Bioinformatics approaches for assessing microbial communities in the surface ocean

Kara Martin

Supervisors:

Prof.Vincent Moulton

Prof.Thomas Mock

Dr.Richard Leggett

A thesis submitted for the degree of Doctor of Philosophy

University of East Anglia,

School of Computing Sciences,

Norwich, United Kingdom.

September, 2018

Abstract

Microbes are vital for life on Earth. Within the oceans, they are the major primary producers of oxygen and contribute greatly to the other biogeochemical cycling of the elements which in turn influence the global climate. These microbes can be found inhabiting the oceans throughout the world and they cover over ∼70% of the surface of the Earth.

Microbes have evolved in different environments in the oceans and in different ways. To gain an understanding of the microbial communities in the surface oceans in the Arctic and Atlantic oceans environmental scientists based at the University of East Anglia, the University of Groningen and Royal Netherlands Institute for Sea Research collected ocean samples from 68 stations along a transect of the Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean. In addition, they recorded environmental data at the time of sampling, such as temperature and salinity. Genomic DNA from filtered samples was sequenced using high-throughput sequencing.

This thesis contains a comprehensive analysis of this sequencing data with the aim of understanding the composition and distribution of microbial communities in the surface of the ocean. To this end, we designed bioinformatic pipelines in order to analyse metatranscriptome, 18S and 16S rDNA datasets from the set of stations. In addition, we developed a novel methodology for normalising 18S and 16S rDNA copy numbers. This enabled us to perform additional analyses such as biodiversity, co-occurrence and breakpoint analyses. The breakpoint analysis is the first of this type performed for microbes in the ocean across a temperature gradient.

In our results, we observed a greater diversity of 18S and 16S rDNA taxa in the tropical regions of the South Atlantic Ocean, versus the polar regions of the Arctic Ocean. Moreover, in the co-occurrence analysis of the 18S and 16S rDNA datasets, we found two community networks, one positively correlated to temperature and the other

negatively. We also performed a breakpoint analysis on our metatranscriptome, 18S and 16S rDNA datasets and found a shift in diversity occurring in the North Atlantic Ocean. In particular, the shift occurs in the temperate region of the North Atlantic Ocean, between the polar Arctic Ocean and tropical South Atlantic Ocean.

These results are important because the co-occurrence analysis enables us to hypothesise that different microbial communities have different preferences for temperature. Moreover, as global warming is predicted to raise the temperatures in the ocean, our results could potentially enable forecasts of how climate change will affect these microbial communities using climate models underpinned by genetic information.

Acknowledgements

I would like to thank my supervisors Prof.Vincent Moulton, Prof.Thomas Mock and Dr.Richard Leggett for their continuous support and guidance. I would also like to thank Dr.Andrew Toseland for all of his support and advice. I would like to thank Dr.Katrin Schmidt, Dr.Willem van de Poll and Dr.Klaas Timmermans for collecting the samples during the three expeditions, and I would like to thank the Joint Genome Institute for sequencing the samples. Finally, I would like to thank the University of East Anglia and the Earlham Institute for funding my PhD.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Marine microbes

The surface of the planet is made up of about 70% water, therefore providing the largest habitat for life on the surface of the planet, in particular, microbes [Das et al., 2006]. Among these marine microbes inhabiting this vast marine ecosystem are prokaryote species such as alphaproteobacteria and eukaryote species such as diatoms [Aryal et al., 2015], [Brierley, 2017]. Phytoplankton is a class of unicellular photosynthetic organisms, that is composed of a wide range of organisms of both prokaryotes and eukaryotes, and include thousands of different species [Collins et al., 2014], [Brierley, 2017]. Diatoms are eukaryotic phytoplankton and are the most diverse of the eukaryotic phytoplankton as they have about 200,000 different species [Armbrust, 2009]. (Figure 1.1).

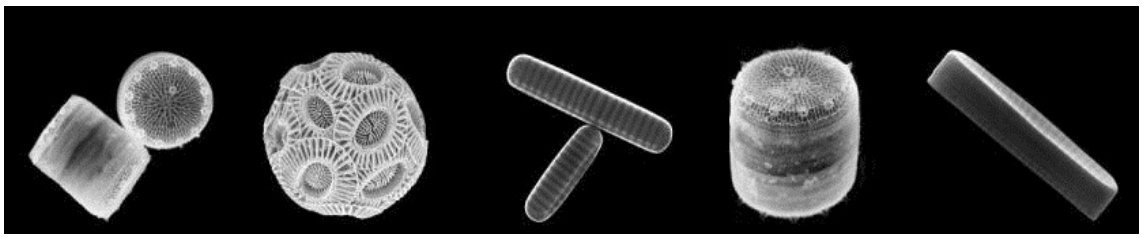

Figure 1.1: Diatoms species, starting from the left is *Thalassiosira pseudonana*, *Emiliania huxleyi*, *Fragilariopsis cylindrus*, *Thalassiosira pseudonana* and *Fragilariopsis cylindrus*. (Photo from MOCK RESEARCH LAB, http://mocklab.com/)

Communities can be defined as a large number of species that interact with each other in a countless number of ways [Godfray and May, 2014]. Communities of marine prokaryote and eukaryote species have co-existed throughout their evolution in

common habitats [Cordero and Datta, 2016], [Brussaard et al., 2016]. In the environment, two principal forces can be seen to act on these microbes - biotic, which refers to the interactions between the microbes, and abiotic, which refers to environmental influences such as temperature and salinity [Bijlsma and Loeschcke, 2005]. The interaction within marine microbial communities occur under these biotic and abiotic influences and therefore shape their evolution and adaptation [Brussaard et al., 2016]. The microbes have a range of interactions such as endosymbiosis (which refers to the merging of distinct cells so one is inside the other), mutualism (which refers to the interaction between different species that results in a mutually beneficial outcome, be it for reproduction and/or survival), parasitism (which refers to a species that obtains sustenance from its host) and commensalism (which refers to when one species gains from its association with another species who is not affected in a positive or negative manner) [Baron, 1996], [Holland and Bronstein, 2008], [Boon et al., 2014], [Haque and Haque, 2017].

A notable example, as outlined next, is how microbial interactions resulted in phytoplankton acquiring the ability to convert light into energy in order to thrive and the subsequent diversification of phytoplankton species [Wernegreen, 2012]. Phytoplankton obtained their photosynthesis capabilities through the process of endosymbiosis from a class of bacteria called cyanobacteria [Simon et al., 2009]. About 1.5 billion years ago an endosymbiosis event is hypothesized to have produced the first photosynthetic eukaryote, a process by which a heterotrophic eukaryote engulfed a cyanobacterium [Shemi et al., 2015]. An intracellular gene transfer occurred between the primitive host of heterotrophic eukaryote and its symbiont cyanobacterium, after which the heterotrophic eukaryote retained the cyanobacterium by converting it into a plastid [Simon et al., 2009]. This primary endosymbiosis event resulted in 3 distinct clades of unicellular algae. One of these clades is the viridiplantae, also known as the green plastid lineage, which is thought to be the source from which all land plants and green algae evolved. Another is the rhodophyta, also known as the red plastid lineage. This clade includes an ancient group of marine red microalgae and seaweed. The other clade is the glaucophyta, which is made up of a small group of freshwater algae. About 1 billion years ago, a second endosymbiosis event is hypothesized to have occurred, of a second heterotroph engulfing and retaining a member of one of these clades. During

this secondary endosymbiosis event, the clades rhodophyta and chlorophytes gave rise to the dominant lineages of algae, such as stramenopiles, dinoflagellates, cryptophytes, and haptophytes. In present-day oceans, the most abundant, diverse, and ecologically important microbes are coccolithophores, diatoms, and dinoflagellates [Shemi et al., 2015].

These microbial community interactions that support their evolution and adaptation, are also the engines that drive the biogeochemical cycle of elements that occur in the oceans, such as the nitrogen and carbon cycles [Brussaard et al., 2016], [Falkowski et al., 2008]. Nitrogen is vital for life on earth and while nitrogen (N_2) is abundant in the atmosphere, this form is inaccessible for use, therefore a conversion into ammonia (NH_3) is required so it may be utilised and this is achieved through a process called nitrogen fixation [Wernegreen, 2012], [Voss et al., 2013]. There are certain bacteria and archaeal groups able to perform nitrogen fixation, and also some eukaryotic lineages have this ability as they required it through endosymbiosis in order to live in nitrogen-poor habitats. For example, microbial communities of cyanobacteria that live within the eukaryote phytoplankton species called diatoms perform this nitrogen fixation by their endosymbiosis mutualism interactions [Foster et al., 2011], [Wernegreen, 2012].

The world's surface ocean is divided into latitudinal temperature zones, ranging from about 30°C in the tropics to about -1.8°C in the ocean's polar sea and ice interface, and in the sea ice, the temperature can even be well below -1.8°C [Toseland et al., 2013]. Phytoplankton species can be found inhabiting all these temperature zones and due to their photosynthetic capabilities can be found inhabiting the surface layers of the ocean [Simon et al., 2009], [Aryal et al., 2015]. In the ocean, mixing occurs mainly during storms by wind and waves, and stratification of the water column occurs as the surface waters warm and storm-driven mixing decreases. Stratification layers consist of warm, low nutrient and illuminated water over deeper, darker, and cooler high nutrient water. These layers are divided by rapid changes over depth of water density, called the pycnocline, and temperature called the thermocline [Brierley, 2017]. In the ocean, the supply of nutrients is dependent on temperature-driven stratification and mixing of the ocean [Toseland et al., 2013]. The growth and diversity of phytoplankton is dependent on their optimal growth temperature and on the supply of nutrients [Toseland et al., 2013]. Rapid phytoplankton growth episodes known as blooms are so big that they

can even be seen from space as shown in figure 1.2 [Brierley, 2017].

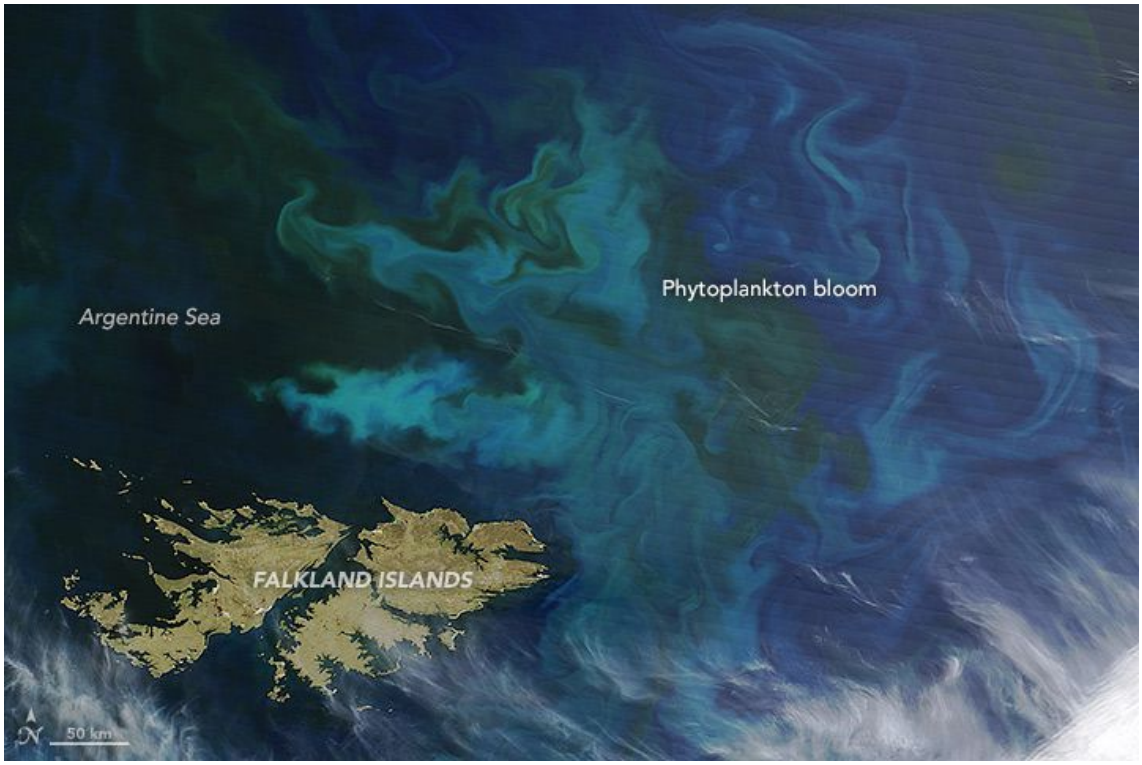

Figure 1.2: The Argentine Sea northeast off the coast of the Falkland Islands. The spiralling pattern of greens and blues are phytoplankton growing on the surface of the Argentine Sea. The image was captured with the Moderate Resolution Imaging Spectroradiometer (MODIS) on NASA's Terra satellite on December 2, 2015. (Photo taken by Jeff Schmaltz, LANCE/EOSDIS Rapid Response, NASA's Earth Observatory)

In recent work [Collins et al., 2014], the effect of recent human activity has been observed. In particular, the amount of carbon dioxide and bicarbonate in the world's oceans is increasing because of anthropogenic (pollution from human activity) carbon dioxide and the marine ecosystems will likely be affected by this increase in the oceans and atmosphere. This is causing ocean acidification, which is a decrease in the oceans pH. The mean pH of the surface oceans has decreased by about 0.1 pH units since the industrial revolution and is likely to decrease by an extra 0.3 pH units by the end of the 21st century, therefore this will result in an increase in acidity of about 150%. It is likely that the biogeochemical cycle of elements will be affected by ocean acidification, as it can change the community composition and can push for physiological and evolutionary change. Additionally, since the industrial revolution, the average ocean surface water temperature has already increased by 0.7°C and is likely to increase by an extra 3°C by the end of the 21st century. An increase of stratification of the oceans surface water is caused by an increase in temperature, which affects the light regime and also decreases

the amounts of nutrient supplied from below [Collins et al., 2014].

Phytoplankton divide their cells asexually at a rapid rate in an order of hours to days and have enormous population sizes. It is these features that enable phytoplankton to evolve in response to changing environmental conditions on time scales of weeks, months or years. For many years phytoplankton such as coccolithophores, diatoms, and dinoflagellate have been used as environmental indicators, as they are abundant in the oceans and have a substantial fossil record. Each year marine phytoplankton are responsible for about 50% of the carbon dioxide that is fixed in the atmosphere [Toseland et al., 2013]. Also, phytoplankton contribute to the base of the marine food web [Sarmento et al., 2010]. Therefore it is crucial to know how these marine microbes will respond to changing environmental conditions as this will affect the marine food web and biogeochemical cycles [Ribeiro et al., 2013], [Collins et al., 2014]. Marine microbes are extremely difficult to isolate from the environment and one possible reason is that during laboratory culturing, community interactions which are important for growth may be destroyed [Joint et al., 2010], [Kazamia et al., 2016]. We have known for some time that standard laboratory culturing techniques can only isolate a very small proportion of microbes [Joint et al., 2010]. Genomic data analysis of microbes has resulted in many insights being discovered such as how they evolved to be the biogeochemical engineers of life [Falkowski et al., 2008].

1.2 Thesis scope

We use multiple sequencing data types and develop bioinformatic approaches to assess microbial communities in the surface ocean. This thesis is a collaboration between the University of East Anglia (UEA), Earlham Institute (EI) and Joint Genome Institute (JGI). Dr.Katrin Schmidt, an environmental scientist who was based at UEA, sampled stations from a transect of the Arctic Ocean and the South Atlantic Ocean. In figure 1.3 are photos taken while Dr.Schmidt was on the expedition to collect the samples. Dr.Willem van de Poll of the University of Groningen, Netherlands and Dr.Klaas Timmermans of the Royal Netherlands Institute for sea research sampled stations from the North Atlantic Ocean.

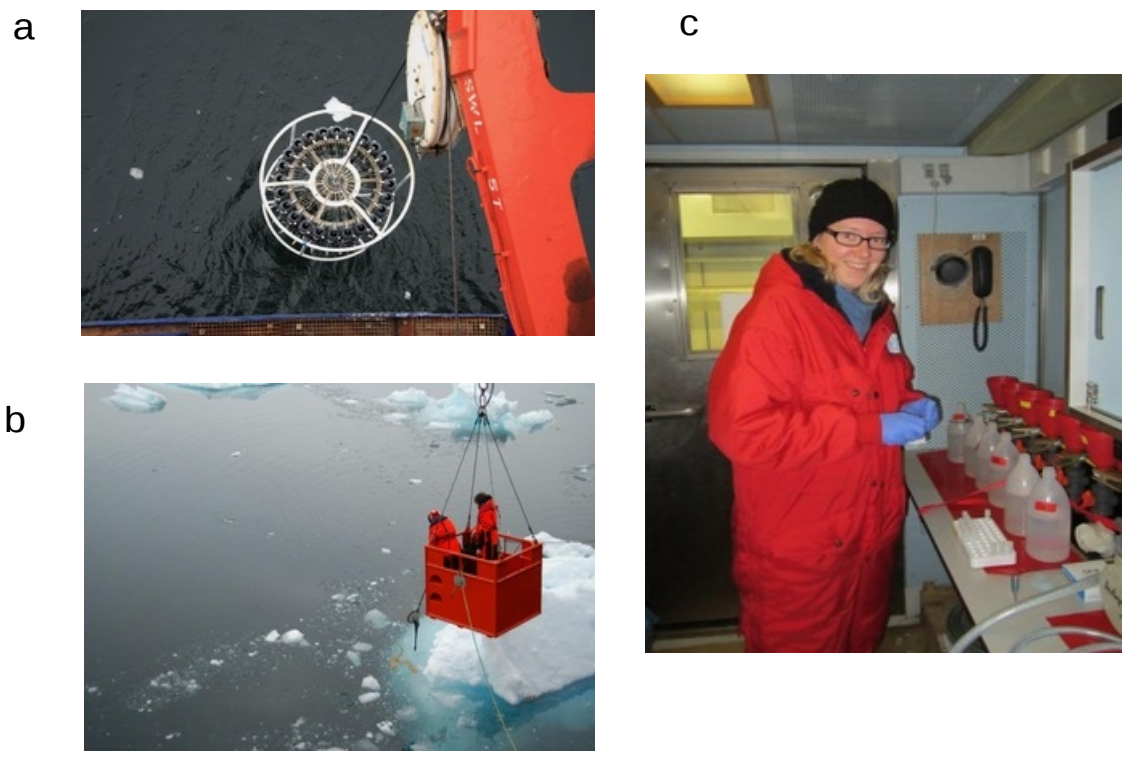

Figure 1.3: Photos taken on the boat during an expedition to collect the Arctic Ocean samples. (**a**) photo of the conductivity, temperature, and depth (CTD) rosette sampler that was used to collect water. (**b**) two crew who are helping by positioning the CTD in the water. (**c**) Dr.Schmidt working in the lab preparing the samples

The samples were sequenced with Illumina technology for metatranscriptome, metagenome, 18S and 16S rDNA analysis. This provided us with large amounts of sequence data, which enabled analysis across a wide range of environmental conditions. This could potentially enable us to forecast how climate change will affect these phytoplankton communities. The main challenge of this project was to develop a strategy to analyse the huge amounts of sequence data and to develop computational tools when necessary to extract pertinent information from the data. To address this challenge, we developed pipelines to analyse 18S rDNA data, 16S rDNA data and metatranscriptome data.

There have been other studies in the areas of phylogenetics and metatranscriptomics of marine samples, for example, [Alemzadeh et al., 2014] and [Alexander et al., 2015]. However, these focus on single locations such as the Persian Gulf and Narragansett Bay, United States of America, respectively. Most recent and most notable is the Tara Oceans project [Bork et al., 2015]. This paper describes data taken from sample locations from India around South Africa and over to South America into the Pacific

Ocean [Bork et al., 2015]. These are tropical and temperate regions in the oceans and sampled mainly in the open ocean. The novelty of our work is that we sampled close to the coast for a transect of the Arctic Ocean down through the North Atlantic Ocean and to Cape Town in the South Atlantic Ocean as shown in figure 3.1**a**. No other study has done this before. Additionally, we developed novel analysis pipelines such as the 18S rDNA taxonomic classification pipeline as well as our development of new approaches to normalising 18S and 16S rDNA copy numbers in our 18S and 16S rDNA datasets. In addition, environmental data at the time of sample collection, such as temperature and salinity, was recorded, which we included in our analysis, including co-occurrence analysis to identify community networks and breakpoint analysis which have not been used in this context before.

1.3 Summary of thesis

In chapter 2 we will summarise bioinformatics tools and methods that we use for the analysis of our 18S and 16S rDNA datasets. We will discuss these tools under their respective headings of sequencing processing, database searching and phylogenetic analysis, and in terms of what these are and their importance. Also, we will give a brief background on next-generation sequencing technology.

In chapter 3 we will discuss the computational pipeline and the analysis for the 18S and 16S rDNA datasets. We describe the sampling and preparation that was performed by Dr.Katrin Schmidt, Dr.Klaas Timmermans and Dr.Willem van de Poll. We also describe the sequencing and preprocessing of the 18S rDNA and 16S rDNA datasets, and the computational pipeline of the 16S rDNA dataset that was performed by JGI. Our contribution was the implementation of the computational pipeline for the 18S rDNA dataset, and the analysis of 18S and 16S rDNA datasets, as well as additional methods for the normalisation of the 18S and 16S rDNA copy number.

In chapter 4 we will discuss the computational pipeline and the analysis for the metatranscriptomic datasets. We also describe the sequencing, preprocessing and the computational pipeline of the metatranscriptomic dataset that was performed by JGI. Our contribution was the implementation of the computational pipeline for the analysis of the metatranscriptomic dataset.

In chapter 5 we will discuss our conclusions on our analyses of the metatranscriptomic, 18S and 16S rDNA datasets. We also discuss future work, such as additional analyses to be performed on our metatranscriptomic dataset, and the analysis of associated metagenomic datasets.

Chapter 2

Sequencing and sequence analysis

2.1 Summary

In this chapter, we shall summarise a number of the bioinformatics tools and methods that have been used for the analysis of our 18S and 16S rDNA datasets. We also give some background on 18S and 16S rDNA, metatranscriptomics and next-generation sequencing technology. We discuss sequence processing, database searching and phylogenetic analysis, describing what they are, why they are important and the tools that are involved in these approaches that are applied in this thesis.

2.2 Next-generation sequencing technologies

All organisms on Earth can be defined by their genome. The genome contains the biological blueprints for the construction and maintenance of that organism. For cellular organisms, the genome is made up of deoxyribonucleic acid (DNA) but the genome of a few viruses consist of ribonucleic acid (RNA). DNA and RNA are polymeric molecules that consist of a chain of monomeric subunits called nucleotides. There are five chemically distinct nucleotides, for DNA, these are Adenine, Thymine, Cytosine and Guanine and in the case of RNA, these are Adenine, Cytosine, Guanine and Uracil instead of Thymine [Brown, 2002]. Segments of DNA that encode for proteins with function or phenotype are called genes [Wain et al., 2002]. These protein coding genes are expressed by the process of transcription into messenger RNA (mRNA) transcripts and these, in turn, are translated into proteins. These proteins specify the type of the

biochemical processes that the cell is able to carry out [Brown, 2002].

Sequencing is a method for determining the DNA or RNA sequences within organisms [Sanger et al., 1977], [Wang et al., 2009]. Sequencing technology has revolutionised biological research, making possible the sequencing of entire genomes or a particular area of interest within a genome and have facilitated the further development of fields such as phylogenetics [Behjati and Tarpey, 2013], [van Dijk et al., 2014]. Additionally, sequencing technology has enabled the study of the transcriptome, which is the full set of mRNA transcripts in a cell. This consists of quantifying the types and amounts of transcripts in a cell under varying conditions [Wang et al., 2009].

In the 1970's, Sanger et al. developed the first generation DNA sequencing technology known as chain termination [Sanger et al., 1977]; and Maxam and Gilbert developed an alternative method called the chemical degradation method [Maxam and Gilbert, 1977] [Brown, 2002], [van Dijk et al., 2014]. Due to the radioisotopes and level of toxic chemicals involved in Maxam and Gilbert's chemical degradation method, the chain termination method became the predominant DNA sequencing method for the next 30 years [van Dijk et al., 2014]. Since the 1990's, Sanger sequencing biochemistry has mostly been carried out by a capillary based method and is semi-automated [Shendure and Ji, 2008].

Within Sanger sequencing high throughput pipelines, DNA can be prepared by a method called targeted resequencing, which consist of Polymerase chain reaction (PCR) amplification with primers that flank the target DNA. This results in many PCR amplicons within a single reaction volume. Cycles of template denaturation, primer annealing and primer extension are performed during the cycle sequencing reaction. The primer sequence is complementary and known to the region flanking the sequence of interest. Each round of primer extension is randomly terminated when fluorescently labelled dideoxynucleotides (ddNTPs) are incorporated. This results in a mixture of end labeled extension products. The product's label on the terminating ddNTP each corresponds to the nucleotide identity of its terminal position. The sequences of the single stranded products in a capillary based polymer gel are determined using high resolution electrophoresis. As they exit the capillary the fluorescent labels are excited by a laser. Also attached is a colour detection of emission spectrum in order to provide an output in the form of a plot called a Sanger sequencing trace. The Sanger

sequencing trace is then converted into DNA sequence with error probabilities for each base call [Shendure and Ji, 2008].

Though accurate, Sanger sequencing is low yielding and expensive [Reuter et al., 2015]. By comparison, the second generation of DNA sequencing technologies often referred to as next-generation sequencing (or NGS for short) can perform reactions in parallel producing much more data at a lower cost [van Dijk et al., 2014]. Our samples were sequenced with Illumina NGS [Bennett, 2004] by Joint Genome Institute (JGI), for metatranscriptome, metagenome, 18S and 16S rDNA sequences. Illumina HiSeq2000, for example, can produce 150-200 Gb (gigabase) of data per run and at a cost of $0.02 per million bases [Pillai et al., 2017].

The Illumina methodology differs from Sanger sequencing in that it uses the technology of sequencing by synthesis (SBS) [Liu et al., 2012]. For Illumina, sequencing occurs within a flow cell which contains one, two or eight separate lanes [Buermans and den Dunnen, 2014]. Adapters which are essentially sequencing primer annealing sequences are ligated to each of the DNA fragments [Kozarewa et al., 2009]. A library of DNA molecules with attached sequencing adapters is denatured to produce single stands. These are passed through a flowcell and attached to the complementary oligonucleotides that are spread over the flowcell. This is followed by a procedure called bridge amplification, which is a solid phase PCR which forms clusters of clonal DNA fragments [Liu et al., 2012], [Heather and Chain, 2016]. SBS involves additions of fluorescent reversible-terminator dNTPs, this results in no more nucleotides being able to bind to the DNA molecule. Before polymerisation can advance, these must be cleaved off thus allowing the sequencing to occur at the same time throughout. In cycles, the altered dNTPs and DNA polymerase are passed through the flowcell containing the primed single stranded clusters. For each cycle, the incorporating nucleotide is identified with a charge coupled device (CCD) by exciting the fluorophores with suitable lasers, before enzymatic removal of the blocking fluorescent component and then continues to the next position [Heather and Chain, 2016].

While Illumina's second generation technology still dominates the market, two third generation technologies from Pacific Biosciences and Oxford Nanopore Technologies have begun to gain traction. These technologies are both single molecule sequencers, requiring no artificial amplification and are characterised by long reads (in the thou-

sands or tens of thousands of bases) and a higher error rate than second generation technologies [Liu et al., 2012], [van Dijk et al., 2014]. Nanopore sequencing technology was first developed by David Deamer (University of California Santa Cruz), George Church and Daniel Branton (Harvard University) [Jain et al., 2016]. A number of companies have developed nanopore based sequencing technologies, but Oxford Nanopore Technologies (ONT) is the only one thus far to bring a product to market (in 2014) with the release of MinION.

The MinION is the smallest portable sequencing device to date, weighing 90g and measuring 10 x 3 x 2 cm and powered from a standard USB3 port [Jain et al., 2016], [Lu et al., 2016]. There is no charge for the device itself, with labs paying only for consumables. It can output read lengths of tens of kilobases limited only by the length of DNA molecules inserted into it and with a single read accuracy of 95% [Laver et al., 2015], [Jain et al., 2016], [Carter and Hussain, 2017]. The MinION contains a flowcell with 2048 nanopores arranged in groups of 4 under 512 current sensors. Before sequencing, adapters are ligated to both ends of the DNA or cDNA fragments which enables the strands to be captured and loading of the processive enzyme (motor protein) at the $5'$ -end of each strand, thus ensuring unidirectional single nucleotide displacement along the strand. The adapters increase the DNA capture rate by several thousand fold by concentrating the DNA substrates at the membrane surface near to the nanopore. When the DNA molecule is captured in the nanopore, the motor protein advances the template strand through the nanopore. Once the enzyme passes through the hairpin, this is repeated for the complementary strand. Each sensor monitors the changes in ionic current as the DNA moves through the pore. The changes in the ionic current are divided into distinct events to which a duration, mean amplitude, and variance can be associated. Using probabilistic models this is interpreted computationally as a sequence of 3 to 6 nucleotide kmers [Jain et al., 2016].

In figure 2.1 we give a high-level overview of how a typical analysis of data proceeds in this thesis. After samples have been sequenced as outlined here in section 2.2, sequence preprocessing takes place, which we outline in section 2.3. Once the reads have been prepared the next step is to identify the reads by phylogenetic analysis, which includes searching databases. We outline database searching in section 2.4 and phylogenetic analysis in section 2.5.

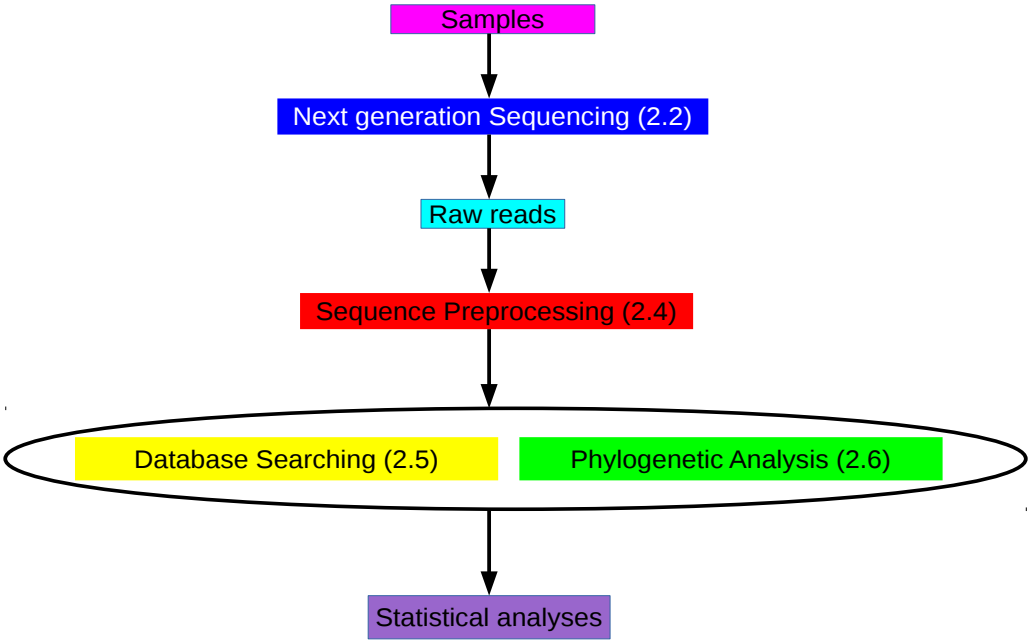

Figure 2.1: Workflow diagram of a typical computational analysis. We begin by sequencing the samples (pink) on next-generation sequencing (blue) platform which results in raw reads (turquoise). The raw reads (turquoise) undergo sequencing preprocessing (red) to prepare them for analysis. We conduct phylogenetic analysis (green) which includes database searching (yellow). Finally, we perform statistical analyses (purple)

2.3 Datasets

2.3.1 18S rDNA and 16S rDNA

Within eukaryotic and prokaryotic cells, the sites of protein synthesis are the ribosomes. In prokaryote cells, the ribosomes are called 70 Svedberg units (S) and in eukaryote cells, the ribosomes which are slightly larger than the prokaryote ribosomes are called 80S. Overall the structure of the ribosomes for eukaryotes and prokaryotes cells are similar. In eukaryotes and prokaryotes, the ribosome is composed of a large subunit and a small subunit. For prokaryote ribosome 70S, the smaller subunit referred to as 30S consists of 16S ribosomal RNA (rRNA) and 21 proteins. The larger subunit referred to as 50S consists of 23S and 5S rRNAs and 34 proteins. For eukaryotic ribosome 80S, the smaller subunit referred to as 40S consists of 18S rRNA and 30 proteins. The larger subunit referred to as 60S consists of 28S, 5.8S, and 5S rRNAs and about 45 proteins [Cooper, 2000]. Eukaryotic and prokaryotic cells typically contain many ribosomes and within each small ribosomal subunit a single RNA species exists of 18S

rRNA and 16S rRNA, respectively [Cooper, 2000], [Amit Roy et al., 2014].

The ribosomal DNA (rDNA) sequences encode for rRNAs, and these are tandemly repeated [Gibbons et al., 2014]. The rDNA copy number can vary greatly among prokaryotes and eukaryotes. For prokaryotes, the 16S rDNA copy number can vary as much as from one to fifteen and for eukaryotes, the 18S rDNA copy number can vary even greater from one to thousands [de Vargas et al., 2015], [Perisin et al., 2016]. The sequence that is composed of repeated rRNA gene clusters separated by intergenic spacers (IGS) is referred to as the ribosomal gene array. For eukaryotes, each cluster is made up of sequences of rRNA encoding highly conserved genes for 18S rRNA, 5.8S rRNA and 28S rRNA and these are separated by internal transcribed spacers which are referred to as ITS1 and ITS2, and also external transcribed spacers called 5ETS and 3ETS, which are located downstream of the 18S rRNA gene and upstream of the 28S rRNA gene respectively [Dyomin et al., 2016], [Fernández-Pérez et al., 2018]. The 5S rDNA sequence which codes for 5S rRNA is clustered in tandem arrays and is separated by flanking DNA sequences called non-transcribed spacers (NTSs) [Fernández-Pérez et al., 2018]. All these components are transcribed into a single RNA precursor, pre-rRNA [Dyomin et al., 2016]. The ETS, ITS and NTSs sequences are lost during maturation [Mandal, 1984], [Fernández-Pérez et al., 2018]. For prokaryotes the rDNA encoding the rRNA are traditionally arranged in a single operon in the order of 16S rDNA, 23S rDNA and 5S rDNA and these are separated by ITSs. All the components are transcribed into a single RNA precursor, which is then separated and processed by RNases [Brewer et al., 2019].

For taxonomic classification, the small subunit rRNA gene is a standard reference sequence, the 18S rDNA for eukaryotes and 16S rDNA for prokaryotes [Wang et al., 2014]. The 16S and 18S rDNA molecular markers are the most popular and ideal for phylogenetic studies [Fu and Gong, 2017]. Features that make rDNA so ideal for phylogenetic studies, for example, are that rDNA contains highly conserved and variable domains, it evolves slower than protein coding genes and is universal among organisms [Johnston, 2006], [Amit Roy et al., 2014]. The 18S and 16S rDNA are highly conserved within eukaryotes and prokaryotes, respectively [Amit Roy et al., 2014], [Wu et al., 2015]. The 18S rDNA is about 1800bp long and 16S rDNA is about 1550 bp long. Each of the 18S and 16S rDNA sequences contains both variable

and conserved regions with characteristic oligonucleotide signature sequences that are distinctive to a specific phylogenetic group [Iwen and Hinrichs, 2002], [Amit Roy et al., 2014]. SILVA [Quast et al., 2013a] is an online web resource for databases of ribosomal RNA (rRNA) gene sequences. SILVA contain sequences from the Bacteria, Archaea and Eukaryota domains [Quast et al., 2013a]. The SILVA database taxonomic rank assignments are manually curated [Balvočit and Huson, 2017]. SILVA is a popular choice for the thousands of researchers around the world who included SILVA in their work, such as [Lambrechts et al., 2019], in which samples were taken from the Antarctic to study the composition of prokaryotic communities [Quast et al., 2013a].

2.3.2 Metagenomes and Metatranscriptomes

Metagenomics, the direct genetic analysis of genomes within an environmental DNA sample, enables the novel exploration of functional gene composition and diversity of these microbial communities [Thomas et al., 2012b]. The same holds for metatranscriptomics which targets the mRNA within an environmental sample, thus enabling the investigation of taxonomic composition and the activity of biochemical functions of microbial community [Gilbert and Hughes, 2011], [Jiang et al., 2016a]. There are three main aims for the analysis of metatranscriptomic datasets: who is there, what are they doing and how do different samples compare? [Huson et al., 2011]. Metatranscriptomic analysis gives a new insight into microbial communities, and in particular how these microbial communities react to changes in the environment [Gilbert and Hughes, 2011], [Narayanasamy et al., 2016]. There have been a number of advances in microbial ecology, evolution, and diversity over the past decade, therefore a substantial number of research laboratories are actively engaged in the field [Thomas et al., 2012b]. Even so, there are a number of limitations, such as the small number of reference genomes, pipelines and computational tools, thus making it challenging to analyse and interpret metatranscriptomics datasets [Jiang et al., 2016b].

Metatranscriptomic approaches are being applied to investigate microbial communities in numerous different habitats including, for example, marine, soil and the human gut [Yu and Zhang, 2013]. Of most relevance to this thesis, the Tara Oceans project has sampled a large number of stations around the world's open oceans, upon which they

have published a number of papers including their metatranscriptomic analysis [Carradec et al., 2018], and in which they have characterised the changing genetic activity in the different organisms' size fractions across the open oceans. Another example is [Toseland et al., 2013], which included samples from the Arctic Ocean, North Atlantic Ocean, the North Pacific Ocean, the Equatorial Pacific and the Southern Ocean. This analysis included environmental data, which enabled the authors/researchers to investigate and identify that cellular resource allocation is significantly impacted by temperature [Toseland et al., 2013].

2.3.3 Dataset applications

It has been estimated that approximately 99% of microbes in the environment cannot be cultured under laboratory conditions. This greatly limits our understanding of microbial diversity, genetics and community ecology [Singh et al., 2009]. Advances in high-throughput sequencing technologies have vastly altered our ability to investigate these unculturable microbes [Bik, 2014]. Taxonomic classification involving the 18S and 16S rDNA provides identification of the reads in a sample for eukaryote and prokaryote species, respectively. These samples can also be compared for differences and similarity in composition, distribution and abundance [Wang et al., 2014]. While with metatranscriptomics, this provides insight into microbial communities, with a particular interest in how these microbial communities react to changes in the environment [Gilbert and Hughes, 2011], [Narayanasamy et al., 2016].

The 18S and 16S rDNA genes have been employed for decades for taxonomic classification studies [Wang et al., 2014]. In that time there have been numerous studies such as [Hunt et al., 2013] which employed 16S rDNA to investigate Bacterioplankton in the surface ocean, specifically looking at the relationship between abundance and their activity. Another example is [Stecher et al., 2016] which employed 18S rDNA to examine protist diversity in sea ice from the central Arctic Ocean. Default parameters were chosen for these and many other studies. When constructing our 18S rDNA pipeline we were confident that the default parameters would provide the best output as these parameters worked well in other studies. For instance in our 18S rDNA pipeline we employed a multiple sequence aligner called ClustalW [Thompson et al., 1994] and

used the default parameters. Likewise in the 18S rDNA study of [Shull et al., 2001] ClustalW was employed with default parameters to investigate the phylogenetic relationship in the suborder group of Adephaga [Shull et al., 2001]. Aligning protein coding genes such as 18S and 16S rDNA genes which are highly conserved within eukaryotes and prokaryotes, respectively, is not as difficult or problematic as aligning noncoding DNA sequences [Wang et al., 2006], [Amit Roy et al., 2014], [Wu et al., 2015]. There is a minimum amount of difficulty in producing convincing alignments of protein coding DNA sequences when sequence divergence is low, this is because indels mostly occur in multiples of three base pairs, and rarely within codon regions [Keightley and Johnson, 2004].

2.4 Sequence preprocessing methods

Sequence preprocessing is a common first step that usually involves quality control (QC), as well as identifying and filtering unwanted data. One source of the unwanted data is low quality sequences, which may occur as a result of sequencing instrument limitations or sample preparation problems [Zhou et al., 2014]. A major step in Illumina sequencing involves the ligation of the target DNA fragments to specific adapters for clonal amplification [Aird et al., 2011]. Artifacts such as adapter sequences which have not been fully removed are another source of unwanted data [Schmieder and Edwards, 2011]. Raw sequencing files can contain millions of reads and downstream analysis is computationally intensive, the demand on computational resources is one of the major bottlenecks for analysis [Schmieder and Edwards, 2011], [Yu and Zhang, 2013]. Unwanted data, if present, can drain resources and result in misassembly or erroneous conclusions (see for example [Schmieder and Edwards, 2011]).

Another aspect of sequence preprocessing is merging paired end reads. Illumina performs paired-end sequencing which can produce reads from both ends of target DNA fragments. If the insert size is sufficiently small, the two paired-end reads can be merged and this results in an increased overall read length that can have a significant and favourable affect on the quality of the analysis [Magoč and Salzberg, 2011], [Zhang et al., 2014]. A large number of short reads are generated from NGS technologies such as Illumina. Short read lengths are problematic for analysis such as *de novo* assemblies,

even for very deep genome coverage [Magoč and Salzberg, 2011]. Illumina sequences can produce single-end reads that range from 75 to 300 bp, and there is an increase in error rates as the reads get longer. When paired-end reads are merged sequencing errors can be corrected when the paired-end reads overlap and therefore potentially give higher quality reads [Zhang et al., 2014].

There are numerous tools available for sequencing preprocessing of raw sequencing files, such as those reviewed in [Zhou et al., 2015]. We do not describe all of the tools that are available, but only those used in our own pipelines.

2.4.1 Adapter searching and trimming

Duk

Duk [Li et al., 2011] (which stands for Decontamination Using kmers) is a matching tool for DNA sequences [Li et al., 2011], [DOE Joint Genome Institute, 2017]. This tool was implemented in our 18S rDNA pipeline. Duk (otherwise known as BBDuk) is part of a suite of bioinformatics tools called BBtools develped by the Joint Genome Institute [DOE Joint Genome Institute, 2017]. Duk is used to screen for contamination such as adapter sequences in raw reads after sequencing. Besides contamination removal, Duk has a range of other applications, such as organelle genome separation and assembly refinement [Li et al., 2011]. The matching technique used by Duk is similar to aligning sequences, but instead of creating an alignment between sequences, Duk takes a query sequence to search against a reference sequence for partial or total matches. Many traditional contamination searching tools are usually performed with an alignment technique. But since it is only necessary to know if a match is present or not, and not which bases of a query sequence match to which position of a reference sequence, Duk's performance is much faster than tools that employ alignment techniques [Li et al., 2011].

Duk uses a kmer (substring of size k) hashing method to index reference sequences to identify matching DNA in the query sequences [Li et al., 2011], [Mapleson et al., 2017]. Duk estimates the occurrence of the match by calculating p-values from a Poisson distribution. In order to calculate this Duk assumes that the kmers in the DNA sequence are randomly distributed. Assuming that there are M kmers in reference

sequences, N kmers in the query sequence and that the kmer size is k, the probability of having c or more common kmers between the reference sequence and the query sequence is given by the formula 1-pois(c,u), where pois is the Poisson distribution function and u=M*N/4 ^{k} [Li et al., 2011]. An output with a high p-value implies that the match between the query sequence and the reference sequence is probably the result of random sequence variation, as opposed to the match event being the result of sequence homology [Li et al., 2011].

Cutadapt

Cutadapt [Martin, 2011] is a software tool that searches and removes adapter sequences from reads that were left on after sequencing. This tool was implemented in our 18S rDNA pipeline. Adapter sequences are considered a form of contamination and must be removed so only the relevant part of the read is used for further analysis. A read which contains an adapter sequence can be trimmed or completely discarded. After trimming if a read falls outside a specified length range, the read can be discarded [Martin, 2011].

For each read, Cutadapt begins by calculating the optimal alignments between the read and all given adapter sequences. Cutadapt's algorithm is called regular semi-global alignment and it does not penalise initial or trailing gaps, thus allowing the read and the given adapter sequence to shift freely relative to one another. The user can tell Cutadapt to look for the adapter at the 3' end of the molecule by giving the "-a" parameter to provide the adapter sequence. This results in Cutadapt removing the adapter sequence and all the nucleotides after it. The adapter sequence must begin at the start or within the read. This is done by penalising initial gaps in the read sequence. If the location of the adapter sequence is unknown, the user can provide the "-b" parameter. Therefore, if the adapter sequence is overlapping the beginning of the read, all nucleotides before the first non-adapter nucleotide are deleted. After Cutadapt has aligned all adapters to the read, the alignment with the greatest number of matching nucleotides between the read and adapter is taken to be the best one. An error rate of e/l is calculated, where e is the number of errors and l is the length of the matching piece between read and adapter. The read is trimmed when the error rate is below the allowed maximum [Martin, 2011].

2.4.2 Read merging

USEARCH merging of paired reads

USEARCH merging of paired reads [Edgar, 2010b] is a tool that merges overlapping paired end reads into single sequences [Edgar, 2010e]. This tool was implemented in the Joint Genome Institute (JGI) 16S rDNA pipeline. The output gives increased length and higher quality of reads, thus improving the quality of the analysis [Zhang et al., 2014].

USEARCH merging of paired reads is accomplished by aligning the forward and reverse reads and also rectifying any mismatches that may have occurred during aligning. The reverse read is the reverse complement to the forward read. Therefore the reverse read is positioned on the same strand as the forward read. Generally, the alignment between the forward and reverse reads is covered between them partially. This, therefore, results in unaligned segments at the start of both reads. The alignment is staggered in the event that the sequencing construct is shorter than the read length, which results in the unaligned segments at the end of both reads rather than the start. By default, USEARCH with the command "fastq mergepairs" will trim the unaligned segments in the event that the alignment is staggered. When mismatches occur in the alignment, the base call with the biggest Q (Phred) score is selected, but if both bases have the same Q score then the forward read base is chosen [Edgar, 2010e].

FLASH

FLASH [Magoč and Salzberg, 2011] is a software tool that locates the correct overlap between paired-end reads and produces a merged read of greater length than a single read [Magoč and Salzberg, 2011]. This tool was implemented in our 18S rDNA pipeline. The output of increased read length results in a significant and beneficial affect on the quality of the analysis such as genome assembly [Magoč and Salzberg, 2011].

Paired-end reads are produced from both ends of the target DNA fragments. FLASH takes each read pair separately and looks for the right overlap between the paired-end reads. The two reads are merged when the right overlap is found, thus producing a longer read. The new longer read is of the same length as the original target DNA fragment from which the paired-end reads were generated. FLASH examines and scores

every possible ungapped alignment overlap between paired-end reads, in order to find the right overlap. FLASH's scoring system accounts for the number of bases overlapping between the paired-end reads, this parameter is called min-olap and the default is set to 10 base pairs (bp) [Magoč and Salzberg, 2011].

2.4.3 Read clustering

CD-HIT

CD-HIT [Li and Godzik, 2006] is a tool for clustering raw sequencing data. This tool was implemented in our 18S rDNA pipeline. Clustering is a data reduction strategy that reduces sequence redundancy and therefore improves the performance of downstream analysis, for example reducing storage space and computational time [Fu et al., 2012]. At first, CD-HIT was designed to cluster protein sequences to build reduced redundancy reference databases such as UniProt, but it was later expanded to support clustering nucleotide sequences [Li and Godzik, 2006], [Fu et al., 2012]. Since CD-HIT's release, it has become very popular for a large range of applications including, for example, protein family classification, metagenomics annotation and identifying artifacts in datasets [Fu et al., 2012].

CD-HIT implements a greedy incremental algorithm. Firstly, CD-HIT orders the sequences by length. The longest sequence becomes the seed for the first cluster and each remaining sequence is compared with established seeds. If there is a similarity with any seed which meets a pre-defined cutoff, it is grouped into that cluster; else, it begins a new cluster [Li et al., 2012]. The similarity between seeds is determined by common word counting, by using word indexing and counting tables. This results in filtering out the undesirable sequence alignment, which can be used to calculate exact similarity. Therefore, a big redundant dataset can be represented by a smaller non-redundant dataset, then each cluster can be represented by a single entry [Fu et al., 2012].

USEARCH cluster otus

USEARCH cluster otus [Edgar, 2010b] is a tool that is employed to cluster sequences, specifically reads from a marker gene amplicon sequencing experiment such as 16S

rDNA [Edgar, 2010a]. This tool was implemented in the JGI 16S rDNA pipeline. As stated above clustering is a data reduction strategy [Fu et al., 2012].

USEARCH cluster otus performs operational taxonomic unit (OTU) clustering, with identity threshold fixed at 97% using the UPARSE-OTU algorithm [Edgar, 2010a]. The aim of the UPARSE-OTU algorithm is to find a set of OTU representative sequences and these must satisfy four criteria points. The criteria points are as follows, one, all pairs of OTU sequences must possess 97% or greater pairwise sequence identity. Two, the OTU sequence must be the most abundant within the 97% group [Edgar, 2010c]. Three, a chimeric amplicon occurs when a partially complete DNA strand anneals to a different template. A new template is synthesised by the primers based on the two different biological sequences and therefore any chimeric sequences identified are discarded [Edgar, 2010c], [Edgar, 2016]. Four, all non-chimeric input sequences must match at least one OTU with 97% or greater pairwise sequence identity. UPARSE-OTU is a greedy algorithm. The input to the UPARSE-OTU algorithm is a set of sequences and each sequence is labelled with a number representing its abundance of reads having a given unique sequence. The sequences are ordered in decreasing abundance, as OTU centroids (representative sequences) are chosen from the more abundant reads [Edgar, 2010c].

Every input sequence is compared to the current OTU database and using the UPARSE-REF [Edgar, 2010d] algorithm a maximum parsimony model of the sequence is determined [Edgar, 2010c]. There are three cases as follows, one, the UPARSE-REF model has 97% or greater sequence identity to an existing OTU, and therefore that input sequence becomes a member of the OTU [Edgar, 2010c]. Two, the model is chimeric, and then the input sequence is discarded [Edgar, 2010c]. Three, the model is greater than 97% and possesses the highest sequence identity to any current OTU. Therefore the input sequence is added to the database and also becomes the centroid of a new OTU [Edgar, 2010c].

2.5 Database searching

In a typical analysis, the datasets of sequences first undergo a similarity search in order to classify these sequences. This is achieved by comparing the sequences to public

databases of annotated sequences such as GenBank [NCBI Resource Coordinators, 2016] or SILVA [Quast et al., 2013b] [Bazinet and Cummings, 2012], [Yu and Zhang, 2013]. Sequence alignments are used to identify similarity between sequences [Eric et al., 2014]. For instance, metagenomic environmental samples contain DNA sequences from many different species and the typical aims of a metagenomic analysis are to try to identify what species and genes are present [Thomas et al., 2012a], [Suzuki et al., 2015]. There are a number of other fields where sequence homology searches are common, including phylogenetic analysis [Pearson, 2013], [Suzuki et al., 2015]. A well founded evolutionary hypothesis based on molecular sequences is dependent largely on the ability to align sequences of nucleotides or amino acids in the same position [Barta, 1997]. There are a number of tools and databases available and we outline the tool we used in the next section.

HMMER

HMMER [Eddy, 1996] is a tool for sequence searching, which uses probabilistic methods known as Hidden Markov Models (HMMs) in the analysis of homologous amino acid and nucleotide sequences with a high degree of sensitivity [Finn et al., 2011], [Ferreira et al., 2014], [Jiang et al., 2016b]. This tool was implemented in our 18S rDNA pipeline. HMMER, which employs Profile Hidden Markov Models, provides a sensitive approach for the detection of distant homologs since sequence profiles can give an improved representation of a set of homologous sequences compared with a single sequence [Sinha and Lynn, 2014].

