of nitrite and nitrate environmental measures may be useful for characterizing the different physiochemical contexts of the summer stages.

B. Limitations of this Study

16S rRNA copy numbers were not accounted for in this study. Gammaproteobacteria can have an average of 5 or more copies of the 16S rRNA gene and shows a large variation within this class [67]. Alphaproteobacteria show lower variability within the class and are reported to have an average of 2 16S rRNA gene copies in the genome. Variation in the copy number occurs at different taxonomic levels and can influence the results in the relative abundance results. Gammaproteobacteria abundance is likely overestimated due to this variability. The dynamics of SAR11 relative abundance was not fully captured by this study. The missing portion of SAR11 abundances likely leads to an overestimation of relative abundance of all other taxa due to the compositional nature of relative abundance data. The water samples for this study only were collected only at 10m in depth, but prokaryotic composition has been shown to vary with water column depth [32], [33]. Additional sampling in a range of depths can be used to explore how community composition changes with depth.

C. Future Research

Further research in the characterization and dynamics of bacteria and other microbes present in the marine ecosystem of the WAP is needed to help predict the effects of climate change on this environment. Phylogenetic composition data of phytoplankton blooms could be integrated with the results presented in this study to elucidate how changes in phytoplankton diversity may alter bacterial diversity. 16S rRNA gene copy number databases could be utilized to correct for the variation present in bacterial rRNA gene copy numbers. The current water samples or any future water samples can be reamplified and sequenced with new universal V4 primers to generate a more accurate snapshot of prokaryotic diversity. The data from this study can also be reanalyzed by categorizing the samples into different stages of phytoplankton blooms to investigate how bacterial community composition changes through a phytoplankton bloom.

V. CONCLUSION

Marine bacterial community composition dynamics in the WAP have previously been investigated, but most studies have only sampled a single season or year. This study analyzed a molecular survey consisting of 15 different 16S rRNA V4 gene amplicon samples that covered 3 different austral summer seasons in the surface waters of the WAP. The aim the study was to contribute to the baseline profiling of marine bacterial community composition and dynamics in the WAP. The results from this study showed how 16S rRNA sequence data of the microbes in environmental water samples can be used to identify a core group of bacterial taxa and detail their temporal dynamics. The classes of Alphaproteobacteria, Gammaproteobacteria, and Bacteroidia form the bulk of the bacterial community abundances. Flavobacteriales are the primary taxa in the early summer but are replaced by different members of the Alphaproteobacteria and Gammaproteobacteria as the summer progresses. Alpha diversity analysis of the bacterial composition helped revealed a pattern of how bacterial richness and

evenness measures are lowest in the early and mid-summer but rises towards the end of the summer.

REFERENCES

[1] M. R. Stukel *et al.*, "The imbalance of new and export production in the western Antarctic Peninsula, a potentially 'leaky' ecosystem," *Glob. Biogeochem. Cycles*, vol. 29, no. 9, pp. 1400–1420, Sep. 2015, doi: 10.1002/2015GB005211.

[2] R. C. Smith *et al.*, "The Palmer LTER: A Long-Term Ecological Research Program at Palmer Station, Antarctica," *Oceanography*, vol. 8, no. 3, pp. 77–86, 1995.

[3] H. W. Ducklow *et al.*, "Marine pelagic ecosystems: the West Antarctic Peninsula," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 362, no. 1477, pp. 67–94, Jan. 2007, doi: 10.1098/rstb.2006.1955.

[4] H. J. Venables, A. Clarke, and M. P. Meredith, "Wintertime controls on summer stratification and productivity at the western Antarctic Peninsula," *Limnol. Oceanogr.*, vol. 58, no. 3, pp. 1035–1047, May 2013, doi: 10.4319/lo.2013.58.3.1035.

[5] H. U. Sverdrup, "On Conditions for the Vernal Blooming of Phytoplankton," *ICES J. Mar. Sci.*, vol. 18, no. 3, pp. 287–295, Jan. 1953, doi: 10.1093/icesjms/18.3.287.