There are two main stages for profile HMM homology detection; first model building and then database searching. Building the model entails the conversion of a multiple sequence alignment into a probabilistic model, and database searching is the scoring of a sequence to the profile HMM [Wistrand and Sonnhammer, 2005]. As compared to searching with a single query sequence, a previously built sequence consensus is used. This is called a consensus profile, which gives a more flexible method for the identification of homologs of a particular family, by highlighting the features they have in common and by lessening the value of the divergences between the sequences [Ferreira et al., 2014]. The profile HMM is mainly comprised of three types of states, one for each of the labels we could assign to a nucleotide therefore corresponding to matches

or mismatches, insertions and deletions, all with precise transitions between the three states [Eddy, 2004], [Ferreira et al., 2014]. The most common algorithms to process HMMs are the Forward algorithm and the Viterbi algorithm. The Forward algorithm computes a full probability for all possible model state paths and the Viterbi algorithm provides the best possible sequence of model states for the generation of the query sequence. The Viterbi algorithm obtains the entire path of states, which corresponds to an optimal alignment of the query sequence to the profile model [Ferreira et al., 2014]. Profile HMMs have a reputation for generating good results, and therefore are employed by a number of databases, such as Pfam [El-Gebali et al., 2019] and Superfamily [Gough et al., 2001]. Within such databases there are a large collection of protein families where each family is represented by a profile HMM and this is what is commonly used to represent the family in database searches [Wistrand and Sonnhammer, 2005].

USEARCH oligodb

USEARCH oligodb [Edgar, 2010b] is a tool that is employed to search for matches to nucleotide sequences in a database of short nucleotide sequences. This tool was implemented in the JGI 16S rDNA pipeline. USEARCH oligodb is typically employed in the search of matches of primers or probes to genome sequences or to gene databases [Edgar, 2010f]. The success of any polymerase chain reaction (PCR) based method is greatly dependant on the correct nucleic acid sequence. The sensitivity and specificity of primers or probes are predicted by searching a database to find sequences that contain the ideal number of mismatches and similarity [Kalendar et al., 2017]. The USEARCH oligodb algorithm is not heuristic, it is therefore exact. The alignments are global, with no gaps allowed except in the case of terminal gaps in the query sequence [Edgar, 2010f].

2.6 Phylogenetic analysis

Phylogenetic analysis is the study of the evolution of species, by examining the relationships between molecules, phenotypes and organisms [Singh et al., 2009], [Rokas, 2011]. The goal of a phylogenetic study is to establish which tree out of all possible

trees gives the best estimate for the true evolutionary relationships of the dataset analysed according to some optimisation criterion. Unfortunately, due to computational expense, it is not typically possible to determine the best tree amidst all possible trees for the sequence data. This is even true for a small number of sequences, as the number of alternative trees are extraordinarily large since the number of all possible trees grows exponentially with the number of sequences. For instance, the number of different phylogenetic trees that portray the evolutionary relationships of 50 sequences is roughly the number of atoms in the known universe [Rokas, 2011].

Phylogenetic analysis is a standard and very important tool for any bioinformatician, since it helps in understanding big evolutionary questions, such as the origins and history of macromolecules, developmental systems, phenotypes, and of course life [Rokas, 2011]. In our approach, phylogenetic analysis is achieved by cloning and sequencing the ribosomal RNA (rRNA) genes, in particular the 16S/18S small subunit rRNA [von Mering et al., 2007]. This can closely estimate the level of species diversity and unusual organisms can also be identified by this approach [von Mering et al., 2007], [Medlar et al., 2014]. Also, determination of the taxonomic composition of environmental samples can provide important indicators into the underlying communities' ecology and function [von Mering et al., 2007]. Phylogenetic analysis is also essential to gene discovery and annotation, to prediction of gene function, and the identification and construction of gene families [Rokas, 2011].

There are a wide numbers of tools available for phylogenetic analysis [Pavlopoulos et al., 2010]. In the following sections we describe the tools that we used in our bioinformatics pipelines. These sections are ordered as in a typical phylogenetic analysis. First sequences are aligned, then a tree is built, so that new sequences can be placed onto this tree, and finally this tree is visualised.

2.6.1 Alignments

ClustalW

ClustalW [Thompson et al., 1994] is a heuristic multiple sequence alignment (MSA) program. This tool was implemented in our 18S rDNA pipeline. MSAs are vital to many areas in bioinformatics, for example in homology modelling and phylogenetic

analysis [Capella-Gutierrez et al., 2009]. The aim of MSA is to output an arrangement of a set of sequences, with the aim that similar sequence features are aligned together, so that patterns can be identified that may be common among many sequences, or changes revealed that may clarify functional and phenotypic variability. A feature can be defined as any relevant biological information, that is, structure, function or homology to the common ancestor [Kemena and Notredame, 2009]. The quality of MSAs for these applications is critical for the reliability and accuracy of the analyses. A large number of algorithms for MSA are presently available, which apply different heuristic algorithms to find the optimal solutions to the alignment problem. 80-90% accuracies have been reported for the best MSA algorithms, but even these algorithms can fail at specific regions in the alignment. For large scale analysis the problem gets worse, due to the implementation of faster algorithms that are less reliable [Capella-Gutierrez et al., 2009].

The basic ClustalW algorithm consists of three key parts. First all pairs of sequences are aligned separately and from this a distance matrix is calculated, thus giving the divergence of each pair of sequences. Second, from the distance matrix and using the Neighbor-Joining [Saitou and Nei, 1987] method a guide tree is calculated. Initially an unrooted tree is constructed with branch lengths proportional to the approximate divergence for each branch. By employing the mid-point method a root is placed at a point on the tree, in which the branch lengths on either side of the root are equal. Third, according to the branching order in the guide tree, from the leaves of the rooted tree towards the root, the sequences are progressively aligned. A dynamic programming algorithm at each stage of the alignment is performed with a residue weight matrix and also penalties for opening and extending gaps. Each part is made up of aligning two existing alignments or sequences and gaps that are introduced in previous alignments remain unaltered. New gaps that are added at each stage get full gap opening and extension penalties, regardless of whether or not they are added inside an original gap location. The score at a position from one sequence or alignment and another sequence or alignment is calculated based on the average of all the pairwise weight matrix scores from the sets of sequences used. If any set of sequences has one or more gaps in one of the locations being considered, this gets scored a zero if it is a gap versus a residue. The default amino acid weight matrices used are adjusted to

be assigned positive values. Consequently, this treatment of gaps results in the score of a residue versus a gap ending up with the worst possible score. Therefore when the sequences are weighted, each weight matrix value is multiplied by the weights from the two sequences [Thompson et al., 1994].

trimAL

trimAl [Capella-Gutierrez et al., 2009] is an automated trimming tool for multiple sequence alignment. This tool was implemented in our 18S rDNA pipeline. It has been reported that the removal of poorly aligned regions from an alignment increases the quality of further analyses [Capella-Gutierrez et al., 2009]. trimAl firstly reads all the columns in the alignment and calculates a score, a gap score, a similarity score or a consistency score for each of the columns. The score for each column is calculated on information from that column or, if a window size is given, it relates to the average value of the given window size columns around the position being considered. The gap score for a given column is the fraction of sequences with no gap in that specified position. The residue similarity score uses the mean distance score between pairs of residues, as defined by a given scoring matrix. The consistency score is only calculated when more than one alignment for the same set of sequences is given. The consistency score is the level of consistency for all residue pairs located in a given column as compared with other alignments. The alignment with the highest consistency score is trimmed to remove the columns that are less conserved.

trimAL can proceed in two ways after all the column scores have been calculated. A conservation threshold relates to the minimum percentage of columns, from the initial alignment, that the user would like to have in the trimmed multiple sequence alignment. If a score and a minimum conservation threshold are provided, trimAL will output a trimmed alignment. This alignment will consist only of the columns with scores greater than the score threshold. If the number falls below the conservation threshold, in a decreasing order of scores trimAl will add more columns to the trimmed alignment until the conservation threshold is hit. Alternatively, trimAl has three modes for the automated selection of parameters- gappyout, strict and strictplus- these are based on the use of gap and similarity scores. trimAl will calculate the specific score thresholds based on the characteristics of each alignment. trimAL also has an option

which implements a heuristic in order to decide the appropriate mode depending on the alignment characteristics [Capella-Gutierrez et al., 2009].

2.6.2 Tree building and placing

RAxML

RAxML [Stamatakis, 2014] (Randomised Axelerated Maximum Likelihood), is a tool for phylogenetic analysis of large datasets under maximum likelihood [Stamatakis, 2014]. This tool was implemented in our 18S rDNA pipeline. There has been a staggering accumulation of genetic information from various organisms in recent years [Stamatakis et al., 2005]. There are a number of approaches such as maximum likelihood and Bayesian methods which can be used to compute these relationships. Maximum likelihood is incorporated in RAxML and is considered to represent one of the more accurate approaches available for phylogenetic reconstruction [Stamatakis et al., 2005].

RAxML has a fast maximum likelihood tree search algorithm, and has been shown to return trees with "good" likelihood scores [Stamatakis, 2014]. RAxML generates a starting tree, which is built by adding sequences one at a time at random, using the parsimony optimality criterion to identify their optimal location on the tree. Due to the random order in which the sequences are added, this is likely to generate several different starting trees each time a new analysis is run. This can result in a improved exploration of the tree space. Moreover, if multiple analyses are run which use different starting trees and all result in the same tree, this gives more confidence that this is close to the true tree. The next step in the search strategy is the implementation of the method called "lazy subtree rearrangement". This means that all possible subtrees are cut and reinserted at all possible locations of a tree. The number of branches between the cut and insertion points must be smaller than N branches. RAxML automatically estimates the N value for a data set or the user can give the value of N. The lazy subtree rearrangement method is applied on the starting tree, and there after multiple times on the currently best tree as the program continues. When RAxML reaches the point where a better tree cannot be found, RAxML ends the search and the tree is returned [Rokas, 2011].

Pplacer

Pplacer [Matsen et al., 2010] is a tool for performing phylogenetic placement, thereby assigning query sequences to taxa and providing an evolutionary understanding of the sequence data. This tool was implemented in our 18S rDNA pipeline. Phylogenetic placement is an alternative to the classic phylogenetic analysis for dealing with datasets with a very large number of sequences. Pplacer assigns the unknown query sequences to a fixed reference tree via a reference alignment according to the maximum likelihood criterion. Since pplacer uses a fixed reference tree, there are just two tree searches needed to precalculate the information required from the reference tree. Therefore, all likelihood calculations are at once performed on the set of three taxon trees. Hence the query sequence placement part of pplacer has linear time and space complexity for the number of taxa in the reference tree.

A problem with likelihood based phylogenetic analysis is that it cannot be applied to the very large number of short reads that are produced by next-generation sequencers [Matsen et al., 2010]. This is due to computational complexity because the maximum likelihood phylogenetic problem is NP-hard (nondeterministic polynomial time) and therefore it is probably not possible to find maximum likelihood trees in a reasonable amount of time [Matsen et al., 2010], [Papadimitriou, 2014]. Also the lack of phylogenetic signal is a problem, because employing the classic method of maximum likelihood phylogeny to a single alignment of shotgun reads with the full- length reference sequences can result in the inaccurate grouping of a short read because of its position in the alignment.

Pplacer can apply the inferential strength of likelihood based methodologies, that enables the fast placement of the large amount of short query sequences and avoid some of the problems associated in applying phylogenetics to a very large number of taxa. This is accomplished as the computing complexity is greatly reduced, resulting in a program that can assign large amounts of query sequences per processor each hour on to a fixed reference phylogeny. Pplacer performs the calculation in parallel, because each query sequence can be processed independently and also the relationships between the query sequences are not investigated, thus reducing exponential time to a linear time. Short length query sequences are less of an issue for pplacer, as they are

compared to the full length of the reference sequences [Matsen et al., 2010].

Output from pplacer is a series of assignments of every query sequence to branches on the reference tree and a confidence score [Matsen et al., 2010], [Matsen et al., 2012]. A query sequence can be assigned to more than one branch to show placement uncertainty for that sequence [Matsen et al., 2012]. Pplacer computes edge uncertainty using a likelihood weight ratio, which measures the uncertainty edge by edge. It does this by comparing the best placement locations for each of the edges. Pplacer applies expected distance between placement locations (EDPL) to overcome the difficult situation in distinguishing between uncertainty of local (where placements are all located in a small area of the reference tree) and global (where possible placements are spread throughout the tree) placements. This is a problem because it depends on the confidence scores computed on an edge by edge basis.

A naive algorithm would place the query sequence onto each edge of the tree and execute a complete branch length optimization by carrying out the cached likelihood vectors. However, Pplacer achieves linear time and space scaling for the size of the reference tree because it performs an initial calculation of likelihood vectors at both ends of each edge of the reference tree. Pplacer executes a two part search algorithm for the query sequences to speed up placements. In particular, an initial quick evaluation of the tree is performed, then a more complete search is made on high scoring regions of the tree. The first part is carried out by calculating likelihood vectors for the middle of each edge. The calculated likelihood vectors are used to rapidly sort the edges in a rough order of fit for each query sequence [Matsen et al., 2010].

2.6.3 Visualisation

MEGAN

Earlier metagenomic projects were focused on the identification of species and their function from individual data sets, but over the past few years there has been a growing emphasis on comparative analysis. MEGAN [Huson et al., 2007] (MEtaGenomic ANalyzer) facilitates interactively examining, analysing and comparing multiple environmental datasets [Huson et al., 2009]. MEGAN was implemented in our 18S rDNA pipeline and the JGI 16S rDNA pipeline. MEGAN establishes its taxonomic classifi-

cation based on the NCBI taxonomy. This is a hierarchically structured classification of all species that are currently represented in NCBI databases. MEGAN performs the taxonomic analysis by assigning each read onto different taxa in the NCBI taxonomy. For determining gene content a sequence comparison of the query reads to one or more databases of determined reads is performed using a comparison tool such as BLAST. Since different projects need to use different alignment tools and databases, MEGAN gives the users the freedom to choose which ever suits their project needs. The results of the sequence comparison are processed by MEGAN. This entails the collection of all matching reads to the sequences of the NCBI database and assigning a taxon ID to each sequence based on the NCBI taxonomy.

MEGAN employs the Lowest Common Ancestor (LCA) algorithm to assign reads, which involves assigning reads to the node representing the lowest ancestor of all high-quality matches for the sequence. As a result, the species-specific sequences are assigned to taxa closer to the leaves and generally conserved sequences end up being assigned to higher order taxa that are nearer the root of the NCBI tree. A MEGAN file is generated that consists of all the information for analysing the output and to produce graphical and statistical outputs [Huson et al., 2011]. The end result is represented on a rooted tree, so that each node displays different taxa, and nodes are scaled and labelled by the number of reads that are assigned to that taxon as shown in figure 2.2 [Mitra et al., 2011].

MEGAN's functional analysis of a microbial community can help to understand the biochemical processes and to estimate the impact of environmental changes in the various ecosystems [Mitra et al., 2011]. MEGAN accomplishes this by using the SEED classification of subsystems and the Kyoto Encyclopedia for Genes and Genomes (KEGG) classification of pathways and enzymes [Huson et al., 2011]. Based on a BLAST file imported into MEGAN, a SEED classification is determined by assigning each read of the highest scoring gene in a BLAST comparison against a protein database to the functional role. Different functional roles are then grouped into assigned subsystems. A rooted tree is used to display the SEED classification, the internal nodes are the different subsystems and the leaves are the functional roles. The tree can be multi-labelled, which means that different leaves on the tree can have the same functional role, as they may occur in different subsystems. The nodes of the tree display the numbers of reads

Figure 2.2: 18S rDNA dataset displayed on MEGAN's taxonomy tree at the taxonomic rank of class

assigned to each functional role. For a comparative analysis, it is possible to map a number of datasets onto the SEED hierarchy and, also based on their SEED content, calculate the distance matrices on the datasets. For the KEGG analysis, MEGAN aims to match each read to a KEGG orthology (KO) accession number, taking the highest scoring match to a reference sequence that a KO accession number is known. A KEGG analysis window in MEGAN presents which KEGG pathways are in the dataset. The user can examine these pathways; for example, visualising reads that are mapped to a pathway of interest [Mitra et al., 2011].

2.7 Discussion

In this chapter, we have given a summary of 18S and 16S rDNA, metatranscriptomics, next-generation sequencing technology and bioinformatic tools. In the next chapter, we will describe in more detail the methodology that we developed, specifically to analyse our 18S and 16S rDNA datasets.

Chapter 3

18S rDNA and 16S rDNA analysis

3.1 Summary

This chapter outlines the computational pipeline and analysis of the 18S rDNA and 16S rDNA data collected in the expedition mentioned in chapter 1. As explained in the next section, samples were collected from a range of latitudes, from the South Atlantic Ocean, spanning the West African and European coasts up to the Arctic Ocean as shown in figure 3.1**a**. Also at the time of sampling environmental data was recorded, such as temperature and salinity as shown in figure 3.1**b-f**. This enabled us to study the composition and distribution of the marine microbial communities in the upper ocean.

Two different pipelines were implemented to taxonomically classify the 18S and 16S rDNA data. Our 18S rDNA pipeline which is outlined in section 3.3.1 is based on pplacer. We choose pplacer for our 18S rDNA pipeline because for unknown sequences likelihood-based phylogenetic inference is commonly regarded to be the most reliable classification method [Matsen et al., 2010]. While JGI's 16S rDNA pipeline which is outlined in section 3.3.2 is based on USEARCH. USEARCH is a popular software package for the analysis of operational taxonomic units (OTUs). USEARCH performs a blast like mapping against a reference database such as SILVA. This works well for microbes that are represented well in ribosomal RNA databases [Hugerth and Andersson, 2017].

In the next section, we describe the sampling and sequencing of the 18S rDNA and 16S rDNA datasets. Then, in Section 3.3 we describe the pipelines for the 18S and

16S analysis, as well as additional methods, after which we present the results of our analysis in Section 3.4. In Section 3.5, we end with a discussion of the results that we obtained.

3.2 Sampling and sequencing

3.2.1 Sampling

Samples were collected during three field expeditions across latitude ranges as shown in figure 3.1**a**. The first set of samples was collected from April to May 2011 in the

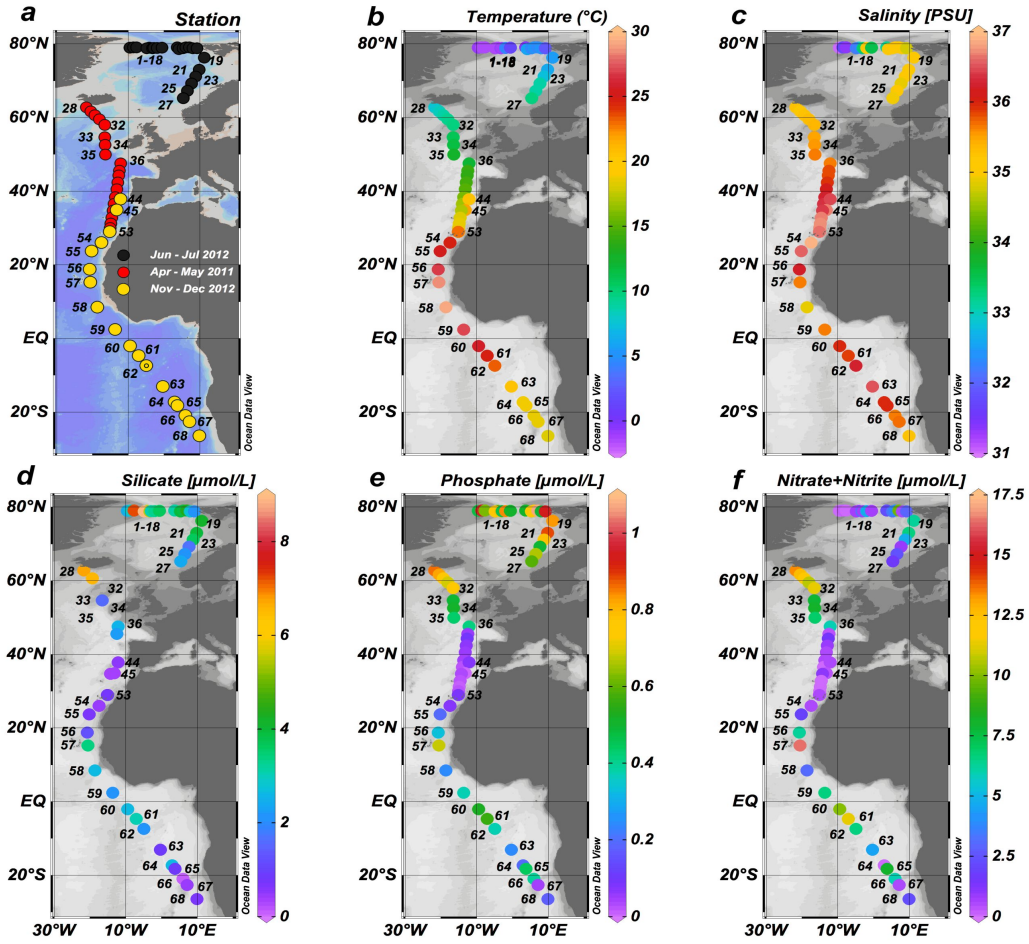

Figure 3.1: Arctic and Atlantic Ocean sampling sites and measured metadata, (**a**) Sites of three expeditions, April to May 2011 in red, June to July 2012 in black and November to December 2012 in yellow. At each station, microbial communities were sampled at the deep chlorophyll maximum (DCM), corresponding to 68 samples altogether. (**b**) Isosurface plot of temperature (°C) measured at sampling depth. (**c**) Salinity (practical salinity unit(PSU)) measured at sampling depth for all stations. (**d**) Dissolved silicate (mol/L) concentrations measured at sampling depth for each station. (**e**) Concentration of dissolved phosphate (mol/L) measured at sampling depth for each station. (**f**) Nitrate and Nitrite (mol/L) concentrations measured at sampling depth for each station. (Figure was generated with Ocean Data view, R. Schlitzer, www.odv.awi.de, 2016)(Plot generated by Dr.Katrin Schmidt)

North Atlantic Ocean spanning the Canary Islands to Iceland, by Dr.Willem van de Poll of the University of Groningen, Netherlands and Dr.Klaas Timmermans of the Royal Netherlands Institute for Sea Research. The second collection of samples was carried out from June to July 2012 by Dr.Katrin Schmidt in the Arctic Ocean, spanning the West Spitsbergen current, east Greenland current and Norwegian Atlantic current. The third set of samples was also collected by Dr.Schmidt from November to December 2012, spanning the Canary Islands down to Cape Town in the South Atlantic Ocean.

The samples were taken either at the chlorophyll maximum (10-110m) and/or surface of the ocean (0-10m). The samples were filtered and frozen in liquid nitrogen and stored at -80°C until further analysis. A full description of the materials and methods used for sampling can be found in [Schmidt, 2016]. Also at the time of sampling environmental data was collected (see Appendix A.A) as shown in figure 3.1**b-f**.

3.2.2 Sequencing and preprocessing of the 18S rDNA and 16S rDNA

All extracted DNA samples were sequenced and preprocessed by the Joint Genome Institute (JGI) (Department of Energy, Walnut Creek, CA, USA). Amplicon sequencing was performed with primers for the V4 region of the 16S and 18S rRNA gene on an Illumina MiSeq instrument with a 2x300 bp read configuration [Tremblay et al., 2015].

18S sequences were preprocessed, this consisted of scanning for contamination with the tool Duk [DOE Joint Genome Institute, 2017] and quality trimming of reads with cutadapt [Martin, 2011]. Paired end reads were merged using FLASH [Magoc and Salzberg, 2011] with a max mismatch set to 0.3 and min overlap set to 20. A total of 54 18S samples out of 68 passed quality control after sequencing. After read trimming, there was an average of 142,693 read pairs per 18S sample with an average length of 367bp and 2.8 Gb of data over all samples.

16S sequences were preprocessed, this consisted of merging the overlapping read pairs using USEARCH's merge pairs [Edgar, 2017] with the parameter minimum number of differences (merge max diff pct) set to 15.0 into unpaired consensus sequences. Any reads that could not be merged are discarded. JGI then applied the tool USEARCH's search oligodb with the parameters length mean (len mean) set to 292, length

standard deviation (len stdev) set to 20, primer trimmed max difference (primer trim max diffs) set to 3, a list of primers and length filter max difference (len filter max diffs) set to 2.5 to ensure the Polymerase Chain Reaction (PCR) primers were located with the correct direction and inside the expected spacing. Reads that did not pass this quality control step were discarded. With a max expected error rate (max exp err rate) set to 0.02, JGI evaluated the quality score of the reads and those with too many expected errors were discarded. Any identical sequence was de-replicated. These are then counted and sorted alphabetically for merging with other such files later. A total of 57 16S samples passed quality control after sequencing. There was an average 393,247 read pairs per sample and an average base length of 253bp for each sequence with a total of 5.6 Gb.

3.3 Methods

In this section, we describe the pipelines that we developed for the 18S and 16S rDNA analyses, as well as further analysis methods that were employed.

3.3.1 Computational pipeline for 18S rDNA analysis

We first describe the computational pipelines that we developed for taxonomically classifying the 18S rDNA data. Also, we describe how we account for 18S rDNA copy number variation in order to give an estimate of abundance for the species in our dataset. An overview of the 18S rDNA classification pipeline is shown in figure 3.2. This consists of the construction of the 18S reference database highlighted in box (**a**), phylogenetic analysis highlighted in box (**b**), phylogenetic placement highlighted in box (**c**) and normalising the 18S rDNA copy number in box (**d**) of the figure. Each of these is explained in the following subsections.

Constructing the 18S reference database (box a)

We began by compiling a reference dataset of 18S rRNA gene sequences that represent algae taxa for the construction of a phylogenetic tree. We retrieved sequences of algae and outgroup taxa from the SILVA database [Quast et al., 2013b] and Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) database [Keeling et al.,

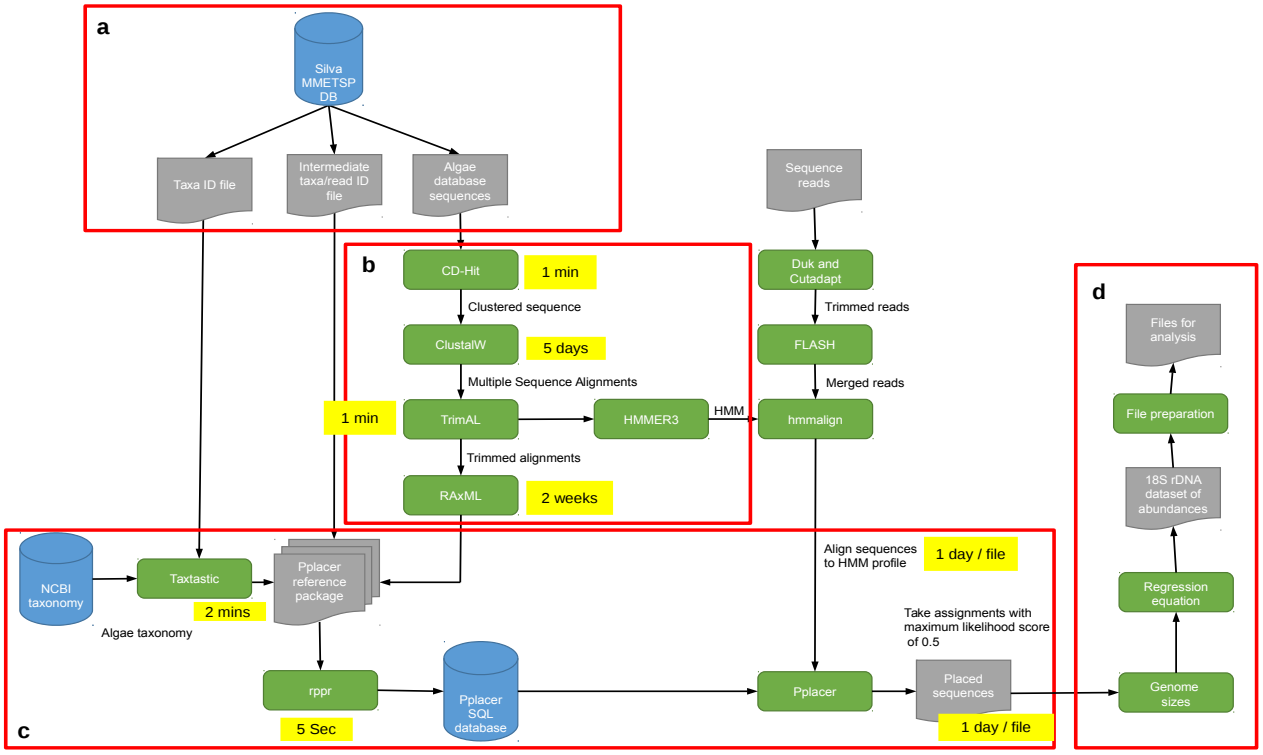

Figure 3.2: Pipeline diagram of Pplacer 18S rDNA classification analysis. The pipeline at various stages incorporates databases (blue), software tools (green), processed files (grey) and runtimes (yellow). Boxes **a**, **b**, **c** and **d** refer to sections in the text

2014]. For our outgroup species, these species came from marine invertebrates, plants, fish, zooplankton and fungi, for example, *Aurelia aurita*, *Zea mays*, *Salmo salar*, *Acartia tonsa* and *Saccharomyces cerevisiae*, respectively and also *Homo sapiens*. The inclusion of outgroups is important for the identification of potential sources of contamination that may have occurred during the earlier stages. The algae reference database consists of 1636 species from the following groups: Opisthokonta, Cryptophyta, Glaucocystophyceae, Rhizaria, Stramenopiles, Haptophyceae, Viridiplantae, Alveolata, Amoebozoa and Rhodophyta as shown in figure 3.3. A full list of the algae reference database taxa IDs can be found in Appendix A.B.

An accurate estimate of phylogeny is essential, not just for ecology but for other areas of research such as genomics. Under certain conditions, phylogenetic methods could return incorrect estimates [Wiens and Tiu, 2012]. A possible common scenario is when too few taxa are included in the reference database, thus leading to such situations as conflicting phylogenetic signals which can cause a lack of resolution and

incongruent phylogenies [Nabhan and Sarkar, 2011], [Wiens and Tiu, 2012]. Adding more taxa even with highly incomplete character data to the reference database can improve phylogenetic accuracy in scenarios where the analysis was misled by a limited number of taxa in the reference database [Wiens and Tiu, 2012].

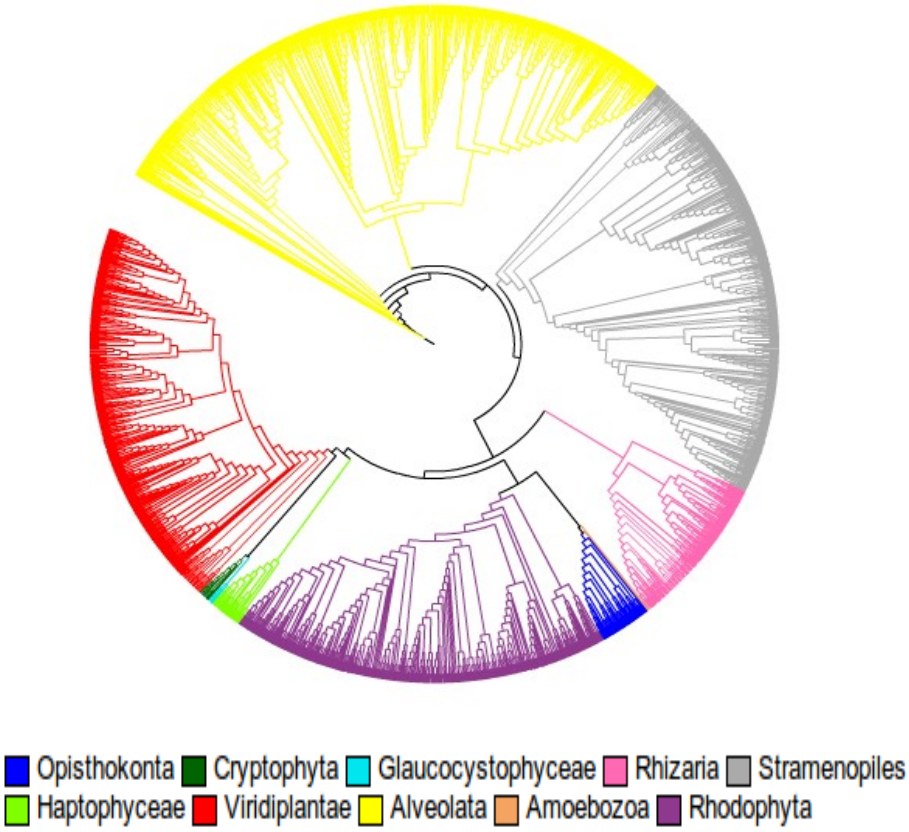

Figure 3.3: The reference tree consists of 1636 species from the groups: Opisthokonta, Cryptophyta, Glaucocystophyceae, Rhizaria, Stramenopiles, Haptophyceae, Viridiplantae, Alveolata, Amoebozoa and Rhodophyta

In order to construct the algae 18S reference database, we first retrieved all eukaryote species from the SILVA database with a sequence length of $\geq$ 1500 base pairs (bp) and converted all base letters of U to T. Under each genus, we took the first encountered species to represent that genus. Using a custom script by Dr.Andrew Toseland, the species of interest (as stated above) were selected from the SILVA databases, classified with NCBI taxa IDs and a sequence information file was produced that describes each of the algae sequences by their sequence ID and NCBI species ID. The taxonomy database from the NCBI, eukaryote sequences from the SILVA database and a list of algal taxa including outgroups were used as input for the script. This information was

combined with the MMETSP database, excluding duplications.

Phylogenetic analysis (box b)

As depicted in figure 3.2 box (**b**), we clustered the algae reference database to remove closely related sequences with CD-HIT [Li and Godzik, 2006] using a similarity threshold of 97%. Then using ClustalW [Thompson et al., 1994], we aligned the algae reference sequences of the database with the addition of the parameter iteration numbers set to 5. The alignment was examined by colour coding each species to their groups and visualising in iTOL [Letunic and Bork, 2007]. We observed that a few species were misaligning to other groups and these were then deleted using Jalview [Waterhouse et al., 2009]. The resulting alignment was tidied up with TrimAL [Capella-Gutierrez et al., 2009] by applying parameters to delete any positions in the alignment that contain gaps in 10% or more of the sequence, except if this results in less than 60% of the sequence remaining [Capella-Gutierrez et al., 2009]. Our algae reference phylogenetic tree is displayed in figure 3.3. We constructed a maximum likelihood phylogenetic reference tree and statistics file based on our algae reference alignment by employing RAxML [Stamatakis, 2014] with a general time reversible model of nucleotide substitution along with the GAMMA model of rate heterogeneity. Based on the algae reference multiple sequence alignment, with HMMER3 [Eddy, 2009] for the 18S rDNA gene we created a Profile HMM (pHMM). Our pHMM differs from other pHMMs constructed by other phylogenetic groups in that our pHMM is based on a reference database composed of an update SILVA database at the time of our analysis and we included the MMETSP database. A full description of our reference database is outlined above in section "Constructing the 18S reference database (box a)"

Phylogenetic placement (box c)

As depicted in figure 3.2 box (**c**), we used the NCBI taxtastic tool [NCBI Resource Coordinators, 2016] to create a taxtastic file, which is a description of the lineages of all species back to the root in our algae reference database. We created this taxtastic file by submitting the taxa IDs for each species to extract a subset of the NCBI taxonomy. A pplacer reference package using the NCBI taxtastic tool was generated, which produced an organized collection of all the files and taxonomic information into one directory.

With the reference package, an SQLite database was created using pplacer's Reference Package PReparer (rppr). With hmmalign, we aligned the query sequences to the reference set and created a combined Stockholm format alignment. Pplacer [Matsen et al., 2010] was used to place the query sequences on the phylogenetic reference tree by means of the reference alignment according to a maximum likelihood model. The placefiles were converted to CSV with pplacer's guppy tool. With the use of our custom made script, we took reads that had a taxonomic assignment with a maximum likelihood score of ≥ 0.5 and counted the number of reads assigned to each classification. This resulted in 6,053,291 reads that were taxonomically assigned for further analysis.

Pplacer's reference tree is fixed, in regards to the topology and branch length. Also, only two tree searches are necessary to precalculate the information that is required from the reference tree. From here all the likelihood calculations are performed on a set of three taxon trees, the number of this is linear to the number of reference taxa in the reference database. Therefore the placement part of the pplacer algorithm has linear time and space complexity for the number of taxa n in the reference tree [Matsen et al., 2010]. In figure 3.2 are the runtimes highlighted in the yellow boxes for the tools as described above that were implemented in our 18S rDNA pipeline. It took a day for an 18S rDNA file containing an average of 142,693 read pairs with an average length of 367bp to run with our pplacer (18S rDNA) pipeline. We were able to perform our pplacer pipeline even faster due to parallel computing on the Earlham Institute cluster thus enabling us to run all 54 files within 3 days. If another similar dataset was to be submitted to our pplacer pipeline, this new dataset would also finish in this time. If the number of taxa in our reference database was to be altered then this would affect the time and space complexity.

Ribosomal RNA (rRNA) gene copy number

The rRNA gene is a marker for taxonomic diversity but the relationship between amplicon and species abundance is indeterminate due to rDNA copy number variation within the genomes of different species. For bacteria, the 16S rDNA copy number can vary as much as from one to fifteen [Perisin et al., 2016]. For eukaryotes, the 18S rDNA copy number can vary even more greatly from one to thousands [de Vargas et al., 2015]. This is a hindrance to effectively analyse our rDNA copy number datasets. In order to

get an estimate of abundances of species in the samples, we had to normalise the data, that is we had to adjust for the rDNA copy number.

Even though there is a 16S rDNA copy number database called the Ribosomal RNA Operon Copy Number Database (rrnDB) [Klappenbach et al., 2001] it only contained 2,876 species in 2014. While this is a considerable amount, it compares little to the actual number of bacteria species in the world. This was evident to us when we attempted to apply the rrnDB database to our 16S rDNA dataset. In order to fill in the missing copy numbers in our 16S rDNA dataset, we used averages between closely related known taxa. We detail this in the section "Normalising 16S rDNA copy number".

There is no database of 18S rDNA copy number available for eukaryote species. Previous work has explored the link between rDNA copy number and genome size [Prokopowich et al., 2003]. We decided to build on this approach in order to get a rough estimate of read count to 18S rDNA copy number. The regression line is an imperfect approach, due to rDNA copy number variation within species that can exist thus being a source of error. In the following sections, we detail our work in determining a regression equation based on 18S rDNA copy and genome size. We also explain how we retrieved where possible the genome sizes for our 18S rDNA dataset, and determined the genome sizes when they were unknown, in order to apply the regression equation to our 18S rDNA dataset.

18S rDNA copy number and genome size regression equation (box d)

In order to address the varying copy number among eukaryotes, we investigated the gene copy number and related genome sizes of 185 species across the eukaryote microbial tree [Godhe et al., 2008], [Carlton et al., 2013], [Torres-Machorro et al., 2010], [Oliver et al., 2007], [Moreau et al., 2012], [Boucher et al., 1991], [Hauser et al., 2010], [Prokopowich et al., 2003], [Rödström, 2017], [NCBI Resource Coordinators, 2016] and [Nordberg et al., 2014]. Species with single genomes were investigated for their 18S rDNA copy number. Outliers were found but all 185 species were included, as we researched as many published 18S rDNA copy numbers as we could find with known genome size, but published 18S rDNA copy numbers are very few especially in comparison to 16S rDNA copy numbers and with known genome size.

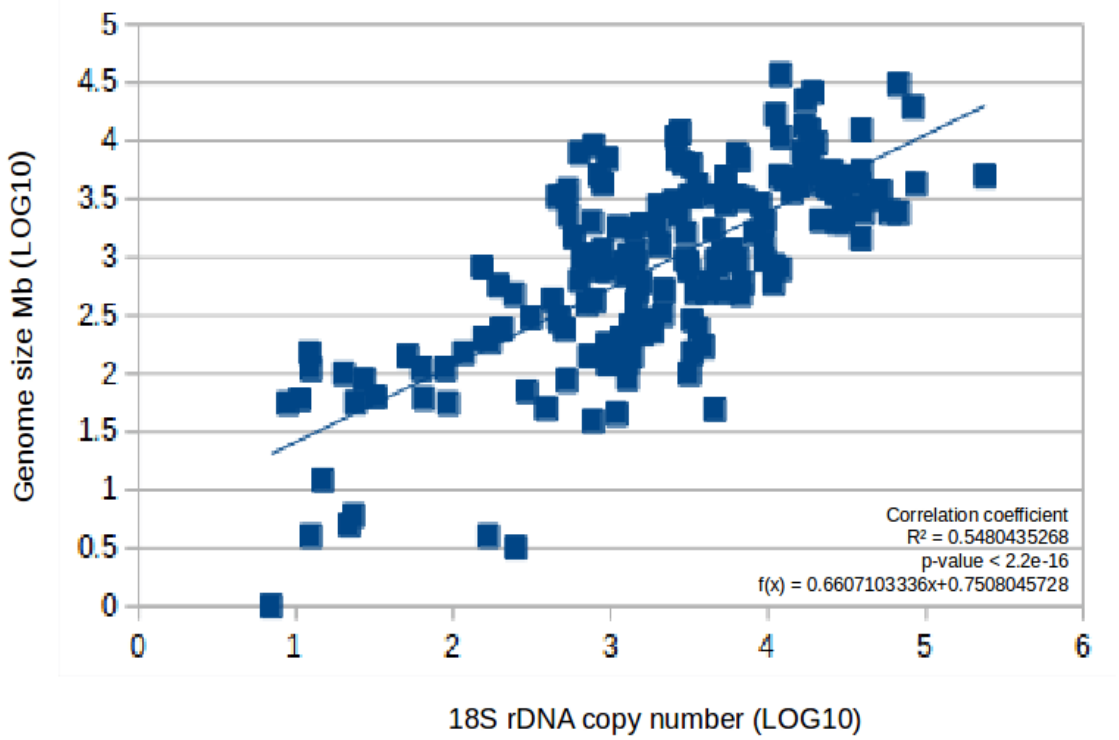

Figure 3.4: The graph of 18S rDNA copy number and their related genome size (Mb) for 185 species across the eukaryote tree. We investigated 18S rDNA gene copy number and their related genome sizes. We observed a significant correlation of R^2 0.5480435268 with a p-value < 2.2e-16 between genome size and 18S rDNA copy number. Based on the log10 transformed data a regression equation was determined, f(x)=0.66X+0.75

We performed a log10 transform of the data of 185 18S rDNA copy number and related genome sizes so that the data's scale and distribution complied better to the assumptions of a parametric statistical test. Log transformation ($x'_i = log_b(x_i + c)$ where log_b can be 2, e, or 10 and c is a small number when $x_i = 0$) is commonly used [Paliy and Shankar, 2016]. Based on the log10 transformed data, a significant correlation with an R^2 of 0.55 and a p-value < 2.2e-16 between genome size and 18S copy number was observed. A regression equation was determined (f(x)=0.66X+0.75) as shown in figure 3.4.

Genome sizes for the 18S rDNA dataset to normalise the 18S rDNA copy number (box d)

In order to apply the equation of the line to our 18S rDNA dataset, we retrieved the genome sizes for the species in the dataset from the NCBI genome database. The NCBI genome database consisted at the time of 2,477 eukaryote entries. Firstly, since multiple entries of a species are in the NCBI genomes database due to different strains,

we calculated an average genome size for each species in the database, which resulted in 2,059 species entries.

The higher taxonomic levels for the NCBI genomes species needed to be established so that we could calculate the average of genome sizes. For a description of the lineages of all species back to the root in NCBI genomes database, we submitted the species names for each entry to extract a subset of the NCBI taxonomy with the NCBI taxtastic tool, thus producing a taxtastic file. The taxtastic file based on species from the NCBI genomes database was used to calculate the average genome sizes for higher taxonomic levels from the known genome size species level, with the assistance of the parent id and taxa id layout in the taxtastic file.

Using the taxtastic file based on our algae reference database, we assigned our algae entries a genome size from species to root from the prepared average genome sizes NCBI genomes taxtastic file. Not all genome sizes in the algae reference database were known. We, therefore, took the average of closely related species from the above taxonomic level of those we could get and took that as the genome size for those that were missing in our dataset.

As depicted in figure 3.2 box (**d**) the 18S rDNA dataset was normalised by assigning their genome sizes and applying the equation of the line. A normalisation procedure called the hits per million reads method was applied to the files, which entails scaling the files to a common value [Robinson and Oshlack, 2010].

File preparation for 18S rDNA analysis (box d)

In our 18S rDNA dataset, we had a total abundance of 53,750,176 taxa from the eukaryote node down to the species nodes. We employed MEGAN to cut out specific taxonomic levels. In MEGAN, we extracted the classifications at the taxonomic rank of species. This consisted of a file being generated for each station that contained the species names and their assigned abundances. The files were further normalised to hits per million.

In MEGAN, we extracted the leaves of the taxonomy tree at the rank of class and above but excluded assignments to the eukaryote node. Firstly, this consisted of a file being generated for each station that contained all assignments to the class nodes as well as any assignments under their respective lineages down to species being summed

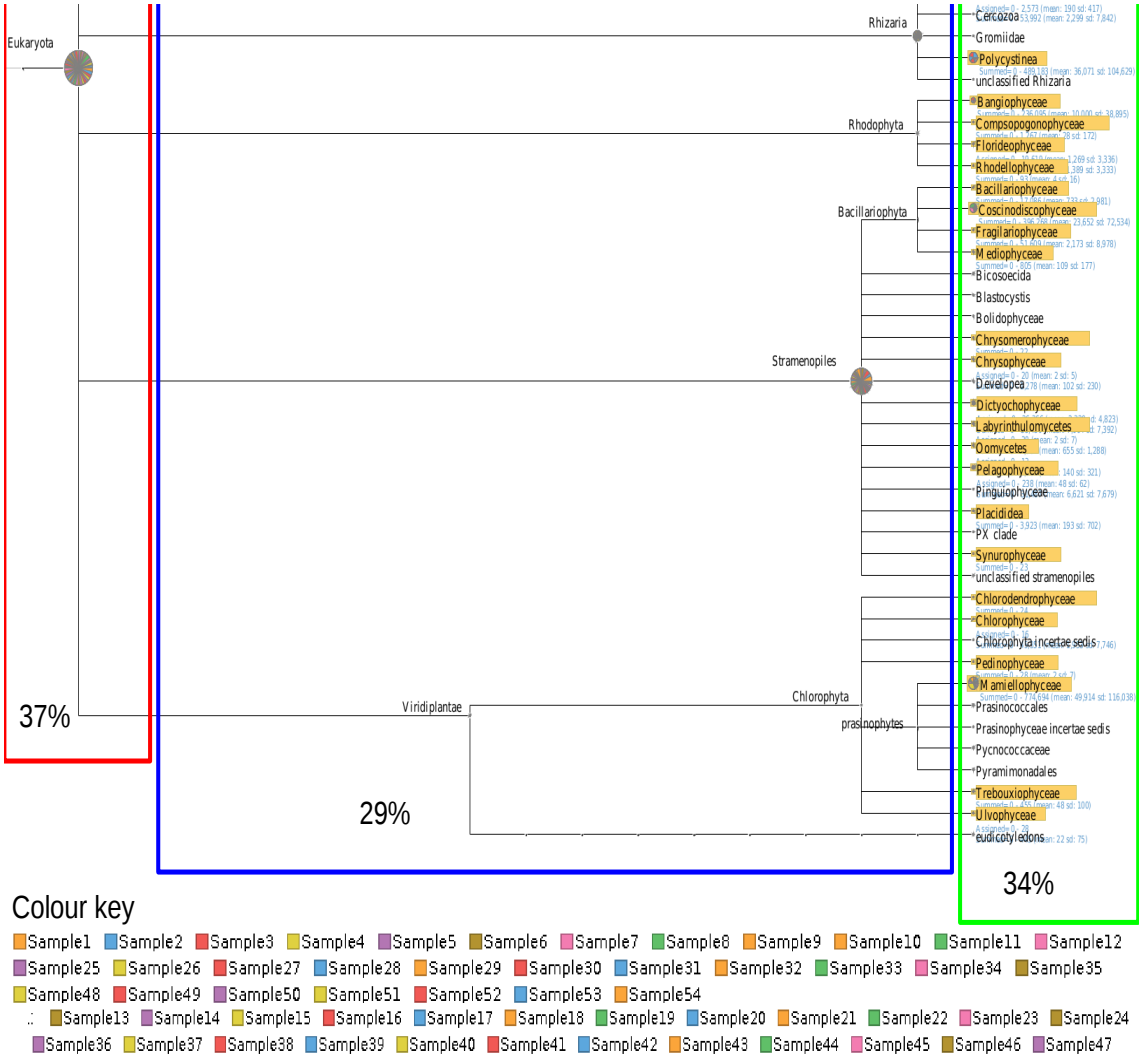

Figure 3.5: Part of the 18S rDNA dataset displayed on MEGAN's taxonomy tree. All the nodes highlighted in yellow at the leaves of the tree are class nodes. Nodes at the leaves of the tree that are not highlighted do not have a taxonomic classification of class in their lineage. The nodes between the leaves of the tree and the eukaryotic node are the internal nodes that are representing higher taxonomic levels such as phylum. The colour key represent each colour on the nodes and corresponds to a sample in the 18S rDNA dataset

up under the individual class node. These are displayed in figure 3.5 as the highlighted nodes on the leaves of the tree.

Secondly, we included nodes that were not highlighted on the leaves of the tree, as displayed in figure 3.5. In NCBI taxonomy there are species that do not have a taxonomy designation at every taxonomy level. In our 18S rDNA dataset, we had species that do not have a taxonomic rank of class and these are displayed in figure 3.5 as the leaves of the tree that are not highlighted. We took the nodes that were not highlighted on leaves of the tree and summed them together within their respective lineages and placed them under a new name. For example, under the phylum Rhizaria, on the leaves of the tree, there is Cercozoa, Gromiidae and unclassified Rhizaria which

are not highlighted. Their abundance was summed together and renamed Nc.Rhizaria, "Nc." standing for "No class". The abundances assigned to Rhizaria were not included in this calculation. The leaves of the tree as displayed in figure 3.5, made up 34% of the total 18S rDNA dataset, as it resulted in an abundance of 18,332,601.

The internal nodes between the leaves of the tree and the eukaryote node as displayed in figure 3.5 was given a "U." in front of their name, "U." standing for "Unknown". This was done to highlight that while they are of course associated with the lower lineages they are in fact considered separate, as those assignments to those nodes could not be determined any lower. The internal nodes made up 29% of the total 18S rDNA dataset, as it resulted in an abundance of 15,678,138.

The abundance assigned to the eukaryote node was excluded from our analysis as these sequences could not be classified lower. This comprised of a total abundance of 19,739,437, which is 37% of the 18S rDNA dataset. The assignments to the eukaryote node were excluded as these reads could not be classified to lower taxonomic ranks. Our algae reference database contains an up to date SILVA database at the time of our analysis, but still 37% of the 18S rDNA dataset is essentially unclassifiable. We have lost valuable information that could have potentially provided more insight and understanding into our analysis.

A file was generated for each station that contained the class nodes, "Nc." nodes and "U." nodes with their respective abundances. The files were further normalised to hits per million.

3.3.2 Computational pipeline for 16S rDNA analysis

In this section, we describe JGI's computational pipeline for taxonomically classifying the 16S rDNA dataset. Also, we describe how we account for 16S rDNA copy number variation in order to give an estimate of abundance for the species in our dataset. An overview of JGI's 16S rDNA classification pipeline is shown in figure 3.6 and a description of the pipeline is below.

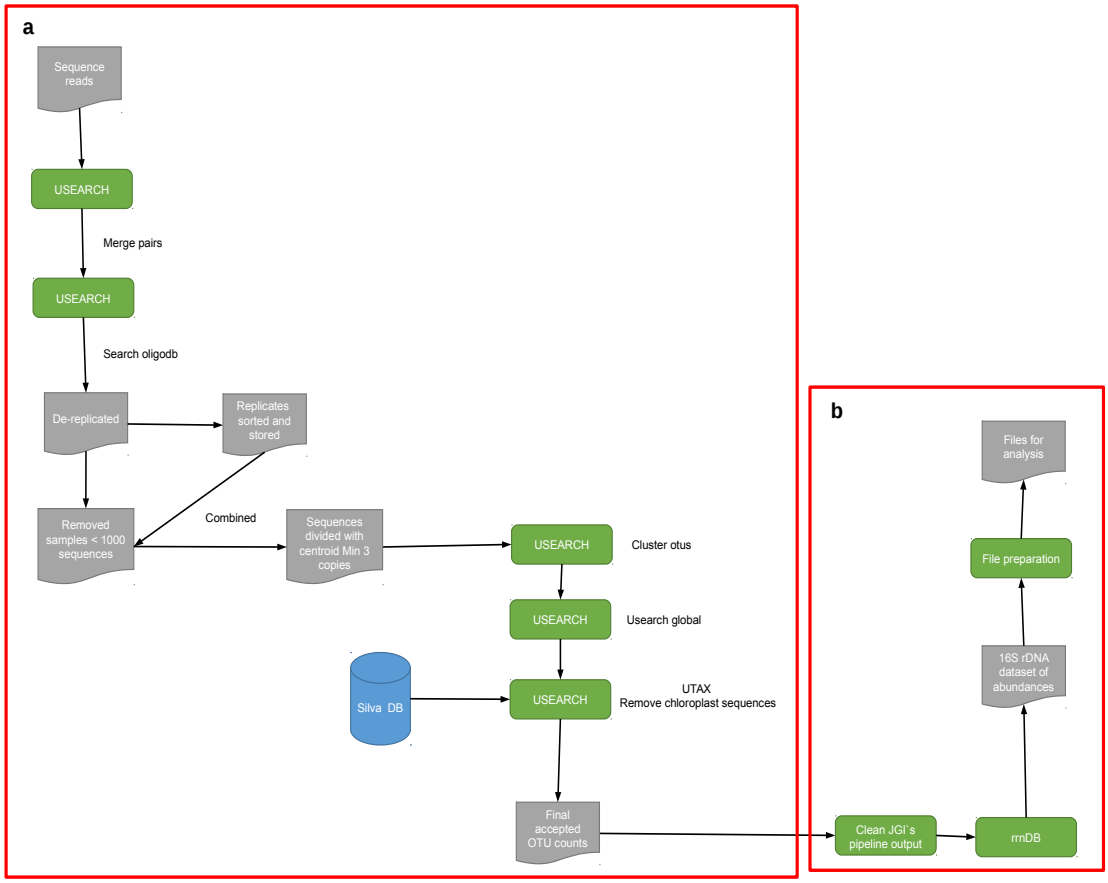

Figure 3.6: JGI's pipeline diagram of 16S rDNA classification analysis. The pipeline at various stages incorporates databases (blue), software tools (green) and processed files (grey). Box **a** and **b** refer to sections in the text

JGI's computational pipeline for taxonomical classifying the 16S rDNA dataset (box a)

The JGI's 16S rDNA classification pipeline consists of firstly removing samples with less than 1000 sequences. The remaining samples and the de-replicated identical sequences from the preprocessing step (as outlined in Section 3.2.2) are then combined and their sequences organized by decreasing abundance. The sequences are divided out based on the criterion as to whether they contained a cluster centroid with a minimum size of at least 3 copies. The low abundance sequences are put aside and not used for clustering. USEARCH's cluster otus command is employed to incrementally cluster the clusterable sequences. This begins at 99% identity and the radius is increased by 1% for each iteration until a OTU clustering identity of 97% is reached.

At each step, the sequences are sorted by decreasing abundance. Once clustering is complete, USEARCH's usearch global is used to map the low-abundance sequences to the cluster centroids. These are added to OTU counts if they were in the prescribed

percent identity threshold. If they do not fall within this prescribed percent identity threshold they are discarded. USEARCH's UTAX along with the SILVA database is used to evaluate the clustered centroid sequences. The predicted taxonomic classifications are then filtered with a cutoff of 0.5. Any chloroplast sequences identified are removed. The final accepted OTUs and read counts for each sample are finally placed in a taxonomic classification file.

Cleaning the JGI's pipeline output (box b)

The JGI output taxonomic classification file contained a matrix of the 57 station's names to their counts of read assignments. Also given were the taxonomic assignments in the form of the SILVA taxonomic path from domain to the taxonomic level it was assigned. Our SILVA taxonomic assignments had a number of issues that needed resolving before we could account for the 16S rDNA copy number variation in our datasets. SILVA contains entries without a formal taxonomic name and these are entered in SILVA's database as for example "unknown". When we had such an assignment we moved the assignment up a taxonomic level until a taxonomic assignment with a "real" taxonomic name was given.

In addition, SILVA taxonomic names are formatted slightly differently from NCBI entries, for example, some entries have numbers attached. We identified such entries in our dataset and edited the names. SILVA taxonomic names are also not updated with NCBI entries. We identified such entries in our dataset and edited the names to ensure they were up to date with the NCBI taxonomy. We created a custom made script to accomplish this, to tidy up SILVA taxonomic names of any format issues, to have "real" up to date taxonomic names for each assignment and to have a single taxonomic name rather than a taxonomic path for each assignment. We combined any duplicated taxonomic assignments for each station.

Normalising 16S rDNA copy number (box b)

In order to normalise the 16S copy number, the 16S copy numbers for the species in the dataset were retrieved from the Ribosomal RNA Operon Copy Number Database (rrnDB) [Klappenbach et al., 2001]. The rrnDB database consisted at the time of 3,021 bacterial entries. Firstly, since multiple entries of a species are in the rrnDB

database due to the presence of different strains, we obtained an average copy number for each species in the rrnDB database, which resulted in 2,876 species entries. The higher taxonomic levels for the rrnDB species needed to be established so that we could calculate their average copy number. For a description of the lineages of all species back to the root in the rrnDB database, we submitted the species names for each entry to extract a subset of the NCBI taxonomy with the NCBI taxtastic tool [NCBI Resource Coordinators, 2016], thus producing a Taxtastic file. The Taxtastic file based on species from the rrnDB database was used to calculate the average copy number for higher taxonomic levels from the known copy number species level, with the assistance of the parent id and taxa id layout in the Taxtastic file. A Taxtastic file based on 16S rDNA species from our dataset was generated and we assigned our 16S species entries a copy number from species to root from the prepared average copy number rrnDB Taxtastic file. Not all copy numbers in the 16S rDNA dataset were known. We therefore took the average of closely related species from the above taxonomic level of those we could get and took that as the copy number for those that were missing in our dataset. The 16S dataset was normalised by dividing by the assigned copy number. The files were normalised to hits per million.

File preparation for 16S rDNA analysis (box b)

In our 16S rDNA dataset, we had a total abundance of 56,999,957 taxonomic assignments to nodes from the bacteria node down to the genus nodes. We prepared the 16S rDNA taxonomic levels in the same manner as the 18S rDNA taxonomic levels as outlined in Section 3.3.1.

We extracted the leaves of the tree that include class nodes and "Nc." nodes with their respective abundances. This step resulted in an abundance of 53,723,979 (94%). Also, we extracted the internal nodes and placed "U." in front of their names. This resulted in an abundance of 1,627,260 (3%). The abundance assigned to the bacteria node was excluded from our analysis and this comprised of a total of 1,648,718 (3%). We generated a file for each station that contained the class nodes, "Nc." nodes and "U." nodes with their respective abundances. The files were further normalised by applying the hits per million reads method.

We extracted the classifications at the taxonomic rank of genus. This consisted of a

file being generated for each station that contained the genus names and their assigned abundances. The files were further normalised by applying the hits per million reads method.

3.3.3 Further analysis methods

In this section, we describe the methodology that we used for our analysis of the 18S and 16S rDNA datasets.

Evenness and occupancy

Alpha diversity determines the diversity of individuals in local communities [Marcon et al., 2014]. Indices are used to describe the general properties of a community and then this enables us to compare different regions or taxa [Morris et al., 2014]. There are a number of alpha diversity indices available, such as the Shannon diversity ($H' = - \sum P_i \ln(P_i)$, where P_i is the proportion of individuals belonging to species i) [MacArthur and MacArthur, 1961] which is sensitive to the different number of species present in a community [Morris et al., 2014], [Johnson and Burnet, 2016]. Taxonomy evenness is the similarity in relative abundance across the sample locations [Zhang et al., 2012]. An evenness value ranges between 1 and 0, with an evenness value of 1 corresponding to complete evenness and 0 no evenness. The occupancy refers to how many sample sites that species occurs in.

We produced an abundance, species evenness and occupancy plot for each 18S rDNA class level (n=54) and 16S rDNA class level (n=57). The x-axis represents the number of times that taxon occurs across the stations. The y-axis represents the evenness of that taxon across stations it occurs in. Each circle represents a taxon abundance. The size of each circle is resized by replacing the area of the circle which represented the total abundance for that taxon with the square root of the abundance divided by π. The evenness and occupancy plot was calculated using a Dispersion index, which is a variant of Pielou's evenness [Pielou, 1966], and based on Shannon diversity index [Payne et al., 2005].

There are multiple indices available to quantify biodiversity, but there is no consensus regarding which is more appropriate and informative [Morris et al., 2014]. For our

analysis, we choose the Shannon diversity index which is equally sensitive to rare and abundant species [Morris et al., 2014]. We also choose the Dispersion index which is also sensitive to samples containing rare species [Payne et al., 2005]. It is important for us to consider both abundant and rare species in our analysis. Rare species are defined as those with very low abundance and are not present in every sample [Chapman et al., 2018]. Our 18S and 16S rDNA datasets contain a considerable portion of rare species, ∼42% and ∼40% respectively.

The presence and abundances of some species in a particular location may not be independent of some other species [Schluter, 1984]. In this circumstance, there would be a potential need to adjust the Shannon index to account for non-independent species. We would then combine the non-independent species into modules instead of considering them separately. We would then apply the combined presence and abundance in the Shannon diversity index.

Breakpoint analysis

We performed a breakpoint analysis based on the methodology from [Castro-Insua et al., 2016]. This approach plots beta diversity against temperature. Beta diversity is a measure of the amount of variation in species composition for a community among samples [Ricotta, 2017]. Beta diversity enables us to compare and contrast communities. There are a number of different indices for beta diversity. The beta diversity indices that we use in our breakpoint analysis is the Sørensen indices ($\beta_{sor} = \frac{b+c}{2a+b+c} = \frac{b}{b+a} + (\frac{c-b}{2a+b+c})\,(\frac{a}{b+a})$), where "a" is the number of species two sites have in common, "b" is the number of unique species in the poorest site and "c" is the number of unique species in the richest site [Baselga and Orme, 2012].

The breakpoint analysis enabled us to search for temperature breakpoints in the 18S and 16S rDNA datasets. This analysis provided insight about changes in the biodiversity of the identified prokaryotic and eukaryotic species across the different temperatures of the polar region in the Arctic Ocean, through the temperate region in the North Atlantic Ocean and into the tropical region in the South Atlantic Ocean.

A breakpoint was determined and plotted for each of the 18S rDNA class level (n=54) and 16S rRNA class level (n=57) datasets. This was calculated by firstly producing a presence absence matrix for each dataset. A multiple-site dissimilarity was

performed on the presence absence matrix with beta.pair, a function from the betapart R package and a dissimilarity index set to sorensen. These values were then plotted against temperature, to enable us to get a range of values in which the breakpoint might be located. We then searched through these possible breakpoints for the one with the lowest mean squared error. In each of the 18S rDNA and 16S rDNA datasets plots the y-axis represents the beta diversity and the x-axis represents temperature with piecewise regression lines and breakpoints shown.

Co-occurrence analysis

We also undertook a co-occurrence analysis with weighted Gene Co-Expression Network Analysis (WGCNA) [Langfelder et al., 2008], a method for finding modules (a WGCNA term used to describe networks) of highly correlated individuals. The prokaryotes at the taxonomic rank of genus and eukaryotes at the taxonomic rank of species normalised files were combined for each station (n=50). Using the R package WGCNA on samples of combined prokaryotes and eukaryotes we obtained modules derived from their log10-scaled abundances.

A network can be described by its adjacency matrix. The adjacency matrix (a_{ij}) is calculated by first defining a co-expression similarity ($s_{ij} = |\mathrm{cor}(x_i, x_j)|$, where cor is correlation, and x_i and x_j are gene (species) expression profiles, consisting of expression of genes (species) i and j across a number of samples). We performed a signed co-expression measure to keep track of the sign of positively correlated genes (species) of the co-expression measure. Also, we transformed the co-expression similarity into a weighted adjacency (an adjacency that keeps track of the correlation values determined between genes (species)), as we wanted the adjacency to keep track of the connection strength (correlation coefficient value) between the genes (species) [Langfelder et al., 2008].

To construct a signed weighted network we first determined a soft threshold which is a power called beta ($\beta \geq 1$) [Langfelder et al., 2008]. The soft threshold was determined by using the pickSoftThreshold function, which is part of the WGCNA package, in R [Langfelder et al., 2008]. We use the power to raise the correlation calculation ($a_{ij} = |(1+\mathrm{cor}(x_i, x_j))/2|\ )^{\beta}$) so the network is constructed with the emphasis on high correlations at the expense of low correlations, therefore the network is more

robust [Zhang and Horvath, 2005], [Langfelder et al., 2008]. We choose the lowest power that results in approximate scale free topology as measured by the scale free topology fitting index [Zhang and Horvath, 2005]. For our data, we got a power of 11. Therefore a signed weighted adjacency measure for each pair of species was determined by raising the absolute value of the correlation coefficient to the power of 11. A topological overlap measure (TOM) was calculated from the resulting adjacency matrix. Hierarchical clustering was carried out on the TOM measure. This resulted in two modules being found ((a) n=70 and (b) n=51).

Highly correlated modules are not distinct, therefore using the modules' eigengene, which is the first principal component of a module, to be a representative of that module, in order to further examine the modules, to merge highly correlated ($\geq$0.75) modules based on their eigengene [Langfelder and Horvath, 2007]. This did not result in the two modules being merged and therefore the two modules were taken for analysis. When incorporating environmental data, latitude values were redefined, so that the North pole is 0°, the Equator is 90° and the South pole is 180°.

3.4 Results

3.4.1 Rarefaction curves

Samples taken from a population should be a representative of that population, but we are unable to visually confirm the organisms we are attempting to capture. Rarefaction curves are a standard tool for analysis, in order to generate a rarefaction curve, the number of species is plotted as a function of the number of individuals sampled. Rarefaction curves are a means of determining if sequencing depth is sufficient [Wooley et al., 2010]. Also with rarefaction curves, we can compare the diversity among the samples [Hughes et al., 2001]. Initially, if the curve begins with a steep slope and then at some point, the curve begins to level off this indicates that fewer species are being discovered in the sample. If the slope increases more gently, this indicates less contribution of the sampling to the total number of species. Interpreting a comparison of the diversity among the samples in the rarefaction curves plot can be achieved by taking the highest shared sample size (x-axis) between the curves and examining the

curves position to each others in relation to the number of leaves in taxonomy (y-axis), the curves that are highest along the y-axis are more diverse [Wooley et al., 2010].

The rarefaction curves in figure 3.7**a** and 3.7**b** were generated using MEGAN and are based on the taxonomic classification of 18S and 16S rDNA species and genus level, respectively. In each rarefaction curve, the individual curves represent a single sample, as we ran rarefaction curves on each sample and then the curves are plotted together per dataset. The samples are coloured by region, which corresponds to the map of the region sampling sites in figure 3.1**a**, where the polar Arctic Ocean samples are coloured black, the temperate North Atlantic Ocean samples are coloured red and the tropical South Atlantic Ocean samples are coloured yellow.

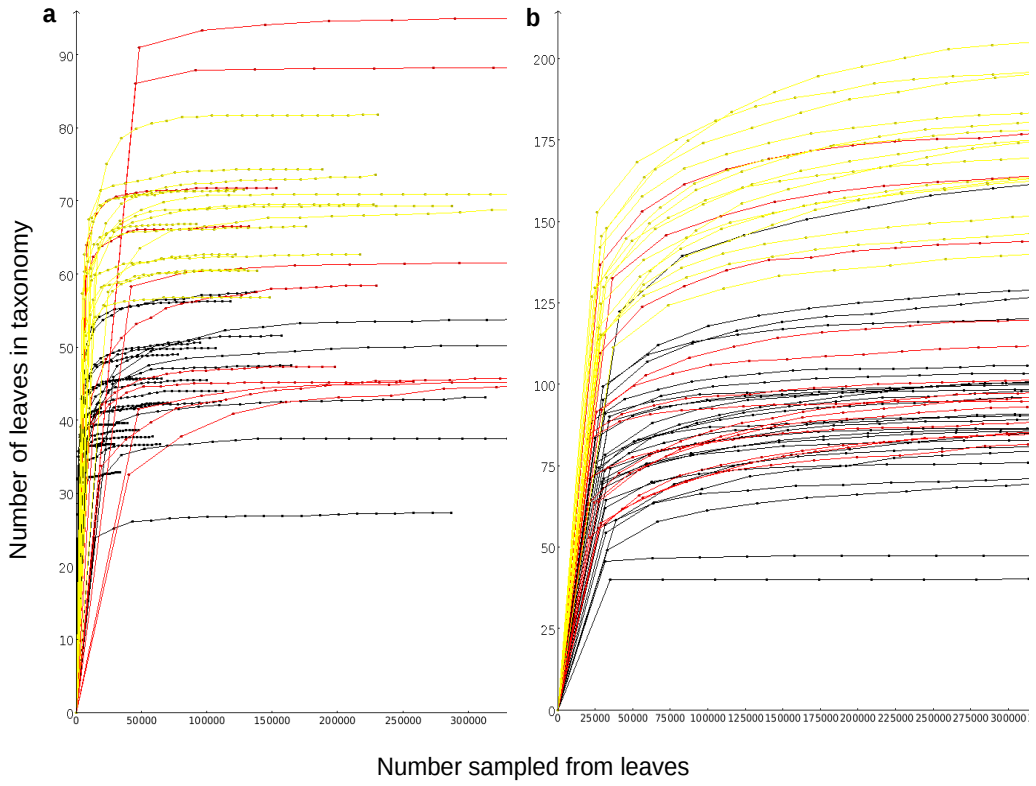

Figure 3.7: (**a**) rarefaction curves for 18S rDNA species level (n=54) and (**b**) the rarefaction curves for 16S rDNA genus level (n=57). In each panel the colours correspond to the sample sites of the three expeditions, April to May 2011 in red (North Atlantic Ocean), June to July 2012 in black (Arctic Ocean) and November to December 2012 in yellow (South Atlantic Ocean). The rarefaction curves were generated using MEGAN

As seen in figure 3.7**a** for the 18S rDNA species, there is a sharp rise at first in the curves of the samples from the Arctic Ocean coloured in black and the curves of the samples from the South Atlantic Ocean coloured in yellow. The curves of the samples from the North Atlantic Ocean coloured in red rise slightly less sharply at

first compared to the other samples from the Arctic Ocean and South Atlantic Ocean. Then all the curves begin to rise more slowly as rarer species are added and then the curves level off. The levelling off of all of the curves happens quite quickly, therefore we conclude that sufficient sampling was achieved during the three expeditions.

In figure 3.7**a** for the 18S rDNA species, the curves for the samples from the South Atlantic Ocean coloured in yellow contain the greatest amount of diversity. The curves for the samples from the Arctic Ocean and North Atlantic Ocean coloured in black and red, respectively, contain in general about an equal amount of diversity, with the exception of several red samples which have a higher amount of diversity.

For the 16S rDNA species as shown in figure 3.7**b** there is an even steeper rise at first in the curves of the samples from all three expeditions. Then all curves begin to rise more slowly as rarer species are added before the curves level off. The levelling off of the curves happens quite quickly. Therefore we conclude that we adequately sampled during the three expeditions.

In figure 3.7**b** for the 16S rDNA species, we see a similar pattern of diversity among the samples as we did in 3.7**a** for the 18S rDNA species. The curves for the samples from the South Atlantic Ocean coloured in yellow contain the greatest amount of diversity. The curves for the samples from the Arctic Ocean and North Atlantic Ocean coloured in black and red, respectively, contain in general about an equal amount of diversity, with the exception of a few red samples which have a higher amount of diversity.

3.4.2 Principal Coordinates Analysis

A clustering analysis was performed in MEGAN for each of our 18S and 16S rDNA datasets. Principal Coordinates Analysis (PCoA) is a multidimensional scaling technique which enables us to compare a large number of samples at once. PCoA attempts to order the objects across the axes of principal coordinates. PCoA accomplishes this by employing a linear transformation of the distance or dissimilarities between the objects onto the plot, while also seeking to explain as much of the variance in the initial dataset [Ramette, 2007], [Paliy and Shankar, 2016]. PCoA summarises the community compositional differences between the samples, the principal coordinates from a PCoA are plotted against each other and each point in the plot represents a sample. The

relative positioning of the point represents the relationships among the variables measured in the samples [Goodrich et al., 2014], [Paliy and Shankar, 2016]. The distance of the groups from one another cannot be quantified with a high degree of reliability as a reflection of the dataset. This is due to the fact that PCoA summarises the dataset in a two dimensional scatterplot, the two principal coordinates used in the plot only explain a fraction of the variability in the dataset [Goodrich et al., 2014]. The groups within the PCoA can be determined by using betadisper, a function from the vegan R package, which calculates how compact the objects are by calculating the average distance of group members to the centroid of that group [Simpson, 2006].

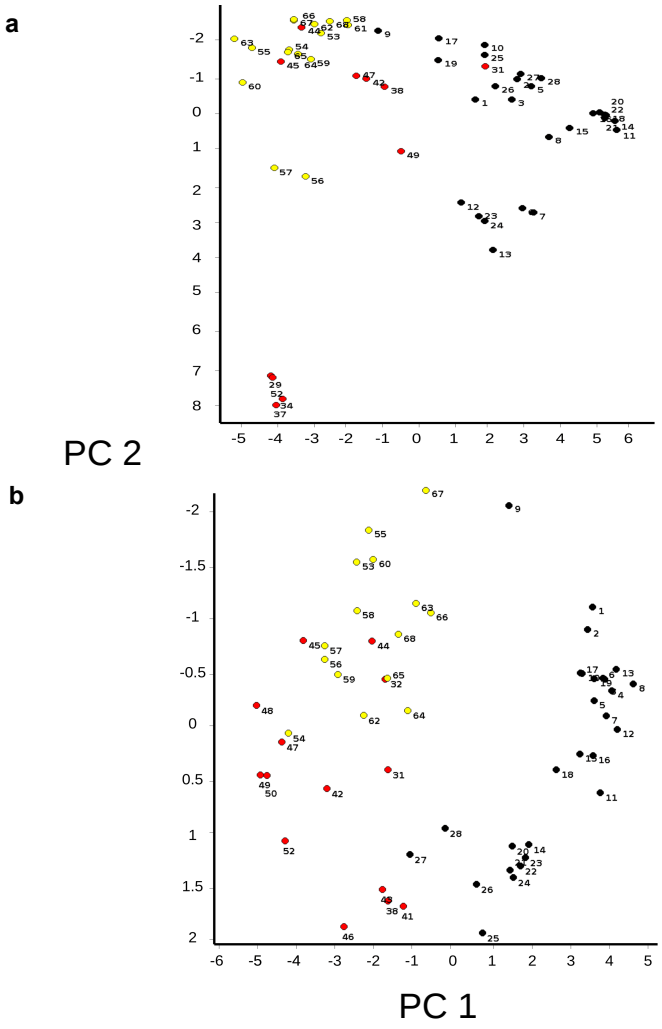

Figure 3.8: Principal coordinates analysis (PCoA) of **a** represents the eukaryotic communities (n=54) and **b**, represents the prokaryotic communities (n=57) at the taxonomic rank of class. In MEGAN, communities are clustered according to their similarity based on Bray-Curtis distances. The samples are numbered and coloured by region; these numbers and colours correspond to the map of the region sampling sites in figure 3.1**a**, where the Arctic Ocean samples are coloured black, the North Atlantic Ocean samples are coloured red and the South Atlantic Ocean samples are coloured yellow

Any distance or dissimilarity matrix can be used in a PCoA [Paliy and Shankar, 2016]. Multivariate analyses of ecological data are often based on a dissimilarity matrix, such as Bray Curtis [Anderson and Santana-Garcon, 2015]. Bray Curtis accounts for both species presence and abundances at each site [Pos et al., 2014]. A distance matrix based on Bray Curtis ($\Sigma \mid x_i - x_j \mid / \Sigma (x_i + x_j)$, where x are the counts of species in samples i and j) was used to calculate the distance between samples to be plotted [Ricotta and Podani, 2017].

Displayed in figure 3.8**a** is the 18S rDNA (n=54) dataset at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the eukaryote node. Displayed in figure 3.8**b** is the 16S rDNA (n=57) dataset at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the prokaryotes node. In each PCoA plot the samples are numbered and coloured by region; these numbers and colours correspond to the map of the region sampling sites in figure 3.1**a**, where the Arctic Ocean samples are coloured black, the North Atlantic Ocean samples are coloured red and the South Atlantic Ocean samples are coloured yellow.

The 18S rDNA are presented in figure 3.8**a** with PC1 accounting for 24% of sample variation, while PC2 accounts for 19.9% of sample variation. Overall the 18S rDNA samples are to a reasonable extent clustering well by region, as the matching colours are grouped together. Also for the 18S rDNA samples, there is a transition of the samples from black to red to yellow, which coincides with how the samples are positioned by latitude as can be seen in figure 3.1**a**. There are 4 samples that are outliers, these are samples 29, 52, 34 and 37 which are shown in figure 3.8**a** at the bottom left side of the plot. These samples came from the North Atlantic Ocean expedition as explained in section 3.2.1. Dr.Willem van de Poll and Dr.Klaas Timmermans noted that at the time of sampling a bloom may have been occurring, which could explain why these samples are clustering differently since their composition and abundance is remarkably different from the other samples in that region of the North Atlantic Ocean.

The 16S rDNA are shown in figure 3.8**b** with PC1 accounting for 51.6% of sample variation, while PC2 accounts for 26.3% of sample variation. The 16S rDNA samples are also to a reasonable extent clustering well by region, as the matching colours are grouping together. The 16S rDNA samples transition from black to red to yellow in

the shape of a horseshoe, this shape is indicative of an underlying linear gradient. Also this transition of colour coincides with how the samples are positioned by latitude as can be seen in figure 3.1**a**.

3.4.3 Heatmaps

Figure 3.9**a** and 3.9**b** display heatmaps arranged by latitude for the 18S and 16S rDNA datasets, respectively. The numbers at the bottom of the heatmap correspond to sample site numbers as shown in figure 3.1**a**. The heatmaps are arranged by placing the most abundant taxa at the top of the plot down to the least abundant at the bottom. Heatmaps enabled us to overview the distribution, composition and abundance of our datasets. The heatmaps were generated on log10-scaled abundances of the 18S and 16S rDNA datasets, using the heatmap.2 function, which is part of the gplots package, in R.

In figure 3.9a is the 18S rDNA at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the eukaryote node. The most abundant and constant eukaryotic taxon across the samples is the U.Stramenopiles. The U.Stramenopiles represent Stramenopiles species that could not be identified to the taxonomic level of class or below. Stramenopiles are a very diverse group of marine microbes and have been found inhabiting the oceans of the world, both in open ocean and coastal areas [Lin et al., 2012]. Therefore this is not a surprising result, that Stramenopiles are the most abundant and constant group throughout our samples. Also for example, in figure 3.9**a**, Dinophyceae abundance increases from the polar regions of the Arctic Ocean as we move down into the tropical region of the South Atlantic Ocean. This result is as expected, as Dinophyceae are found in the polar and tropical regions, with a higher abundance found in the tropical regions [Okolodkov and Dodge, 1996], [Taylor et al., 2007].

The taxonomy entries at the bottom of the heatmap from around the entry U.Rhodophyta and below are the ones that are driving diversity. There are a greater number of taxonomy entries in samples from around 29 to 68 which are located in the tropical regions compared to those samples around 1 to 28 located in the colder regions, which corresponds to what one would expect as the tropical regions are more diverse then the

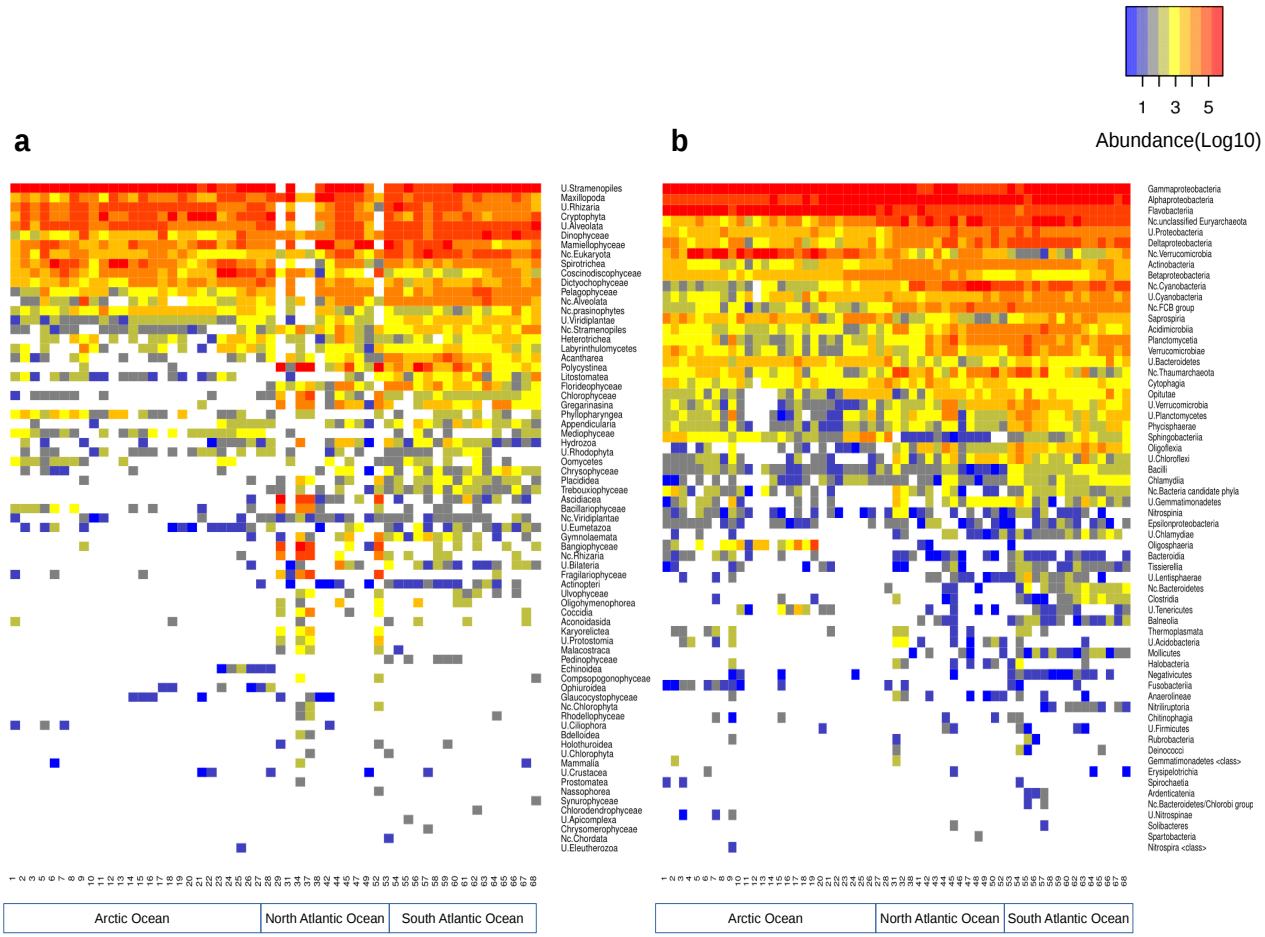

Figure 3.9: Panel **a** and **b** represents heatmaps of abundances arranged by latitude to the taxonomic rank of class, **a** represents the 18S rDNA taxonomy and **b** represents the 16S rDNA taxonomy. The taxonomy names in the dataset are displayed along the right side. The numbers at the bottom correspond to sample locations as shown in figure 3.1 **a**. The three regions of the Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean are displayed underneath their corresponding sample numbers. The colours correspond to log10-scaled abundances of the 18S and 16S rDNA, where red colours are high values and blue colours are low values. The heatmaps were generated using the heatmap.2 function, which is part of the gplots package, in R

colder regions.

In our PCoA plots in figure 3.8**a** at the bottom left side of the plot we identified 4 samples that are outliers, these are samples 29, 52, 34 and 37. In the heatmap in figure 3.9**a** it can be seen clearly how these samples' composition and abundance are different from those of the surrounding samples. These samples came from the North Atlantic Ocean expedition as explained in section 3.2.1. As noted before, Dr.Willem van de Poll and Dr.Klaas Timmermans noted that at the time of sampling a bloom was occurring, which could explain why these samples are clustering differently, since their composition and abundance are remarkably different from the other samples in

that region of the North Atlantic Ocean.

Displayed in figure 3.9**b** is the 16S rDNA at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the bacteria node. Likewise, for 16S rDNA, the top entries are the most abundant and constant throughout our samples going down to our least abundant and constant. The most abundant and constant taxon across our 16S rDNA samples is the Gammaproteobacteria. The Gammaproteobacteria are a large class of bacteria and can be found throughout the world's oceans [Williams et al., 2010], [Franco et al., 2017]. Therefore it is somewhat reassuring that Gammaproteobacteria are the most abundant and constant group throughout our 16S rDNA samples.

Also, the taxonomy entries at the bottom of the heatmap from around the entry U.Chlamydiae and below are the ones that are driving diversity. There are a greater number of taxonomy entries in samples from around 1 to 13 and 48 to 68 which are located in the polar and tropical regions, respectively. The pattern appears "n" shaped because diversity is decreasing in taxonomy entries from the polar regions as we pass into the temperate regions and then increasing as we pass into the tropical regions. This pattern is different to what is shown in the 18S rDNA heatmap in figure 3.9**a**.

3.4.4 Evenness and occupancy

For each of the species in our 18S and 16S rDNA datasets an evenness score (J) was calculated (J = H' /log(number of species), where H' is the Shannon diversity index) [Morris et al., 2014]. The occupancy refers to how many sample sites that species occurs in. A description of the methodology for the evenness and occupancy plots can be found in Section 3.3.3.

In figure 3.10**a** and 3.10**b**, we present an evenness and occupancy plot of the 18S and 16S rDNA datasets, respectively. In figure 3.10**a** is the 18S rDNA at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the eukaryote node. Displayed in figure 3.10**b** is the 16S rDNA at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the prokaryotes node. The numbers in the figure 3.10**a** and 3.10**b** correspond to taxonomy names which can be found in the Appendix A.C.

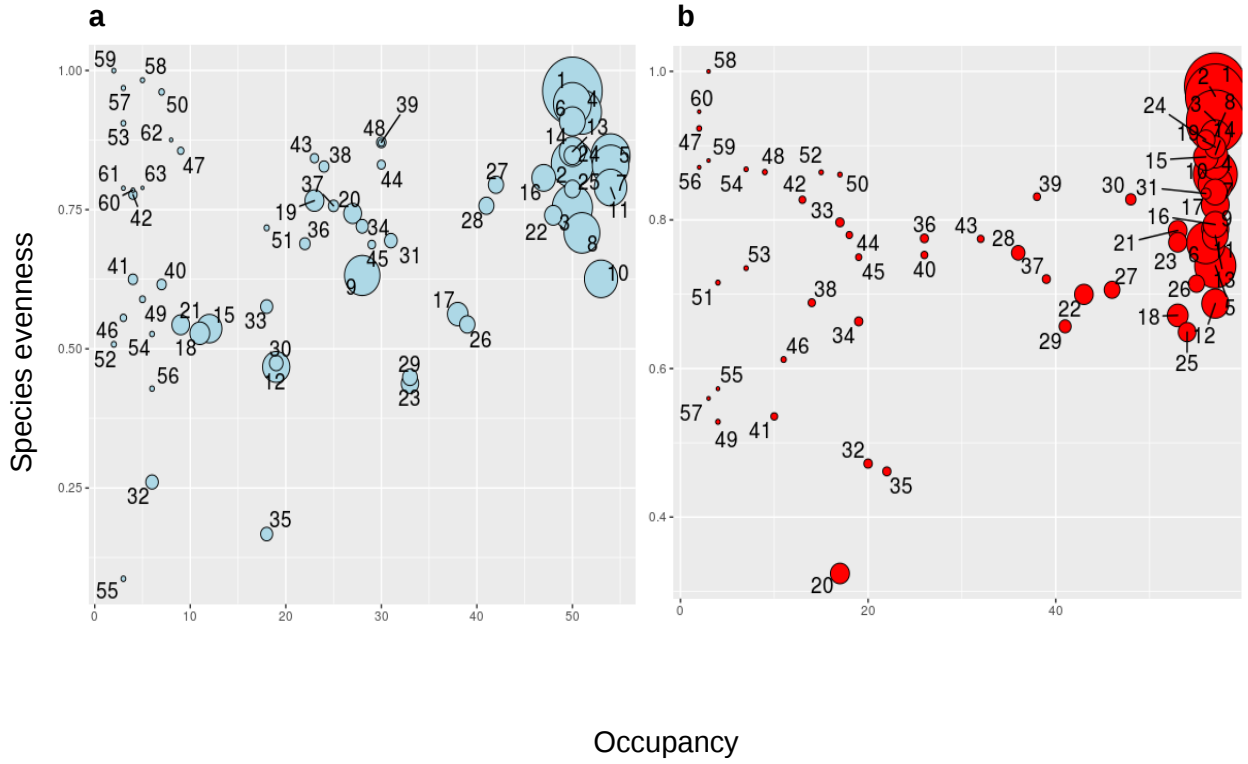

Figure 3.10: Panels **a** and **b** represent abundance taxonomy evenness and occupancy plots for the 18S and 16S rDNA datasets, respectively. The numbers in the plots correspond to taxon names which can be found in the Appendix A.C. The x-axis represents the number of times that class taxonomy occurs across the stations. The y-axis represents the evenness of that class taxonomy across stations it occurs in. Each circle represents a class taxonomy abundance. The size of each circle corresponds to the total abundance for that class, calculated by taking the square root of the abundance divided by π

The class evenness occupancy plot displayed in figure 3.10**a** for 18S rDNA shows that the more then 50% of the 18S rDNA classes have an evenness value between 0.5 to 0.85. This result indicates that the majority of the 18S rDNA classes are moderately to constantly present throughout our samples. A gradient pattern of high to low abundance and occupancy is observed of the 18S rDNA classes. The largest abundance is indicated by the greatest size of the circles and is located on the upper right hand corner of the plot. As the occupancy goes from high on the right hand side of the plot to low on the left hand side, the sizes of the circles decrease, indicating abundance also decreases. The Stramenopiles (1), Cryptophyta (2) and Mamiellophyceae (3) are the most abundant and widespread classes. From the 18S rDNA heatmap in figure 3.10**a**,

Stramenopiles are shown to dominate communities in the polar and temperate regions and decline only slightly in the tropical region. To a lesser degree, Cryptophyta also dominate communities in the polar and temperate regions and decline in the tropical region. Whereas Mamiellophyceae are more abundant in communities from the tropical region than in the polar and temperate regions.

The class evenness occupancy plot displayed in figure 3.10**b** of 16S rDNA shows the vast majority of the 16S rDNA taxonomy have an evenness value around 0.6 to 1. Therefore as we observed with the 18S rDNA species, the majority of the 16S rDNA species are also moderately to constantly present throughout our samples. Also, a gradient pattern of high to low abundance and occupancy is observed for the 16S rDNA species. As the occupancy goes from high on the right hand side of the plot to low on the left hand side, the sizes of the circles decrease, indicating abundance also decreases. The Gammaproteobacteria (1), Alphaproteobacteria (2) and Flavobacteriia (3) are the most abundant and widespread. From the 16S rDNA heatmap in figure 3.10**b**, Gammaproteobacteria are shown to dominate communities in the polar and temperate regions and decline very slightly in the tropical region. To a lesser degree, Flavobacteriia also dominate communities in the polar and temperate regions and decline in the tropical region. Whereas Alphaproteobacteria are more abundant in communities from the tropical and temperate regions than in the polar region.

Note that in panel **a** of figure 3.10 the 18S rDNA (n=54) for Prostomatea, Nassophorea, Synurophyceae, Chlorodendrophyceae, Apicomplexa, Chrysomerophyceae, Cephalochordata and Eleutherozoa are excluded from the analysis due to insufficient data to calculate the Shannon index. Also for 16S rDNA (n=57) in panel **b** Spartobacteria and Nitrospira were excluded from the analysis due to insufficient data to calculate the Shannon index.

3.4.5 Environmental plots

In this section, we present environmental plots. These plots were generated in collaboration with Dr.Schmidt.

Correlation heatmaps

In figure 3.11 panel **a** is the correlation heatmap for the Arctic Ocean samples, in panel **b** is the correlation heatmap for the North Atlantic Ocean samples and in panel **c** is the correlation heatmap for the South Atlantic Ocean samples. The correlation heatmaps enabled us to understand the statistical relationship of each taxon to the environmental variables. We produced the correlation heatmap with hierarchical clustering dendrograms on log10-scaled abundances of the 18S and 16S rDNA using the cor function, which is part of the WGCNA package and heatmap.2 function, which is part of the gplots package, in R. Displayed in figure 3.11 is the 18S rDNA at the taxonomic rank of class along with higher-level taxonomic assignments but excluding those assigned to the eukaryote node. Displayed in figure 3.12 is the 16S rDNA at the taxonomic rank of class along with higher-level taxonomic assignments but excluding those assigned to the prokaryotes node.

In figure 3.11**a**, which represents the correlation heatmap for the 18S rDNA of the Arctic Ocean samples, we observe that under temperature, salinity, longitude and latitude the heatmap is divided into two parts. We also can see according to the dendrogram on the left-hand side of the heatmap that the taxa form two large clusters. The cluster of taxa represented on the top part of the correlation heatmap is predominantly taxa that are positively correlating with temperature, salinity, longitude and latitude, while the cluster of taxa represented in the bottom half of the correlation heatmap is predominantly taxa that are negatively correlating with temperature, salinity, longitude and latitude. Under phosphate and silicate, there is no discernible pattern that can be determined for either of the clusters of taxa.

In figure 3.11**b**, which represents the correlation heatmap for the 18S rDNA of the North Atlantic Ocean samples, we can see according to the dendrogram on the left-hand side of the heatmap that the taxa form three clusters. The correlation heatmap is roughly divided into eight blocks of positively correlating and negatively correlating variables. The cluster of taxa represented on the top part of the correlation heatmap is predominantly taxa that are positively correlating with temperature, salinity, longitude and latitude, and negatively correlating with phosphate and silicate. The largest cluster of taxa represented in the middle part of the correlation heatmap is formed by two

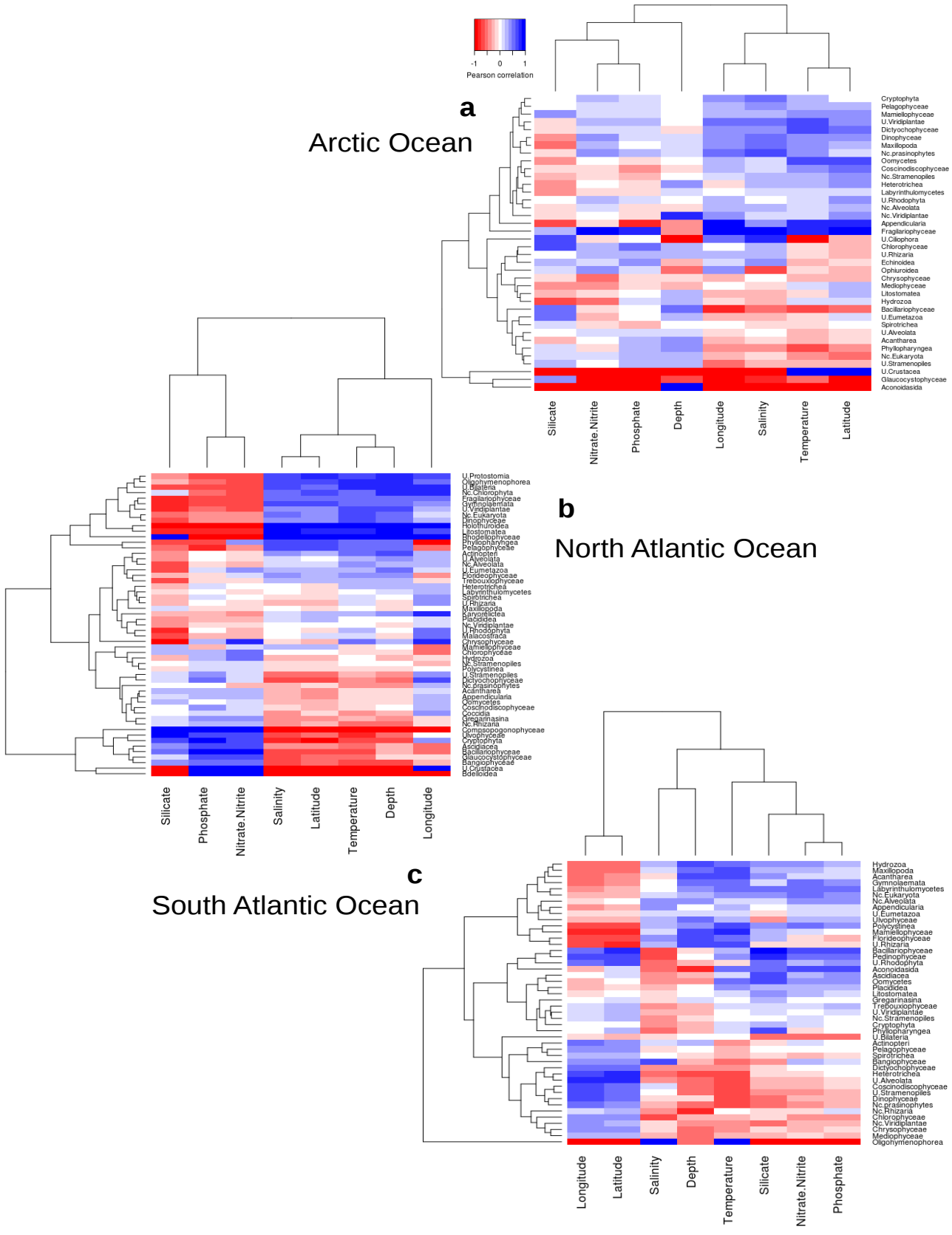

Figure 3.11: Correlation heatmaps of identified taxonomic classes and environmental data, **a** represents the Arctic 18S rDNA classes, **b** represents the North Atlantic 18S rDNA classes and **c** represents the South Atlantic 18S rDNA classes. The class names in the dataset are displayed along the right side and hierarchical clustering dendrogram on the opposite side. The environmental parameters are displayed at the bottom and hierarchical clustering dendrogram on the opposite side. The colours correspond to the Pearson correlation coefficient, where blue indicates a positive and red a negative correlation. The grey colour corresponds to no results, due to insufficient abundance data across the samples. The actual coefficients and p-values are given in Appendix A.D

smaller clusters, each with their own preference for the environmental variables. For the cluster of taxa at the top of the middle cluster of the correlation heatmap, the majority of the taxa are correlating positively with temperature, salinity, longitude and latitude, and negatively correlating with phosphate and silicate. For the cluster of taxa at the bottom of the middle cluster of the correlation heatmap, this consists predominantly of taxa that are negatively correlating with temperature, salinity, longitude and latitude, and positively correlating with phosphate and silicate. The cluster of taxa represented on the bottom part of the correlation heatmap is predominantly taxa that are negatively correlating with temperature, salinity, longitude and latitude, and positively correlating with phosphate and silicate.

In figure 3.11**c**, which represents the correlation heatmap for the 18S rDNA of the South Atlantic Ocean samples, we can see according to the dendrogram on the left-hand side of the heatmap that the taxa form three large clusters plus a singleton. The cluster of taxa represented on the top part of the correlation heatmap is predominantly taxa that are positively correlating with temperature, salinity, phosphate and silicate, while negatively correlating with longitude and latitude. The cluster of taxa represented in the middle of the correlation heatmap is predominantly taxa that are positively correlating with temperature, phosphate and silicate, longitude and latitude, while negatively correlating with salinity. The cluster of taxa represented at the bottom of the correlation heatmap is predominantly taxa that are negatively correlating with temperature, salinity, phosphate and silicate, while positively correlating with longitude and latitude. The singleton at the very bottom of the correlation heatmap is correlating negatively with phosphate, silicate, longitude and latitude, while correlating positively with salinity and temperature.

Specifically in the example in figure 3.11**a**, Chlorophyceae have a slightly negative correlation relationship with temperature and a slightly positive correlation relationship with salinity. In figure 3.11**b**, Chlorophyceae have a slightly negative correlation relationship with temperature and no correlation relationship with salinity. In figure 3.11**c**, Chlorophyceae have a negative correlation relationship with temperature and salinity. These results are not what is observed elsewhere in our analysis, for example, Chlorophyceae are more abundant in a tropical region, according to the 18S rDNA heatmap in figure 3.9**a**. It has been observed that species belonging to the group Chlorophyta,

of which Chlorophyceae is a member, can live even in extremely hot environments such as hot springs where temperatures can reach as high as 60°C [Mezhoud et al., 2014]. Therefore other relationships, potentially even positive correlations with temperature and salinity, would be possible in regions such as the South Atlantic Ocean.

In figure 3.12**a**, which represents the correlation heatmap for the 16S rDNA of the Arctic Ocean samples, we observe that under temperature, salinity, longitude and latitude the heatmap is divided into two parts. We also can see according to the dendrogram on the left-hand side of the heatmap that the taxa form two large clusters. The cluster of taxa represented on the top part of the correlation heatmap is predominantly taxa that are positively correlating with temperature, salinity, longitude and latitude, while the cluster of taxa represented in the bottom half of the correlation heatmap is predominantly taxa that are negatively correlating with temperature, salinity, longitude and latitude. Under phosphate and silicate, there is no discernible pattern that can be determined for either of the clusters of taxa.

In figure 3.12**b**, which represents the correlation heatmap for the 16S rDNA of the north Atlantic Ocean samples, we observe that the correlation heatmap is divided into four parts. We also can see according to the dendrogram on the left-hand side of the heatmap that the taxa form two large clusters. The cluster of taxa represented on the top part of the correlation heatmap is predominantly taxa that are positively correlating with temperature, salinity and latitude, while negatively correlating with phosphate, silicate and longitude. The cluster of taxa represented in the bottom half of the correlation heatmap consists predominantly of taxa that are negatively correlating with temperature, salinity and latitude, while positively correlating with phosphate, longitude and silicate.

In figure 3.12**c**, which represents the correlation heatmap for the 16S rDNA of the South Atlantic Ocean samples, we can see according to the dendrogram on the left-hand side of the heatmap that the taxa form five clusters, three small clusters and two larger clusters. The two small clusters of taxa represented on the top part of the correlation heatmap, are predominantly positively correlating with latitude, salinity and temperature, while negatively correlating with longitude, silicate and phosphate. The larger cluster of taxa represented in the middle part of the correlation heatmap, are predominantly negatively correlating with latitude, salinity and temperature, while

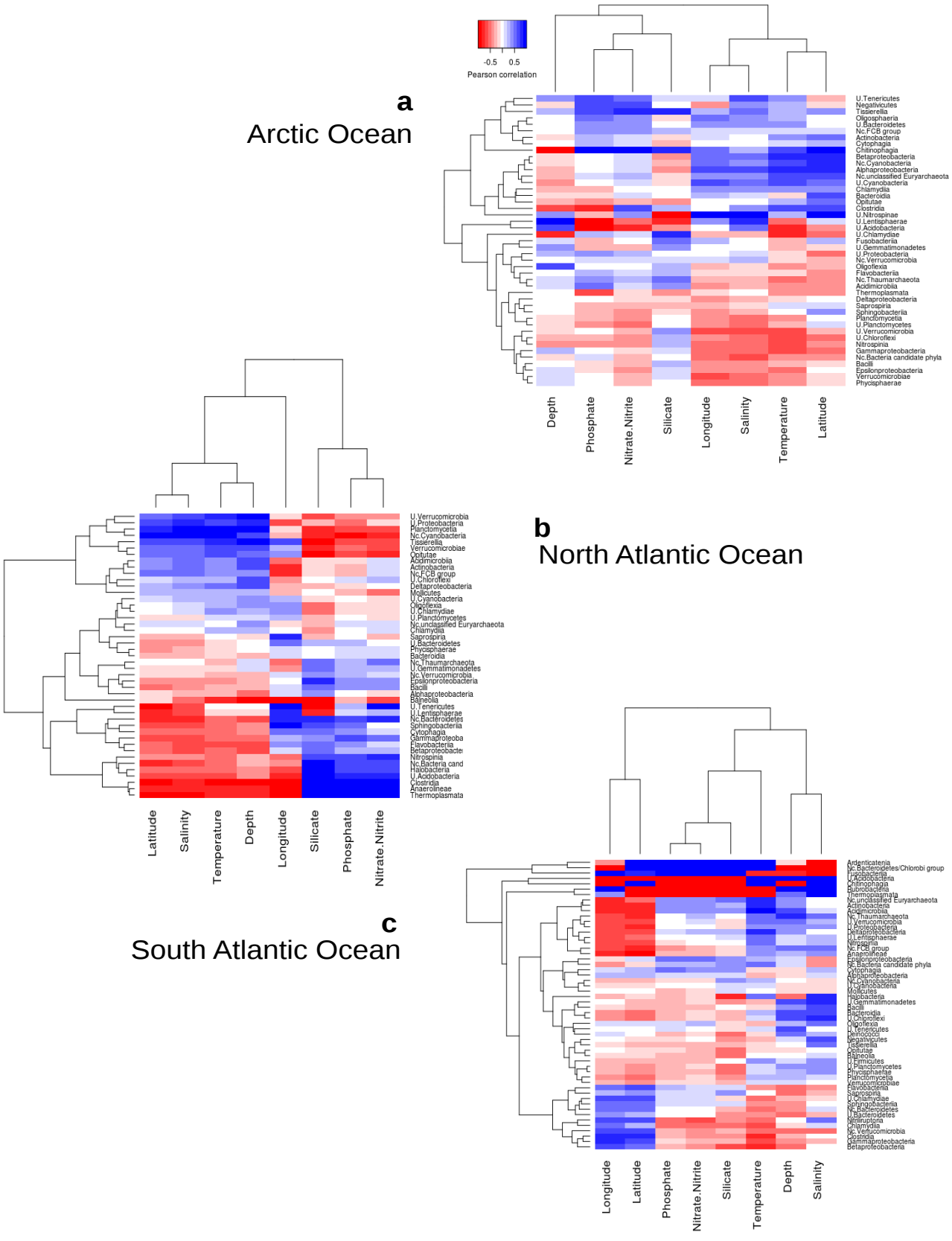

Figure 3.12: Correlation heatmaps of identified taxonomic classes and environmental data, **a** represents the Arctic 16S rDNA classes, **b** represents the North Atlantic 16S rDNA classes and **c** represents the South Atlantic 16S rDNA classes. The class names in the dataset are displayed along the right side and hierarchical clustering dendrogram on the opposite side. The environmental parameters are displayed at the bottom and hierarchical clustering dendrogram on the opposite side. The colours correspond to the Pearson correlation coefficient, where blue indicates a positive and red a negative correlation. The grey colour corresponds to no results, due to insufficient abundance data across the samples. The actual coefficients and p-values are given in Appendix A.D

positively correlating with longitude, silicate and phosphate. The two small clusters of taxa represented on the bottom part of the correlation heatmap, are predominantly negatively correlating with latitude, salinity and temperature, while negatively correlating with longitude, silicate and phosphate.