[6] G. K. Saba *et al.*, "Winter and spring controls on the summer food web of the coastal West Antarctic Peninsula," *Nat. Commun.*, vol. 5, no. 1, p. 4318, Sep. 2014, doi: 10.1038/ncomms5318.

[7] L. Legendre and F. Rassoulzadegan, "Plankton and nutrient dynamics in marine waters," *Ophelia*, vol. 41, no. 1, pp. 153–172, Feb. 1995, doi: 10.1080/00785236.1995.10422042.

[8] H. W. Ducklow, O. Schofield, M. Vernet, S. Stammerjohn, and M. Erickson, "Multiscale control of bacterial production by phytoplankton dynamics and sea ice along the western Antarctic Peninsula: A regional and decadal investigation," *J. Mar. Syst.*, vol. 98–99, pp. 26–39, Sep. 2012, doi: 10.1016/j.jmarsys.2012.03.003.

[9] B. C. Cho and F. Azam, "Major role of bacteria in biogeochemical fluxes in the ocean's interior," *Nature*, vol. 332, no. 6163, pp. 441–443, Mar. 1988, doi: 10.1038/332441a0.

[10] A. Prieto, E. Barber-Lluch, M. Hernández-Ruiz, S. MartÍnez-GarcÍa, E. Fernández, and E. Teira, "Assessing the role of phytoplankton–bacterioplankton coupling in the response of microbial plankton to nutrient additions," *J. Plankton Res.*, vol. 38, no. 1, pp. 55–63, Jan. 2016, doi: 10.1093/plankt/fbv101.

[11] C. L. Sabine, "The Oceanic Sink for Anthropogenic CO2," *Science*, vol. 305, no. 5682, pp. 367–371, Jul. 2004, doi: 10.1126/science.1097403.

[12] T. L. Frölicher, J. L. Sarmiento, D. J. Paynter, J. P. Dunne, J. P. Krasting, and M. Winton, "Dominance of the Southern Ocean in Anthropogenic Carbon and Heat Uptake in CMIP5 Models," *J. Clim.*, vol. 28, no. 2, pp. 862–886, Jan. 2015, doi: 10.1175/JCLI-D-14-00117.1.

[13] S. Khatiwala, F. Primeau, and T. Hall, "Reconstruction of the history of anthropogenic CO2 concentrations in the ocean," *Nature*, vol. 462, no. 7271, pp. 346–349, Nov. 2009, doi: 10.1038/nature08526.

[14] J. A. Fuhrman and F. Azam, "Thymidine incorporation as a measure of heterotrophic bacterioplankton production in marine surface waters: Evaluation and field results," *Mar. Biol.*, vol. 66, no. 2, pp. 109–120, 1982, doi: 10.1007/BF00397184.

[15] D. G. Vaughan *et al.*, "Recent Rapid Regional Climate Warming on the Antarctic Peninsula," *Clim. Change*, vol. 60, no. 3, pp. 243–274, 2003, doi: 10.1023/A:1026021217991.

[16] M. Montes-Hugo *et al.*, "Recent Changes in Phytoplankton Communities Associated with Rapid Regional Climate Change Along the Western Antarctic Peninsula," *Science*, vol. 323, no. 5920, pp. 1470–1473, Mar. 2009, doi: 10.1126/science.1164533.

[17] J. A. Cram *et al.*, "Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years," *ISME J.*, vol. 9, no. 3, pp. 563–580, Mar. 2015, doi: 10.1038/ismej.2014.153.

[18] M. T. Suzuki *et al.*, "Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample.," *Appl. Environ. Microbiol.*, vol. 63, no. 3, pp. 983–989, 1997, doi: 10.1128/AEM.63.3.983-989.1997.

[19] J.-C. Auguet, A. Barberan, and E. O. Casamayor, "Global ecological patterns in uncultured Archaea," *ISME J.*, vol. 4, no. 2, pp. 182–190, Feb. 2010, doi: 10.1038/ismej.2009.109.

[20] J. G. Caporaso *et al.*, "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample," *Proc. Natl. Acad. Sci.*, vol. 108, no. Supplement_1, pp. 4516–4522, Mar. 2011, doi: 10.1073/pnas.1000080107.