Specifically for the example in figure 3.12**a**, **b** and **c**, Nc.Cyanobacteria have a positive correlation relationship with temperature and salinity. Cyanobacteria are more abundant in the tropical regions, according to the 16S rDNA heatmap in figure 3.9**a**. Cyanobacteria can be found throughout the oceans of the world with a higher abundance for some of the species under Cyanobacteria in the tropical regions of the world [Flombaum et al., 2013].

Note that in figure 3.11 and figure 3.12 a number of the taxa are excluded from the analysis due to insufficient data to calculate the correlation coefficient values. In figure 3.11 panel **a** which represents the 18S rDNA from the Arctic Ocean the taxa excluded are Actinopteri, Ascidiacea, Bangiophyceae, Gregarinasina, Gymnolaemata, Mammalia, Nc.Rhizaria, U.Bilateria and U.Eleutherozoa. In figure 3.11 panel **b** which represents the 18S rDNA from the North Atlantic Ocean the taxa excluded are Aconoidasida, Echinoidea, Mammalia, Mediophyceae, Nassophorea, Ophiuroidea, Prostomatea, U.Chlorophyta and U.Ciliophora. In figure 3.11 panel **c** which represents the 18S rDNA from the South Atlantic Ocean the taxa excluded are Chlorodendrophyceae, Chrysomerophyceae, Coccidia, Compsopogonophyceae, Fragilariophyceae, Holothuroidea, Malacostraca, Mammalia, Nc.Chordata, Rhodellophyceae, Synurophyceae, U.Apicomplexa and U.Chlorophyta. In figure 3.12 panel **a** which represents the 16S rDNA from the Arctic Ocean the taxa excluded are Anaerolineae, Erysipelotrichia, Gemmatimonadetes, Halobacteria, Nitrospira, Rubrobacteria and U.Firmicutes. In figure 3.12 panel **b** which represents the 16S rDNA from the North Atlantic Ocean the taxa excluded are Deinococci, Erysipelotrichia, Fusobacteriia, Gemmatimonadetes, Negativicutes, Oligosphaeria, Rubrobacteria, Solibacteres, Spartobacteria and U.Firmicutes. In figure 3.12 panel **c** which represents the 16S rDNA from the South Atlantic Ocean the taxa excluded are Erysipelotrichia, Oligosphaeria, Solibacteres and Spirochaetia.

Non-metric multidimensional scaling

In figure 3.13**a** and 3.13**b** we present the non-metric multidimensional scaling (NMDS) plots for the 18S and 16S rDNA datasets, respectively, with environmental factors fitted. The numbers correspond to sample locations as displayed in figure 3.1**a**. NMDS plots allow us to visualise the distance matrix that is based on Bray-Curtis. Also, environmental factors that significantly correlate to 18S and 16S rDNA communities were fitted the the plot. Only environmental variables with a p-value of < 0.05 were selected. We performed NMDS with the metaMDS function and fitted the environmental variables with the envfit function, both functions are part of the vegan package, in R. This allowed us to investigate which environmental variables are driving the diversity of 18S and 16S rDNA communities. Figure 3.13**a** shows the 18S rDNA dataset at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the eukaryote node. Similarly, figure 3.13**b** shows the 16S rDNA dataset at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the bacteria node. We log10 transformed the 18S and 16S rDNA datasets for our NMDS analysis.

In figure 3.13**a** which displays the 18S rDNA dataset, temperature, salinity, phosphate and silicate are the environmental factors fitted, each with a p-value of 0.001. The majority of the samples cluster together, distinct from a sub-population on the right hand side of the plot. This sub-population is composed of samples 29, 34, 37 and 52. These are the same samples that cluster separately in the ordination analysis of the PCoA plots in figure 3.8**a**. In figure 3.13**a**, the short distance between the vectors representing the environmental variables temperature and salinity suggests that they are strongly positively correlated with one another. Likewise, for phosphate and silicate, the represented vectors are also positioned very close together, indicating they have a very strong positive correlation. Also given the opposing orientation of the pairs of environmental variables of temperature and salinity to phosphate and silicate this suggests a strong negative correlation between the two pairs. In the larger cluster, the samples are grouped by their regions of Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean. Samples from the polar region of the Arctic Ocean correlate strongly to phosphate and silicate, and samples from the tropical region of the South Atlantic

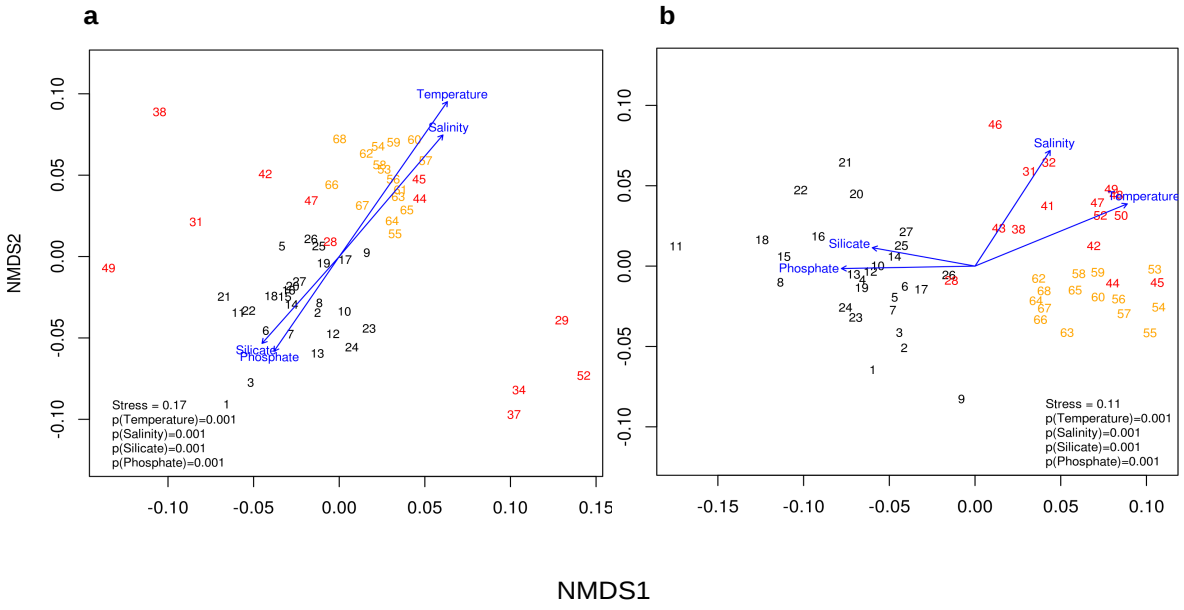

Figure 3.13: Panel **a**, NMDS of sampled stations with each number representing one sample of the 18S rDNA community. Significant environmental vectors for temperature (p=0.001), salinity (p=0.001), phosphate (p=0.001) and silicate (p=0.001) were fixed. Panel **b**, NMDS of sampled stations with each number representing one sample of 16S rDNA community. Significant environmental vectors for temperature (p=0.001), salinity (p=0.001), phosphate (p=0.001) and silicate (p=0.001) were fixed. NMDS was performed with the metaMDS function and the environmental variables were fitted with the envfit function (permutation test, 999 permutations). Both functions are part of the vegan package, in R. The numbers correspond to sample locations and the colours of the numbers correspond to ocean regions, black corresponds to the Arctic Ocean, red corresponds to the North Atlantic Ocean and yellow corresponds to the South Atlantic Ocean as shown in figure 3.1**a**

Ocean correlate strongly to temperature and salinity.

In figure 3.13**b**, which displays the 16S rDNA dataset, temperature, salinity, phosphate and silicate are again the environmental factors fitted, each with a p-value of 0.001. In figure 3.13**b**, the distance between the vectors representing the environmental variables temperature and salinity suggests that they are positively correlated, though not so strongly as in the 18S rDNA dataset. Likewise, for phosphate and silicate, the vectors are positioned even closer together, indicating they are strongly positively correlated. Also given the position of the vector representing temperature to those of phosphate and silicate this suggests a negative correlation, while salinity has a negative correlation to phosphate but a very weak correlation to silicate. The samples are

grouped by their regions of Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean. Samples from the polar region of the Arctic Ocean correlate with phosphate and silicate, while samples from the tropical region of the South Atlantic Ocean correlate with temperature and salinity.

Alpha diversity versus temperature

We investigated alpha-diversity (Shannon index) (as explained in section 3.3.3) in relation to the environmental covariates. To determine which environmental covariates were significant in the 18S and 16S rDNA datasets, the environmental covariates were related to the datasets' Shannon index by fitting generalized linear models. We then used a step-by-step backwards selection of the environmental covariates for model building and removed non-significant environmental covariates until the remaining environmental covariates were significant.

For the 18S and 16S rDNA datasets, temperature was the only significant environmental covariate with a p-value of 5.73e-2 and a p-value of 2.6e-05, respectively, that explained significant amounts of variation in the diversity for each of the 18S and 16S rDNA datasets. In figure 3.14**a** and 3.14**b** we present alpha diversity plotted against temperature for 18S and 16S rDNA respectively. The y-axis represents the alpha diversity for the stations, the x-axis represents the temperature. The numbers in the plots correspond to sample locations, according to the map in figure 3.1**a**. For each of the datasets, a significant positive correlation was observed, for the 18S rDNA dataset a R^2 of 0.4 with a p-value of 1.99e-07 and for the 16S rDNA dataset a R^2 of 0.7 with a p-value $<$ 2.2e-16. The alpha diversity was lower in the polar and temperate communities and highest in the tropical communities.

This relationship is also observed in the heatmaps in figure 3.9**a** and 3.9**b** for the 18S and 16S rDNA datasets, respectively. In the heatmaps, diversity increases as we move from the cold regions in the Arctic Ocean to the tropical regions of the Atlantic Ocean.

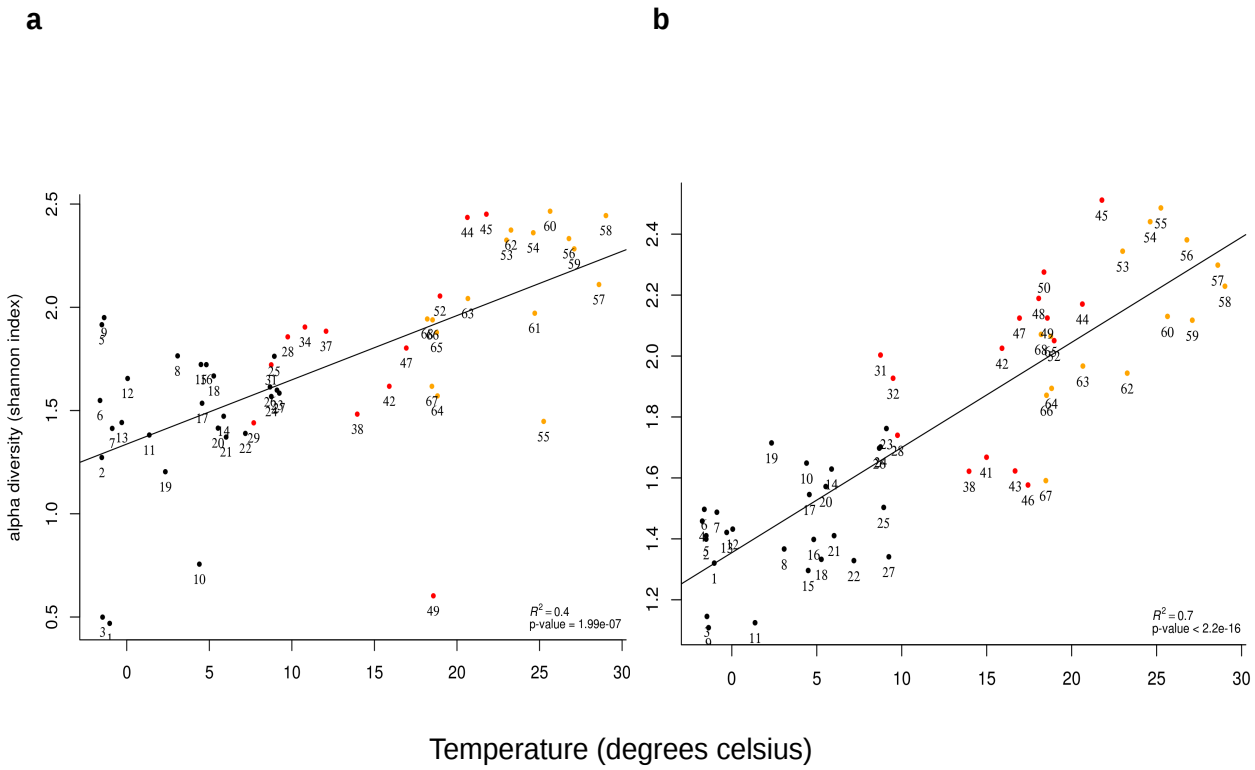

Figure 3.14: Panel **a**, a positive correlation of 18S rDNA diversity based on Shannon index with temperature. Based on backward model selection temperature was the only significant environmental covariate determined. Panel **b**, a positive correlation of 16S rDNA diversity, based on Shannon index with temperature. Based on backward model selection where temperature was the only significant environmental covariate determined. In panels **a** and **b**, the numbers correspond to sample locations and the colours of the numbers correspond to ocean regions, black corresponds to the Arctic Ocean, red corresponds to the North Atlantic Ocean and yellow corresponds to the South Atlantic Ocean as shown in figure 3.1**a**

3.4.6 Breakpoint analysis

In our previous analysis we investigated alpha-diversity in relation to the environmental covariates and determined temperature to be the only significant environmental covariate with a p-value of 5.73e-2 for our 18S rDNA dataset and a p-value of 2.6e-05 for our 16S rDNA dataset. We, therefore, chose to perform a breakpoint analysis in relation to temperature alone. The breakpoint analysis enabled us to investigate how increasing temperature affects the changing diversity of 18S and 16S rDNA across the Arctic Ocean down to the South Atlantic Ocean.

Displayed in figure 3.15**a** is the 18S rDNA dataset at the taxonomic rank of class

along with higher level taxonomic assignments but excluding those assigned to the eukaryote node. In figure 3.15**b** is the 16S rDNA dataset at the taxonomic rank of class along with higher level taxonomic assignments but excluding those assigned to the prokaryote node. In figure 3.15**a** and 3.15**b** is the breakpoint analysis for the 18S and 16S rDNA datasets, respectively. The numbers in the plots correspond to sample location as shown in figure 3.1**a**. The y-axis represents the beta diversity across the stations, the x-axis represents the temperature and the horizontal line marks the breakpoint.

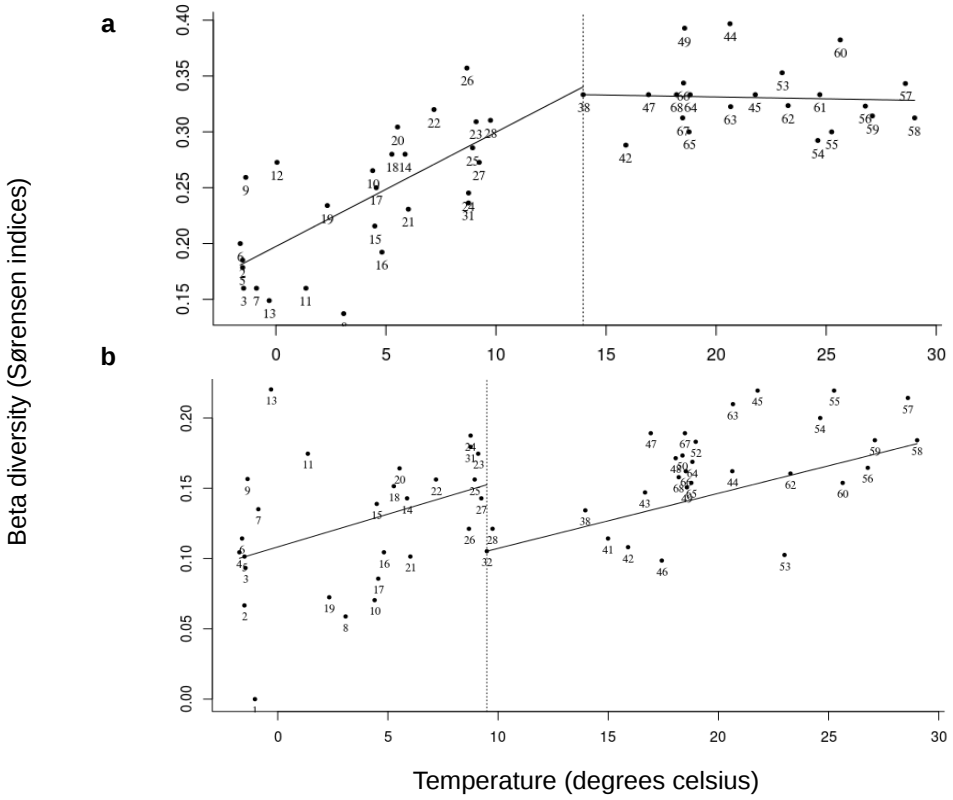

Figure 3.15: Panels **a** and **b** represent breakpoint analysis to the taxonomic rank of class, **a** represents the 18S rDNA dataset and **b** represents the 16S rDNA dataset. The numbers correspond to sample locations as shown in figure 3.1**a**. The breakpoint analysis was generated using piecewise regression in R as detailed in [Castro-Insua et al., 2016]. The y-axis represents the beta diversity across the stations. The x-axis represents the temperature. In each plot, the horizontal line marks the breakpoint. For the 18S rDNA dataset in panel **a** the breakpoint is 13.96°C with a p-value of 8.407e-11. For the 16S rDNA dataset in panel **b** the breakpoint is 9.49°C with a p-value of 1.413e-4

The breakpoint for the 18S rDNA dataset was determined to be 13.96°C with a p-value of 8.407e-11 and for the 16S rDNA dataset, the breakpoint was determined to be 9.49°C with a p-value of 1.413e-4. In figure 3.15**a** and 3.15**b**, the left hand side of the plots represent the lower temperatures, containing sample sites from the Arctic

Ocean and North Atlantic Ocean. The right hand side of the plots represent the higher temperatures and therefore contain sample sites from the North Atlantic Ocean and South Atlantic Ocean. There is a clear shift in the diversity for both the 18S and 16S rDNA dataset at their respective breakpoints. For both datasets of 18S and 16S rDNA, this shift in diversity occurs around the sample sites numbered in the 20's moving into the 30's, which positions the breakpoints in the North Atlantic Ocean as shown in figure 3.1**a**.

But the results are inconclusive due to the sparsity of the data points within the temperature range 9°C to 14°C, therefore we cannot conclusively determine breakpoints to exist within the temperate region of the North Atlantic Ocean. A change in diversity for both the 18S and 16S rDNA can be seen in the heatmaps in figure 3.9**a** and 3.9**b**. At the sample sites moving from the 20's to the 30's in the heatmaps which correspond to the temperature range 9°C to 14°C, the diversity can be seen to begin to change as it moves through the North Atlantic Ocean. Therefore potentially breakpoints within this region of the North Atlantic could exist but without greater sampling across this region, and at sufficient depth, we cannot determine this from our breakpoint analysis.

3.4.7 Co-occurrence analysis

In our co-occurrence analysis using the WGCNA package in R on the log10-scaled abundances of 18S rDNA species level and 16S rDNA genus level, two modules (a WGCNA term used to describe networks of highly correlating individuals) were found. In figure 3.16 we present the correlation heatmap between the two modules' eigengenes and the environmental variables. The module represented as blue (n=51) has a strong positive relationship with phosphate and a moderate positive relationship with silicate. Also, the blue module has a strong negative relationship with temperature and moderate negative relationship with salinity. The module represented as turquoise (n=70) has a strong negative relationship with phosphate and a moderate negative relationship with silicate. This turquoise module also has a strong positive relationship with temperature and a moderate positive relationship with salinity. Both the turquoise and the blue modules exhibit a very weak relationship with the environmental variable nitrate/nitrite. Above all the environmental variables, both the turquoise and blue

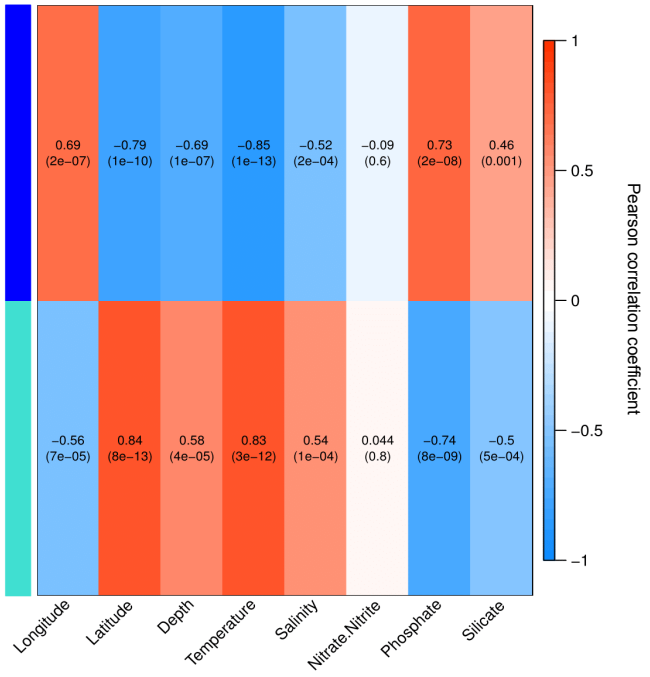

Figure 3.16: In the co-occurrence analysis with WGCNA on the log10-scaled abundances of 18S rDNA species level and 16S rDNA genus level, two modules were found. Depicted is the correlation heatmap between the modules' eigengene and environmental variables. Along the left side, the two modules are displayed in turquoise (n=70) and blue (n=51). The environmental variables are displayed at the bottom. The colours correspond to the correlation values; red is positively correlated and blue is negatively correlated. The values in each of the squares correspond to the assigned Pearson correlation coefficient value on top and p-value in brackets below

modules have the highest correlation to temperature. Also, the modules exhibit the opposite preference to temperature, the turquoise module has a strong positive correlation to temperature, while the blue module has a strong negative correlation to temperature.

We further examined each module's relationship to the environmental variables as depicted in figure 3.17. The correlation heatmap shows how each individual species of the turquoise module **a** (n=70) and the blue module **b** (n=51) correlate to the environmental variables. The taxa of the turquoise module represented in figure 3.17 **a** (n=70) have an overall negative relationship with phosphate and silicate and an overall positive relationship with temperature and salinity in varying degrees. The taxa of the blue module represented in figure 3.17 **b** (n=51) have an overall positive relationship with phosphate and silicate and an overall negative relationship with temperature and salinity in varying degrees. Both modules exhibit a very weak relationship with the environmental variable nitrate/nitrite.

In figure 3.18**a1** (turquoise (n=70)) and 3.18**b1** (blue (n=51)) the modules are

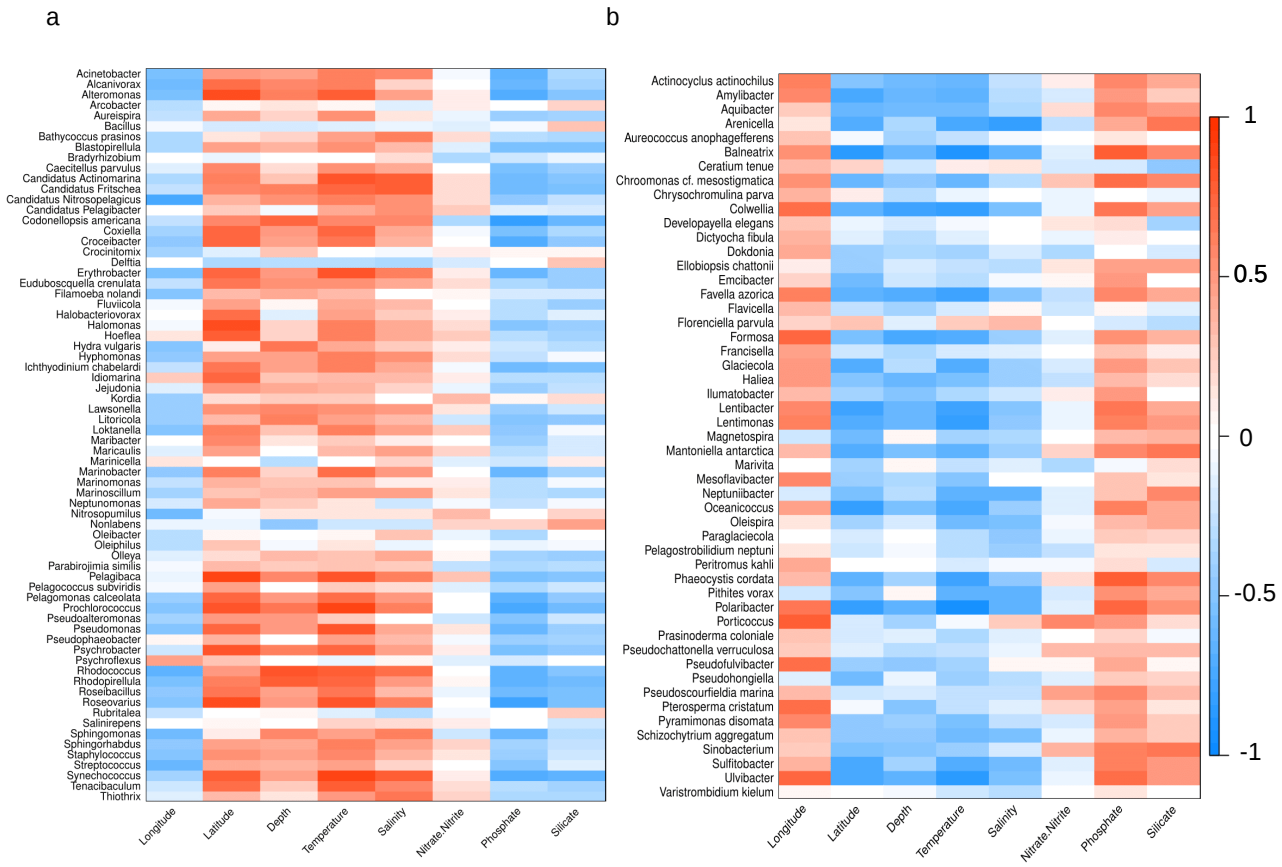

Figure 3.17: In the co-occurrence analysis with WGCNA on the log10-scaled abundances of 18S rDNA species level and 16S rDNA genus level, two modules were found. Depicted is the correlation heatmap for each of the two modules **a** (turquoise (n=70)) and **b** (blue (n=51)) of their species log10-scaled abundances and environmental variables. Along the left side on each of the two modules **a** and **b** are displayed the species name and environmental variables are displayed at the bottom. The colours correspond to the correlation values; red is positively correlated and blue is negatively correlated. The Pearson correlation coefficient values and p-values can be found in Appendix A.F

depicted as network diagrams. The edge distance between the nodes is an indication of the correlation strength. The numbers on the nodes correspond to the taxa names; the list of numbers to species can be found in the Appendix A.E. Degrees of connectivity refers the nodes (species) with the highest number of connections to other nodes (species). This does not automatically indicate the most abundant species or species that appear in all the stations. We could have connectivity between species that are lowly abundant and appear in few stations. The top five species/nodes with the greatest degrees of connectivity are coloured in orange. For the module in figure 3.18**a1** the five most highly connected taxa are Erythrobacter(1), Alteromonas(2), Roseovarius(3), Marinobacter(4) and Pelagomonas calceolata(5). For the module in

figure 3.18**b1** the five most highly connected taxa are Colwellia(1), Polaribacter(2), Balneatrix(3), Ulvibacter(4) and Amylibacter(5).

We also depict the two modules as word clouds in figure 3.18**a2** (turquoise (n=70)) and 3.18**b2** (blue (n=51)). These consist of the member taxa names with the size of a word reflecting the number of connections that taxa possess. Displayed in figure 3.18**a3** (turquoise (n=70)) and **b3** (blue (n=51)) are pie charts of the modules' taxa

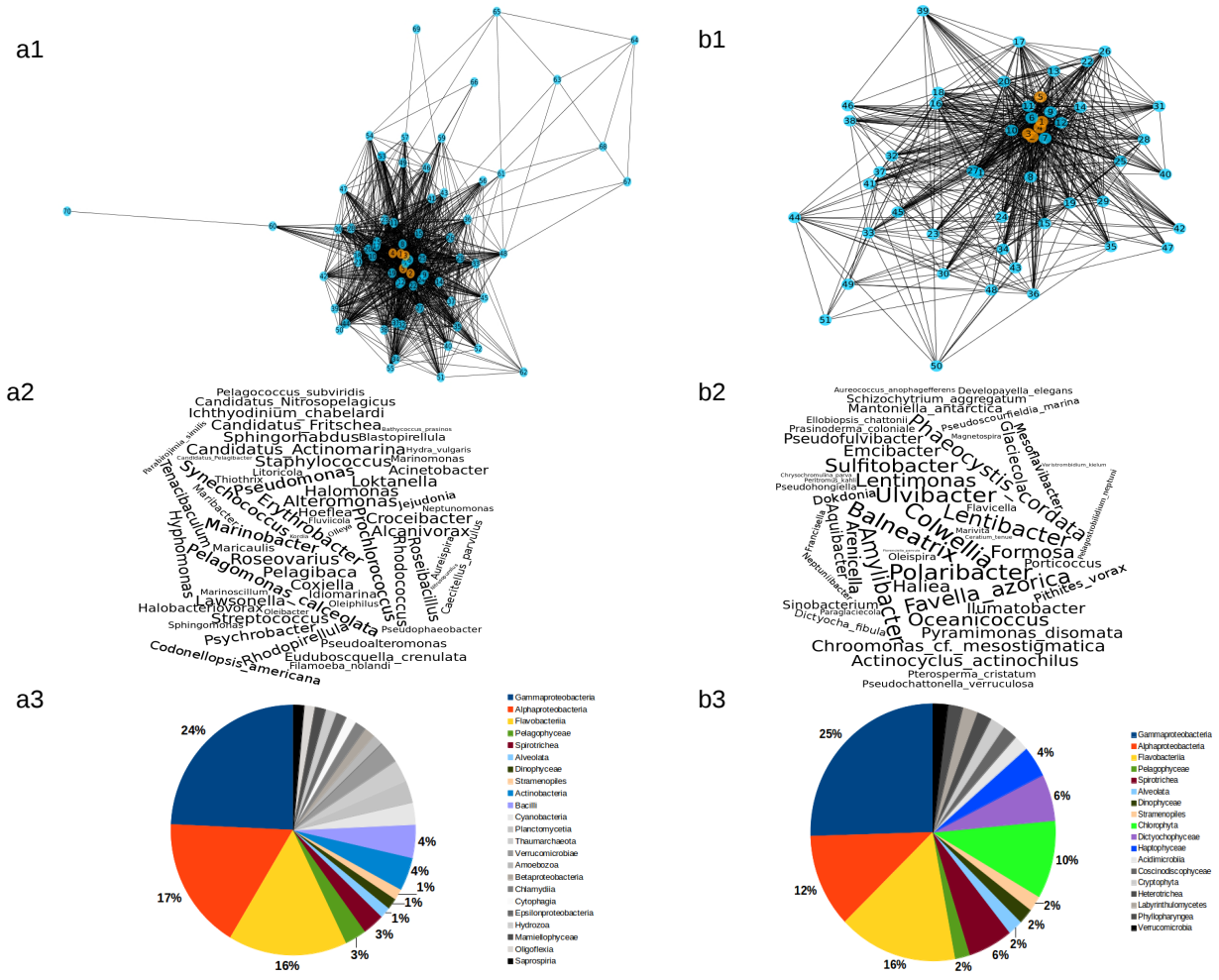

Figure 3.18: Co-occurrence analysis with WGCNA on the log10-scaled abundances of 18S rDNA species level and 16S rDNA genus level. In panels **a1** (turquoise (n=70)) and **b1** (blue (n=51)) are the two modules that we found depicted as network diagrams. These were generated with Cytoscape [Shannon et al., 2003]. The edge distance indicates the correlation strength between the nodes and the top five most highly connected nodes are coloured in orange. The numbers on the nodes correspond to the taxa names, the list of names to numbers of the taxa can be found in the Appendix A.E. In panels **a2** (turquoise (n=70)) and **b2** (blue (n=51)) the modules are depicted as word clouds. These consist of the member species and genus names, generated in WordCloud.com. The larger the name, the more connections that taxa has. In panel **a2**, taxa Crocinitomix, Salinirepens, Bacillus, Bradyrhizobium, Rubritalea, Arcobacter, Delftia, Marinicella, Nonlabens and Psychroflexus were removed as they are unreadable due to their frequency being too low to display with the others. In panels **a3** (turquoise (n=70))and **b3** (blue (n=51)) are pie charts of the classes of the modules taxa. The list of names of the taxa to their class can be found in the Appendix A.E

class membership; the list of names of the taxa to their class can be found in the Appendix A.E. Gammaproteobacteria, Alphaproteobacteria, Flavobacteriia, Pelagophyceae, Spirotrichea, Alveolata, Dinophyceae and Stramenopiles are the greatest contributors in each module and are similarly proportionate. In figure 3.18**a3** (turquoise (n=70)), Actinobacteria and Bacilli contribute significantly but are unique to that module. Also in figure 3.18**b3** (blue (n=51)) Chlorophyta, Dicictyochophyceae and Haptophyceae contribute significantly and are unique to that module. These results are consistent with what we see in the heatmaps in figure 3.9**a** and 3.9**b**. The heatmaps show the distribution and abundance across the samples from the polar Arctic Ocean down to the tropical South Atlantic Ocean.

3.5 Discussion

In this chapter, we have presented a phylogenetic analysis of phytoplankton 18S and 16S rDNA. We described the methodology of the phylogenetic analysis and normalised copy number for both the 18S and 16S rDNA dataset, and their analysis with rarefaction curves, heatmaps, PCoA, evenness and occupancy plots, environmental plots, breakpoint analysis and co-occurrence analysis. This was a unique large scale examination of the 18S and 16S rDNA species communities taken from a transect of the Arctic and Atlantic oceans that was sampled in close proximity to the coast. There are significant differences between coastal waters and the open ocean. Coastal waters have a lower temperature and higher nutrient content in comparison to the open ocean [Toseland et al., 2013]. Therefore coastal waters are more productive than the open ocean for phytoplankton, in terms of, for example, Chlorophyll a concentrations and primary productivity [Trimborn et al., 2015]. This has therefore given us a new understanding of how the environmental conditions affect phytoplankton species communities.

While to the best of our abilities we have extensively sampled and precisely designed our experimental approaches there are limitations in our analysis. One such limitation is that the North Atlantic Ocean samples were acquired from two other collaborators; Dr.Willem van de Poll of the University of Groningen, Netherlands and Dr.Klaas Timmermans of the Royal Netherlands Institute for Sea Research. Their sampling procedures were slightly different to those performed by Dr.Katrin Schmidt

who performed an additional pre-filtering step with a 100m mesh to remove larger organisms. Also another limitation is that a number of the samples from the North Atlantic Ocean failed to pass quality control steps during sequencing. From our 18S rDNA samples a total of 13 out of 24 samples failed from the North Atlantic Ocean and also station 4 from the Arctic Ocean. From our 16S rDNA samples, a total of 10 samples failed out of 24 from the North Atlantic Ocean and also station 61 from the South Atlantic Ocean. In addition, only a single sample was obtained at each station; we did not obtain replicate samples. However, we regard these as minimal limitations given the number of high quality samples that we obtained and the relatively close proximity of the samples to one another.

For our analysis of the 18S and 16S rDNA datasets, we identified a gradient of increasing diversity, as we moved from the cold temperatures of the Arctic Ocean down to the tropical temperatures of the South Atlantic Ocean. These findings are what we expected to observe, as it has been known for years that diversity is higher in the warm tropical regions than in the cold polar regions of the world [Brown, 2014]. We also identified a number of 18S and 16S rDNA species that were occurring continuously throughout the samples, which included, for example, Stramoenopiles and Gammaproteobacteria, and these findings are consistent with previewed published works have found [Lin et al., 2012], [Franco et al., 2017].

We further identified in both 18S and 16S rDNA datasets, that the samples from the polar region of the Arctic Ocean are correlated strongly to phosphate and silicate, and samples from the tropical region of the South Atlantic Ocean are correlated strongly to temperature and salinity. In addition, for both the 18S and 16S rDNA datasets we found that temperature was the only significant environmental covariate for alpha diversity. In our co-occurrence analysis with WGCNA on the 18S rDNA species level and 16S rDNA genus level, we found two modules. The larger of the modules (turquoise (n=70)) was found to have a strong positive correlation relationship to temperature, indicating that this module is likely to be found in a warm climate. In contrast, the smaller module (blue (n=51)) was found to have an overall strong negative correlation relationship to temperature, indicating that this module is likely to be found in a cold climate. These co-occurrence analysis results are important as we can hypothesise that different microbial communities have different preferences for temperature. Moreover,

as global warming is predicted to raise the temperatures in the ocean, our results could potentially enable us to forecast how climate change will affect these microbial communities using climate models underpinned by genetic information.

In our breakpoint analysis of the 18S and 16S rDNA datasets, we identified a putative breakpoint of 13.96°C for the 18S rDNA dataset and 9.49°C for the 16S rDNA dataset. This positioned the breakpoint in the North Atlantic Ocean off the coast of France. It is interesting to note that the breakpoints were all located in the temperate region of the North Atlantic Ocean. Therefore as you move from the cold Arctic Ocean to the warm tropical regions in the South Atlantic Ocean there is a radical shift in the diversity and interactions of the 18S and 16S rDNA species communities. There have been various types of studies that have included a breakpoint analysis, such as [Campra and Morales, 2016] which looks at surface air temperature records in southeastern Spain over a number of years. Also, another breakpoint study is [Castro-Insua et al., 2016], which is an analysis of various terrestrial animals such as bats and birds across latitudes in America. They determined a number of breakpoints for each of their subject animals, and these were found in the range of 25° to 58° latitude [Castro-Insua et al., 2016]. Our breakpoints are also located in this latitude range, as we can determine from our metadata that the 18S rDNA breakpoint of 13.96°C is located at 59.5° latitude and the 16S rDNA breakpoint is located at 45.5° latitude. But our results are currently inconclusive due to the sparsity of the data points within the temperature range 9 to 14°C, therefore we cannot conclusively determine breakpoints to exist within the temperate region of the North Atlantic Ocean. To explicitly determine breakpoints in our 18S and 16S rDNA datasets we require further sampling to be uniformly distributed across the entire sampling range and of sufficient sampling depth throughout.

Chapter 4

Metatranscriptomics analysis

4.1 Summary

In the last chapter, we analysed 18S and 16S rDNA datasets from the Arctic Ocean, North Atlantic Ocean and the South Atlantic Ocean and found a greater diversity of microbes in the tropical regions of the South Atlantic Ocean, versus the polar regions of the Arctic Ocean. Additionally, in our co-occurrence analysis on the 18S and 16S rDNA datasets, we found two community networks, one positively correlated to temperature and the other negatively correlated to temperature. Based on these results, we can hypothesise that different microbial communities have different preferences for temperature. We also performed a breakpoint analysis on our 18S and 16S rDNA datasets and found a shift in diversity occurring in the North Atlantic Ocean. In particular, the shift occurs in the temperate region of the Ocean, between the polar Arctic Ocean and the tropical South Atlantic Ocean.

Now we focus on the question of what are the microbes doing, and in particular gaining an understanding of their genetic activity. Moreover, it will be of interest to see if the different data types (16S/18S rDNA data and metatranscriptome data) are in agreement or not. To address this question, we shall perform a metatranscriptomic analysis. In the next section, we describe the sequencing of the metatranscriptomic dataset. In Section 4.3 we describe our pipelines for the metatranscriptomic analysis, as well as additional methods, then we will present the results of our analysis in Section 4.4. In Section 4.5, we end with a discussion of the results presented in this chapter.

4.2 Sequencing and preprocessing

The samples were collected as described in chapter 3. All samples were sequenced and preprocessed by the Joint Genome Institute (JGI) (Department of Energy, Walnut Creek, CA, USA). Metatranscript sequencing was performed on an Illumina HiSeq-2000 instrument [Huntemann et al., 2016]. A total of 65 samples passed quality control after sequencing with 5.7 Gb of sequence read data over all samples for analysis. Here we describe the JGI's computational pipeline for preprocessing the metatranscript reads.

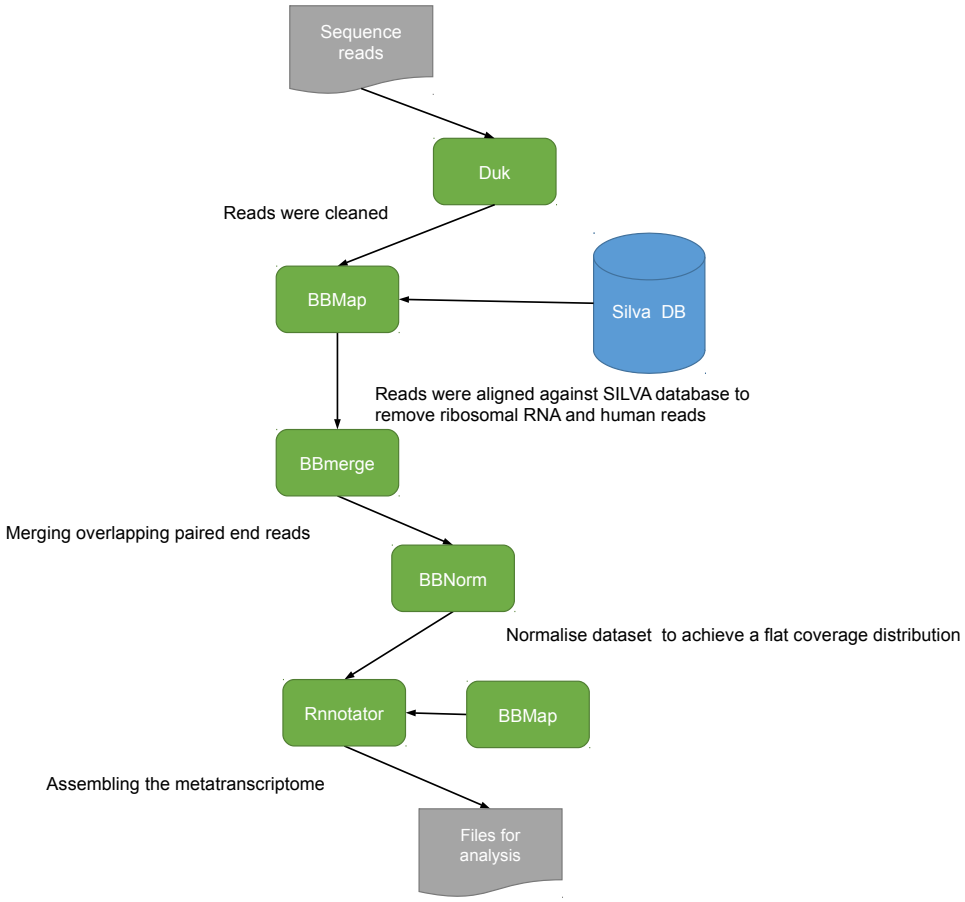

Figure 4.1: Diagram of JGI's computational pipeline for preprocessing the metatranscript reads. The pipeline at various stages incorporates databases (blue), BBTools tools (green) and processed files (grey)

In figure 4.1 we provide an overview of the JGI pipeline. JGI employed their suite of tools called BBTools [DOE Joint Genome Institute, 2017] for preprocessing the sequences. As shown in figure 4.1, first the sequences were cleaned using a tool called Duk [DOE Joint Genome Institute, 2017]. Duk (see Section 2.4.1) is a tool in the BBTools suite, that performs various data quality procedures such as quality trimming and filtering by kmer matching [DOE Joint Genome Institute, 2017]. In our dataset, Duk identified and removed adapter sequences, and also quality trimmed the raw reads

to a phred score of Q10. In Duk the parameters were; kmer-trim (ktrim) was set to r, kmer (k) was set to 25, shorter kmers (mink) set to 12, quality trimming (qtrim) was set to r, trimming phred (trimq) set to 10, average quality below (maq) set to 10, maximum Ns (maxns) set to 3, minimum read length (minlen) set to 50, the flag "tpe" was set to t, so both reads are trimmed to the same length and the "tbo" flag was set to t, so to trim adapters based on pair overlap detection. The reads were further filtered to remove process artefacts also using Duk with the kmer (k) parameter set to 16.

BBMap [DOE Joint Genome Institute, 2017] is also a tool in the BBTools suite, that performs other operations such as making sequence alignments of DNA and RNA reads to a database. BBMap aligns the reads by using a multi-kmer-seed-and-extend approach [DOE Joint Genome Institute, 2017]. To remove ribosomal RNA reads, the reads were aligned against a trimmed version of the SILVA database using BBMap with parameters set to; minratio (minid) set to 0.90, local alignment converter flag (local) set to t and fast flag (fast) set to t. Also any human reads identified were removed using BBMap [DOE Joint Genome Institute, 2017].

BBmerge [DOE Joint Genome Institute, 2017] is a tool in the BBTools suite that performs the merging of overlapping paired end reads [DOE Joint Genome Institute, 2017]. For assembling the metatranscriptome, the reads were first merged with the tool BBmerge, and then BBNorm was used to normalise the coverage so as to generate a flat coverage distribution. This type of operation can speed up assembly and can even result in an improved assembly quality [DOE Joint Genome Institute, 2017].

Finally as shown in figure 4.1, Rnnotator [Martin et al., 2010] was employed for assembling the metatranscriptome. Rnnotator assembles the transcripts by using a de novo assembly approach of RNA-Seq data and it accomplishes this without a reference genome [Martin et al., 2010]. The tool BBMap was used for reference mapping, the cleaned reads were mapped to metagenome/isolate reference(s) and the metatranscriptome assembly [DOE Joint Genome Institute, 2017].

4.3 Methods

4.3.1 Computational pipeline for taxonomic classification analysis

PhymmBL is a hybrid taxonomic classifier, that combines Phymm composition-based taxonomic predictions and BLAST based homology results to label each of the input sequences [Mande et al., 2012]. Phymm employs interpolated Markov models (IMMs) which is a form of the Markov chain that uses a variable number of states to calculate the probability of the next state. Phymm builds IMMs to characterize the variable length oligonucleotides that are distinct for a particular phylogenetic clade, whether it be a species, genus or a higher taxonomic level. During construction of the IMMs, the IMM algorithm builds a probability distribution based on the observed patterns of nucleotides that describe each species in the reference database. During classification for each of the input sequences, each of the IMMs is used as a scoring method by inspecting the nucleotides in the input sequence and then outputs a score that corresponds to the probability that the input sequence was generated from the same distribution as that used to construct the IMM. The input sequence is classified with the clade labels that belong to the organism whose IMM produced the best score for that input sequence. During BLAST, each input sequence is submitted as a BLASTN query, searching against the same reference database used to generate the IMMs, and clade labels are assigned for the best BLAST hit. The combined score of Phymm and BLAST for PhymmBL is determined by using the function score=IMM+1.2(4-log(E)), where IMM is the score from the best matching IMM and E is the smallest E-value given by BLAST. It has been demonstrated by [Brady and Salzberg, 2009a] that PhymmBL's hybrid method outperforms Phymm and BLAST separately and also that BLAST outperforms Phymm [Brady and Salzberg, 2009a].

The metatranscriptomic dataset was taxonomically classified using Dr.Andrew Toseland's taxonomic classification pipeline as described in [Toseland et al., 2013]. This pipeline employ's PhymmBL [Brady and Salzberg, 2009b] which contains phytoplankton in its database, and the contents of PhymmBL's default database were reduced by taking the first occurrence of a species under each genus. We ran the pipeline with

Dr.Toseland's assistance.

The PhymmBL taxonomic classification pipeline contains a representative set of 44 eukaryote organisms. This set consists of genomes and expressed sequence tags (EST) of the major eukaryote groups with a focus on algal species. From NCBI-dbEST the EST sequences were downloaded. Then with CD-HIT-est these EST sequences were clustered with a 95% similarity in order to ensure non-redundancy of the sequences. The genome sequences were downloaded from NCBI GenBank, JGI and separately four genomes; *Cyanidioschyzon merolae, Danio rerio, Homo sapiens* and *Strongylocentrotus purpuratus* were downloaded. These four genomes were obtained from specific locations as outlined in Appendix B.A. Species such as *Danio rerio* and *Homo sapiens* were included to check for contamination. Taxonomic classifications were taken from the NCBI taxonomy [NCBI Resource Coordinators, 2016] and AlgaeBase [Guiry, M.D. & Guiry, 2008] to produce a PhymmBL configuration file. In batch mode, the sequence files and taxonomic details were added to PhymmBL and interpolated Markov models (IMMs) were created for each new organism [Toseland et al., 2013].

For each sequence, PhymmBL outputs a taxonomic label of genus, family, order, class and phylum with a confidence score between 0 and 1 [Brady and Salzberg, 2011]. For our PhymmBL results, we took a confidence score cutoff of ≥ 0.9 at the phylum level to order to ensure our results were as accurate as possible. For each sample, the number of sequences under each taxon was summed up. The files were further normalised by applying hits per million.

4.3.2 Computational pipeline for functional analysis

JGI performed the functional analysis on the metatranscriptomic dataset. A functional analysis for a metatranscriptomic dataset is a standard pipeline but we liaised with JGI, and they updated their pipeline based on our feedback. For example, functional analysis results obtained using JGI's Integrated Microbial Genomes (IMG) standard pipeline were different to those obtained using the standard JGI pipeline. It was found that this was due to different versions of databases being employed. This feedback induced JGI to update their databases in their standard pipeline.

Our datasets were submitted to JGI's IMG. JGI's annotation system is called the

Metagenome Annotation Pipeline (MAP) (v4.15.2). JGI used HMMER 3.1b2 [Eddy, 1996] and the Pfam v30 [Finn et al., 2016] database for the functional analysis of our metatranscriptomic dataset. There are other databases that perform similar roles but we only discuss Pfam because that is the one we use for this thesis.