[21] D. J. Lane, B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace, "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses," *Proc. Natl. Acad. Sci.*, vol. 82, no. 20, pp. 6955–6959, Oct. 1985, doi: 10.1073/pnas.82.20.6955.

[22] A. E. Parada, D. M. Needham, and J. A. Fuhrman, "Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples: Primers for marine microbiome studies," *Environ. Microbiol.*, vol. 18, no. 5, pp. 1403–1414, May 2016, doi: 10.1111/1462-2920.13023.

[23] A. Apprill, S. McNally, R. Parsons, and L. Weber, "Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton," *Aquat. Microb. Ecol.*, vol. 75, no. 2, pp. 129–137, Jun. 2015, doi: 10.3354/ame01753.

[24] E. Bolyen *et al.*, "Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2," *Nat. Biotechnol.*, vol. 37, no. 8, pp. 852–857, Aug. 2019, doi: 10.1038/s41587-019-0209-9.

[25] M. Martin, "Cutadapt removes adapter sequences from high-throughput sequencing reads," *EMBnet.journal*, vol. 17, no. 1, p. 10, May 2011, doi: 10.14806/ej.17.1.200.

[26] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes, "DADA2: High-resolution sample inference from Illumina amplicon data," *Nat. Methods*, vol. 13, no. 7, pp. 581–583, Jul. 2016, doi: 10.1038/nmeth.3869.

[27] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, "Exact sequence variants should replace operational taxonomic units in marker-gene data analysis," *ISME J.*, vol. 11, no. 12, pp. 2639–2643, Dec. 2017, doi: 10.1038/ismej.2017.119.

[28] J. A. Gilbert *et al.*, "Defining seasonal marine microbial community dynamics," *ISME J.*, vol. 6, no. 2, pp. 298–308, Feb. 2012, doi: 10.1038/ismej.2011.107.

[29] J. J. Grzymski *et al.*, "A metagenomic assessment of winter and summer bacterioplankton from Antarctica Peninsula coastal surface waters," *ISME J.*, vol. 6, no. 10, pp. 1901–1915, Oct. 2012, doi: 10.1038/ismej.2012.31.

[30] J. F. Ghiglione and A. E. Murray, "Pronounced summer to winter differences and higher wintertime richness in coastal Antarctic marine bacterioplankton: Temporal variation in Southern Ocean coastal bacterioplankton," *Environ. Microbiol.*, vol. 14, no. 3, pp. 617–629, Mar. 2012, doi: 10.1111/j.1462-2920.2011.02601.x.

[31] R. E. Jamieson, A. D. Rogers, D. S. M. Billett, D. A. Smale, and D. A. Pearce, "Patterns of marine bacterioplankton biodiversity in the surface waters of the Scotia Arc, Southern Ocean," *FEMS Microbiol. Ecol.*, vol. 80, no. 2, pp. 452–468, Apr. 2012, doi: 10.1111/j.1574-6941.2012.01313.x.

[32] C. Luria, H. Ducklow, and L. Amaral-Zettler, "Marine bacterial, archaeal and eukaryotic diversity and community structure on the continental shelf of the western Antarctic Peninsula," *Aquat. Microb. Ecol.*, vol. 73, no. 2, pp. 107–121, Oct. 2014, doi: 10.3354/ame01703.

[33] C. N. Signori, F. Thomas, A. Enrich-Prast, R. C. G. Pollery, and S. M. Sievert, "Microbial diversity and community structure across environmental gradients in Bransfield Strait, Western Antarctic Peninsula," *Front. Microbiol.*, vol. 5, Dec. 2014, doi: 10.3389/fmicb.2014.00647.

[34] Y.-X. Zeng, Y. Yu, Z.-Y. Qiao, H.-Y. Jin, and H.-R. Li, "Diversity of bacterioplankton in coastal seawaters of Fildes Peninsula, King George Island, Antarctica," *Arch. Microbiol.*, vol. 196, no. 2, pp. 137–147, Feb. 2014, doi: 10.1007/s00203-013-0950-2.

[35] C. M. Luria, L. A. Amaral-Zettler, H. W. Ducklow, and J. J. Rich, "Seasonal Succession of Free-Living Bacterial Communities in Coastal Waters of the Western Antarctic Peninsula," *Front. Microbiol.*, vol. 7, Nov. 2016, doi: 10.3389/fmicb.2016.01731.

[36] C. N. Signori, V. H. Pellizari, A. Enrich-Prast, and S. M. Sievert, "Spatiotemporal dynamics of marine bacterial and archaeal communities in surface waters off the northern

Antarctic Peninsula," *Deep Sea Res. Part II Top. Stud. Oceanogr.*, vol. 149, pp. 150–160, Mar. 2018, doi: 10.1016/j.dsr2.2017.12.017.

[37] S. Cao, J. He, F. Zhang, L. Lin, Y. Gao, and Q. Zhou, "Diversity and community structure of bacterioplankton in surface waters off the northern tip of the Antarctic Peninsula," *Polar Res.*, vol. 38, no. 0, Mar. 2019, doi: 10.33265/polar.v38.3491.

[38] S. Fuentes *et al.*, "Summer phyto- and bacterioplankton communities during low and high productivity scenarios in the Western Antarctic Peninsula," *Polar Biol.*, vol. 42, no. 1, pp. 159–169, Jan. 2019, doi: 10.1007/s00300-018-2411-5.

[39] J. M. Gonzalez *et al.*, "Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria)," *Proc. Natl. Acad. Sci.*, vol. 105, no. 25, pp. 8724–8729, Jun. 2008, doi: 10.1073/pnas.0712027105.

[40] R. M. Morris *et al.*, "SAR11 clade dominates ocean surface bacterioplankton communities," *Nature*, vol. 420, no. 6917, Dec. 2002, doi: 10.1038/nature01240.

[41] H.-A. Giebel, T. Brinkhoff, W. Zwisler, N. Selje, and M. Simon, "Distribution of *Roseobacter* RCA and SAR11 lineages and distinct bacterial communities from the subtropics to the Southern Ocean," *Environ. Microbiol.*, vol. 11, no. 8, pp. 2164–2178, Aug. 2009, doi: 10.1111/j.1462-2920.2009.01942.x.

[42] D. Kelley and C. Richards, "oce: Analysis of Oceanographic Data." [Online]. Available: https://dankelley.github.io/oce

[43] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahé, "VSEARCH: a versatile open source tool for metagenomics," *PeerJ*, vol. 4, p. e2584, Oct. 2016, doi: 10.7717/peerj.2584.

[44] C. Quast *et al.*, "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D590–D596, Nov. 2012, doi: 10.1093/nar/gks1219.

[45] M. S. Robeson *et al.*, "RESCRIPt: Reproducible sequence taxonomy reference database management for the masses," Bioinformatics, preprint, Oct. 2020. doi: 10.1101/2020.10.05.326504.

[46] K. Katoh and D. M. Standley, "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability," *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.

[47] B. Q. Minh *et al.*, "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Mol. Biol. Evol.*, vol. 37, no. 5, pp. 1530–1534, May 2020, doi: 10.1093/molbev/msaa015.

[48] J. Bisanz, "qiime2R." [Online]. Available: https://github.com/jbisanz/qiime2R

[49] P. J. McMurdie and S. Holmes, "phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data," *PLoS ONE*, vol. 8, no. 4, p. e61217, Apr. 2013, doi: 10.1371/journal.pone.0061217.

[50] L. Lahti and S. Shetty, "Tools for microbiome analysis in R. Microbiome package version 1.12.0." [Online]. Available: https://github.com/microbiome/microbiome

[51] S. Smith, "phylosmith." [Online]. Available: https://github.com/schuyler-smith/phylosmith/

[52] N. J. Gotelli and R. K. Colwell, "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness," *Ecol. Lett.*, vol. 4, no. 4, pp. 379–391, Jul. 2001, doi: 10.1046/j.1461-0248.2001.00230.x.

[53] P. J. McMurdie and S. Holmes, "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible," *PLoS Comput. Biol.*, vol. 10, no. 4, p. e1003531, Apr. 2014, doi: 10.1371/journal.pcbi.1003531.