4.3.3 Further analysis

In this section, we describe the methodology that we developed for our analysis of the metatranscriptomic dataset. From JGI's IMG, we downloaded a Pfam file for each sample. This resulted in 7,453 Pfam functional assignments and their gene counts across the 65 samples. The files were further normalised by applying hits per million.

Canonical Correspondence Analysis (CCA)

We used the R package VEGAN to perform a Canonical Correspondence Analysis (CCA) between the Pfam dataset and the environmental data. CCA uses a dataset of measured variables such as Pfam gene counts in our case, and a dataset of additional explanatory variables such as temperature and salinity during the analysis, in order to find the relationship between the two datasets, and therefore enable us to determine how might the environmental variables determine the response variable values [Paliy and Shankar, 2016].

Environmental variables such as temperature and salinity can influence microbial communities [Hou et al., 2017]. Therefore in chapter 3 for the analysis of our 18S and 16S rDNA datasets, we used NMDS plots to visualise the distance matrix based on Bray-Curtis [Paliy and Shankar, 2016] with environmental factors fitted that significantly correlate to 18S and 16S rDNA communities. To analyse the metatranscriptome data we instead performed a CCA on our Pfam protein families datasets, because not all genes are affected by environmental variables such as housekeeping genes which are constantly expressed [Eisenberg and Levanon, 2013]. CCA enables us to determine what proportion of our dataset is affected by environmental variables and identify the environmental variables that significantly explain the variation of our dataset [Paliy and Shankar, 2016].

We log10 transformed the Pfam normalised gene counts, as explained in chapter

3 so that the data complies better to the assumptions of a parametric statistical test [Paliy and Shankar, 2016]. The environmental data consisted of temperature, salinity, nitrate/nitrite, phosphate and silicate.

Co-occurrence analysis

The methodology of how we performed the co-occurrence analysis is outlined in section 3.3.3. Based on the Pfam log10-scaled gene counts dataset the power beta ($\beta \geq 1$) was determined to be 15.

The co-occurrence analysis resulted in fifteen modules being found, that included a grey module which represents those taxa that could not be assigned to a module. The modules were further examined to determine if any modules were highly correlated (≥ 0.75) to one another based on their eigengene. This resulted in two modules being merged and therefore thirteen modules were taken for analysis.

4.4 Results

4.4.1 Taxonomic classification heatmap

In figure 4.2 we present a heatmap that we generated, arranged by latitude for the taxonomic classified metatranscriptomic dataset at the taxonomic rank of phylum. The numbers at the bottom correspond to sample site numbers as shown in figure 3.1**a**. The heatmap is arranged by placing the most abundant read counts at the top of the plot down to the least abundant read counts at the bottom. Heatmaps enabled us to overview the distribution, composition and abundant read counts of our dataset. In figure 4.2 we basically observe no gradient of increasing diversity of taxa across the samples as we move from the polar Arctic Ocean through the temperate North Atlantic Ocean and into the tropical South Atlantic Ocean. This is in contrast to what we observed in the heatmaps of the 18S and 16S rDNA datasets in figure 3.9**a** and 3.9**b**, respectively. In figure 4.2 the majority of the taxa display a consistent abundance across the samples. There is variation between taxa abundance, as the heatmap is roughly divided into four blocks of colour.

It is not meaningful to compare the read counts of the metatranscriptomic dataset

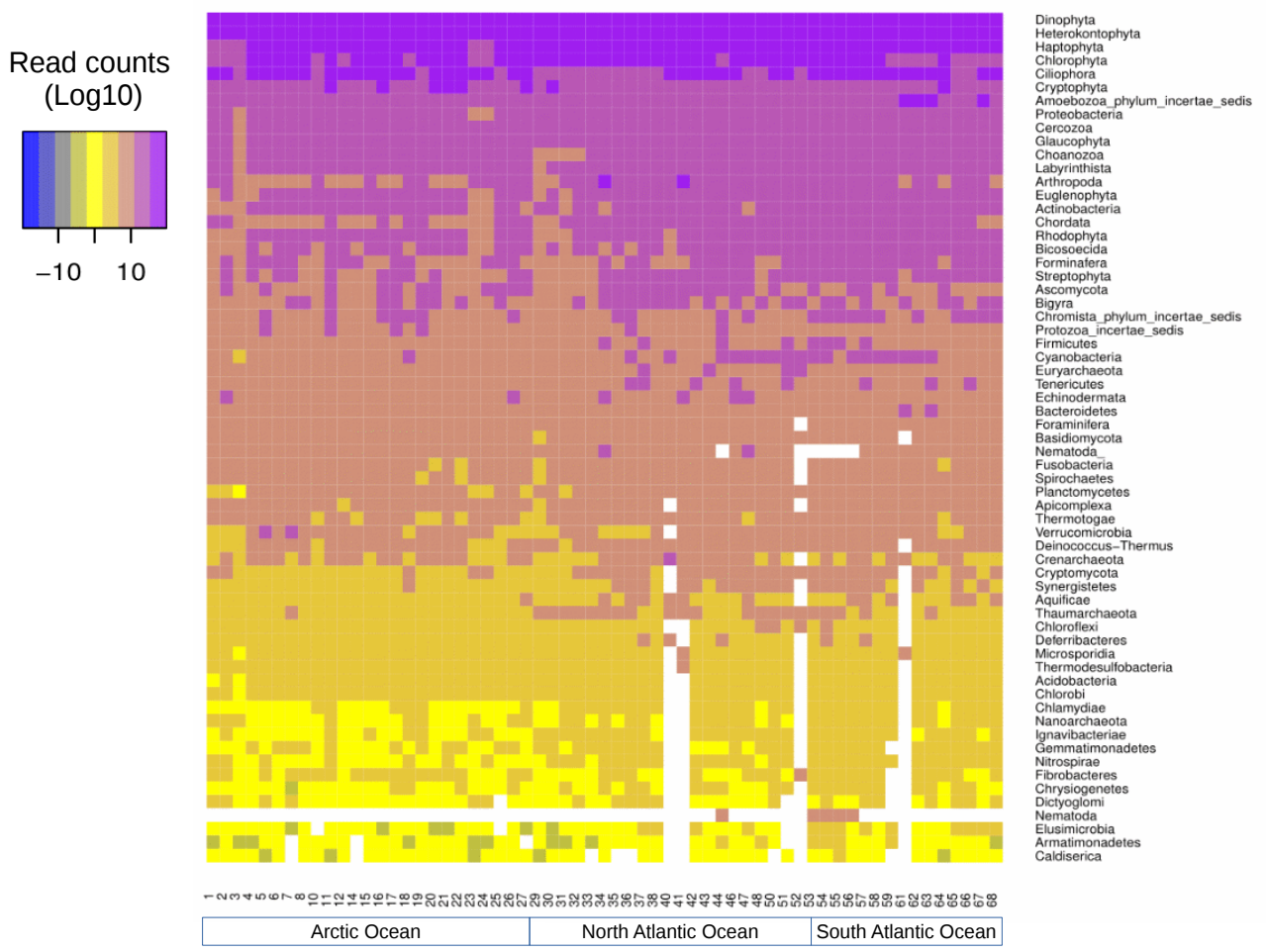

Figure 4.2: A heatmap of the metatranscriptomic dataset taxonomically classified and arranged by latitude versus the taxonomic rank of phylum. The taxonomy names in the dataset are displayed along the right side. The numbers at the bottom correspond to sample locations as shown in figure 3.1 **a**. The three regions of the Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean are displayed underneath their corresponding sample numbers. The colours correspond to log10-scaled read counts, where purple colours are high values and blue colours are low values. The heatmap was generated using the heatmap.2 function, which is part of the gplots package, in R

in figure 4.2 with the abundances of the 18S and 16S rDNA datasets in figure 3.9. This is because abundance gives an indication on the density of a species, whereas read counts only give an indication of presence or absence. However with metagenomic data that we will have in the near future for these sample sites, we can use this data to compare with the abundances of the 18S and 16S rDNA datasets. We do not see a gradient on species in the metatranscriptomic dataset as we did in the 18S and 16S rDNA datasets because the 18S and 16S rDNA analysis targets specific genes and then taxonomically classifies the samples against a reference database based on the SILVA database as described in Section 3.3.1 and Section 3.3.2, respectively [Reller et al., 2007]. In contrast the metatranscriptomic analysis employs direct cDNA sequencing of

the sample and then taxonomically classifies it against a reference databases of genomes and NCBI expressed sequence tags (EST) as described in Section 4.4.1 [Leimena et al., 2013].

However, we can compare identified taxa entities between the metatranscriptomic dataset and the 18S and 16S rDNA datasets, as a means of confirmation between the two analyses. In figure 4.3 **a** and **b** we represent Venn diagrams, comparing the number of taxa entities at the taxonomic rank of phylum between the 18S rDNA dataset to the metatranscriptomic dataset and the 16S rDNA dataset to the metatranscriptomic dataset, respectively. In figure 4.3 **a**, in the comparison of the 18S rDNA taxa entities to the metatranscriptomic taxa entities we see they share 27 taxa entities such as Dinophyta and Chlorophyta, while the 18S rDNA dataset contains 27 unique taxa entities such as Rotifera and the metatranscriptomic dataset contains 64 unique taxa entities such as Crenarchaeota. In figure 4.3 **b**, the comparison of the 16S rDNA taxa entities to the metatranscriptomic taxa entities we see they share 18 taxa entities such as Cyanobacteria and Proteobacteria, while the 16S rDNA dataset contains 13 unique taxa entities such as Balneolaeota and the metatranscriptomic dataset contain 56 unique taxa entities such as Ignavibacteriae.

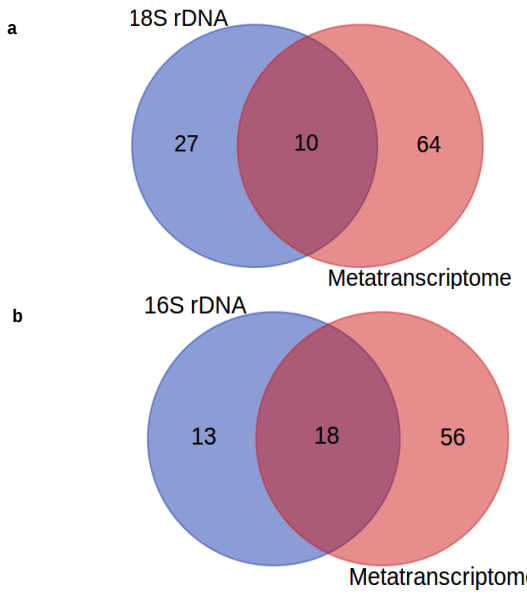

Figure 4.3: Panel **a** and **b** represents Venn diagrams of the number of taxa entities in the 18S, 16S rDNA and metatranscriptome datasets at the taxonomic rank of phylum. Panel **a** is comparing the 18S rDNA dataset and metatranscriptome dataset. Panel **b** is comparing the 16S rDNA dataset and metatranscriptome dataset. (Figure was generated with http://bioinformatics.psb.ugent.be/webtools/Venn/)

4.4.2 Rarefaction curves

The rarefaction curves in figure 4.4 are based on the Pfam protein families. These curves were generated using the rarecurve function, which is part of the VEGAN package, in R. The numbers displayed in the plot correspond to sample location, as shown in figure 3.1**a**. As outlined in section 3.4, rarefaction curves enable us to investigate if we have sufficient data to justify the results of our analyses.

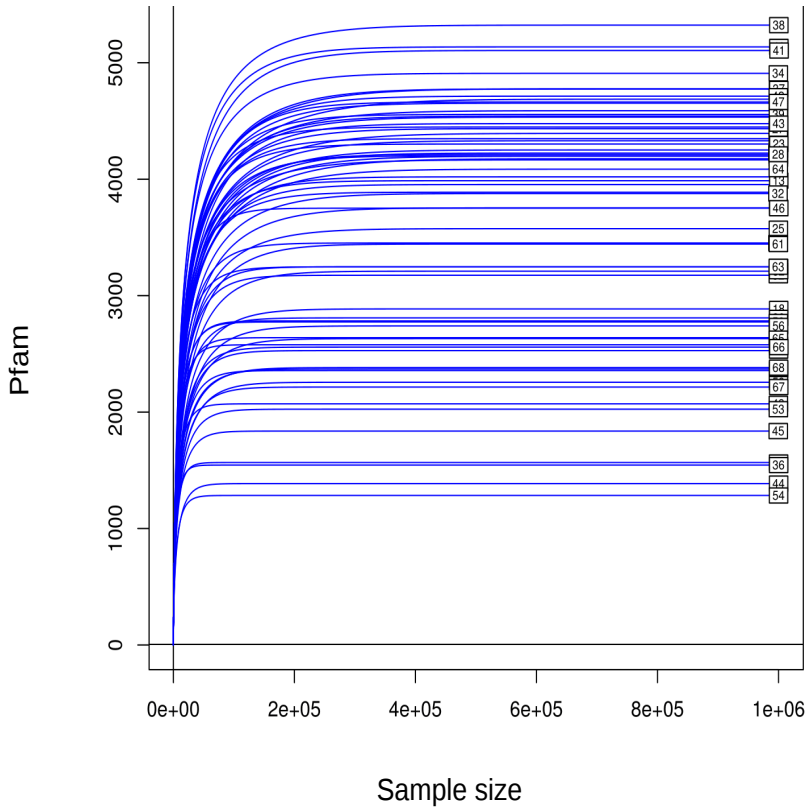

Figure 4.4: Rarefaction curves for Pfam protein families. The numbers displayed in the plot correspond to sample location. The x-axis represents the random subsample size taken from the dataset and the y-axis indicates the number of unique Pfam protein families found. The R package VEGAN using the rarecurve function was employed to perform the rarefaction curves analysis

In figure 4.4, there is a sharp rise at first in all the curves for the 65 samples. The majority of the curves, for example samples 47, 64 and 25, rise more slowly as more rare species are added, and then the curves level off. There are a number of curves, for example samples 53, 66 and 45, which level off immediately. The levelling off of all of the curves happens quite quickly, therefore we conclude that sufficient sampling was achieved.

4.4.3 Principle Components Analysis (PCA)

The Principle Components Analysis (PCA) in figure 4.5 is based on the Pfam protein families (log10 transformed) gene counts. This PCA was generated using the princomp function, which is part of the VEGAN package, in R. The numbers and colours displayed in the plot correspond to sample location, as shown in figure 3.1**a**.

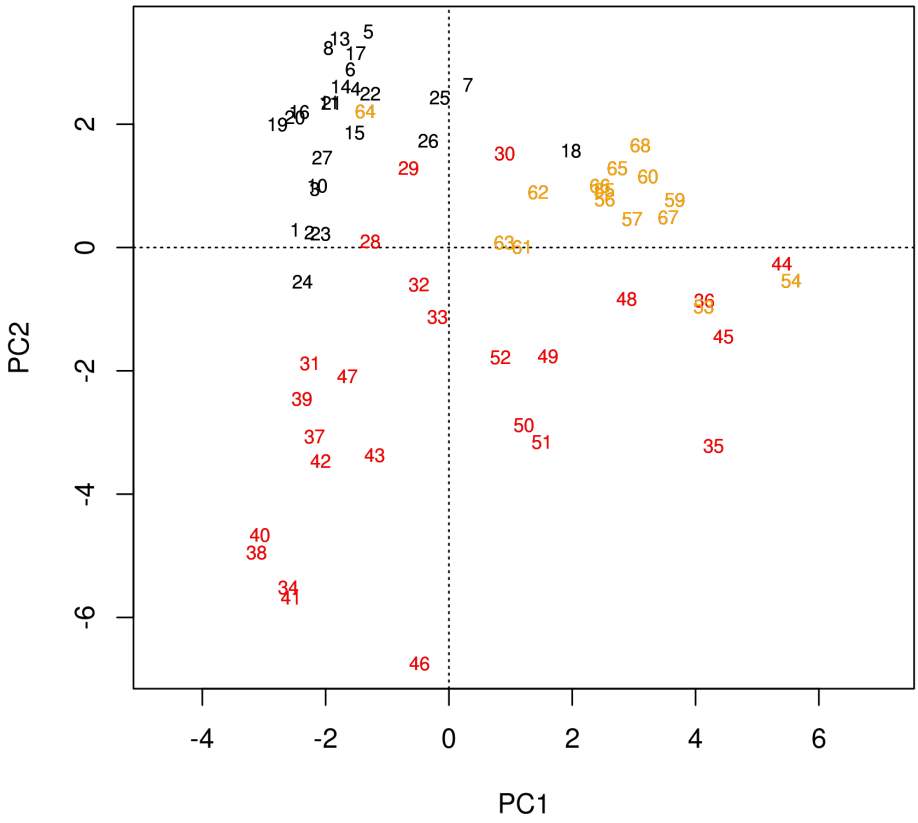

Figure 4.5: A Principle Components Analysis (PCA) for the Pfam protein families (log10 transformed) gene counts. The samples are numbered and coloured by region; these numbers and colours correspond to the map of the region sampling sites in figure 3.1**a**, where the Arctic Ocean samples are coloured black, the North Atlantic Ocean samples are coloured red and the South Atlantic Ocean samples are coloured yellow. The R package VEGAN using the princomp function was employed to perform the PCA

PCA is one of the most popular methods for exploratory analyses, as it is a simple visualization tool to summarize dataset variance [Ramette, 2007], [Paliy and Shankar, 2016]. The general principle of PCA is to calculate new synthetic variables called principal components. Principal components are linear combinations of the initial variables and these attempts to account for as much of the variance of the initial dataset

as possible. The objective is to represent the objects and variables of the dataset in a new system of coordinates. This is generally on two axes where the maximum amount of variation from the initial dataset can be displayed [Ramette, 2007]. The largest gradient of variability in the dataset is represented by the first principal component axis of the PCA. The second principal component represents the second largest and so on, till all the dataset variability has been accounted for [Paliy and Shankar, 2016].

Displayed in figure 4.5 is the Pfam protein families (log10 transformed) gene counts for the first two principal components. In the PCA plot the samples are numbered and coloured by region; these numbers and colours correspond to the map of the region sampling sites in figure 3.1**a**, where the Arctic Ocean samples are coloured black, the North Atlantic Ocean samples are coloured red and the South Atlantic Ocean samples are coloured yellow.

The Pfam protein families dataset is derived from the metatranscriptomic dataset. The functional analysis which generated the Pfam protein families dataset is outlined in Section 4.4.2. The activity of the organisms within each sample is reflected in the functional composition of transcripts, any changes may indicate a metabolic response to conditions [Klingenberg and Meinicke, 2017]. Therefore the Pfam protein families dataset represented in figure 4.5 is a based on similarity of activity level.

The Pfam protein families dataset displayed in figure 4.5 has PC1 accounting for 26.97% of sample variation, while PC2 accounts for 8.4% of sample variation. Overall the Pfam protein families samples are to a reasonable extent clustering well by region, as the matching colours are grouped together. Also for the Pfam protein families samples, there is a general transition of the samples from black to red to yellow, which coincides with how the samples are positioned by latitude as can be seen in figure 3.1**a**.

4.4.4 Canonical Correspondence Analysis (CCA)

In figure 4.6 we present a CCA for the Pfam protein families dataset. The numbers in the CCA plot correspond to sample site numbers as shown in figure 3.1**a**. The y-axis represents the CCA2, the x-axis represents the CCA1, and the arrows represent the direction and the length of the vectors for the environmental variables. A CCA enables us to find the relationship between the Pfam protein families and the environmental

variables, and therefore enable us to determine how the environmental variables might determine the response variable values of the Pfam protein families [Paliy and Shankar, 2016].

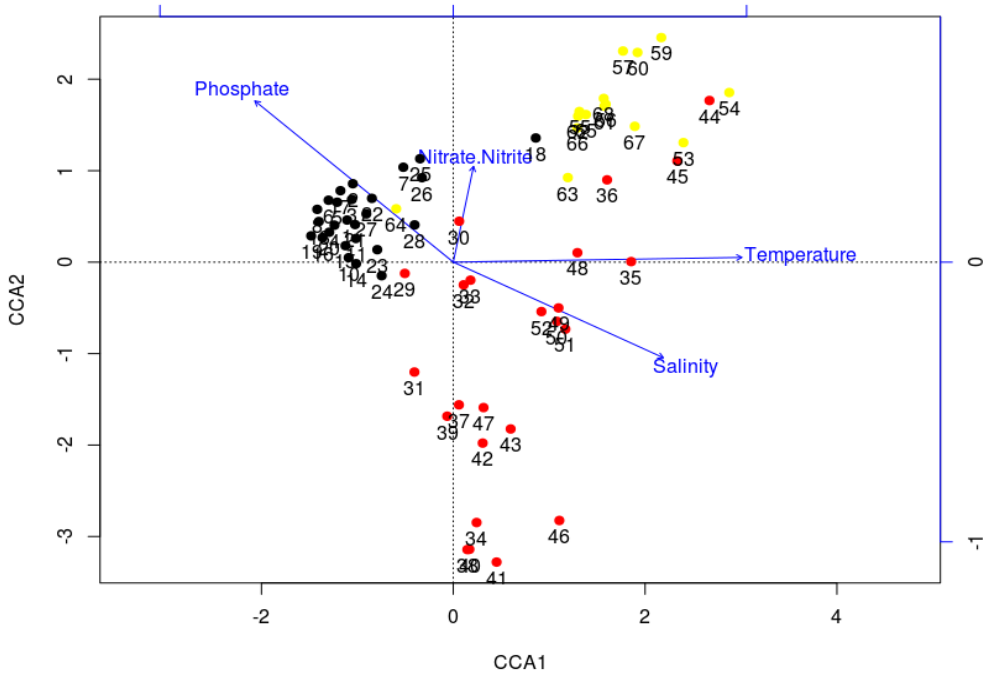

Figure 4.6: A CCA for the Pfam protein families. The numbers in the plots correspond to sample locations as given in figure 3.1**a**. The numbers are coloured by region, red is for the North Atlantic Ocean, black is for the Arctic Ocean and yellow is for South Atlantic Ocean. We used the R package VEGAN to perform a Canonical Correspondence Analysis (CCA) between the Pfam dataset and the environmental data. The y-axis represents the CCA2 and the x-axis represents the CCA1. The arrows represent the direction and the length of the vector. Each vector represents an environmental factor variable

CCA captured 13.3% of the total variability within the dataset. CCA1 accounts for approximately 45.7% of the constrained variability, with CCA2 accounting for 28.9%, CCA3 accounting for 12.8% and CCA4 accounting for 12.6%. In figure 4.6, the first axis CCA1 is associated with increasing temperature, while the second axis CCA2 is associated with decreasing salinity, increasing nitrate/nitrite and increasing phosphate. The samples are plotted in relation to the arrows, indicating how they are influenced by these environmental variables. The samples from the tropical region of the South Atlantic Ocean, plotted on the right hand side of the figure are strongly influenced by temperature. The samples from the polar region of the Arctic Ocean, plotted on the left hand side of the figure, are poorly influenced by temperature.

4.4.5 Breakpoint analysis

In figure 4.7 we present a breakpoint analysis for the Pfam protein families dataset. The breakpoint analysis was generated using piecewise regression in R as outlined in section 3.3.3. The numbers in the plot correspond to sample location as shown in figure 3.1**a**. The y-axis represents the beta diversity across the stations, the x-axis represents the temperature and the horizontal line marks the breakpoint. The breakpoint analysis enabled us to investigate how increasing temperature affects the changing diversity of the Pfam protein families from the polar Arctic Ocean through the temperate North Atlantic Ocean and down to the tropical South Atlantic Ocean.

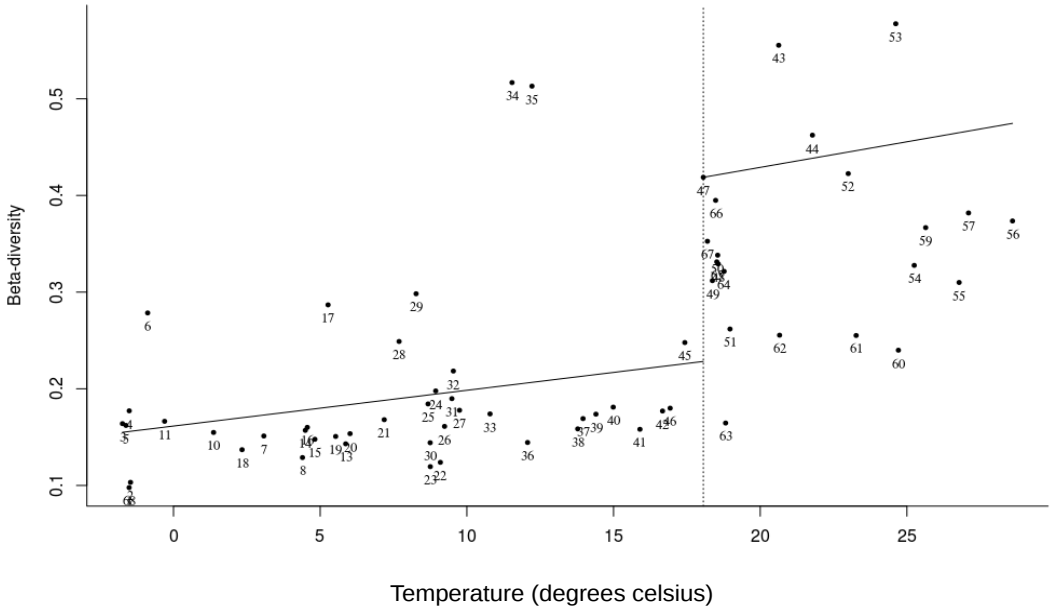

Figure 4.7: A breakpoint analysis for the Pfam protein families. The numbers in the plots correspond to sample locations as given in figure 3.1**a**. The breakpoint analysis was generated using piecewise regression in R as outlined in section 3.3.3. The y-axis represents the beta diversity across the stations. The x-axis represents the temperature. In the plot, the horizontal line marks the breakpoint. The Pfam protein families breakpoint is 18.06°C with a p-value of 1.24e-07

In figure 4.7, the left hand side of the plot represents the lower temperatures containing sample sites from the Arctic Ocean and North Atlantic Ocean. As we move across to the right hand side of the plot, which represents the higher temperatures, we move to sample sites from the North Atlantic Ocean and South Atlantic Ocean. The breakpoint for the Pfam protein families was determined to be 18.06°C with a p-value of 1.24e-07. There is a clear shift in the diversity for the Pfam protein families at the

breakpoint. This shift in diversity occurs around the samples located in the temperate region of the North Atlantic Ocean just before we move into the tropical region of the South Atlantic Ocean.

The location of the Pfam protein families breakpoint in the North Atlantic Ocean is consistent with the breakpoint results for the 18S and 16S rDNA datasets. These were also located in the North Atlantic Ocean, with breakpoints determined at 13.96°C (p-value of 3.121e-06) and 9.49°C (p-value of 8.114e-03), respectively (see chapter 3).

In figure 4.2 the heatmap of taxonomically classified sequences belonging to the metatranscriptomic dataset, we observed no gradient of increasing diversity of taxa across the samples as we move from the polar Arctic Ocean through the temperate North Atlantic Ocean and into the tropical South Atlantic Ocean. In figure 4.7 the breakpoint analysis which is based on the Pfam protein families dataset, we observed changes in the diversity of the Pfam as we move across the polar Arctic Ocean through the temperate North Atlantic Ocean and down to the tropical South Atlantic Ocean. The Pfam protein families dataset is derived from the metatranscriptomic dataset. The functional analysis which generated the Pfam protein families dataset is outlined in Section 4.3.2. The activity of the organisms within each sample is reflected in the functional composition of transcripts, any changes may indicate a metabolic response to conditions [Klingenberg and Meinicke, 2017].

4.4.6 Co-occurrence analysis

In our co-occurrence analysis using WGCNA on the Pfam protein family (log10 transformed) gene counts, thirteen modules (networks) were found and a grey module which represents those protein families that could not be assigned to a module. We call the modules black (n=174), blue (n=547), brown (n=515), cyan (n=83), green (n=403), greenyellow (n=116), pink (n=162), purple (n=132), red (n=205), salmon (n=85), tan (n=100), turquoise (n=768), yellow (n=264) and grey (n=7).

In figure 4.8 we present a correlation heatmap generated between each module's eigengene and environmental parameters. A number of modules were highly correlated, either positively or negatively with the environmental variables. For example, for temperature, the highly correlated modules are tan, blue, turquoise, yellow and pink.

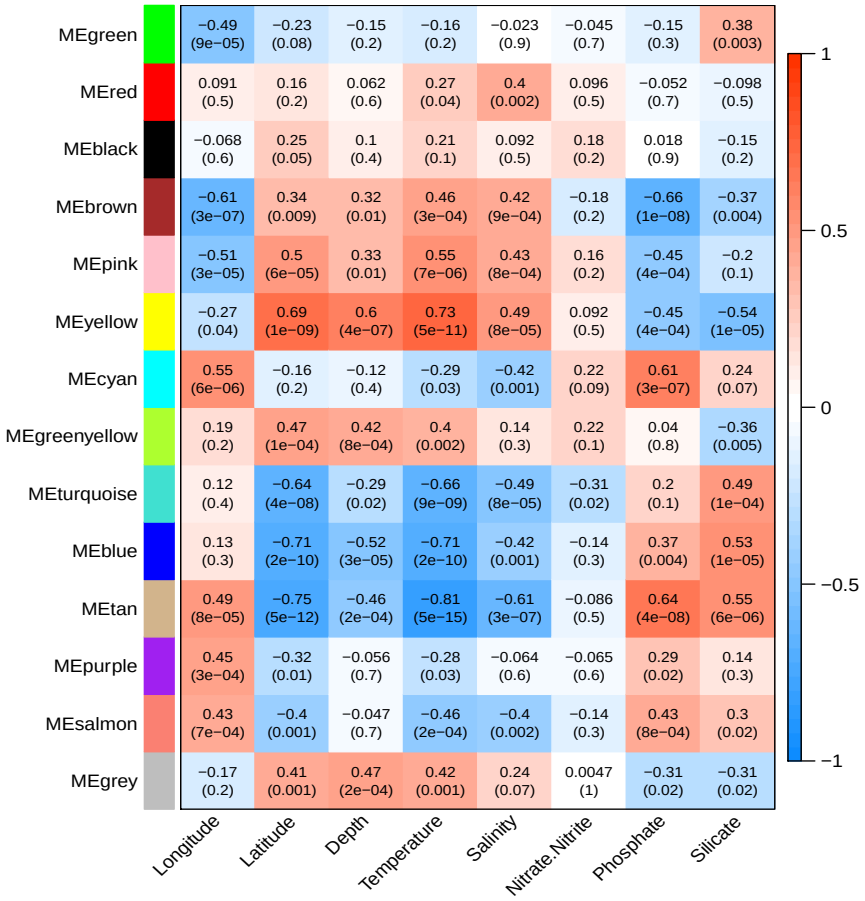

Figure 4.8: In the WGCNA analysis of the log10-scaled gene counts of Pfam protein families, thirteen modules (and a grey module) were found. In the figure, we present a correlation heatmap for the modules. The fourteen modules are displayed as coloured blocks labelled along the left hand side of the plot. The environmental parameters are displayed at the bottom. The colours correspond to the correlation values, red is positively correlated and blue is negatively correlated. The values in each of the squares correspond to the assigned Pearson correlation coefficient value on top and p-value in brackets below

For salinity only the tan module was highly correlated. For phosphate, the highly correlated modules are tan, cyan and brown. For silicate, the highly correlated modules are tan, blue and yellow. No modules correlated significantly with nitrate/nitrite.

The module tan (n=100) will be an ideal module for further analysis as it had the highest correlation to the environmental variables of temperature, phosphate and salinity as shown in figure 4.8.

4.5 Discussion

In this chapter, we have presented a metatranscriptomic analysis. We described the dataset and the methodology of the metatranscriptomic analysis, including heatmaps, rarefaction curves, canonical correspondence analysis, breakpoint analysis and co-

occurrence analysis. This was a unique large-scale examination of the protein composition of the dataset, taken from a transect of the Arctic and Atlantic oceans, the sampling of which is described in section 3.2.1. This has given us new insights into what these marine microbial communities are probably doing in response to environmental conditions.

From each sample metatranscriptomic, metagenomic, 18S and 16S rDNA sequencing was performed. As explained in section 3.5, only a single sample was obtained at each station; we did not obtain replicate samples. In addition due to time constraints, a full and extensive analysis of the metatranscriptomic dataset was not performed. Further analysis is still required, as well as a more in-depth examination of the Pfam proteins families co-occurrence analysis.

For our metatranscriptomic analysis, we performed a CCA on the Pfam protein families dataset. With the CCA we captured 13.3% of the total variability in the Pfam protein families dataset, and of this CCA1 accounts for approximately 45.74% of the constrained variability. From the plot in figure 4.6 we identified CCA1 to have an association with increasing temperature for about half of the samples.

We performed a breakpoint analysis on our Pfam protein families dataset and determined the breakpoint to be 18.06°C with a p-value of 1.24e-07. This positions the breakpoint in the temperate region of the North Atlantic Ocean off the coast of France and is in agreement with our 18S and 16S rDNA datasets, as we identified breakpoint of 13.96°C for the 18S rDNA dataset and 9.49°C for the 16S rDNA dataset. These results indicate that as you move from the cold Arctic Ocean to the warm tropical regions in the South Atlantic Ocean there is a radical shift in the diversity of the 18S and 16S rDNA species communities and a radial shift in activity according to the Pfam protein families.

In our co-occurrence analysis with WGCNA on the Pfam protein families dataset, we found thirteen modules for further analysis. A number of modules were highly correlated either positively or negatively with the environmental variables but the tan module (n=100) had the highest correlations to the environmental variables (temperature, phosphate and salinity). For future work we will continue the examination of the thirteen modules, beginning with the tan module. For each environmental variable temperature, phosphate and salinity, we will plot gene significance against module

membership, to identify Pfam protein families in the tan module that have a high significance to that environmental variable and analyse them to understand their connection to that environmental variable.

Chapter 5

Discussion and future work

5.1 Summary

In chapter 3 we outlined the computational pipelines and analysis of the 18S rDNA and 16S rDNA datasets that were derived from samples collected from a transect of the Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean. Firstly, this involved constructing a computational pipeline for taxonomically classifying the 18S rDNA dataset. Then we devised a methodology to normalise the 18S and 16S rDNA copy number, in order to interpret the data in terms of species abundance rather than read counts and therefore conduct various analyses that also included the environmental data that was recorded during the expeditions.

From our analysis of the 18S rDNA and 16S rDNA datasets, we observed a greater diversity of microbes in the tropical regions of the South Atlantic Ocean, in comparison to the polar regions of the Arctic Ocean. From analyses that included environmental data, we identified temperature to be the driving force of diversity. Furthermore, a breakpoint analysis was performed on our 18S and 16S rDNA datasets in which we found a shift in diversity occurring in the temperate region of the North Atlantic Ocean, between the polar Arctic Ocean and tropical South Atlantic Ocean. In addition, from our co-occurrence analysis on the 18S and 16S rDNA datasets, we identified two community networks. Each of these networks was found to have a temperature preference, as one positively correlated to temperature and the other negatively correlated to temperature.

In chapter 4 we outlined the computational pipeline of the metatranscriptomic

dataset that was also derived from the samples that were collected from a transect of the Arctic Ocean, North Atlantic Ocean and South Atlantic Ocean. We also outlined the analyses of the Pfam protein families dataset. For our analysis of this dataset, we performed a canonical correspondence analysis (CCA) and we observed which environmental variables and by how much they explained the variation in our dataset. Furthermore, a breakpoint analysis was performed in which we found a shift in diversity occurring in the temperate region of the North Atlantic Ocean, between the polar Arctic Ocean and tropical South Atlantic Ocean. The Pfam protein families breakpoint is in the same region as that of the 18S and 16S rDNA datasets as described in chapter 3. In addition, from our co-occurrence analysis on the Pfam protein families dataset we identified thirteen networks for further analysis.

5.2 Future work

For future work on our 68 samples, we will continue the examination of our metatranscriptomic dataset. As mentioned in chapter 4, we performed a co-occurrence analysis on Pfam protein families and this resulted in several modules being found. These modules will be taken for further analysis and research. We aim to determine Pfam protein families' relationships with one another and how environmental factors may be affecting these relationships. Also, we want to perform a GO enrichment analysis which provides defined GO terms to genes. GO terms cover for the highest level cellular components, molecular functions, and biological processes, and at the lowest level, these can be assigned to genes when relevant. Then we will perform various analyses such as heatmaps in order to examine their composition and distribution.

Our 18S, 16S rDNA and metatranscriptomic analyses have given us new insights into microbial communities, but studies of metatranscriptomes have found that functional genes predicted from metagenomic studies are not necessarily expressed [Heintz-Buschart et al., 2016], [Narayanasamy et al., 2016]. Ideally, we therefore need to integrate metatranscriptomics with metagenomics analysis for a more conclusive link between the genetic potential and the actual phenotype *in situ* [Narayanasamy et al., 2016].

Bearing this in mind, from the 68 stations sampled, we have chosen 11 samples for

metagenomic analysis. These are samples 1, 7, 18, 21, 22 and 23 which come from the Arctic Ocean, sample 45 from the North Atlantic Ocean and samples 54, 57, 59 and 64 from the South Atlantic Ocean, as shown in figure 3.1**a**. These samples were selected based on our 18S rDNA analysis as we were attempting to avoid samples that contain large amounts of dinoflagellates, as they possess some of the largest nuclear genomes. These samples have been sequenced by JGI. From the metagenomic analysis, this should enable us to understand their genetic potential. These metagenomic samples will be analysed by taxonomic classification and functional analysis. They will then be compared to their corresponding metatranscriptomic samples.

In related work, Emma Langan, a PhD student in the Environmental Sciences department at UEA will be making an expedition to the Antarctic Ocean at the end of December 2018, just before the ocean begins to freeze. The ship will float with the ice during which time a number of samples will be collected periodically, thus enabling the *in situ* real-time sequencing on Oxford nanopore sequencing instrument of polar microbes. The aim of this study is to monitor polar microbes in relation to their composition, distribution and abundance, and therefore monitor how climate change and different oceanographic features may affect these polar microbes. This will directly build on the results presented in this thesis since we examined the composition, distribution and abundance of microbes in the polar regions of the Arctic Ocean, the temperate regions of the North Atlantic Ocean and the tropical regions of South Atlantic Ocean. Sequencing *in situ* overcomes a number of common problems for researchers such as samples degrading. Also, the sampling can be directed based on real-time results rather than best guess of where to sample. Furthermore, if high quality genome assemblies can be achieved from the nanopore long reads, they can be used to produce reference genomes, which will be a great benefit to the research community as there are currently only a few phytoplankton reference genomes available. This will also enable them to perform comparative genomic analysis between temperate and polar species, in order to find the evolutionary mechanisms for polar adaptations.

Recently, more samples have been collected and sequenced by JGI. These samples cover the gap in our data between Cape Town in the South Atlantic Ocean and the Antarctic Ocean. Metatranscriptomic and metagenomic analysis will be performed on these samples. The Antarctic Ocean has a seasonal temperature that rises above

0°C and therefore these polar microbes experience reduced kinetic energy that imposes constraints on cellular activity. But despite this, the Antarctic Ocean contributes high levels of microbial primary production. For example, it has been estimated that while the Antarctic Ocean represents only about 10% of the total surface area of the world's ocean, it contributes about 30% of the global ocean uptake of carbon dioxide. Also, the Antarctic Ocean forms a significant amount of the oceanic food web [Wilkins et al., 2013]. This study will, therefore, provide an important and interesting insight into how these polar microbes live and function and how they compare with the microbes we have analysed in this thesis.

An exciting possibility is that our data from our metatranscriptomic, 18S and 16S rDNA datasets may be used to generate models by T. M. Lenton from the University of Exeter. These models could potentially enable us to forecast how climate change will affect these microbial communities [Toseland et al., 2013].

5.3 Conclusions

High-throughput sequencing technology has enabled us to study species that cannot be grown in the laboratory. In the past, culturing a species was necessary in order to study that species, but due to the advances of high-throughput sequencing we can examine these species, study their expression and genetics and compare and contrast multiple samples in metatranscriptomic and metagenomic analyses. However, there are challenges with implementing these analyses.

The first difficulty we had was preparing the 18S rDNA datasets for analysis. In the NCBI taxonomy not every species has a taxonomy designation at every rank. This caused issues when trying to describe and analyse our data at the taxonomic rank of class. We overcame this problem by placing those that do not have a taxonomic rank of class into a new name to represent them at the class level. This situation occurs because it is possible that there is currently no assignment at a particular taxonomic rank that the species fits into or there is no agreement among scientists into which higher taxonomy a species should be placed. The creation of a temporary name until a permanent rank can be assigned would be one possible solution to this problem.

An important additional issue was caused by the 18S and 16S rDNA copy number.

The relationship between amplicon and species abundance is indeterminate due to rDNA copy number variation within the genomes of different species [Perisin et al., 2016]. While there is a 16S rDNA database available, it is not very large. The 18S rDNA gene has no database available at all. NCBI could provide a solution to this, by creating an 18S rDNA database. NCBI actively collects all kinds of information and would be well positioned to accomplish this.

We were not able to taxonomically classify about a third of our 18S rDNA dataset. We lost a considerable amount of information due to the fact that those species were not in the NCBI database. The 16S rDNA species are much better represented in databases such as SILVA. Only 3% of our 16S rDNA dataset could not be taxonomically classified. This is changing, as there is an increase in the number of research studies that are focusing on eukaryote species.

Another difficulty we had was due to the size of our datasets. We had difficulties with the time it took to analyse our datasets. This was very apparent when we tried to perform a GO enrichment analysis with the tools online such as WEGO 2.0 [Ye et al., 2018]. While there are a number of websites available, very few take multiple files such as in this study. It is therefore important to develop fast new bioinformatics tools to handle large datasets.

Despite these difficulties, we thoroughly analysed the metatranscriptomic, 18S and 16S rDNA datasets. In the surface ocean, we identified temperature to be the important environmental factor that is driving marine microbial diversity and affecting how these microbial communities interact with each other. While diversity increases with increasing temperature, the composition of the microbial communities appear to change once a threshold is reached. The breakpoint analyses of the 16S, 18S rDNA and metatranscriptomic datasets identified breakpoints to be 9.49°C, 13.96°C and 18.06°C, respectively. These breakpoints are located in the North Atlantic Ocean and this might be the location of the threshold for the microbial communities to change. Since microbial community composition on either side of the temperature threshold is different, this may have consequences in terms of how global warming will affect these interactions and therefore have implications for the biogeochemical cycle of elements. As more data is collected and analysed and our understanding of ocean microbes improves, it will be interesting to see how these predictions of our breakpoint analysis will work

out.

Bibliography

[Aird et al., 2011] Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2):R18.

[Alemzadeh et al., 2014] Alemzadeh, E., Haddad, R., and Ahmadi, A.-R. (2014). Phytoplanktons and DNA barcoding: Characterization and molecular analysis of phytoplanktons on the Persian Gulf. *Iranian Journal of Microbiology*, 6(4):296–302.

[Alexander et al., 2015] Alexander, H., Jenkins, B. D., Rynearson, T. A., and Dyhrman, S. T. (2015). Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proceedings of the National Academy of Sciences*, 112(17):E2182–E2190.

[Amit Roy et al., 2014] Amit Roy, S. R., Ray, S., and Roy, A. (2014). Molecular Markers in Phylogenetic Studies-A Review. *Journal of Phylogenetics & Evolutionary Biology*, 02(02):1–9.

[Anderson and Santana-Garcon, 2015] Anderson, M. J. and Santana-Garcon, J. (2015). Measures of precision for dissimilarity-based multivariate analysis of ecological communities. *Ecology Letters*, 18(1):66–73.

[Armbrust, 2009] Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature*, 459(7244):185–192.

[Aryal et al., 2015] Aryal, S., Karki, G., and Pandey, S. (2015). Microbial Diversity in Freshwater and Marine Environment. *Nepal Journal of Biotechnology*, 3(1):68.

[Balvočit and Huson, 2017] Balvočit, M. and Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT how do these taxonomies compare? *BMC Genomics*, 18(S2):114.

[Baron, 1996] Baron, S. (1996). *Introduction to Parasitology*. University of Texas Medical Branch at Galveston.

[Barta, 1997] Barta, J. R. (1997). Investigating Phylogenetic Relationships within the Apicomplexa Using Sequence Data: The Search for Homology. *Methods*, 13(2):81–88.

[Baselga and Orme, 2012] Baselga, A. and Orme, C. D. L. (2012). betapart : an R package for the study of beta diversity. *Methods in Ecology and Evolution*, 3(5):808–812.

[Bazinet and Cummings, 2012] Bazinet, A. L. and Cummings, M. P. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13(1):92.

[Behjati and Tarpey, 2013] Behjati, S. and Tarpey, P. S. (2013). What is next generation sequencing? *Archives of disease in childhood. Education and practice edition*, 98(6):236–8.

[Bennett, 2004] Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, 5(4):433–438.

[Bijlsma and Loeschcke, 2005] Bijlsma, R. and Loeschcke, V. (2005). Environmental stress, adaptation and evolution: an overview. *Journal of Evolutionary Biology*, 18(4):744–749.

[Bik, 2014] Bik, H. M. (2014). Deciphering diversity and ecological function from marine metagenomes. *The Biological Bulletin*, 227(2):107–16.

[Boon et al., 2014] Boon, E., Meehan, C. J., Whidden, C., Wong, D. H.-J., Langille, M. G. I., and Beiko, R. G. (2014). Interactions in the microbiome: communities of organisms and communities of genes. *FEMS Microbiology Reviews*, 38(1):90–118.

[Bork et al., 2015] Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara Oceans studies plankton at planetary scale. *Science*, 348(6237):873–873.

[Boucher et al., 1991] Boucher, N., Vaulot, D., and Partensky, F. (1991). Flow cytometric determination of phytoplankton DNA in cultures and oceanic populations. *Marine Ecology Progress Series*, 71(1):75–84.

[Brady and Salzberg, 2011] Brady, A. and Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nature Methods*, 8(5):367.

[Brady and Salzberg, 2009a] Brady, A. and Salzberg, S. L. (2009a). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, 6(9):673–6.

[Brady and Salzberg, 2009b] Brady, A. and Salzberg, S. L. (2009b). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, 6.

[Brewer et al., 2019] Brewer, T. E., Albertsen, M., Edwards, A., Kirkegaard, R. H., Rocha, E. P. C., and Fierer, N. (2019). Unlinked rRNA genes are widespread among Bacteria and Archaea. *bioRxiv*, page 705046.

[Brierley, 2017] Brierley, A. S. (2017). Plankton. *Current Biology*, 27(11):R478–R483.

[Brown, 2014] Brown, J. H. (2014). Why are there so many species in the tropics? *Journal of Biogeography*, 41(1):8–22.

[Brown, 2002] Brown, T. A. (2002). *Genomes. 2nd edition.* Wiley-Liss.

[Brussaard et al., 2016] Brussaard, C. P. D., Bidle, K. D., Pedrós-Alió, C., and Legrand, C. (2016). The interactive microbial ocean. *Nature Microbiology*, 2(1):16255.

[Buermans and den Dunnen, 2014] Buermans, H. and den Dunnen, J. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10):1932–1941.

[Campra and Morales, 2016] Campra, P. and Morales, M. (2016). Trend analysis by a piecewise linear regression model applied to surface air temperatures in Southeastern Spain. (May):1–25.

[Capella-Gutierrez et al., 2009] Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

[Carlton et al., 2013] Carlton, J. M., Perkins, S. L., and Deitsch, K. W. (2013). *Malaria parasites : comparative genomics, evolution and molecular biology.*

[Carradec et al., 2018] Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M. B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler, C., and Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1):373.

[Carter and Hussain, 2017] Carter, J.-M. and Hussain, S. (2017). Robust long-read native DNA sequencing using the ONT CsgG Nanopore system. *Wellcome Open Research*, 2:23.

[Castro-Insua et al., 2016] Castro-Insua, A., Gómez-Rodríguez, C., and Baselga, A. (2016). Break the pattern: breakpoints in beta diversity of vertebrates are general across clades and suggest common historical causes. *Global Ecology and Biogeography*, 25(11):1279–1283.

[Chapman et al., 2018] Chapman, A. S. A., Tunnicliffe, V., and Bates, A. E. (2018). Both rare and common species make unique contributions to functional diversity in an ecosystem unaffected by human activities. *Diversity and Distributions*, 24(5):568–578.

[Collins et al., 2014] Collins, S., Rost, B., and Rynearson, T. A. (2014). Evolutionary potential of marine phytoplankton under ocean acidification. *Evolutionary Applications*, 7(1):140–155.

[Cooper, 2000] Cooper, G. M. (2000). *The Cell: A Molecular Approach.* Sinauer Associates.

[Cordero and Datta, 2016] Cordero, O. X. and Datta, M. S. (2016). Microbial interactions and community assembly at microscales. *Current Opinion in Microbiology*, 31:227–234.

[Das et al., 2006] Das, S., Lyla, P. S., and Khan, S. A. (2006). Marine microbial diversity and ecology: importance and future perspectives.

[de Vargas et al., 2015] de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., Boss, E., Follows, M., Karp-Boss, L., Krzic, U., Reynaud, E. G., Sardet, C., Sullivan, M. B., and Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237):1261605–1261605.

[DOE Joint Genome Institute, 2017] DOE Joint Genome Institute (2017). BBDuk Guide - DOE Joint Genome Institute.

[Dyomin et al., 2016] Dyomin, A. G., Koshel, E. I., Kiselev, A. M., Saifitdinova, A. F., Galkina, S. A., Fukagawa, T., Kostareva, A. A., and Gaginskaya, E. R. (2016). Chicken rRNA Gene Cluster Structure. *PLOS ONE*, 11(6):e0157464.

[Eddy, 1996] Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology*, 6(3):361–365.

[Eddy, 2004] Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, 22(10):1315–1316.

[Eddy, 2009] Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome informatics. International Conference on Genome Informatics*, 23(1):205–11.

[Edgar, 2010a] Edgar, R. (2010a). *USEARCH cluster otus.* https://www.drive5.com/usearch/manual/cmd_cluster_otus.html.

[Edgar, 2010b] Edgar, R. C. (2010b). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461.

[Edgar, 2010c] Edgar, R. C. (2010c). *UPARSE-OTU algorithm.*

[Edgar, 2010d] Edgar, R. C. (2010d). *UPARSE-REF algorithm.*

[Edgar, 2010e] Edgar, R. C. (2010e). *USEARCH Merging paired reads.* https://www.drive5.com/usearch/manual/merge_pair.html.

[Edgar, 2010f] Edgar, R. C. (2010f). *USEARCH oligodb.* https://www.drive5.com/usearch/manual/cmd_search_oligodb.html.

[Edgar, 2016] Edgar, R. C. (2016). UCHIME2: improved chimera prediction for amplicon sequencing. *doi: 10.1101/074252.*

[Edgar, 2017] Edgar, R. C. (2017). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *bioRxiv*, page 192211.

[Eisenberg and Levanon, 2013] Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *Trends in Genetics : TIG*, 29(10):569–74.

[El-Gebali et al., 2019] El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C., and Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, 47(D1):D427–D432.

[Eric et al., 2014] Eric, S. D., Nicholas, T. K. D. D., and Theophilus, K. A. (2014). Bioinformatics with basic local alignment search tool (BLAST) and fast alignment (FASTA). *Journal of Bioinformatics and Sequence Analysis*, 6(1):1–6.

[Falkowski et al., 2008] Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879):1034–1039.

[Fernández-Pérez et al., 2018] Fernández-Pérez, J., Nantón, A., and Méndez, J. (2018). Sequence characterization of the 5S ribosomal DNA and the internal transcribed spacer (ITS) region in four European Donax species (Bivalvia: Donacidae). *BMC Genetics*, 19(1):97.

[Ferreira et al., 2014] Ferreira, M., Roma, N., and Russo, L. M. S. (2014). Cache-Oblivious parallel SIMD Viterbi decoding for sequence search in HMMER. *BMC Bioinformatics*, 15(1):165.

[Finn et al., 2011] Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue):W29–37.

[Finn et al., 2016] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285.

[Flombaum et al., 2013] Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K. W., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A., and Martiny, A. C. (2013). Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proceedings of the National Academy of Sciences of the United States of America*, 110(24):9824–9.

[Foster et al., 2011] Foster, R. A., Kuypers, M. M. M., Vagner, T., Paerl, R. W., Musat, N., and Zehr, J. P. (2011). Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. *The ISME Journal*, 5(9):1484–93.

[Franco et al., 2017] Franco, D. C., Signori, C. N., Duarte, R. T. D., Nakayama, C. R., Campos, L. S., and Pellizari, V. H. (2017). High Prevalence of Gammaproteobacteria in the Sediments of Admiralty Bay and North Bransfield Basin, Northwestern Antarctic Peninsula. *Frontiers in Microbiology*, 8:153.

[Fu et al., 2012] Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

[Fu and Gong, 2017] Fu, R. and Gong, J. (2017). Single Cell Analysis Linking Ribosomal (r)DNA and rRNA Copy Numbers to Cell Size and Growth Rate Provides Insights into Molecular Protistan Ecology. *Journal of Eukaryotic Microbiology*, 64(6):885–896.

[Gibbons et al., 2014] Gibbons, J. G., Branco, A. T., Yu, S., and Lemos, B. (2014). Ribosomal DNA copy number is coupled with gene expression variation and mitochondrial abundance in humans. *Nature Communications*, 5:1–12.

[Gilbert and Hughes, 2011] Gilbert, J. A. and Hughes, M. (2011). Gene Expression Profiling: Metatranscriptomics. In *Methods in Molecular Biology (Clifton, N.J.)*, volume 733, pages 195–205.

[Godfray and May, 2014] Godfray, H. C. J. and May, R. M. (2014). Open questions: are the dynamics of ecological communities predictable? *BMC Biology*, 12:22.

[Godhe et al., 2008] Godhe, A., Asplund, M. E., Härnström, K., Saravanan, V., Tyagi, A., and Karunasagar, I. (2008). Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Applied and Environmental Microbiology*, 74(23):7174–82.

[Goodrich et al., 2014] Goodrich, J. K., Di Rienzi, S. C., Poole, A. C., Koren, O., Walters, W. A., Caporaso, J. G., Knight, R., and Ley, R. E. (2014). Conducting a microbiome study. *Cell*, 158(2):250–262.

[Gough et al., 2001] Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov mod-

els that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4):903–919.

[Guiry, M.D. & Guiry, 2008] Guiry, M.D. & Guiry, G. (2008). AlgaeBase.

[Haque and Haque, 2017] Haque, S. Z. and Haque, M. (2017). The ecological community of commensal, symbiotic, and pathogenic gastrointestinal microorganisms - an appraisal. *Clinical and Experimental Gastroenterology*, 10:91–103.

[Hauser et al., 2010] Hauser, P. M., Burdet, F. X., Cissé, O. H., Keller, L., Taffé, P., Sanglard, D., Pagni, M., Jr, C. T., Limper, A., Jr, C. T., Limper, A., Davis, J., Fei, M., Huang, L., Wakefield, A., Demanche, C., Berthelemy, M., Petit, T., Polack, B., Wakefield, A., Wakefield, A., Stringer, J., Tamburrin, E., Dei-Cas, E., Gigliotti, F., Harmsen, A., Haidaris, C., Haidaris, P., Aliouat-Denis, C., Chabé, M., Demanche, C., El, M. A., Viscogliosi, E., Keely, S., Renauld, H., Wakefield, A., Cushion, M., Smulian, A., Kutty, G., Maldarelli, F., Achaz, G., Kovacs, J., Joffrion, T., Cushion, M., Kaneshiro, E., Rodrigues, M., Fonseca, A., Keeling, P., Fast, N., Law, J., Williams, B., Slamovits, C., Payne, S., Loomis, W., Gardner, M., Shallom, S., Carlton, J., Salzberg, S., Nene, V., Omsland, A., Cockrell, D., Howe, D., Fischer, E., Virtaneva, K., Ewann, F., Hoffman, P., Stanke, M., Schöffmann, O., Morgenstern, B., Waack, S., Sugiyama, J., Cushion, M., Smulian, A., Slaven, B., Sesterhenn, T., Arnold, J., Andersson, J., Andersson, S., Sakharkar, K., Dhar, P., Chow, V., Cushion, M., Corradi, N., Pombert, J., Farinelli, L., Didier, E., Keeling, P., Nahimana, A., Francioli, P., Blanc, D., Bille, J., Wakefield, A., Choi, M., Chung, B., Chung, Y., Yu, J., Cho, S., Ambrose, H., Keely, S., Aliouat, E., Dei-Cas, E., Wakefield, A., Basselin, M., Qiu, Y., Lipscomb, K., Kaneshiro, E., Reichard, U., Jousson, O., Monod, M., Atzori, C., Angeli, E., Mainini, A., Agoston, F., Micheli, V., Stringer, J., Cushion, M., Nowrousian, M., Nowrousian, M., Stajich, J., Chu, M., Engh, I., Espagne, E., Korf, I., Birney, E., Clamp, M., Durbin, R., Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y., Borodovsky, M., Cantarel, B., Korf, I., Robb, S., Parra, G., Ross, E., Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Hau, J., Muller, M., Pagni, M., Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Schnoes, A., Brown, S., Dodevski, I., and Babbitt, P. (2010). Comparative Ge-

nomics Suggests that the Fungal Pathogen Pneumocystis Is an Obligate Parasite Scavenging Amino Acids from Its Host's Lungs. *PLoS ONE*, 5(12):e15152.

[Heather and Chain, 2016] Heather, J. M. and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8.

[Heintz-Buschart et al., 2016] Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., Wampach, L., Schneider, J. G., Hogan, A., de Beaufort, C., and Wilmes, P. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology*, 2:16180.

[Holland and Bronstein, 2008] Holland, J. and Bronstein, J. (2008). Mutualism. *Encyclopedia of Ecology*, pages 2485–2491.

[Hou et al., 2017] Hou, D., Huang, Z., Zeng, S., Liu, J., Wei, D., Deng, X., Weng, S., He, Z., and He, J. (2017). Environmental Factors Shape Water Microbial Community Structure and Function in Shrimp Cultural Enclosure Ecosystems. *Frontiers in Microbiology*, 8:2359.

[Hugerth and Andersson, 2017] Hugerth, L. W. and Andersson, A. F. (2017). Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. *Frontiers in Microbiology*, 8:1561.

[Hughes et al., 2001] Hughes, J. B., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, 67(10):4399–406.

[Hunt et al., 2013] Hunt, D. E., Lin, Y., Church, M. J., Karl, D. M., Tringe, S. G., Izzo, L. K., and Johnson, Z. I. (2013). Relationship between abundance and specific activity of bacterioplankton in open ocean surface waters. *Applied and Environmental Microbiology*, 79(1):177–84.

[Huntemann et al., 2016] Huntemann, M., Ivanova, N. N., Mavromatis, K., Tripp, H. J., Paez-Espino, D., Tennessen, K., Palaniappan, K., Szeto, E., Pillay, M., Chen, I.-M. A., Pati, A., Nielsen, T., Markowitz, V. M., and Kyrpides, N. C. (2016). The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Standards in Genomic Sciences*, 11(1):17.

[Huson et al., 2007] Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research*, 17(3):377–86.

[Huson et al., 2011] Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9):1552–60.

[Huson et al., 2009] Huson, D. H., Richter, D. C., Mitra, S., Auch, A. F., and Schuster, S. C. (2009). Methods for comparative metagenomics. *BMC Bioinformatics*, 10(Suppl 1):S12.

[Iwen and Hinrichs, 2002] Iwen, P. C. and Hinrichs, S. H. (2002). Utilization of the internal transcribed spacer regions as molecular targets to detect and identify human fungal. (July 2001):87–109.

[Jain et al., 2016] Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1):239.

[Jiang et al., 2016a] Jiang, Y., Xiong, X., Danska, J., and Parkinson, J. (2016a). Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality. *Microbiome*, 4(1):2.

[Jiang et al., 2016b] Jiang, Y., Xiong, X., Danska, J., and Parkinson, J. (2016b). Metatranscriptomic analysis of diverse microbial communities reveals core metabolic pathways and microbiome-specific functionality. *Microbiome*, 4(1):2.

[Johnson and Burnet, 2016] Johnson, K. V.-A. and Burnet, P. W. J. (2016). Microbiome: Should we diversify from diversity? *Gut Microbes*, 7(6):455–458.

[Johnston, 2006] Johnston, J. S. (2006). Characteristics of the nuclear (18S, 5.8S, 28S and 5S) and mitochondrial (12S and 16S) rRNA genes of. *Insect Molecular Biology*, 15:657–686.

[Joint et al., 2010] Joint, I., Mühling, M., and Querellou, J. (2010). Culturing marine bacteria - an essential prerequisite for biodiscovery. *Microbial Biotechnology*, 3(5):564–75.

[Kalendar et al., 2017] Kalendar, R., Khassenov, B., Ramankulov, Y., Samuilova, O., and Ivanov, K. I. (2017). FastPCR: An in silico tool for fast primer and probe design and advanced sequence analysis. *Genomics*, 109(3-4):312–319.

[Kazamia et al., 2016] Kazamia, E., Helliwell, K. E., Purton, S., and Smith, A. G. (2016). How mutualisms arise in phytoplankton communities: building eco-evolutionary principles for aquatic microbes. *Ecology Letters*, 19(7):810–822.

[Keeling et al., 2014] Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. B., Schroeder, D. C., Simpson, A. G. B., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., and Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, 12(6):e1001889.

[Keightley and Johnson, 2004] Keightley, P. D. and Johnson, T. (2004). MCALIGN: stochastic alignment of noncoding DNA sequences based on an evolutionary model of sequence evolution. *Genome research*, 14(3):442–50.

[Kemena and Notredame, 2009] Kemena, C. and Notredame, C. (2009). Upcoming challenges for multiple sequence alignment methods in the high-throughput era. *Bioinformatics (Oxford, England)*, 25(19):2455–65.

[Klappenbach et al., 2001] Klappenbach, J. A., Saxman, P. R., Cole, J. R., and Schmidt, T. M. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research*, 29(1):181–4.

[Klingenberg and Meinicke, 2017] Klingenberg, H. and Meinicke, P. (2017). How to normalize metatranscriptomic count data for differential expression analysis. *PeerJ*, 5:e3859.

[Kozarewa et al., 2009] Kozarewa, I., Ning, Z., Quail, M. A., Sanders, M. J., Berriman, M., and Turner, D. J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, 6(4):291–5.

[Lambrechts et al., 2019] Lambrechts, S., Willems, A., and Tahon, G. (2019). Uncovering the Uncultivated Majority in Antarctic Soils: Toward a Synergistic Approach. *Frontiers in Microbiology*, 10:242.

[Langfelder and Horvath, 2007] Langfelder, P. and Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology*, 1:54.

[Langfelder et al., 2008] Langfelder, P., Horvath, S., Fisher, R., Zhou, X., Kao, M., Wong, W., Steffen, M., Petti, A., Aach, J., D'haeseleer, P., Church, G., Stuart, J., Segal, E., Koller, D., Kim, S., Zhang, B., Horvath, S., Carey, V., Gentry, J., Whalen, E., Gentleman, R., Schaefer, J., Strimmer, K., Chuang, C., Jen, C., Chen, C., Shieh, G., Cokus, S., Rose, S., Haynor, D., Gronbech-Jensen, N., Pellegrini, M., Horvath, S., Zhang, B., Carlson, M., Lu, K., Zhu, S., Felciano, R., Laurance, M., Zhao, W., Shu, Q., Lee, Y., Scheck, A., Liau, L., Wu, H., Geschwind, D., Febbo, P., Kornblum, H., Cloughesy, T., Nelson, S., Mischel, P., Horvath, S., Dong, J., Langfelder, P., Horvath, S., Carlson, M., Zhang, B., Fang, Z., Horvath, S., Mishel, P., Nelson, S., Ghazalpour, A., Doss, S., Zhang, B., Plaisier, C., Wang, S., Schadt, E., Thomas, A., Drake, T., Lusis, A., Horvath, S., Fuller, T., Ghazalpour, A., Aten, J., Drake, T., Lusis, A., Horvath, S., Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G., Gunnarsdottir, S., Mouy, M., Steinthorsdottir, V., Eiriksdottir, G., Bjornsdottir, G., Reynisdottir, I.,

Gudbjartsson, D., Helgadottir, A., Jonasdottir, A., Jonasdottir, A., Styrkarsdottir, U., Gretarsdottir, S., Magnusson, K., Stefansson, H., Fossdal, R., Kristjansson, K., Gislason, H., Stefansson, T., Leifsson, B., Thorsteinsdottir, U., Lamb, J., Gulcher, M., null Reitman, Kong, A., Schadt, E., Stefansson, K., van Nas, A., Guhathakurta, D., Wang, S., Yehya, S., Horvath, S., Zhang, B., Drake, L. I., Chaudhuri, G., Schadt, E., Drake, T., Arnold, A., Lusis, A., Oldham, M., Horvath, S., Geschwind, D., Miller, J., Oldham, M., Geschwind, D., Oldham, M., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., Geschwind, D., Keller, M., Choi, Y., Wang, P., Davis, D. B., Rabaglia, M., Oler, A., Stapleton, D., Argmann, C., Schueler, K., Edwards, S., Steinberg, H., Neto, E. C., Kleinhanz, R., Turner, S., Hellerstein, M., Schadt, E., Yandell, B., Kendziorski, C., Attie, A., Presson, A., Sobel, E., Papp, J., Suarez, C., Whistler, T., Rajeevan, M., Vernon, S., Horvath, S., Weston, D., Gunter, L., Rogers, A., Wullschleger, S., Wilcox, R., Yip, A., Horvath, S., Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., Barabási, A., Li, A., Horvath, S., Kaufman, L., Rousseeuw, P., Langfelder, P., Zhang, B., Horvath, S., Dudoit, S., Fridlyand, J., Hastie, T., Tibshirani, R., Sherlock, G., Eisen, M., Brown, P., Botstein, D., Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R., Dong, J., Horvath, S., Watts, D., Strogatz, S., Dudoit, S., Yang, Y., Callow, M., Speed, T., Hu, Z., Snitkin, E., DeLisi, C., Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., Frohlich, H., Speer, N., Poustka, A., BeiSZbarth, T., Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H., Lempicki, R., Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G., Zhang, B., Kirov, S., Snoddy, J., Liu, M., Liberzon, A., Kong, S., Lai, W., Park, P., Kohane, I., Kasif, S., Henegar, C., Clement, K., Zucker, J., Gentleman, R., Huber, W., Carey, V., Irizarry, R., Dudoit, S., Opgen-Rhein, R., Strimmer, K., Aten, J., Fuller, T., Lusis, A., Horvath, S., Neto, E. C., Ferrara, C., Attie, A., and Yandell, B. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

[Laver et al., 2015] Laver, T., Harrison, J., O'Neill, P., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D. (2015). Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3:1–8.

[Leimena et al., 2013] Leimena, M. M., Ramiro-Garcia, J., Davids, M., van den Bogert, B., Smidt, H., Smid, E. J., Boekhorst, J., Zoetendal, E. G., Schaap, P. J., and Kleerebezem, M. (2013). A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC Genomics*, 14(1):530.

[Letunic and Bork, 2007] Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–128.

[Li et al., 2011] Li, M., Copeland, A., and Han, J. (2011). DUK A Fast and Efficient Kmer Matching Tool DUK A Fast and Efficient Kmer Based Sequence Matching Tool.

[Li et al., 2012] Li, W., Fu, L., Niu, B., Wu, S., and Wooley, J. (2012). Ultrafast clustering algorithms for metagenomic sequence analysis. *Briefings in Bioinformatics*, 13(6):656–68.

[Li and Godzik, 2006] Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

[Lin et al., 2012] Lin, Y.-C., Campbell, T., Chung, C.-C., Gong, G.-C., Chiang, K.-P., and Worden, A. Z. (2012). Distribution patterns and phylogeny of marine stramenopiles in the north pacific ocean. *Applied and Environmental Microbiology*, 78(9):3387–99.

[Liu et al., 2012] Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012:1–11.

[Lu et al., 2016] Lu, H., Giordano, F., and Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, 14(5):265–279.

[MacArthur and MacArthur, 1961] MacArthur, R. H. and MacArthur, J. W. (1961). On Bird Species Diversity. *Ecology*, 42(3):594–598.

[Magoč and Salzberg, 2011] Magoč, T. and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics (Oxford, England)*, 27(21):2957–63.

[Magoc and Salzberg, 2011] Magoc, T. and Salzberg, S. L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963.

[Mandal, 1984] Mandal, R. K. (1984). The Organization and Transcription of Eukaryotic Ribosomal RNA Genes. *Progress in Nucleic Acid Research and Molecular Biology*, 31:115–160.

[Mande et al., 2012] Mande, S. S., Mohammed, M. H., and Ghosh, T. S. (2012). Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics*, 13(6):669–681.

[Mapleson et al., 2017] Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., and Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics (Oxford, England)*, 33(4):574–576.

[Marcon et al., 2014] Marcon, E., Scotti, I., Hérault, B., Rossi, V., and Lang, G. (2014). Generalization of the partitioning of shannon diversity. *PloS one*, 9(3):e90289.

[Martin et al., 2010] Martin, J., Bruno, V. M., Fang, Z., Meng, X., Blow, M., Zhang, T., Sherlock, G., Snyder, M., and Wang, Z. (2010). Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11(1):663.

[Martin, 2011] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *journal.embnet.org*, 17.

[Matsen et al., 2012] Matsen, F. A., Hoffman, N. G., Gallagher, A., and Stamatakis, A. (2012). A Format for Phylogenetic Placements. *PLoS ONE*, 7(2):e31009.

[Matsen et al., 2010] Matsen, F. A., Kodner, R. B., and Armbrust, E. V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1):538.

[Maxam and Gilbert, 1977] Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2):560–4.

[Medlar et al., 2014] Medlar, A., Aivelo, T., and Löytynoja, A. (2014). S{é}ance: reference-based phylogenetic analysis for 18S rRNA studies. *BMC Evolutionary Biology*, 14(1):235.

[Mezhoud et al., 2014] Mezhoud, N., Zili, F., Bouzidi, N., Helaoui, F., Ammar, J., and Ouada, H. B. (2014). The effects of temperature and light intensity on growth, reproduction and EPS synthesis of a thermophilic strain related to the genus Graesiella. *Bioprocess and Biosystems Engineering*, 37(11):2271–2280.

[Mitra et al., 2011] Mitra, S., Rupek, P., Richter, D. C., Urich, T., Gilbert, J. A., Meyer, F., Wilke, A., and Huson, D. H. (2011). Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics*, 12 Suppl 1(Suppl 1):S21.

[Moreau et al., 2012] Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., Van Bel, M., Poulain, J., Katinka, M., Hohmann-Marriott, M. F., Piganeau, G., Rouzé, P., Da Silva, C., Wincker, P., de Peer, Y., and Vandepoele, K. (2012). Gene functionalities and genome structure in Bathycoccus prasinos reflect cellular specializations at the base of the green lineage. *Genome Biology*, 13(8):R74.

[Morris et al., 2014] Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Müller, C., Obermaier, E., Prati, D., Socher, S. A., Sonnemann, I., Wäschke, N., Wubet, T., Wurst, S., and Rillig, M. C. (2014). Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution*, 4(18):3514–24.

[Nabhan and Sarkar, 2011] Nabhan, A. R. and Sarkar, I. N. (2011). The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics*, 13(1):122–34.

[Narayanasamy et al., 2016] Narayanasamy, S., Jarosz, Y., Muller, E. E. L., Heintz-Buschart, A., Herold, M., Kaysen, A., Laczny, C. C., Pinel, N., May, P., and Wilmes, P. (2016). IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biology*, 17(1):260.

[NCBI Resource Coordinators, 2016] NCBI Resource Coordinators, N. R. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(D1):D7–19.

[Nordberg et al., 2014] Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., and Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, 42(D1):D26–D31.

[Okolodkov and Dodge, 1996] Okolodkov, Y. B. and Dodge, J. D. (1996). Biodiversity and biogeography of planktonic dinoflagellates in the Arctic Ocean. *Journal of Experimental Marine Biology and Ecology*, 202(1):19–27.

[Oliver et al., 2007] Oliver, M. J., Petrov, D., Ackerly, D., Falkowski, P., and Schofield, O. M. (2007). The mode and tempo of genome size evolution in eukaryotes. *Genome Research*, 17(5):594–601.

[Paliy and Shankar, 2016] Paliy, O. and Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, 25(5):1032–57.

[Papadimitriou, 2014] Papadimitriou, C. (2014). Algorithms, complexity, and the sciences. *Proceedings of the National Academy of Sciences*, 111(45):15881–15887.

[Pavlopoulos et al., 2010] Pavlopoulos, G. A., Soldatos, T. G., Barbosa-Silva, A., and Schneider, R. (2010). A reference guide for tree analysis and visualization. *BioData Mining*, 3(1):1.

[Payne et al., 2005] Payne, L. X., Schindler, D. E., Parrish, J. K., and Temple, S. A. (2005). QUANTIFYING SPATIAL PATTERN WITH EVENNESS INDICES. *Ecological Applications*, 15(2):507–520.

[Pearson, 2013] Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, Chapter 3:Unit3.1.

[Perisin et al., 2016] Perisin, M., Vetter, M., Gilbert, J. A., and Bergelson, J. (2016). 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *The ISME Journal*, 10(4):1020–4.

[Pielou, 1966] Pielou, E. (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, 13:131–144.

[Pillai et al., 2017] Pillai, S., Gopalan, V., and Lam, A. K.-Y. (2017). Review of sequencing platforms and their applications in phaeochromocytoma and paragangliomas. *Critical Reviews in Oncology/Hematology*, 116:58–67.

[Pos et al., 2014] Pos, E., Guevara Andino, J. E., Sabatier, D., Molino, J.-F., Pitman, N., Mogollón, H., Neill, D., Cerón, C., Rivas, G., Di Fiore, A., Thomas, R., Tirado, M., Young, K. R., Wang, O., Sierra, R., García-Villacorta, R., Zagt, R., Palacios, W., Aulestia, M., and Ter Steege, H. (2014). Are all species necessary to reveal ecologically important patterns? *Ecology and Evolution*, 4(24):4626–36.

[Prokopowich et al., 2003] Prokopowich, C. D., Gregory, T. R., and Crease, T. J. (2003). The correlation between rDNA copy number and genome size in eukaryotes. *Genome*, 46(1):48–50.

[Quast et al., 2013a] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013a). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue):D590–6.

[Quast et al., 2013b] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013b). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue):D590—-6.

[Ramette, 2007] Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, 62(2):142–60.

[Reller et al., 2007] Reller, L. B., Weinstein, M. P., and Petti, C. A. (2007). Detection and Identification of Microorganisms by Gene Amplification and Sequencing. *Clinical Infectious Diseases*, 44(8):1108–1114.

[Reuter et al., 2015] Reuter, J. A., Spacek, D. V., and Snyder, M. P. (2015). High-throughput sequencing technologies. *Molecular cell*, 58(4):586–97.

[Ribeiro et al., 2013] Ribeiro, S., Berge, T., Lundholm, N., Ellegaard, M., and Richardson, B. (2013). Hundred Years of Environmental Change and Phytoplankton Ecophysiological Variability Archived in Coastal Sediments. *PLoS ONE*, 8(4):e61184.

[Ricotta, 2017] Ricotta, C. (2017). Of beta diversity, variance, evenness, and dissimilarity. *Ecology and Evolution*, 7(13):4835–4843.

[Ricotta and Podani, 2017] Ricotta, C. and Podani, J. (2017). On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecological Complexity*, 31:201–205.

[Robinson and Oshlack, 2010] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3):R25.

[Rödström, 2017] Rödström, E. M. (2017). *Skeletonema marinoi*. http://cemeb.science.gu.se/research/target-species-imago+/skeletonema-marinoi.

[Rokas, 2011] Rokas, A. (2011). Phylogenetic Analysis of Protein Sequence Data Using the Randomized Axelerated Maximum Likelihood (RAXML) Program. In *Current Protocols in Molecular Biology*, volume Chapter 19, page Unit19.11. John Wiley & Sons, Inc., Hoboken, NJ, USA.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–25.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7.

[Sarmento et al., 2010] Sarmento, H., Montoya, J. M., Vázquez-Domínguez, E., Vaqué, D., and Gasol, J. M. (2010). Warming effects on marine microbial food web processes: how far can we go when it comes to predictions? *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 365(1549):2137–49.

[Schluter, 1984] Schluter, D. (1984). A Variance Test for Detecting Species Associations, with Some Example Applications. *Ecology*, 65(3):998–1005.

[Schmidt, 2016] Schmidt, K. (2016). Thermal adaptation of Thalassiosira pseudonana using experimental evolution approaches. (June).

[Schmieder and Edwards, 2011] Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–504.

[Shemi et al., 2015] Shemi, A., Ben-Dor, S., and Vardi, A. (2015). Elucidating the composition and conservation of the autophagy pathway in photosynthetic eukaryotes. *Autophagy*, 11(4):701–715.

[Shendure and Ji, 2008] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

[Shull et al., 2001] Shull, V. L., Vogler, A. P., Baker, M. D., Maddison, D. R., and Hammond, P. M. (2001). Sequence alignment of 18S ribosomal RNA and the basal relationships of adephagan beetles: Evidence for monophyly of aquatic families and the placement of trachypachidae. *Systematic Biology*, 50(6):945–969.

[Simon et al., 2009] Simon, N., Cras, A.-L., Foulon, E., and Lemée, R. (2009). Diversity and evolution of marine phytoplankton. *Comptes Rendus Biologies*, 332(2-3):159–170.

[Simpson, 2006] Simpson, G. L. (2006). *betadisper function — R Documentation*. https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/betadisper.

[Singh et al., 2009] Singh, J., Behal, A., Singla, N., Joshi, A., Birbian, N., Singh, S., Bali, V., and Batra, N. (2009). Metagenomics: Concept, methodology, ecological inference and recent advances. *Biotechnology Journal*, 4(4):480–494.

[Sinha and Lynn, 2014] Sinha, S. and Lynn, A. M. (2014). HMM-ModE: implementation, benchmarking and validation with HMMER3. *BMC Research Notes*, 7(1):483.

[Stamatakis, 2014] Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9):1312–3.

[Stamatakis et al., 2005] Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463.

[Stecher et al., 2016] Stecher, A., Neuhaus, S., Lange, B., Frickenhaus, S., Beszteri, B., Kroth, P. G., and Valentin, K. (2016). rRNA and rDNA based assessment of sea ice protist biodiversity from the central Arctic Ocean. *European Journal of Phycology*, 51(1):31–46.

[Suzuki et al., 2015] Suzuki, S., Kakuta, M., Ishida, T., and Akiyama, Y. (2015). Faster sequence homology searches by clustering subsequences. *Bioinformatics*, 31(8):1183–1190.

[Taylor et al., 2007] Taylor, F. J. R., Hoppenrath, M., and Saldarriaga, J. F. (2007). Dinoflagellate diversity and distribution. pages 173–184. Springer, Dordrecht.

[Thomas et al., 2012a] Thomas, M. K., Kremer, C. T., Klausmeier, C. A., and Litchman, E. (2012a). A Global Pattern of Thermal Adaptation in Marine Phytoplankton. *Science*, 338(6110):1085–1088.

[Thomas et al., 2012b] Thomas, T., Gilbert, J., and Meyer, F. (2012b). Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation*, 2(1):3.

[Thompson et al., 1994] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–80.

[Torres-Machorro et al., 2010] Torres-Machorro, A. L., Hernández, R., Cevallos, A. M., and López-Villaseñor, I. (2010). Ribosomal RNA genes in eukaryotic microorganisms: witnesses of phylogeny? *FEMS Microbiology Reviews*, 34(1):59–86.

[Toseland et al., 2013] Toseland, A., Daines, S. J., Clark, J. R., Kirkham, A., Strauss, J., Uhlig, C., Lenton, T. M., Valentin, K., Pearson, G. A., Moulton, V., and Mock, T. (2013). The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nature Climate Change*, 3(11):979–984.

[Tremblay et al., 2015] Tremblay, J., Singh, K., Fern, A., Kirton, E. S., He, S., Woyke, T., Lee, J., Chen, F., Dangl, J. L., and Tringe, S. G. (2015). Primer and platform effects on 16S rRNA tag sequencing. *Frontiers in Microbiology*, 6:771.

[Trimborn et al., 2015] Trimborn, S., Hoppe, C. J., Taylor, B. B., Bracher, A., and Hassler, C. (2015). Physiological characteristics of open ocean and coastal phytoplankton communities of Western Antarctic Peninsula and Drake Passage waters. *Deep Sea Research Part I: Oceanographic Research Papers*, 98:115–124.

[van Dijk et al., 2014] van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426.

[von Mering et al., 2007] von Mering, C., Hugenholtz, P., Raes, J., Tringe, S. G., Doerks, T., Jensen, L. J., Ward, N., and Bork, P. (2007). Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. *Science*, 315(5815):1126–1130.

[Voss et al., 2013] Voss, M., Bange, H. W., Dippner, J. W., Middelburg, J. J., Montoya, J. P., and Ward, B. (2013). The marine nitrogen cycle: recent discoveries, uncertainties and the potential relevance of climate change. *Philosophical transactions of the Royal Society of London. Series B, Biological Sciences*, 368(1621):20130121.

[Wain et al., 2002] Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., and Povey, S. (2002). Guidelines for Human Gene Nomenclature. *Genomics*, 79(4):464–470.

[Wang et al., 2006] Wang, J., Keightley, P. D., and Johnson, T. (2006). MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics*, 7(1):292.

[Wang et al., 2014] Wang, Y., Tian, R. M., Gao, Z. M., Bougouffa, S., and Qian, P.-Y. (2014). Optimal eukaryotic 18S and universal 16S/18S ribosomal RNA primers and their application in a study of symbiosis. *PloS one*, 9(3):e90053.

[Wang et al., 2009] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1):57–63.

[Waterhouse et al., 2009] Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191.

[Wernegreen, 2012] Wernegreen, J. J. (2012). Endosymbiosis. *Current Biology*, 22(14):R555–R561.

[Wiens and Tiu, 2012] Wiens, J. J. and Tiu, J. (2012). Highly Incomplete Taxa Can Rescue Phylogenetic Analyses from the Negative Impacts of Limited Taxon Sampling. *PLoS ONE*, 7(8):e42925.

[Wilkins et al., 2013] Wilkins, D., Yau, S., Williams, T. J., Allen, M. A., Brown, M. V., DeMaere, M. Z., Lauro, F. M., and Cavicchioli, R. (2013). Key microbial drivers in Antarctic aquatic environments. *FEMS Microbiology Reviews*, 37(3):303–335.

[Williams et al., 2010] Williams, K. P., Gillespie, J. J., Sobral, B. W. S., Nordberg, E. K., Snyder, E. E., Shallom, J. M., and Dickerman, A. W. (2010). Phylogeny of gammaproteobacteria. *Journal of Bacteriology*, 192(9):2305–14.

[Wistrand and Sonnhammer, 2005] Wistrand, M. and Sonnhammer, E. L. L. (2005). Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics*, 6(1):99.

[Wooley et al., 2010] Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A Primer on Metagenomics. *PLoS Computational Biology*, 6(2):e1000667.

[Wu et al., 2015] Wu, S., Xiong, J., and Yu, Y. (2015). Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study of Subclass Copepoda. *PLOS ONE*, 10(6):e0131498.

[Ye et al., 2018] Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., Xu, H., Huang, X., Li, S., Zhou, A., Zhang, X., Bolund, L., Chen, Q., Wang, J., Yang, H., Fang, L., and Shi, C. (2018). WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Research*, 46(W1):W71–W75.

[Yu and Zhang, 2013] Yu, K. and Zhang, T. (2013). Construction of Customized Sub-Databases from NCBI-nr Database for Rapid Annotation of Huge Metagenomic Datasets Using a Combined BLAST and MEGAN Approach. *PLoS ONE*, 8(4):e59831.

[Zhang and Horvath, 2005] Zhang, B. and Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1):Article17.

[Zhang et al., 2012] Zhang, H., John, R., Peng, Z., Yuan, J., Chu, C., Du, G., and Zhou, S. (2012). The relationship between species richness and evenness in plant communities along a successional gradient: a study from sub-alpine meadows of the Eastern Qinghai-Tibetan Plateau, China. *PloS one*, 7(11):e49024.

[Zhang et al., 2014] Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford, England)*, 30(5):614–20.

[Zhou et al., 2014] Zhou, Q., Su, X., and Ning, K. (2014). Assessment of quality control approaches for metagenomic data analysis. *Scientific Reports*, 4(1):6957.

[Zhou et al., 2015] Zhou, Q., Su, X., and Ning, K. (2015). Assessment of quality control approaches for metagenomic data analysis. *Scientific Reports*, 4(1):6957.

Appendices

Appendix A

A.A Metadata

Station	Longitude	Latitude	Depth meter	Temperature degrees celsius	Salinity	Nitrate.Nitrite	Phosphate	Silicate
1	-9.52472	79.0225	17	-1.0337	31.0274	0	0.47	2.48
2	-8.52472	79.07611	26	-1.5122	31.708	0.14	0.95	2.86
3	-7.67278	79.04278	20	-1.4645	31.3282	0	0.62	7.17
4	-4.08556	79.00056	35	-1.7398	33.9123	0.81	0.58	5.49
5	-4.08556	79.00056	10	-1.5083	32.9517	0.81	0.58	5.49
6	-4.78556	78.85611	20	-1.6191	32.2039	0.6	0.74	8.97
7	-3.22861	78.86694	15	-0.8805	32.3454	3.3	0.79	5.47
8	-2.83722	78.89667	110	3.0802	34.9887	10.44	1.05	4.46
9	-2.83722	78.89667	25	-1.3686	33.3389	0.7	0.46	3.03
10	-1.84	78.93667	16	4.3973	35.062	4.77	0.84	3.26
11	-0.55917	78.98167	10	1.3665	33.5306	0.69	0.5	3.64
12	3.73583	79.07278	10	0.0522	33.586	2.8	0.61	2.49
13	3.73583	79.07278	5	-0.3027	33.1465	2.21	0.51	3.12
14	4.1527	78.62153	15	5.87	35.05	NA	NA	NA
15	5.32861	78.83889	15	4.4939	35.0994	3.32	0.73	3.95
16	6.10556	79.08944	7	4.8199	35.1105	3.9	0.76	3.88
17	7.07639	79.06667	18	4.5614	35.0886	9.69	1.07	4.31
18	8.11222	78.86972	10	5.269	35.0693	0.51	0.49	2.75
19	9.23306	78.85139	25	2.3374	34.7579	2.93	0.94	2.16
20	11.30917	76.25389	15	5.5282	35.1514	6.26	0.83	4.08
21	9.85667	73.01889	20	6.0186	35.1528	6.55	0.88	3.9
22	8.86667	71.20083	10	7.1834	35.1344	4.83	0.77	3.62
23	7.73028	69.23028	10	9.0976	34.7321	2.46	0.46	1.52
24	7.73028	69.23028	5	8.7545	34.8661	0.86	0.49	1.68
25	6.53028	67.23028	15	8.9384	35.091	4.16	0.74	2.48
26	6.53028	67.23028	20	8.6781	35.1	2.38	0.66	2.12
27	5.41917	65.24611	20	9.24	34.93	2.85	0.6	2.44
28	5.41917	65.24611	5	9.75	34.94	1.58	0.59	2.35
29	-21.7407	62.8	10	7.69	35.2	13.35	0.862	6.69
30	-20.4903	61.7103	11	8.27	35.18	12.55	0.82	NA
31	-19.3398	60.6801	10	8.75	35.25	11.75	0.755	6.57
32	-18.0699	59.5	10	9.49	35.33	10.61	0.69	NA
33	-16.5201	58.0002	11	9.54	35.37	11.23	0.729	NA
34	-16.509	54.6333	10	10.79	35.45	7.79	0.488	1.59
35	-16.4972	52.6218	26	11.54	35.55	7.95	0.512	NA
36	-16.3567	49.9151	50	12.22	35.64	6.86	0.428	NA
37	-12.1104	47.5701	40	12.07	35.64	5.92	0.412	2.54
38	-12.4301	45.5298	16	13.96	35.81	0.33	0.049	2.18
39	-12.6098	44.2798	20	13.78	35.8	1.47	0.087	NA
40	-12.8799	42.3401	21	14.4	35.9	0.76	0.062	NA
41	-13.1901	40.5296	25	14.99	36.06	0.45	0.047	NA
42	-13.576	38.4205	67	15.9	36.26	0.52	0.045	NA
43	-13.94	36.5297	25	16.67	36.32	0.02	0.019	NA
44	-12.0872	37.833	80	20.63	36.44	0.426330313	0.065720386	0.478187565
45	-13.1352	34.876	80	21.78	36.65	0.269427308	0	0.292683967
46	-14.2597	34.7198	21	17.43	36.46	0.04	0.015	NA
47	-14.2601	34.7197	48	16.93	36.42	1.02	0.031	0.2926
48	-14.5898	32.8199	73	18.06	36.66	0.05	0.022	NA
49	-14.8699	31.2202	75	18.57	36.75	0.02	0.015	NA
50	-14.8704	31.2193	72	18.37	36.71	0.03	0.014	NA
51	-15.0705	30.0186	30	18.55	36.71	0.03	0.021	NA
52	-15	29	81	18.97	36.83	0.02	0.023	0.75105
53	-15.1548	28.937	90	23	36.6	0.543261784	0.076334058	0.751057865
54	-17.4585	26.049	80	24.62	36.9	0.516096977	0.066033049	0.405116155
55	-20.1823	23.69	60	25.25	36.57	1.954317104	0.194632089	0.689491487
56	-20.7015	18.755	45	26.78	36.17	6.127505044	0.331454155	1.187283885
57	-20.515	15.249	55	28.6	35.64	16.42369279	0.682339671	3.367016894
58	-18.6202	8.472	46	29.02	34.88	3.106489549	0.241043741	2.590527277
59	-13.602	2.405	80	27.1	35.61	6.458482733	0.3717252	2.064853304
60	-9.4218	-2.045	63	25.64	35.94	10.02985752	0.51653754	2.617416148
61	-7.0612	-4.668	47	24.71	35.83	11.32366081	0.55562155	3.066000484
62	-4.9043	-7.394	45	23.27	36.23	6.480985475	0.366034527	2.020429743
63	-0.3178	-13.104	75	20.66	36.5	4.395764083	0.236727944	0.826883927
64	2.9768	-17.283	30	18.82	35.97	0	0.19	2.63
65	3.7993	-18.25	43	18.77	35.99	8.005879218	0.444372737	0.85904911
66	5.9978	-20.987	30	18.52	35.62	5.951945303	0.392739675	0.065775733
67	7.1922	-22.644	40	18.48	35.7	0.502052705	0.044543045	0.43401924
68	9.9828	-26.441	30	18.2	35.2	2.283057742	0.173644994	0.688728588

A.B 18S rDNA reference databases taxa IDs

1032745	127567	498767	650286	1072577	670782	858356	1171960	434030
639210	127563	67961	110372	1072585	651459	210735	35113	181586
156996	119481	101922	238096	1072581	110316	2923	408072	197858
157007	342563	221933	66473	75742	348448	425803	5932	223571
296670	402028	48970	632150	203142	650289	425804	37092	197862
44058	1003327	179290	160611	1072566	650283	1003175	104774	274053
1104430	397339	707317	1434936	237895	693284	387435	863766	524897
96791	512344	710654	2957	5875	650331	225107	5927	524900
89044	396045	71746	43686	27996	650333	189332	289478	524895
1117030	49980	138175	217026	5865	651455	326572	289481	693931
135473	219168	332216	8030	476202	113556	510609	210825	197857
413849	6009	44656	7739	300409	240359	1162085	5928	394799
416808	219837	889457	201616	86281	133427	376358	311390	1004046
262226	219841	170400	248472	84966	66800	107758	1081457	240170
31296	653507	1082342	261852	88456	259926	258031	230077	1166918
45884	1071533	56002	31408	84969	2925	39447	122942	455295
1054987	652930	51328	436072	700707	58156	673113	1002335	155143
1054391	223996	56009	143010	482538	73915	271680	1002334	168254
278983	708628	464287	42008	162052	66791	143672	278832	155140
1213618	515478	243130	173452	99158	497816	414396	1002222	168251
178368	497703	1085968	31461	29176	160621	459342	1002226	346245
432302	33640	173495	197877	5811	2916	400756	692909	703568
507873	913975	63592	38533	94643	66465	79894	693871	168244
1247838	303408	3190	309167	689363	325318	410738	880990	168257
985835	97221	926286	522302	62968	326149	71001	35100	446414
614059	265540	204415	536595	333133	685761	858351	211662	1126575
502092	285082	204408	536601	294749	244960	424504	881006	172019
46462	211638	515480	221821	31383	505693	66468	223573	346241
247150	35117	515477	38836	282351	358189	210733	274051	168242
1170558	279580	44430	67781	110459	1118043	39450	435890	168248
693439	216975	643632	36881	110473	158379	160623	669200	151077
696972	425790	652932	94289	541018	5992	1165720	693053	1170556
358025	363325	1050086	552938	46026	696187	327389	35097	37474
601995	142111	197538	578126	244810	651135	339968	1071701	5965
693140	382377	329754	81603	5817	47934	566470	385028	211642
153251	216766	97102	332487	169156	622439	5960	57509	159163
47889	197897	395882	47887	346247	182090	467015	39461	415605
47894	860613	435893	324857	346240	181653	467023	394801	273907
358022	1173301	641229	338344	497731	641228	358021	497729	264369
693421	375585	513281	686996	1173416	513283	197901	990178	1176393
550469	375583	211016	692885	990236	1049793	1170553	375587	1292461
550467	5997	211025	754198	990234	513204	482403	990189	1176361
692521	993385	1169006	268522	614050	1071509	754193	1229648	5974

89957	9606	1134686	142497	257561	35206	332489	400980	132687
584794	60559	101920	173468	164319	81598	332491	400982	864286
122588	7604	182762	123987	257567	479265	332493	356898	592675
160615	82378	1094580	257592	105601	504345	332497	152445	5936
261834	7668	257374	40391	257545	397053	332495	652658	152452
302480	206668	257375	40385	257551	81617	189623	330974	5940
160619	34765	326073	680465	88417	81584	284005	693924	138849
261837	569450	2786	31369	152008	81611	332501	99922	331608
79898	425018	1094582	291393	239144	1074216	332498	382958	59999
388225	97517	35688	367044	257547	1074204	332500	385034	311320
402582	230730	101926	257542	257549	39628	189644	860611	1004043
388228	491138	2771	173450	257554	322167	332483	423615	125641
388232	6669	45157	173447	28020	81592	536594	170499	94673
107036	202087	82540	159503	31421	197880	332484	346229	70075
373098	6661	31354	326253	305493	42024	31367	1174532	1001742
66792	194544	35151	122415	38544	327986	31472	211028	1287487
72554	439682	468936	122413	88422	239160	42480	1292964	5950
261842	178832	82561	35157	88412	931254	257814	1229645	114681
2866	136180	204479	536613	31447	31455	191046	181651	558275
393032	299778	204480	339588	41686	257607	225046	1229734	114678
1132513	560977	753684	339604	196371	164055	31474	1292960	99924
160617	231624	110510	257588	257575	81613	282360	513210	268526
2966	1234259	139984	142502	29232	81600	110475	1176324	278822
763934	104782	362230	31417	29222	81590	332481	1170544	686991
340370	1435206	139907	142499	29241	91060	189642	247154	71585
412152	322853	139980	142505	95361	81594	536593	1071508	40806
418113	136452	282340	35173	29227	81587	536603	526559	1144346
88552	1197702	101924	142512	1269974	81607	536604	717095	163346
95749	314080	37198	128537	257571	38331	536610	1170546	163350
103983	6087	82843	142492	197835	31500	536616	1170548	163357
51511	209422	436068	142489	197840	81615	536591	238901	71594
223366	201645	82839	128534	197837	406714	189638	1297756	5987
350068	201679	70838	217487	38372	81619	536586	692525	993383
1176391	99915	488251	139859	467021	513215	414913	398671	375580
180940	5947	1071522	39464	758568	211022	57513	414910	857030
385029	552936	488247	864284	124797	402891	469766	694303	692523
157072	33674	278986	227086	188973	197904	40330	459245	1170439
101203	87102	366611	552664	188950	527218	502090	55726	1030620
135477	463366	366598	552666	149084	394493	39564	587064	485330
92981	126728	1054399	67809	188943	693161	352448	197894	40328
92972	42467	987159	238782	686740	92969	463362	366596	332299
92976	463364	1054387	238789	45107	687159	985833	54108	88563
92966	4773	366601	238793	563755	391043	350847	74790	1170431
986738	1127205	123356	160259	987156	332305	188946	984054	459518
1054397	278125	983661	37095	1408141	332301	999463	340085	1003332

536599	159727	643655	31496	164601	327057	42696	238774	60001
536597	94299	320783	31481	77548	50045	162320	238761	696133
536609	282348	433419	79259	52965	50044	162317	70182	1005899
536614	282353	1196376	498007	710656	470007	3092	167538	1005901
173553	103713	177373	2801	87090	3097	183309	37360	88570
1261578	110477	109050	700918	647327	52964	1035567	167535	459270
335259	159725	244126	38265	327050	104531	75803	933485	5982
173540	189624	31363	131155	52035	104533	113520	188971	459237
1261580	189634	122417	1070855	269637	104536	113536	5855	459276
1261571	94295	389187	35153	158507	3175	113522	110365	941344
173548	189646	89943	2762	132247	51320	91193	2969	459265
173550	1206573	231748	38271	52685	55999	3088	672928	941345
1261582	158647	871653	38269	142647	33095	55409	340708	459235
1261569	131068	389190	77922	52679	329040	91197	230744	459268
89212	131064	48608	55529	69401	337949	91196	251331	63136
155556	131094	48948	3032	436124	163323	202681	663229	1170433
110469	282363	1034349	478117	361666	993091	795116	340200	485360
121067	189632	464653	464988	579148	204991	183317	340204	909417
228269	35161	1034350	77928	55997	1008953	326141	283649	459257
110466	76903	48942	46947	361668	44654	651586	283647	485357
155558	189640	48951	437768	56006	156110	132186	326279	164621
228262	189636	48615	2898	332218	1148060	875623	1127215	1044902
110471	189648	464689	195067	47790	132188	875621	1127417	523209
1261574	189628	35170	57475	47788	271407	875620	1127213	127229
35163	189652	464649	40526	51718	165818	876697	1127100	188964
945030	268567	48962	193549	56011	55410	145388	1127415	151026
159598	31486	871656	167772	1034561	271398	191687	1127103	361427
282366	67958	231757	551846	361670	271396	307507	1127203	981202
282357	111861	348030	3046	179866	302391	117505	85466	361426
282355	35202	48959	3156	3192	160063	82291	310810	188939
110462	31490	1034343	52028	1158268	132190	889455	135480	87111
860634	209631	348081	71744	34115	299577	31300	114742	33676
99899	209632	48975	107616	3095	170393	356782	100874	12968
98045	877583	279125	1127141	100933	98068	159317	492103	1127216
98064	78393	220110	1127137	687161	64707	377434	35217	1128110
420600	136833	299204	278118	1044904	438413	1127223	39714	178372
52559	280858	534455	64708	42384	536091	1127220	178364	432305
52556	281460	366605	98060	114254	862249	1127211	707168	420607
98055	33653	413940	98066	230409	626141	1127108	707170	3000
67593	127148	413938	2996	42750	1108494	1127133	178375	308878
143451	357350	629711	52241	104657	941460	1069743	39717	44056
126844	272144	629713	590968	159342	210589	44432	27967	88167
707994	167961	1191181	137466	382380	216777	643630	2876	1117027
65357	1486930	366592	216741	443638	178366	515472	29207	210620
1054395	98058	515487	278979	278121	72520	2885	55585	55587

179863	41300	3159	262509	221848	195969	70452	505996	2850
164533	63684	185965	173376	111460	195967	561169	517775	431345
52677	3072	138176	374114	221826	156128	119497	232859	431593
164532	155715	230557	173371	204410	88271	216820	505992	431588
3081	188049	44573	231078	204422	631452	1158023	915350	210449
558848	154508	138169	231080	204419	81844	515468	196682	37319
76111	173497	73033	31301	34119	687948	432268	1115528	1003042
202684	1034627	398706	445995	34121	70448	216617	487668	431324
133488	37433	189347	507620	34140	41875	1003029	1115533	431370
532145	152768	230559	3075	204412	156133	1003039	216745	432551
247495	1293078	230561	38881	162054	156131	33645	265532	1003108
797674	187458	241068	120749	162063	41886	216618	186025	1003036
797671	1034482	63412	163307	35845	539821	426669	941921	1003053
55402	413988	163314	120751	219609	221441	515488	431322	1003046
798524	312843	160070	1065486	219614	35139	426676	216747	515482
3080	306440	348768	202679	3140	127547	426664	216734	1003145
247497	312849	43941	577621	35855	418912	186020	221720	2853
3099	200484	160076	577622	35853	424526	515470	515474	216749
247499	82176	53263	120745	35857	53265	515464	431333	431368
221620	1065496	173490	1212498	35859	38817	515479	216798	197754
75799	312850	173499	1034818	2788	2903	1158036	303458	265561
3074	577485	216028	183676	168183	424539	515471	1003069	197753
797666	63683	153906	51329	3702	127562	1003094	431358	216896
797637	480381	216033	926342	4081	97492	196633	1003020	3003
91190	160068	63410	926344	3659	156173	426662	303409	502569
163309	34148	262505	162066	4558	373042	210441	265537	210618
163317	3093	262511	162062	33129	97495	196638	303462	232512
160069	1232723	439779	162076	4577	1136786	515462	1051708	451786
63675	415184	1081497	926348	4530	284051	210602	718187	29205
634749	34116	1108641	162059	3760	259385	528331	1003112	83371
190057	60131	220632	333302	3871	13221	210616	49249	94617
191674	3171	1081495	162067	284941	37099	455043	449131	423563
249350	284972	220631	185010	1486889	118079	515484	449130	425076
86898	92126	96789	74375	1003075	44451	1003142	216908	441170
142656	92128	37475	2835	59812	658122	1003199	2829	246117
3012	82369	96788	73013	210444	162275	210609	90936	246119
185805	308770	1104315	195974	327391	1278069	374047	375454	243268
2880	240564	98049	73019	134680	475233	233771	310674	1034831
27963	240566	45116	236787	45653	1003208	246123	308883	214821
49246	216801	112064	479719	157126	1003064	1003062	66222	116065
265543	216773	243166	77034	38822	1003144	1003137	62316	157001
1003027	216790	88149	240383	215586	1158022	91992	105404	258581
1003085	265556	64927	82161	35684	265572	3005	105414	196139
1090589	265552	62313	82166	35682	265563	188541	38824	1003033
655750	1003084	3022	420584	172671	172669	124430	157128	

A.C Evenness and occupancy taxonomy names to numbers

Table A.1: Evenness and occupancy taxonomy names to numbers

18S rDNA taxonomy	Number	16S rDNA taxonomy	Number
Stramenopiles	1	Gammaproteobacteria	1
Cryptophyta	2	Alphaproteobacteria	2
Mamiellophyceae	3	Flavobacteriia	3
Rhizaria	4	Marine Group II	4
Maxillopoda	5	unclassified Verrucomicrobia	5
Alveolata	6	Synechococcales	6
Dinophyceae	7	Deltaproteobacteria	7
Spirotrichea	8	Proteobacteria	8
Polycystinea	9	Candidatus Marinimicrobia	9
Coscinodiscophyceae	10	Actinobacteria	10
Amoebozoa	11	Planctomycetia	11
Ascidiacea	12	Acidimicrobiia	12
Pelagophyceae	13	Betaproteobacteria	13
Dictyochophyceae	14	Cyanobacteria	14
Bangiophyceae	15	Verrucomicrobiae	15
unclassified Alveolata	16	Saprospiria	16
Acantharea	17	Marine Group III	17
Haptophyceae	18	unclassified Thaumarchaeota	18
Cercozoa	19	Nitrosopumilales	19
Gregarinasina	20	Verrucomicrobia	20
Florideophyceae	21	Bacteroidetes	21
Fragilariophyceae	22	Oligosphaeria	22
Chlorophyceae	23	Opitutae	23
Bicosoecida	24	Oligoflexia	24
Viridiplantae	25	Planctomycetes	25

Litostomatea	26	Cytophagia	26
Heterotrichea	27	Sphingobacteriia	27
Labyrinthulomycetes	28	Phycisphaerae	28
Hydrozoa	29	Chloroflexi	29
Pyramimonadales	30	Gemmatimonadetes	30
Bacillariophyceae	31	Oscillatoriophycideae	31
Phyllopharyngea	32	Patescibacteria group	32
Prasinococcales	33	Chlamydiia	33
Coccidia	34	Bacilli	34
Ellobiopsidae	35	Tenericutes	35
Gymnolaemata	36	Bacteroidetes Order II.Incertae sedis	36
Appendicularia	37	Clostridia	37
Bilateria	38	Lentisphaerae	38
PXclade	39	Chlamydiae	39
Placididea	40	Nitrospinia	40
Blastocystis	41	Acidobacteria	41
unclassified Rhizaria	42	Archaea	42
Chrysophyceae	43	Epsilonproteobacteria	43
Pycnococcaceae	44	Tissierellia	44
Oomycetes	45	Halobacteria	45
Mediophyceae	46	Thermoplasmata	46
Perkinsea	47	Bacteroidia	47
Oligohymenophorea	48	Balneolia	48
Karyorelictea	49	Mollicutes	49
Developea	50	Candidatus Saccharibacteria	50
Protostomia	51	Anaerolineae	51
Trebouxiophyceae	52	Candidatus Hydrogenedentes	52
Rhodophyta	53	Gemmatimonadetes <class>	53
Eumetazoa	54	Nitriliruptoria	54
Gromiidae	55	Deinococci	55
unclassified stramenopiles	56	Negativicutes	56

Compsopogonophyceae	57	Rubrobacteria	57
Ulvophyceae	58	Fusobacteriia	58
eudicotyledons	59	Firmicutes	59
Malacostraca	60	Chitinophagia	60
Aconoidasida	61	Erysipelotrichia	61
Actinopteri	62	Chlorobi	62
Bdelloidea	63	Nitrospinae	63
Chlorophyta incertae sedis	64	Spirochaetia	64
Echinoidea	65	Nostocales	65
Mammalia	66	Ardenticatenia	66
Ophiuroidea	67	Solibacteres	67
Prasinophyceae incertae sedis	68		
Rhodellophyceae	69		
Pedinophyceae	70		
Chlorophyta	71		
Ciliophora	72		
Holothuroidea	73		
Pinguiophyceae	74		
Bolidophyceae	75		
Glaucocystophyceae	76		
Crustacea	77		

A.D 18S and 16S rDNA correlation heatmap coefficient values and p-values

Table A.2: Arctic 18S rDNA correlation heatmap coefficients values and p-values

Taxonomy	Longitude correlation value	Longitude p-value	Latitude correlation value	Latitude p-value	Depth correlation value	Depth p-value	Temperature correlation value	Temperature p-value	Salinity correlation value	Salinity p-value	Nitrate.Nitrite correlation value	Nitrate.Nitrite p-value	Phosphate correlation value	Phosphate p-value	Silicate correlation value	Silicate p-value
Acantharea	-0.3306590312	0.0089630265	-0.1895514322	0.3536991718	0.3444051012	0.0849027225	-0.2864100723	0.156043455	-0.1041407717	0.6126600112	0.0408177114	0.8430635591	0.2619883622	0.19604354	-0.2700277216	0.1821707301
Aconoidasida	-1	3.80E-160	-1	0	1	3.22E-159	-1	5.86E-160	-1	0	-1	3.48E-160	-1	8.71E-156	-1	5.15E-160
Actinopteri	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Appendicularia	0.8675559817	9.62E-009	0.861165873	1.64E-008	-0.4358133757	0.0260458265	0.7951928397	1.21E-006	0.4460630086	0.0223680361	-0.1642120671	0.422782146	-0.7712803984	3.98E-006	-0.7156091394	3.96E-005
Ascidiacea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bacillariophyceae	-0.7962469394	1.14E-006	-0.5188402405	0.0066116078	0.4915966463	0.0107539635	-0.6519338957	0.0003080299	-0.576241719	0.0020633559	-0.1232648824	0.5485639967	0.0121422847	0.9530550629	0.4673369132	0.0160751537
Bangiophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Chlorophyceae	0.06463254 71	0.7537622152	-0.2011958576	0.324342501	0.3321 424703	0.0973670895	-0.1213233316	0.554927312	0.2782076699	0.1687701077	0.2760593661	0.1722204201	0.4887107008	0.0112974745	0.6224029705	0.0006855486
Chrysophyceae	-0.2705068369	0.1813660718	-0.2681467241	0.1855375 05	-0.1476887 02	0.4715287757	-0.2 448499313	0.2279916 79	-0.0427455302	0.8357481109	-0.5618270232	0.0028188007	-0.0938015339	0.6485482386	-0.1485911456	0.4687939044
Coscinodiscophyceae	0.2395563426	0.2385228 308	0.4770260023	0.013736641	-0.1161369812	0.5720887621	0.4440273073	0.0230627003	0.1956045044	0.3382481708	-0.1278607026	0.5336373504	-0.4357502171	0.0260699917	-0.1575583416	0.4420705678
Cryptophyta	0.3877526973	0.0503158107	0.0663319575	0.7474918647	0.0273558609	0.8944683222	0.2959135707	0.1421679418	0.5066550733	0.0082581424	0.2765718386	0.1713929089	0.1992851262	0.3290547368	0.0229234397	0.8872652 96
Dictyochophyceae	0.4192018024	0.0330318507	0.5171679403	0.0068197337	-0.1594105871	0.4366543411	0.6371117575	0.0004649849	0.4614984291	0.0176357991	0.1057672016	0.6070901711	0.0748291749	0.7163807213	-0.122665045	0.5505263229
Dinophyceae	0.3550857729	0.075064828	0.3494001589	0.0801871036	0.0811186062	0.6936304471	0.3757191143	0.0585509413	0.5256378448	0.0058197048	0.3334705437	0.0959547005	0.0706318081	0.731697109	-0.3877198363	0.0503369897
Echinoidea	0.0781413333	0.7043691058	-0.0860414056	0.6760010359	-0.3212836513	0.1095045504	-0.3298182634	0.0998761841	0.4197041867	0.03280059	0.1723882387	0.3997256518	0.4360003661	0.0259746084	0.3265374483	0.1034996792
Fragilariophyceae	0.9787547505	5.01E-018	0.905579365	2.03E-010	-0.3968691342	0.0447047411	0.8190644212	3.12E-007	0.9714323695	1.69E-016	0.9380824585	1.52E-012	0.8495140547	4.04E-008	0.2165094128	0.2880740709
Glaucocystophyceae	-0.9752218177	3.12E-017	-0.9803711448	1.95E-018	-0.6122074312	0.0008872826	-0.5828294681	0.0017806451	-0.8375693329	9.47E-008	-0.9986851018	1.75E-032	-0.9973948557	6.37E-029	0.3527341553	0.0771522276
Gregarinasina	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Gymnolaemata	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Heterotrichea	-0.1622005425	0.4285640956	0.3815003861	0.0544729481	0.3516787897	0.0781032762	0.3300718709	0.0996000815	0.2096105458	0.3040830216	0.0159018748	0.9385433528	-0.1628435346	0.4267112082	-0.3577954754	0.0727133842
Hydrozoa	-0.1775965925	0.3854155557	0.1735068022	0.3966274357	0.2439457933	0.2297680542	0.0747211411	0.7167736343	-0.2784184369	0.1684342368	-0.4690161079	0.0156479505	0.1464342296	0.4753441094	-0.6300013994	0.0005623456
Labyrinthulomycetes	0.0316658053	0.8779553509	0.1964276388	0.3361788779	0.0842578866	0.6823695245	0.1419333156	0.4891625289	0.0778213759	0.7055264847	-0.0747910492	0.7165193742	-0.194485356	0.3410733596	-0.43590883	0.0260094516
Litostomatea	-0.2830409682	0.1611856189	0.2573450933	0.2043777264	0.2775584746	0.1698075972	-0.1061739808	0.6057004394	-0.2877649707	0.1540085642	-0.1336688018	0.514991769	-0.0468786372	0.8201097847	-0.2378945496	0.2418939984
Mamiellophyceae	0.2873798391	0.1545850235	0.372586509	0.0608575595	-0.0651012	0.7520314798	0.3454583289	0.0838913734	0.2802891016	0.1654737673	0.1658992322	0.4179657036	0.1814717098	0.3749614438	0.3730147823	0.0605381204
Mammalia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Maxillopoda	0.4030937998	0.0411660133	0.2582601967	0.2027164407	0.1273906713	0.5351550856	0.4037116288	0.0408272539	0.4955635731	0.0100424971	0.2184550218	0.2836573097	0.033301092	0.8717025806	-0.5696871317	0.0023819989
Mediophyceae	-0.0385939031	0.8515179246	0.2882351394	0.1533068971	-0.2912712295	0.1488306297	0.1854723511	0.3643427669	-0.1381373121	0.5009721055	-0.3421855602	0.0870641848	-0.1380901239	0.5011197945	-0.433363282	0.0269936538
Nc.Alveolata	0.3032576458	0.1320697731	0.2949245599	0.1435690685	-0.1850842652	0.3653650 63	0.1531283877	0.4551690881	0.2479727669	0.2212673777	0.0829356749	0.6871045345	-0.1506541479	0.4625729241	-0.193896547	0.342566188
Nc.Eukaryota	-0.2849678195	0.1582302069	-0.508369174	0.0080076534	0.3126738959	0.119896502	-0.4629154943	0.0172460531	-0.1687462521	0.4099071641	0.0734329257	0.7214641885	0.2961501179	0.1418342876	0.1071298204	0.6024401431
Nc.prasinophytes	0.5912056033	0.0014697465	0.082954327	0.6870376585	0.0918353796	0.6554634252	0.4572879359	0.0188366445	0.7306202956	2.25E-005	0.3774381083	0.0573143948	0.2642064905	0.1921456738	-0.1935707942	0.3433934763
Nc.Rhizaria	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nc.Stramenopiles	0.0960549048	0.6406576482	0.3592106775	0.0713079304	-0.0620327768	0.7633839949	0.3014551026	0.134498666	0.2303708691	0.2575480398	-0.156992248	0.4437330529	-0.2437982524	0.2300588021	-0.3180466913	0.1133303077
Nc.Viridiplantae	0.4077967471	0.0386426043	0.4412411342	0.0240417572	0.7904442547	1.55E-006	0.1540221175	0.4525101622	0.1859811509	0.3630050242	0.0271443287	0.8952799383	-0.1694698099	0.4078729929	-0.0948316189	0.6449365596
Oomycetes	0.2655292184	0.1898468257	0.631959058	0.0005339159	0.0261610616	0.8990539168	0.6600696884	0.0002434107	0.1740251501	0.3951963196	0.0439337657	0.8312457177	-0.165323412	0.4196061184	-0.438403975	0.0250730984
Ophiuroidea	0.3746408714	0.0593370817	-0.2706649235	0.1811011132	-0.5157274475	0.0070033965	-0.1398033246	0.4957717086	-0.6677514991	0.0001936324	0.3754594039	0.0587395506	0.0836101987	0.6846875291	0.1041236094	0.6127188973
Pelagophyceae	0.2142989157	0.2931445254	0.2129819479	0.2961918697	0.0008232325	0.996815476	0.2345829453	0.2487051742	0.3663699059	0.0656430035	0.0796643046	0.6988688917	0.1071112096	0.6025035521	0.0069549759	0.9731006555
Phylloplaryngea	-0.3764170842	0.0580463862	-0.453210058	0.0200626119	0.4185425845	0.0333372845	-0.6610157857	0.0002367295	-0.4638981911	0.0169799455	-0.1042531263	0.6122745652	0.2201467808	0.2798518958	0.188654175	0.356024462
Placididea	-1	0	-1	3.20E-179	1	0	-1	0	-1	0	-1	0	-1	0	1	2.32E-177
Polycystinea	-1	2.32E-177	-1	0	1	7.27E-177	-1	2.32E-177	-1	1.77E-168	1	1.70E-173	-1	0	1	5.04E-168
Spirotrichea	0.0072813826	0.971838712	-0.0583696051	0.7769993934	0.0224992313	0.9131226936	-0.1292691636	0.5291016455	-0.0782014696	0.7041516463	-0.0909430678	0.6586109738	-0.2356386139	0.2465204341	0.105959812	0.6064319656
U.Alveolata	-0.0738649147	0.7198901595	-0.1576038518	0.4419370598	0.1723711054	0.3997732135	-0.2280745924	0.2624537495	-0.0074687283	0.971114431	0.1104373873	0.5912158766	0.1584731377	0.4393910909	-0.0073679951	0.9715038642
U.Bilateria	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Ciliophora	0.5200859685	0.0064600642	-0.2413649732	0.2348892132	-0.9913177727	1.16E-022	-0.9266579027	1.10E-011	0.8338189025	1.22E-007	-0.1362824512	0.5067936366	-0.0293322993	0.886890135	0.627655021	0.0005981485
U.Crustacea	-1	0	1	2.87E-170	-1	2.32E-189	1	1.59E-178	-1	0	-1	0	-1	0	-1	0
U.Eleutherozoa	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Eumetazoa	-0.2806967307	0.1648335563	0.113767885	0.5800056885	0.3060600407	0.1283569872	-0.1488198201	0.4681022094	-0.3025608877	0.1330048419	-0.2439905272	0.2296799487	-0.0159330302	0.9384231691	0.4684420101	0.0157929428
U.Rhizaria	0.2234579199	0.2724982918	-0.3126123458	0.1199733036	0.2014591164	0.3236964933	-0.0753518867	0.7144806397	0.2806630516	0.1648863855	0.3195123445	0.1115860353	0.3030285081	0.1323767548	-0.0470899633	0.8193119395
U.Rhodophyta	0.2408323203	0.2359555031	0.3871431789	0.0507097843	0.02238182	0.9135786115	0.1409906161	0.4920821534	0.0655220404	0.7504783098	0.2503366857	0.2174082827	0.1060429823	0.6061478407	0.001462821	0.9943413757
U.Stramenopiles	-0.4905559148	0.010947402	-0.3266656739	0.1033562503	0.2072613352	0.3096581072	-0.3248857839	0.105360432	-0.3198081684	0.1112363928	-0.0308308235	0.8811508064	0.3316808464	0.0978616438	0.2583732946	0.202511763
U.Viridiplantae	0.5247139033	0.0059224	0.4118513924	0.0365671633	-0.0370955439	0.8572233913	0.6037788174	0.0010910417	0.5703430636	0.0023483236	0.2689313079	0.1840214338	0.3295516358	0.100167077	-0.0794809492	0.6995303036

Table A.3: North Atlantic 18S rDNA correlation heatmap coefficients values and p-values

Taxonomy	Longitude correlation value	Longitude p-value	Latitude correlation value	Latitude p-value	Depth correlation value	Depth p-value	Temperature correlation value	Temperature p-value	Salinity correlation value	Salinity p-value	Nitrate.Nitrite correlation value	Nitrate.Nitrite p-value	Phosphate correlation value	Phosphate p-value	Silicate correlation value	Silicate p-value
Acantharea	0.1401339837	0.6640105487	-0.3410312765	0.2780016819	-0.1622069494	0.6144952157	-0.085820593	0.7908612089	-0.2978423736	0.347090103	0.2405401614	0.4513993974	0.2753090127	0.3864471445	0.3324061517	0.2911180692
Aconoidasida	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Actinopteri	0.2015995466	0.529796471	0.2240637866	0.4838755978	0.6173542095	0.0324532708	0.557124628	0.0598688306	0.3434633373	0.274365417	-0.0706610322	0.8272600817	-0.0515858485	0.8734998074	-0.4328908851	0.1598296228
Appendicularia	0.301736781	0.34051432	-0.3352565021	0.2867454771	-0.1850264366	0.5648160274	-0.0858261672	0.7908478908	-0.300694683	0.3422673058	0.2809963616	0.3763052226	0.2244010362	0.4832003941	0.1450519139	0.6528626201
Ascidiacea	-0.5647223086	0.0557432	-0.4137680173	0.1811898775	-0.2601663069	0.4141209728	-0.5157951879	0.0860680407	-0.3813867203	0.2212328358	0.6837982552	0.0142035151	0.6881889352	0.0133546612	0.4451183528	0.1470495126
Bacillariophyceae	-0.5415864495	0.068957932	-0.6636185665	0.0186252002	-0.3316034515	0.2923562112	-0.7169593139	0.0068666627	-0.626476354	0.0292784342	0.873359526	0.0002066205	0.8708931383	0.0002265591	0.4910425121	0.1049970004
Bangiophyceae	-0.2504417805	0.432396597	-0.5811880725	0.0474912889	-0.6109931374	0.034808451	-0.628830479	0.0284968583	-0.6375159015	0.0257431386	0.5894872861	0.0437278766	0.598781134	0.0396684254	0.3722956369	0.233356852
Bdelloidea	-1	0	-1	1.03E-075	-1	0	-1	0	-1	0	1	1.33E-079	1	2.14E-074	-1	1.33E-079
Chlorophyceae	-0.5181557695	0.0843934377	-0.0168308597	0.9585959467	-0.0754247157	0.8157852273	-0.1967540628	0.5399382588	-0.043305 4808	0.8936939509	0.5058251157	0.0933890336	0.2654960669	0.4042705517	0.2640144665	0.4069969239
Chrysophyceae	0.8171684152	0.0011731892	-0.2838074562	0.3713439034	0.2084101257	0.5156810913	0.4734887093	0.1199818209	-0.1090479658	0.7358494174	0.8649562158	0.0002807272	0.2708072585	0.3945730616	-0.9805520021	2.12E-008
Coccidia	0.3651821799	0.2431134714	-0.1700650522	0.5972054793	-0.2924224378	0.3563536986	-0.4090611948	0.1867066569	-0.359258 4894	0.2514189469	0.1473864062	0.6475933727	0.1888486924	0.5566572384	0.0063483895	0.9843778497
Compsopogonophyceae	-1	0	-1	0	-1	0	-1	0	-1	0	1	0	1	0	1	0
Coscinodiscophyceae	0.2646023703	0.4059139897	-0.2144324867	0.5033379868	-0.0304836096	0.925074612	-0.1705636174	0.5961148253	-0.1765882234	0.5829964082	0.1819085678	0.5715068036	0.355525878	0.2567365812	0.0220082154	0.94587412
Cryptophyta	0.4443590819	0.1478233402	-0.8863426899	0.0001230674	-0.6096088879	0.0353367278	-0.6578924689	0.0200451208	-0.8080548395	0.0014719725	0.6445719825	0.0236521622	0.9203822702	2.20E-005	0.6842631156	0.0141118079
Dictyochophyceae	0.6253500861	0.0296577684	-0.682566859	0.0144485688	-0.5326965751	0.0745641191	-0.3595927724	0.250936695	-0.6618514701	0.0190552721	0.1424358988	0.6587846396	0.4828892079	0.1117927005	0.0919230953	0.776312371
Dinophyceae	0.224208297	0.4835862217	0.3930374246	0.2062613002	0.5162577995	0.085738104	0.6412588107	0.0246179876	0.4492278966	0.1429063282	-0.4622555153	0.1302723272	-0.4318010954	0.161001737	-0.6048606379	0.0371923798
Echinoidea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Florideophyceae	-0.3981574503	0.1998827925	0.2963347484	0.3496538478	0.3974088072	0.2008078023	0.4182798303	0.1759981108	0.3485158995	0.26689892	0.0943596815	0.7705214825	-0.1376162035	0.6697422896	-0.3303598632	0.2942802664
Fragilariophyceae	0.5573771886	0.0597283775	0.5810495227	0.0475568926	0.5304698995	0.0760157327	0.5451983718	0.066765526	0.5152642521	0.0864477656	-0.6161085259	0.0329052036	-0.6112887168	0.034696384	-0.7416453204	0.0057620529
Glaucocystophyceae	-0.1519979961	0.6372281283	-0.4841489051	0.1107242443	-0.1943989131	0.5448969487	-0.598555744	0.0397624155	-0.620254605	0.0314182903	0.6322486044	0.0272041205	0.6515429259	0.0217103413	0.1478387882	0.6465739949
Gregarinasina	-0.136632165	0.6719866814	-0.3648032488	0.2436398485	-0.3374124565	0.2834630301	-0.2395954892	0.4532328206	-0.3613875253	0.2484149839	0.3967859627	0.2015793728	0.4048773372	0.1916969414	0.1739197396	0.5887929921
Gymnolaemata	0.3873197833	0.2135292338	0.7095140113	0.0097543203	0.5018524662	0.096419297	0.5853168818	0.0455654123	0.6838947123	0.0114184501	-0.6900772583	0.0130013685	-0.7182025115	0.0085172698	-0.7874769119	0.0023600515
Heterotrichea	0.3231990311	0.3054975095	-0.1340290447	0.677936425	0.222748554	0.4865129078	0.2493932477	0.4343902245	-0.0364919917	0.910354751	0.0020966993	0.9948401844	0.1519906261	0.6372446466	-0.2082577776	0.5159950363
Holothuroidea	1	4.15E-076	1	1.33E-079	1	0	1	0	1	0	-1	0	-1	0	-1	0
Hydrozoa	-0.159769653	0.6198951891	-0.2625125591	0.4097700401	-0.2238514397	0.4843009538	-0.0241846499	0.9405294787	-0.2265258551	0.4789562398	0.5826132697	0.0468201519	0.2757072444	0.3857324779	-0.2230286219	0.4859507676
Karyorelictea	0.8120026039	0.0013361367	0.3013346146	0.3411902592	0.3516968599	0.2622591161	0.2163816784	0.4993713785	0.135718495	0.6740731665	-0.3457695107	0.2709427217	-0.3080817338	0.3299458098	-0.2626211277	0.4095692636
Labyrinthulomycetes	0.261198079	0.4122048143	-0.0688816785	0.8315543096	0.0853690146	0.7919403034	0.0334936798	0.9176973131	-0.0648349918	0.8413360288	-0.1845603151	0.5658142811	0.0469932259	0.884692454	-0.0820507814	0.7998797474
Litostomatea	0.7082724448	0.0099414496	0.8629164134	0.0003014965	0.9639438446	4.52E-007	0.8538286953	0.0004089927	0.9089366249	4.23E-005	-0.8292868227	0.0008508353	-0.79354668	0.002064286	-0.7493903222	0.0050197589
Malacostraca	0.4762969784	0.1174954036	0.0972552463	0.7636537494	-0.1337083383	0.6786705911	0.130287091	0.6865184138	0.0022128143	0.9945544378	-0.2878967215	0.3641880508	-0.3191021933	0.3120205724	-0.7019004241	0.0109442166
Mamiellophyceae	-0.5519608862	0.0627910255	0.3217155451	0.3078507009	-0.0519196709	0.8726870574	-0.1070756397	0.7404804066	0.2414961639	0.1495475824	0.0783115331	0.808847395	-0.2235012184	0.4850028626	0.2798866341	0.3782732006
Mammalia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Maxillopoda	0.0351305361	0.9136880519	-0.0835600101	0.7962664565	-0.0832711428	0.7969577608	0.04334240471	0.8936046374	-0.0518996985	0.8727356805	0.1109386458	0.7314176944	-0.1020073238	0.7524167351	0.1900722771	0.5540556937
Mediophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nassophorea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nc.Alveolata	0.2647436741	0.4056539202	0.1156031635	0.7205165606	0.2915761031	0.3578119306	0.3477034697	0.2680914944	0.1488746607	0.6442419069	-0.2397843942	0.4528659107	-0.1635267887	0.6115783292	-0.6181673563	0.032160675
Nc.Chlorophyta	0.7846333782	0.0025092528	0.6885055424	0.0132493367	0.7979042233	0.0018700949	0.5934952005	0.0419146332	0.5784115933	0.0488181687	-0.6609395007	0.0192800453	-0.599086317	0.0395414126	0.0699428267	0.8289928567
Nc.Eukaryota	0.0866438293	0.7888948668	0.4426511972	0.1495734986	0.5671054456	0.0544911349	0.6359572323	0.0262224946	0.5159695231	0.0859436033	-0.4453921144	0.1467711416	-0.4557328321	0.136503267	-0.5216181435	0.0819774393
Nc.prasinophytes	0.2762629089	0.3847364177	-0.1858075481	0.5631447846	-0.4176851151	0.1766770601	-0.4073508448	0.188736828	-0.316113269	0.3168278141	-0.2456601904	0.4415241355	-0.0622005751	0.8477151557	0.0655852161	0.8395209782
Nc.Rhizaria	-0.1763691875	0.583471362	-0.2217230768	0.4885737192	-0.4692925242	0.1237611	-0.5447910051	0.0670103533	-0.344590644	0.2726892454	0.330353245	0.294290525	0.3146264065	0.3192343198	0.1770568692	0.5819807136
Nc.Stramenopiles	-0.2125356506	0.5072114231	-0.1481516129	0.6458694151	-0.0436632018	0.8928202616	0.0057095517	0.9859497609	-0.1012588841	0.7541836685	0.1896401866	0.5549738181	0.1143738888	0.7233848733	-0.0500129644	0.8773307729
Nc.Viridiplantae	0.1390328608	0.6665152491	-0.020338831	0.9499750046	-0.0742301257	0.8186597466	0.0402905424	0.9010617893	-0.0171604922	0.9577856786	-0.0981907225	0.7614382897	-0.0765648419	0.8130437056	-0.39214313	0.2073879827
Oligohymenophorea	0.5644251563	0.0559007151	0.6016440945	0.0384881211	0.8206834507	0.0010713825	0.7656206832	0.0037091175	0.666016394	0.0180539421	-0.6187282681	0.0319599479	-0.5881001207	0.0442987362	-0.3065430485	0.3324922156
Oomycetes	0.2205176981	0.4910011073	-0.3133077813	0.3213769124	-0.1687320782	0.6001252111	-0.1120115484	0.7289061656	-0.3203518057	0.3100228039	0.0753112862	0.8160580815	0.0490077242	0.8797804382	0.2380579632	0.4562243909
Ophiuroidea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Pelagophyceae	-0.5052947759	0.0937898095	0.7238135546	0.0077832748	0.5503539613	0.0637201988	0.5438276185	0.0675918126	0.6496254755	0.0222324354	-0.4299067336	0.1630521434	-0.7525793238	0.0047360905	-0.5551752338	0.0609606543
Phyllopharyngea	-0.9916144955	3.22E-010	0.6991693781	0.0113962461	0.5180160739	0.0844919175	0.5437047952	0.0676661927	0.6760159635	0.0158049934	0.3509443114	0.2633525414	-0.6017132278	0.0384600621	-0.6396955218	0.0250834969
Placididea	0.1414666292	0.6609834266	0.2510770058	0.4311909819	-0.0422761836	0.8962084845	-0.0087581124	0.978449037	0.1482430765	0.6456634608	-0.2007874762	0.5314904672	-0.2215915527	0.4888383174	-0.4052726648	0.1912219294
Polycystinea	-0.0459211433	0.887308078	-0.1540206406	0.6327005559	-0.1285669035	0.6904750799	-0.1317931225	0.683060252	-0.1673495304	0.6031592466	0.1887321217	0.5569053478	0.1982342238	0.5368315845	-0.1470520708	0.6483471067
Prostomatea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rhodellophyceae	1	1.58E-072	1	5.56E-068	1	4.92E-070	1	4.15E-071	1	0	-1	4.25E-073	-1	3.38E-071	1	1.66E-070
Spirotrichea	0.3973136137	0.2009256091	-0.1678529555	0.6020537962	0.0629607166	0.8458736282	0.1263812658	0.695512611	-0.1166685459	0.7180333345	-0.0299396929	0.9264082765	0.1376752917	0.6696075882	-0.3065192309	0.3325316549
Trebouxiophyceae	-0.1446987716	0.653660984	0.2434514092	0.4457715033	0.2649559746	0.4052633387	0.2813841698	0.3756187425	0.2344817778	0.4632184196	-0.0763963766	0.8134486729	-0.1886372317	0.5571073455	-0.6675402252	0.0176961934
U.Alveolata	0.2796586201	0.3786782163	0.0597630838	0.8536249259	0.3465409942	0.2698032426	0.2896397156	0.3611601602	0.1306953634	0.6855803869	-0.0849997658	0.7928229118	0.04152725119	0.8888910857	-0.4180898449	0.1762148274
U.Bilateria	0.8449746498	0.0005402853	0.5861594035	0.0451793159	0.7611327157	0.0040349382	0.7352579599	0.00643404	0.6211687778	0.0310970498	-0.6966584244	0.0118239402	-0.6397217316	0.0250756401	-0.6465891684	0.023077786
U.Chlorophyta	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Ciliophora	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Crustacea	1	0	-1	1.33E-079	-1	0	-1	0	-1	0	1	0	1	0	-1	0
U.Eumetazoa	0.0784053705	0.8086220867	0.2019483447	0.5290695798	0.4779507393	0.1160471765	0.2953614926	0.3513142347	0.2768647648	0.3836590496	0.4526628334	0.1395015647	0.1042626906	0.7470988369	-0.6620327555	0.0190108201
Ulvophyceae	0.0271732179	0.9331942013	-0.4673898252	0.1255001464	-0.5380238762	0.0711684713	-0.7922934277	0.0021228918	-0.6043706801	0.0373877334	0.6192477293	0.0317748579	0.6899060642	0.0130331095	0.8935280657	8.99E-005
U.Protostomia	0.6344353605	0.0266967332	0.7616299882	0.0039967502	0.8628664182	0.0003020202	0.7000803744	0.0111233383	0.700513919	0.0111720063	-0.695131395	0.0120897932	-0.6592770552	0.0196947688	-0.3541154893	0.2587628095
U.Rhizaria	0.3779278884	0.2257999478	-0.244298319	0.444140634	-0.0682076164	0.8331821782	0.1079723623	0.7383739265	-0.2122053741	0.5078872066	-0.1191949187	0.7121548266	0.0241874397	0.9405226293	-0.2060709626	0.5205105833
U.Rhodophyta	0.5115697238	0.0891215406	-0.1002022596	0.7566800133	0.0010362118	0.9974499512	0.2492250594	0.4347104239	-0.059758704	0.8536355511	0.3207318408	0.3297318408	0.0030152739	0.9925796893	-0.7456845468	0.0053653029
U.Stramenopiles	0.379989304	0.2230712636	-0.5312090979	0.0755317095	-0.3657376259	0.2423431099	-0.3185470368	0.3129103878	-0.5367670989	0.0719597715	0.1991627753	0.5348865373	0.3778639233	0.2258849366	0.1263201153	0.6956537164
U.Viridiplantae	0.2549773221	0.4238244987	0.4382673312	0.15412642	0.6176560035	0.0323444539	0.6973804122	0.0116997642	0.4530704897	0.139101002	-0.7164695049	0.0087540872	-0.5961703051	0.0407668948	-0.8368531089	0.0006874628

Table A.4: South Atlantic 18S rDNA correlation heatmap coefficients values and p-values

Taxonomy	Longitude correlation value	Longitude p-value	Latitude correlation value	Latitude p-value	Depth correlation value	Depth p-value	Temperature correlation value	Temperature p-value	Salinity correlation value	Salinity p-value	Nitrate.Nitrite correlation value	Nitrate.Nitrite p-value	Phosphate correlation value	Phosphate p-value	Silicate correlation value	Silicate p-value
Acantharea	-0.5608185541	0.023823551	-0.4580468169	0.0743885553	0.6336740462	0.0083980961	0.6830084566	0.0035434223	-0.1007676719	0.7103978157	0.1836667895	0.4959341214	0.1658195665	0.5393820906	0.3408227633	0.1964266912
Aconoidasida	-0.2109833599	0.4328250327	0.1939212756	0.4717471845	-0.7875637952	0.0002923218	0.5620906723	0.0234470319	-0.4876135433	0.0553717778	0.7123183925	0.0019607309	0.7063794021	0.0022227451	0.4714591628	0.0652524855
Actinopteri	0.5139494089	0.0416969049	0.4255470981	0.1003135394	0.2066145251	0.4426304082	-0.4512029959	0.0793918089	0.0959501975	0.7237249885	0.1218756301	0.6529700151	0.0525884205	0.8466277783	-0.1895199065	0.4819109464
Appendicularia	-0.3423540452	0.1942935909	-0.3156310287	0.2337100255	0.0231906794	0.9320637081	0.2750177177	0.3025931654	0.4084294378	0.1162740263	0.1528989159	0.57185215	0.1651908099	0.5409429766	0.0310631147	0.9090793222
Ascidiacea	-0.0036325847	0.9893473008	0.1108192644	0.6828465394	-0.3163847253	0.2325342903	0.1318457622	0.6264446996	-0.344017455	0.1919935968	0.3189427031	0.2285717777	0.418266721	0.1068997974	0.6272942229	0.0092931396
Bacillariophyceae	0.5341581432	0.0330574972	0.8441118166	3.93E-005	-0.0996656923	0.7134395807	0.2836459651	0.2870518947	-0.7007440805	0.0024969323	0.7024249859	0.0024124319	0.7096679108	0.0020743751	0.9599108039	4.01E-009
Bangiophyceae	0.4529275808	0.0781087961	0.4337232038	0.0932651127	-0.2905979268	0.2748852387	-0.5404663329	0.0306614363	0.7104159075	0.0020417757	0.2069979646	0.4417653342	0.198783796	0.4604839758	-0.4449174472	0.0841967128
Chlorodendrophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Chlorophyceae	0.3516048202	0.1817284249	0.3592651824	0.1717380065	-0.3049644428	0.2507501378	-0.2289874262	0.3936207216	-0.6047257411	0.0130797634	-0.4049224202	0.1197502891	-0.3579383916	0.1734416455	-0.0682288771	0.80176077339
Chrysomerophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Chrysophyceae	0.3800227862	0.1465231047	0.3362609471	0.2028713601	-0.5829681072	0.0177829342	-0.4177027214	0.1074224054	-0.1819571456	0.5000228364	-0.2927901788	0.2711144772	-0.1423563645	0.5989393475	-0.1525441243	0.5727552422
Cocidia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Compsopogonophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Coscinodiscophyceae	0.613225222	0.0115334758	0.5715838648	0.020717311	-0.4687036965	0.0670586511	-0.6905577693	0.0030617243	0.0795840876	0.7695388247	-0.2395032558	0.3716403453	-0.1129770566	0.6769793407	-0.2828350441	0.2884917637
Cryptophyta	-0.0480355401	0.8597779942	-0.009183823	0.9730719055	-0.2367527924	0.3773229292	0.1843906576	0.4942077374	-0.4175003462	0.1076103627	0.0731299262	0.7878134939	0.0523898736	0.8472004809	0.2129217707	0.4285105511
Dictyochophyceae	0.5377108975	0.0316912037	0.5255497286	0.0365532454	-0.4456689056	0.0836115492	-0.4063399154	0.1183366391	-0.3955661134	0.1293763729	-0.0384234023	0.8876508497	-0.0017839396	0.9947684215	-0.0794521557	0.7699112818
Dinophyceae	0.6396621384	0.0076214826	0.5644961893	0.0227246639	-0.1704777535	0.5278811894	-0.6261141004	0.009466662	-0.0686546029	0.8005469643	-0.3066707132	0.2479740065	-0.2523478151	0.3457344309	-0.5058301262	0.0456076159
Florideophyceae	-0.6076707697	0.0125266427	-0.6623139838	0.00518316	0.6288279161	0.0090713952	0.4795643014	0.0601460022	0.4503093628	0.0800625773	-0.17832578	0.5087598164	-0.2344794861	0.3820553497	0.2273825525	0.3970354644
Fragilariophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Gregarinasina	0.0327966913	0.9040264331	0.0737622218	0.7860184252	0.1590518602	0.556286967	-0.0164967173	0.9516477643	-0.1251913678	0.6441033238	-0.0548604521	0.8400793553	-0.0185319808	0.9456901049	0.311065575	0.2409117791
Gymnolaemata	-0.4782571509	0.0609489855	-0.421157808	0.1024289775	0.4870649096	0.0556879901	0.5504765095	0.0271334217	0.0018507871	0.9945723871	0.5129185103	0.0421789678	0.3653620284	0.1640525297	0.0772557376	0.7761188627
Heterotrichea	0.7214012473	0.0016089694	0.740442529	0.0010366306	-0.6267894193	0.0093670543	-0.6944166196	0.0028367691	-0.478032004	0.0610880845	-0.1461991938	0.5890061391	-0.0027215104	0.99201897	-0.1491418212	0.5814458644
Holothuroidea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hydrozoa	-0.5064854411	0.0452821857	-0.4869168828	0.0557735354	0.7110461686	0.0020146314	0.5366646276	0.0320890274	0.2344720554	0.3820708712	0.3452684648	0.1902755858	0.2531190978	0.3442122221	0.3500228589	0.1838382186
Labyrinthulomycetes	-0.4633987833	0.0706382956	-0.2555920297	0.3393573058	0.3572551207	0.1743233428	0.5942999245	0.0151932199	0.0472458805	0.8620624055	0.5474661558	0.0280977411	0.5268525424	0.0360070174	0.5020930657	0.0474969448
Litostomatea	-0.07678954	0.7774380931	-0.0570787205	0.8336953564	-0.054014518	0.8425163706	0.1878128432	0.4860848652	-0.1174172592	0.6649610314	0.2089959944	0.4372715423	0.173593604	0.5202508262	0.3969186099	0.1279529543
Malacostraca	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Mamiellophyceae	-0.7933379508	0.0002450573	-0.7504731304	0.0008100814	0.6773202973	0.0039452115	0.81865362	0.0001338586	0.0842538002	0.7563867099	0.1730338763	0.5216178037	0.0507175829	0.852026975	0.2508663897	0.3486688426
Mammalia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Maxillopoda	-0.5822193821	0.0179653863	-0.5252581474	0.0366763484	0.5142392017	0.0415621413	0.6068598975	0.0126770711	0.1924452303	0.4751926435	0.2334270185	0.3842572006	0.1937495925	0.4721473061	0.2703772657	0.3111528213
Mediophyceae	0.2090033306	0.4372550857	0.2337412093	0.3835991665	-0.5418921464	0.0301386823	-0.2281593144	0.3953807459	-0.2006178743	0.4562706529	-0.1780121866	0.5095176308	-0.2378136588	0.3751255224	-0.2685663705	0.3145312034
Nc.Alveolata	-0.0846275008	0.7553368462	0.0110105682	0.9677180461	0.3443621668	0.1915191974	0.1161410139	0.6684077743	0.3564421098	0.1753763211	0.3568404349	0.1748600132	0.359451632	0.1714994938	0.1823741629	0.4990240525
Nc.Chordata	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nc.Eukaryota	-0.3019216406	0.2557484258	-0.2453919272	0.3596339051	0.2290164152	0.3935591865	0.357842113	0.1735657047	0.0783954443	0.7728961801	0.4801534722	0.0597866291	0.4009561208	0.1237680952	0.421091154	0.1043095656
Nc.prasinophytes	0.5436226383	0.0295133745	0.3983362211	0.1264726445	-0.4717772109	0.0650463246	-0.632715817	0.0085279834	-0.2425852764	0.3653291389	-0.37655591	0.1505481243	-0.319342811	0.2279558643	-0.5524975252	0.0264604941
Nc.Rhizaria	0.099539499	0.7137881682	0.2344657866	0.3820839661	-0.846144455	3.61E-005	-0.0443366557	0.8704872373	-0.3435716042	0.1926083222	-0.1032117896	0.7036660434	0.0728062874	0.7887326831	-0.0779432244	0.774174483
Nc.Stramenopiles	0.1243636387	0.6463126204	0.2202549481	0.3994441623	-0.1792206223	0.5066002728	0.0349284983	0.8978175627	-0.4046948338	0.1199783467	0.1428814225	0.5975781709	0.0131470123	0.9614581703	-0.0025898881	0.9924049517
Nc.Viridiplantae	0.3373797794	0.2012782498	0.2059341733	0.4441674579	-0.3378203274	0.2006531807	-0.4253990986	0.1004444974	-0.0939424707	0.7293014945	-0.322393039	0.2232949637	-0.228579705	0.3944867431	-0.5679427687	0.0217306469
Oligohymenophorea	-0.98979375	3.01E-013	-0.9965785956	1.46E-016	-0.5691770366	0.0213829536	0.9024105355	1.74E-006	0.9544602365	9.65E-009	-0.9754896181	1.34E-010	-0.9896068498	3.42E-013	-0.9778201882	6.68E-011
Oomycetes	-0.2736517737	0.305098163	-0.017902623	0.9475320978	-0.3698870061	0.1584995495	0.5006084359	0.0482634922	-0.4343175902	0.0927668469	0.3553622808	0.1767813706	0.4257256352	0.1001557209	0.6120059303	0.0117459976
Pelinophyceae	0.6723306089	0.0043270841	0.7051366742	0.002281005	-0.013294277	0.961026752	0.3932594061	0.1318291462	-0.6612616057	0.005280439	0.5772080456	0.0192242652	0.5261742482	0.0362906289	0.830641016	4.76E-005
Pelagophyceae	0.4260810366	0.0098420854	0.4244339744	0.1013014577	-0.0225674074	0.9338857927	-0.262736954	0.3255510718	-0.1068023865	0.6938139738	0.0283877778	0.9168835304	0.0366123524	0.8929172594	0.0569883516	0.8339552433
Phyllopharyngea	0.0490976985	0.8567069159	0.2528657355	0.3447118402	-0.2924095077	0.2717669776	0.1953988522	0.4683104453	-0.5904489344	0.0160378943	-0.1269736863	0.6393556164	-0.0398536389	0.8834949198	0.6109131668	0.0119390688
Placididea	-0.2942238501	0.2686656125	-0.1121008268	0.6793597981	0.043836399	0.8719372732	0.4292718844	0.0970572902	-0.2023569617	0.4522934078	0.3030533823	0.25388151	0.2978041033	0.2626092574	0.2261038609	0.3997674945
Polycystinea	-0.7255474137	0.0014664671	-0.645325192	0.0069405529	0.4462016011	0.0831985105	0.7078871571	0.0021536862	0.3007716464	0.2576533579	0.4959197766	0.0507449104	0.37955777775	0.1470586872	0.366214179	0.1629969814
Rhodellophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Spirotrichea	0.3315336746	0.2096923736	0.2972959156	0.2634637664	-0.1083989282	0.6894479126	-0.2961388288	0.265415734	0.0517936446	0.8489207166	-0.2402772371	0.3700498107	-0.1808683363	0.5026350387	-0.0888930363	0.7433822048
Synurophyceae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Trebouxiophyceae	0.1159654051	0.6688825272	0.2349545388	0.3810637554	-0.3136115076	0.2368788013	0.1102088474	0.6845094191	-0.4189859055	0.1062360065	0.155007519	0.566497329	0.2175522473	0.4182946762	0.1936670423	0.4723397553
U.Alveolata	0.7921205461	0.000254455	0.7785231906	0.0003813119	-0.5295532489	0.0348943719	-0.6564695212	0.0057420319	-0.3895997516	0.1357858239	-0.2049142649	0.4464766913	-0.1514111176	0.5756432328	-0.2874448671	0.2803640621
U.Apicomplexa	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Blasteria	-0.1389895807	0.6076969723	-0.3114753006	0.2402598589	0.0274342179	0.9196669098	-0.0223164612	0.9346195007	-0.1402592258	0.6043884324	-0.5569293697	0.0250296223	-0.557703314	0.0247859604	-0.4832631098	0.0579159791
U.Chlorophyta	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Crustacea	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Eumetazoa	-0.1867594741	0.4885782779	-0.1461927368	0.5890227728	0.0821036868	0.7624348887	0.1868666139	0.4883243897	0.1829177112	0.4977236335	0.145130318	0.5917622563	0.0721486739	0.7906012207	-0.0791796869	0.7706806379
Ulvophyceae	-0.2588257105	0.3330680873	-0.2528461445	0.3447504896	0.4871962794	0.0556121522	0.3005802863	0.2579711859	0.2054565264	0.445248171	0.3092261958	0.2438520557	0.2735464991	0.3052917311	-0.3973235279	0.1275289138
U.Rhizaria	-0.7039010748	0.0023401514	-0.7357120253	0.0011601495	0.6163453293	0.0110033876	0.5984818237	0.0143157002	0.3238972667	0.2210188852	-0.140363399	0.6041172887	-0.1967579946	0.4651600857	-0.0790347415	0.7710899944
U.Rhodophyta	0.5900014012	0.01613838	0.6771359393	0.0039588211	-0.2925871017	0.2714624487	-0.0698939312	0.7970158911	-0.499647495	0.0487645359	0.3120702075	0.2393152604	0.3972819861	0.1275723728	0.5371929138	0.0318876856
U.Stramenopiles	0.6573488711	0.0056550026	0.5806303399	0.0183574225	-0.5030042783	0.0470309827	-0.7279881971	0.001387507	0.0609669853	0.822528923	-0.3864585361	0.1392462613	-0.2782404132	0.2967313935	-0.3458158749	0.1895269991
U.Viridiplantae	0.180684582	0.5030765213	0.3242699932	0.2204571962	-0.2249770164	0.4021833868	-0.0517953329	0.8489158446	-0.2964842264	0.2648321362	-0.0480548853	0.8597220434	0.1040854709	0.7012646519	0.1475649812	0.5854921119

Table A.5: Arctic 16S rDNA correlation heatmap coefficients values and p-values

Taxonomy	Longitude correlation value	Longitude p-value	Latitude correlation value	Latitude p-value	Depth correlation value	Depth p-value	Temperature correlation value	Temperature p-value	Salinity correlation value	Salinity p-value	Nitrate,Nitrite correlation value	Nitrate,Nitrite p-value	Phosphate correlation value	Phosphate p-value	Silicate correlation value	Silicate p-value
Acidimicrobiia	-0.3565515032	0.067916724	-0.4061800309	0.0355267612	0.1613444591	0.4214039353	-0.4168431789	0.0305379429	-0.2947986035	0.1355058259	0.1331339377	0.5079669276	0.5555357101	0.0026274545	0.3379228467	0.084718382
Actinobacteria	0.1244392701	0.536298397	0.5139433752	0.0061020117	-0.1646467806	0.4118341443	0.3538619787	0.0701666045	0.0178753947	0.9294832684	0.1570911706	0.4339079051	0.2102537229	0.2925068836	-0.0969261827	0.6305544689
Alphaproteobacteria	0.6996988068	4.87E-005	0.8242951881	1.26E-007	-0.2553024621	0.1987077366	0.8600695547	8.95E-009	0.6159847281	0.0006244876	0.1590229439	0.4282040455	-0.0478840625	0.8125208916	-0.4025909411	0.0373454816
Anaerolineae	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bacilli	-0.3919625352	0.0431705985	-0.076170627	0.70571654	-0.0174436039	0.9311826222	-0.3253878821	0.0976848035	-0.4636490868	0.0148556907	-0.3023288009	0.1253411351	-0.1170944071	0.5607979746	0.2563835475	0.1967474137
Bacteroidia	0.2270338905	0.2547741344	0.6398412517	0.0003257578	-0.1270602722	0.5276800024	-0.1225797364	0.5424529557	0.1345364102	0.5034671349	0.0688242609	0.7330252793	-0.144447702	0.4722397687	0.039023138	0.8467642602
Betaproteobacteria	0.529301589	0.0045253985	0.8383612716	4.80E-008	-0.1255095919	0.5327709431	0.7589525236	4.46E-006	0.475811586	0.0121208776	0.1311984572	0.5142092137	-0.0087532748	0.9654370108	-0.3965042075	0.0405991218
Chitinophagia	0.5307197465	0.0043991288	0.9087790957	5.59E-011	-0.9970855417	1.83E-029	0.62494431564	0.0004921382	0.2336166352	0.2408828902	0.9975698635	1.89E-030	0.9694373286	9.02E-017	0.7377982425	1.12E-005
Chlamydiia	0.4157205109	0.0310349699	0.3848344699	0.0474650601	-0.3145403513	0.1109601039	0.4067658498	0.0352367538	0.3926976249	0.0427458579	0.0379131547	0.8510743676	-0.3311565492	0.0912332958	-0.0333455296	0.8688533179
Clostridia	-0.0271276408	0.89315449	0.6780133897	0.0001018609	-0.6911247959	6.57E-005	0.6397865386	0.0003262646	0.387707918	0.0456950872	0.7233163053	2.02E-005	-0.864461971	6.16E-009	0.2826020855	0.1532157944
Cytophagia	-0.0031389608	0.9876023958	0.221636744	0.2665458419	-0.004371981	0.9827330862	0.1863120439	0.3521230179	0.0149470839	0.9410134614	0.0951400802	0.6368960831	0.2564050283	0.1967085979	-0.2439266084	0.2201499943
Deltaproteobacteria	-0.3556849448	0.0686362711	0.0162426032	0.9359108009	-0.039693835	0.8441619943	-0.0991090666	0.6228386504	-0.2794550426	0.1580417654	-0.1793040336	0.3708470026	0.0564717759	0.7796508035	-0.1233393098	0.5399349838
Epsilonproteobacteria	-0.458161806	0.0162466064	0.0409096677	0.8394487184	0.1117042083	0.5790945662	-0.5690644085	0.0019504144	-0.4625746216	0.015120004	-0.3759754499	0.0532642528	-0.1448819021	0.4708950642	0.1995075164	0.3184276337
Erysipelotrichia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Flavobacteriia	-0.096038491	0.6337030906	-0.4073672684	0.0349409985	0.0283304618	0.8864891112	-0.153638539	0.4442042134	-0.1334806908	0.5068525406	0.0793062973	0.6941655001	0.2063737602	0.3017076279	0.2662989272	0.1793861843
Fusobacteriia	0.2555044683	0.1983404265	0.3094721818	0.1162249025	0.1142394219	0.5704562502	-0.253088553	0.2027639546	0.0600942197	0.7658923222	0.0141539302	0.9441385899	-0.2232033037	0.2630934599	0.5727124719	0.0017959122
Gammaproteobacteria	-0.5784158289	0.0015754963	-0.5381444827	0.003785805	0.2540587638	0.2009794475	-0.6773597536	0.0001040551	-0.5204733406	0.0053827319	-0.1395335207	0.4875962327	0.0527597577	0.7938169932	0.1943434554	0.3313702446
Gemmatimonadetes <class>	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Halobacteria	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nc.Bacteria candidate phyla	-0.4751036343	0.0122676714	-0.4291737066	0.0254894677	-0.0856239318	0.6710965334	-0.4363091091	0.0228923052	-0.5988919712	0.0009649231	-0.2299886823	0.2484755038	0.1025307887	0.6108220078	-0.054135755	0.7885580629
Nc.Cyanobacteria	0.5288566184	0.0045656467	0.7909170257	9.09E-007	-0.1349727049	0.5020713368	0.6663734695	0.0001477375	0.424920203	0.0271481297	0.1255420943	0.5326639975	-0.0545820626	0.7868542467	-0.3216295761	0.1018481595
Nc.FCB group	0.1138013478	0.5719447877	0.1128217428	0.5752796528	-0.0312499634	0.8770314717	0.1014587334	0.6145765476	0.1916654083	0.33820602	0.343077051	0.0797833423	0.3883533827	0.0453047665	0.2262105643	0.2565475852
Nc.Thaumarchaeota	-0.3140648857	0.1106279203	-0.3765088836	0.0529000661	0.1145457479	0.5694164119	-0.4770934339	0.011858801	-0.2427668728	0.2224201356	0.2510177014	0.2066088862	0.4358686264	0.0230461225	0.5025328661	0.0075538533
Nc.unclassified Euryarchaeota	0.425887713	0.0267633765	0.622148283	0.0005305115	-0.2243303299	0.2606277191	0.6063849433	0.000799776	0.4377992339	0.0223781402	0.2952146776	0.1349290663	0.1984028957	0.3211696411	-0.0832288521	0.6798093719
Nc.Verrucomicrobia	0.1470357743	0.464254009	-0.3263574078	0.0966317796	0.053747975	0.7900392165	-0.1357363326	0.4996329655	0.2316164055	0.2450497937	0.1792327582	0.3710403625	0.0972981406	0.6292370082	0.1300496	0.5179319854
Negativicutes	-0.3389006414	0.0837647995	-0.0943193394	0.6398185765	-0.1573362497	0.4331819974	0.2555066588	0.198336646	0.4529263337	0.0176720562	0.7024882283	4.41E-005	0.6674434474	0.0001428687	-0.0329883599	0.8702463004
Nitrospinia	-0.474177816	0.0124618657	-0.3917492581	0.0432944557	-0.3683107107	0.0587171092	-0.6316017732	0.0004103322	-0.5802926444	0.0015082659	-0.3996499279	0.0388905593	-0.3407667161	0.0819675884	0.3101730053	0.1153576028
Nitrospira <class>	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Oligoflexia	-0.2413970896	0.2251216298	-0.3085415008	0.1173840794	0.6754688433	0.00011064	-0.3839478758	0.0480219588	-0.1303756228	0.5168742996	-0.0053068174	0.9790417248	0.0268271559	0.894331431	0.2042905402	0.3067214127
Oligosphaeria	0.5520899343	0.0028290338	0.0173370803	0.9316019015	-0.0220693776	0.9129942414	0.2167837383	0.2774257661	0.4244055663	0.0273546041	0.4070825052	0.0350807856	0.4988583525	0.0080791205	-0.1219894698	0.5444134638
Opitutae	0.0501756461	0.8037168657	0.5619657652	0.0022840595	-0.207457913	0.2991187248	0.3198712416	0.1038408893	-0.0308730246	0.8785038457	-0.210755815	0.2913293159	-0.3359258059	0.0866915428	-0.4006156199	0.0383777116
Phycisphaerae	-0.4843884949	0.0104554171	-0.157271725	0.4333730511	0.0897401543	0.6562195369	-0.4590439567	0.0160160259	-0.482908754	0.0107282766	-0.2927128004	0.138424104	-0.0137554301	0.9457090396	0.0599669665	0.7663744992
Planctomycetia	-0.4149675269	0.0313719643	0.0347486208	0.8633849148	-0.1430881304	0.4764631288	-0.3671500212	0.0595794286	-0.5025606527	0.0075499934	-0.4573728247	0.0164551449	-0.2468670329	0.214464306	0.0542992424	0.7879338253
Rubrobacteria	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Saprospiria	-0.2110152803	0.2907219603	0.1800594082	0.3688813877	-0.0241632204	0.9047748242	0.1365218132	0.4971309862	-0.1202182952	0.5503159892	-0.3295533264	0.0932209694	-0.2283765516	0.2518992352	-0.268642012	0.1754449509
Sphingobacteriia	-0.3577067918	0.0669664987	0.334915378	0.0877030554	0.0155946196	0.9384627168	0.0499113039	0.8047246114	-0.3247416931	0.0983913959	-0.4529050976	0.0176780408	-0.3235152505	0.0997430037	-0.1325754434	0.5097643369
Spirochaetia	1	0	-1	1.63E-129	1	0	-1	0	1	0	NA	1	0	1	0	0
Thermoplasmata	-0.1369521927	0.4957627671	-0.3837623048	0.0481391711	0.0409387551	0.839336024	-0.3529319503	0.0709612651	0.0134184178	0.9470373264	-0.1716013181	0.3920815058	-0.7222511506	2.10E-005	-0.343162124	0.0797037683
Tissierellia	0.2754626627	0.1643182265	0.3700439707	0.0574475419	0.2892441971	0.1433772353	0.2997609654	0.1287426064	0.4998912151	0.0079284674	0.7361866778	1.20E-005	0.6152219266	0.000637071	0.7583331329	4.59E-006
U.Acidobacteria	-0.0001842599	0.999272222	-0.3513369881	0.0723357051	0.7210713534	2.20E-005	-0.7870142384	1.12E-006	0.4839994805	0.0105265828	-0.8191245811	1.75E-007	-0.9846083258	1.85E-020	-0.4453366255	0.019919629
U.Bacteroidetes	0.3719445754	0.0560801219	0.012609871	0.9502246579	-0.0141450161	0.9441737174	0.3530885866	0.0708273766	0.4455582287	0.019850841	0.3535015621	0.070475308	0.4017891649	0.0377617541	-0.0660008714	0.743607762
U.Chlamydiae	-0.2265858422	0.2557382389	-0.5913619301	0.0011599355	-0.8266723706	1.07E-007	-0.8631995415	6.86E-009	-0.2449007888	0.2182551667	0.1350485145	0.5018290022	0.2525834949	0.2036971159	0.8255625738	1.16E-007
U.Chloroflexi	-0.3725580679	0.0556442069	-0.4799333628	0.0112949227	-0.2568530326	0.1959002535	-0.6515281932	0.0002320871	-0.4722722288	0.0128696125	-0.3478303861	0.0754293132	-0.2150368123	0.2814106592	0.3274685337	0.0954354975
U.Cyanobacteria	0.6918710101	6.40E-005	0.4703540513	0.0132911718	-0.3637157965	0.0621886482	0.6604879872	0.0001772265	0.597775333	0.0009919078	0.0994933575	0.621484286	-0.0017684082	0.9930153348	-0.1959391207	0.3273374395
U.Firmicutes	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
U.Gemmatimonadetes	0.0388295678	0.8475155946	-0.08226168	0.6833392411	0.3509662653	0.0726580804	-0.3248625415	0.0982589608	-0.0125871473	0.9503142467	-0.2119606853	0.2885157262	-0.2524753513	0.2038973152	0.3998760281	0.0387699979
U.Lentisphaerae	0.4166443967	0.0306254773	0.1969319766	0.324843306	0.9370095332	6.38E-013	-0.5112672596	0.006419387	0.7666592689	3.11E-006	-0.4685895646	0.013689001	-0.9307558647	2.01E-012	-0.8323091862	7.33E-008
U.Nitrospinae	0.9631736072	8.97E-016	0.8757768064	2.21E-009	0.3383408988	0.0843096786	0.3260823132	0.0969297081	0.9853059088	1.04E-020	0.371961786	0.0560678568	-0.3218280047	0.1016250858	-0.9607010904	1.99E-015
U.Planctomycetes	-0.3891563499	0.0448228869	0.0957258243	0.6348135998	-0.1236423819	0.5389318399	-0.2838283302	0.1513641346	-0.4780057672	0.0116751572	-0.4978292244	0.0082316116	-0.4598101288	0.0158179561	-0.013176458	0.9479910628
U.Proteobacteria	0.0384138645	0.8491295677	-0.4863656445	0.0100998995	0.2713605419	0.1709486381	-0.1834309871	0.3597515056	0.1645822393	0.4120200107	0.2848242878	0.1498720293	0.4059377529	0.0356472569	-0.0388818075	0.8473128156
U.Tenericutes	0.0849600047	0.6735076765	-0.2888740029	0.1439133043	0.418103896	0.0299874638	0.3984678523	0.0395257389	0.6575811034	0.0001936172	0.5348079901	0.0040517443	0.6132442558	0.0006707301	0.1946662681	0.3305519614
U.Verrucomicrobia	-0.6070610596	0.0007861632	-0.318440344	0.1054838691	-0.1183306686	0.5566389565	-0.6310165235	0.000417011	-0.6898967815	6.85E-005	-0.3300366519	0.0927131371	0.0036498365	0.9855843836	0.3648158067	0.0613434204
Verrucomicrobiae	-0.6405516157	0.0003192399	-0.1519859673	0.449178352	0.1841137108	0.3571385702	-0.4216711692	0.0284730737	-0.5461576614	0.003207135	-0.2796370802	0.1577597037	0.045666048	0.8210635196	0.1326956652	0.5093771619

Table A.6: North Atlantic 16S rDNA correlation heatmap coefficients values and p-values

Taxonomy	Longitude correlation value	Longitude p-value	Latitude correlation value	Latitude p-value	Depth correlation value	Depth p-value	Temperature correlation value	Temperature p-value	Salinity correlation value	Salinity p-value	Nitrate.Nitrite correlation value	Nitrate.Nitrite p-value	Phosphate correlation value	Phosphate p-value	Silicate correlation value	Silicate p-value
Acidimicrobiia	-0.5806598701	0.0232265271	0.4327345271	0.1071583055	0.6244420637	0.0128310595	0.3883477785	0.1525864988	0.5215060084	0.0461810862	0.1313177153	0.6408590873	-0.1371847135	0.6258783039	-0.1430164294	0.6111209557
Actinobacteria	-0.7777589116	0.0006411037	0.4574959204	0.0864039911	0.611118627	0.015505443	0.4623351972	0.0827082746	0.5348096385	0.0399657187	0.1234582836	0.6611297916	-0.2429140639	0.3830133676	-0.1170908732	0.6777148788
Alphaproteobacteria	0.1454419856	0.605023254	-0.0956578965	0.7345196879	-0.5794523955	0.0235835519	-0.2895026134	0.2952870009	-0.210365247	0.4517254794	-0.1423278026	0.6128564737	-0.0494770407	0.8609968026	0.3722313982	0.1718719267
Anaerolineae	-0.898973701	5.19E-006	-0.8452397003	7.26E-005	-0.7475429238	0.0013561346	-0.842546714	8.06E-005	-0.818302204	0.0001922776	0.868928243	2.62E-005	0.8821878057	1.35E-005	0.9933561337	1.36E-013
Bacilli	0.1439726613	0.6087141864	-0.4900973498	0.0636533493	-0.2916715046	0.2915192604	-0.2451777678	0.3784560184	-0.4595303155	0.0848363544	0.3432604125	0.2103522068	0.3947349688	0.1453690541	0.5056191516	0.0544662984
Bacteroidia	0.1119391455	0.6912351526	-0.3215101342	0.2425927169	-0.2002802968	0.4741822245	-0.1476734559	0.5994347725	-0.3292033968	0.2308594797	0.1164154862	0.6794822868	0.1137110702	0.6918357571	0.0237621223	0.9330109424
Balneolia	-0.8887232481	9.49E-006	-0.167099984	0.5516828808	-0.9214265092	1.07E-006	-0.7925170351	0.0004262258	-0.562864362	0.0289227832	-0.6703003616	0.0062465346	-0.2647634188	0.3402690358	-1	0
Betaproteobacteria	0.1183322691	0.6744703628	-0.5344755866	0.0401137141	-0.7300309763	0.0020018581	-0.5028270956	0.0560761447	-0.6028978134	0.0173580447	0.268085988	0.3340149377	0.3072251632	0.2655338022	0.5218232553	0.0460250417
Chitinophagia	1	4.01E-096	-1	6.51E-091	1	0	1	1.35E-098	-1	0	1	0	1	0	NA	NA
Chlamydiia	-0.0442944764	0.8754462396	-0.0288253096	0.9187766774	0.2473340536	0.3741423641	0.2178390703	0.4354317083	0.0267682913	0.9245570935	0.1058546381	0.7073153751	0.0413997361	0.8835322282	-0.4496873807	0.0926117789
Clostridia	-0.9673165651	4.00E-009	-0.8798095981	1.53E-005	-0.9247387042	8.16E-007	-0.8914594496	8.13E-006	-0.8628212685	3.46E-005	0.9201376289	1.19E-006	0.8919753653	7.89E-006	0.9145436397	1.82E-006
Cytophagia	0.5258093941	0.0440974272	-0.5397253506	0.0378345911	-0.3802480356	0.1620783072	-0.5540240616	0.0321167276	-0.583267956	0.0224694253	0.1435545419	0.6097661075	0.4115047632	0.127522838	0.4740874803	0.0742038259
Deinococci	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Deltaproteobacteria	-0.2753487387	0.3205729313	0.2874057989	0.2989566375	0.7131273644	0.0023394755	0.4703485045	0.0768381748	0.392978631	0.1473303374	-0.00079114436	0.9776758978	-0.1528607291	0.5865240928	-0.2950976173	0.2856256867
Epsilonproteobacteria	0.000858271	0.99757779006	-0.3973908805	0.1424367121	-0.2817365195	0.3090114515	-0.4108583216	0.1281815584	-0.3979746163	0.141797609	0.3372797769	0.2189302577	0.3501482302	0.2007415421	0.8583155068	4.23E-005
Erysipelotrichia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Flavobacteriia	0.4086582551	0.1304408115	-0.5962757438	0.0189703832	-0.6901771784	0.0044015857	-0.6612063389	0.0072715586	-0.6772274189	0.0055451205	0.19127362	0.4946829259	0.3957409231	0.1442536635	0.5095147783	0.052370627
Fusobacteriia	1	0	-1	3.88E-103	-1	4.90E-100	1	2.35E-087	-1	0	-1	0	-1	0	NA	NA
Gammaproteobacteria	0.2071933635	0.4587308268	-0.731071863	0.0019576392	-0.5453081452	0.035518413	-0.5134582251	0.0502716973	-0.7130248027	0.0028452971	0.552777736	0.0325874974	0.6387471098	0.0103732138	0.421893633	0.1172522386
Gemmatimonadetes <class>	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Halobacteria	-0.7259252958	0.0021841636	-0.5972408031	0.0187285317	-0.4496161326	0.0926698284	-0.5826631746	0.0226432877	-0.5582170248	0.0305704052	0.6259643769	0.012549905	0.6573145632	0.00774865	0.9194557276	1.25E-006
Mollicutes	-0.2484054717	0.3720089769	0.2581605989	0.3528918877	0.2662304029	0.3374997119	0.2975146338	0.2815108753	0.3072985836	0.2652179315	-0.5747145152	0.025024927	-0.2677916833	0.3345662737	-0.0214800495	0.9394329966
Nc.Bacteria candidate phyla	-0.3883069488	0.1526333914	-0.7840037384	0.0005414307	-0.2423161923	0.3842219295	-0.5428252449	0.0365350193	-0.7264069926	0.0021621102	0.6874683003	0.0046237288	0.695677596	0.0039765893	0.9167947172	1.54E-006
Nc.Bacteroidetes	0.7723569059	0.0007389092	-0.7693407094	0.0007085786	-0.6596250503	0.0074625391	-0.4934770188	0.0615732382	-0.7361778512	0.0017519445	0.8168705624	0.0002015986	0.7090801817	0.0030765448	0.8620495733	3.59E-005
Nc.Cyanobacteria	-0.2645753516	0.3406251334	0.9517729333	4.83E-008	0.7262288706	0.0021702442	0.8838493466	1.24E-005	0.9366797368	2.74E-007	-0.7643737424	0.0009053192	-0.9182875382	1.37E-006	-0.7935914707	0.0004132321
Nc.FCB group	-0.8425927321	8.05E-005	0.5375593301	0.0387629138	0.459381312	0.0849504809	0.3406454707	0.214076113	0.5587934962	0.0303627264	0.0768576012	0.7854326007	-0.3244103948	0.2381270072	-0.0855260919	0.7618441594
Nc.Thaumarchaeota	-0.5916396763	0.0201657688	-0.0665569429	0.8136878427	0.1230821621	0.6621055028	-0.2332334379	0.4028327192	-0.0166286925	0.9530964119	0.5559476846	0.031400184	0.3038535187	0.2708883234	0.4770448165	0.0721665767
Nc.unclassified Euryarchaeota	-0.1134290636	0.6873315876	0.0844532287	0.7647535098	0.4463527785	0.0953562198	0.0522258056	0.853349301	0.1417039822	0.6144303125	0.134063734	0.6338309594	0.0792616157	0.7788725833	-0.2708406886	0.3288794226
Nc.Verrucomicrobia	0.0989141803	0.7257985438	-0.1609470224	0.5666280684	-0.2854144878	0.3024662707	-0.3065800056	0.266394694	-0.1589039656	0.5716279684	0.1966243253	0.4824539254	0.2990879243	0.2788515535	0.4576011504	0.0863224044
Negativicutes	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Nitrospinia	-0.5041588561	0.0553234088	-0.4242783357	0.11497773386	-0.2923381907	0.2903668457	-0.5028220423	0.0560790151	-0.3715325155	0.1727201295	0.8463724713	6.94E-005	0.6342088478	0.0111092817	0.7048012352	0.0033441039
Oligoflexia	0.362567372	0.1841352053	0.0329550306	0.9071834293	0.4195903521	0.1194783462	0.3312574841	0.2277878089	0.0906752183	0.7479227062	-0.090237985	0.7491020928	-0.0879376557	0.7553154152	-0.5185862539	0.0476356385
Oligosphaeria	1	0	-1	1.83E-088	-1	3.88E-103	-1	7.71E-091	-1	3.07E-078	-1	2.88E-093	1	7.96E-092	NA	NA
Opitutae	0.3955391099	0.1444769682	0.5491564877	0.0339847709	0.5264381044	0.0437989441	0.7056308484	0.0032908313	0.5649304893	0.0282122045	-0.7793160691	0.000614957	-0.6193418332	0.0138087593	-0.9157466664	1.66E-006
Phycisphaerae	0.0921880919	0.7438459181	-0.3360240144	0.2207590089	-0.0375414489	0.8943285482	-0.1318353062	0.6395321822	-0.3267814571	0.2345142935	0.1961361149	0.4835636986	0.1828320025	0.5142674005	0.0494689317	0.8610193806
Planctomycetia	-0.1349808998	0.6314899948	0.8377061506	9.70E-005	0.9316575702	4.44E-007	0.9108190326	2.38E-006	0.8903423056	8.66E-006	-0.6375601538	0.0105619429	-0.7248721186	0.0022330097	-0.8259376295	0.0001483325
Rubrobacteria	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Saprospiria	0.761614764	0.0009694272	-0.2682120865	0.3337788686	-0.1372711594	0.6256585674	0.0044255234	0.9875113025	-0.2863048736	0.3008940247	-0.2961329652	0.2838587243	0.013295181	0.9624921972	-0.2003208455	0.4740908684
Solibacteres	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Spartobacteria	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Sphingobacteriia	0.9139389912	1.90E-006	-0.7068450322	0.0032140861	-0.4647711408	0.0808911652	-0.5725007001	0.0257209071	-0.7104309016	0.0029957324	0.0389089689	0.8927941213	0.5885209583	0.02100017636	0.7027520844	0.0034786375
Thermoplasmata	-0.9097759329	2.56E-006	-0.9535363947	3.81E-008	-0.7693837522	0.0007977005	-0.8199030404	0.0001822768	-0.9111659558	2.32E-006	0.9429288484	1.41E-007	0.9476431938	8.16E-008	0.9555668589	2.86E-008
Tissierellia	0.7042945734	0.0033769752	0.6190488858	0.0138666176	0.937714269	2.46E-007	0.8032015934	0.0003107479	0.6947707045	0.0040443172	-0.6148451196	0.0147177223	-0.6315639044	0.0115566413	-0.9137675338	1.92E-006
U.Acidobacteria	-0.853658793	5.15E-005	-0.6853702916	0.0048181104	-0.4663474393	0.0797305892	-0.6825478704	0.0050499077	-0.6219919503	0.0132937993	0.762469568	0.000949184	0.7857188734	0.000516389	0.9686377339	3.07E-009
U.Bacteroidetes	0.5744837399	0.0250968029	-0.4188437526	0.1202061271	0.0624065815	0.8251350188	-0.1958094741	0.48430688	-0.3742621892	0.1693390805	0.0436549564	0.8772318989	0.3158730289	0.2514195073	0.2308836603	0.407723542
U.Chlamydiae	0.3554333898	0.1935612479	0.0171707109	0.9515691982	0.2777400935	0.3162160601	0.3359454879	0.2208736847	0.0822701087	0.7706825984	-0.0944657729	0.7377200982	-0.0804381584	0.775667074	-0.4731884549	0.0748312409
U.Chloroflexi	-0.403680525	0.1356522216	0.1066543813	0.7051950738	0.6221739364	0.0132589905	0.2080286538	0.4568808527	0.2638725395	0.3419572445	0.2222230276	0.4260150387	0.1991424925	0.4767491559	-0.0143458279	0.9595303108
U.Cyanobacteria	0.2436685509	0.3814911431	0.1140540943	0.6856739036	0.4612166106	0.0835523472	0.2550475586	0.3589320799	0.2095209847	0.4535848827	-0.1290247215	0.6467495468	0.0243821479	0.9312667551	-0.1323748024	0.6381501988
U.Firmicutes	1	0	-1	5.00E-090	NA	NA	-1	0	-1	1.18E-079	1	0	1	0	1	0
U.Gemmatimonadetes	-0.4162654131	0.122742904	-0.1979569019	0.4794309918	0.1315276112	0.6403208726	-0.1881344991	0.5019244355	-0.1458610296	0.6039722379	0.3281044125	0.2325134605	0.3028300796	0.2725867996	0.513441 9952	0.0502802069
U.Lentisphaerae	0.8303443065	0.0001269819	-0.7957532371	0.0003880674	-0.1951621712	0.4857812459	-0.124414654 9	0.6586511145	-0.6676014757	0.0065380201	0.263698029	0.3422884651	0.1781498306	0.525280324	-0.7849651377	0.000527275
U.Planctomycetes	0.2526092658	0.3637027657	-0.1639179118	0.5593906307	0.1279464675	0.6495262356	0.1123485112	0.6901575734	-0.1492173145	0.5955804246	-0.0745059629	0.7918626879	-0.0125874431	0.9644876448	-0.3295745326	0.230302581
U.Proteobacteria	-0.7143638583	0.002770036	0.7188397836	0.0025299388	0.776866209	0.0006564961	0.692983695	0.0041804557	0.7726754912	0.0007328241	-0.1934564142	0.489676657	-0.5428108682	0.0365409684	-0.3222882933	0.2413894949
U.Tenericutes	0.9108288795	2.38E-006	-0.9311565234	4.65E-007	-0.0060233062	0.9830029264	-0.0151147196	0.9573630318	-0.7255283621	0.0022024716	0.8962919133	6.12E-006	0.3215378813	0.2425497496	-1	0
U.Verrucomicrobia	-0.1440369844	0.608552423	0.5663790416	0.027721946	0.9163080816	1.59E-006	0.7574771013	0.0010724169	0.6636814494	0.0069803615	-0.3389343657	0.2165353398	-0.4315398169	0.108239952	-0.6367227428	0.0106966956
Verrucomicrobiae	0.2852267779	0.3027983412	0.4830780694	0.0681355997	0.6674187467	0.0065581357	0.7181300547	0.0025668555	0.5067710801	0.0538684364	-0.6616859609	0.0072143983	-0.5877223249	0.0212200398	-0.781423797	0.0005809537

Table A.7: South Atlantic 16S rDNA correlation heatmap coefficients values and p-values

Taxonomy	Longitude correlation value	Longitude p-value	Latitude correlation value	Latitude p-value	Depth correlation value	Depth p-value	Temperature correlation value	Temperature p-value	Salinity correlation value	Salinity p-value	Nitrate,Nitrite correlation value	Nitrate,Nitrite p-value	Phosphate correlation value	Phosphate p-value	Silicate correlation value	Silicate p-value
Acidimicrobiia	-0.8642133826	3.25E-005	-0.7744082461	0.000700436	0.6286554727	0.0120646072	0.8972702433	5.76E-006	0.1257366678	0.6552302114	0.4598523656	0.0845900581	0.3665476601	0.1790082621	0.415534829	0.1234683364
Actinobacteria	-0.8076409733	0.000271004	-0.7440618479	0.0014687904	0.398170468	0.1415836149	0.8027254736	0.0003152792	-0.0593116494	0.8336926566	0.3710872786	0.1732757862	0.3454003913	0.2073355309	0.3303518126	0.2291389745
Alphaproteobacteria	0.2945956686	0.2864846764	0.2899828826	0.2944502308	-0.1299293592	0.6444232793	-0.2550254913	0.3589751532	0.0862395753	0.7599109963	0.1066941684	0.7050896415	0.186591416	0.5055020844	0.1550581733	0.5810892787
Anaerolineae	-0.7568900904	0.0010877186	-0.9550148583	3.10E-008	0.2356953725	0.3977419222	0.4567479093	0.0869855134	0.3557059884	0.1931954597	-0.2800725442	0.3119996037	-0.3139449521	0.2544831224	-0.1788423267	0.5236448519
Ardenticatenia	-0.4545275723	0.0887279858	0.9522257956	4.55E-008	-0.1202600251	0.6694425223	0.9867475894	1.19E-011	-0.9912491008	8.08E-013	0.9996435347	7.60E-022	0.9993783116	2.82E-020	0.999014111	1.04E-012
Bacilli	-0.0954894146	0.7349717549	-0.2475499534	0.3737119345	0.4278966167	0.1115851684	-0.075035929	0.7904124989	0.6231576958	0.0130720462	-0.2071652381	0.4587931822	-0.2620723775	0.3453826045	-0.0288022098	0.9188415694
Bacteroidia	-0.4163682909	0.122640987	-0.5414956339	0.0370882801	0.6286444887	0.0120665579	0.2621750259	0.3451867695	0.655792411	0.0079417961	-0.0812740034	0.7733918394	-0.2128933536	0.4461803198	-0.3032172643	0.2719434855
Balneolia	-0.1024139165	0.7164601276	-0.1723829744	0.5389884279	-0.0650551593	0.8178260171	-0.0237271424	0.9331093524	0.1213347227	0.6666451991	-0.3012302491	0.2752546371	-0.2902907019	0.2939146547	-0.5358157966	0.0395224143
Betaproteobacteria	0.6777817446	0.0054918038	0.494660504	0.0608566443	-0.5045659478	0.0550947951	-0.8032741348	0.0003100622	0.0176903137	0.9501052908	-0.3605254473	0.1868020514	-0.2860274501	0.3013833858	-0.6728382797	0.0059818765
Chitinophagia	-1	0	1	1.98E-087	-1	1.35E-090	1	0	1	4.16E-078	-1	0	-1	0	-1	0
Chlamydiia	0.4997943315	0.0578182127	0.2771645655	0.317261493	-0.0780796688	0.7820961491	-0.6452387246	0.0093870648	0.1258183109	0.6550191557	-0.5341499048	0.0402583937	-0.4876918022	0.0651646269	-0.5325927049	0.0409555296
Clostridia	0.8223611442	0.0001677544	0.7650439667	0.0008902766	-0.2182810909	0.4344775024	-0.8158675037	0.0002083474	-0.054528209	0.8469529345	-0.3437768048	0.2096217215	-0.2669294182	0.3361845645	-0.345130208	0.2077148718
Cytophagia	0.181604355	0.5171446963	0.2560904817	0.3569022178	-0.0982971574	0.7274487221	-0.1609527554	0.5666140642	0.2887262967	0.2966425285	0.4191372269	0.1199196874	0.5137879185	0.0500990629	0.3361031993	0.2206434096
Deinococci	0.174305423	0.5344012233	0.0239458486	0.9324940741	0.4786225158	0.0710963329	-0.1588844076	0.571675921	0.2907444856	0.293126162	-0.1104091474	0.6952674449	-0.2662567712	0.3374500491	-0.5624260487	0.0290752487
Deltaproteobacteria	-0.7177345606	0.0025876135	-0.6524610908	0.00837776875	0.3620719288	0.1847799879	0.7438694025	0.0014752317	0.0483791077	0.8640546696	0.2175676184	0.4360182275	0.0767021936	0.7858571393	0.2665898916	0.3368229971
Epsilonproteobacteria	-0.1415808004	0.6147412732	0.0681409262	0.8093281178	0.1449006541	0.606382049	0.3157034721	0.2516880133	-0.3396386054	0.2155210424	0.5687065442	0.0269477383	0.5498274364	0.0337225526	0.3712855243	0.1730282304
Erysipelotrichia	-1	1.26E-086	-1	3.59E-084	NA	NA	1	0	1	2.13E-082	-1	6.72E-087	1	3.20E-087	1	5.45E-087
Flavobacteriia	0.5625013561	0.0290490103	0.714685954	0.0027521723	-0.5938440838	0.0195903963	-0.3932848516	0.1469871123	-0.4439637723	0.0973572425	0.1253930835	0.6561186781	0.2956134361	0.2847445569	0.1774694777	0.5268893515
Fusobacteriia	0.9414677521	1.66E-007	0.8317062006	0.0001209205	-0.7449945437	0.0014378926	-0.7735830753	0.0007157116	-0.9985583642	6.67E-018	0.8712263815	2.35E-005	0.8838307679	1.24E-005	0.9516439416	4.92E-008
Gammaproteobacteria	0.7671712557	0.0008438675	0.7189763633	0.0025228836	-0.3567601324	0.1917851496	-0.7022507858	0.0035121968	-0.2553233999	0.3583946244	-0.2136250221	0.4445818613	-0.190069769	0.4974542212	-0.3139993449	0.2543963836
Halobacteria	-0.2874476037	0.2988832147	-0.1056965578	0.707734719	-0.5195295289	0.0471620772	0.5533936714	0.032354199	0.8260472443	0.0001477678	-0.1726398463	0.5383744968	-0.2194596054	0.4319385977	-0.8650191619	3.14E-005
Mollicutes	0.0227066494	0.9359807284	0.036820709	0.8963471106	-0.1641166536	0.558907884	-0.0420774674	0.8816388761	-0.1776972001	0.5263505442	-0.1822300411	0.515677331	-0.2150972779	0.4413741846	0.0735862742	0.7943880476
Nc.Bacteria candidate phyla	-0.3400167172	0.2149777009	-0.1480591644	0.5984708947	0.112602926	0.6894881517	0.5351024201	0.0398363418	-0.3536458264	0.1959709699	0.2774669103	0.3167120428	0.2756488248	0.3200243141	0.3575925635	0.1906761762
Nc.Bacteroidetes	0.5273118766	0.043386606	0.5058277545	0.054390583	-0.3419393945	0.2122282678	-0.5661643002	0.0277942148	0.1102294201	0.6957416142	0.0306293168	0.9137103419	0.0635115461	0.8220841093	-0.3249008559	0.2373768916
Nc.Bacteroidetes/Chlorobi group	-1	1.23E-082	1	0	-1	9.85E-091	1	0	-1	3.11E-083	1	0	1	0	1	0
Nc.Cyanobacteria	-0.2079304823	0.4570980885	-0.2009925335	0.4725787955	-0.0263193864	0.9258190338	0.2086328277	0.4555450484	-0.177471805	0.5268838436	-0.1229002905	0.662577484	-0.0723216744	0.7978463726	0.3405816054	0.2141675818
Nc.FCB group	-0.6389505674	0.0103411293	-0.7967610221	0.0003767714	0.5517953629	0.0329622042	0.4067785825	0.1323924218	0.5406192609	0.0374563514	-0.2825398803	0.3075747778	-0.3850103117	0.1564513896	-0.1511385039	0.5907979859
Nc.Thaumarchaeota	-0.7296269334	0.0020192372	-0.781105443	0.0005859891	0.7364625943	0.0017410072	0.5996470714	0.0181357986	0.4984901335	0.0585793752	0.2201491103	0.4304566915	0.0608262522	0.8295024828	0.0151854978	0.9571635426
Nc.unclassified Euryarchaeota	-0.7555242038	0.0011240075	-0.5865869576	0.0215333269	0.3930376624	0.1472641306	0.8457207589	7.12E-005	0.0041013329	0.9884261	0.4351389059	0.1050043096	0.4168966403	0.122118487	0.5916468753	0.020163869
Nc.Verrucomicrobia	0.6975981416	0.0038361362	0.7074865653	0.003174117	-0.5103532117	0.0519190768	-0.5308031242	0.0417677377	-0.5871620185	0.0213742095	-0.358159942	0.1899226903	-0.3105137772	0.2599912562	-0.2452901139	0.3782306079
Negativicutes	0.0524513672	0.8527222815	-0.1577689404	0.5744136164	0.1311136543	0.6413824993	-0.2635326352	0.3426025488	0.628089231	0.0121654883	-0.0548737624	0.8459936989	-0.1965413671	0.4826424157	-0.4027401066	0.13665241
Nitriliruptoria	0.6094593658	0.0158664723	0.6374376587	0.0105815712	-0.037311951	0.8949712338	-0.5712795518	0.0261110412	0.4968141069	0.0595682375	-0.6336748597	0.0111984945	-0.5839700678	0.022268861	-0.402892609	0.1364898765
Nitrospinia	-0.7216632059	0.0023872627	-0.6767986542	0.0055866404	0.1915920111	0.4939511968	0.5498439965	0.0337160999	0.3073589572	0.2651192056	-0.0071811005	0.9797363328	-0.1134108573	0.6873637239	0.0642441581	0.8200626006
Oligoflexia	0.1428982628	0.61141863	0.1086825267	0.6998270475	0.2596457996	0.3500301943	-0.2149678843	0.4116556327	0.5213895036	0.0462384895	0.2285189341	0.4126764619	0.1691227828	0.5468070508	-0.1470569341	0.6009767305
Oligosphaeria	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Opitutae	-0.0122701921	0.9653821896	-0.1360292663	0.6288181303	-0.0930707529	0.7414702892	-0.1345101277	0.6326911846	0.0194378187	0.9451830528	-0.2227164304	0.4249618058	-0.1727978887	0.5379969242	-0.5385455108	0.0383381706
Phycisphaerae	-0.3261016148	0.235546639	-0.4586330899	0.0855252162	0.2709941963	0.328594573	0.1111653212	0.6932736028	0.368171713	0.1769434216	-0.2029852058	0.4681067612	-0.3138473827	0.2546387589	-0.5055355293	0.0545530837
Planctomycetia	-0.4128807737	0.1261283999	-0.5116889362	0.0512056167	0.3130813547	0.2558626995	0.2680761334	0.3340333905	0.1357165746	0.6296146102	-0.1307797846	0.6422392059	-0.234923877	0.3993335423	-0.3622590804	0.1845362516
Rubrobacteria	1	2.70E-085	-1	6.15E-089	1	3.51E-101	-1	0	1	0	-1	3.88E-103	-1	3.88E-103	-1	4.90E-100
Saprospiria	0.2020898264	0.4701136532	0.3971397169	0.1427122953	-0.4690267391	0.0777852524	-0.0434258524	0.8778717342	-0.2506080938	0.3676440466	0.0948582776	0.7366659243	0.1493077377	0.5953549854	0.3509545331	0.1996352397
Solibacteres	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Sphingobacteriia	0.500393356	0.0574710357	0.5730909801	0.0255339166	-0.4289972588	0.1105670971	-0.4188285167	0.1202210107	-0.0157300775	0.9556287184	0.1608165405	0.5669468399	0.1686159294	0.5480270075	0.0783609214	0.7813287784
Spirochaetia	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Thermoplasmata	0.4591138536	0.0851556091	-0.9375729312	2.50E-007	0.5377437762	0.0386832083	-0.9571657113	2.27E-008	0.9205678414	1.15E-006	-0.9741683495	8.82E-010	-0.9410592675	1.73E-007	-0.9729583476	1.18E-009
Tissierellia	-0.1022667677	0.7168520278	-0.2124578791	0.4471330538	0.0423885438	0.8807695736	-0.0696049242	0.8053032572	0.5939886283	0.0195531129	-0.3148854865	0.2529858238	-0.3075551324	0.2647985638	-0.2403046952	0.388303074
U.Acidobacteria	-1	4.43E-098	-1	4.43E-098	1	3.49E-093	1	0	1	6.93E-086	-1	1.21E-097	-1	3.18E-099	-1	0
U.Bacteroidetes	0.3447331527	0.2082731448	0.3005752635	0.2763513758	-0.5403453654	0.0375719433	-0.3383537258	0.2173738862	-0.231529271	0.4063765944	0.0223946519	0.9368587482	0.0482821532	0.8643247856	-0.3885426368	0.1523628381
U.Chlamydiae	0.6609146228	0.0073064984	0.713554968	0.0028153064	-0.3014004048	0.2749701484	-0.4887645777	0.0644874847	-0.1295131759	0.6454931121	0.1901076082	0.4973670037	0.1828516928	0.514221311	-0.1874755665	0.5034507325
U.Chloroflexi	-0.3720826859	0.172035134	-0.5476911818	0.034562751	0.7053034902	0.0033117708	0.1170162616	0.6779100524	0.7691058368	0.0008033844	-0.1709490266	0.5424212584	-0.2911022388	0.2925054122	-0.3022801769	0.2735020438
U.Cyanobacteria	-0.1273861424	0.6504559919	-0.1714501476	0.5412205606	-0.119592609	0.6711817696	0.1301648928	0.6438181078	-0.1782256351	0.525102602	-0.0473600463	0.8668944742	0.0015614409	0.9955935209	0.1353177473	0.6306310423
U.Firmicutes	-0.2039649857	0.4659154851	-0.3032210345	0.2719549741	0.1159280558	0.6807587979	0.3341637807	0.2234856779	0.3676639205	0.1775873521	-0.2139167802	0.4439452665	-0.3050111287	0.2689748803	0.062440163	0.8250422625
U.Gemmatimonadetes	-0.0422556515	0.8811409256	-0.2552712745	0.3584961628	0.6962818948	0.0039319622	-0.2091150109	0.4544803488	0.8147707019	0.0002159379	-0.1572815266	0.5756115876	-0.2328962462	0.4035326258	-0.4475675452	0.0943499073
U.Lentisphaerae	-0.6150340133	0.0146786333	-0.489775197	0.0638542525	0.1546907849	0.5819964756	0.5315764147	0.0414153224	0.2064422264	0.4603975672	-0.0340841363	0.9040166544	-0.0559108889	0.8431159073	0.0944421939	0.7377834395
U.Planctomycetes	-0.2836751543	0.3055514126	-0.4422201974	0.0988361321	0.3491787283	0.2020769589	0.0906177257	0.7480777562	0.4270009828	0.1124183903	-0.1364914091	0.6276416675	-0.2933865663	0.2885601113	-0.4999107699	0.0577506089
U.Proteobacteria	-0.6736426419	0.0058898701	-0.7632565901	0.00093085	0.3829162175	0.1589094148	0.4464738093	0.0952556209	0.4482186157	0.0938136483	0.0593633911	0.8335494432	0.0212345013	0.9401242145	-0.073157672	0.7955549713
U.Tenericutes	-0.0493635765	0.861312729	-0.0434582274	0.8777813141	0.6205423668	0.0135735982	0.0673505988	0.811502757	0.0340893208	0.9040021168	0.0228543166	0.9355651891	0.0707896163	0.8020495839	-0.1320592489	0.6389583926
U.Verrucomicrobia	-0.6757072431	0.0056934373	-0.7313793793	0.0019447265	0.4807465305	0.069673595	0.5509882299	0.0332724827	0.2362932773	0.396510741	0.0859771966	0.7606217515	0.007949807	0.9775676711	-0.0877242854	0.7558924529
Verrucomicrobiae	-0.3001726301	0.2770268616	-0.3977364367	0.1420581458	0.0906509951	0.7479880319	0.1663694115	0.5534483102	-0.0054666686	0.9845735123	-0.1776091897	0.5265587538	-0.1160861072	0.6961197875	-0.2573948093	0.3543724814

A.E Co-occurrence node numbers to taxon and class membership

Table A.8: Co-occurrence module (n=70) node numbers to taxon and class membership

Node number	Species/Genus	Class
1	Erythrobacter	Alphaproteobacteria
2	Alteromonas	Gammaproteobacteria
3	Roseovarius	Alphaproteobacteria
4	Marinobacter	Gammaproteobacteria
5	Pelagomonas calceolata	Pelagophyceae
6	Prochlorococcus	Cyanobacteria
7	Pseudomonas	Gammaproteobacteria
8	Synechococcus	Cyanobacteria
9	Pelagibaca	Alphaproteobacteria
10	Staphylococcus	Bacilli
11	Candidatus actinomarina	Actinobacteria
12	Alcanivorax	Gammaproteobacteria
13	Croceibacter	Flavobacteriia
14	Halomonas	Gammaproteobacteria
15	Loktanella	Alphaproteobacteria
16	Rhodococcus	Actinobacteria
17	Sphingorhabdus	Alphaproteobacteria
18	Streptococcus	Bacilli
19	Coxiella	Gammaproteobacteria
20	Hyphomonas	Alphaproteobacteria
21	Lawsonella	Actinobacteria
22	Roseibacillus	Verrucomicrobiae
23	Candidatus fritschea	Chlamydiia
24	Psychrobacter	Gammaproteobacteria
25	Rhodopirellula	Planctomycetia

26	Tenacibaculum	Flavobacteriia
27	Ichthyodinium chabelardi	Alveolata
28	Acinetobacter	Gammaproteobacteria
29	Hoeflea	Alphaproteobacteria
30	Candidatus nitrosopelagicus	Thaumarchaeota
31	Codonellopsis americana	Spirotrichea
32	Eudubosquella crenulata	Dinophyceae
33	Halobacteriovorax	Oligoflexia
34	Pseudoalteromonas	Gammaproteobacteria
35	Idiomarina	Gammaproteobacteria
36	Jejudonia	Flavobacteriia
37	Blastopirellula	Planctomycetia
38	Maricaulis	Alphaproteobacteria
39	Pelagococcus subviridis	Pelagophyceae
40	Maribacter	Flavobacteriia
41	Caecitellus parvulus	Stramenopiles
42	Litoricola	Gammaproteobacteria
43	Aureispira	Saprospiria
44	Filamoeba nolandi	Amoebozoa
45	Marinomonas	Gammaproteobacteria
46	Thiothrix	Gammaproteobacteria
47	Sphingomonas	Alphaproteobacteria
48	Oleiphilus	Gammaproteobacteria
49	Pseudophaeobacter	Alphaproteobacteria
50	Marinoscillum	Cytophagia
51	Neptunomonas	Gammaproteobacteria
52	Fluviicola	Flavobacteriia
53	Hydra vulgaris	Hydrozoa
54	Oleibacter	Gammaproteobacteria
55	Parabirojimia similis	Spirotrichea
56	Olleya	Flavobacteriia

57	Bathycoccus prasinos	Mamiellophyceae
58	Candidatus pelagibacter	Alphaproteobacteria
59	Kordia	Flavobacteriia
60	Nitrosopumilus	Thaumarchaeota
61	Nonlabens	Flavobacteriia
62	Psychroflexus	Flavobacteriia
63	Arcobacter	Epsilonproteobacteria
64	Delftia	Betaproteobacteria
65	Marinicella	Gammaproteobacteria
66	Bradyrhizobium	Alphaproteobacteria
67	Rubritalea	Verrucomicrobiae
68	Bacillus	Bacilli
69	Salinirepens	Flavobacteriia
70	Crocinitomix	Flavobacteriia

Table A.9: Co-occurrence module (n=51) node numbers to taxon and class membership

Node number	**Species/Genus**	**Class**
1	Colwellia	Gammaproteobacteria
2	Polaribacter	Flavobacteriia
3	Balneatrix	Gammaproteobacteria
4	Ulvibacter	Flavobacteriia
5	Amylibacter	Alphaproteobacteria
6	Lentibacter	Alphaproteobacteria
7	Favella azorica	Spirotrichea
8	Phaeocystis cordata	Haptophyceae
9	Lentimonas	Verrucomicrobia
10	Sulfitobacter	Alphaproteobacteria
11	Oceanicoccus	Gammaproteobacteria
12	Formosa	Flavobacteriia
13	Haliea	Gammaproteobacteria

14	Actinocyclus actinochilus	Coscinodiscophyceae
15	Chroomonas cf. mesostigmatica	Cryptophyta
16	Arenicella	Gammaproteobacteria
17	Emcibacter	Alphaproteobacteria
18	Ilumatobacter	Acidimicrobiia
19	Pyramimonas disomata	Chlorophyta
20	Aquibacter	Flavobacteriia
21	Glaciecola	Gammaproteobacteria
22	Pseudofulvibacter	Flavobacteriia
23	Pithites vorax	Phyllopharyngea
24	Dokdonia	Flavobacteriia
25	Mantoniella antarctica	Chlorophyta
26	Mesoflavibacter	Flavobacteriia
27	Porticoccus	Gammaproteobacteria
28	Sinobacterium	Gammaproteobacteria
29	Schizochytrium aggregatum	Labyrinthulomycetes
30	Dictyocha fibula	Dictyochophyceae
31	Flavicella	Flavobacteriia
32	Oleispira	Gammaproteobacteria
33	Pseudochattonella verruculosa	Dictyochophyceae
34	Pterosperma cristatum	Chlorophyta
35	Prasinoderma coloniale	Chlorophyta
36	Pseudoscourfieldia marina	Chlorophyta
37	Developayella elegans	Stramenopiles
38	Ellobiopsis chattonii	Alveolata
39	Pseudohongiella	Gammaproteobacteria
40	Francisella	Gammaproteobacteria
41	Neptuniibacter	Gammaproteobacteria
42	Marivita	Alphaproteobacteria
43	Paraglaciecola	Gammaproteobacteria
44	Aureococcus anophagefferens	Pelagophyceae

45	Pelagostrobilidium neptuni	Spirotrichea
46	Magnetospira	Alphaproteobacteria
47	Peritromus kahli	Heterotrichea
48	Ceratium tenue	Dinophyceae
49	Chrysochromulina parva	Haptophyceae
50	Varistrombidium kielum	Spirotrichea
51	Florenciella parvula	Dictyochophyceae

A.F Co-occurrence module correlation heatmaps

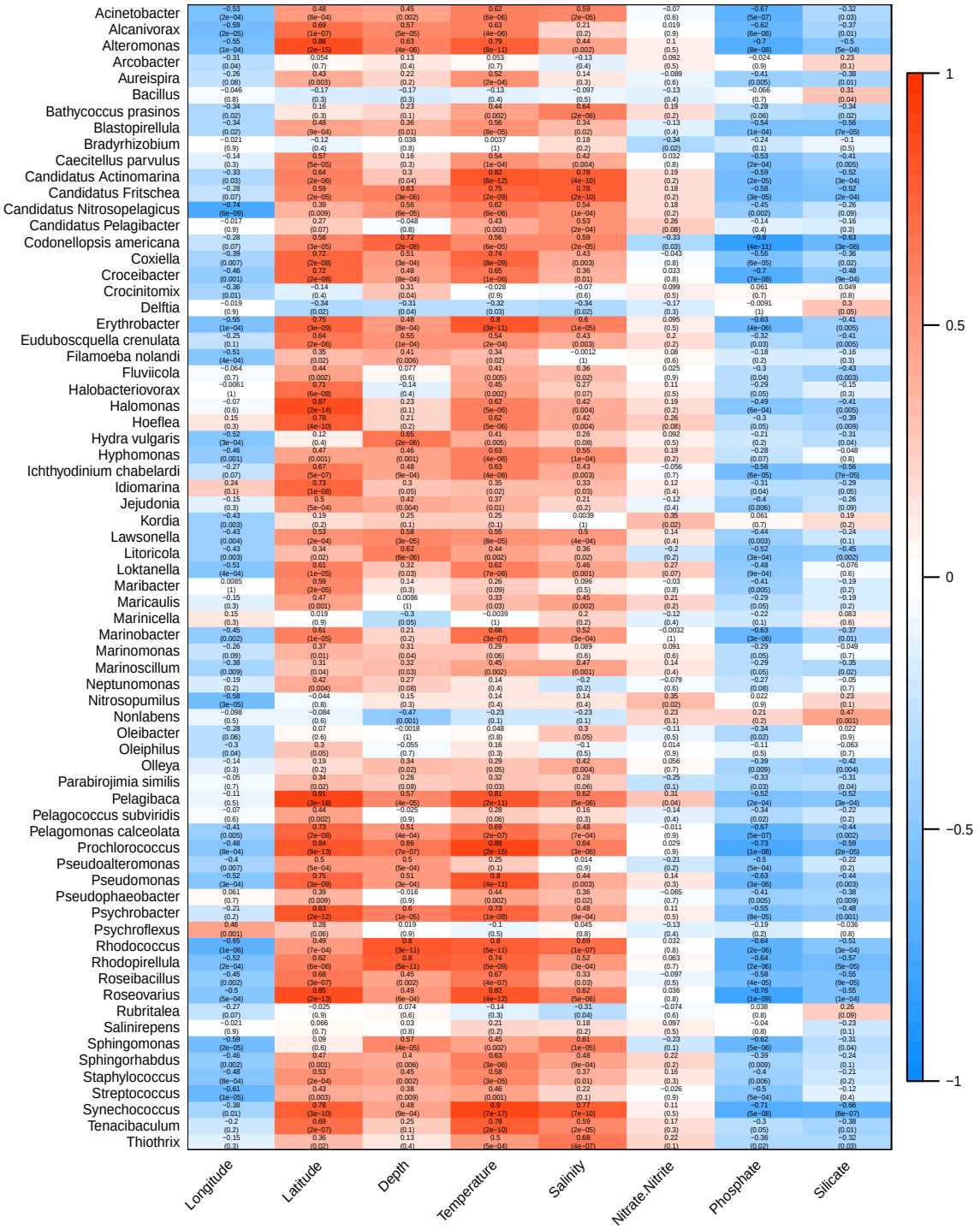

Figure A.1: In the co-occurrence analysis with WGCNA on the log10-scaled abundances of 18S rDNA species level and 16S rDNA genus level, two modules were found. Depicted is the correlation heatmap for the turquoise module in figure 3.17 **a** (n=70). Along the left hand side is the species/genus name and environmental variables are displayed at the bottom. The colours correspond to the correlation values, red is positively correlated and blue is negatively correlated. The Pearson correlation coefficient values and p-values in brackets are displayed in each square

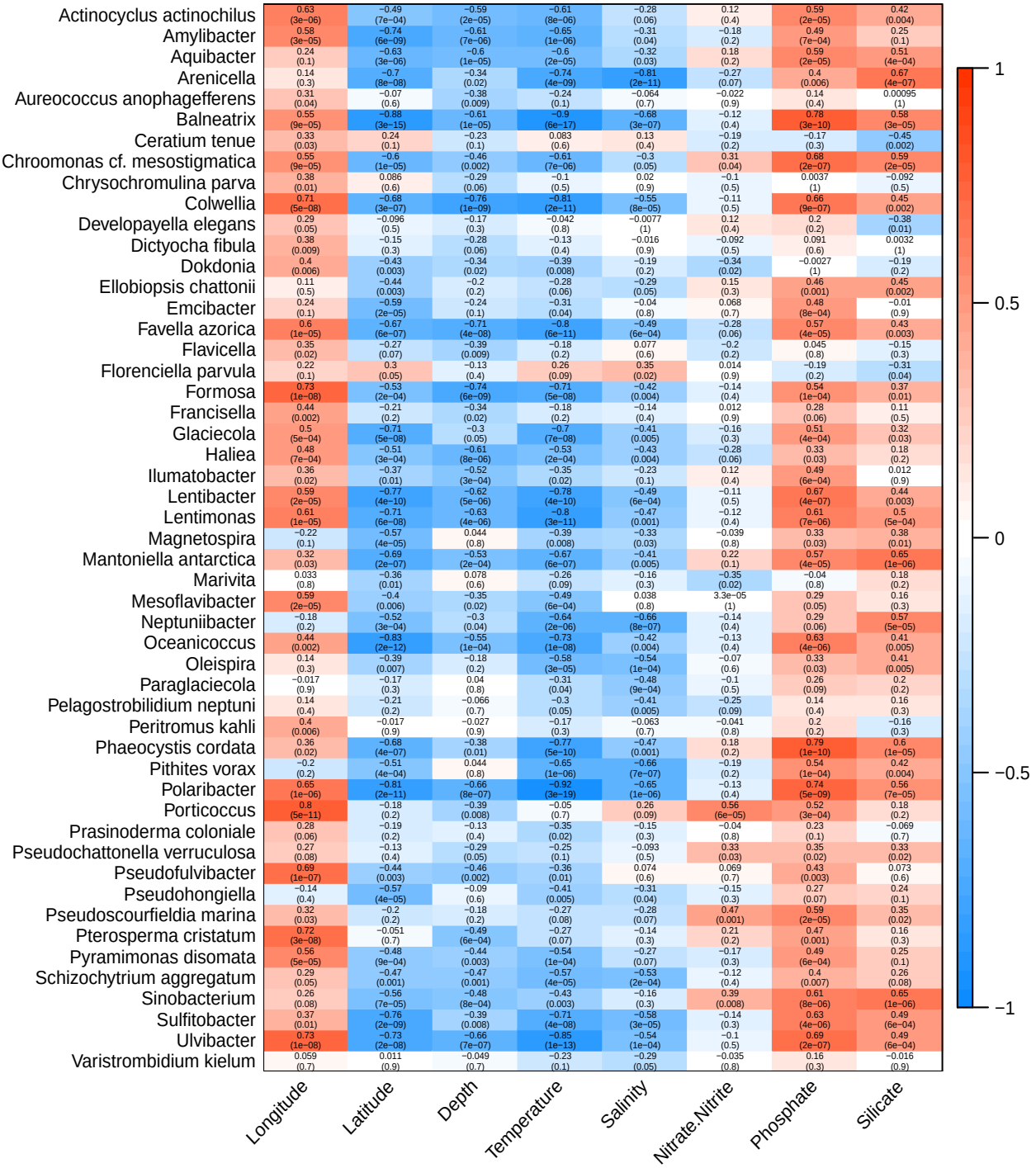

Figure A.2: In the co-occurrence analysis with WGCNA on the log10-scaled abundances of 18S rDNA species level and 16S rDNA genus level, two modules were found. Depicted is the correlation heatmap for the blue module in figure 3.17 **b** (n=51). Along the left hand side is the species/genus name and environmental variables are displayed at the bottom. The colours correspond to the correlation values, red is positively correlated and blue is negatively correlated. The Pearson correlation coefficient values and p-values in brackets are displayed in each square

Appendix B

B.A PhymmBL four genomes locations online

Cyanidioschyzon merolae from Cyanidioschyzon merolae Genome Project http://merolae.biol.s.u-tokyo.ac.jp/download;

Danio rerio from UCSC http://genome.ucsc.edu/cgi-bin/hgGateway?db=danRer5;

Homo sapiens from Genome Reference Consortium http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/

Strongylocentrotus purpuratus from Sea Urchin Genome Project http://www.hgsc.bcm.tmc.edu/projectspecies-o-Strongylocentrotus.