[54] S. Weiss *et al.*, "Normalization and microbial differential abundance strategies depend upon data characteristics," *Microbiome*, vol. 5, no. 1, p. 27, Dec. 2017, doi: 10.1186/s40168-017-0237-y.

[55] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome Datasets Are Compositional: And This Is Not Optional," *Front. Microbiol.*, vol. 8, p. 2224, Nov. 2017, doi: 10.3389/fmicb.2017.02224.

[56] P. I. Costea, G. Zeller, S. Sunagawa, and P. Bork, "A fair comparison," *Nat. Methods*, vol. 11, no. 4, pp. 359–359, Apr. 2014, doi: 10.1038/nmeth.2897.

[57] J. N. Paulson, H. C. Bravo, and M. Pop, "Reply to: 'A fair comparison,'" *Nat. Methods*, vol. 11, no. 4, pp. 359–360, Apr. 2014, doi: 10.1038/nmeth.2898.

[58] A. Chao, "Nonparametric Estimation of the Number of Classes in a Population," *Scand. J. Stat.*, vol. 11, no. 4, pp. 265–270, 1984.

[59] B.-R. Kim *et al.*, "Deciphering Diversity Indices for a Better Understanding of Microbial Communities," *J. Microbiol. Biotechnol.*, vol. 27, no. 12, pp. 2089–2093, Dec. 2017, doi: 10.4014/jmb.1709.09027.

[60] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/j.2517-6161.1995.tb02031.x.

[61] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight, "Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities," *Appl. Environ. Microbiol.*, vol. 73, no. 5, pp. 1576–1585, Mar. 2007, doi: 10.1128/AEM.01996-06.

[62] J. Oksanen *et al.*, *vegan: Community Ecology Package*. 2020. [Online]. Available: https://CRAN.R-project.org/package=vegan

[63] D. Borcard, F. Gillet, and P. Legendre, *Numerical Ecology with R*. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-319-71404-2.

[64] K. Clarke and M. Ainsworth, "A method of linking multivariate community structure to environmental variables," *Mar. Ecol. Prog. Ser.*, vol. 92, pp. 205–219, 1993, doi: 10.3354/meps092205.

[65] P. Legendre and E. D. Gallagher, "Ecologically meaningful transformations for ordination of species data," *Oecologia*, vol. 129, no. 2, pp. 271–280, Oct. 2001, doi: 10.1007/s004420100716.

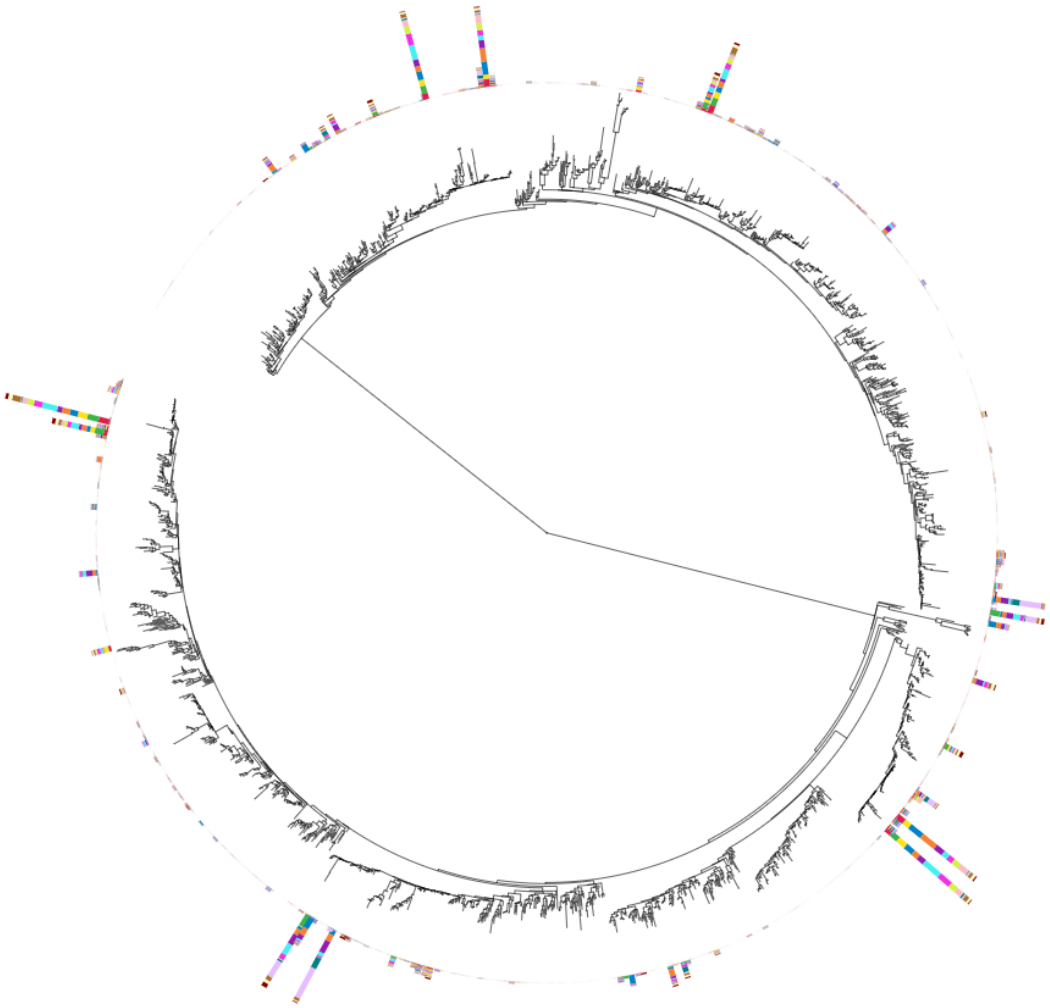
[66] O. Paliy and V. Shankar, "Application of multivariate statistical techniques in microbial ecology," *Mol. Ecol.*, vol. 25, no. 5, pp. 1032–1057, Mar. 2016, doi: 10.1111/mec.13536.

[67] T. Větrovský and P. Baldrian, "The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses," *PLoS ONE*, vol. 8, no. 2, p. e57923, Feb. 2013, doi: 10.1371/journal.pone.0057923.

[68] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation," *Nucleic Acids Res.*, p. gkab301, Apr. 2021, doi: 10.1093/nar/gkab301.

APPENDICES

Appendix A. Mid-point Rooted Phylogenetic Tree with Abundances



A *de novo* phylogenetic tree was created for the ASVs with the align-to-tree-mafft-iqtree pipeline in the phylogeny plugin for QIIME 2. The tree and ASV abundances were visualized with iTOL [68].

Appendix B. Taxa Names of Core Microbiome

Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Flavobacteriaceae;Polaribacter
Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Nitrincolaceae;uncultured
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Planktomarina
Bacteria;Proteobacteria;Gammaproteobacteria;Thiomicrospirales;Thioglobaceae;SUP05_cluster
Bacteria;Proteobacteria;Gammaproteobacteria;Cellvibrionales;Porticoccaceae;SAR92_clade
Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Flavobacteriaceae;NS5_marine_group
Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;NS9_marine_group;NS9_marine_group
Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Cryomorphaceae;uncultured
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Sulfitobacter
Bacteria;Proteobacteria;Gammaproteobacteria;Burkholderiales;Methylophilaceae;OM43_clade
Bacteria;Proteobacteria;Gammaproteobacteria;SAR86_clade;SAR86_clade;SAR86_clade
Bacteria;Proteobacteria;Gammaproteobacteria;Nitrosococcales;Methylophagaceae;uncultured
Bacteria;Bacteroidota;Bacteroidia;Chitinophagales;Saprospiraceae;uncultured
Bacteria;Proteobacteria;Alphaproteobacteria;Puniceispirillales;SAR116_clade;SAR116_clade
Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Pseudohongiellaceae;Pseudohongiella
Archaea;Crenarchaeota;Nitrososphaeria;Nitrosopumilales;Nitrosopumilaceae;Candidatus_Nitrosopumilus
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Ascidiaceihabitans
Bacteria;Proteobacteria;Alphaproteobacteria;SAR11_clade;Clade_I;Clade_Ia
Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Colwelliaceae;Colwellia
Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Flavobacteriaceae;NS4_marine_group
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Yoonia-Loktanella
Bacteria;Proteobacteria;Gammaproteobacteria;Granulosicoccales;Granulosicoccaceae;Granulosicoccus
Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Alteromonadaceae;Paraglaciecola
Bacteria;Bacteroidota;Bacteroidia;Cytophagales;Cyclobacteriaceae;Marinoscillum
Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Flavobacteriaceae;Ulvibacter
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;Amylibacter
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;Magnetospiraceae;Magnetospira
Bacteria;Bacteroidota;Bacteroidia;Chitinophagales;Saprospiraceae;Lewinella
Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Thiotrichaceae;Cocleimonas
Bacteria;Proteobacteria;Alphaproteobacteria;Rhodobacterales;Rhodobacteraceae;uncultured
Bacteria;Proteobacteria;Gammaproteobacteria;OM182_clade;OM182_clade;OM182_clade
Bacteria;Proteobacteria;Alphaproteobacteria;SAR11_clade;Clade_IV;Clade_IV
Bacteria;Bacteroidota;Bacteroidia;Chitinophagales;Saprospiraceae;Portibacter
Bacteria;SAR324_clade;SAR324_clade;SAR324_clade;SAR324_clade;SAR324_clade
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Hellea
Bacteria;Actinobacteriota;Acidimicrobiia;Microtrichales;Microtrichaceae;Sva0996_marine_group
Bacteria;Proteobacteria;Gammaproteobacteria;Arenicellales;Arenicellaceae;Arenicella
Bacteria;Proteobacteria;Alphaproteobacteria;SAR11_clade;Clade_II;Clade_II
Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Thiotrichaceae;Leucothrix

Bacteria;Proteobacteria;Alphaproteobacteria;Rhodospirillales;AEGEAN-169_marine_group;AEGEAN-169_marine_group
Bacteria;Proteobacteria;Alphaproteobacteria;Parvibaculales;OCS116_clade;OCS116_clade
Bacteria;Proteobacteria;Alphaproteobacteria;Defluviicoccales;uncultured;uncultured
Bacteria;Proteobacteria;Gammaproteobacteria;Cellvibrionales;Halieaceae;OM60(NOR5)_clade
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Litorimonas
Bacteria;Campilobacterota;Campylobacteria;Campylobacterales;Arcobacteraceae;uncultured
Bacteria;Proteobacteria;Alphaproteobacteria;Thalassobaculales;Nisaeaceae;OM75_clade
Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Cryomorphaceae;Vicingus
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Robiginitomaculum
Bacteria;Proteobacteria;Gammaproteobacteria;Thiotrichales;Thiotrichaceae;uncultured
Bacteria;Proteobacteria;Gammaproteobacteria;Oceanospirillales;Kangiellaceae;uncultured
Bacteria;Bacteroidota;Bacteroidia;Sphingobacteriales;NS11-12_marine_group;NS11-12_marine_group
Bacteria;Verrucomicrobiota;Verrucomicrobiae;Verrucomicrobiales;Rubritaleaceae;Rubritalea
Bacteria;Proteobacteria;Gammaproteobacteria;Alteromonadales;Psychromonadaceae;Psychromonas
Bacteria;Fusobacteriota;Fusobacteriia;Fusobacteriales;Fusobacteriaceae;Psychrilyobacter
Bacteria;Actinobacteriota;Actinobacteria;PeM15;PeM15;PeM15
Bacteria;Actinobacteriota;Acidimicrobiia;Microtrichales;Microtrichaceae;uncultured
Bacteria;Bacteroidota;Bacteroidia;Flavobacteriales;Crocinitomicaceae;Crocinitomix
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Hyphomonadaceae;Fretibacter
Bacteria;Proteobacteria;Alphaproteobacteria;SAR11_clade;Clade_III;Clade_III
Bacteria;Proteobacteria;Gammaproteobacteria;Cellvibrionales;Spongiibacteraceae;BD1-7_clade
Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiaceae;Clostridium_sensu_stricto_1
Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas