Explorative analysis of the mechanisms of *Phaeocystis globosa* blooms in the Beibu Gulf using amplicon sequencing data

by

Kathryn Gibson

BSc, Simon Fraser University, 2018

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Molecular Biology and Biochemistry
Faculty of Science

SIMON FRASER UNIVERSITY
Summer 2021

Declaration of Committee

Name:	**Kathryn Gibson**
Degree:	**Master of Science**
Title:	**Explorative analysis of the mechanisms of *Phaeocystis globosa* blooms in the Beibu Gulf using amplicon sequencing data**
Committee:	**Chair:** Michel Leroux Professor, Molecular Biology and Biochemistry
	Jack Chen Supervisor Professor, Molecular Biology and Biochemistry
	Fiona Brinkman Committee Member Professor, Molecular Biology and Biochemistry
	Ryan Morin Committee Member Associate Professor, Molecular Biology and Biochemistry
	Lynne Quarmby Examiner Professor, Molecular Biology and Biochemistry

Abstract

Phaeocystis is an ecologically important cosmopolitan genus with several species that form harmful algal blooms. Previous studies of the mechanisms of *Phaeocystis* blooms have been hindered by the small size of *Phaeocystis* cells and the complex *Phaeocystis* life cycle, which includes multiple free-living stages and a colonial stage that dominates during blooms. In this thesis, I apply 16S amplicon sequencing to explore the mechanisms underlying a *P. globosa* bloom in the Beibu Gulf. Using the spatial-temporal dynamics of *P. globosa*, bacteria, archaea, phytoplankton and environmental variables, I develop a model for the development and progression of the *P. globosa* bloom. After, I identify bacteria that interact with *P. globosa* during the bloom by studying the *P. globosa* colony microbiome. While *P. globosa* colonies had different bacterial compositions compared to seawater samples collected from the same locations, I did not find evidence for a core *P. globosa* colony microbiome.

Keywords: *Phaeocystis globosa*; harmful algal blooms; phytoplankton; Beibu Gulf; 16S amplicon sequencing

Acknowledgements

I am incredibly grateful to my senior supervisor, Dr. Jack Chen, for the opportunity to complete my master's thesis in his lab. I have learned so much and become a better scientist under his supervision. Thank you to my committee members, Dr. Ryan Morin and Dr. Fiona Brinkman, for their support and feedback throughout my degree. Thank you to Dr. Lynne Quarmby and Dr. Michel Leroux for serving as the internal examiner and chair for my defense respectively. I would also like to thank Mimi Fourie, Christine Beauchamp and the other MBB office staff for their help and for answering my many administrative questions. Thank you to Sharon Ruschkowski for the administrative support with the Bioinformatics program.

I am very grateful to have worked with the previous members of the Chen lab. Thank you to Matt Douglas, Shinta Thio, Marija Jovanovic and Dr. Michelle Hu for their support and friendship, both in and out of the lab. Special thanks to Dr. Jiarui Li for his supervision during my internship at Applied Biological Materials and for his perspectives and career advice. Thank you to the members of the MBB Graduate Caucus and other MBB graduate students for the friendship and memories.

Finally, I would like to thank my family and friends. Thank you to my mom and dad for supporting me throughout my many years of school. I am especially grateful for all the times you drove me to SFU for my 8:30am classes. Thank you to all my other family and friends for their ongoing support. Most importantly, thank you to my cat, Cali, for being the best (and most fluffy) study buddy.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AIC	Akaike Information Criterion
ANCOM	Analysis of Composition of Microbiomes
ANOSIM	Analysis of Similarities
ANOVA	Analysis of Variance
ASV	Amplicon Sequence Variant
CCA	Canonical Correspondence Analysis
CLR	Centred-Log Ratio
DGGE	Denaturing Gradient Gel Electrophoresis
DMSP	Dimethylsulfoniopropionate
DNA	Deoxyribonucleic Acid
GLM	Generalized Linear Model
HAB	Harmful Algal Bloom
ITS	Internal Transcribed Spacer
LSU rRNA	Large Subunit Ribosomal Ribonucleic Acid
ML	Maximum Likelihood
NB	Negative Binomial
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
PCA	Principle Component Analysis
PCoA	Principle Coordinate Analysis
PCR	Polymerase Chain Reaction
PERMANOVA	Permutational Multivariate Analysis of Variance
PID	Percentage Identity
RDP	Ribosomal Database Project
RUBISCO	Ribulose Bisphosphate Carboxylase/Oxygenase
SSU rRNA	Small Subunit Ribosomal Ribonucleic Acid
ARISA	Automated Ribosomal Intergenic Spacer Analysis
T-RFLP	Terminal Restriction Fragment Length Polymorphism

Glossary

Algae	A diverse group of photosynthetic aquatic organisms
Commensalism	An association between two organisms in which one benefits and the other derives neither benefit nor harm
Cryptic species	One of two or more morphologically indistinguishable biological groups that are incapable of interbreeding
Cyanobacteria	A group of aquatic photosynthetic bacteria
Diatom	A major group of unicellular algae with silica cell walls
Dinoflagellate	A major group of flagellated unicellular algae
Epipelagic	The part of the oceanic zone into which enough light penetrates for photosynthesis
Haptophyte	The clade of algae that includes *Phaeocystis*
Heterotroph	An organism that eats other plants or animals for energy and nutrients
Interpolation	A procedure that predicts values for cells in raster from a limited number of sample points
Phytoplankton	A diverse, polyphyletic group of organisms that are mostly unicellular autotrophs found in marine and fresh water
Symbiosis	An interaction between two different organisms living in close physical association
Zooplankton	Small protozoans or metazoans that feed on other plankton

Chapter 1. Introduction

1.1. Phytoplankton

The term phytoplankton is used to describe a diverse, polyphyletic group of organisms that are mostly unicellular autotrophs found in marine and fresh water (Falkowski and Raven 1997). Phytoplankton taxa can be distributed into at least eight major groups or phyla, including one major prokaryotic group (the cyanobacteria) and a handful of eukaryotic groups (Falkowski et al. 2004). The two major groups of eukaryotic phytoplankton are the diatoms, a monophyletic group with cells walls composed of intricate and striking patterns of silica, and the dinoflagellates, a monophyletic group with cells that contain a flagellum for swimming. Phytoplankton possess a complex evolutionary history with photosynthesis initially spreading from the cyanobacteria to a variety of eukaryotic clades via endosymbiosis (Delwiche 1999). Following this event, two major plastid lineages evolved: “green” and “red”, which differ based on their pigment composition and were both subsequently diversified by secondary and tertiary endosymbiotic events (Falkowski et al. 2004; Figure 1). Recognizing this complex evolutionary history and the resulting diversity is crucial for appreciating the importance of phytoplankton.

Phytoplankton are key contributors to primary production and a variety of biogeochemical cycles. In addition to being responsible for most of the ocean’s primary production, phytoplankton also account for >45% of the planet’s annual net primary production (Field et al. 1998). The evolutionary diversity of phytoplankton translates into diverse contributions to the planet’s biogeochemical cycles (Litchman et al. 2015). For example, coccolithophorids contribute to the marine calcium cycle via the formation of calcium carbonate plates, which contribute to calcium carbonate rock formations. In contrast, diatoms contribute significantly to the global silica cycle due to their unique silica cell walls (Litchman et al. 2015). Despite their important contributions to our planet’s biogeochemical cycles, phytoplankton are more well-known for their ability to form harmful algal blooms.

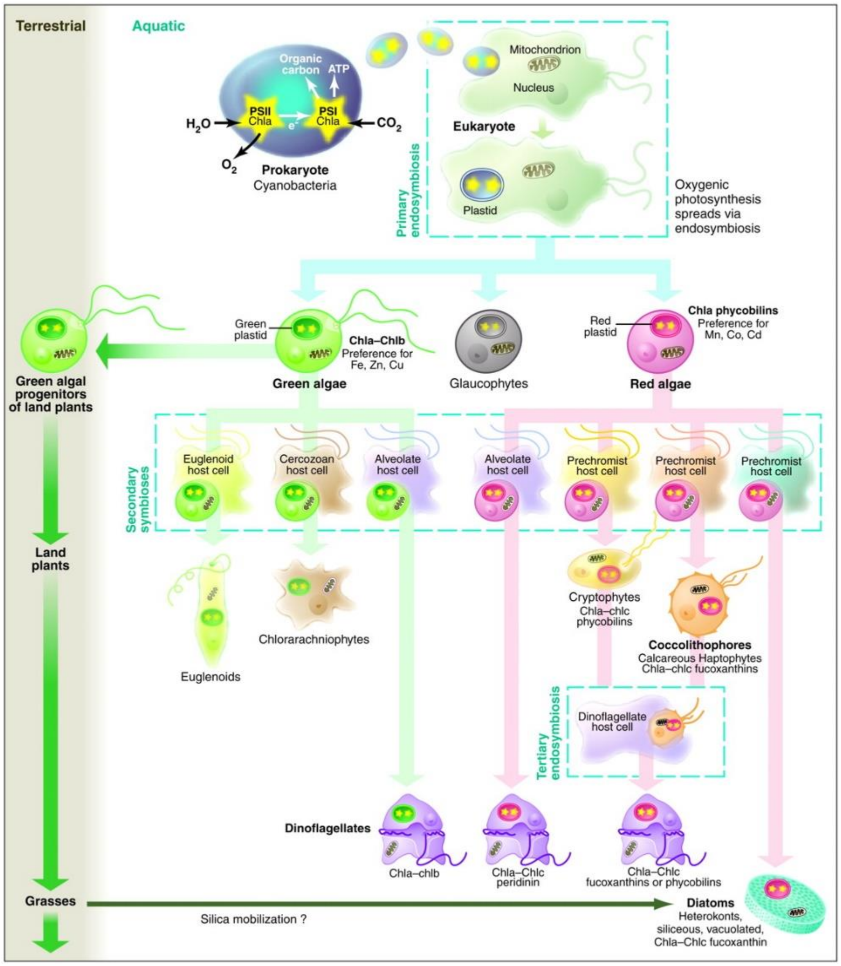


Figure 1 The evolutionary history of phytoplankton.
Source: From (Falkowski et al. 2004). Reprinted with permission from AAAS.

1.2. Harmful algal blooms

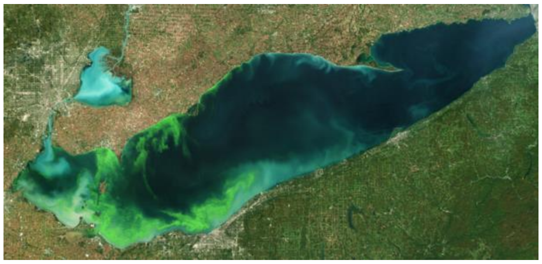


Figure 2 An algal bloom in Lake Eeerie.
Source: National Oceanic and Atmospheric Administration

Harmful algal blooms (HABs) are a global phenomenon (Zohdi and Abbaspour 2019) that is increasingly important due to the growing impacts on human and ecosystem health and the resulting economic losses. Algal blooms are rapid accumulations in the algae population that can often be easily recognized by water discolouration due to the algae pigments (Zohdi and Abbaspour 2019; Figure 2). Algal blooms that have associated negative impacts such as natural toxin production or dissolved oxygen depletion are labelled as HABs (Zohdi and Abbaspour 2019). Of the 5,000+ phytoplankton species, only about 300 are known to form algal blooms and about 75 can form HABs (Malaei Tavana et al. 2008), most of which are cyanobacteria, diatoms and dinoflagellates (Zohdi and Abbaspour 2019). The causes of HABs are believed to be multi-faceted (Zohdi and Abbaspour 2019) with contributions from human activities i.e. eutrophication (Malaei Tavana et al. 2008) and natural factors i.e. current systems and weather patterns (Anderson 1994). The consequences of HABs are extensive and include the direct or indirect poisoning of humans and aquatic animals, decreased water quality, ecosystem damage and economic losses due to human health, fisheries, tourism, and recreation (Zohdi and Abbaspour 2019). Alarmingly, the number of HABs, the scale of the phenomenon and the downstream consequences have all increased in recent decades (Sellner et al. 2003, Zohdi and Abbaspour 2019; Figure 3), a trend that many researchers attribute to increased human activities such as eutrophication (Sellner et al. 2003). One example that follows this trend is HABs caused by members of the genus *Phaeocystis*.

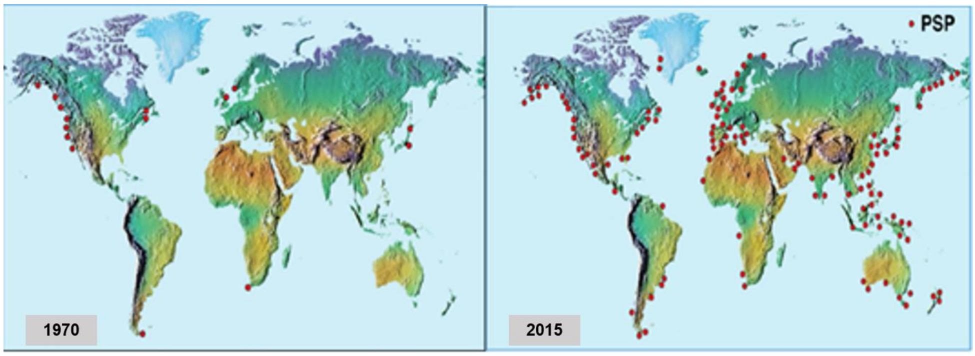


Figure 3 Comparison of harmful algal blooms (HABs) from 1970 and 2015. Outbreaks are measured by occurrences of paralytic shellfish poisoning (PSP), a syndrome of shellfish poisoning caused by toxins produced by some HAB species.

Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *International Journal of Environmental Science and Technology*, Harmful algal blooms (red tide): a review of causes, impacts and approaches to monitoring and prediction, E. Zohdi and M. Abbaspour, *Copyright © 2019, Islamic Azad University (IAU)* (2019)

1.3. *Phaeocystis*

Phaeocystis is a cosmopolitan haptophyte genus that is regarded as ecologically important and a nuisance. There are at least six species in the genus (Medlin and Zingone 2007), three of which have been identified as forming blooms: *P. globosa*, *P. pouchetti* and *P. antarctica* (Schoemann et al. 2005). While the genus has a global distribution, *P. globosa* blooms are restricted to temperate and tropical waters and *P. pouchetti* and *P. antarctica* bloom in the Arctic and Antarctic waters respectively (Schoemann et al. 2005). *Phaeocystis* blooms are considered harmful for many reasons including the production of toxic haemolytic substances (van Rijssel et al. 2007), the resulting fish mortality and the formation of odorous foams on beaches (Schoemann et al. 2005). Despite their label as a nuisance, *Phaeocystis* are ecologically important: they are key organisms in driving biogeochemical cycles (Schoemann et al. 2005) and are significant producers of DMSP (Liss et al. 1994). DMSP produced by *Phaeocystis* is subsequently cleaved into DMS (Stefels and Dijkhuizen 1996), which reduces the effect of greenhouse gases such as carbon dioxide (Verity et al. 2007). The observed correlation between *Phaeocystis* blooms and DMS in the atmosphere (Turner et al. 1995) suggests that these harmful blooms may paradoxically be important for global climate regulation. The global distribution, bloom-formation and ecological importance of *Phaeocystis* have resulted in extensive study of the genus, but our understanding of its blooms has been hindered by its complex life cycle (Verity et al. 2007).

The formation of blooms by *Phaeocystis* is a life cycle event, which involves the transition from a free-living to a colonial morphotype. *Phaeocystis* has a complex, polymorphic life cycle with alteration between three types of free-living cells and colonial cells (Rousseau et al. 2007; Figure 4). *Phaeocystis* colonies consist of thousands of cells embedded in a polysaccharidic matrix (Schoemann et al. 2005; Figure 4a) that can reach up to 3 cm in size (Chen et al. 2002). Critically, the majority of *Phaeocystis* blooms consist of the colony form (Verity et al. 2007), which suggests that the success of the genus can largely be attributed to their ability to form colonies (Schoemann et al. 2005). Hypotheses that have been proposed to explain the function of colony formation in *Phaeocystis* include 1) a defensive role by reducing viral infections (Brussaard et al. 2005) and predation (Noordkamp et al. 2000, Nejstgaard et al. 2007), 2) a reproductive role by facilitating sexual reproduction (Rousseau et al. 2013) and 3) a means of sequestering micronutrients by promoting symbiotic relationships with bacteria (Bertrand et al. 2007, Delmont et al. 2014, Bender et al. 2018). Potential triggers of colony formation, which may be the same as the triggers of bloom formation (Bender et al. 2018), include abiotic factors, such as nutrient availability (Bender et al. 2018) and light (Cariou et al. 1994), and biotic factors, such as grazing cues (Long et al. 2007) and viral infection (Brussaard et al. 2005). Overall, however, the mechanisms of colony and bloom formation in *Phaeocystis* remain unclear. One strategy that can be used to further our understanding of the bloom mechanisms is studying geographical regions that are recurrently impacted by *Phaeocystis* blooms.

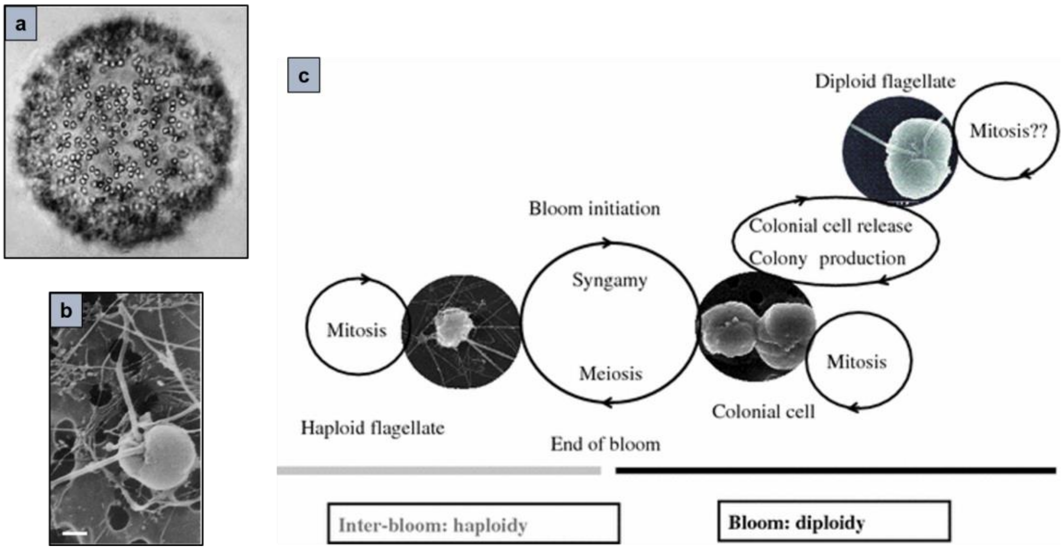


Figure 4 ***P. globosa*** **colony (a) and free-living (b) stages. The** ***P. globosa*** **life cycle (c) is complex, with two haploid flagellate, one diploid flagellate and one diploid colonial stage. The diploid colonial stage is associated with the formation of blooms. The white scale bar in (b) is 1 µm.**

Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Biogeochemistry*, A taxonomic review of the genus *Phaeocystis*, L. Medlin and A. Zingone, *Copyright © 2007, Springer Science Business Media B.V.* (2007). Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Biogeochemistry*, The life cycle of *Phaeocystis*: state of knowledge and presumptive role in ecology, V. Rousseau et al., *Copyright © 2007, Springer Science Business Media, Inc.* (2007)

1.4. *P. globosa* blooms in the Beibu Gulf

The Beibu Gulf is a coastal region in southwest China (Figure 5) that has recently been impacted by recurrent *P. globosa* blooms (Xu et al. 2019). The region contains some of the most abundant fishing grounds in the coastal waters of China (Zhong 2015) but, like most of China’s coastal waters, has been subject to extensive eutrophication

(Han et al. 2012). While HAB occurrences from 1985-2010 in the Beibu Gulf were largely dominated by the cyanobacterium *Microcystis aeruginosa* and co-occurrences of cyanobacteria, dinoflagellates and diatoms, *P. globosa* is now the major species in blooms (Xu et al. 2019). The first documented *P. globosa* bloom occurred in 2011 and has been followed by large recurrent blooms in the subsequent years (Xu et al. 2019). The impact of these blooms on the local aquaculture industry is substantial and several of the blooms have also posed a threat to the cooling system of nuclear power plants (Xiaokun et al. 2019). While human activities have been identified as key contributors to HABs in the Beibu Gulf (Xu et al. 2019), the mechanisms of *P. globosa* blooms in the Beibu Gulf are currently unknown. Tracking the microbial community composition during a *P. globosa* bloom may provide insight into the mechanisms underlying this phenomenon.

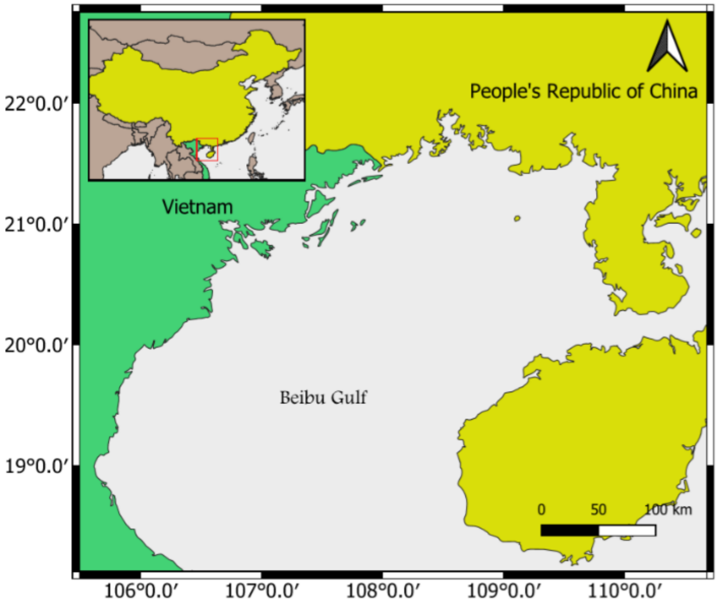


Figure 5 The Beibu Gulf.

1.5. Methods for studying microbial community composition

1.5.1. Culturing and microscopy

Culturing and microscopy are fundamental, but limited, techniques in molecular ecology for determining microbial community composition. Since the first intentional isolation of bacteria for scientific purposes was achieved by Robert Kock and Julius Petri in the 1870’s, culturing, a method of multiplying microbial organisms via reproduction in a controlled laboratory setting, has become the gold standard for microbial characterization. Culturing facilitates the production of a large number of cells from a clonal population that are amenable to numerous functional tests on their biochemistry, physiology and genetics (Hugerth and Andersson 2017). However, the number of species that can be characterized using culturing is limited because most bacteria cannot be cultivated with standard techniques. Growth requirements differ between organisms, with many having narrow windows of growth that prevent them from growing fast enough in the lab to be distinguishable (Lagier et al. 2015). Additionally, some microbes may fail to grow if important pathways are missing in their environment (Nye et al. 1999) or if they are dependent on molecules produced by other members of their community (D’Onofrio et al. 2010). Despite advances in culturing techniques, isolating and culturing bacteria remains a complex and time-consuming endeavour. An alternative to culturing is performing microscopy directly on environmental samples. Microscopy has vastly improved since the first observations of microbial organisms by Antony van Leeuwenhoek in the 1670’s, with techniques such as electron microscopy, confocal microscopy and photoswitchable fluorophores now available to be used on images of live or fixated bacteria (Coltharp and Xiao 2012). However, there are also several limitations to microscopy: the taxonomic resolution obtained is typically inadequate for the diversity of microbes found in an environmental sample, years of training are required to be able to correctly visually identify microbes, cryptic species cannot be differentiated and biases due to observer effects are common (Hugerth and Andersson 2017). Thus, while culturing and microscopy are still important tools in molecular ecology, the field has moved beyond the limitations of these techniques to molecular fingerprinting.

1.5.2. Molecular fingerprinting

Molecular fingerprinting, which uses molecular techniques to identify microbes, is an integral tool in molecular ecology for determining microbial community composition. The small subunit (SSU) of the ribosomal RNA (rRNA) gene (16S in prokaryotes and 18S in eukaryotes) was one of the first genes to be established as suitable for inferring phylogenetic relationships between prokaryotes (Woese and Fox 1977) and was soon applied to study the composition of natural communities (Pace et al. 1985). The advantages of the SSU rRNA gene are that it is universal i.e. found in all cellular life forms, highly conserved, rarely transferred horizontally and has both conserved regions that can be targeted by PCR primers and variable regions that can be used as molecular markers (Hugerth and Andersson 2017). While other genes do share these properties e.g. the large subunit rRNA, the length of the SSU rRNA gene was most suitable for early molecular techniques and the wealth of knowledge that has since been accumulated in databases makes it impractical for the field to switch to a different gene (Hugerth and Andersson 2017). Today, the SSU rRNA and the internal transcribed spacer (ITS) are the most commonly used genes for analyzing the phylogenetic composition of communities (Hugerth and Andersson 2017).

High-throughput environmental fingerprinting approaches such as denaturing gradient gel electrophoresis (DGGE), automated ribosomal intergenic spacer analysis (ARISA) and terminal restriction fragment length polymorphism (T-RFLP) first arose in the 1990s. To perform DGGE, primers are designed to amplify the target gene sequence using PCR and the PCR products are subsequently run on a polyacrylamide gel to denature and separate the fragments (Muyzer et al. 1993). The resulting banding pattern can be used to compare changes in taxonomic composition between samples. Similarly, ARISA involves PCR amplification of the intergenic spacer region between the small and large subunits of the rRNA gene operon followed by running the resulting PCR fragments on a polyacrylamide gel (Fisher and Triplett 1999). Finally, T-RFLP uses the size of the terminal fragments of the amplified region following digestion with restriction enzymes to determine taxonomic composition (Liu et al. 1997). While DGGE, ARISA and T-RFLP were commonly used in the 1990s and early 2000s, these techniques have since been replaced with higher throughput and resolution molecular methods.

1.5.3. DNA arrays

Microarrays are an alternative to molecular fingerprinting that also arose in the 1990s. DNA arrays are a group of technologies that use oligonucleotide probes attached to a surface in an array to quantify the relative concentration of nucleic acid species in a solution via the binding of labeled nucleic acids to the probes (Bumgarner 2013). While microarrays are higher throughput than techniques like DGGE, identification is restricted to sequences that are previously known, which limits its application as a general environmental survey tool (Hugerth and Andersson 2017).

1.5.4. Amplicon sequencing

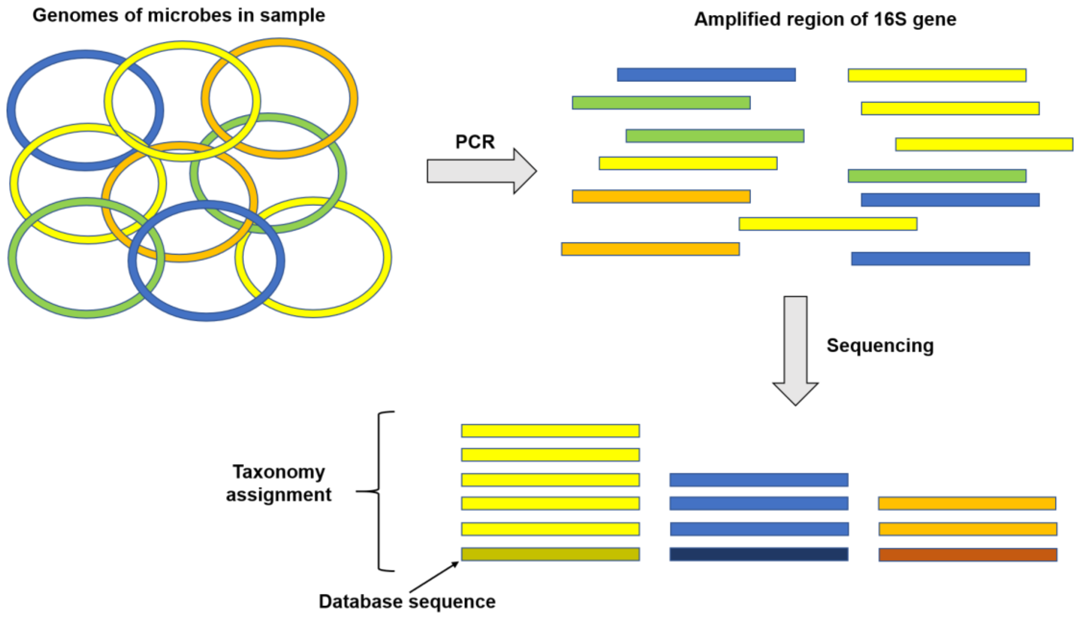


Figure 6 **Illustration of a workflow for amplicon sequencing of environmental samples. Following DNA extraction, PCR is used to amplify a region of the 16S gene. The amplified regions are sequenced and assigned taxonomy by comparing the sequences to a database of sequences with known taxonomy.**

High throughput DNA sequencing i.e. next generation sequencing (NGS) has transformed the field of molecular ecology with the application of amplicon sequencing. While the exact sequencing approach differs between NGS platforms, all platforms facilitate the sequencing of millions of small fragments of DNA in parallel. Moreover, NGS technologies have become increasingly rapid, sensitive and cost-efficient (Sboner et al. 2011). Amplicon sequencing is a method of targeted NGS that uses PCR to amplify targeted regions in the genome (i.e. the SSU rRNA), which can then be pooled together from hundreds of samples and sequenced simultaneously. Taxonomic characterization of the generated sequences is based on similarity to the reference gene sequences available in public databases (Gupta et al. 2019; Figure 6). The first study that used 454 pyrosequencing to assess microbial communities was performed in a marine water community (Sogin et al. 2006) and demonstrated a much larger microbial diversity than previously expected with an underappreciated number of low abundance organisms. The benefits of amplicon sequencing include increased throughput compared to historical techniques, increased phylogenetic resolution and the ability to determine the relative abundance of all microbes in a sample (amplicon sequencing does not rely on whether the microbes can be cultured or identified based on morphology). However, some of the recognized limitations of amplicon sequencing include PCR biases due to variable primer binding efficiencies, the challenge of identification beyond the genus level due to a high similarity between SSU rRNA gene sequences from closely related species (Gupta et al. 2019) and the reality that amplicon sequencing only gives information on the relative abundance of gene copies in each sample. Amplicon sequencing cannot provide information on the absolute abundance of organisms and variation in gene copy number between genomes can make results misleading (Větrovský and Baldrian 2013). Despite theses limitations, amplicon sequencing has been used in major projects such as the Earth Microbiome project (Gilbert et al. 2014) and the Tara Oceans global ocean survey (Lima-Mendez et al. 2015) to gain insight into the microbial community composition of environmental samples at a depth and resolution that has not been previously achieved. Amplicon sequencing has also been used to further our understanding of the mechanisms of phytoplankton blooms caused by a number of different species by observing the microbial community composition during the bloom (e.g. Needham and Fuhrman 2016).

1.6. Thesis aims

In this thesis, I report the application of 16S amplicon sequencing to explore the mechanisms of *P. globosa* blooms in the Beibu Gulf. The 16S rRNA gene was chosen to allow for simultaneous study of the bacteria, archaea and photosynthetic eukaryotes. The 16S rRNA gene is advantageous for studying photosynthetic eukaryote communities due to the low copy number variation in the chloroplast 16S rRNA gene compared to the 18S rRNA gene (Needham and Fuhrman 2016). First, I explore the composition of *P. globosa* and other microbes in the Beibu Gulf epipelagic layer. Second, I explore the spatial-temporal distribution of microbial organisms and environmental variables at three time points of a *P. globosa* bloom in the Beibu Gulf. These spatial-temporal patterns will provide an initial insight into the roles of different microbes and environmental factors in the community at different stages of the bloom. Third, I identify microbes with putative interactions with *P. globosa* colonies with a focus on the microbiome of *P. globosa* colonies sampled from the same bloom. Identifying microbes that are associated with *P. globosa* will facilitate the identification of microbes that are candidates for forming symbiotic relationships with *P. globosa* colonies during blooms. Collectively, this explorative approach will further our understanding of the mechanisms underlying *P. globosa* blooms in the Beibu Gulf as well as the global mechanisms of *Phaeocystis* blooms.

The explorative analysis of the mechanisms of *P. globosa* blooms in the Beibu Gulf reported here has three aims. In chapter two I present the methods that were used for data collection and the construction of a bioinformatics pipeline to address the three aims.

The first aim, which is presented in chapter three, is to explore the composition of *P. globosa* and other microbes in the Beibu Gulf using 16S amplicon sequencing.

The second aim, which is presented in chapter four, is to explore the spatial-temporal dynamics of *P. globosa*, bacteria, archaea, phytoplankton and environmental factors during a *P. globosa* bloom in the Beibu Gulf using 16S amplicon sequencing. In this chapter I use the spatial-temporal distribution patterns of the microbes and environmental data to develop a model for the development and progression of the bloom.

The third aim, which is presented in chapter five, is to explore microbes that potentially interact with *P. globosa* colonies during a bloom in the Beibu Gulf using 16S amplicon sequencing. In addition to the field samples, I use direct sequencing of *P. globosa* colonies to explore the *P. globosa* colony microbiome.

Finally, in chapter six I present my conclusions on the mechanisms underlying *P. globosa* blooms in the Beibu Gulf and discuss future experiments that can be used to further this understanding.

Chapter 2. Data collection and bioinformatics pipeline

In this chapter I present the methods that were used to collect 16S amplicon sequencing data during a *P. globosa* bloom in the Beibu Gulf and the bioinformatics pipeline that was used for the analysis of the amplicon sequencing data. My contribution to this chapter is the development and application of the bioinformatics pipeline.

2.1. Data collection

2.1.1. Sample collection

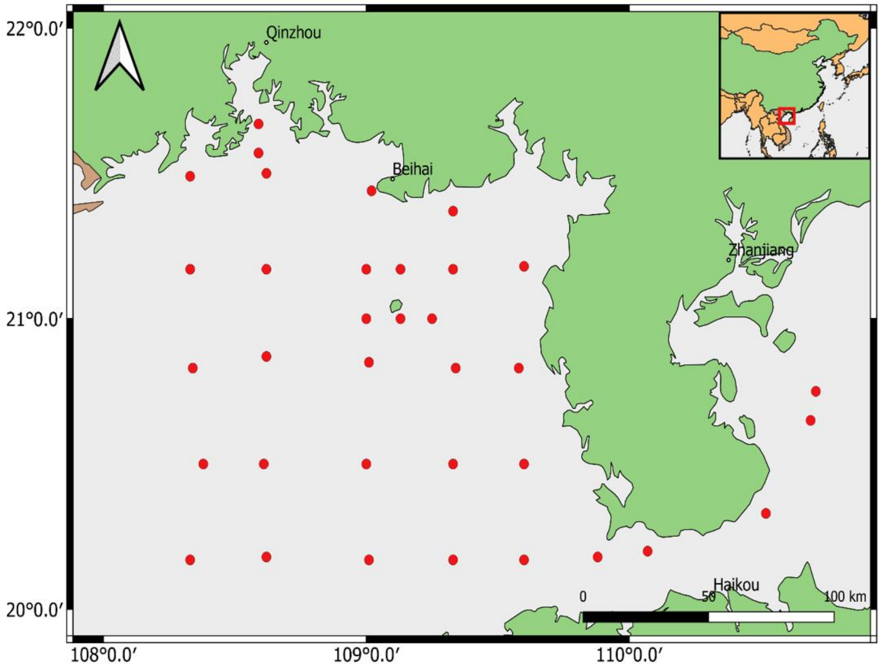


Figure 7 Locations of thirty-five sampling sites in the Beibu Gulf.

Three expedition voyages were conducted in 2019 in the Beibu Gulf during a *P. globosa* bloom to collect field and colony samples. The first expedition (January) was during the peak of the bloom, the second expedition (Feb-March) was during the initial decay of the bloom and the third expedition (April) was at a later decay stage of the bloom. Field samples were collected at thirty-five unique locations (Figure 7) at a range of water sampling depths (0-72m). Every location had a sample collected at the surface

(0m) and some locations had additional samples collected at additional depths below the surface (5-72m) where the water depth permitted. For each field sample (*n* = 231), 0.5-4 L of seawater was collected and filtered using 200 µm mesh (Hebei Anping Wire Mesh Co., Ltd, China) to remove larger zooplankton and phytoplankton. A second filtration was performed using 10 µm polycarbonate membranes (Millipore, USA) followed by a final filtration using 0.2 µm polycarbonate membranes (Millipore, USA). The materials captured by the 10 µm membranes were 10-200 µm in size and represent the large filtration size fraction. The materials captured by the 0.2 µm membranes were 0.2-10 µm in size and represent the small filtration size fraction (Figure 8). The 10 µm and 0.2 µm membranes (*n* = 455) were transferred into liquid nitrogen for storage. Samples of whole *P. globosa* colonies (*n* = 6) were also collected during the January and Feb-March expeditions. Each colony sample was separated from the seawater using a Dispette, washed with sterile seawater three times and placed in a 2 mL cryopreservation tube for storage in liquid nitrogen.

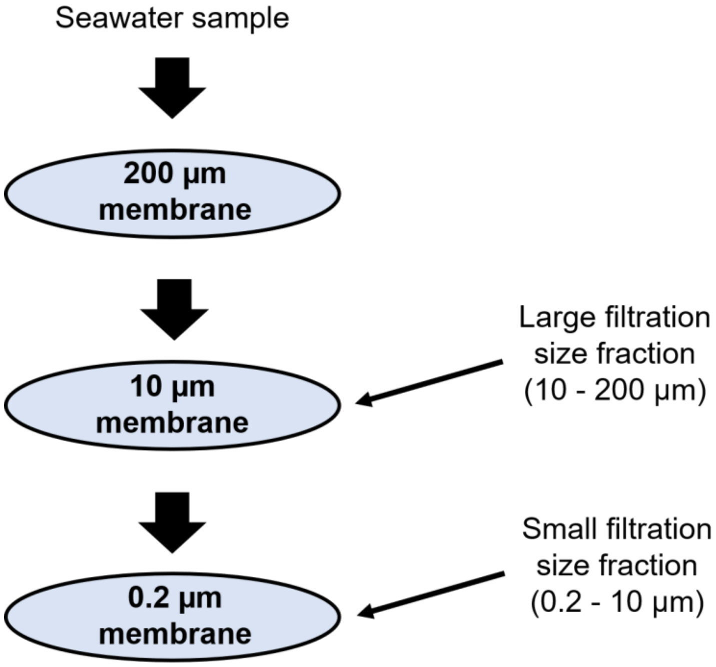

Seawater sample
200 µm membrane
10 µm membrane
Large filtration size fraction (10 - 200 µm)
0.2 µm membrane
Small filtration size fraction (0.2 - 10 µm)

Figure 8 **Filtration strategy for processing seawater samples. Field samples were filtered through 200 µm, 10 µm and 0.2 µm polycarbonate membranes. The materials captured by the 10 µm membranes were 10-200 µm in size and represent the large filtration size fraction. The materials capture by the 0.2 µm membranes were 0.2-10 µm in size and represent the small filtration size fraction**

A number of environmental variables were also measured during the expeditions to allow for the study of their role in bloom development and progression. Water depth, temperature and salinity were determined using a Conductivity-Temperature-Depth profile (Sea-Bird, America) and NO_3, NO_2, PO_4, NO_2, NH_4 and SiO_3 were measured with Skalar San++ CC Continuous Flow Analyzers (Netherlands). Chlorophyll a was measured from 0.3-0.5 L seawater samples that were collected and filtered using GF/F filters and stored at 0°C away from light. The filters were extracted using a buffered acetone solution (90%) for 24 h and the Chlorophyll a concentration in the extract was determined by a Fluorometer (Trilogy, Turner Design).

2.1.2. Sample processing

DNA was extracted from the field samples using the HP Plant DNA kit (Omega, USA) according to the manufacturer’s instructions with some modifications. Briefly, after the samples were taken out of liquid nitrogen, 500 µL CSPL buffer (Omega, USA) was immediately added so that the membranes were completely immersed. The membranes were cut with scissors 50 times and then crushed by a cell crusher (MP, USA) for 5 s at a speed of 4 m/s after adding 20 mg glass beads. The remaining procedure followed the manufacturer’s instructions. DNA was extracted from the colony samples using the HP Plant DNA kit (Omega, USA) according to the manufacturer’s instructions. The V3-V4 region of the 16S rRNA gene was amplified from the extracted DNA using the 341F forward primer, CCTAYGGGRBGCASCAG, and the 806R reverse primer, GGACTACNNGGGTATCTAAT, with a unique barcode sequence at the 5’ end. PCR reactions were performed in 50 µL reactions with 50 ng of DNA template. The reaction conditions consisted of an initial denaturation at 94°C for 4 min, followed by 30 cycles of denaturation at 94°C for 1 min, annealing at 52°C for 90 s and elongation at 72°C for 2 min, with a final extension at 72°C for 10 min. The quality of the libraries was assessed with a Qubit 2.0 Fluorometer (Thermo Scientific). Finally, the prepared libraries were sequenced with 2x250 paired-end reads using the NovaSeq Illumina platform (Illumina, USA; Biomarker Technologies, China) with an average of 89,072 reads/sample (range: 54,470 - 112,635 reads/sample; Figure 9).

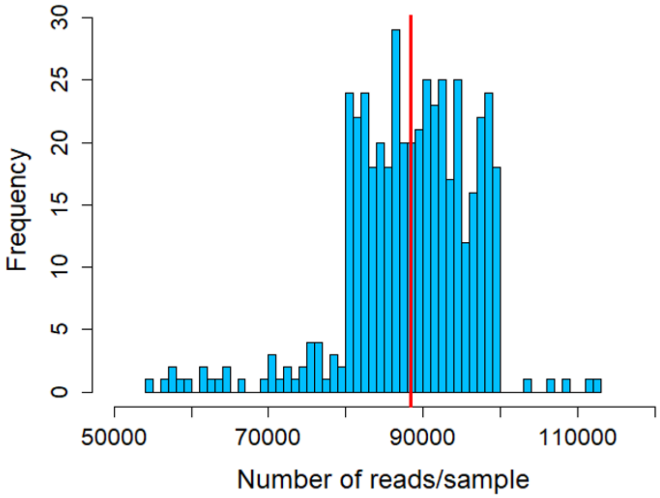


Figure 9 **Histogram of the number of reads/sample. The red vertical line represents the mean.**

2.2. Bioinformatics pipeline

2.2.1. Operational taxonomic units

In the context of molecular ecology, the analysis of amplicon sequencing data typically begins with the construction of operational taxonomic units (OTUs). OTUs are defined by clustering sequences based on a similarity threshold, which is typically set to 97% for 16S rRNA sequences. This similarity threshold is based on the DNA reassociation value that was previously accepted as the definition for bacterial species (Stackebrandt and Goebel 1994). Following this definition, OTUs are often used as the working definition of a species. OTUs can be used to merge variation within strains into a single cluster, merge variation between strains of a single species into a single cluster and to merge variation due to experimental error into a single cluster. Dealing with experimental error is particularly important with high-throughput sequencing where an error rate of ~0.1%/nucleotide (the standard Illumina error rate) results in many sequences having at least one error, which can obscure the underlying biology and artificially inflate sample diversity. OTU picking procedures can be divided into three main approaches: 1) closed reference methods, where reads are mapped and assigned to sequences in a databases and reads that fail to map are discarded, 2) open reference methods, where reads that fail to map to sequences in a reference database are submitted to a *de novo* approach and 3) *de novo* methods, which employ either hierarchical clustering e.g. MOTHUR (Schloss et al. 2009) or heuristic methods e.g. Usearch (Edgar 2013) (Hugerth and Andersson 2017). While OTUs are widely used in molecular ecology, there are several recognized shortcomings: a reduction in phylogenetic resolution i.e. OTUs cannot differentiate between strains of the same species, the use of arbitrary identity thresholds (there is now evidence that the optimal identity threshold for the 16S rRNA gene is > 97% (Edgar 2018) and that the optimal identity threshold differs between genes), dataset dependency of *de novo* OTUs, and the loss of biological variation that is not represented in the reference database for closed reference OTUs (Callahan et al. 2017). Newer methods that remove the need to cluster sequences based on similarity and are independent of datasets and databases have thus been recently developed to overcome these shortcomings.

2.2.2. Amplicon sequencing variants

Methods that resolve amplicon sequencing variants (ASVs) by inferring biological sequences and correcting or removing sequencing errors have recently been developed (Eren et al. 2015, Callahan et al. 2016, Edgar 2016, Amir et al. 2017) with the intention of replacing OTUs (Callahan et al. 2017). ASVs are inferred *de novo* and biological sequences are discriminated from sequencing errors on the basis of the expectation that biological sequences are more likely to be observed repeatedly than sequences that are errors (Callahan et al. 2017). In addition to having better sensitivity and resolution than OTU methods (Callahan et al. 2016), ASV labels are consistent because they represent a biological reality and can thus be compared even if they are inferred independently from different samples. This consistent labelling allows for data from studies to be merged for meta-analysis without reprocessing, facilitates replication and facilitates the use of ASVs as predictive biomarkers (Callahan et al. 2017). ASVs are also independent of the reference database, which allows for applications in less-studied environments, e.g. the ocean, and means that ASVs remain consistent even with changing reference databases (Callahan et al. 2017). DADA2, which uses an error model based on a matrix of nucleotide transition probability parameters and the abundance of each sequence in an iterative partitioning algorithm, is one of the most popular tools for generating ASVs (Callahan et al. 2016). Less popular tools that also use sequencing error models include UNOISE2 (Edgar 2016) and Deblur (Amir et al. 2017). Alternatively, Minimum-Entropy

Decomposition (MED) uses Shannon entropy to identify sequencing errors (Eren et al. 2015). While these tools are able to produce similar microbial compositions based on relative abundance (Nearing et al. 2018), I elected to use DADA2 in this analysis pipeline because it tends to find more ASVs, suggesting that it could be better at finding rare organisms (Nearing et al. 2018).

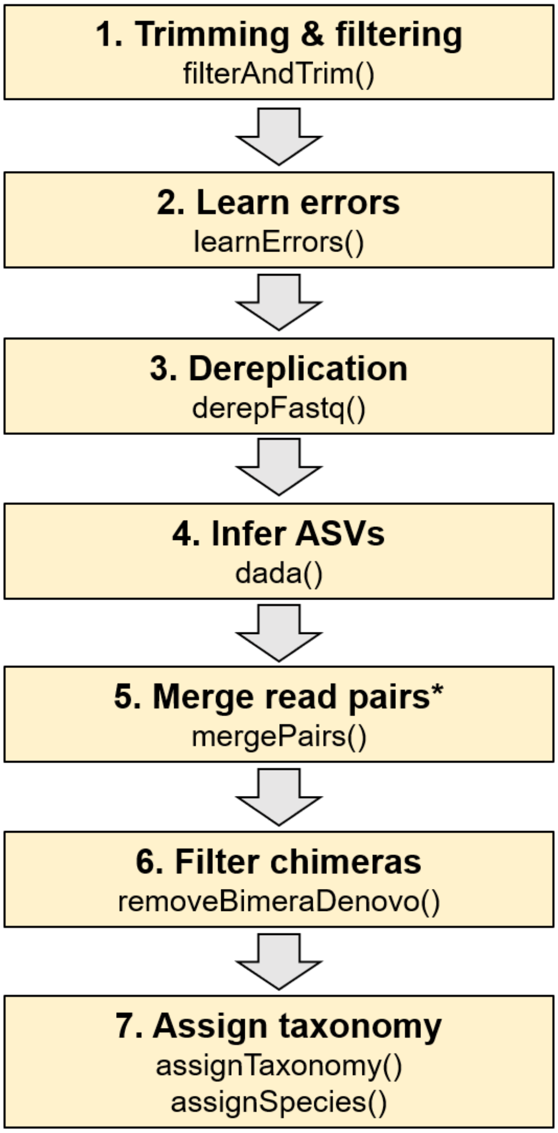


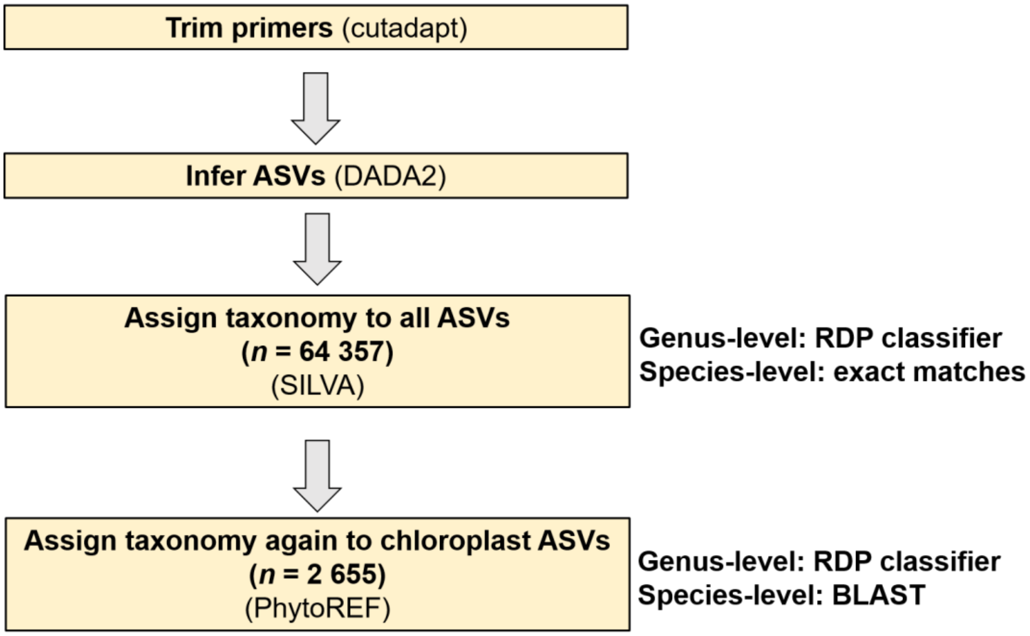
Figure 10 The DADA2 pipeline. Non-bolded text indicates the names of the functions run in the "dada2" R package (Callahan et al. 2016). *Only for paired-end reads.

The full DADA2 pipeline (Figure 10) is well-documented and can be run in R using the "dada2" package (Callahan et al. 2016). The first step in the pipeline is filtering and trimming reads, which is necessary to e.g. remove reads that are too short for downstream analysis, trim low-quality bases from the ends of reads and remove sequences with many expected errors. Next, a subset of reads is used to learn the model error rates, followed by dereplication of sequences and ASV inference using the core partitioning algorithm and the learned error model. If paired-end reads are used, read pairs are merged based on their exact overlap. PCR chimeras are removed using a global alignment that searches for a combination of left- and right-parents that exactly match the child sequence and finally, taxonomy is assigned to the remaining sequences (Callahan et al. 2016). Taxonomic classification of environmental 16S rRNA gene sequences is typically carried out using either homology-based approaches, which require alignment of query 16S rRNA sequences with 16S rRNA sequences present in the reference database e.g. UCLUST (Edgar 2010), or prediction-based approaches, the most common of which is the Ribosomal Database (RDP) Classifier (Wang et al. 2007). The RDP classifier uses a naïve Bayesian approach to classify sequences based on exact matches of 8-letter words and bootstrapping to give probability estimates. The DADA2 pipeline uses the RDP classifier to assign taxonomy to the genus-level and then species are assigned using exact matches (Callahan et al. 2016).

2.2.3. Taxonomy reference databases

The choice of a taxonomy reference database is an important component of analyzing amplicon sequencing data and depends upon the goals of the analysis. Reference taxonomy for 16S rRNA reads is usually based on one of thee reference databases: SILVA (Quast et al. 2013), RDP (Cole et al. 2014) or Greengenes (Mcdonald et al. 2011), all of which cover a large number of species and are manually curated and revised. The SILVA database contains taxonomic information for the domains of Bacteria, Archaea and Eukarya and is based primarily on phylogenies for the SSU rRNA genes (16S/18S) that are constructed from guide trees (Quast et al. 2013). In contrast, the RDP database contains taxonomic information for the 16S rRNA sequences from Bacteria and Archaea as well as the LSU rRNA (28S) sequences from Fungi (Cole et al. 2014). Finally, the Greengenes databases is dedicated to Bacteria and Archaea 16S rRNA sequences with classification based on *de novo* tree construction, but has not been updated since 2013 (Mcdonald et al. 2011). The ideal database for assigning taxonomy depends largely on the targeted genes, targeted groups and the databases that are used most often in the area of study. I chose to use the SILVA database to assign taxonomy because it is actively maintained, has SSU data for all three domains of life and is frequently used in studying ocean microbiome data (e.g. Sunagawa et al. 2015), which facilitates the comparison of this data with other studies. To improve the taxonomic assignment for photosynthetic Eukaryotes, which have 16S plastidial genes, the PhytoREF database (Decelle et al. 2015) was used. The PhytoREF database contains plastidial 16S rDNA reference sequences that originate from a large diversity of eukaryotes, with a focus on marine microalgae. Sequences in the database undergo stringent quality filtering and are assigned to taxonomy using phylogeny-based methods (Decelle et al. 2015).

2.2.4. Pipeline summary



Trim primers (cutadapt)
Infer ASVs (DADA2)
Assign taxonomy to all ASVs
(*n* = 64 357)
(SILVA)
Genus-level: RDP classifier
Species-level: exact matches
Assign taxonomy again to chloroplast ASVs
(*n* = 2 655)
(PhytoREF)
Genus-level: RDP classifier
Species-level: BLAST

Figure 11 Analysis pipeline.

Prior to processing the raw reads with DADA2 as part of the analysis pipeline (Figure 11), cutadapt v. 2.1 (Martin 2011) was used with the settings -minimum-length 210, -maximum-length 250 and --discard-untrimmed to trim the primers from the ends of the reads and discard any read pairs that did not contain both primers or were too short or too long after trimming. Next, the trimmed reads were ran through the DADA2 pipeline using the “dada2” v. 1.12.1 (Callahan et al. 2016) R package in R v. 3.6.0 (R Core Team 2019). The filterAndTrim() function was run with the settings minLen=220 and truncLen =c(220,220), which were selected based on the distribution of quality scores observed using the plotQualityProfile() function (Figure A1) and the read length required to create a large enough overlap for the read pairs to merge downstream. The dada() function was run with pool=TRUE to pool samples for the ASV inference step and facilitate the discovery of rare ASVs. Finally, the mergePairs() function was run with minOverlap=10 (instead of the default of 12) to allow read pairs from species with longer amplicons to merge. After processing reads though the DADA2 pipeline, the number of reads/sample ranged from 41,652 – 84,784 with a mean of 67,307 reads/sample (Figure 12).

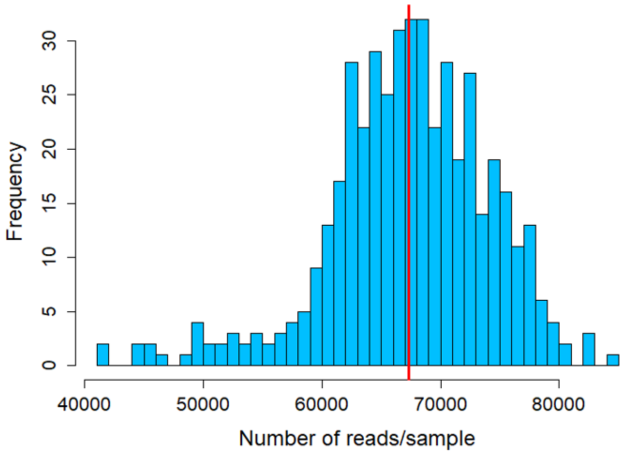


Figure 12 Histogram of the number of reads/sample after ASV inference using DADA2. The red vertical line represents the mean.

The 64,357 ASVs generated were assigned taxonomy using the SILVA nr v. 132 (Quast et al. 2013) database (Figure 11) using the assignTaxonomy() function to assign taxonomy using the RDP classifier up to the genus-level with the setting minBoot=80. The assignSpecies() function was used to assign species with the setting allowMultiple=TRUE to allow multiple species to be assigned to each ASV. To ensure the taxonomy assignment was not specific to the SILVA database, which could occur if the database representation of marine microbes was incomplete, I assigned taxonomy again using the same methods with the RDP v. 11.5 (Cole et al. 2014) database. While an accurate comparison between the two databases is challenging due to the different taxonomy systems implemented by each database, a comparison of the taxonomy of the top ten ASVs revealed similar taxonomy assignments from the two databases (Table 1). This provides some support that the results are independent of the SILVA database. However, it is important to consider that taxonomy assignments from any database are unlikely to be 100% accurate due to the incompleteness of taxonomy databases, which could result in an ASV being incorrectly assigned to the most similar sequence in the databases.

For the 2,655 ASVs that were assigned as “Chloroplast” by SILVA, taxonomy was assigned again using the PhytoREF v. 1.0 (Decelle et al. 2015) database (Figure 11) with the assignTaxonomy() function and minBoot=80. Due to the smaller size of the PhytoREF database, which makes it challenging to assign species using exact matches, species were assigned using blastn from BLAST+ v. 2.10.0 (Camacho et al. 2009) against the PhytoREF database with the setting -qcov_hsp_perc 95 and the results were parsed to find the best hit(s) with PID > 90.

After removing singletons, 55,985 ASVs remained, which were subsequently classified as abundant, intermediate or rare: 2,425 ASVs were classified as abundant (≥0.1% of reads in at least one sample), 15,236 were intermediate (<0.1% and >0.001% of reads in at least one sample) and 38,324 were rare (<0.001% of reads in all samples). Of the abundant ASVs, the majority (87.7%) could not be assigned to a classified species (this excludes species assignments such as “uncultured_bacterium”) using SILVA (Figure 13a), which reflects the incompleteness of the reference database for marine microbiome data. Only 6.3% of abundant ASVs were assigned to a single classified species, while 6.1% were assigned to >1 classified species (Figure 13a), demonstrating the occurrence of multiple species with the same V3-V4 16S rRNA sequence. A smaller proportion of abundant chloroplast ASVs (42.4%) could not be assigned to a classified species using PhytoREF (Figure 13b), likely because these ASVs were assigned to species less stringently using BLAST. Additionally, 21.1% of abundant chloroplast ASVs were assigned to >1 classified species (Figure 13b), likely due to the increased ambiguity from assigning species using BLAST. The rarefaction curve suggests that the sample depth was adequate for ASV discovery as the number of ASVs plateaued with the number of reads in each sample (Figure 14).

Table 1 Comparison of taxonomy assigned from the Phylum to the Genus levels for the first ten ASVs using the SILVA (S) (Quast et al. 2013) and RDP (R) (Wang et al. 2007) databases (D).

ASV	D	Phylum	Class	Order	Family	Genus
ASV_1	S	Cyanobacteria	Oxyphotobacteria	Synechococcales	Cyanobiaceae	Synechococcus_CC9902
ASV_1	R	Cyanobacteria/Chloroplast	Cyanobacteria	Family_II	GpIIa	NA

ASV	D	Phylum	Class	Order	Family	Genus
ASV_2	S	Proteobacteria	Alphaproteobacteria	SAR11_clade	Clade_I	Clade_Ia
ASV_2	R	Proteobacteria	Alphaproteobacteria	SAR11	Candidatus_Pelagibacter	NA
ASV_3	S	Proteobacteria	Alphaproteobacteria	SAR11_clade	Clade_I	Clade_Ia
ASV_3	R	Proteobacteria	Alphaproteobacteria	SAR11	Candidatus_Pelagibacter	NA
ASV_4	S	Actinobacteria	Acidimicrobiia	Actinomarinales	Actinomarinaceae	Candidatus_Actinomarina
ASV_4	R	NA	NA	NA	NA	NA
ASV_5	S	Thaumarchaeota	Nitrososphaeria	Nitrosopumilales	Nitrosopumilaceae	Candidatus_Nitrosopumilus
ASV_5	R	Thaumarchaeota	Nitrosopumilales	Nitrosopumilaceae	Nitrosopumilus	NA
ASV_6	S	Cyanobacteria	Oxyphotobacteria	Chloroplast	NA	NA
ASV_6	R	Cyanobacteria/Chloroplast	Chloroplast	Chloroplast	Bacillariophyta	NA
ASV_7	S	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Alteromonas
ASV_7	R	Proteobacteria	Gammaproteobacteria	Alteromonadales	Alteromonadaceae	Alteromonas
ASV_8	S	Proteobacteria	Alphaproteobacteria	SAR11_clade	Clade_I	Clade_Ia
ASV_8	R	Proteobacteria	Alphaproteobacteria	SAR11	Candidatus_Pelagibacter	NA
ASV_9	S	Cyanobacteria	Oxyphotobacteria	Chloroplast	NA	NA
ASV_9	R	Cyanobacteria/Chloroplast	Chloroplast	Chloroplast	Bacillariophyta	NA
ASV_10	S	Cyanobacteria	Oxyphotobacteria	Chloroplast	NA	NA
ASV_10	R	Cyanobacteria/Chloroplast	NA	NA	NA	NA

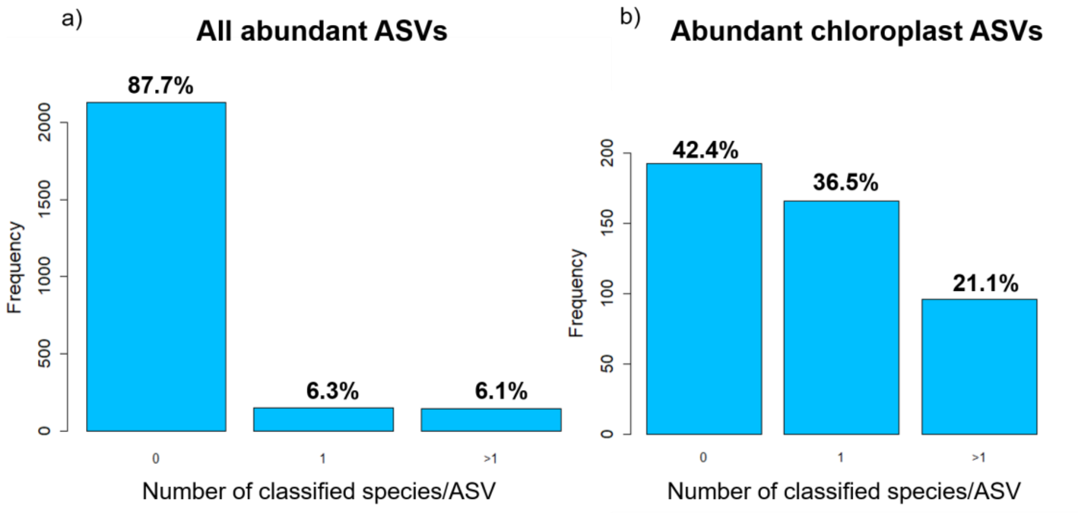


Figure 13 The number of classified species assigned to each ASV for a) all abundant ASVs and b) abundant chloroplast ASVs. Abundant ASVs are defined as making up ≥0.1% of reads in at least one sample. The ASVs in a) were assigned to species using exact matches against the SILVA (Quast et al. 2013) database, while the ASVs in b) were assigned to species using BLAST against the PhytoREF (Decelle et al. 2015) database.

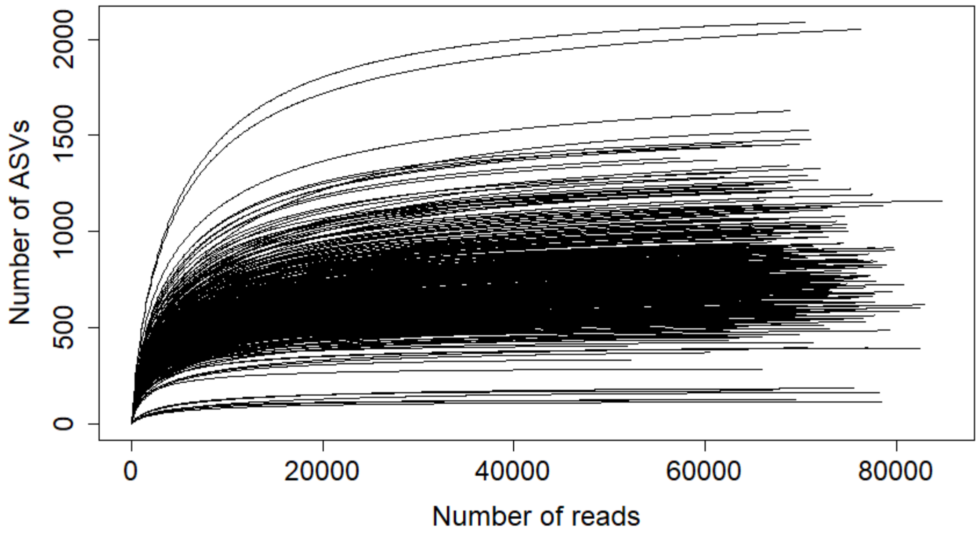


Figure 14 Rarefaction curve generated using the “vegan” v. 2.5-6 (Oksanen et al. 2019) R package.

Chapter 3. Exploration of the microbial composition of the Beibu Gulf

The aim of this chapter is to explore the composition of *P. globosa* and other microbes in the Beibu Gulf epipelagic layer using 16S amplicon sequencing data. First, I identify ASVs that represent *P. globosa*, explore the sequence conservation of the *P. globosa* 16S rRNA gene and explore the composition of the putative *P. globosa* ASVs in the field and colony samples. After, I explore the composition of the other microbial organisms detected in addition to *P. globosa* by the analysis of the field samples. My contribution to this chapter is the data analysis and interpretation.

3.1. *P. globosa* ASVs

ASV_10 was identified as a perfect match to the 16S V3-V4 sequence from a published *P. globosa* chloroplast genome (NCBI reference sequence: NC_021637.1:39760-41245). To explore the level of conservation of the 16S rRNA gene across *P. globosa* strains, additional copies of the 16S rRNA gene were assembled from whole-genome sequencing of 49 different *P. globosa* strains collected from the Beibu Gulf (*n* = 15), the East China Sea (*n* = 3), Vietnam (*n* = 27) and Thailand (*n* = 4) (Appendix B). 47/49 strains had complete assemblies of the 16S rDNA sequence, 44 of which were identical to the full NC_021637.1 16S rDNA sequence. The other three strains, one from the Beibu Gulf, one from the East China Sea and one from Vietnam, each differed from the NC_021637.1 16S rDNA sequence by one nucleotide (PID = 99.9%) at different positions. Only one of the three strains differed in the amplified V3-V4 region, but this sequence was not identical to any of the ASVs. Given that the NC_021637.1 strain was collected from the North Sea, this provides evidence that the *P. globosa* 16S rRNA gene is highly conserved locally within the Beibu Gulf and internationally between strains in the Beibu Gulf, the East China Sea, Vietnam, Thailand and the North Sea. Next, I performed a phylogenetic analysis of additional ASVs in the Beibu Gulf that may represent *P. globosa* to further explore the level of conservation of the *P. globosa* 16S rRNA gene.

3.1.1. Methods for building phylogenies from marker gene sequencing data

Many sophisticated model-based approaches exist for building phylogenetic trees from molecular data. In the case of marker gene sequencing data, phylogenies are used to represent hypotheses about the evolutionary relationships among OTUs/ASVs. The first step in building a phylogeny from this type of data is aligning the OTUs/ASVs using a multiple-sequence alignment program, the most common of which are Clustal (Larkin et al. 2007) and MAFFT (Katoh and Standley 2013). Both programs rely on similar methods: a pairwise distance matrix is built from the input sequences, which is used to construct an initial guide tree that is used to align the sequences. This initial alignment is scored and used to produce a new guide tree and subsequent alignment, which is done iteratively until a best-scoring alignment is reached. The next step after sequence alignment is selecting a model that best describes the evolutionary process that generated the data at hand. Model-based approaches for building phylogenies are now used more frequently than simpler approaches such as Maximum Parsimony and Neighbor-Joining methods. Model selection is performed within either a maximum likelihood (ML) framework, which typically involves comparing the ML score of a set of models using e.g. the Akaike information criterion (AIC), or a Bayesian inference framework. The selected model of evolution and the aligned OTUs/ASVs are then used to build the phylogeny using either ML software, e.g. RAxML (Stamatakis 2014), or Bayesian inference software such as MrBayes (Ronquist et al. 2012) and BEAST (Suchard et al. 2018). Finally, the reliability of the phylogenetic tree can be assessed using either ML or Bayesian inference. In the ML approach, branch support values are estimated using nonparametric bootstrapping. The bootstrapping procedure involves the production of pseudo-replicates by randomly resampling characters from the original data, which are subject to the same phylogenetic analysis as the original data. The bootstrap support for each clade is calculated as the proportion of times that the clade is obtained in the pseudo-replicates (Felsenstein 1985). In contrast, the Bayesian approach uses posterior probabilities to assess reliability.

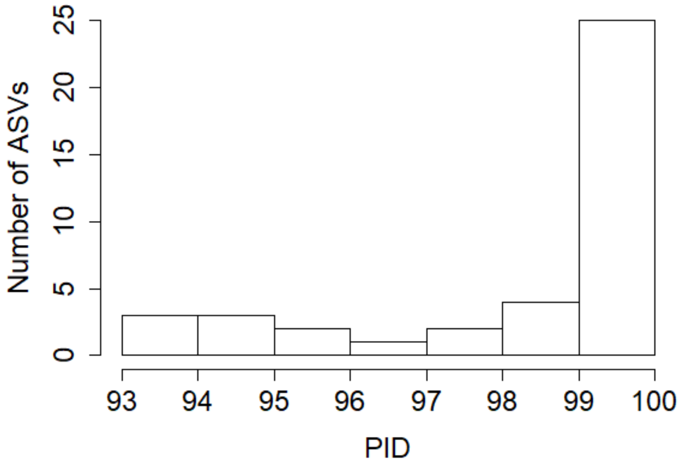


Figure 15 **Histogram of the percentage identity (PID) to the *P. globosa* NC_021637.1 reference sequence of ASVs assigned to *P. globosa* and with at least one read in one the six *P. globosa* colony samples.**

3.1.2. Phylogeny of putative *P. globosa* ASVs

A phylogeny was constructed of the additional putative *P. globosa* ASVs in the data to generate a hypothesis of their evolutionary relationship to the other haptophytes. Of the 99 chloroplast ASVs assigned to *P. globosa* using PhytoREF (Decelle et al. 2015), 40 ASVs had at least one read in one of the six *P. globosa* colony samples, providing some further support that these ASVs represented *P. globosa*. Additionally, most of the 40 putative *P. globosa* ASVs had PID > 99% to the *P. globosa* NC_021637.1 reference sequence (Figure 15). The phylogeny was constructed using the top 20 most abundant putative *P. globosa* ASVs and 25 other haptophyte sequences from PhytoREF that were trimmed to the same primer sequences. The sequences (*n* = 45) were aligned using Clustal Omega v. 1.2.4 (Sievers et al. 2011) with default settings and manually edited using MEGA-X v. 10.0.5 (Kumar et al. 2018). Model selection, performed in MEGA using ML, selected K2+I as the model with the lowest AIC. The phylogeny was subsequently constructed in MEGA using the K2+I model with 1,000 bootstrap replicates. While most of the putative *P. globosa* ASVs were closely related to the *P. globosa* reference sequence, supporting a high level of conservation of the *P. globosa* 16S V3-V4 region, one group of ASVs formed a separate clade (Figure 16).

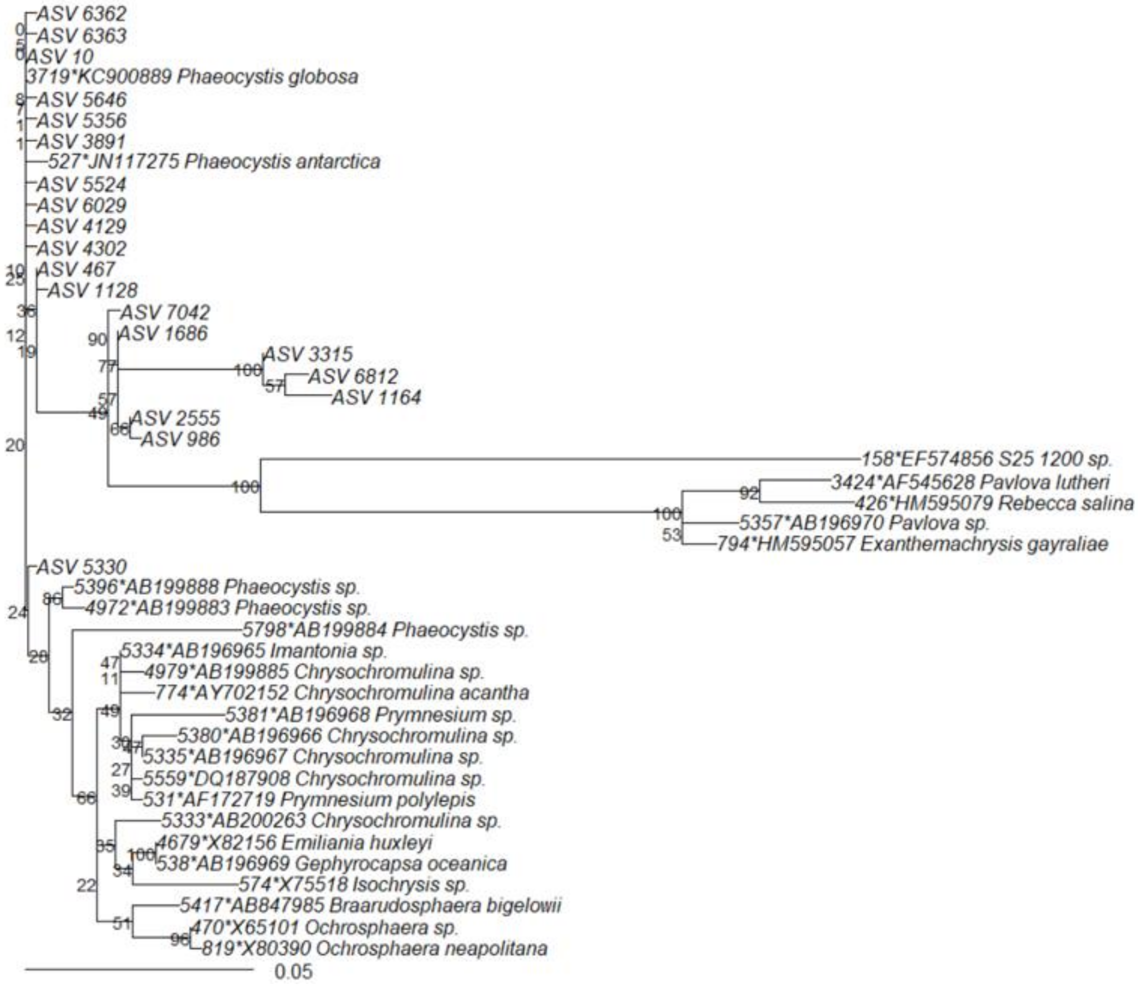

ASV 6362
ASV 6363
ASV 10
3719*KC900889 Phaeocystis globosa
ASV 5646
ASV 5356
ASV 3891
527*JN117275 Phaeocystis antarctica
ASV 5524
ASV 6029
ASV 4129
ASV 4302
ASV 467
ASV 1128
ASV 7042
ASV 1686
ASV 3315
ASV 6812
ASV 1164
ASV 2555
ASV 986
158*EF574856 S25 1200 sp.
3424*AF545628 Pavlova lutheri
426*HM595079 Rebecca salina
5357*AB196970 Pavlova sp.
794*HM595057 Exanthemachrysis gayraliae
ASV 5330
5396*AB199888 Phaeocystis sp.
4972*AB199883 Phaeocystis sp.
5798*AB199884 Phaeocystis sp.
5334*AB196965 Imantonia sp.
4979*AB199885 Chrysochromulina sp.
774*AY702152 Chrysochromulina acantha
5381*AB196968 Prymnesium sp.
5380*AB196966 Chrysochromulina sp.
5335*AB196967 Chrysochromulina sp.
5559*DQ187908 Chrysochromulina sp.
531*AF172719 Prymnesium polylepis
5333*AB200263 Chrysochromulina sp.
4679*X82156 Emiliania huxleyi
538*AB196969 Gephyrocapsa oceanica
574*X75518 Isochrysis sp.
5417*AB847985 Braarudosphaera bigelowii
470*X65101 Ochrosphaera sp.
819*X80390 Ochrosphaera neapolitana
90
77
100
57
57
49
66
20
12
19
36
100
92
100
53
86
24
28
32
66
47
11
49
39
27
39
35
100
34
22
51
96
0.05

Figure 16 **Maximum likelihood phylogeny of 20 putative *P. globosa* ASVs and 25 haptophyte sequences from PhytoREF. The phylogeny was constructed in MEGA-X (Kumar et al. 2018) using the K2+I model with 1000 bootstrap replicates. The scale bar represents the mean number of nucleotide substitutions per site and the node values represent the bootstrap support.**

3.1.3. ASV_10

To further validate the utility of ASV_10 for tracking *P. globosa* throughout the bloom, I compared the relative abundance of ASV_10 between the field and colony samples (Figure 17a). As expected, ASV_10 made up most of the reads in the six colony samples (range: 0.80-0.94) (Figure 17a), but a lower fraction in the field samples (mean = 0.0065, median = 0.0022, range: 0.00-0.14) (Figure 17). Additionally, when considering only the 40 putative *P. globosa* ASVs, both the colony and field samples were dominated by ASV_10 (Figure 18). The high relative abundance of ASV_10 in the colony samples and its dominance over the other putative *P. globosa* ASVs supports its application as a marker for *P. globosa*. The lower relative abundance of ASV_10 in the field samples (Figure 17b) is likely because most of the colonies were too large to pass through the filtration procedure, thus diluting their presence in the field samples to only very small colonies and free-living *P. globosa* cells.

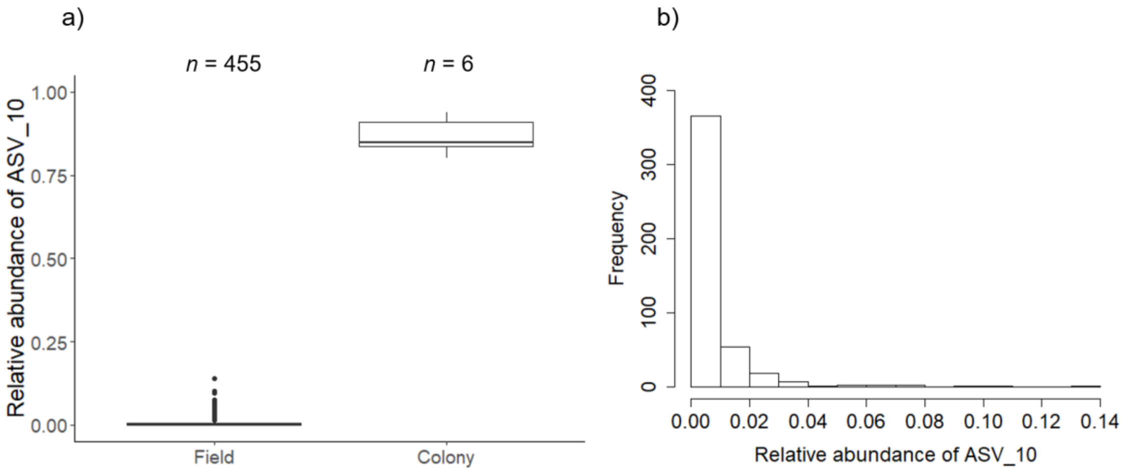


Figure 17 a) Comparison of the relative abundance of ASV_10 (*P. globosa*) in the field vs. colony samples and b) histogram of the relative abundance of ASV_10 in the field samples.

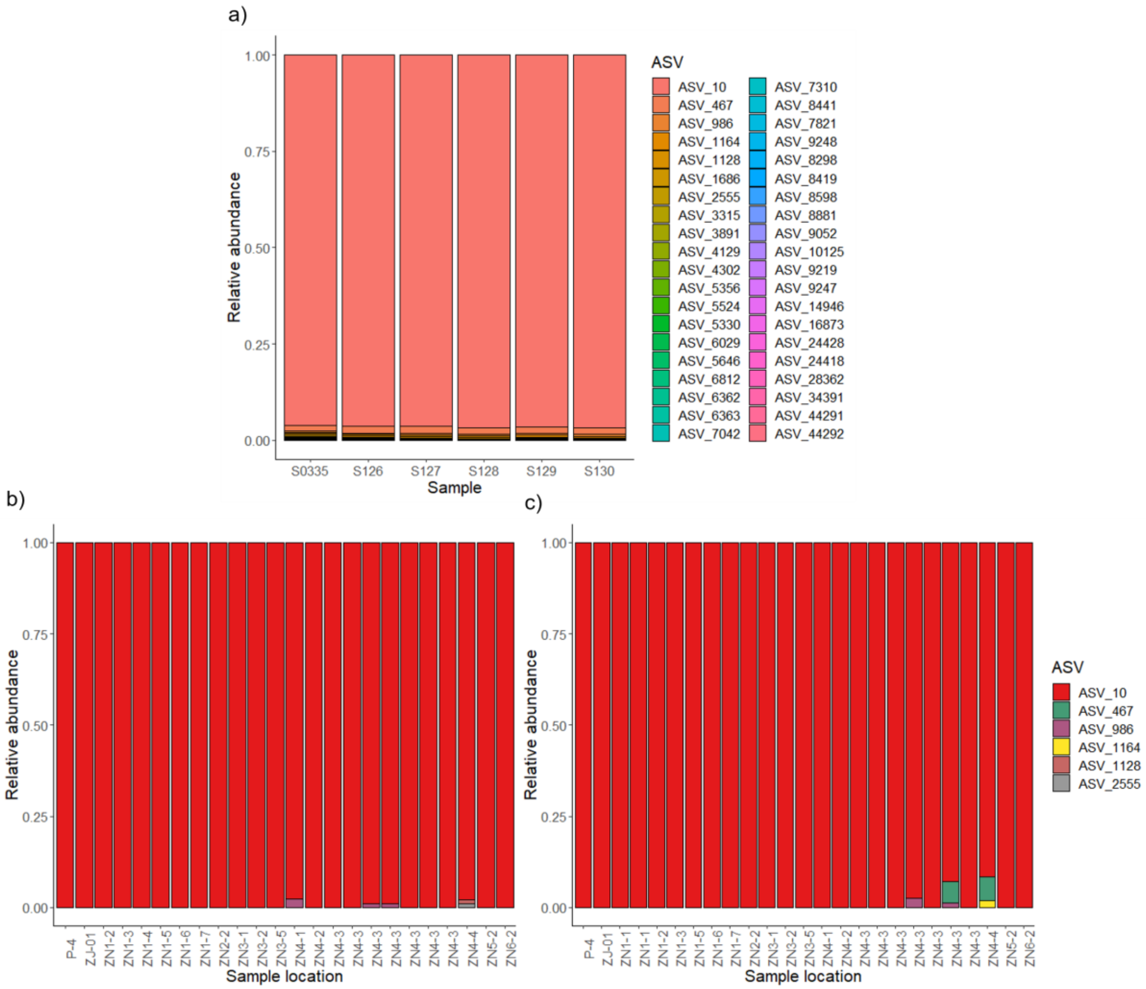


Figure 18 The relative abundance of ASV_10 and other putative *P. globosa* ASVs in a) the colony samples, b) the small filtration size surface field samples from the January expedition and c) the large filtration size surface field samples from the January expedition when considering only putative *P. globosa* ASVs. Only ASV_10 was detected in the Feb-March and April expeditions.

3.2. Composition of other microbial organisms in the Beibu Gulf

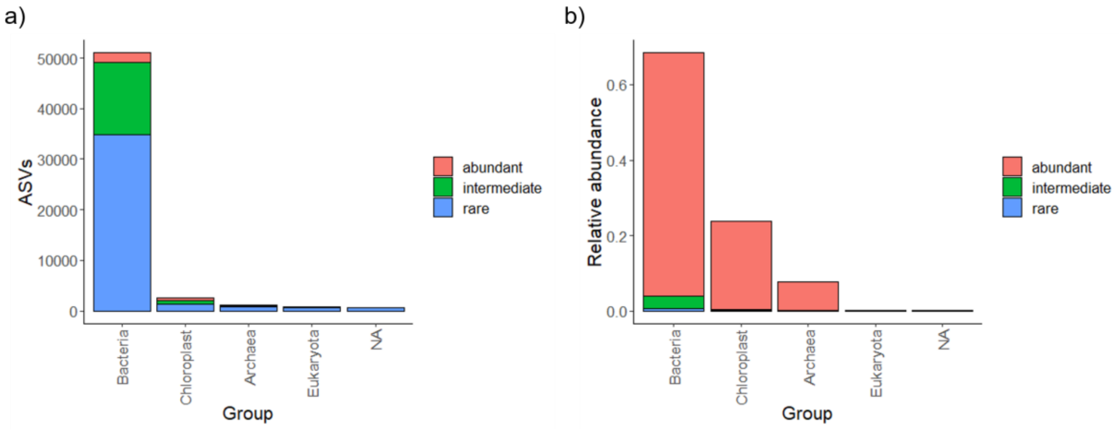


Figure 19 a) The number of ASVs and b) the relative abundance of reads from the field samples assigned by SILVA (Quast et al. 2013) to the kingdoms Bacteria, Archaea and Eukaryota and the order Chloroplast. See text for the definitions of abundant, intermediate and rare.

After excluding singletons, the majority of ASVs (*n* = 51,014) from the field samples were classified as Bacteria with only 2,546 ASVs classified as Chloroplasts and 1,101 ASVs classified as Archaea (Figure 19a). The small number of ASVs (*n* = 718) classified as Eukaryota were from reads assigned to 18S rRNA sequences in the SILVA database. Most of the Eukaryota ASVs (666/718) were rare and only 0.017% of reads were classified as Eukaryota, which suggests that the amplification of 18S rRNA was rare. Like the number of ASVs, most of the reads (68.5%) from the field samples were classified as Bacteria, followed by 23.8% of reads as Chloroplasts and 7.7% of reads as Archaea (Figure 19b).

3.2.1. Bacteria

The majority of bacteria ASVs (47,014/51,014) were classified at the class level with most ASVs belonging to the Gammaproteobacteria (*n* = 18,204), Bacteroidia (*n* = 7,823), Deltaproteobacteria (*n* = 5,933), Alphaproteobacteria (*n* = 5,671), Verrucomicrobiae (*n* = 1,310), Acidimicrobiia (*n* = 702) and Oxyphotobacteria (*n* = 652) (Figure 20a). Similarly, more reads were assigned to a class of Bacteria (67.9%) than were unclassified (0.6%). However, the Alphaproteobacteria were the class of bacteria with the most reads (28.5%), followed by the Gammaproteobacteria (18.7%), Bacteroidia

(6.2%), Acidimicrobiia (5.9%), Oxyphotobacteria (5.1%), Deltaproteobacteria (1.6%) and Verrucomicrobiae (1.0%) (Figure 20b).

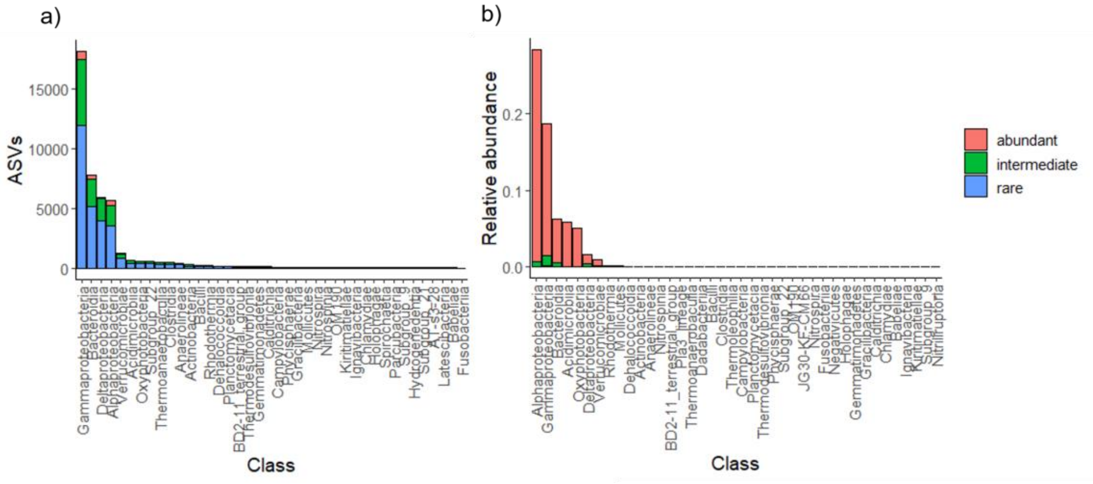


Figure 20 **a) The number of ASVs and b) the relative abundance of reads from the field samples assigned by SILVA (Quast et al. 2013) to the different classes of Bacteria. Only the top 40 classes are shown. See text for the definitions of abundant, intermediate and rare.**

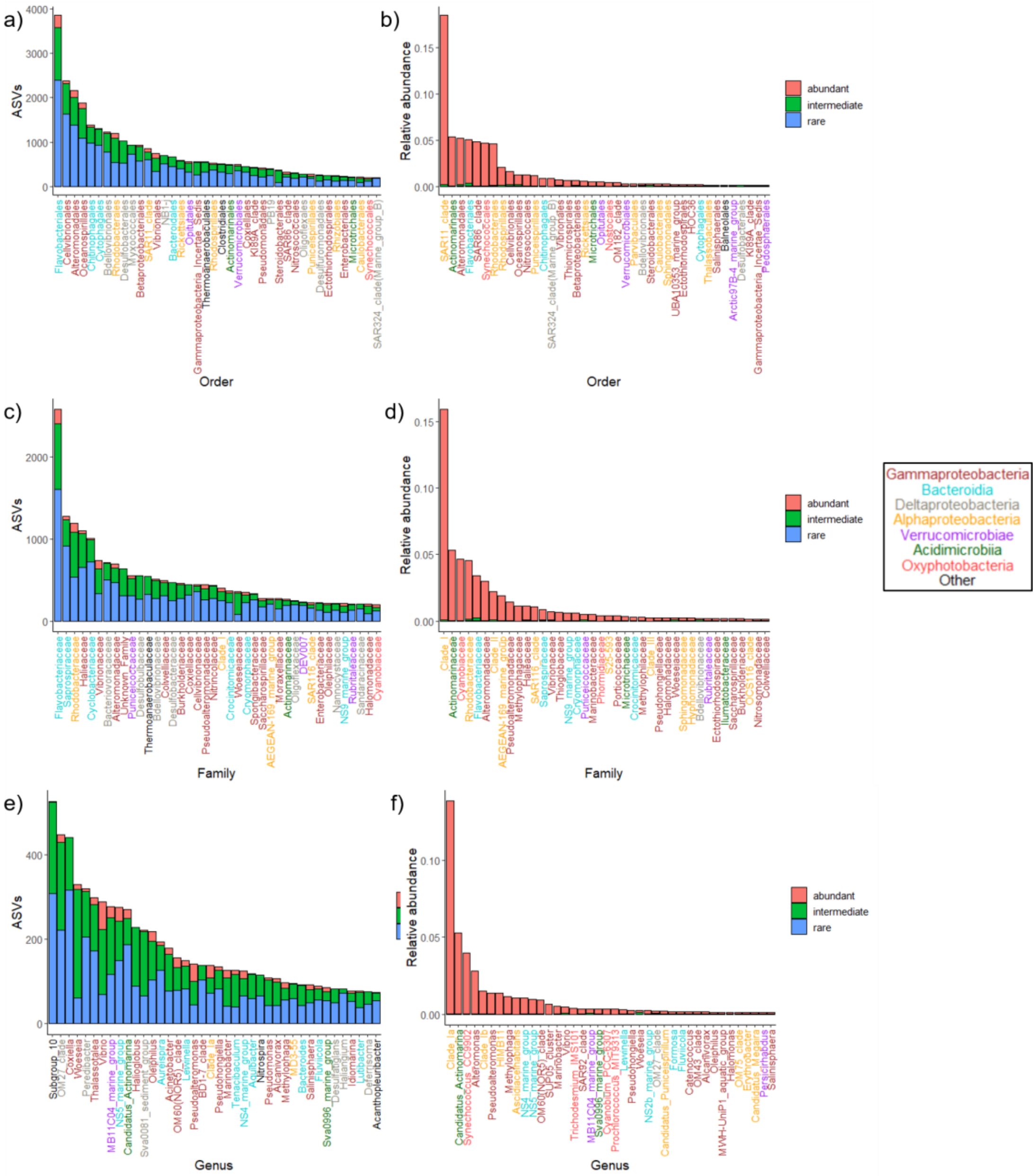


Figure 21 **a, c and e) The number of ASVs and b, d and f) the relative abundance of reads from the field samples assigned by SILVA (Quast et al. 2013) to the different a-b) orders, c-d) families and e-f) genera of Bacteria. Only the top 40 groups are shown for each level. See text for the definitions of abundant, intermediate and rare.**

Of the 38,143 ASVs assigned to an order of bacteria, the order with the most ASVs assigned (*n* = 3,863) was the Flavobacteriales from the class Bacteroidia (Figure 21a). In contrast, when the number of reads was considered, the most common order of bacteria was the SAR11_clade (18.4% of reads) from the Alphaproteobacteria (Figure

21b). Of the 31,780 ASVs assigned to a family of bacteria, the family with the most ASVs (*n* = 2,584) was the Flavobacteriaceae (Figure 21c). Like the order level, the family of bacteria with the most reads assigned (15.9%) was Clade_I from the SAR11_clade (Figure 21d). Finally, of the 16,577 ASVs assigned to a genus of bacteria, the genus with the most ASVs (*n* = 527) was Subgroup_10 (*n* = 527 ASVs) from the class Thermoanaerobaculia. However, most of these ASVs were intermediate or rare and the genus Vibrio from the Gammaproteobacteria had the most abundant ASVs (Figure 21e). The genus of bacteria with the most reads (13.9%) was Clade_Ia from the SAR11_clade (Figure 21f).

3.2.2. Archaea

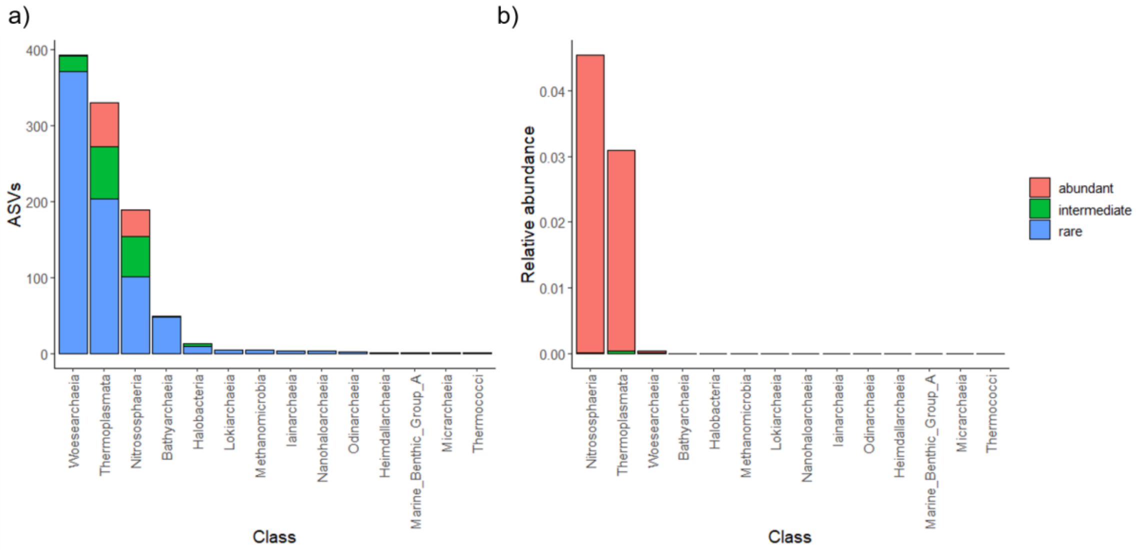


Figure 22 a) The number of ASVs and b) the relative abundance of reads from the field samples assigned by SILVA (Quast et al. 2013) to the different classes of Archaea. Abundant ASVs make up ≥0.1% of reads in at least one sample, See text for the definitions of abundant, intermediate and rare.

Most Archaea ASVs (996/1,101) were classified at the class level with the most ASVs assigned to the Woesearchaeia (*n* = 393), Thermoplasmata (*n* = 330) and Nitrososphaeria (*n* = 189) (Figure 22a). However, the majority of ASVs assigned to the Woesearchaeia were intermediate or rare and thus the classes of Archaea with the most reads were the Nitrososphaeria (4.5% of reads) and Thermoplasmata (3.1% of reads) (Figure 22b).

Of the 511 ASVs assigned to an order of Archaea, most were assigned to Marine_Group_II (*n* = 257) from the class Thermoplasmata or the Nitrosopumilales (*n* = 188) from the class Nitrososphaeria (Figure 23a). Similarly, most reads were assigned to either the Nitrosopumilales (4.5% of reads) or Marine_Group_II (3.1% of reads) (Figure 23b). Of the 206 ASVs assigned to a family of Archaea, almost all (*n* = 188) were assigned to the Nitrosopumilaceae (from the Nitrososphaeria) as the taxonomy for the Marine_Group_II does not go beyond the order level (Figure 23c). Likewise, the Nitrosopumilaceae were the order of Archaea with the most reads (4.5%) (Figure 23d). Of the 149 ASVs assigned to an Archaea genus, most were assigned to Candidatus_Nitrosopumilus (*n* = 104) and Candidatus_Nitrosopelagicus (*n* = 30), both from the Nitrososphaeria (Figure 23e). These two genera also had the most reads: 3.6% of reads were assigned to Candidatus_Nitrosopumilus and 0.9% to Candidatus_Nitrosopelagicus (Figure 23f).

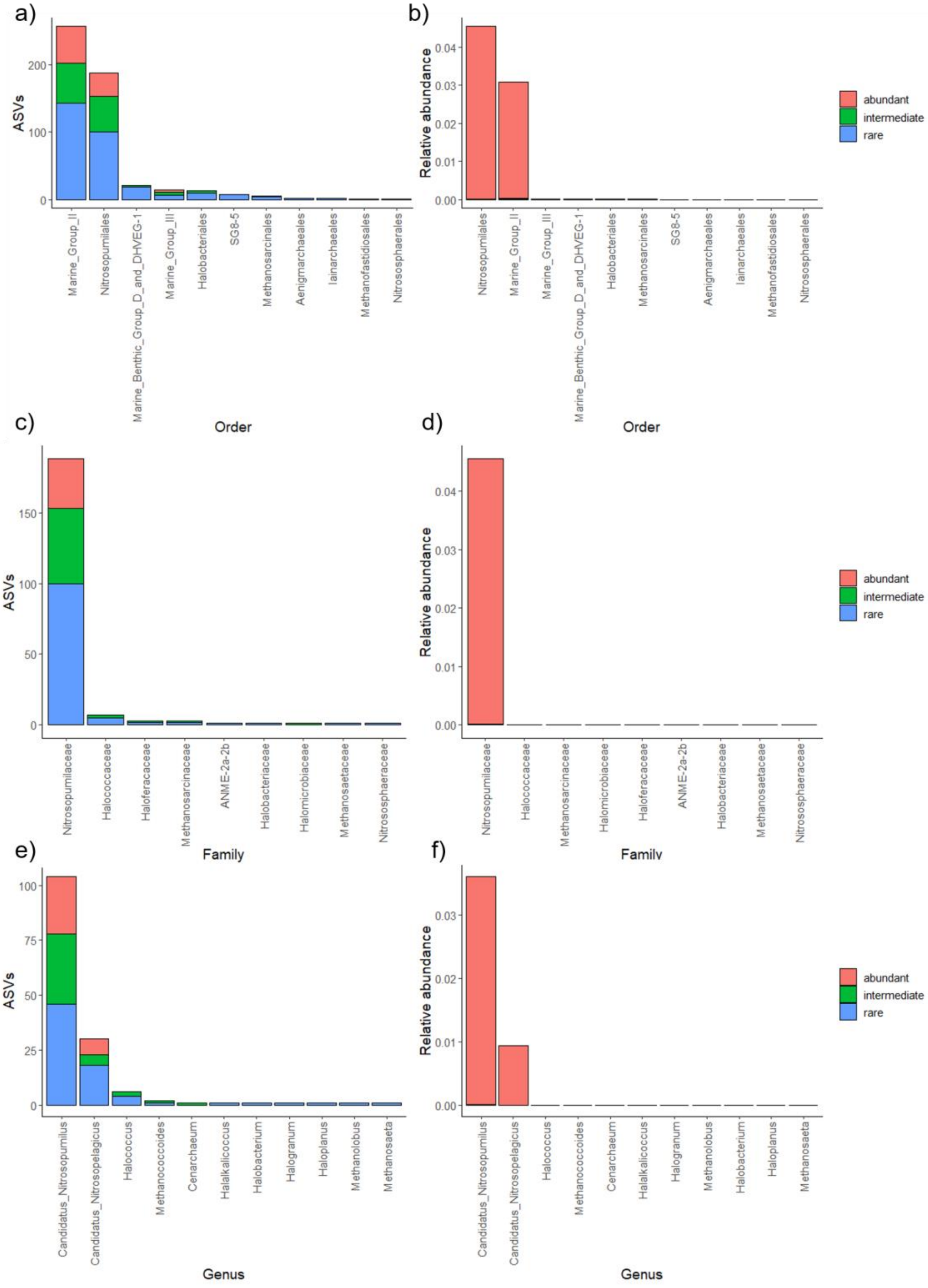


Figure 23 a, c and e) The number of ASVs and b, d and f) the relative abundance of reads from the field samples assigned by SILVA (Quast et al. 2013) to the different a-b) orders, c-d) families and e-f) genera of Archaea. See text for the definitions of abundant, intermediate and rare.

3.2.3. Chloroplasts

Most Chloroplast ASVs (2,307/2,557) were classified at the class level with the most ASVs assigned to the Bacillariophyta i.e. the diatoms (*n* = 1,518), Prymnesiophyceae (*n* = 304), Dictyophyceae (*n* = 105) and Cryptophyceae (*n* = 104) (Figure 24a). The three classes of Eukaryotes from the Chloroplasts with the most reads were the Bacillariophyta (20.5% of reads), Prymnesiophyceae (1.5% of reads) and the Cryptophyceae (1.1% of reads) (Figure 24b).

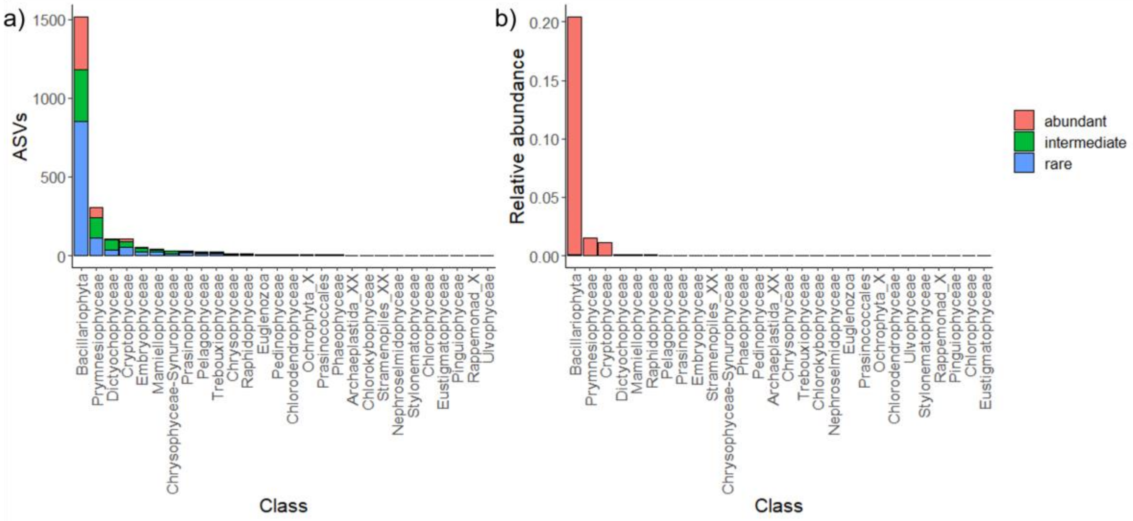


Figure 24 **a) The number of ASVs and b) the relative abundance of reads from the field samples assigned by PhytoREF (Decelle et al. 2015) to the different classes of Eukaryota. See text for the definitions of abundant, intermediate and rare.**

Of the 1,029 ASVs assigned to an order of Eukaryotes from the Chloroplasts, most were assigned to the Pyrenomonadales (*n* = 104) from the Cryptophyceae, followed by the Chaetocerotales (*n* = 96) from the Bacillariophyta and the Phaeocystales (*n* = 82) from the Prymnesiophyceae (Figure 25a). The three orders with the most reads were all from the Bacillariophyta: the Thalassiosirales (4.7% of reads), Chaetocerotales (1.2% of reads) and Coscinodiscales (1.1% of reads) (Figure 25b). Of the 842 Chloroplast ASVs assigned to a family, the most ASVs (*n* = 82) were assigned to the Phaeocystaceae (*n* = 82) (Figure 25c). The family with the most reads (1.1%) was the Coscinodiscaceae, while the Phaeocystaceae had the third-most reads (0.7%) (Figure 25d). Finally, of the 603 ASVs assigned to a genus from the Chloroplasts, the most ASVs (*n* = 68) were assigned to an unclassified group of Dictyophyceae, followed by the genus *Phaeocystis* (*n* = 40) (Figure 25e). The genus with the most reads (1.0%) was an unclassified group of Coscinodiscaceae (Figure 25f).

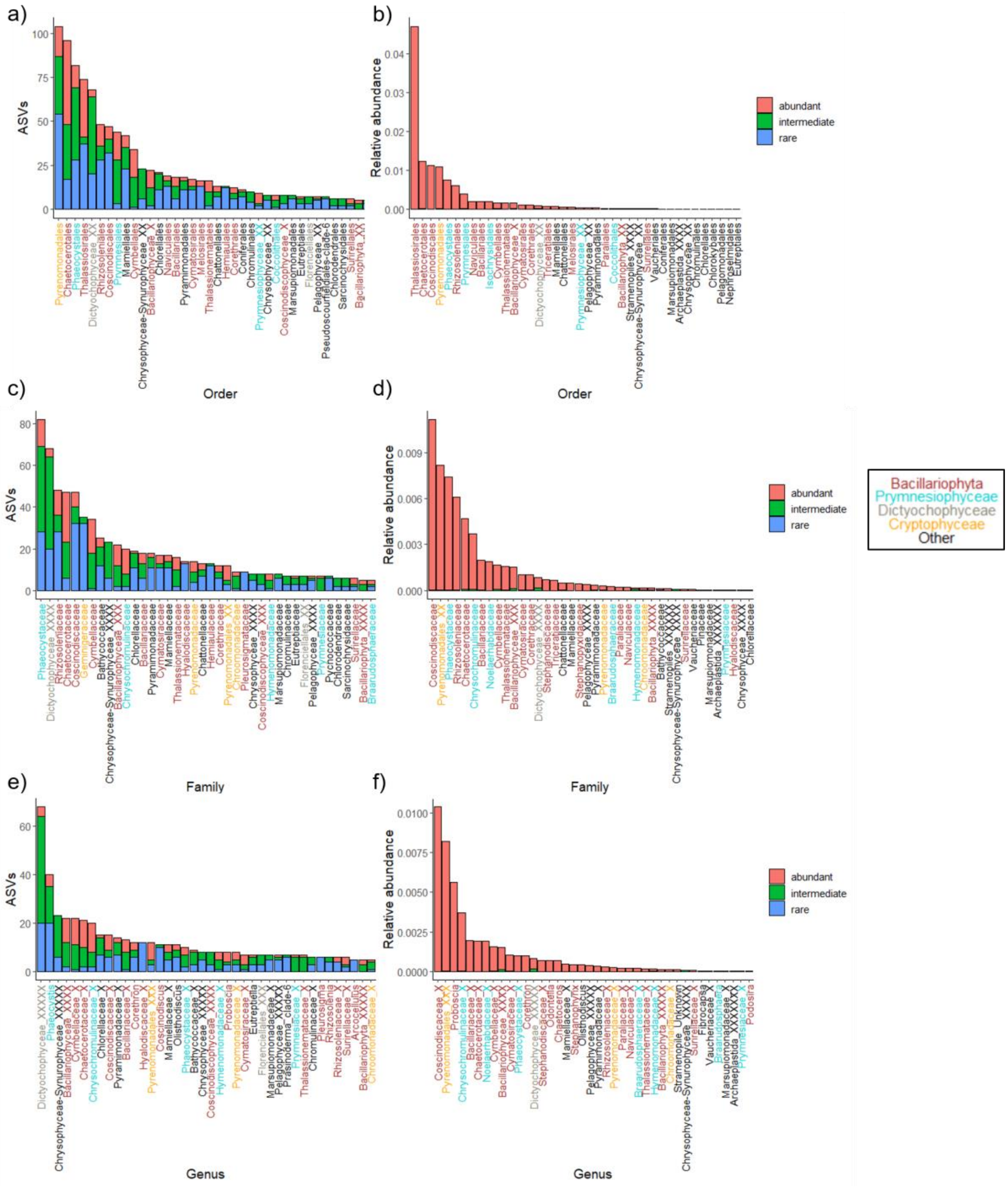


Figure 25 **a, c and e) The number of ASVs and b, d and f) the relative abundance of reads from the field samples assigned by PhytoREF (Decelle et al. 2015) to the different a-b) orders, c-d) families and e-f) genera of Eukaryota. Only the top 40 groups are shown for each level. See text for the definitions of abundant, intermediate and rare.**

3.3. Discussion

ASV_10 was identified as a suitable molecular marker for *P. globosa* due to the low level of intraspecific variation in the *P. globosa* 16S rRNA gene. To my knowledge, no previous studies have investigated the conservation of the 16S rRNA gene in *P. globosa* as previous work has focused largely on the conservation of the 18S SSU rRNA, 28S LSU rRNA, ITS and other plastid regions such as the RUBISCO spacer regions (Lange et al. 2002, Medlin and Zingone 2007, Xiaokun et al. 2019, Qingchun et al. 2020). I found a high level of conservation of the *P. globosa* 16S rRNA gene in both 1) the 16S rRNA sequences assembled from whole-genome sequencing of strains from the Beibu Gulf, the East China Sea, Vietnam and Thailand and 2) the ASVs from the 16S V3-V4 region in the Beibu Gulf. This high level of conservation may be explained at least partly by the mutation rates of the *P. globosa* chloroplast genome, which are lower than in the mitochondria and nucleus (Smith et al. 2014). Despite the high level of conservation of the 16S rRNA, I did find some evidence for intraspecific variation in this gene. While ASV_10 dominated the field and colony samples when considering only putative *P. globosa* ASVs, other putative *P. globosa* ASVs were also present in low levels (Figure 18). Interestingly, the phylogenetic analysis identified an additional clade of putative *P. globosa* ASVs that was separate from ASV_10 (Figure 16). While these ASVs may represent *P. globosa*, it is also possible that they represent other species such as cryptic *P. globosa*-like species. This would be consistent with previous suggestions that *P. globosa* colonies are actually complexes of up to three or four cryptic species (Medlin and Zingone 2007). Alternatively, some of the observed variation may be due to sequencing errors that were not corrected by DADA2. Overall, while I identified some level of intraspecific variation in the *P. globosa* 16S rRNA gene, the dominance of ASV_10 in the colony samples (Figure 17a) and over the other *P. globosa* ASVs makes it a suitable molecular marker for *P. globosa* for the remaining analyses.

The remaining composition of microbes in the Beibu Gulf epipelagic layer was generally consistent with studies from surrounding regions. At the class level, I found that the Alphaproteobacteria were the class of Bacteria with the most reads (28.5%), followed by the Gammaproteobacteria (18.7%), Bacteroidia (6.2%), Acidimicrobiia (5.9%) and Oxyphotobacteria (5.1%) (Figure 20b). Similarly, a previous study of free-living bacteria from the Beibu Gulf during a *P. globosa* bloom found that the bacterial community was dominated by the phyla Proteobacteria (Alphaproteobacteria and Gammaproteobacteria), Bacteroidetes (Bacteroidia) and Actinobacteria (Acidimicrobiia) (Li et al. 2020b). My results are also consistent with the bacterial composition of nearby areas, e.g. the South China Sea, which is mainly composed of Alphaproteobacteria (mostly SAR11), Gammaproteobacteria, Cyanobacteria and Bacteroidetes (Zhang et al. 2018) and the western North Pacific ocean, which is mainly composed of Alphaproteobacteria (mostly SAR11 and *Roseobacter*), Gammaproteobacteria (mostly SAR86 and *Alteromonas*) and the Bacteroidetes during phytoplankton blooms (Tada et al. 2011). For the archaea, I found that most reads were assigned to either the Nitrosopumilales (4.5% of reads) from the class Nitrososphaeria or Marine_Group_II (MGII) from the class Thermoplasmata (3.1% of reads) (Figure 23b). Similarly, a recent study in the South China Sea found that the archaea communities were composed mainly of Nitrososphaeria and MGII (Li et al. 2020b). In contrast, another study in the South China Sea found that archaea communities were dominated by MGII at all depths (Liu et al. 2017). Finally, I found that the three classes of Eukaryotes from the Chloroplasts with the most reads were the Bacillariophyta (20.5% of reads), Prymnesiophyceae (1.5% of reads) and the Cryptophyceae (1.1% of reads) (Figure 24b). The dominance of diatoms (i.e. Bacillariophyta) is not surprising given that other typically abundant groups of Eukaryotes like the Dinophyceae mostly do not have chloroplasts and cannot be detected using the 16S rRNA gene. Overall, the microbial composition of the Beibu Gulf epipelagic layer was mostly consistent with other studies in nearby regions and my next aim involved exploring the spatial-temporal dynamics of these microbes during the *P. globosa* bloom to better understand their role in the bloom.

Chapter 4. Spatial-temporal dynamics of microbes and environmental variables during a *P. globosa* bloom in the Beibu Gulf

In the previous chapter I explored the composition of *P. globosa* and other microbes in the Beibu Gulf. The aim of this chapter is to explore the spatial-temporal dynamics of *P. globosa*, bacteria, archaea, phytoplankton, and environmental factors during a *P. globosa* bloom in the Beibu Gulf. First, I explore the spatial-temporal dynamics of *P. globosa* during its bloom. Next, I explore the temporal dynamics of the alpha and beta diversity of the bacteria, archaea and chloroplasts during the bloom. After, I explore the spatial-temporal dynamics of different groups of bacteria, archaea and chloroplasts. Finally, I explore the spatial-temporal dynamics of the environmental variables and their relationships with ASV_10 (*P. globosa*) and other microbes. Collectively, this exploration will improve our understanding of the mechanisms of the *P. globosa* bloom in the Beibu Gulf through the establishment of a preliminary model for the development and progression of the bloom. My contribution to this chapter is the data analysis and interpretation.

4.1. Spatial-temporal dynamics of *P. globosa*

The spatial-temporal distribution of *P. globosa* throughout the bloom was explored using ASV_10 as a marker. This was done by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size followed by a centred log-ratio (clr) transformation of the compositional ASV table. After, I used interpolation, a procedure that predicts values for cells in raster from a limited number of sample points, and mapped the interpolated clr values in QGIS v. 3.8 (QGIS.org 2020). I also used linear mixed-effects models with sample location as a random effect to test for the effect of expedition, filtration size and water sampling depth on the relative abundance of ASV_10 using the “nlme” v. 3.1-142 (Pinheiro et al. 2020) and “multcomp” v. 1.4-12 (Hothorn et al. 2008) packages in R.

The relative abundance of ASV_10 varied with expedition ($P < 0.0001$) with the greatest relative abundance in the January expedition (Figure 26). This supports the January expedition as the peak of the bloom and the Feb-March and April expeditions as different stages of decay of the bloom. Interestingly, the relative abundance of ASV_10 also varied with the filtration size (P = 0.0065) as the relative abundance was greater in the 0.22-10 µm size fraction (Figure 26). One possible explanation for this is that more free-living *P. globosa* cells were captured in the smaller size fraction and most of the colonies were too large to be captured in the 10-200 µm size fraction. There was no relationship between the relative abundance of ASV_10 and the water sampling depth (P = 0.49). Mapping the relative abundance of ASV_10 also allowed for exploration of its spatial distribution, which showed that during the January and Feb-March expeditions the bloom appeared to be most intense in the northeast region of the Beibu Gulf near Weizhou Island (Figure 26b).

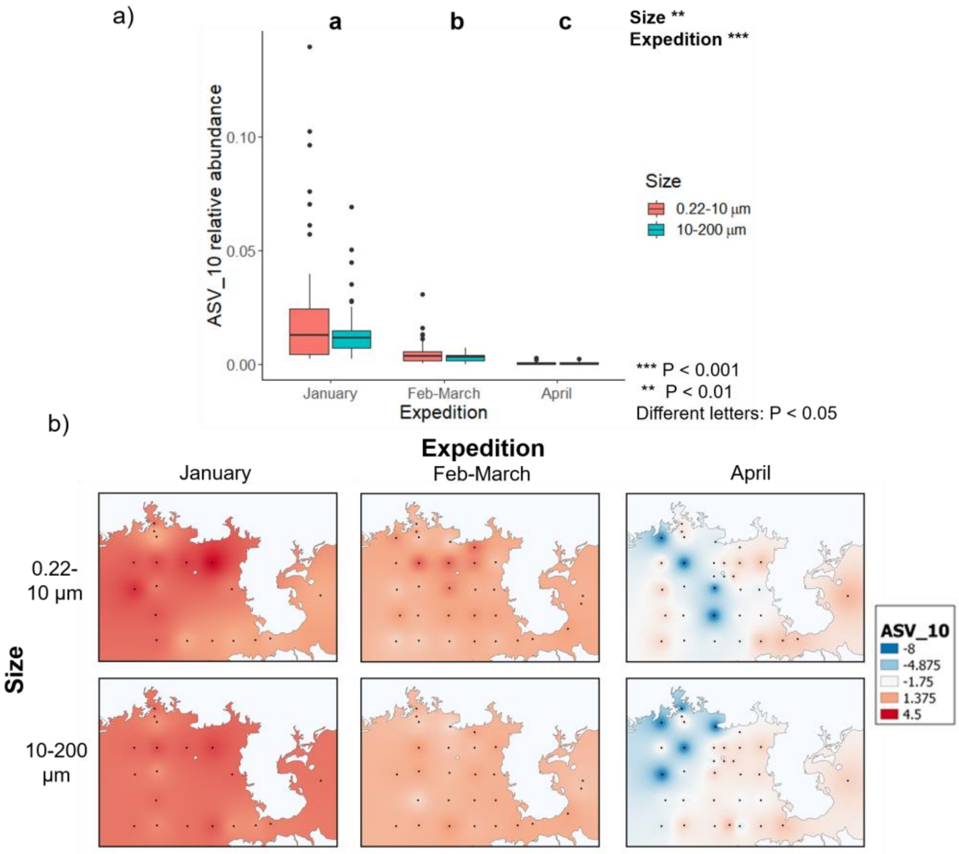


Figure 26 **a) Variation in the relative abundance of ASV_10 (*P. globosa*) with expedition and flitration size fraction. The significance of size and expedition were tested using a linear mixed effects model with size, expedition and water sampling depth as fixed effects and sampling location as a random effect. Pairwise comparisons for expedition were performed using Tukey’s method. b) The spatial distribution of ASV_10 at each expedition and filtration size fraction. The mapped values are interpolated from a centred log-ratio transformation in QGIS (QGIS.org 2020).**

4.2. Temporal dynamics of the alpha and beta diversity of microbial organisms

4.2.1. Alpha diversity

Alpha diversity is a measurement of the diversity within a single sample that can be estimated in many ways. The most naïve way of measuring alpha diversity is counting the observed richness i.e. the number of OTUs/ASVs in each sample. However, it can be challenging to identify every taxon in a sample and thus, techniques that consider the incompleteness of the sampling effort are often used. For example, the Chao1 estimator considers the number of singletons and doubletons in its estimate of alpha diversity. The evenness of the distribution of species in a sample is another important measure of diversity and several metrics have been developed that combine species richness and evenness into a single measure. For example, the calculation of the Simpson index corresponds to the odds that two individual microbes sampled at random will belong to the same OTU/ASV. Similarly, Shannon's diversity index is based on entropy and measures the uncertainty involved in predicting the species of an individual sample at random (Hugerth and Andersson 2017). It is common practice to include multiple measurements of different components of alpha diversity when analyzing microbial data.

The analysis of the alpha diversity of the field samples was performed for 1) all microbes, 2) bacteria, 3) archaea and 4) chloroplasts. For each group, four different metrics of alpha diversity were calculated for each of the field samples: the observed richness, Chao1, Shannon and InvSimpson (the inverse of the Simpson index). Statistical analysis was performed using linear mixed-effects models with sample location as a random effect to test for the effect of expedition, filtration size and water sampling depth on each diversity metric for each group using the "nlme" v. 3.1-142 (Pinheiro et al. 2020) R package. Pairwise comparisons were performed using Tukey's method with the "multcomp" v. 1.4-12 (Hothorn et al. 2008) package.

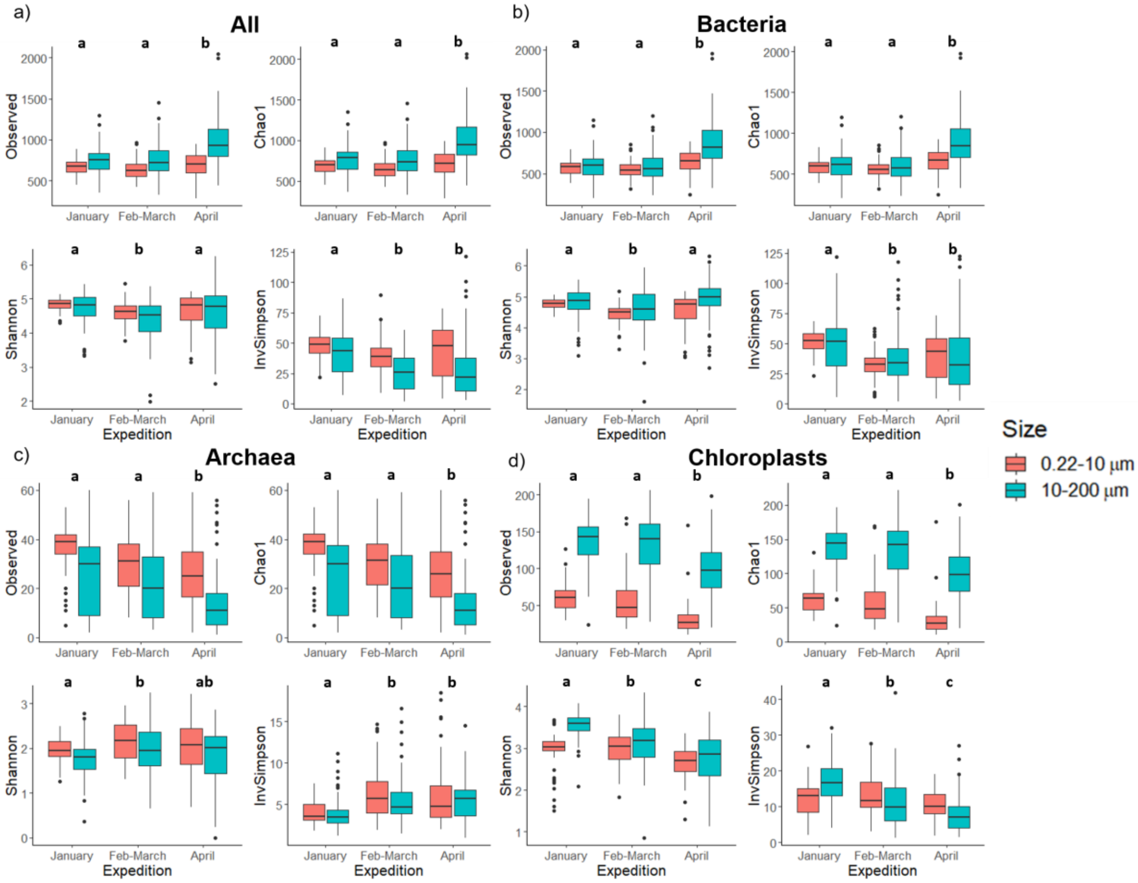


Figure 27 **Variation in alpha diversity of a) all microbes, b) bacteria, c) archaea and d) chloroplasts across different expeditions and flitration size fractions sampled from a *P. globosa* bloom in the Beibu Gulf. Different letters: P < 0.05.**

When all microbes were considered, the observed richness and the Chao1 index were greater in the samples from the large filtration size and the April expedition (Table 2, Figure 27a). However, the Shannon and InvSimpson indices, which also take community evenness into account, were greater in the small filtration size. The diversity was lowest in the Feb-March expedition using the Shannon index and lowest in the Feb-March and April expeditions using the InvSimpson index (Table 2, Figure 27a). This suggests that while the number of ASVs increased in the April expedition, the communities from this expedition did not become more even. The results were the same when only the bacteria where considered except the samples from the large filtration size were more diverse for all metrics other than InvSimpson, which had no relationship with filtration size (Table 2, Figure 27b). When considering only the archaea, the samples from the small filtration size were more diverse for all metrics except InvSimpson, which also had no relationship with filtration size. The alpha diversity of the archaea was lowest in the April expedition for the observed richness and the Chao1 index, but highest in the Feb-March and April expeditions for the indices that take evenness into account (Table 2, Figure 27c). This suggests that the number of archaea ASVs decreased in later expeditions, but the communities became more even. Finally, the alpha diversity of the chloroplasts was greater in the samples from the large filtration size for all indices except the InvSimpson, which had no relationship with filtration size. The diversity of the chloroplasts was lowest in the April expedition for all four metrics (Table 2, Figure 27d) as the number of chloroplast ASVs and the community evenness both decreased in later expeditions.

Table 2 **Results from linear-mixed effects models with sampling location as a random effect, water sampling depth, flitration size and expedition as fixed effects and different alpha diversity metrics as the dependent variable. Variables that are significant (P < 0.05) are bolded.**

Group	Metric	Variable	F-value	P-value
All	Observed	Depth	3.7	0.06
		Filtration size	**105.9**	**< 0.0001**
		Expedition	**27.0**	**< 0.0001**
	Chao1	Depth	4.0	0.05
		Filtration size	**105.0**	**< 0.0001**
		Expedition	**28.9**	**< 0.0001**
	Shannon	Depth	0.8	0.38
		Filtration size	**10.1**	**0.002**
		Expedition	**9.3**	**0.0001**
	InvSimpson	Depth	0.1	0.73
		Filtration size	**45.9**	**< 0.0001**
		Expedition	**11.0**	**< 0.0001**
Bacteria	Observed	**Depth**	**4.9**	**0.03**
		Filtration size	**44.3**	**< 0.0001**
		Expedition	**48.6**	**< 0.0001**
	Chao1	**Depth**	**5.2**	**0.02**
		Filtration size	**43.9**	**< 0.0001**
		Expedition	**50.7**	**< 0.0001**
	Shannon	Depth	0.0	0.91
		Filtration size	**12.5**	**0.0004**
		Expedition	**8.7**	**0.0002**
	InvSimpson	Depth	0.0	0.87
		Filtration size	0.4	0.51
		Expedition	**10.5**	**< 0.0001**
Archaea	Observed	**Depth**	**44.8**	**< 0.0001**

Group	Metric	Variable	F-value	P-value
		Filtration size	**111.4**	**< 0.0001**
		Expedition	**17.0**	**< 0.0001**
	Chao1	**Depth**	**44.6**	**< 0.0001**
		Filtration size	**112.8**	**< 0.0001**
		Expedition	**16.4**	**< 0.0001**
	Shannon	**Depth**	**6.9**	**0.009**
		Filtration size	**25.7**	**< 0.0001**
		Expedition	**8.8**	**0.0002**
	InvSimpson	Depth	0.2	0.69
		Filtration size	**5.0**	**0.03**
		Expedition	**19.2**	**< 0.0001**
Chloroplasts	Observed	Depth	1.2	0.27
		Filtration size	**682.6**	**< 0.0001**
		Expedition	**51.1**	**< 0.0001**
	Chao1	Depth	1.0	0.32
		Filtration size	**666.4**	**< 0.0001**
		Expedition	**50.4**	**< 0.0001**
	Shannon	**Depth**	**4.2**	**0.04**
		Filtration size	**35.7**	**< 0.0001**
		Expedition	**46.7**	**< 0.0001**
	InvSimpson	Depth	2.5	0.12
		Filtration size	0.0	0.98
		Expedition	**31.3**	**< 0.0001**

4.2.2. Beta diversity

Beta diversity measures the degree to which two samples differ and can be measured in many ways. Prior to measuring the beta diversity of microbial sequencing data, samples must be normalized to account for differences in the he read depth of each sample. While there are a number of normalization techniques available for data of this type (see McMurdie and Holmes 2014), normalization is typically accomplished by rarefying, where a minimum library size is selected, libraries that have fewer reads than the minimum sized are discarded and the remaining libraries are subsampled without replacement to the minimum library size (Hughes and Hellmann 2005). The subsequent calculation of beta diversity involves the use of a distance metric to measure the dissimilarity between samples. The most widely known distance metric is the Euclidian distance, which does not perform well in datasets with many zeroes i.e. microbial community composition data. Alternative metrics such as the Jensen-Shannon or Bray-

Curtis dissimilarity are more appropriate for zero-inflated datasets and are thus more commonly used in microbial ecology (Hugerth and Andersson 2017). Phylogenetic information can also be incorporated into distance metrics i.e. the UniFrac distance, which may be more biologically meaningful because it accounts for evolutionary distances between species. However, the calculation of these metrics is more complex as it relies on the construction of an accurate phylogenetic tree and the correct placement of OTUs/ASVs on the tree (Hugerth and Andersson 2017). The Bray-Curtis dissimilarity metric was selected for the analyses here because of its simplicity and common use in microbial ecology.

Beta diversity analysis results in the creation of a highly dimensional matrix of pairwise distance measures between each of the samples, which must be condensed into two- or three-dimensional space to be visualized. One of the most common methods to achieve this is Principal Component Analysis (PCA) (Ringnér 2008). PCA is a dimensionality reduction technique that increases data interpretability by creating new uncorrelated variables (principle components) that successively maximize variance. The first few components often explain a large amount of the variance, allowing a visual inspection of the distance between samples in two- or three-dimensional space. However, because PCA uses Euclidian distance, it is seldom appropriate for microbial data. Instead, Principle Coordinate Analysis (PCoA), which can be used with any of the dissimilarity metrics, is frequently used. Samples plotted on a PCA/PCoA plot can be coloured according to metadata values to explore how samples from, e.g. different time points, differ from each other. Finally, non-parametric statistical tests can used to test hypotheses such as whether *a priori* groupings of samples (e.g. different treatments) correspond to statistically different microbial communities. Common tests for this are PERMANOVA and ANOSIM, which assesses whether ranks of distances of objects within classes that are defined *a priori* are smaller than those between classes (Hugerth and Andersson 2017).

The analysis of the beta diversity of the field samples was performed for 1) all microbes, 2) bacteria, 3) archaea and 4) chloroplasts. For each group, I rarefied the ASV counts for the abundant ASVs to a different number of reads depending on the number of reads in each sample (All: 39,668 reads, Bacteria: 10,000 reads, Archaea: 100 reads, Chloroplasts: 1,000 reads) and ordinated the Bray-Curtis dissimilarity matrix using a PCoA with the “phyloseq” v. 1.28.0 (McMurdie and Holmes 2013) R package. I tested for differences in the community composition of samples from different expeditions, filtration size fractions and water sampling depths using the ANOSIM statistic from the “vegan” v. 2.5-6 (Oksanen et al. 2019) R package.

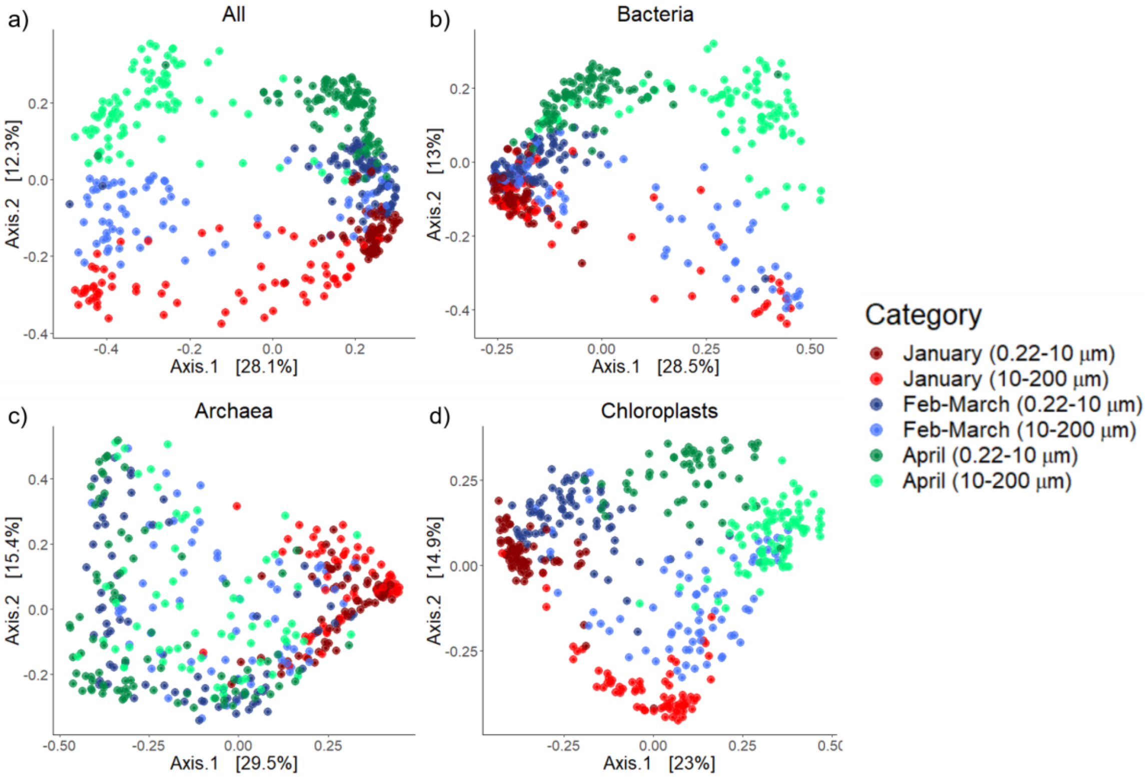


Figure 28 **Variation in the community composition of a) all microbes, b) bacteria, c) archaea and d) chloroplasts between different expedition time points and filtration size fractions visualized using PCoA from the Bray-Curtis dissimilarity matrices.**

When all microbes were included, the community composition differed significantly between samples from different expeditions (ANOSIM statistic R = 0.23, P = 0.001) and filtration sizes (R = 0.26, P = 0.001). The first principal component, which explained 27% of the variance, separated the samples by size and the second principal component, which explained 12% of the variance, separated the samples by expedition as there was a progressive change in community composition with time (Figure 28a). Similarly, for bacteria, community composition differed significantly between samples from different filtration sizes (R = 0.17, P = 0.001) and different expeditions (R = 0.24, P = 0.001) with a progressive change in bacteria community composition from January to April (Figure 28b). For the archaea, the community composition also differed significantly between samples from different filtration sizes (R = 0.075, P = 0.001) and different expeditions (R = 0.27, P = 0.001). However, there was no progressive change across all three expeditions as the Feb-March and April expeditions did not cluster separately in the PCoA plot (Figure 28c). Finally, the community composition of chloroplasts differed significantly between samples from different filtration sizes (R = 0.21, P = 0.001) and different expeditions (R = 0.45, P = 0.001) with a progressive change in community composition from January to April (Figure 28d). There was no variation with community composition and water sampling depth for any of the groups (P > 0.05 in all cases) Overall, the beta diversity analysis provided evidence of change in the microbial community composition in the Beibu Gulf over time during the *P. globosa* bloom and between different filtration sizes.

4.3. Spatial-temporal dynamics of other microbial organisms

4.3.1. Bacteria

Consistent with the observed richness of the bacteria ASVs, the overall relative abundance of bacteria was greater in the 0.22-10 µm filtration size (linear mixed effects model with sampling location as a random effect P < 0.0001) and increased from the January to the April expedition (P < 0.0001; Figure 29). The observed increase in the relative abundance of bacteria at the end of the bloom is consistent with the idea that there is an increase in heterotrophic bacteria during the decay phase of the bloom in response to elevated levels of algal-derived organic material (Teeling et al. 2012).

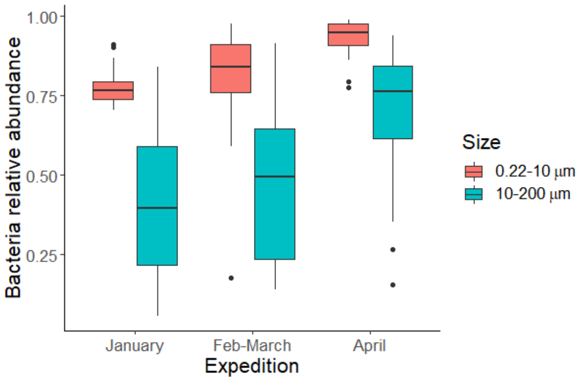


Figure 29 Relative abundance of bacteria across different expeditions and filtration size fractions during a *P. globosa* bloom in the Beibu Gulf.

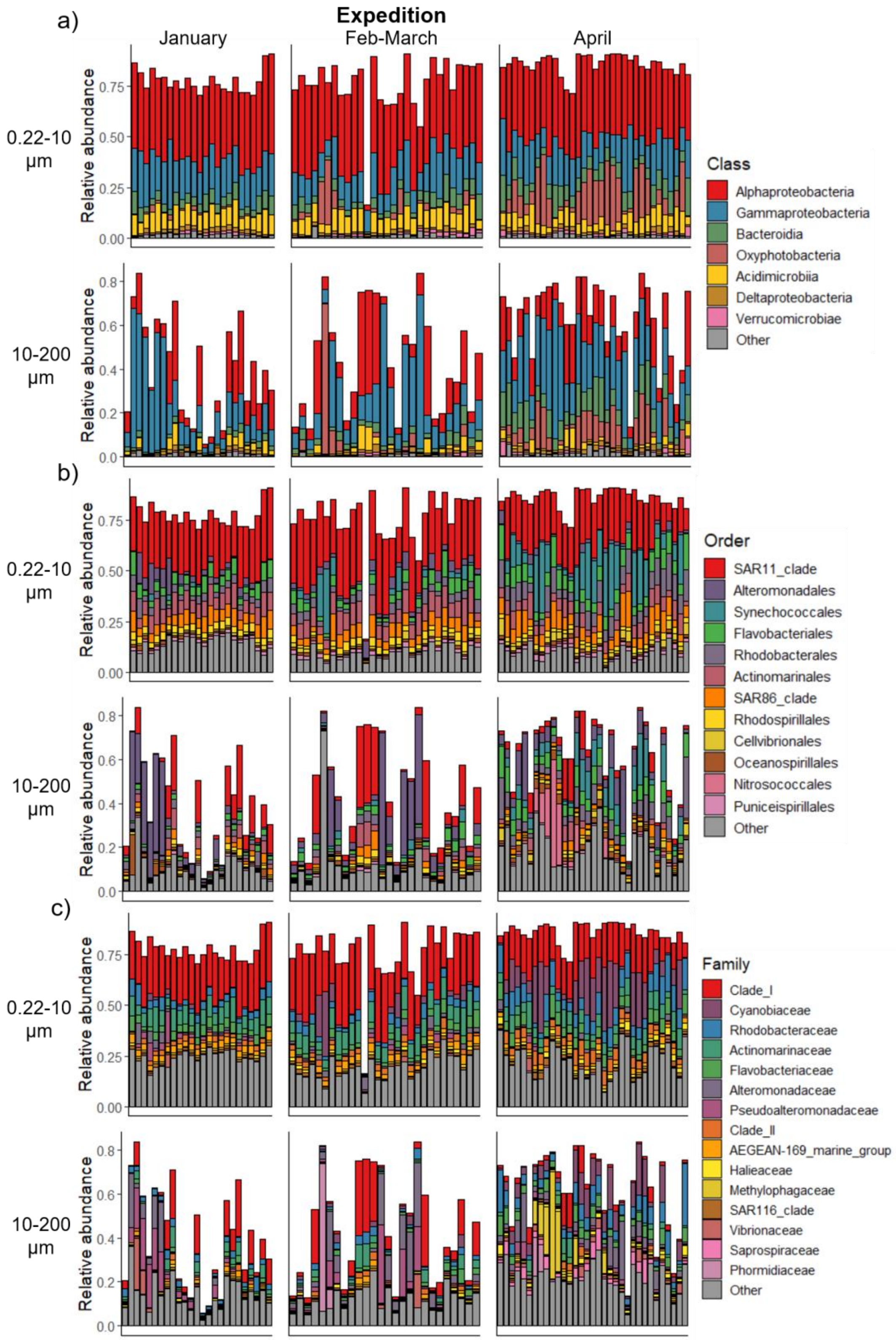


Figure 30 **Composition of the major a) classes, b) orders and c) families of bacteria in individual surface samples grouped by filtration size and expeditions from a *P. globosa* bloom in the Beibu Gulf.**

In agreement with the results from the beta diversity analysis, there was distinct bacterial communities from surface samples across different filtration sizes and expeditions at the class, order and family levels (Figure 30). I decided to further explore the spatial-temporal dynamics of specific groups of bacteria to better understand their role in the bloom. First, I agglomerated samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and performed a centred log-ratio (clr) transformation for each ASV separately from the compositional ASV table. This was followed by interpolation of the clr-transformed values and mapping of the interpolated clr values in R using the “tmap” v. 3.0 (Tennekes 2018) package. Statistical analysis of the temporal changes in different groups was performed using linear mixed-effects models from the R package “nlme” v. 3.1-142 (Pinheiro et al. 2020). In the models, sample location was specified as a random effect, expedition, filtration size and water sampling depth as fixed effects and the relative abundance of the group as the dependent variable.

Alphaproteobacteria

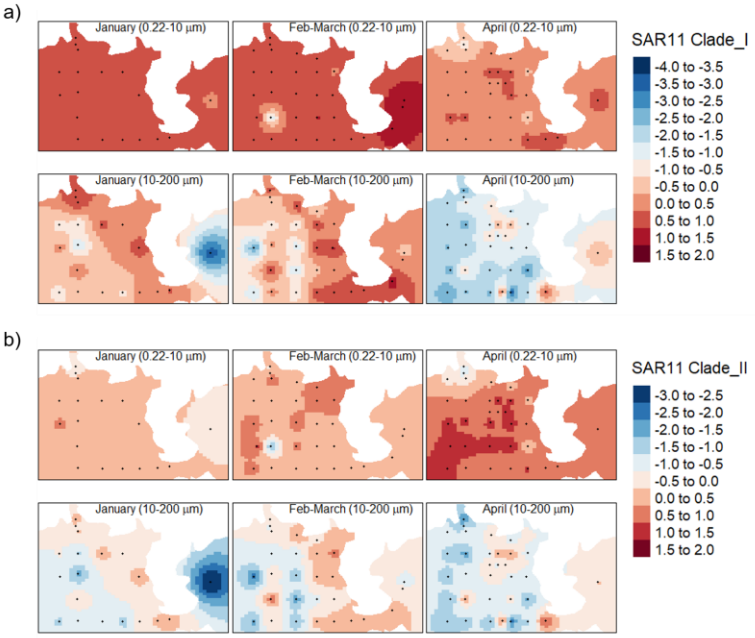


Figure 31 **Spatial-temporal distribution of two clades of SAR11 during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and flitration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.**

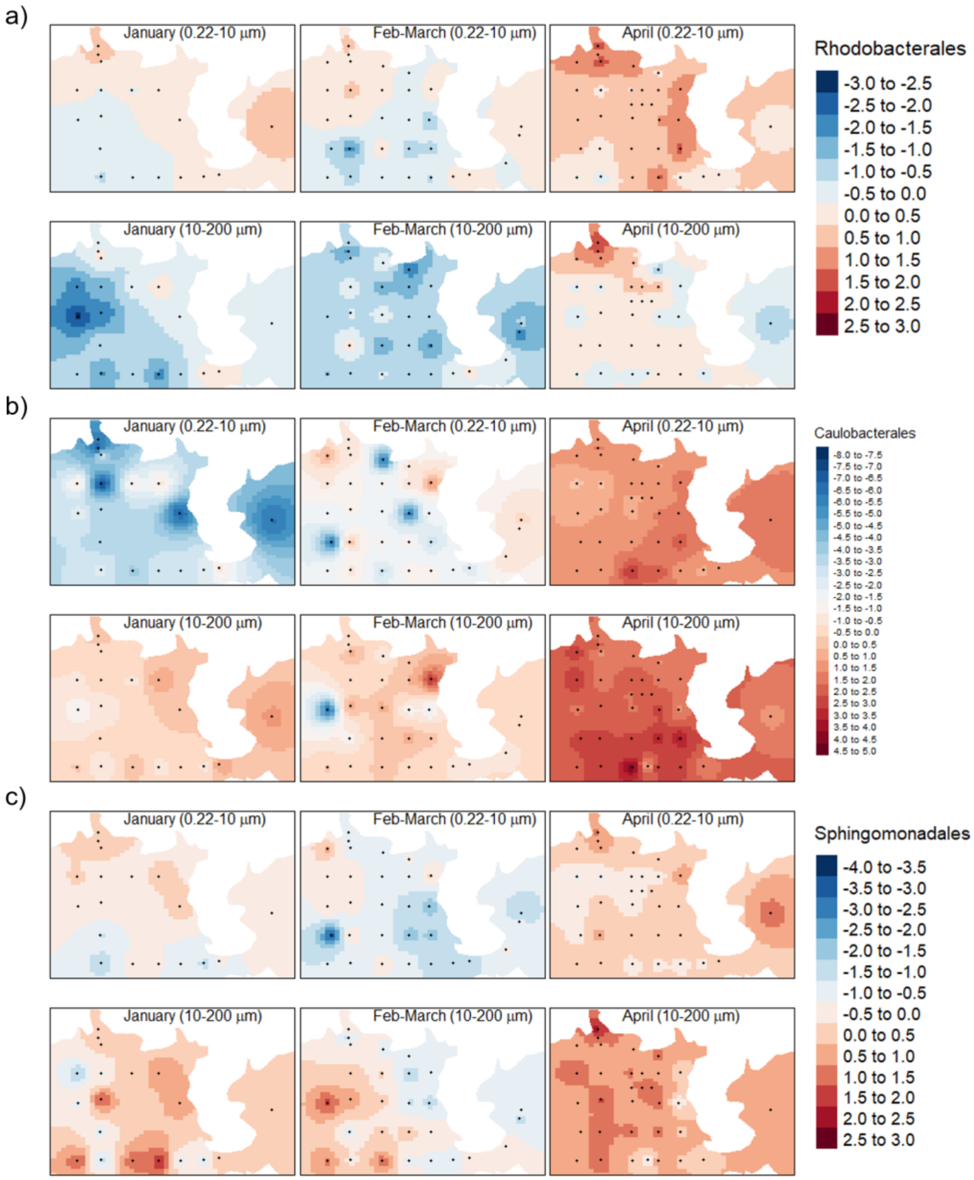


Figure 32 Spatial-temporal distribution of three orders of Alphaproteobacteria during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.

Several groups of microbes from the Alphaproteobacteria displayed interesting spatial-temporal patterns throughout the *P. globosa* bloom in the Beibu Gulf. The most abundant order of bacteria in this study, the chemoheterotrophic SAR11_clade, had contrasting temporal patterns for its different clades. While members of the more abundant Clade_I decreased in relative abundance from January to April (P < 0.0001;

Figure 30c and Figure 31a), Clade_II showed the opposite pattern as the group increased in relative abundance from January to April (P < 0.0001; Figure 30c and Figure 31b). Thus while members of Clade_I may be outcompeted by other groups that are better able to respond to the increase in available substrates at the end of the bloom, some members of Clade_II may be able to increase their growth rates in response to the increase in available substrates. The order Rhodobacterales, which are regarded as important consumers of carbon (Buchan et al. 2014) and amino acids (Alonso-Saez and Gasol 2007) during and after phytoplankton blooms, also increased in relative abundance at the end of the bloom (P < 0.0001; Figure 30b and Figure 32a). Other groups from the Alphaproteobacteria with interesting spatial-temporal patterns included the chemoorganotrophic Caulobacterales and the Sphingomonadales, which degrade a wide range of hydrocarbons (Kertesz et al. 2019). Both groups had the greatest relative abundance during the April expedition (P < 0.0001 for both; Figure 32b-c).

Gammaproteobacteria

A handful of groups of Gammaproteobacteria also exhibited compelling spatial-temporal patterns throughout the *P. globosa* bloom in the Beibu Gulf. The order Nitrosococcales, which derive their energy from the oxidation of ammonium, increased in relative abundance during the April expedition (P < 0.0001; Figure 30b and Figure 33a). Similarly, the Salinisphaerales, a group with little known about its source of energy, distinctly increased in relative abundance at the end of the bloom (P = 0.0005; Figure 33b). In contrast, the Steroidobacterales were one of the few groups of bacteria that had the greatest relative abundance during the peak of the bloom in January (P < 0.0001; Figure 33c). Most members of this order were from the genus *Woeseia*, which belongs to a family that contains a broad range of energy-yielding metabolisms in its genome (Mußmann et al. 2017) that may have allowed them to thrive during the peak of the bloom. The family Colwelliaceae, which has previously been associated with the decay of *Phaeocystis* blooms (Delmont et al. 2014) displayed contrasting temporal patterns between its two most abundant genera (Figure 34). While ASVs from the genus *Colwellia* generally had the greatest relative abundance in the January expedition, ASVs from the genus *Thalassotalea* peaked in relative abundance in the April abundance, suggesting different roles for the two genera in the decay of the bloom. The family Oleiphilaceae, which is represented by one species that feeds almost exclusively on hydrocarbons (Golyshin et al. 2002) exhibited marked variation in temporal patterns at the ASV-level (Figure 35).

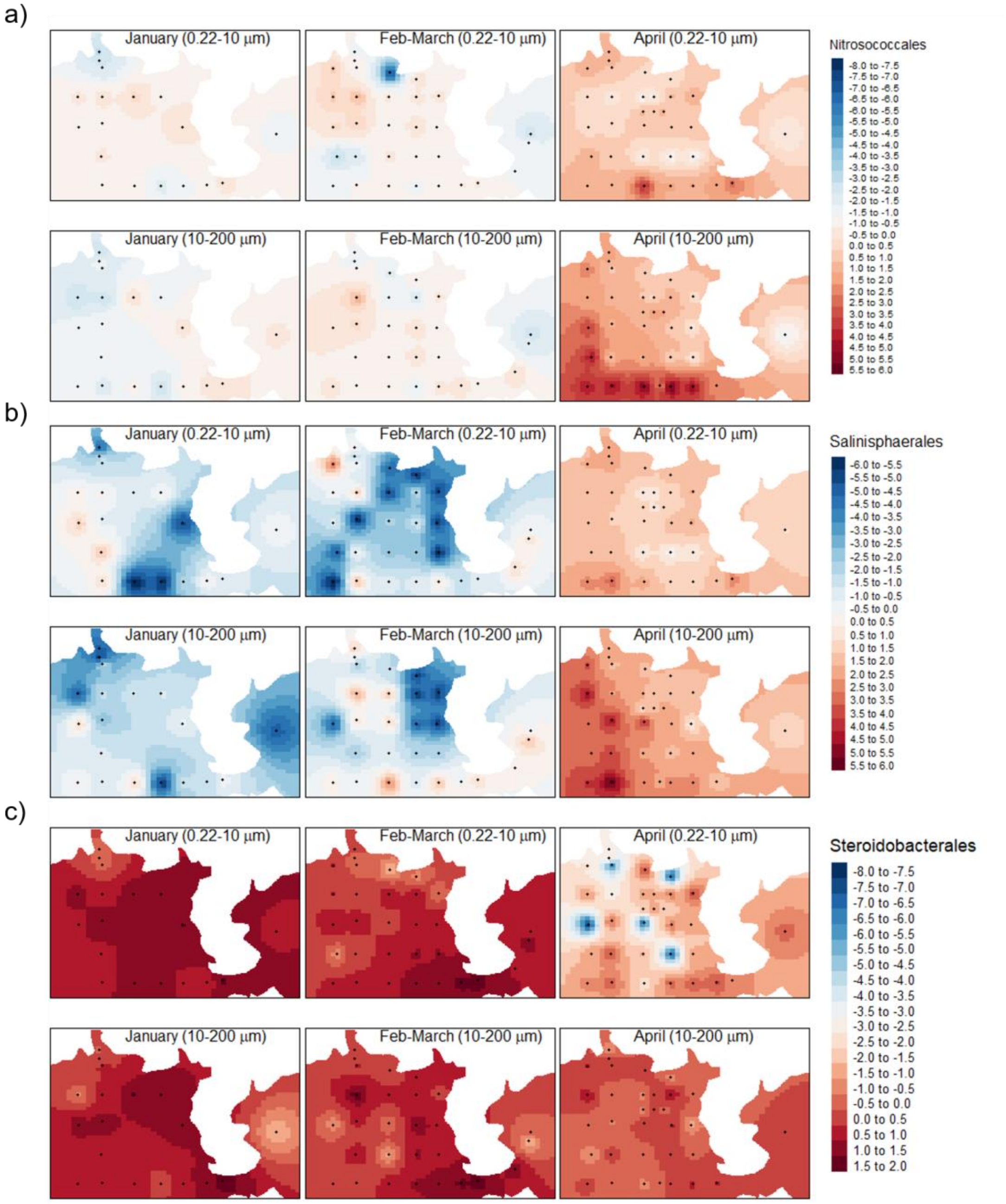


Figure 33 Spatial-temporal distribution of three orders of Gammaproteobacteria during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.

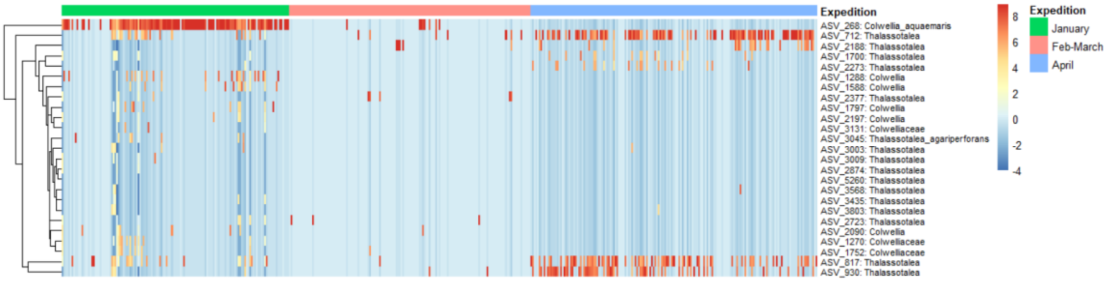


Figure 34 **Heatmap of abundant ASVs from the family Colwelliaceae. Rows represent ASVs and columns represent samples, which are ordered by expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package "microbiomeutilities" v. 0.99.0 (Shetty and Lahti 2019) with transformation set to "clr".**

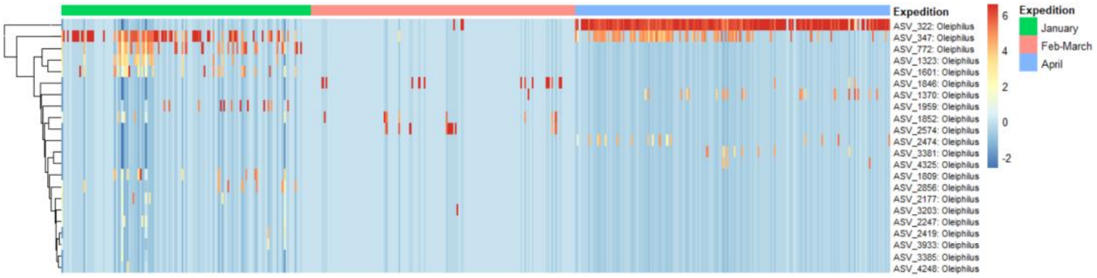


Figure 35 **Heatmap of abundant ASVs from the family Oleiphilaceae. Rows represent ASVs and columns represent samples, which are ordered by expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package "microbiomeutilities" v. 0.99.0 (Shetty and Lahti 2019) with transformation set to "clr".**

Bacteroidia

Consistent with their specialization in the degradation of high molecular weight organic matter, the class Bacteroidia increased in relative abundance during the April expedition, particularly in the larger 10-200 µm filtration size fraction ($P < 0.0001$; Figure 30a). The Flavobacteriales, which were the main order of Bacteroidia observed in the Beibu Gulf, increased in relative abundance at the end of the bloom in April ($P < 0.0001$; Figure 30b and Figure 36a).Other less abundant orders such as the Chitinophagales and the Cytophagales displayed similar patterns ($P < 0.0001$ for both; Figure 36b-c). This supports the role of microbes from these orders in the degradation of organic matter from the bloom.

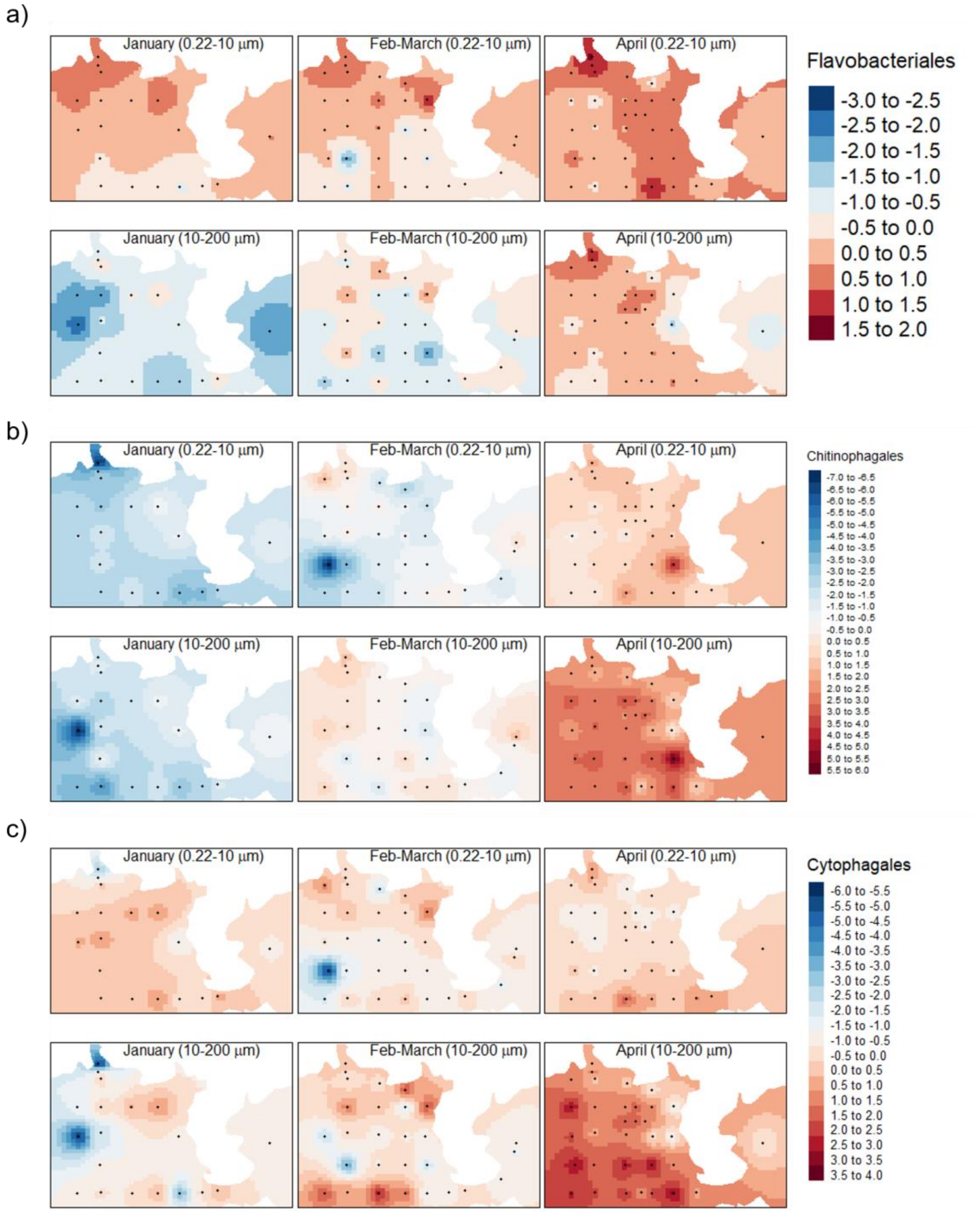


Figure 36 **Spatial-temporal distribution of three orders of Bacteroidia during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.**

Oxyphotobacteria

The dominant order of cyanobacteria in the Beibu Gulf, the Synechococcales, increased in relative abundance from the January to April expedition (P < 0.0001; Figure 30b and Figure 37). Given that only N_2, CO_2, water and mineral elements are needed by cyanobacteria for growth in the light (Mur et al. 1999), this temporal pattern may have occurred because the cyanobacteria were under less competition from *Phaeocystis* and other phytoplankton at the end of the bloom.

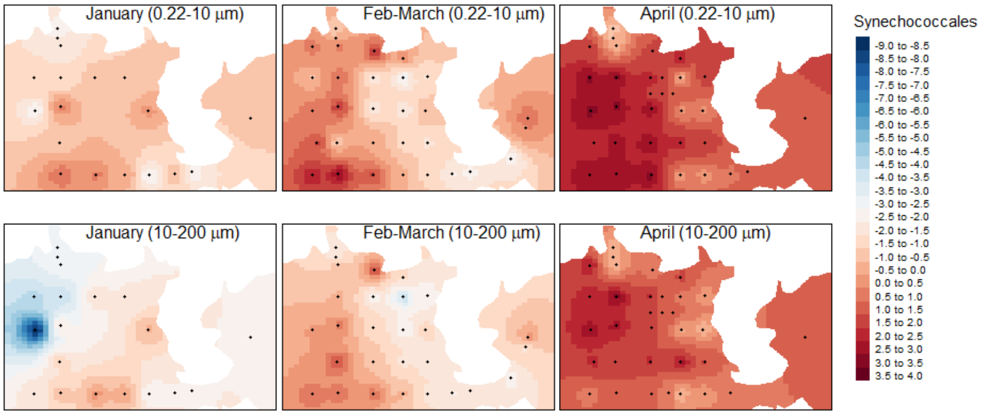


Figure 37 **Spatial-temporal distribution of the Synechococcales during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.**

Verrucomicrobiae

The Verrucomicrobiae, which are predominantly heterotrophic with carbohydrate-degrading metabolisms and genomes enriched in glycoside hydrolases (Martinez-Garcia et al. 2012), had the greatest relative abundance during the April expedition (P < 0.0001; Figure 38). This is consistent with the idea that the Verrucomicrobiae play a role in the degradation of carbohydrates from *P. globosa* and other phytoplankton at the end of the bloom.

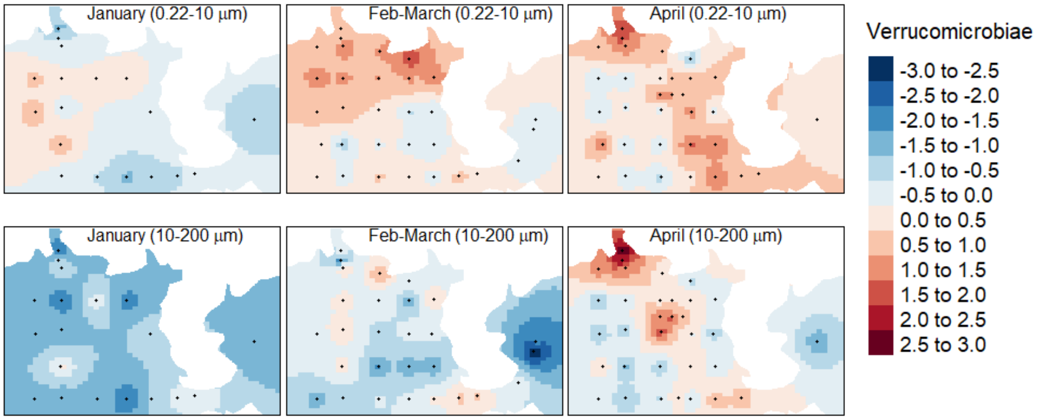


Figure 38 Spatial-temporal distribution of the Verrucomicrobiae during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.

Random forest model analysis

The random forest classifier is an ensemble learning method that is frequently used both for classification and feature importance. Decisions trees, where the leaves of the tree represent class labels and the branches represent conjunctions of features that lead to those class labels, are used by random forest models for classification. More specifically, a random forest classifier uses many decisions trees at training time and classifies cases using the mode of the classes from all the trees in the forest. The training set for each tree uses a subset of cases that are drawn by sampling with replacement and a subset of features is used to split each node in each tree. One of the advantages of using an ensemble of trees is that it corrects for overfitting, which is common with decision trees. In addition to the classification of cases, random forest models can also be used to measure feature importance. This can be accomplished using permutational methods, which randomly permute the values of each feature and compare the number of cases that are correctly classified in the permuted data vs. the untouched data, or the Gini index, which uses the number of nodes that include the feature from all the trees to measure feature importance.

To identify bacteria ASVs with distinct temporal patterns, a random forest classifier was used to measure feature importance. The random forest model was trained using the RandomForestClassifier from the “scikit-learn” v. 0.23.0 (Pedregosa et al. 2011) module in Python v. 3.7.4 with expedition as the model label and the top 500 most abundant ASVs (after excluding the chloroplast ASVs) as the features. The model was trained using 1,000 trees and the feature importance was measured using the default feature_importances_, which uses the Gini index. The input ASVs were ranked according to their feature importance and a heatmap was produced with the top fifty most important ASVs for classifying expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package “microbiomeutilities” v. 0.99.0 (Shetty and Lahti 2019) with transformation set to “clr”.

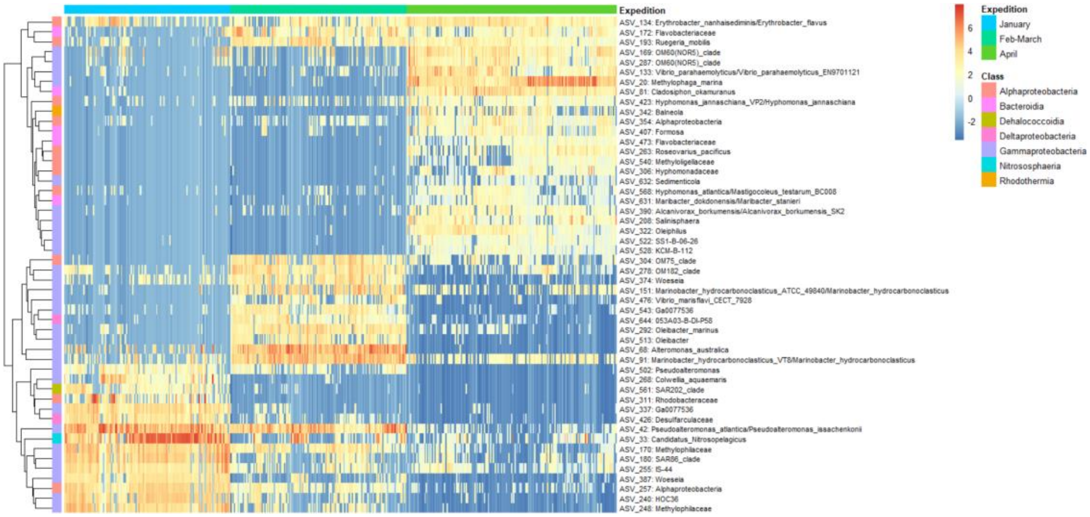


Figure 39 Heatmap of the top fifty most important non-chloroplast ASVs for classifying expedition, identified using a random forest classifier. Rows represent ASVs and columns represent samples, which are ordered by expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package “microbiomeutilities” v. 0.99.0 (Shetty and Lahti 2019) with transformation set to “clr”.

Random forest analysis identified a marked amount of temporal variation at the ASV-level. The random forest classifier picked out ASVs with three distinct temporal patterns: 1) ASVs with the greatest relative abundance at the peak of the bloom (January), 2) ASVs with the greatest relative abundance at the early decay stage of the bloom (Feb-March) and 3) ASVs with the greatest relative abundant at the late decay stage of the bloom (April) (Figure 39). In many cases, closely related ASVs exhibited distinct temporal patterns, supporting a previously underappreciated amount of variation in niches within taxonomic groups. For example, ASV_374, which belongs to the genus *Woeseia*, peaked in relative abundance during the Feb-March expedition (P < 0.0001;

Figure 40a). In contrast, ASV_387, which belongs to the same genus, had the greatest relative abundance at the peak of the bloom in January (P < 0.0001; Figure 40b). Despite being assigned to the same genus, these two ASVs displayed distinct temporal patterns, supporting the use of different ecological niches during the bloom.

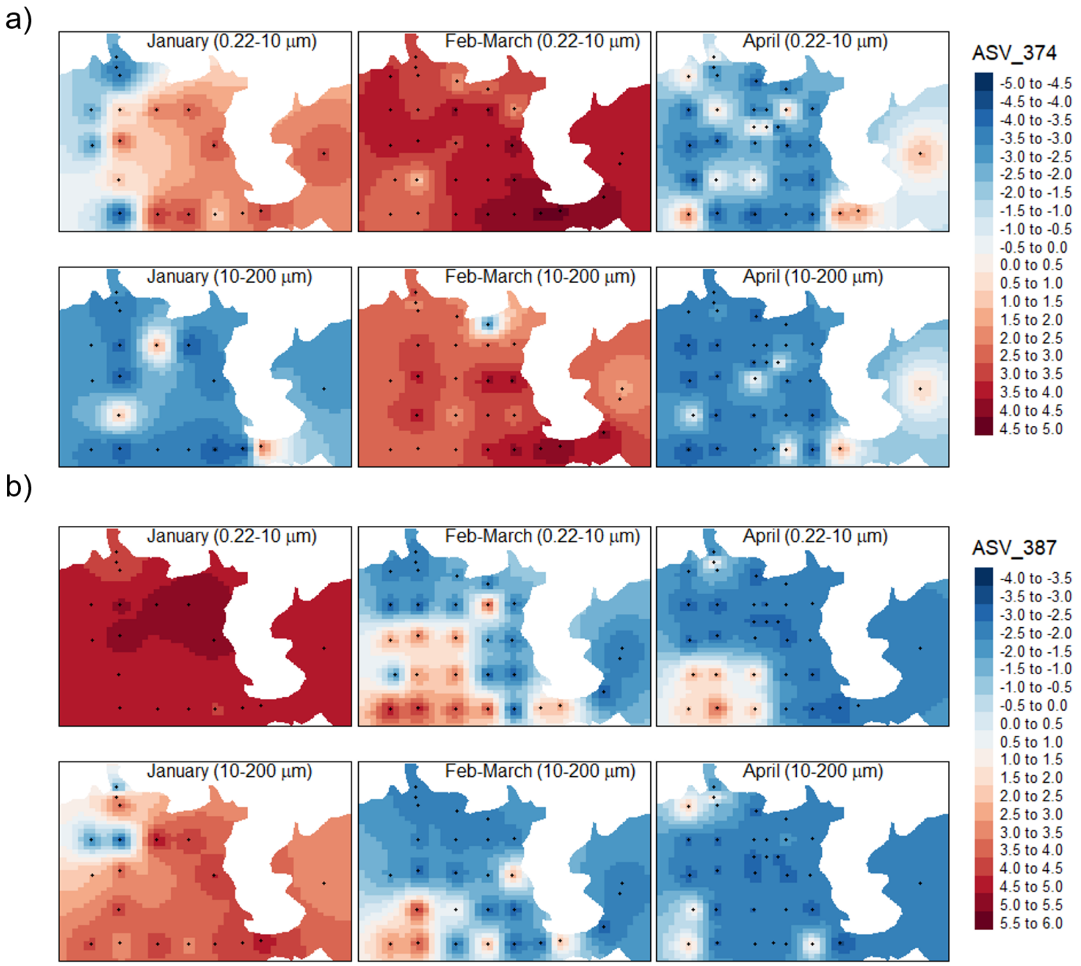


Figure 40 **Spatial-temporal distribution of two ASVs from the genus *Woeseia* with contrasting temporal patterns during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.**

4.3.2. Archaea

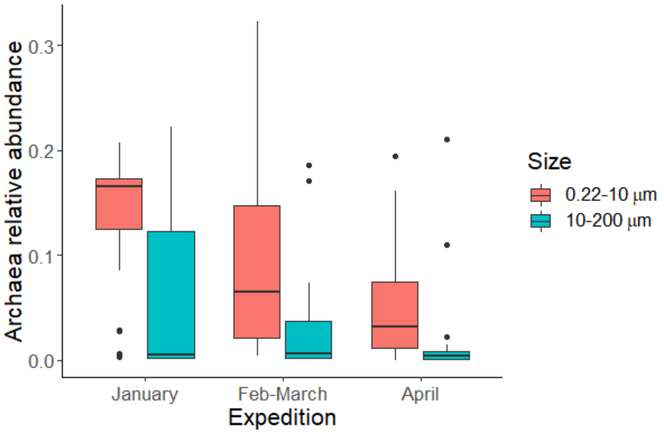


Figure 41 **Relative abundance of archaea across different expeditions and filtration size fractions during a *P. globosa* bloom in the Beibu Gulf.**

Consistent with the observed richness of the archaea ASVs, the overall relative abundance of archaea was greater in the 0.22-10 µm (linear mixed effects model with sampling location as a random effect P < 0.0001) filtration size and decreased from the January to the April expedition (P < 0.0001; Figure 41). The composition of the archaea communities from the surface samples changed across the different filtration sizes and expeditions at the class level (Figure 42). Using the same methods as for the bacteria, I further explored the spatial-temporal dynamics of specific groups of archaea to better understand their role in the bloom.

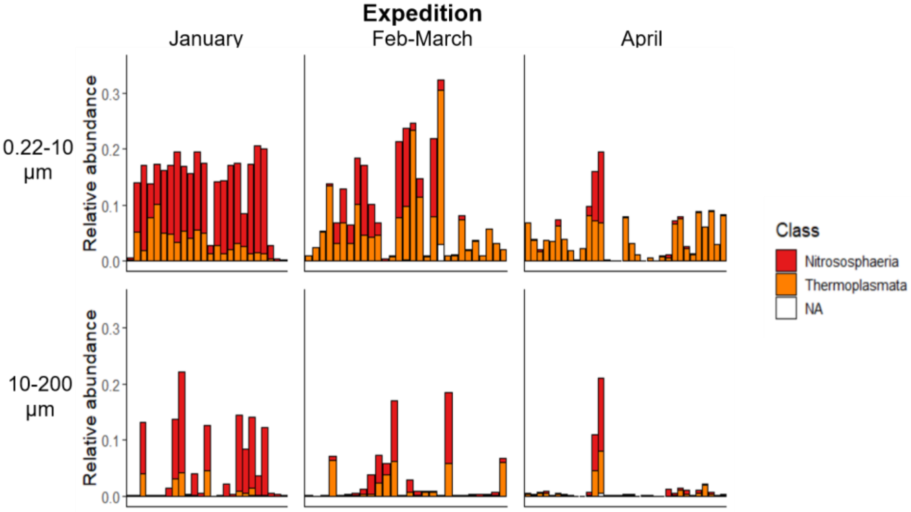


Figure 42 **Composition of the classes of archaea in individual surface samples grouped by filtration size and expeditions from a *P. globosa* bloom in the Beibu Gulf.**

Nitrososphaeria

The Nitrososphaeria, a class from the phylum Thaumarchaeota, decreased in relative abundance from the January to the April expedition (P < 0.0001; Figure 42 and Figure 43).

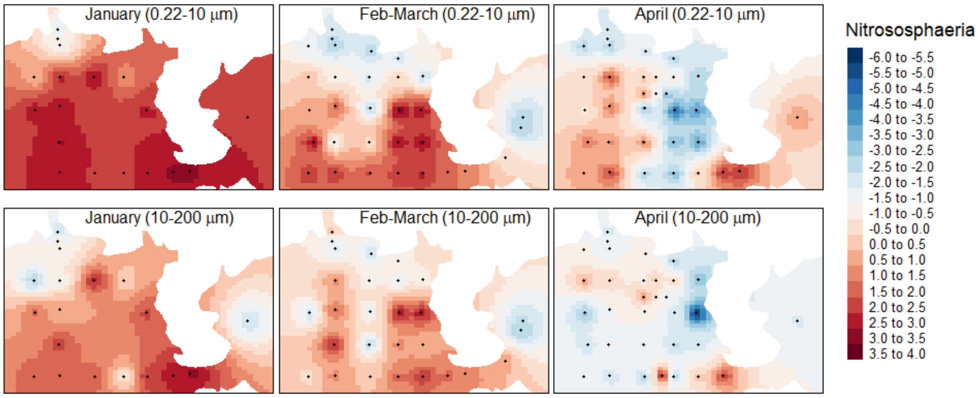


Figure 43 Spatial-temporal distribution of the Nitrososphaeria during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.

Thermoplasmata

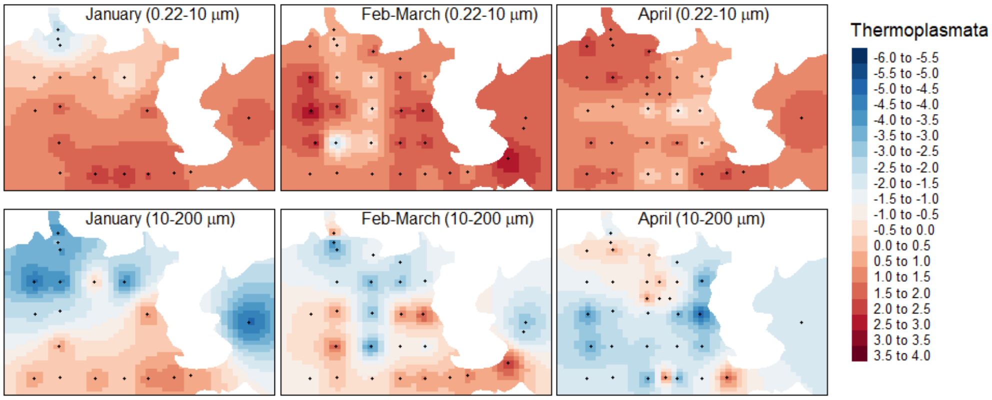


Figure 44 Spatial-temporal distribution of the Thermoplasmata during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.

The Thermoplasmata, which were dominated by the order Marine_Group_II (MGII), peaked in relative abundance during the Feb-March and April expeditions (P = 0.02; Figure 42 and Figure 44). These results are consistent with the speculation of the MGII as facultative colonizers of particles (Santoro et al. 2019) and suggests that they were involved in the degradation of *Phaeocystis* colonies and other phytoplankton. Just as was observed for several groups of bacteria, the temporal dynamics of members of the MGII showed remarkable variation at the ASV level (Figure 45). This suggests that members of the MGII have intra-group variation in their metabolic roles and interactions with other species during the bloom.

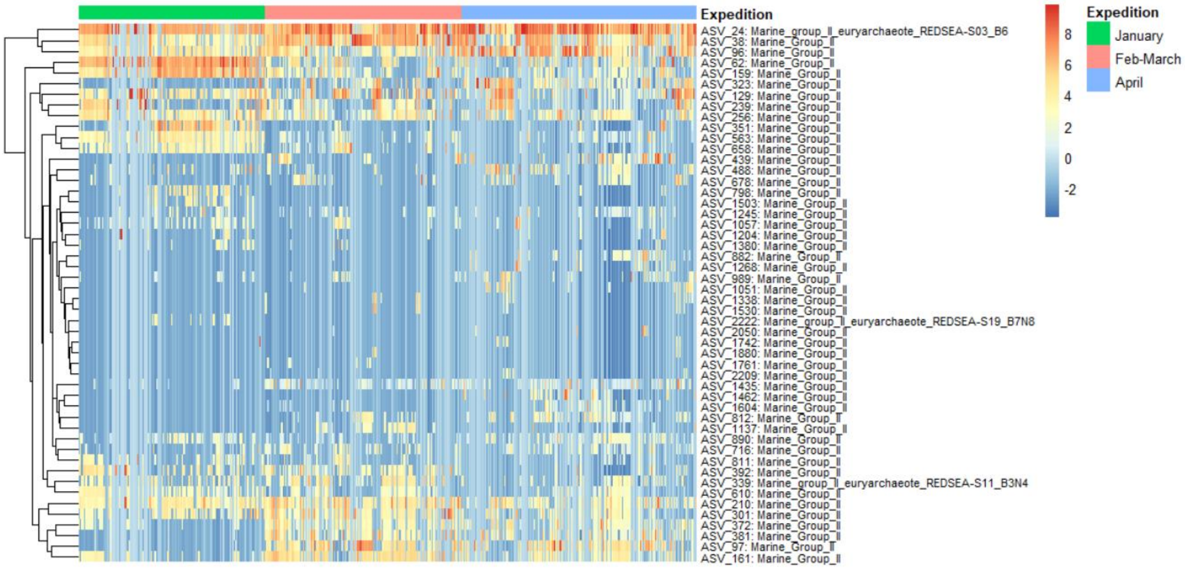


Figure 45 **Heatmap of the top fifty abundant ASVs from the order Marine_Group_II. Rows represent ASVs and columns represent samples, which are ordered by expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package "microbiomeutilities" v. 0.99.0 (Shetty and Lahti 2019) with transformation set to "clr".**

4.3.3. Chloroplasts

Consistent with the observed richness of the chloroplast ASVs, the overall relative abundance of chloroplasts was greater in the 10-200 µm filtration size (linear mixed effects model with sampling location as a random effect P < 0.0001) and decreased from the January to the April expedition (P = 0.0002; Figure 46). Also consistent with the beta diversity analysis, there was distinct chloroplast communities from surface samples across different filtration sizes and expeditions at the at the class and order levels (Figure 47). Using the same methods as for the bacteria and the archaea, I further explored the spatial-temporal dynamics of specific groups of chloroplasts to better understand their role in the bloom.

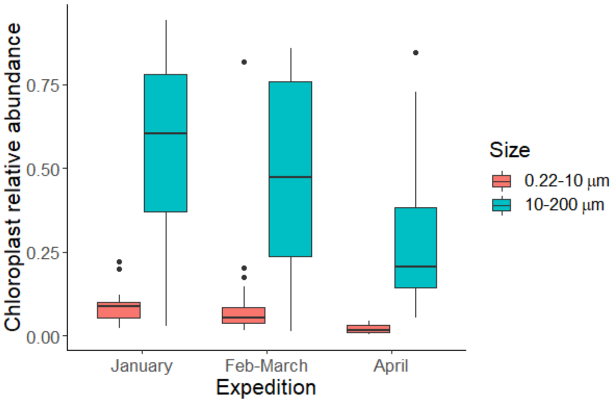


Figure 46 Relative abundance of chloroplasts across different expeditions and filtration size fractions during a *P. globosa* bloom in the Beibu Gulf.

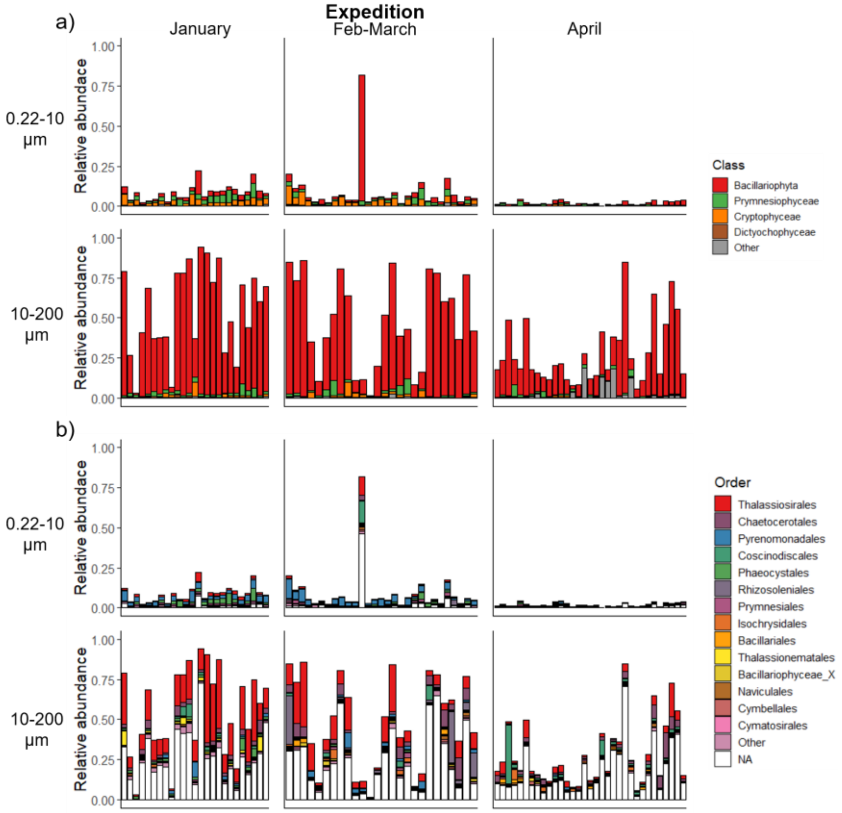


Figure 47 Composition of the major a) classes and b) orders of chloroplasts in individual surface samples grouped by filtration size and expeditions from a *P. globosa* bloom in the Beibu Gulf.

The three most abundant classes of chloroplasts, the Bacillariophyta (P = 0.005), Prymnesiophyceae (P < 0.0001) and Cryptophyceae (P < 0.0001), all decreased in relative abundance from the January to the April expedition (Figure 47a and Figure 48).

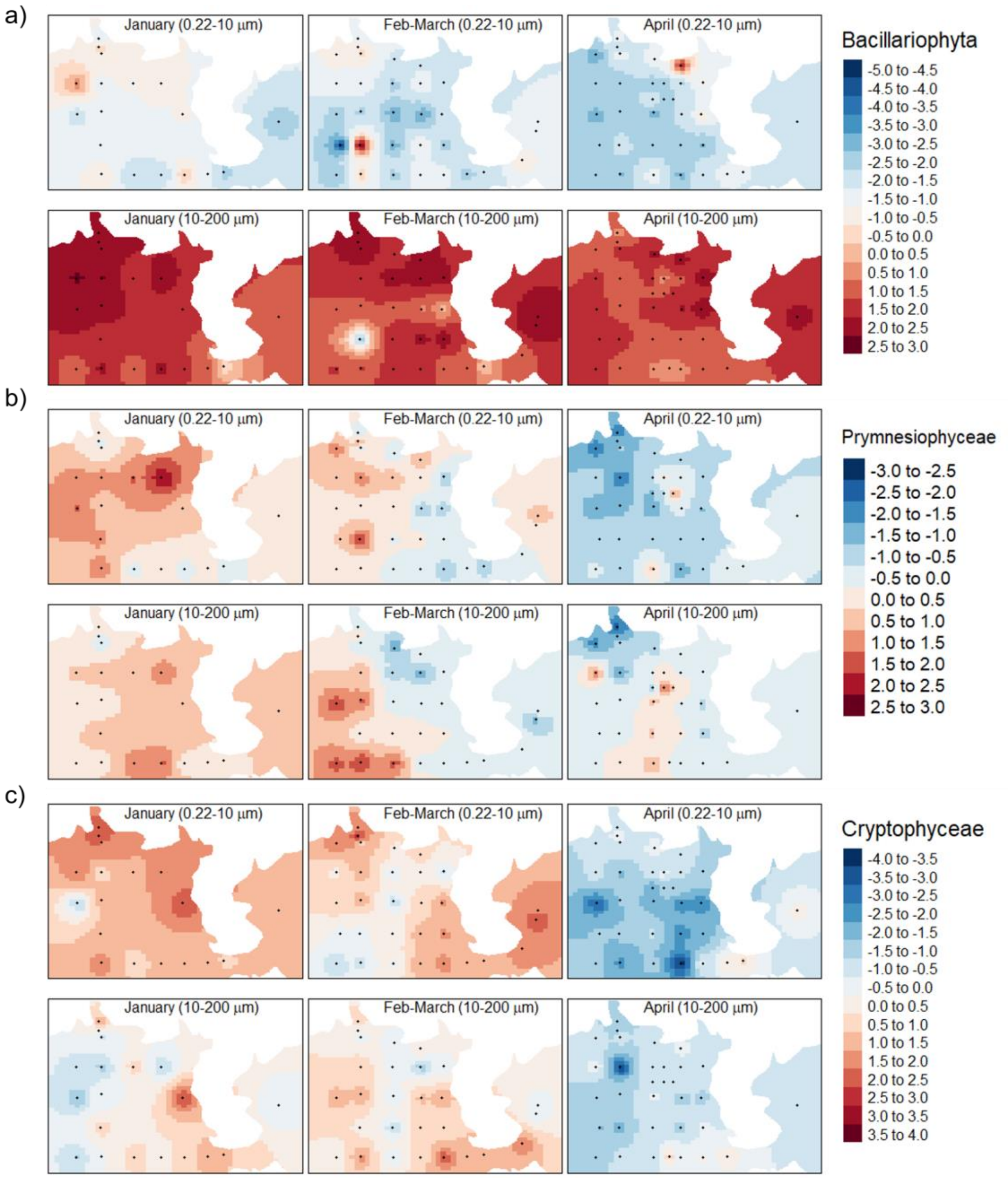


Figure 48 Spatial-temporal distribution of three classes of chloroplasts during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month and filtration size fraction. Mapping was performed by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.

Random forest model analysis

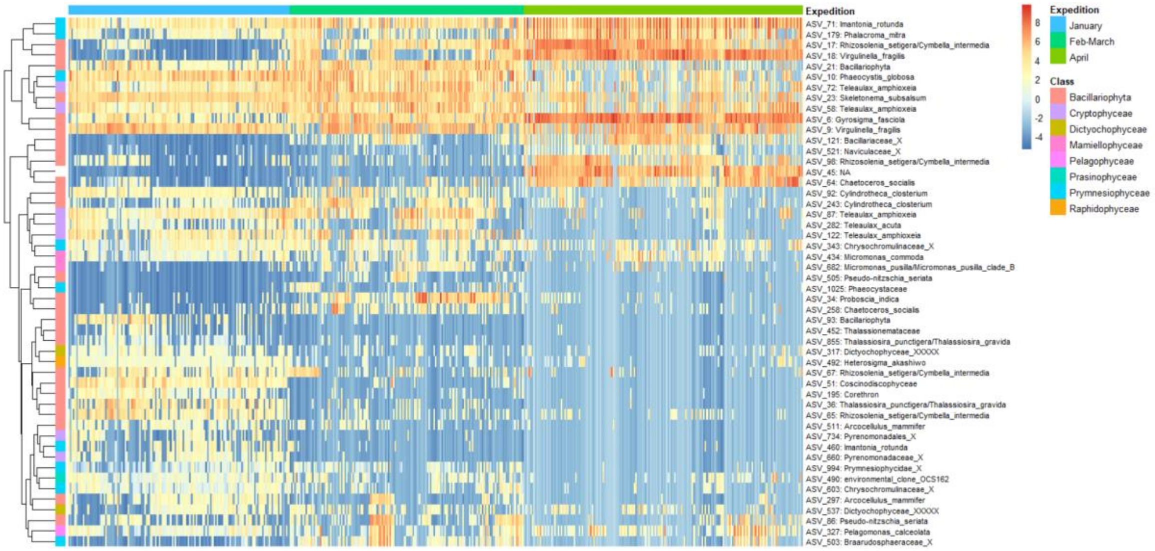


Figure 49 **Heatmap of the top fifty most important chloroplast ASVs for classifying expedition, identified using a random forest classifier. Rows represent ASVs and columns represent samples, which are ordered by expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package "microbiomeutilities" v. 0.99.0 (Shetty and Lahti 2019) with transformation set to "clr".**

To identify chloroplast ASVs with distinct temporal patterns, a random forest classifier was used to measure feature importance. The random forest classifier was trained using the same methods as for the bacteria, except the top 500 most abundant chloroplast ASVs were used instead of the bacteria ASVs. Using the random forest classifier, pronounced variation in temporal patterns at the ASV-level was observed. While the majority of the top fifty most important ASVs for classifying expedition showed the expected temporal pattern for the chloroplasts, i.e. a decrease in relative abundance from the January to April expedition, there were also chloroplast ASVs that peaked in relative abundance in the Feb-March and April expeditions instead (Figure 49). For example, ASV_34 (*Proboscia indica*) had the greatest relative abundance in the Feb-March expedition. Most ASVs from the same order as ASV_34, the Rhizosoleniales, had the same temporal pattern (Figure 50). The ASVs that had the greatest relative abundance in the April expedition included ASV_179 (*Phalacroma mitra*, a dinoflagellate) ASV_17 & ASV_98 (*Rhizosolenia setigara*/*Cymbella intermedia*), ASV_18 (*Virgulinella fragilis*), ASV_6 (*Gyrosigma fasciola*) and ASV_64 (*Chaetoceros socialis*). As was the case for several groups of bacteria and archaea, different ASVs within the same order as ASV_64, the Chaetocerotales, had contrasting temporal patterns (Figure 51). Overall, the variation in temporal pattern for the chloroplasts at the ASV level may reflect different abiotic requirements and/or interactions with other species.

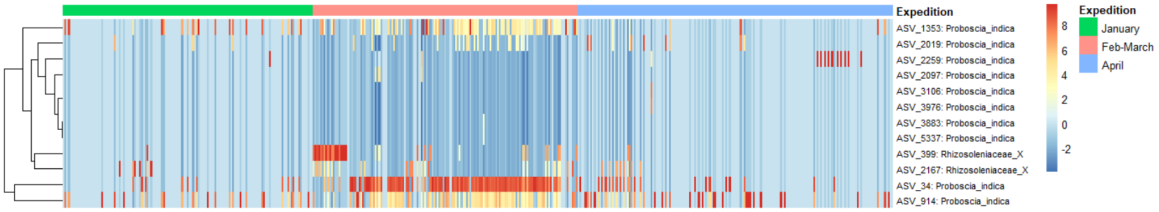


Figure 50 **Heatmap of abundant ASVs from the order Rhizosoleniales. Rows represent ASVs and columns represent samples, which are ordered by expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package "microbiomeutilities" v. 0.99.0 (Shetty and Lahti 2019) with transformation set to "clr".**

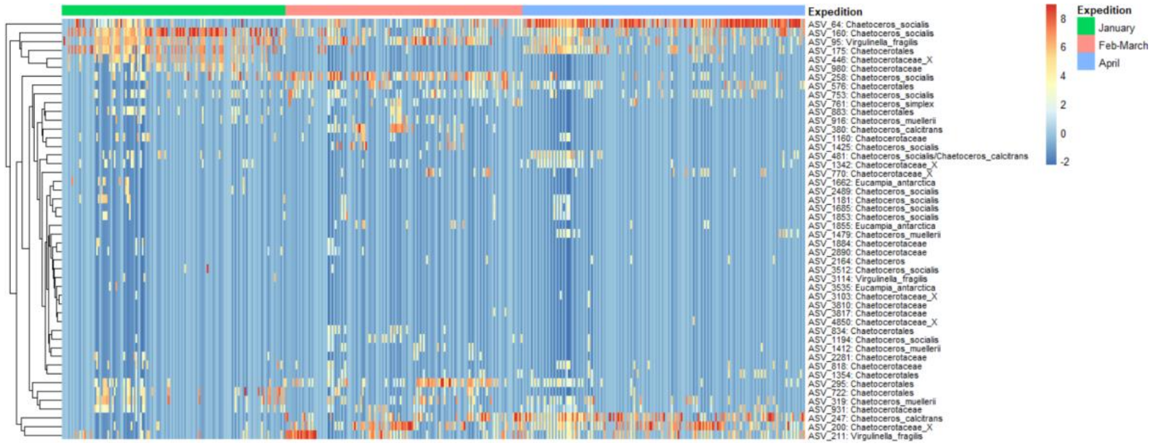


Figure 51 **Heatmap of abundant ASVs from the order Chaetocerotales. Rows represent ASVs and columns represent samples, which are ordered by expedition. The heatmap was generated using the plot_taxa_heatmap() function from the R package "microbiomeutilities" v. 0.99.0 (Shetty and Lahti 2019) with transformation set to "clr".**

4.4. Environmental variables

4.4.1. Geographic distance vs. beta diversity

Prior to exploring the environmental variables driving the *P. globosa* bloom, I examined whether the geographic distance between samples impacted the similarity of the microbial communities. This was done for the surface samples from each expedition and filtration size fraction by regressing the pairwise geographic distances between the samples against the pairwise Bray-Curtis distance metrics calculated using all microbes. The relationship between geographic distance and the Bray-Curtis distance was significant for all the expeditions and filtration sizes ($P < 0.0001$ in all cases) as sampling locations that had less geographic distance between them had more similar microbial communities (Figure 52). One explanation for this observation is that sampling locations that are closer together tend to have more similar abiotic environments, i.e. temperature and nutrient levels, which results in similar microbial community assemblages.

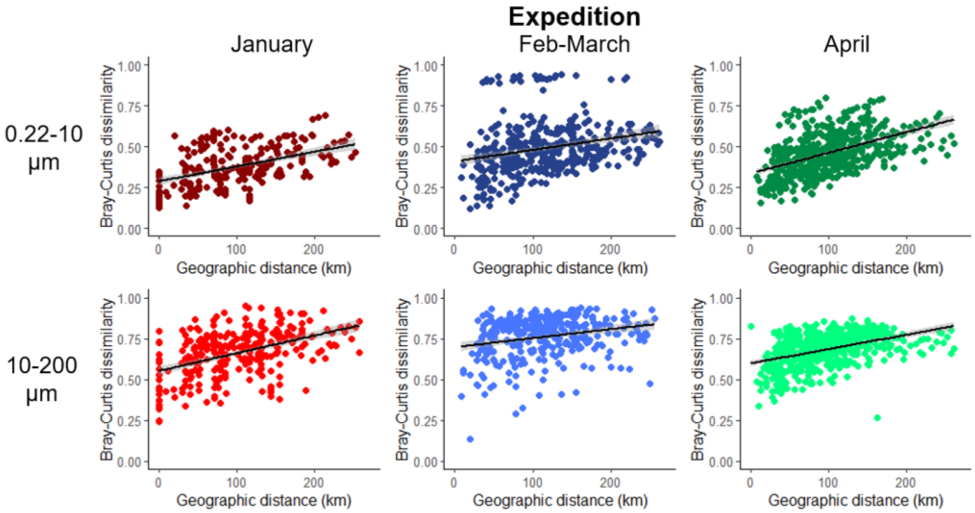


Figure 52 **The relationship between the pairwise geographic distance between samples and the pairwise Bray-Curtis distance (calculated using all microbes) for the surface samples from different expeditions and filtration sizes. The regression slope is significant for all expeditions and filtration sizes ($P < 0.0001$ for all).**

4.4.2. Correlations between environmental variables

From the eight measured environmental variables, seven pairs had moderate or strong correlations (|spearman ρ| ≥ 0.5). Five of these correlations were positive: NH_3 and temperature, PO_4 and NO_3, NO_2 and NO_3, SiO_3 and NO_3, and SiO_3 and PO_4 (Figure 53). The two negative correlations were between SiO_3 and temperature and SiO_3 and NH_3 (Figure 53).

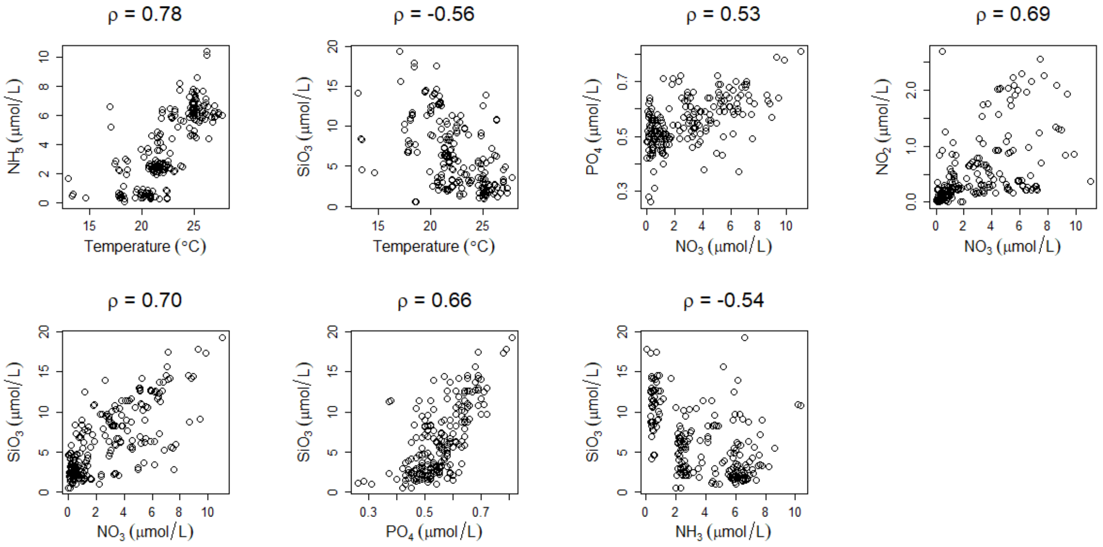


Figure 53 **Pairs of environmental variables from the Beibu Gulf with moderate or strong correlations (|spearman ρ| ≥ 0.5).**

4.4.3. Spatial-temporal dynamics of environmental variables in the Beibu Gulf

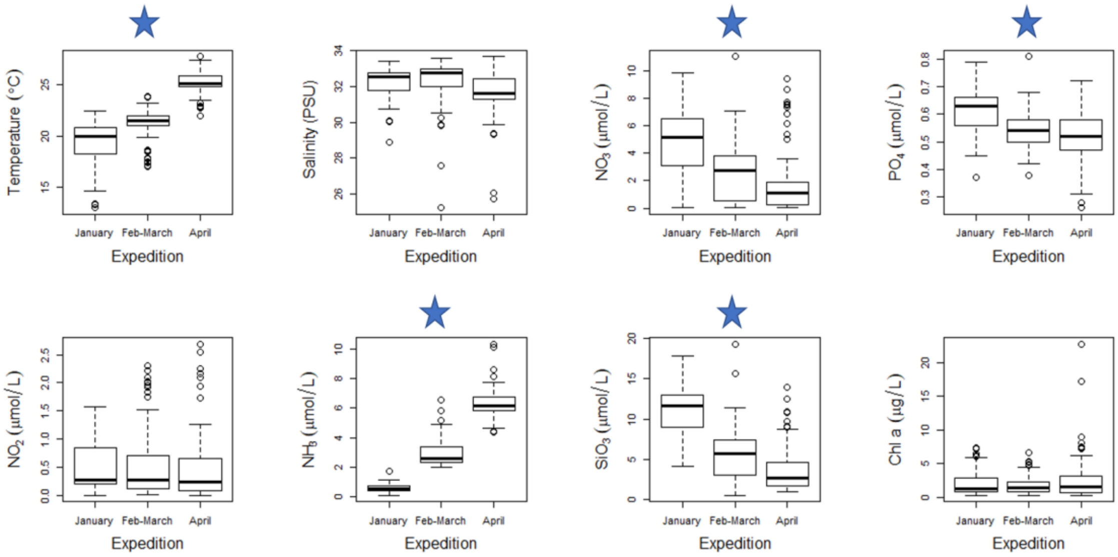


Figure 54 **Relationship of eight environmental variables with expedition month during a *P. globosa* bloom in the Beibu Gulf. Blue stars indicate variables with a significant relationship with expedition (ANOVA corrected for multiple comparisons P < 0.05).**

Like many of the microbes detected in the Beibu Gulf, the environmental variables exhibited distinct spatial-temporal patterns during the *P. globosa* bloom. Five of the environmental variables changed with expedition (ANOVA corrected for multiple comparisons P < 0.05): temperature and NH_3 increased from the January to April expedition, while NO_3, PO_4 and SiO_3 decreased (Figure 54). The directionalities of the loading arrows from a PCA using the eight environmental variables were consistent with this result (Figure 55). None of the environmental variables were related to the water sampling depth (|spearman ρ| < 0.5 for all). To further explore the spatial-temporal distribution of the environmental variables, the measurements from the surface samples were interpolated and in R using the “tmap” v. 3.0 (Tennekes 2018) package. In addition to its temporal pattern, temperature had a distinct spatial pattern as higher temperatures were observed farther away from the coast (Figure 56a). A similar spatial pattern was observed for salinity (Figure 56b), whereas chlorophyll a had the opposite pattern with higher values observed near the coast (Figure 56f). The spatial patterns of NO_3 and NO_2 were also similar to each other as both variables had greater values on the east side of the Beibu Gulf (Figure 56c and e). The spatial patterns for PO_4, NH_3 and SiO_3 were less clear (Figure 56d and f-g).

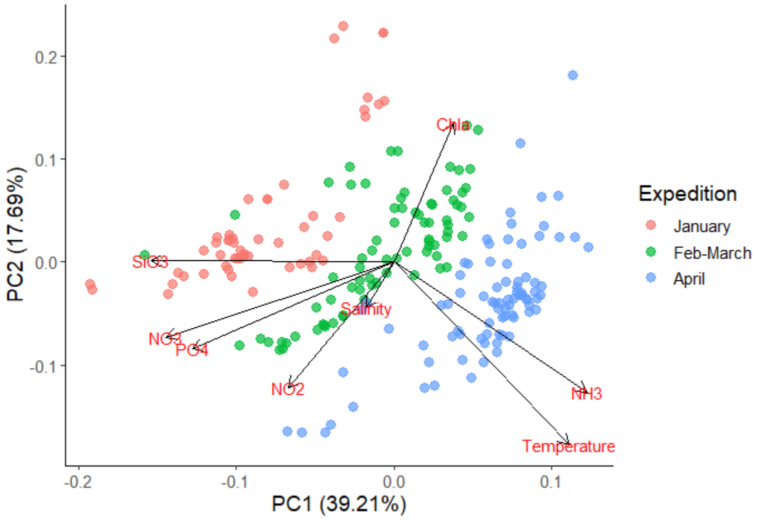


Figure 55 **PCA biplot of eight environmental variables measured during a P*. globosa* bloom in the Beibu Gulf. Observations are shown as points (coloured by expedition month) and variable loadings are shown as arrows.**

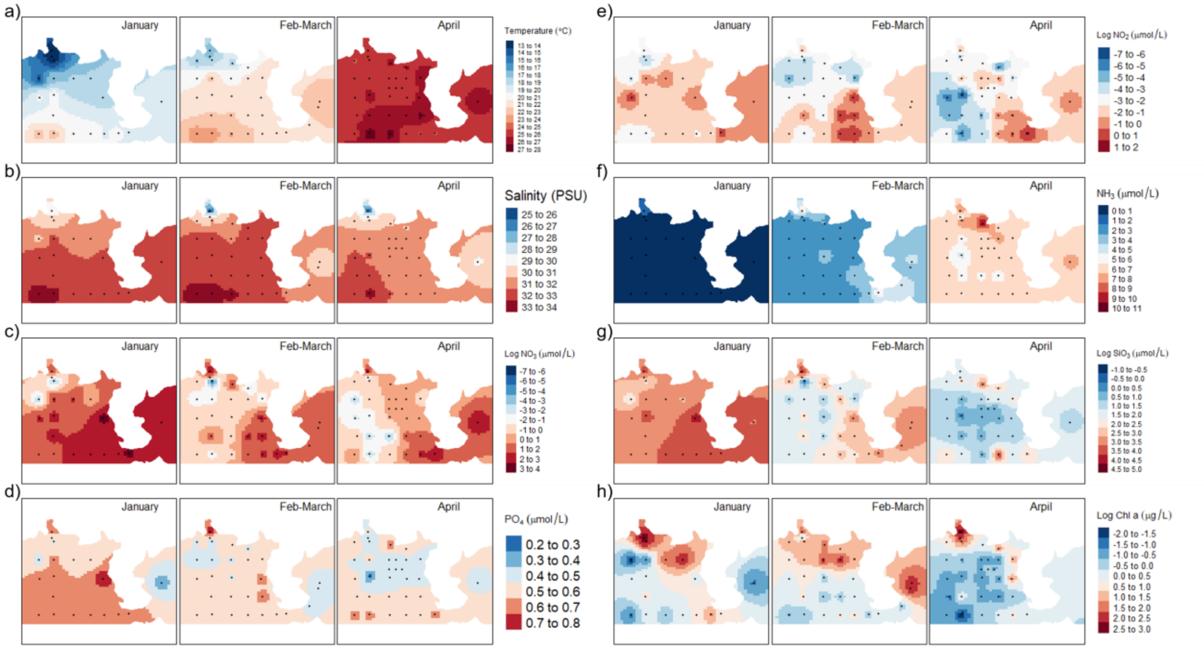


Figure 56 Spatial-temporal distribution of eight environmental variables during a *P. globosa* bloom in the Beibu Gulf. Map titles correspond to the expedition month. Mapping was performed for each variable by interpolating the values from the surface samples.

4.4.4. Correlation of environmental variables with *P. globosa*

To explore potential drivers of the *P. globosa* bloom in the Beibu Gulf, I tested for environmental variables that were correlated with *P. globosa* (ASV_10). Temperature, NH_3 and SiO_3 had moderate-strong correlations ($|\text{spearman } \rho| \geq 0.5$) with the log relative abundance of ASV_10 in the surface samples from both filtration sizes (Figure 57). The directions of these correlations were consistent with the temporal patterns of ASV_10 and the environmental variables. For example, the negative correlation between NH_3 and ASV_10 occurred because NH_3 increased from the January to April expedition, while the relative abundance of ASV_10 decreased from the January to April expedition (Figure 58). Similarly, temperature, which increased from the January to April expedition (Figure 54 and Figure 56a), was negatively correlated with ASV_10 (Figure 57). In contrast, SiO_3, which decreased from the January to April expedition (Figure 54 and Figure 56g), was positively correlated with ASV_10 (Figure 57). Additionally, after controlling for expedition month there was no relationship between any of the environmental variables and the relative abundance of ASV_10 ($P > 0.05$ in all cases) for both filtration sizes. This provides additional support that the observed correlations between ASV_10 and temperature, NH_3 and SiO_3 were due to temporal patterns rather than spatial patterns.

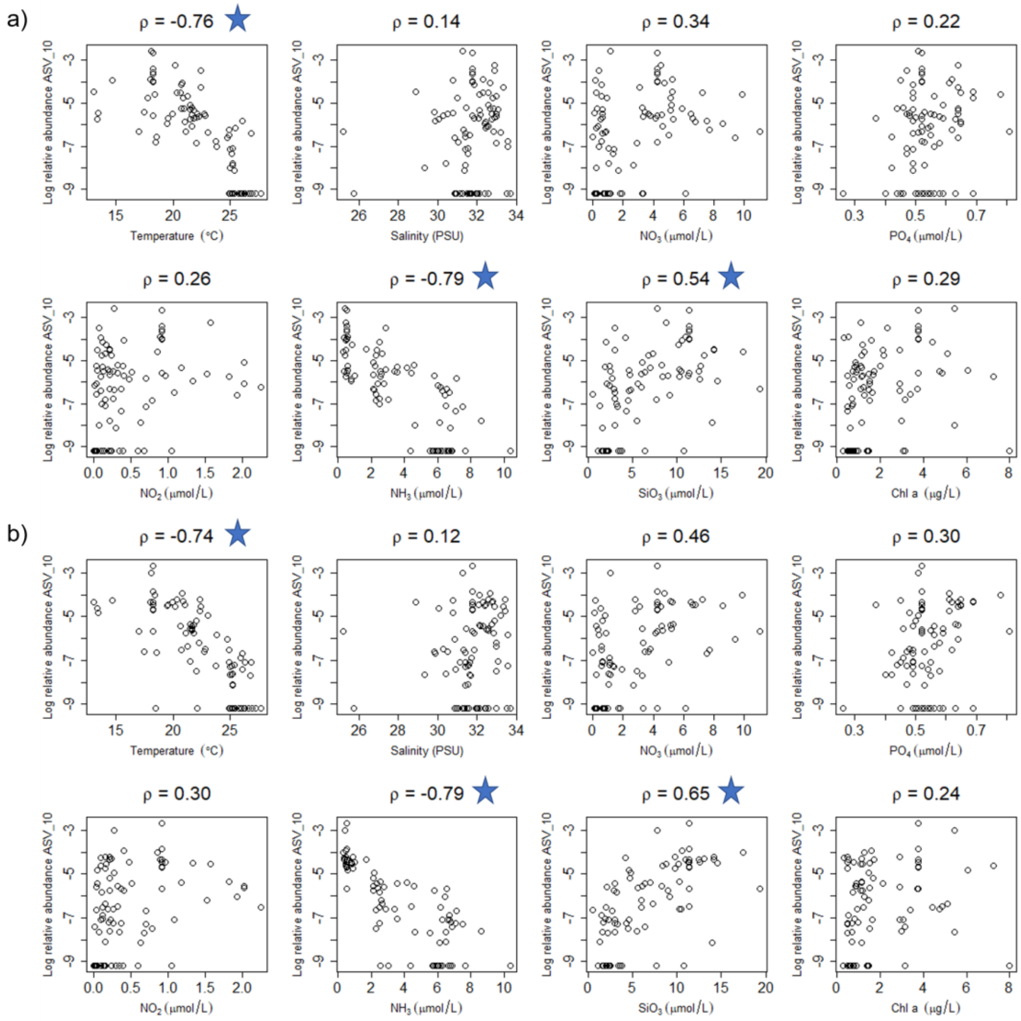


Figure 57 **Relationships of eight environmental variables with the log relative abundance of ASV_10 (*P. globosa*) in the surface samples from a) 0.2-10 μm filtration size fraction and b) 10-200 μm filtration size fraction. Blue stars indicate variables with |spearman ρ| ≥ 0.5.**

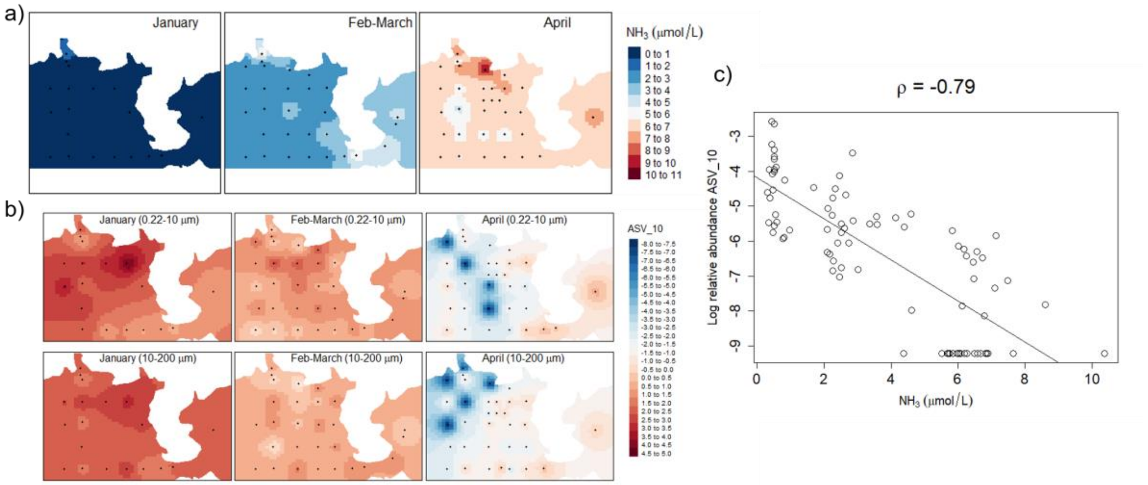


Figure 58 a) Spatial-temporal distribution of NH_3 mapped from the interpolated values from the surface samples, b) spatial-temporal distribution of ASV_10 (*P. globosa*) mapped by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values, and c) the relationship between NH_3 and the log relative abundance of ASV_10 in the surface samples from the 0.22-10 µm filtration size.

4.4.5. Canonical correspondence analysis

A canonical correspondence analysis (CCA) was implemented to explore correlations between the other ASVs and the environmental variables. CCA is similar to PCA, but instead of projecting one set of variables, CCA projects two sets of variables in a way that maximizes the correlations between the projections. CCA is thus often used as an exploratory technique for finding correlations between two sets of variables. A CCA was run on the data using the Bray-Curtis dissimilarity metric calculated for all microbes and the eight environmental variables as the constraining variables. Overall, the CCA explained 9% of the variance in the data. The first component (CCA1) explained 3.21% of the variance and reflected the temporal variation in the data (Figure 59a). In contrast, the second component (CCA2) explained 1.82% of the variance and appeared to reflect the spatial variation in the data as the loading arrows for salinity and chlorophyll a pointed in opposite directions along its axis (Figure 59a), reflecting the opposing spatial distributions of the two variables (Figure 56b and h). Plotting the centroids for the top fifteen most abundant orders provided initial insights into correlations between different groups and environmental variables (Figure 59b). For example, the Nitrosococcales and the Synechococcales, which both increased in relative abundance from the January to the April expedition, appeared to be positively correlated with temperature and NH_3 (Figure 59b).

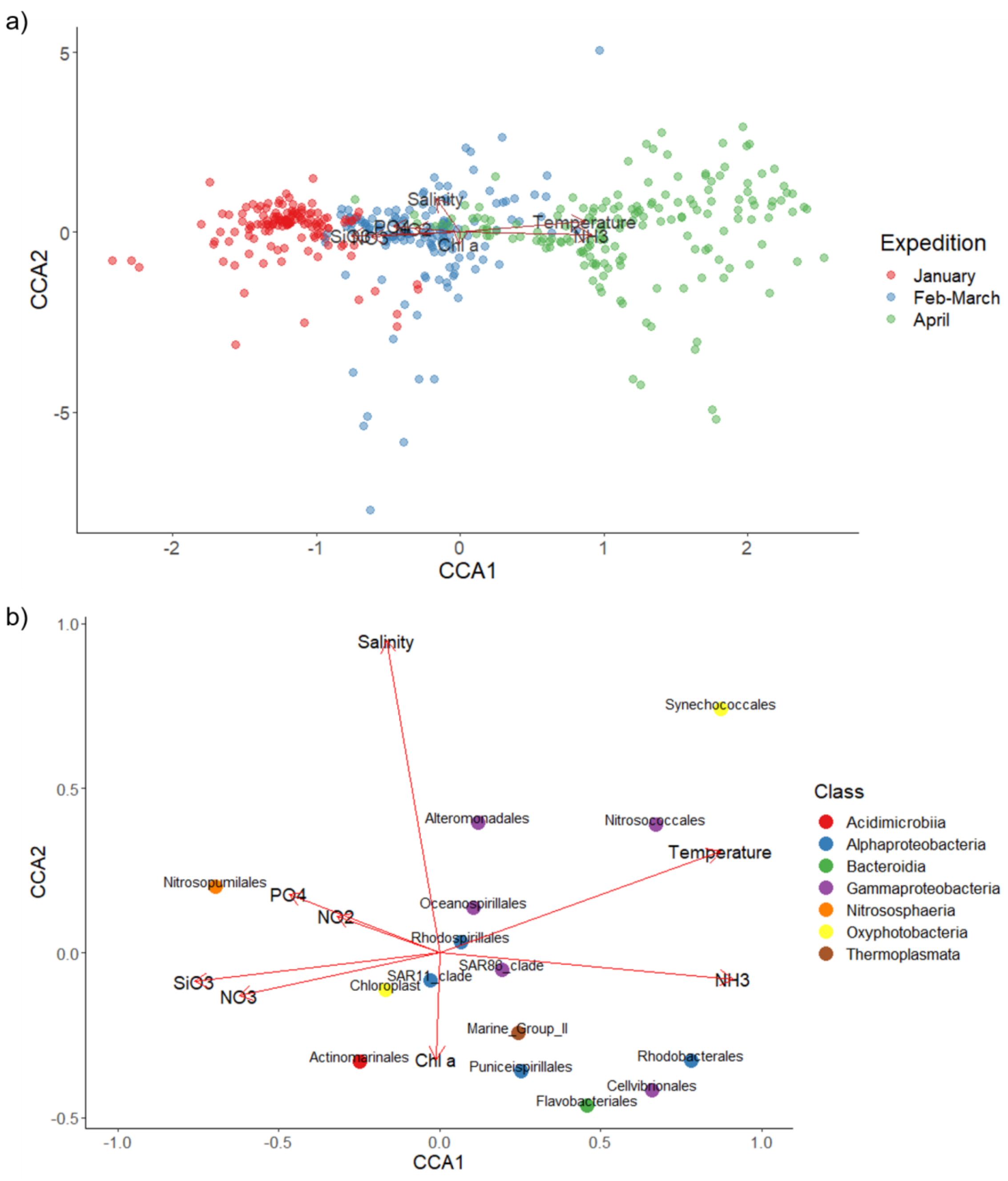


Figure 59 Biplots from a canonical correspondence analysis (CCA) using all microbes and eight environmental variables. The points in a) represent individual samples coloured by expedition and the points in b) represent centroids for the fifteen most abundant orders

4.4.6. Correlation of environmental variables with other microbes

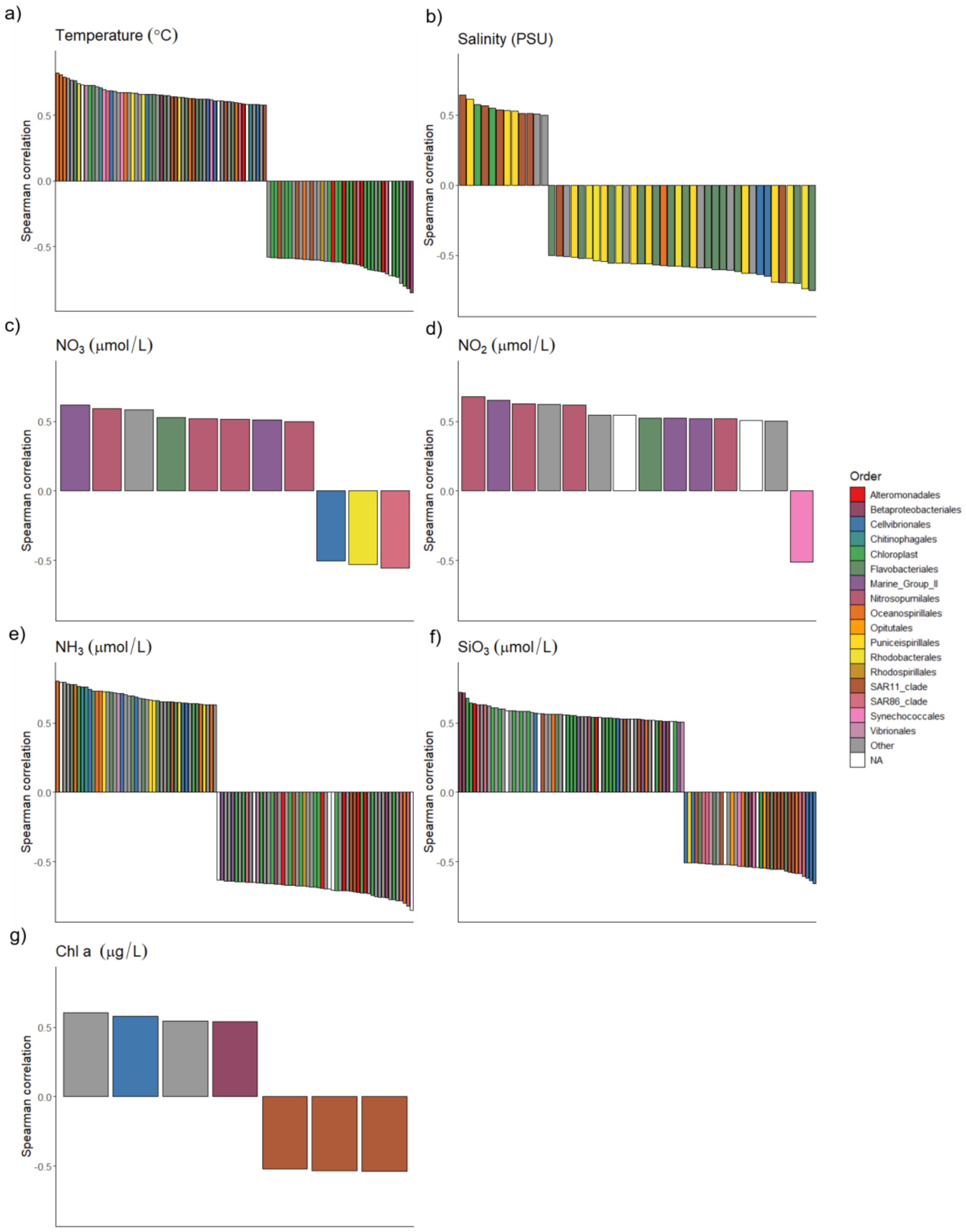


Figure 60 **ASVs that are correlated (P < 0.05 and |spearman ρ| ≥ 0.5) with environmental variables in the surface samples from the 0.2-10 µm filtration size. Each bar represents an individual ASV. Only the top 100 ASVs are shown for temperature, NH_3 and SiO_3.**

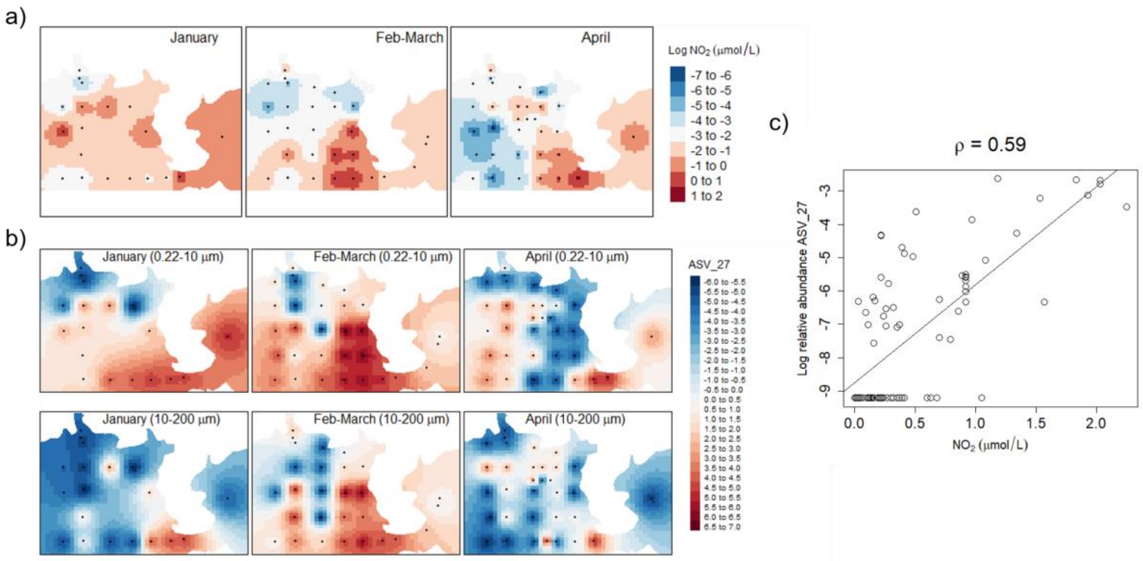


Figure 61 **a) Spatial-temporal distribution of NO_2 mapped from the interpolated values from the surface samples, b) spatial-temporal distribution of ASV_27 (genus *Candidatus_Nitrosopumilus*) mapped by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values, and c) the relationship between NO_2 and the log relative abundance of ASV_27 in the surface samples from the 0.22-10 μm filtration size.**

A spearman correlation matrix was generated to further explore correlations between other ASVs and the environmental variables. The correlation matrix was produced using the surface samples from the small filtration size and subsequently filtered for pairs of ASVs and environmental variables with Bonferroni adjusted $P < 0.05$ and $|\rho| \geq 0.5$. The environmental variables with the most correlated ASVs were NH_3 (n = 246), temperature (n = 181) and SiO_3 (n = 104). The ASVs that were correlated with these three variables had distinct temporal patterns. For example, many of the ASVs that were positively correlated with SiO_3 were chloroplasts that increased in relative abundance from the January to April expedition (Figure 60f). Four of eight of the ASVs that were positively correlated with NO_3 and four of thirteen of the ASVs that were positively correlated with NO_2 were Nitrosopumilales (Figure 60c and d) from the archaea class Nitrososphaeria. For example, ASV_27 from the genus *Candidatus_Nitrosopumilus* was positively correlated with NO_2 ($\rho = 0.59$; Figure 61). The Nitrososphaeria gain energy from the oxidation of NH_3 and their abundance is closely tied to the flux of NH_4, NO_2 and NO_3 (Wuchter et al. 2006, Santoro et al. 2019). Consistent with this, in the Beibu Gulf the Nitrososphaeria had the greatest relative abundance in the January expedition, when NH_3 levels were lowest and the levels of NO_2/NO_3 were highest, and decreased in relative abundance in the Feb-March and April expeditions when NH_3 levels were highest and the levels of NO_2/NO_3 were lowest. The ASV with the third highest positive correlation coefficient for NH_3 was ASV_20 (*Methylophaga marina)* from the Nitrosococcales (ρ = 0.78; Figure 62). The Nitrosococcales gain their energy from the oxidation of NH_3, which explains their overall increase in relative abundance from the January to April expedition and the strong positive correlation of ASV_20 with NH_3.

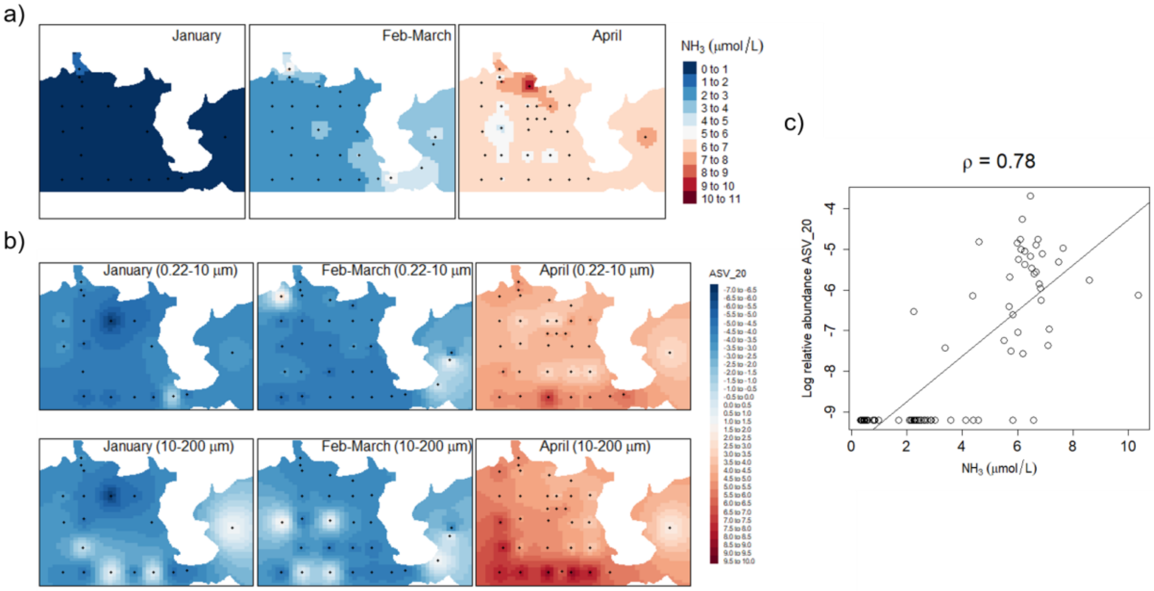


Figure 62 **a) Spatial-temporal distribution of NH_3 mapped from the interpolated values from the surface samples, b) spatial-temporal distribution of ASV_20 (*Methylophaga marina*) mapped by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values, and c) the relationship between NH_3 and the log relative abundance of ASV_20 in the surface samples from the 0.22-10 μm filtration size.**

4.5. Discussion

The aim of this chapter was to explore the spatial-temporal dynamics of *P. globosa*, bacteria, archaea, phytoplankton and environmental factors during a *P. globosa* bloom in the Beibu Gulf. Distinct communities of bacteria, archaea and eukaryotes were observed at the three different time points during the bloom. In many cases different groups had distinct temporal patterns, which could be related to their ecological niches during the bloom using descriptions from the literature as well as the temporal patterns of the environmental variables. Additionally, variation in the temporal response to the bloom at the ASV-level was observed for the bacteria, archaea and chloroplast ASVs, which provides evidence for an underappreciated amount of variation of niches within taxonomic groups.

A distinct succession of bacterial groups was observed across the three expeditions during the *P. globosa* bloom in the Beibu Gulf (Figure 30). Successions of bacteria during phytoplankton blooms are linked to the abilities of different bacteria to degrade algal-derived organic matter (Teeling et al. 2012). Bacteria that reside on algal cells or colonies also alter their structure as a phytoplankton bloom progresses, which is in part driven by responses of bacteria to phytoplankton exudates like DMSP (Delmont et al. 2014). *Phaeocystis* blooms provide ecological niches for microbial heterotrophs (Delmont et al. 2014) as the majority of biomass produced by *Phaeocystis* blooms is remineralized by heterotrophic bacteria at the end of the bloom, which may explain the increase in alpha diversity and relative abundance of bacteria in the April expedition (Figure 27b and Figure 29). The composition of organic matter that can be utilized by bacteria changes throughout *Phaeocystis* blooms, which results in changes in the bacterial community structures (Alderkamp et al. 2007). For example, complex carbohydrates like glucan and mucopolysaccharides that are produced during *Phaeocystis* blooms may shape the composition of bacterial communities that are specialised in the degradation of complex carbohydrates (Arrieta and Herndl 2002). One example of this in the Beibu Gulf is the increase at the end of the bloom in the relative abundance of the different groups of Bacteroidetes (Figure 36), which are specialized in the degradation of high molecular weight organic matter. This observation is consistent with the increase in the contribution of Bacteroidetes that was observed during the decay of a *P. globosa* bloom in the North Sea (Alderkamp et al. 2006). Similarly, I observed an increase at the end of the bloom in the relative abundance of the Rhodobacterales (Figure 32a), which are involved in the degradation or organic carbon after phytoplankton blooms (Buchan et al. 2014) and have been observed to increase in contribution during the decay of experimental (Brussaard et al. 2005) and natural (Alderkamp et al. 2006) *P. globosa* blooms. *Phaeocystis* blooms may also alter the structure of particle-associated bacteria as the bloom progresses because the colonies provide a habitat for bacterial species (Delmont et al. 2014) For example, particle-associated bacteria during a *P. antarctica* bloom were enriched in the genus *Colwellia* (Delmont et al. 2014), which also had the greatest relative abundance during the peak of the bloom in the Beibu Gulf (Figure 34). Strikingly, there was a marked amount of variation in response to the bloom from different groups of bacteria at the ASV-level (e.g. Figure 35 and Figure 40), which suggests that many of these groups have intragroup variation in their ability to utilize organic matter or associate with particles like *P. globosa* colonies.

Like the bacteria, there was distinct changes in the archaea communities across the three expeditions during the *P. globosa* bloom (Figure 42).The MGII have been observed to form blooms that are coincident with or just following phytoplankton blooms (Galand et al. 2010, Needham and Fuhrman 2016), which may be related to their speculated role as facultative colonizers of particles (Orsi et al. 2015, 2016, Santoro et al. 2019). The peak in abundance of MGII during the Feb-March expedition in the Beibu Gulf (Figure 44) suggests that the MGII may have been involved in the early degradation of *Phaeocystis* colonies and other phytoplankton. As was observed for many groups of bacteria, the MGII also had variation in their temporal response to the bloom at the ASV-level (Figure 45), suggesting different metabolic roles within the group. The other major group of archaea in the Beibu Gulf was the Nitrososphaeria, which are known to gain energy from the oxidation of ammonia and have previously had their abundance correlated to changes in ammonium and other nitrogen compounds (Wuchter et al. 2006, Santoro et al. 2019). Consistent with this, four ASVs in the Beibu Gulf from the Nitrososphaeria were positively correlated with NO_2 and NO_3 (e.g. Figure 61) and the relative abundance of the Nitrososphaeria peaked in the January expedition (Figure 43) when NH_3 levels were lowest and NO_2/NO_3 were highest (Figure 54).

I observed distinct changes in the chloroplast ASVs across the three expeditions during the *P. globosa* bloom (Figure 47), which can largely be explained by the temporal patterns of the environmental variables. As expected, the relative abundance of ASV_10 (*P. globosa*) peaked during the January expedition (Figure 26). The development of Phaeocystis blooms has previously been associated with changes in daily irradiance, temperature and increasing nutrient loads, especially NO_3 and PO_4 (Riegman et al. 1992, Peperzak et al. 1998), which were both highest in the January expedition in the Beibu Gulf (Figure 54). This is consistent with the *P. globosa* bloom in the Beibu Gulf being triggered by high levels of N and P coupled with changes in the daily irradiance and temperature. The majority of chloroplast ASVs also had the greatest relative abundance in the January expedition at the peak of the bloom (Figure 46), but some chloroplast ASVs displayed different temporal patterns (Figure 49, Figure 50 and Figure 51), which could be explained by differences in nutrient requirements and/or different interactions with other species in the environment. In addition to N and P, Si is an important nutrient for the diatoms, which is not used by *Phaeocystis*. A previous review from the North Sea found that *Phaeocystis* blooms after the spring diatom bloom once Si has been depleted because *Phaeocystis* cannot compete with diatoms for N and P (Peperzak et al. 1998). In contrast, *Phaeocystis*-diatom blooms occur concurrently in the Dutch coastal zone, presumably because high Si concentrations allow diatoms to bloom at the same time (Peperzak et al. 1998). In the Beibu Gulf, it is likely that the high SiO_3 concentrations in the January expedition (Figure 54) allowed the diatoms and other phytoplankton to thrive alongside *Phaeocystis*.

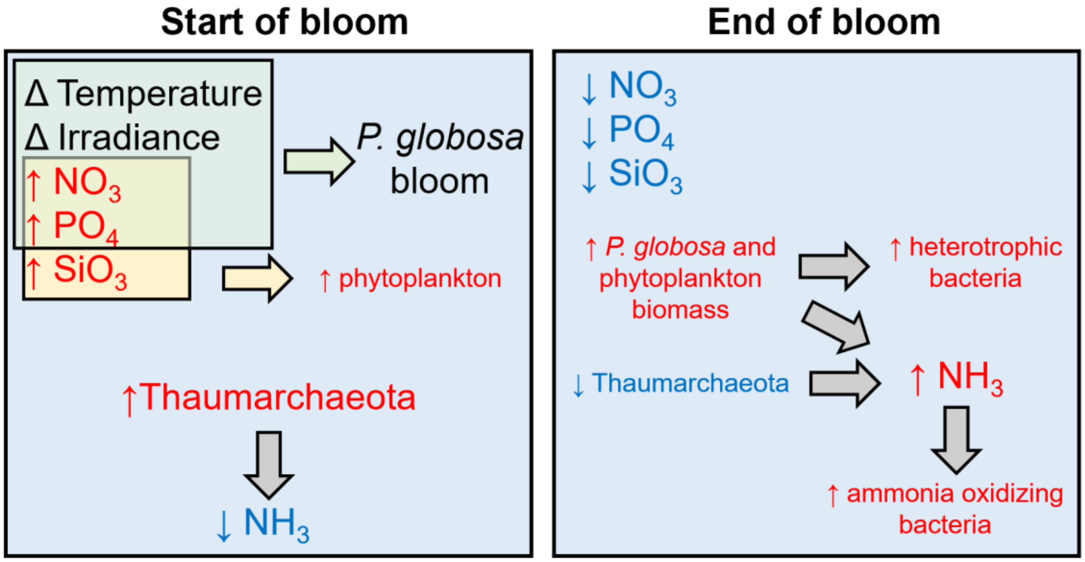


Figure 63 **Model of the development and progression of the *P. globosa* bloom in the Beibu Gulf.**

Based on the observed spatial-temporal patterns of microbes and environmental variables, I developed a preliminary model for the development and progression of the *P. globosa* bloom in the Beibu Gulf (Figure 63). The *P. globosa* bloom was likely triggered in January by changing temperature and irradiance levels and high levels of NO_3 and PO_4. I hypothesize that the high levels of NO_3, PO_4 and SiO_3 allowed the diatoms and other phytoplankton to thrive alongside *Phaeocystis*. As the bloom progressed, NO_3, PO_4 and SiO_3 were depleted by *P. globosa* and other phytoplankton and levels of NH_3 increased from the degradation of phytoplankton biomass and a decrease in the relative abundance of the Nitrososphaeria (Thaumarchaeota), which deplete NH_3. Other groups that gain energy from oxidizing ammonia, e.g. the Nitrosococcales, then increased in relative abundance at the end of the bloom due to the increase in NH_3. Finally, groups of heterotrophic bacteria, e.g. the Flavobacteriales, increased in relative abundance at the end of the bloom due to their ability to thrive off organic materials produced during bloom decay. This model describes an initial understanding of the mechanisms underlying the *P. globosa* bloom in the Beibu Gulf and is based off the spatial-temporal patterns of the microbes and environmental variables during the bloom. Two limitations of the analyses used to develop the preliminary model are 1) the use of observational data, which limits the strength of the evidence supporting the model, and 2) the lack of technical and biological replicates. To address the second limitation, for the temporal analyses, samples collected at the same time points, but different sampling locations, were used as pseudo-replicates. Future studies could benefit from 1) the collection of experimental data using controlled conditions such as mesocosms to strengthen the evidence supporting the model and 2) the use of technical and biological replicates. In the next chapter, our understanding of the *P. globosa* bloom mechanisms is further advanced by exploring microbes that potentially interacted with *P. globosa* during the bloom.

Chapter 5. Exploration of microbes that interact with *P. globosa* during a *P. globosa* bloom in the Beibu Gulf

The aim of this chapter is to explore microbes that potentially interact with *P. globosa* colonies during a bloom in the Beibu Gulf using 16S amplicon sequencing. First, I identify ASVs that potentially interacted with *P. globosa* using interaction networks constructed from the field samples. Next, I identify ASVs that were associated with large particles, such as *P. globosa* colonies, by identifying ASVs enriched in the large filtration size fraction from the field samples. Finally, I explore the *P. globosa* colony microbiome using the 16S gene sequences from the colony samples. My contribution to this chapter is the data analysis and interpretation.

5.1. Interaction networks

Interaction networks are a powerful tool in molecular ecology for generating hypotheses of ecological interactions between taxa. A microbial interaction network is a set of microbial taxa (nodes) connected by edges that represent potential biological interactions between taxa. These edges can represent positive interactions, e.g. mutualism or commensalism, if two taxa co-occur more frequently than expected by chance or negative interactions, e.g. competition, if they co-occur less frequently than expected by chance. While it is challenging to determine whether pattern of co-occurrence are due to interactions between taxa or environmental constraints, interaction networks are still a useful first step in developing interaction hypotheses (Hugerth and Andersson 2017). Microbial relationships depicted from interaction networks can be used to determine drivers in environmental ecology (e.g. Lima-Mendez et al. 2015) and for generating hypotheses for further study (Weiss et al. 2016). However, constructing a microbial interaction network is a challenging task.

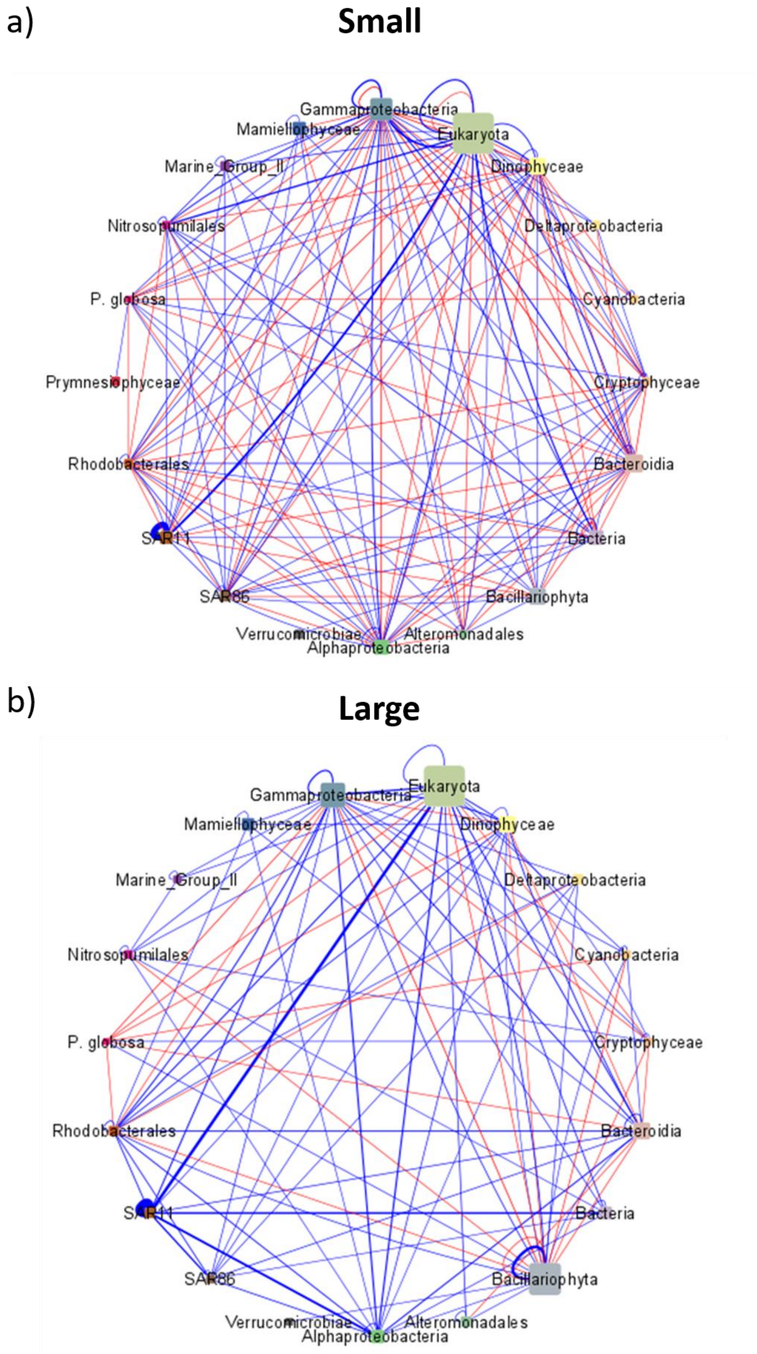


Figure 64 **Interaction networks for the surface samples from a) the small filtration size fraction and b) the large filtration size fraction. Interactions were inferred at the ASV-level using FastSpar (Watts et al. 2019), filtered for correlations with |Pearson's coefficient| ≥ 0.6 and P = 0.001 and collapsed into higher taxonomy levels for visualization. Node size corresponds to the number of ASVs in the group. Blue edges represent positive interactions, red edges represent negative interactions and edge width corresponds to the number of ASVs that are correlated between the two groups.**

A number of methods have been developed to address the challenges of building interaction networks from molecular ecology data. The simplest methods rely upon measures of correlation such as Spearman's and Pearson's. However, simple correlation approaches are impeded by limiting sampling depth and the compositionality of the data, which can induce spurious correlations (Friedman and Alm 2012, Hugerth and Andersson 2017). Sparse Correlations for Compositional data (SparCC) circumvents these challenges by estimating correlations using the log-ratio transformation of compositional data under the assumptions that microbial networks are large-scale and sparse (Friedman and Alm 2012). Another challenge in building molecular interaction networks is that many microbes display diverse types of relationships, e.g. exponential or periodic, that cannot be detected by a single test (Reshef et al. 2011). Maximal Information Coefficient (MIC) is designed to capture a wide range of associations between taxa by implementing a non-parametric approach for detecting associations that uses a measure of the predictability of two variables in relation to each other. (Reshef et al. 2011). The tool CoNet also addresses this challenge by combining information from several different standard comparison metrics (Faust et al. 2012). Finally, Local Similarity Analysis (LSA) is a tool that is optimized to detect non-linear and time-sensitive relationships from time-series data (Ruan et al. 2006).

I elected to use FastSpar, a C++ implementation of the SparCC algorithm that is faster and less memory intensive (Watts et al. 2019), to generate hypotheses for species that interact with *P. globosa* during its bloom in the Beibu Gulf. Two interaction networks were constructed from the surface samples: one for the small filtration size fraction and one for the large. For each network, ASVs with an average of less than two reads per sample were filtered out (as reccomended by Friedman and Alm 2012) and 1,000 bootstrap permutations were used to calculate the P-values. The resulting interactions were filtered for |Pearson's correlation coefficient| ≥ 0.6 and P = 0.001. While it would be ideal to correct the P-values for multiple comparisons, this approach is too computationally intensive in this case due to the large number of permutations required to obtain a significant corrected P-value. The smallest possible P-value for 1,000 bootstrap permutations (P = 0.001) was thus applied as the threshold for significance. Cytoscape (Shannon et al. 2003) was used to visualize the interaction network.

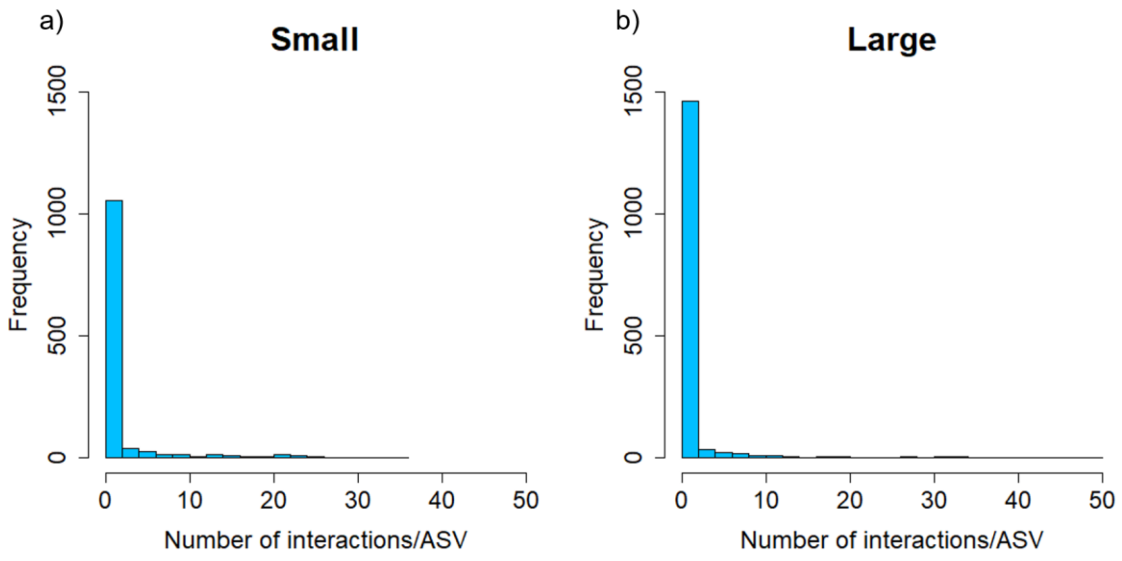


Figure 65 Histograms of the number of interactions/ASV inferred using FastSpar (Watts et al. 2019) for the surface samples from a) the small filtration size fraction and b) the large filtration size fraction.

The interaction network from the small filtration size fraction had 1,099 interactions from 1,232 ASVS, 874 of which were positive and 225 were negative (Figure 64a). The network from the large filtration size fraction had 1,149 interactions from 1,625 ASVs, but less of these were negative as there were 1,096 positive interactions and only 53 negative interactions (Figure 64b). For both interaction networks, the majority of ASVs had zero interactions (Figure 65). The mean interactions/ASV in the small filtration size network was 1.78 with a maximum of 36 interactions from ASV_20 (*Methylophaga marina*). The group with the most interactions was SAR11 (*n* = 502) (Figure 64a). For the large filtration size network, there was a mean of 1.41 interactions/ASV with a maximum of 50 interactions from ASV_16 from the AEGEAN-169_marine_group. Like the small filtration size fraction network, the group with the most interactions was SAR11 (*n* = 822) (Figure 64b).

5.1.1. Microbes with potential interactions with *P. globosa*

The ASVs that were correlated with ASV_10 (*P. globosa*) in the interaction networks generated hypotheses for taxa with biologically meaningful interactions with *P. globosa*. Twenty-six ASVs were correlated with ASV_10 in the surface samples from the small filtration size (four positive and fourteen negative) (Table 3) and eighteen ASVs were correlated with ASV_10 in the samples from the large filtration size (four positive and fourteen negative) (Table 4), many of which were correlated in both networks. The ASVs that were positively correlated with ASV_10 had the same temporal pattern (decrease in relative abundance from the January to the April expedition) and are candidates for forming symbiotic relationships with *P. globosa* colonies. However, these ASVs may also have similar abiotic requirements to *P. globosa* that explain their positive correlations. The ASVs that were negatively correlated with *P. globosa* have the opposite temporal pattern (increase in relative abundance from the January to April expedition) and are candidates for taxa that compete with *P. globosa* or feed on decaying *P. globosa* colonies. However, these negative correlations may also have occurred because these ASVs have opposing abiotic requirements to *P. globosa*. Further investigation is required to determine the role of these taxa correlated with *P. globosa* during its bloom.

Table 3 ASVs that were correlated with ASV_10 (*P. globosa*) in the surface samples from the small filtration size fraction. Correlations were calculated using FastSpar (Watts et al. 2019) and filtered for |Pearson's coefficient| ≥ 0.6 and P = 0.001.

ASV	Correlation coefficient	Order	Family	Genus
ASV_23	0.7307	Thalassiosirales	NA	NA
ASV_30	0.7303	Thiomicrospirales	Thioglobaceae	SUP05_cluster
ASV_72	0.7176	Pyrenomonadales	Pyrenomonadales_XX	Pyrenomonadales_XXX
ASV_42	0.6645	Alteromonadales	Pseudoalteromonadaceae	Pseudoalteromonas
ASV_257	0.6318	NA	NA	NA
ASV_170	0.6259	Betaproteobacteriales	Methylophilaceae	NA
ASV_122	0.6228	Pyrenomonadales	Pyrenomonadales_XX	Pyrenomonadales_XXX
ASV_343	0.6139	Prymnesiales	Chrysochromulinaceae	Chrysochromulinaceae_X
ASV_251	0.6127	Nitrosococcales	Nitrosococcaceae	Cm1-21
ASV_87	0.6104	Pyrenomonadales	NA	NA
ASV_248	0.6093	Betaproteobacteriales	Methylophilaceae	NA
ASV_58	0.6061	Pyrenomonadales	Pyrenomonadales_XX	Pyrenomonadales_XXX
ASV_390	-0.6064	Oceanospirillales	Alcanivoracaceae	Alcanivorax
ASV_528	-0.6073	Acidithiobacillales	Acidithiobacillaceae	KCM-B-112
ASV_522	-0.6130	Oceanospirillales	SS1-B-06-26	NA
ASV_354	-0.6265	NA	NA	NA
ASV_642	-0.6320	Flavobacteriales	Cryomorphaceae	NA
ASV_133	-0.6385	Vibrionales	Vibrionaceae	Catenococcus
ASV_45	-0.6410	NA	NA	NA
ASV_208	-0.6516	Salinisphaerales	Salinisphaeraceae	Salinisphaera
ASV_1	-0.6587	Synechococcales	Cyanobiaceae	Synechococcus_CC9902

ASV	Correlation coefficient	Order	Family	Genus
ASV_35	-0.6614	Rhodobacterales	Rhodobacteraceae	NA
ASV_407	-0.6623	Flavobacteriales	Flavobacteriaceae	Formosa
ASV_322	-0.6679	Oceanospirillales	Oleiphilaceae	Oleiphilus
ASV_81	-0.6861	Chitinophagales	Saprospiraceae	Lewinella
ASV_20	-0.6862	Nitrosococcales	Methylophagaceae	Methylophaga

Table 4 ASVs that were correlated with ASV_10 (*P. globosa*) in the surface samples from the large filtration size fraction. Correlations were calculated using FastSpar (Watts et al. 2019) and filtered for |Pearson's coefficient| ≥ 0.6 and P = 0.001.

ASV	Correlation coefficient	Order	Family	Genus
ASV_92	0.6857	NA	NA	NA
ASV_87	0.6259	Pyrenomonadales	NA	NA
ASV_36	0.6259	Thalassiosirales	NA	NA
ASV_23	0.6226	Thalassiosirales	NA	NA
ASV_169	-0.6025	Cellvibrionales	Halieaceae	OM60(NOR5) clade
ASV_88	-0.6205	Synechococcales	Cyanobiaceae	Cyanobium_PCC-6307
ASV_522	-0.6241	Oceanospirillales	SS1-B-06-26	NA
ASV_528	-0.6366	Acidithiobacillales	Acidithiobacillaceae	KCM-B-112
ASV_263	-0.6374	Rhodobacterales	Rhodobacteraceae	Tropicibacter
ASV_390	-0.6420	Oceanospirillales	Alcanivoracaceae	Alcanivorax
ASV_473	-0.6486	Flavobacteriales	Flavobacteriaceae	NA
ASV_1	-0.6529	Synechococcales	Cyanobiaceae	Synechococcus_CC9902
ASV_208	-0.6568	Salinisphaerales	Salinisphaeraceae	Salinisphaera
ASV_133	-0.6649	Vibrionales	Vibrionaceae	Catenococcus
ASV_322	-0.6766	Oceanospirillales	Oleiphilaceae	Oleiphilus
ASV_45	-0.6818	NA	NA	NA
ASV_20	-0.6910	Nitrosococcales	Methylophagaceae	Methylophaga
ASV_81	-0.7029	Chitinophagales	Saprospiraceae	Lewinella

5.2. Particle-associated microbes

5.2.1. Methods for identifying differentially abundant ASVs

Normalization is a critical step in identifying differentially abundant OTUs/ASVs due to the differences in library size and the sparsity of OTU/ASV tables that is inherent in molecular ecology data (Weiss et al. 2017). Rarefying is commonly used for normalization and is considered the standard in molecular ecology (Weiss et al. 2017), but a number of alternative normalization methods exist. One alternative approach involves scaling the counts of the count matrix by a quantile of the data. For example, the log upper quartile (logUQ) scales each sample by the 75th percentile of its count distribution followed by a log transformation (Bullard et al. 2010). Cumulative sum scaling (CSS) is similar to logUQ, but uses a distribution-dependent threshold for determining the quantile divisor of each sample and only scales the segments of the distribution that are relatively invariant across samples (Paulson et al. 2013). The variance stabilization approach implemented by DESeq2 calculates a scaling factor for each OTU/ASV in each column and divides all the reads for each column by the median of its scaling factors. A mean-variance relation is then fit for all OTUs/ASVs using a negative binomial (NB) generalized linear model (GLM) to adjust the matrix counts so that the variance in the counts across samples is approximately independent of its mean (Love et al. 2014). Finally, edgeR uses a Trimmed Mean by M-Values (TMM) scaling factor, which is calculated as the weighted mean of log-ratios between each pair of samples. The normalization factors for each sample are a product of the TMM scaling factor and the original library size (Robinson et al. 2010).

Following normalization, the identification of differentially abundant OTUs/ASVs is a challenging task due to the zero-inflation and over-dispersion of microbiome data. The simplest approaches for identifying differentially abundant OTUs/ASVs use nonparametric tests, e.g. the Mann-Whitney/Wilcoxon rank-sum test for tests of two groups and the Kruskal-Wallis test for tests of multiple groups, on the rarefied count data (Weiss et al. 2017). These approaches, however, do not account for the compositionality of microbial marker gene data. Alternatively, analysis of composition of microbiomes (ANCOM) accounts for the composition of microbiome data by comparing the log-ratio of the abundance of each OTU/ASV to the abundance of all the remaining OTUs/ASVs one at a time using the Mann-Whitney U (Mandal et al. 2015). More recently, parametric models developed for differential gene expression testing on RNA-Seq data have been applied to microbial marker gene data. These models are composed of GLMs that assume a distribution; often either a Poisson, NB or zero-inflated lognormal (Weiss et al. 2017). Some of the more popular tools are DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010), both of which assume a NB model. In contrast, metagenomeSeq, a tool designed for metagenomics data (Paulson et al. 2013), assumes a zero-inflated Gaussian model. There is currently no consensus in the field as to which of these tools performs best on microbial marker gene data (Weiss et al. 2017) so I elected to use both DESeq2 and edgeR for the analyses to increase the strength of the results.

5.2.2. ASVs with differential abundance between filtration sizes

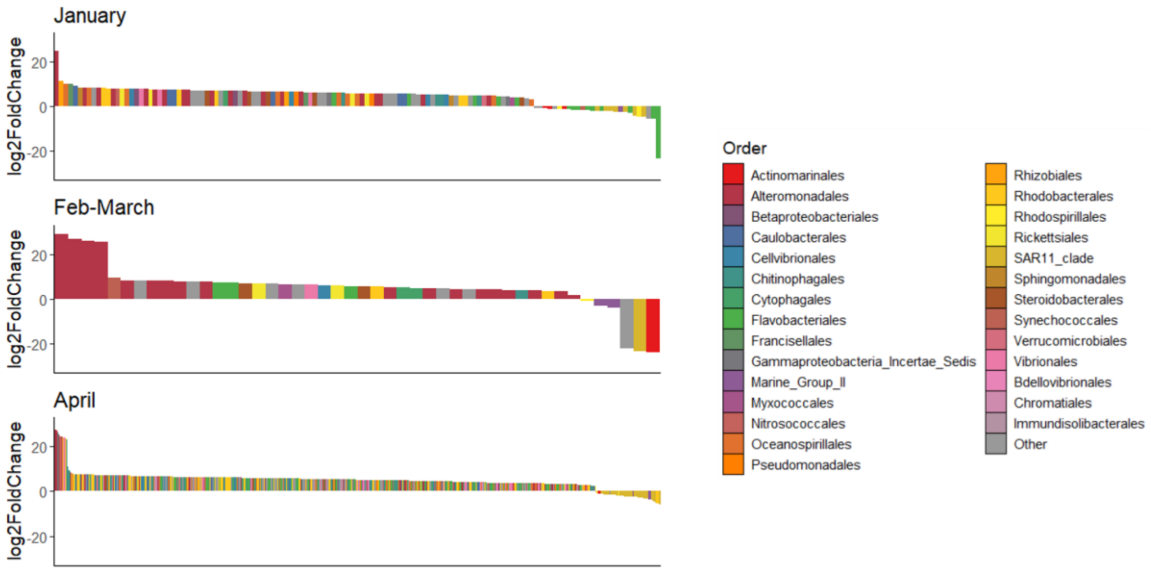


Figure 66 **Bacteria and archaea ASVs identified by DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010) as differentially abundant between the small (0.2-10 µm) and large (2-10 µm) filtration sizes in the surface samples from each expedition. Each bar represents an indiviudal ASV. Log2FoldChange > 0 indicates ASVs that are more abundant in the large filtration size and log2FoldChange < 0 indicates ASVs that are more abundant in the small filtration size.**

DESeq2 and edgeR were used to identify particle-associated microbes that may have interacted with *P. globosa* during its bloom. First, DESeq2 was run on the surface samples from each of the three expedition separately to identify bacteria and archaea ASVs that were differentially abundant (threshold P < 0.05) between the samples from the small and the large filtration sizes in each expedition. Next, the same analysis was performed using edgeR and the differentially abundant ASVs from DESeq2 were filtered to only include ASVs that were also differentially abundant (threshold FDR < 0.05) using edgeR. Bacteria and archaea ASVs that were enriched in the large (10-200 µm) filtration size are more likely to be associated with large particles such as *P. globosa* colonies.

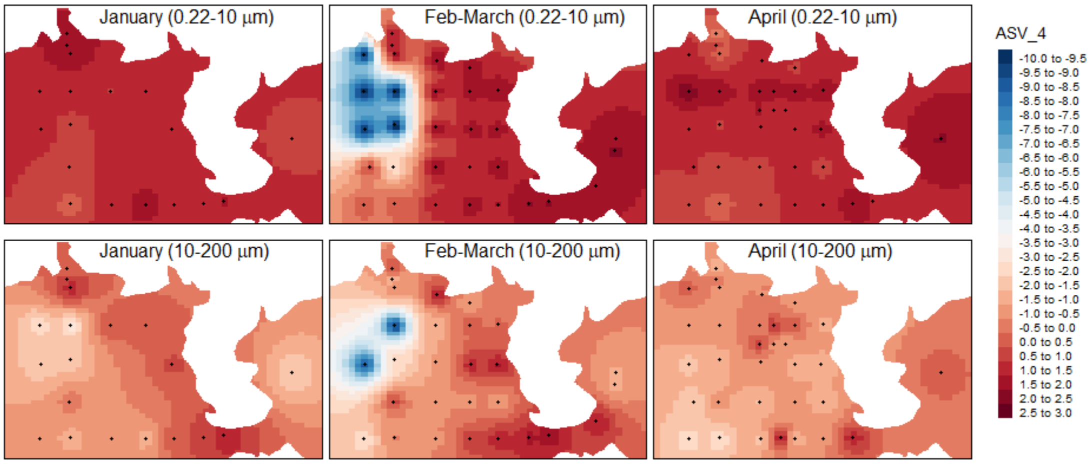


Figure 67 **Spatial-temporal distribution of ASV_4 (*Candidatus_Actinomarina*) mapped by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.**

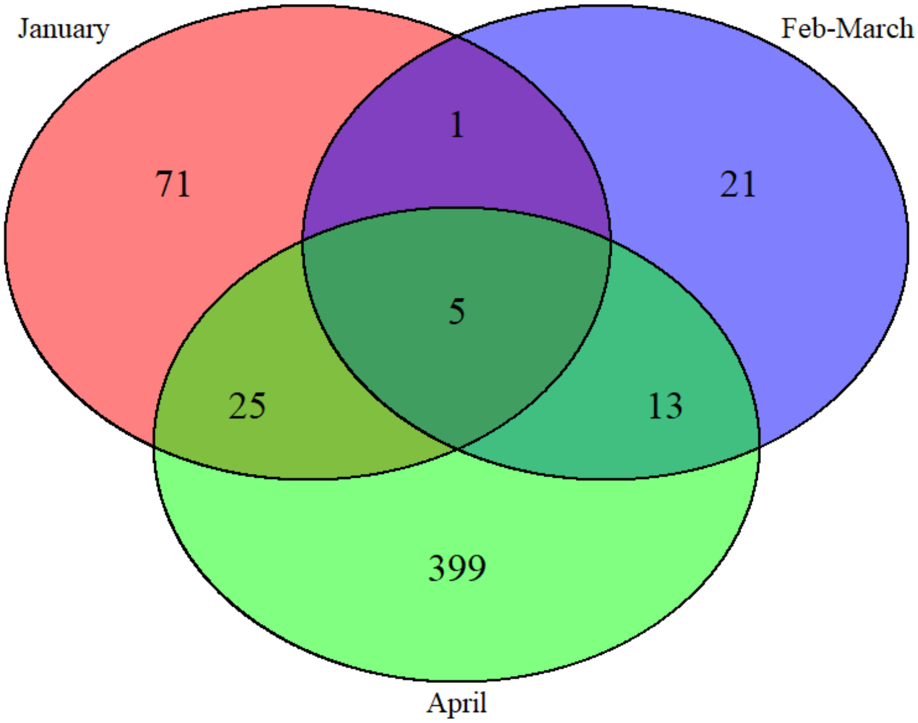


Figure 68 **Venn diagram of ASVs that were identified by DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010) as enriched in the large filtration size fraction in the surface samples from each expedition month.**

The ASVs that were enriched in the large filtration size revealed hypotheses for microbes that interacted with *P. globosa* colonies at different stages of the bloom. In all three expeditions, there were more ASVs that were enriched in the large filtration size than in the small (Figure 66). The ASVs that were enriched in the small filtration size were mostly from the SAR11_clade, the genera *NS4_marine_group*, *NS5_marine_group* and *Candidatus_Actinomarina*, and the archaea Marine_Group_II (Figure 66). These groups represent mostly free-living taxa that have small or very small-sized cells. For example, ASV_4, from the very small-sized *Candidatus_Actinomarina*, was enriched in the small filtration size in the January and April expeditions (Figure 67). Interestingly, only five ASVs were enriched in the large filtration size for all three expeditions (Figure 68). These five ASVs included one Alteromonadaceae, one Rhodobacteraceae, one Cytophagales, one Flavobacteriales (*Aquibacter*) and one Sandaracinaceae. The small number of ASVs enriched in the large filtration size for all three expeditions suggests that the microbes associated with large particles, including *P. globosa*, changed at different stages of the bloom. The 102 ASVs that were enriched in the large filtration size in the January expedition (Figure 68) are candidates for forming symbiotic relationships with *P. globosa* colonies at peak of bloom. Two examples are ASV_268 (*Colwellia aquaemaris)* and ASV_1173 (*Nioella sediminis)*, both of which were enriched in the large filtration size for only the January expedition (Figure 69). The April expedition had the most ASVs enriched in the large filtration size (*n* = 442; Figure 68), which is consistent with more ASVs being involved in the degradation of and associated with organic materials at this time point. These ASVs are candidates for involvement in the degradation of *P. globosa* colonies. For example, ASV_965 (*Roseovarius atlanticus*) is a member of the *Roseobacter* group, which are important organic carbon consumers after phytoplankton blooms (Buchan et al. 2014), and was enriched in the large filtration size for only the April expedition (Figure 70). Overall, many of the ASVs identified by this analysis are candidates for taxa that interact with *P. globosa* during its bloom but require further investigation to determine their role.

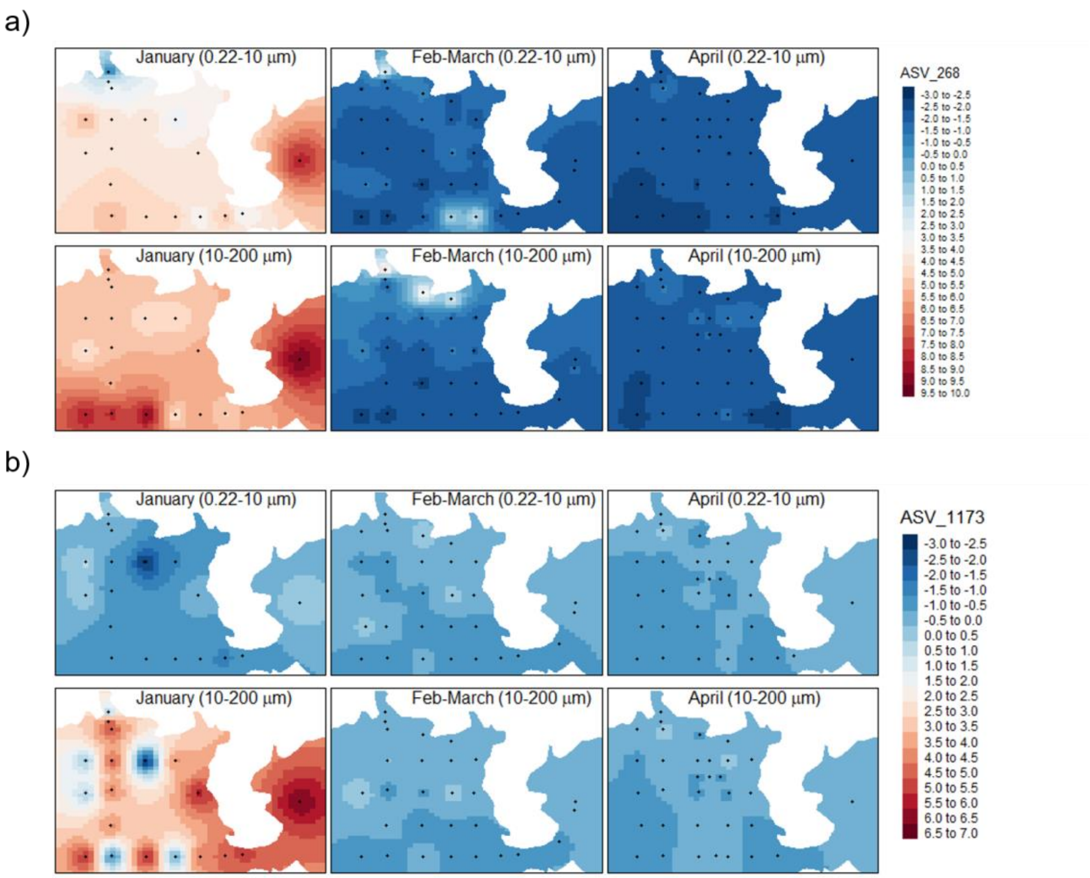


Figure 69 **Spatial-temporal distribution of a) ASV_268 (*Colwellia aquaemaris*) and b) ASV_1173 (*Nioella sediminis)* mapped by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.**

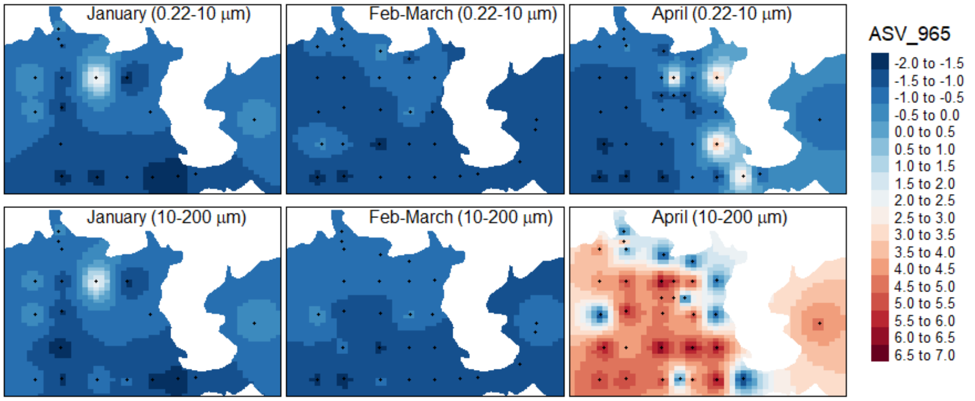


Figure 70 **Spatial-temporal distribution of ASV_965 (*Roseovarius atlanticus*) mapped by agglomerating samples collected from different water sampling depths at the same sampling location, expedition, and filtration size, and interpolating the centered log-ratio (clr) transformed values.**

5.3. The *P. globosa* colony microbiome

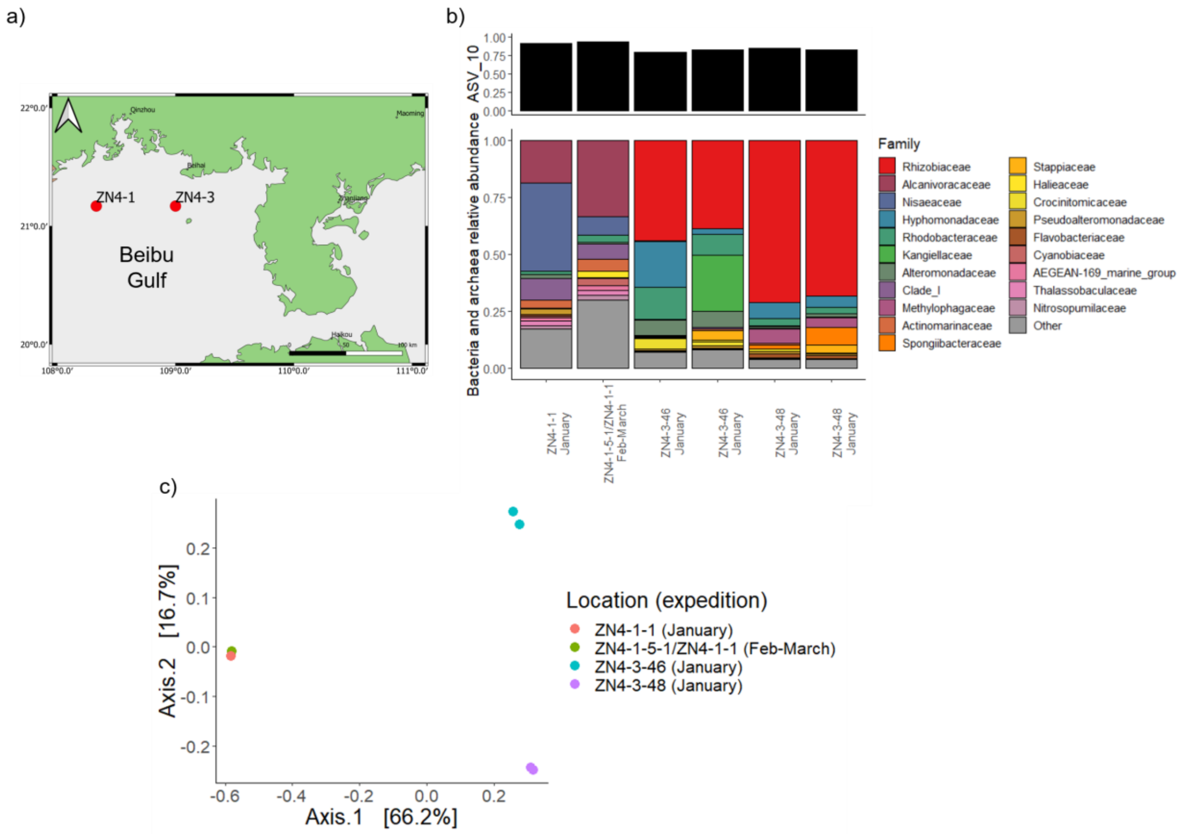


Figure 71 a) The two locations *P. globosa* colonies were sampled in the Beibu Gulf. b) The relative abundance of ASV_10 (*P. globosa*) and the families of bacteria and archaea (after removing chloroplast reads) in the colony samples. Samples are labelled by sampling location and expedition month. c) PCoA plot of the bacteria and archaea reads from the colony samples.

The *P. globosa* colony microbiome showed substantial variation by sampling location (Figure 71). The composition of the bacteria and archaea reads in the colony samples differed considerably between the ZN4-1 and ZN4-3 sampling locations. Colonies sampled from ZN4-3 location were dominated by the family Rhizobiaceae, while the two most abundant families in the colonies from the ZN4-1 location were Alcanivoracaceae and Nisaeaceae (Figure 71b). This was also reflected by the PCoA of the bacteria and archaea reads from the colony samples (Figure 71c), which was performed by rarefying to 837 reads/sample and ordinating the Bray-Curtis dissimilarity matrix. The first axis of the PCoA explained 66.2% of the variation and separated the samples by location (ZN4-1 vs. ZN4-3). The second axis of the PCoA explained 16.7% of the variation and separated the two different samples from the ZN4-3 location (ZN4-3-

46 vs. ZN4-3-48). Interestingly, the two samples from the ZN4-1 location clustered closely together despite being sampled during different expedition months (January vs. Feb-March).

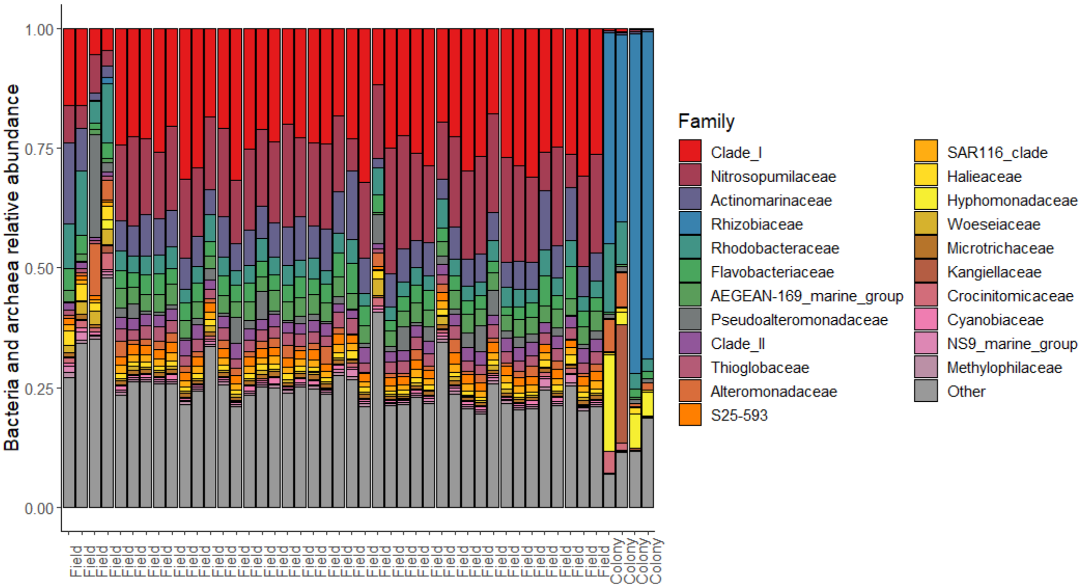


Figure 72 **The relative abundance of families of bacteria and archaea (after removing chloroplast reads) in the field and colony samples from the ZN4-3 location and the January expedition.**

ASVs that were likely to be associated with *P. globosa* colonies were identified by comparing the colony samples with the field samples from the same location and expedition. ASVs enriched in the colony samples represent microbes that were likely either attached to the outside of or located inside a *P. globosa* colony. The composition of the bacteria and archaea in the January expedition at the ZN4-3 location differed substantially between the field and colony samples (Figure 72). Six ASVs were identified by both DESeq2 and edgeR analysis as significantly enriched ($P < 0.05$ for DESeq2 and $FDR < 0.05$ for edgeR) in the two ZN4-3-46 colony samples compared to the ZN4-3 field samples (Table 5). Similarly, four ASVs were enriched in the two ZN4-3-48 colony samples compared to the ZN4-3 field samples (Table 6). Two ASVs, including *Lentilitoribacter donghaensis*, were enriched in both the ZN4-3-46 and ZN4-3-48 colony samples compared to the ZN4-3 field samples (Table 7).

Table 5 ASVs identified by DESeq2 and edgeR as enriched in the two ZN4-3-46 colony samples compared to the ZN4-3 field samples. The log2FoldChange values are from DESeq2.

ASV	log2 FoldChange	Family	Genus	Species
ASV_201	18.57593	Rhizobiaceae	Lentilitoribacter	Lentilitoribacter_donghaensis
ASV_794	16.41617	Hyphomonadaceae	Maricaulis	NA
ASV_1173	15.78868	Rhodobacteraceae	Nioella	Nioella_sediminis
ASV_2624	13.92337	Alteromonadaceae	Aestuariibacter	NA
ASV_630	10.92051	Kangiellaceae	Aliikangiella	NA
ASV_506	8.646819	Rhodobacteraceae	NA	NA

Table 6 ASVs identifed by DESeq2 and edgeR as enriched in the two ZN4-3-48 colony samples compared to the ZN4-3 field samples. The log2FoldChange values are from DESeq2.

ASV	log2 FoldChange	Family	Genus	Species
ASV_201	19.12964	Rhizobiaceae	Lentilitoribacter	Lentilitoribacter_donghaensis
ASV_1407	15.37035	Hyphomonadaceae	NA	NA
ASV_1467	15.21268	Methylophagaceae	Marine_Methylotrophic_Group_3	NA
ASV_506	8.00869	Rhodobacteraceae	NA	NA

Table 7 ASVs identified by DESeq2 and edgeR as enriched in both the ZN4-3-46 and ZN4-3-48 colony samples compared to the ZN4-3 field samples.

ASV	Family	Genus	Species
ASV_201	Rhizobiaceae	Lentilitoribacter	Lentilitoribacter_donghaensis
ASV_506	Rhodobacteraceae	NA	NA

Like the ZN4-3 location, the composition of the bacteria and archaea in the January and Feb-March expeditions at the ZN4-1 location differed between the field and colony samples (Figure 73). Two ASVs were significantly enriched in the ZN4-1 colony samples compared to the ZN4-1 field samples from the same expeditions, *Nisaea denitrificans* and *Alcanivorax borkumensis* (Table 8). Neither of these ASVs were enriched in the ZN4-3 colony samples, suggesting that the *P. globosa* colony microbiome may be location-dependent.

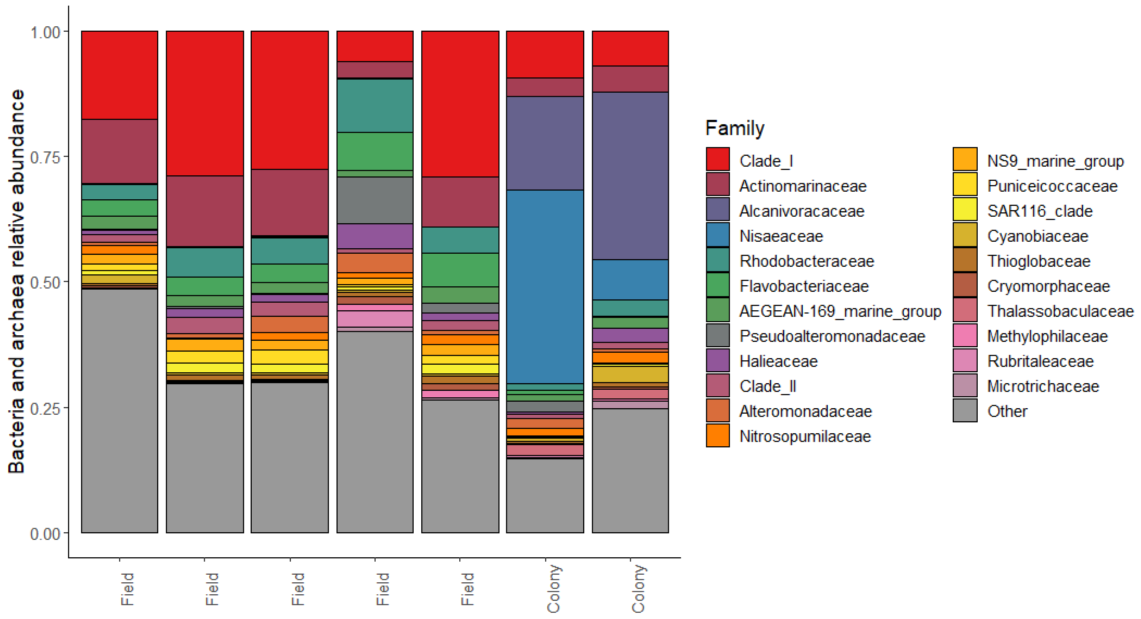


Figure 73 **The relative abundance of families of bacteria and archaea (after removing chloroplast reads) in the field and colony samples from the ZN4-1 location and the January and Feb-March expeditions.**

Table 8 **ASVs identifed by DESeq2 and edgeR as enriched in the two ZN4-1 colony samples compared to the ZN4-1 field samples. The log2FoldChange values are from DESeq2.**

ASV	log2 FoldChange	Family	Genus	Species
ASV_1345	15.67868	Nisaeaceae	Nisaea	Nisaea_denitrificans/ Nisaea_denitrificans_DSM_18348
ASV_390	15.45297	Alcanivoracaceae	Alcanivorax	Alcanivorax_borkumensis/ Alcanivorax_borkumensis_SK2

Finally, seven ASVs were enriched in the colony samples from at least one location and in the large filtration size from at least one expedition (Table 9). The enriched ASVs included the species *Lentilitoribacter donghaensis*, *Nioella sediminis* and *Alcanivorax borkumensis*. This provides support that these microbes are associated with large particles in the environment, but this association may not be specific to *P. globosa* colonies.

Table 9 **ASVs enriched in the colony samples from at least one sampling location and in the large filtration size from at least one expedition.**

ASV	Family	Genus	Species
ASV_201	Rhizobiaceae	Lentilitoribacter	Lentilitoribacter_donghaensis

ASV	Family	Genus	Species
ASV_390	Alcanivoracaceae	Alcanivorax	Alcanivorax_borkumensis/ Alcanivorax_borkumensis_SK2
ASV_630	Kangiellaceae	Aliikangiella	NA
ASV_794	Hyphomonadaceae	Maricaulis	NA
ASV_1173	Rhodobacteraceae	Nioella	Nioella_sediminis
ASV_1407	Hyphomonadaceae	NA	NA
ASV_1467	Methylophagaceae	Marine_Methylotrophic_ Group_3	NA

5.4. Discussion

In this chapter microbes that potentially interacted with *P. globosa* colonies during a bloom in the Beibu Gulf were identified using 16S amplicon sequencing. ASVs with potential interactions with *P. globosa* were identified by building interaction networks, identifying particle-associated microbes and exploring the *P. globosa* colony microbiome through sequencing entire *P. globosa* colonies from a naturally occurring bloom. Interestingly, while the *P. globosa* colonies had different bacterial compositions compared to seawater samples collected from the same locations, there was no core set of bacteria in the colonies sampled from different locations.

ASVs that were positively or negatively correlated with ASV_10 (*P. globosa*) during the bloom were identified using interaction networks. The ASVs that were positively correlated with ASV_10 (Table 3 and Table 4) had the same temporal dynamics as ASV_10 as they decreased in relative abundance from the January to the April expedition. These ASVs represent candidates for forming symbiotic or commensal interactions with *P. globosa*. In contrast, the ASVs that were negatively correlated with ASV_10 (Table 3 and Table 4) had the opposite temporal dynamics and increased in relative abundance from the January to the April expedition. These ASVs are candidates for forming competitive or predator-prey interactions with *P. globosa*. However, the observed correlations may also be explained by the abiotic requirements of the ASVs. For example, many of the ASVs that were positively correlated with *P. globosa* were chloroplasts (Table 3 and Table 4), which have similar nutrient requirements to *P. globosa*. Ultimately, further investigations, such as laboratory observations of interactions between different species, is required to determine the roles of these ASVs in the *P. globosa* bloom.

Particle-associated microbes were identified that were enriched in the large filtration size fraction for each of the three expeditions during the *P. globosa* bloom (Figure 66). Given the large size of the *P. globosa* colonies, the particle-associated microbes may have been associated with *P. globosa* colonies or other large particles in the environment, e.g. other phytoplankton. Some of the particle-associated microbes may also be represented by microbes that are not associated with particles, but instead form large colonies like *P. globosa*. Interestingly, the ASVs that were particle-associated changed with expedition (Figure 68), suggesting that the microbes that were associated with particles such as *P. globosa* colonies changed throughout the bloom. For example, ASV_268 (*Colwellia aquaemaris)* and ASV_1173 (*Nioella sediminis)*, both of which were identified as particle-associated for only the January expedition, are two examples of candidates for forming symbiotic relationships with *P. globosa* colonies at the peak of the bloom (Figure 69). The enrichment of *Colwellia aquaemaris* in the large filtration size during the January expedition is consistent with a previous study of a *P. antarctica* bloom, which found that particle-associated bacteria at 250m depth were enriched in *Colwellia* (Delmont et al. 2014). Overall, while the particle-associated microbes represent interesting hypotheses for microbes that interacted with *P. globosa* colonies, there is no evidence that these interactions were specific to *P. globosa* colonies and further investigation of their role in the bloom is necessary.

A variable *P. globosa* colony microbiome was discovered by sequencing entire colonies from multiple locations during a *P. globosa* bloom in the Beibu Gulf. The interactions between phytoplankton and bacteria are complex, involving the exchange of cofactors, micronutrients, proteins and signalling molecules, which results in mutualistic, commensal, competitive and antagonistic interactions (Behringer et al. 2018). The first step in understanding these interactions is determining the types of bacteria that are associated with the phytoplankton. Cultivation studies are an inherently biased approach for determining the bacteria associated with phytoplankton because most marine bacteria cannot be maintained using current culturing techniques (Rappe and Giovannoni 2003). Direct sequencing of the 16S rRNA genes from *P. globosa* colonies from a naturally occurring bloom was thus used to explore the *P. globosa* colony microbiome, which, to my knowledge, has never been done. In theory, the ASVs identified in the *P. globosa* colony samples represent microbes that were located inside the colonies or attached to the outside, which would provide some evidence for their interaction with *P. globosa*. Interestingly, while the *P. globosa* colonies had different bacterial compositions compared to seawater samples collected from the same locations (Figure 72 and Figure 73), there was no core set of bacteria that were enriched in the colonies sampled from different locations. The bacteria associated with the *P. globosa* colonies were location-dependent as the ZN4-3-46 and ZN4-3-48 colony samples were more similar to each other than the ZN4-1 colony samples (Figure 71). This suggests that the *P. globosa* colony microbiome may be flexible and dependent on the surrounding environment. Consistent with these results, previous studies have found that marine bacteria from diatom cultures and diatom-dominated blooms belong to a small number of genera compared to the total genera found in seawater (e.g. Baker and Kemp 2014). However, I did not identify a core set of bacteria shared by different strains, which has been previously observed for two species of coccolithophores (Green et al. 2015) and two diatom species (Behringer et al. 2018). These results are more consistent with the findings that 13 different cultures of the green algae *Ostreococcus tauri* contained varying bacteria with no core microbiome (Abby et al. 2014). Several species were identified that are worth further investigation of their interactions with *P. globosa*, including *Lentilitoribacter donghaensis*, *Nioella sediminis*, *Nisaea denitrificans* and *Alcanivorax borkumensis*, due to their presence in the *P. globosa* colony microbiome.

Chapter 6. Conclusions and future directions

6.1. Conclusions

In this thesis, the mechanisms underlying the recurrent *P. globosa* blooms in the Beibu Gulf were explored using 16S amplicon sequencing. In chapter two, I developed a bioinformatics pipeline for analyzing the amplicon sequencing data collected from field and colony samples during a *P. globosa* bloom. In chapter three, I identified the 16S rRNA gene as a suitable marker for tracking *P. globosa* due to low levels of intraspecific variation in the gene. Additionally, I found that the composition of the bacteria, archaea and eukaryotes in the Beibu Gulf was generally consistent with other studies in the Beibu Gulf and nearby regions. In chapter four, I observed distinct communities of bacteria, archaea and eukaryotes at three different time points during the *P. globosa* bloom. In many cases different groups had distinct temporal patterns, which could be related to their ecological niches during the bloom using descriptions from the literature as well as the temporal patterns of the environmental variables Additionally, I identified variation in the temporal response to the bloom at the ASV-level for the bacteria, archaea and chloroplast ASVs, which provides evidence for a previously underappreciated amount of variation of niches within taxonomic group. Using the spatial-temporal dynamics of *P. globosa*, other bacteria, archaea and phytoplankton and the environmental variables, I developed a preliminary model for the development and progression of the *P. globosa* bloom in the Beibu Gulf. Finally, in chapter five I identified bacteria that potentially interacted with *P. globosa* during the bloom by studying the *P. globosa* colony microbiome. Interestingly, while the *P. globosa* colonies had different bacterial compositions compared to seawater samples collected from the same locations, there was no core set of bacteria in the colonies sampled from different locations. This suggests that *P. globosa* may not have a core microbiome and that the bacteria that interact with *P. globosa* colonies vary with strain or location.

6.2. Future directions

There are several different types of data that could be collected to further advance our understanding of the mechanisms underlying the *P. globosa* looms in the Beibu Gulf. First, the strength of this analysis could be improved with the use of technical and biological replicates. The amplicon sequencing approach used here could be improved in the future from the use of longer reads, e.g. PacBio, to increase the taxonomic resolution of the ASVs (Martins et al. 2020). The results from this work will also become more useful as taxonomy databases become more complete and more ASVs can be annotated to the species level. Future expeditions will likely be planned to sample future *P. globosa* blooms in the Beibu Gulf at additional time points. Collecting additional samples prior to the start of the bloom (e.g. Lamy et al. 2010) will improve our understanding of the changes that occurred to trigger bloom formation. Additionally, collecting samples at a finer timescale, i.e. daily samples for a week at the peak of the bloom (Needham and Fuhrman 2016), will improve our understanding of the mechanisms underlying the peak of the bloom on a finer timescale. While changes in the community composition were observable using 16S amplicon sequencing data, this data did not provide information on the change in activity of different microbes at different stages of the bloom. Collecting metatranscriptomics data at different time points during the bloom would allow for insight into the functional roles of different species at different time points of the bloom (e.g. Nowinski et al. 2019). The support for the preliminary model could also be strengthened using data from experimental systems such as mesocosms that simulate the conditions during a *P. globosa* bloom in the Beibu Gulf. Additionally, the analysis of the *P. globosa* colony microbiome would benefit from an increase in sample size. Future expeditions will likely plan the collection of a larger sample size of colonies from a future *P. globosa* bloom to further investigate the bacteria identified here as potentially interacting with *P. globosa* colonies. Finally, further investigation of the microbes identified as potentially interacting with *P. globosa* could be performed by observing these interactions in a laboratory environment.

References

Abby, S. S., Touchon, M., Jode, A. De, Grimsley, N. and Piganeau, G. 2014. Bacteria in *Ostreococcus tauri* cultures – friends, foes or hitchhikers? - Front. Microbiol. 5: 1–10.

Alderkamp, A., Sintes, E. and Herndl, G. J. 2006. Abundance and activity of major groups of prokaryotic plankton in the coastal North Sea during spring and summer. - Aquat. Ecosyst. Heal. Manag. 45: 237–246.

Alderkamp, A. C., Buma, A. G. J. and Van Rijssel, M. 2007. The carbohydrates of *Phaeocystis* and their degradation in the microbial food web. - Biogeochemistry 83: 99–118.

Alonso-Saez, L. and Gasol, J. M. 2007. Seasonal variations in the contributions of different bacterial groups to the uptake of low-molecular-weight compounds in Northwestern Mediterranean coastal waters. - Appl. Environmantal Microbiol. 73: 3528–3535.

Amir, A., Mcdonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., Knightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A. and Knight, R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. - mSystems 2: e00191-16.

Anderson, D. M. 1994. Red Tides. - Sci. Am. 271: 52–58.

Arrieta, J. and Herndl, G. 2002. Changes in bacterial beta-glucosidase diversity during a coastal phytoplankton bloom. - Limnol. Oceanogr. 47: 594–599.

Baker, L. J. and Kemp, P. F. 2014. Exploring bacteria–diatom associations using single-cell whole genome amplification. - Aquat. Microb. Ecol. 72: 73–88.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S. O. N., Prjibelski, A. D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M. A. X. A. and Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its application to single-cell sequencing. - J. Comput. Biol. 19: 455–477.

Behringer, G., Ochsenkühn, M. A., Fei, C., Fanning, J., Koester, J. A. and Amin, S. A. 2018. Bacterial communities of diatoms display strong conservation across ctrains and time. - Front. Microbiol. 9: 1–15.

Bender, S. J., Moran, D. M., McIlvin, M. R., Zheng, H., McCrow, J. P., Badger, J., DiTullio, G. R., Allen, A. E. and Saito, M. A. 2018. Colony formation in *Phaeocystis antarctica*: Connecting molecular mechanisms with iron biogeochemistry. - Biogeosciences 15: 4923–4942.

Bertrand, E. M., Saito, M. A., Rose, J. M., Riesselman, C. R., Lohan, M. C., Noble, A. E., Lee, P. A. and DiTullio, G. R. 2007. Vitamin B12 and iron colimitation of phytoplankton growth in the Ross Sea. - Limnol. Oceanogr. 52: 1079–1093.

Brussaard, C. P. D., Kuipers, B. and Veldhuis, M. J. W. 2005. A mesocosm study of *Phaeocystis globosa* population dynamics: I. Regulatory role of viruses in bloom control. - Harmful Algae 4: 859–874.

Buchan, A., LeCleir, G. R., Gulvik, C. A. and González, J. M. 2014. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. - Nat. Rev. Microbiol. 12: 686–698.

Bullard, J. H., Purdom, E., Hansen, K. D. and Dudoit, S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. - BMC Bioinformatics 11: 1–13.

Bumgarner, R. 2013. DNA microarray: types, applications and their future. - Curr. Protoc. Mol. Biol. 101: 1–17.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. and Holmes, S. P. 2016. DADA2: high resolution sample inference from Illumina amplicon data. - Nat. Methods 13: 581–583.

Callahan, B. J., McMurdie, P. J. and Holmes, S. P. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. - ISME J. 11: 2639–2643.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. 2009. BLAST+: architecture and applications. - BMC Bioinformatics 10: 1–9.

Cariou, V., Casotti, R., Birrien, J. L. and Vaulot, D. 1994. The initiation of *Phaeocystis* colonies. - J. Plankton Res. 16: 457–470.

Chen, Y. Q., Wang, N., Zhang, P., Zhou, H. and Qu, L. H. 2002. Molecular evidence identifies bloom-forming Phaeocystis (Prymnesiophyta) from coastal waters of southeast China as *Phaeocystis globosa*. - Biochem. Syst. Ecol. 30: 15–22.

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., Brown, C. T., Porras-alfaro, A., Kuske, C. R. and Tiedje, J. M. 2014. Ribosomal Database Project: data and tools for high throughput rRNA analysis. - Nucleic Acids Res. 42: D633–D642.

Coltharp, C. and Xiao, J. 2012. Superresolution microscopy for microbiology. - Cell. Microbiol. 14: 1808–1818.

D'Onofrio, A., Crawford, J. M., Stewart, E. J., Witt, K., Gavrish, E., Epstein, S., Clardy, J. and Lewis, K. 2010. Siderophores from neighboring organisms promote the growth of uncultured bacteria. - Chem. Biol. 17: 254–264.

Decelle, J., Romac, S., Stern, R. F., Bendif, E. L. M., Zingone, A., Audic, S., Guiry, M. D., Guillou, L., Tessier, D., Le Gall, F., Gourvil, P., Dos Santos, A. L., Robert, I., Vaulot, D., De Vargas, C. and Christen, R. 2015. PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. - Mol. Ecol. Resour. 15: 1435–1445.

Delmont, T. O., Hammar, K. M., Ducklow, H. W., Yager, P. L. and Post, A. F. 2014. *Phaeocystis antarctica* blooms strongly influence bacterial community structures in the Amundsen Sea polynya. - Front. Microbiol. 5: 1–13.

Delwiche, C. F. 1999. Tracing the thread of plastid diversity through the tapestry of life. - Am. Nat. 154: S164-177.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. - Bioinformatics 26: 2460–2461.

Edgar, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. - Nat. Commun. 10: 996–998.

Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. - bioRxiv: 081257.

Edgar, R. C. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. - Bioinformatics 34: 2371–2375.

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H. and Sogin, M. L. 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. - ISME J. 9: 968–979.

Falkowski, P. G. and Raven, J. A. 1997. Aquatic Photosynthesis. - Blackwell Scientific Publications.

Falkowski, P. G., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O. and Taylor, F. J. R. 2004. The evolution of modern eukaryotic phytoplankton. - Science (80-.). 305: 354–360.

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J. and Huttenhower, C. 2012. Microbial co-occurrence relationships in the human microbiome. - PLoS Comput. Biol. 8: e1002606.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. - Evolution (N. Y). 39: 783–791.

Field, C. B., Behrenfeld, M. J., Randerson, J. T. and Falkowski, P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. - Science (80-.). 281: 237–240.

Fisher, M. M. and Triplett, E. W. 1999. Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. - Appl. Environ. Microbiol. 65: 4630–4636.

Friedman, J. and Alm, E. J. 2012. Inferring correlation networks from genomic survey data. - PLoS Comput. Biol. 8: e1002687.

Galand, P. E., Gutie, C., Massana, R., Gasol, J. M. and Casamayor, E. O. 2010. Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). - Limnol. Oceanogr. 55: 2117–2125.

Gilbert, J. A., Jansson, J. K. and Knight, R. 2014. The Earth Microbiome project: successes and aspirations. - BMC Biol. 12: 1–4.

Golyshin, P. N., Chernikova, T. N., Abraham, W.-R., Timmis, K. N. and Yakimov, M. M. 2002. Oleiphilaceae fam. nov., to include *Oleiphilus messinensis* gen. nov., sp. nov., a novel marine bacterium that obligately utilizes hydrocarbons. - Int. J. Syst. Evol. Microbiol. 52: 901–911.

Green, D. H., Echavarri-bravo, V., Brennan, D. and Hart, M. C. 2015. Bacterial diversity associated with the coccolithophorid algae Emiliania huxleyi and Coccolithus pelagicus f . braarudii. - Biomed Res. Int.: 194540.

Gupta, S., Mortensen, M. S., Schjørring, S., Trivedi, U., Vestergaard, G., Stokholm, J., Bisgaard, H., Krogfelt, K. A. and Sørensen, S. J. 2019. Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. - Commun. Biol. 2: 1–7.

Han, Q., Huang, X., Xing, Q. and Shi, P. 2012. A review of environment problems in the coastal sea of South China. - Aquat. Ecosyst. Heal. Manag. 15: 108–117.

Hothorn, T., Bretz, F. and Westfall, P. 2008. Simultaneous inference in general parametic models. - Biometrical J. 50: 346–363.

Hugerth, L. W. and Andersson, A. F. 2017. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. - Front. Microbiol. 8: 1–22.

Hughes, J. B. and Hellmann, J. J. 2005. The application of rarefaction techniques to molecular inventories of microbial diversity. - In: Methods in Enzymology. pp. 292–308.

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L. and Birol, I. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. - Genome Res. 27: 768–777.

Kajitani, R., Yoshimura, D., Okuno, M., Minakuchi, Y., Kagoshima, H., Fujiyama, A., Kubokawa, K., Kohara, Y., Toyoda, A. and Itoh, T. 2019. Platanus-allee is a de novo haplotype assembler enabling a comprehesnive access to divergent heterozygous regions. - Nat. Commun. 10: 1–15.

Katoh, K. and Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. - Mol. Biol. Evol. 30: 772–780.

Kertesz, M. A., Kawasaki, A. and Stolz, A. 2019. Aerobic Hydrocarbon-Degrading *Alphaproteobacteria: Sphingomonadales*. - In: Taxonomy, Genomics and Ecophysiology of Hydrocarbon-Degrading Microbes, Handbook of Hydrocarbon and Lipid Microbiology. Springer Nature, pp. 105–124.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. 2018. MEGA X: Molecular evolutionary genetics analysis across computing platforms. - Mol. Biol. Evol. 35: 1547–1549.

Lagier, J. C., Edouard, S., Pagnier, I., Mediannikov, O., Drancourt, M. and Raoult, D. 2015. Current and past strategies for bacterial culture in clinical microbiology. - Clin. Microbiol. Rev. 28: 208–236.

Lamy, D., Obernosterer, I., Laghdass, M., Artigas, L. F., Breton, E., Grattepanche, J. D., Lecuyer, E., Degros, N., Lebaron, P. and Christaki, U. 2010. Temporal changes of major bacterial groups and bacterial heterotrophic activity during a Phaeocystis globosa bloom in the eastern English Channel. - Aquat. Microb. Ecol. 58: 95–107.

Lange, M., Chen, Y. Q. and Medlin, L. K. 2002. Molecular genetic delineation of *Phaeocystis* species (Prymnesiophyceae) using coding and non-coding regions of nuclear and plastid genomes. - Eur. J. Phycol. 37: 77–92.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., Mcgettigan, P. A., Mcwilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. and Higgins, D. G. 2007. Clustal W and Clustal X version 2.0. - Bioinformatics 23: 2947–2948.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. - Bioinformatics 25: 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. - Bioinformatics 25: 2078–2079.

Li, N., Zhao, H., Jiang, G., Xu, Q., Tang, J., Li, X., Wen, J., Liu, H., Tang, C., Dong, K. and Kang, Z. 2020a. Phylogenetic responses of marine free-living bacterial community to *Phaeocystis globosa* bloom in Beibu Gulf, China. - Front. Microbiol. 11: 1–13.

Li, J., Gu, L., Bai, S., Wang, J., Su, L., Wei, B., Zhang, L. and Fang, J. 2020b. Characterization of particle-associated and free-living bacterial and archaeal communities along the water columns of the South China Sea. - Biogeosciences in press.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J. C., Roux, S., Vincent, F. and Bittner, L. 2015. Determinants of community structure in the grobal plankton interactome. - Science (80-.). 348: 1262073.

Liss, P. S., Malin, G., Turner, S. M. and Holligan, P. M. 1994. Dimethyl sulphide and *Phaeocystis*: A review. - J. Mar. Syst. 5: 41–53.

Litchman, E., de Tezanos Pinto, P., Edwards, K. F., Klausmeier, C. A., Kremer, C. T. and Thomas, M. K. 2015. Global biogeochemical impacts of phytoplankton: a trait-based perspective. - J. Ecol. 103: 1384–1396.

Liu, W. T., Marsh, T. L., Cheng, H. and Forney, L. J. 1997. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. - Appl. Environ. Microbiol. 63: 4516–4522.

Liu, H., Zhang, C. L., Yang, C., Chen, S., Cao, Z., Zhang, Z. and J, T. 2017. Marine Group II dominates planktonic archaea in water column of the Northeastern South China Sea. - Front. Microbiol. 8: 1–11.

Long, J. D., Smalley, G. W., Barsby, T., Anderson, J. T. and Hay, M. E. 2007. Chemical cues induce consumer-specific defenses in a bloom-forming marine phytoplankton. - Proc. Natl. Acad. Sci. U. S. A. 104: 10512–10517.

Love, M. I., Huber, W. and Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. - Genome Biol. 15: 1–21.

Malaei Tavana, H., Behpoor, S., Changizi, M. and H, K. 2008. Investigate the reinforcing factors in forming and occurrence of harmful algal bloom. - Natl. Conf. human, Environ. Sustain. Dev.

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R. and Peddada, S. D. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. - Microb. Ecol. Heal. Dis. 26: 1–7.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. - EMBnet.journal 17: 10–12.

Martinez-Garcia, M., Brazel, D. M., Swan, B. K., Arnosti, C., Chain, P. S. G., Reitenga, K. G., Xie, G., Poulton, N. J., Gomez, M. L., Masland, D. E. D., Thompson, B., Bellows, W. K., Ziervogel, K., Lo, C. C., Ahmed, S., Gleasner, C. D., Detter, C. J. and Stepanauskas, R. 2012. Capturing single cell genomes of active polysaccharide degraders: An unexpected contribution of *Verrucomicrobia*. - PLoS One 7: 1–11.

Martins, L. D. O., Charles, I. G., Page, A. J., Mather, A. E. and Charles, I. G. 2020. Taxonomic resolution of the ribosomal RNA operon in bacteria: implications for its use with long-read sequencing. - NAR Genomics Bioinforma. 2: 1–7.

Mcdonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., Desantis, T. Z., Probst, A., Andersen, G. L., Knight, R. and Hugenholtz, P. 2011. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. - ISME J. 6: 610–618.

McMurdie, P. J. and Holmes, S. 2013. Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. - PLoS One 8: e61217.

McMurdie, P. J. and Holmes, S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. - PLoS Comput. Biol. in press.

Medlin, L. and Zingone, A. 2007. A taxonomic review of the genus *Phaeocystis*. - Biogeochemistry 83: 3–18.

Mur, L. R., Skulberg, O. M. and Utkilen, H. 1999. Cyanobacteria in the environment. - In: Chorus, I. and Bartram, J. (eds), Toxic Cyanobacteria in Water: A guide to their public health consequences, monitoring and management. WHO, in press.

Mußmann, M., Pjevac, P., Krüger, K. and Dyksma, S. 2017. Genomic repertoire of the Woeseiaceae/JTB255, cosmopolitan and abundant core members of microbial communities in marine sediments. - ISME J. 11: 1276–1281.

Muyzer, G., De Waal, E. C. and Uitterlinden, A. G. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analyis of polymerase chain reaction-amplified genes coding for 16S rRNA. - Appl. Environmantal Microbiol. 59: 695–700.

Nearing, J. T., Douglas, G. M., Comeau, A. M. and Langille, M. G. I. 2018. Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. - PeerJ 6: e5364.

Needham, D. M. and Fuhrman, J. A. 2016. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. - Nat. Microbiol. 1: 1–7.

Nejstgaard, J. C., Tang, K. W., Steinke, M., Dutz, J., Koski, M., Antajan, E. and Long, J. D. 2007. Zooplankton grazing on *Phaeocystis*: a quantitative review and future challenges. - Biogeochemistry 83: 147–172.

Noordkamp, D. J. B., Gieskes, W. W. C., Gottschal, J. C., Forney, L. J. and van Rijssel, M. 2000. Acrylate in *Phaeocystis* colonies does not affect the surrounding bacteria. - J. Sea Res. 43: 287–296.

Nowinski, B., Smith, C. B., Thomas, C. M., Esson, K., Iii, R. M., Preston, C. M., Birch, J. M., Scholin, C. A., Huntemann, M., Clum, A., Foster, B., Foster, B., Roux, S., Palaniappan, K., Varghese, N., Mukherjee, S., Reddy, T. B. K., Daum, C., Copeland, A., Chen, I. A., Ivanova, N. N., Kyrpides, N. C., Glavina, T., Whitman, W. B., Kiene, R. P., Eloe-fadrosh, E. A. and Moran, M. A. 2019. Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. - Sci. Data 6: 1–7.

Nye, K. J., Fallon, D., Gee, B., Messer, S., Warren, R. E. and Andrews, N. 1999. A comparison of blood agar supplemented with NAD with plain blood agar and chocolated blood agar in the isolation of *Streptococcus pneumoniae* and *Haemophilus influenzae* from sputum. - J. Med. Microbiol. 48: 1111–1114.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O’Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. and Wagner, H. 2019. vegan: Community Ecology Package. R package version 2.5-6. in press.

Orsi, W. D., Smith, J. M., Wilcox, H. M., Swalwell, J. E., Carini, P., Worden, A. Z. and Santoro, A. E. 2015. Ecophysiology of uncultivated marine euryarchaea is linked to particulate organic matter. - ISME J. 9: 1747–1763.

Orsi, W. D., Smith, J. M., Liu, S., Liu, Z., Sakamoto, C. M., Wilken, S., Poirier, C., Richards, T. A., Keeling, P. J., Worden, A. Z. and Santoro, A. E. 2016. Diverse, uncultivated bacteria and archaea underlying the cycling of dissolved protein in the ocean. - ISME J. 10: 2158–2173.

Pace, N. R., Stahl, D. Q., Lane, D. J. and Olsen, G. J. 1985. Analyzing natural microbial populations by rRNA sequences. - ASM News 51: 4–12.

Paulson, J. N., Stine, O. C., Bravo, H. C. and Pop, M. 2013. Differential abundance analysis for microbial marker-gene surveys. - Nat. Methods 10: 1200–1202.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. 2011. Scikit-learn: machine learning in python. - J. Mach. Learn. Res. 12: 2825–2830.

Peperzak, L., Colijn, F., Gieskes, W. W. C. and Peeters, J. C. H. 1998. Development of the diatom-*Phaeocystis* spring bloom in the Dutch coastal zone of the North Sea: The silicon depletion versus the daily irradiance threshold hypothesis. - J. Plankton Res. 20: 517–537.

Pinheiro, J., Bates, D., DebRoy, S., DD, S. and R Core Team 2020. nlme: Linear and Nonlinear Mixed Effects Models.: R package version 3.1-148.

QGIS.org 2020. QGIS Geographic Information System. - Open Source Geospatial Found. Proj.

Qingchun, Z., Zhuang, N., Jinxiu, W., Chao, L., Fanzhou, K., Xiaokun, H., Jiayu, Z. and Yu, R. 2020. Development of high-resolution chloroplast markers for intraspecific phylogeographic studies of *Phaeocystis globosa**. - J. Oceanol. Limnol. in press.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. - Nucleic Acids Res. 41: D590–D596.

R Core Team 2019. R: A language and environment for statistical computing.

Rappe, M. S. and Giovannoni, S. J. 2003. The uncultured microbial majority. - Annu. Rev. Microbiol. 57: 369–94.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., Mcvean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. 2011. Detecting novel association in large data sets. - Science (80-.). 334: 1518–1524.

Riegman, R., Noordeloos, A. A. M. and Gerhard, C. C. 1992. *Phaeocystis* blooms and eutrophication of the continental coastal zones of the North Sea. - Mar. Biol. 112: 479–484.

Ringnér, M. 2008. What is principal component analysis ? - Comput. Biol. 26: 303–304.

Robinson, M. D., Mccarthy, D. J. and Smyth, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. - Bioinformatics 26: 139–140.

Ronquist, F., TEslenko, M., Mark, P. V. D., Ayres, D., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A. and Huelsenbeck, J. P. 2012. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. - Softw. Syst. Evol. 61: 539–542.

Rousseau, V., Chrétiennot-Dinet, M. J., Jacobsen, A., Verity, P. and Whipple, S. 2007. The life cycle of *Phaeocystis*: state of knowledge and presumptive role in ecology. - Biogeochemistry 83: 29–47.

Rousseau, V., Lantoine, F., Rodriguez, F., LeGall, F., Chrétiennot-Dinet, M. J. and Lancelot, C. 2013. Characterization of *Phaeocystis globosa* (Prymnesiophyceae), the blooming species in the Southern North Sea. - J. Sea Res. 76: 105–113.

Ruan, Q., Dutta, D., Schwalbach, M. S., Steele, J. A., Fuhrman, J. A. and Sun, F. 2006. Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. - Bioinformatics 22: 2532–2538.

Santoro, A. E., Richter, R. A. and Dupont, C. L. 2019. Planktonic Marine Archaea. - Ann. Rev. Mar. Sci. 11: 131–158.

Sboner, A., Mu, X. J., Greenbaum, D., Auerbach, R. and Gerstein, M. . 2011. The real cost of sequencing: higher than you think! - Genome Biol. 12: 1–10.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Horn, D. J. Van and Weber, C. F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. - Appl. Environmantal Microbiol. 75: 7537–7541.

Schoemann, V., Becquevort, S., Stefels, J., Rousseau, V. and Lancelot, C. 2005. *Phaeocystis* blooms in the global ocean and their controlling mechanisms: a review. - J. Sea Res. 53: 43–66.

Sellner, K. G., Doucette, G. J. and Kirkpatrick, G. J. 2003. Harmful algal blooms: causes, impacts and detection. - J. Ind. Microbiol. Biotechnol. 30: 383–406.

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 Res. 13: 2498–2504.

Shetty, S. and Lahti, L. 2019. microbiomeutilities: An R package with utility functions for the microbiome R package. in press.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., Mcwilliam, H., Remmert, M., Soding, J., Thompson, J. D. and Higgins, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. - Mol. Syst. Biol. 7: 1–6.

Smith, D. R., Arrigo, K. R., Alderkamp, A.-C. and Allen, A. E. 2014. Massive difference in synonymous substitution rates among mitochondrial, plastid, and nuclear genes of *Phaeocystis* algae. - Mol. Phylogenet. Evol. 71: 36–40.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. and Herndl, G. J. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere.” - Proc. Natl. Acad. Sci. U. S. A. 103: 12115–12120.

Stackebrandt, E. and Goebel, B. M. 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the pesent species definition in bacteriology. - Int. J. Syst. Bacteriol. 44: 846–849.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. - Bioinformatics 30: 1312–1313.

Stefels, J. and Dijkhuizen, L. 1996. Characteristics of DMSP-lyase in *Phaeocystis* sp. (Prymnesiophyceae). - Mar. Ecol. Prog. Ser. 131: 307–313.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J. and Rambaut, A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. - Virus Evol. 4: 1–5.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-castillo, F. M., Costea, P. I., Cruaud, C., Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F. and Lepoivre, C. 2015. Structure and function of the global ocean microbiome. - Science (80-.). 348: 1261359.

Tada, Y., Taniguchi, A., Nagao, I., Miki, T., Uematsu, M., Tsuda, A., Hamasaki, K. and Icrobiol, A. P. P. L. E. N. M. 2011. Differing growth responses of major phylogenetic groups of marine bacteria to natural phytoplankton blooms in the western North Pacific Ocean. - Appl. Environmantal Microbiol. 77: 4055–4065.

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., Kassabgy, M., Huang, S., Mann, A. J., Waldmann, J., Weber, M., Klindworth, A., Otto, A., Lange, J., Bernhardt, J., Reinsch, C., Hecker, M., Peplies, J., Bockelmann, F. D., Callies, U., Gerdts, G., Wichels, A., Wiltshire, K. H., Glöckner, F. O., Schweder, T. and Amann, R. 2012. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. - Science (80-.). 336: 608–611.

Tennekes, M. 2018. tmap: Thematic Maps in R. - J. Stat. Softw. 84: 1–39.

Turner, S. M., Nightingale, P. D., Broadgate, W. and Liss, P. S. 1995. The distribution of dimethyl sulphide and dimethylsulphoniopropionate in Antarctic waters and sea ice. - Deep. Res. Part II 42: 1059–1080.

van Rijssel, M., Alderkamp, A. C., Nejstgaard, J. C., Sazhin, A. F. and Verity, P. G. 2007. Haemolytic activity of live *Phaeocystis pouchetii* during mesocosm blooms. - Biogeochemistry 83: 189–200.

Verity, P. G., Brussaard, C. P., Nejstgaard, J. C., Leeuwe, M. A. Van, Lancelot, C. and Medlin, L. K. 2007. Current understanding of *Phaeocystis* ecology and biogeochemistry, and perspectives for future research. - Biogeochemistry 83: 311–330.

Větrovský, T. and Baldrian, P. 2013. The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. - PLoS One 8: 1–10.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. - Appl. Environ. Microbiol. 73: 5261–5267.

Watts, S. C., Ritchie, S. C., Inouye, M. and Holt, K. E. 2019. FastSpar: rapid and scalable correlation estimation for compositional data. - Bioinformatics 35: 1064–1066.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L. C., Xu, Z. Z., Ursell, L., Alm, E. J., Birmingham, A., Cram, J. A., Fuhrman, J. A., Raes, J., Sun, F., Zhou, J. and Knight, R. 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. - ISME J. 10: 1669–1681.

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R. and Knight, R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. - Microbiome 5: 1–18.

Woese, C. R. and Fox, G. E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. - Proc. Natl. Acad. Sci. U. S. A. 74: 5088–5090.

Wuchter, C., Abbas, B., Coolen, M. J. L., Herfort, L., Bleijswijk, J. Van, Timmers, P., Strous, M., Teira, E., Herndl, G. J., Middelburg, J. J., Schouten, S. and Damste, J. S. S. 2006. Archaeal nitrification in the ocean. - Proc. Natl. Acad. Sci. 103: 12317–12322.

Xiaokun, H., Qingchun, Z., Zhenfan, C., Fanzhou, K., Jinxiu, W. and Rencheng, Y. 2019. Genetic diversity of *Phaeocystis globosa* strains isolated from the Beibu Gulf, the South China Sea. - Oceanol. Limnol. Sin. 50: 601–610.

Xu, Y., Zhang, T. and Zhou, J. 2019. Historical occurrence of algal blooms in the northern Beibu Gulf of China and implications for future trends. - Front. Microbiol. 10: 1–13.

Zhang, Y., Li, J., Cheng, X., Luo, Y., Mai, Z. and Zhang, S. 2018. Community differentiation of bacterioplankton in the epipelagic layer in the South China Sea. - Ecol. Evol. 8: 4932–4948.

Zhong, C. 2015. Protect the last clean ocean in China: Beibu Gulf. - Contemp. Guangxi in press.

Zohdi, E. and Abbaspour, M. 2019. Harmful algal blooms (red tide): a review of causes, impacts and approaches to monitoring and prediction. - Int. J. Environ. Sci. Technol. 16: 1789–1806.

Appendix A.

DADA2 read quality plots

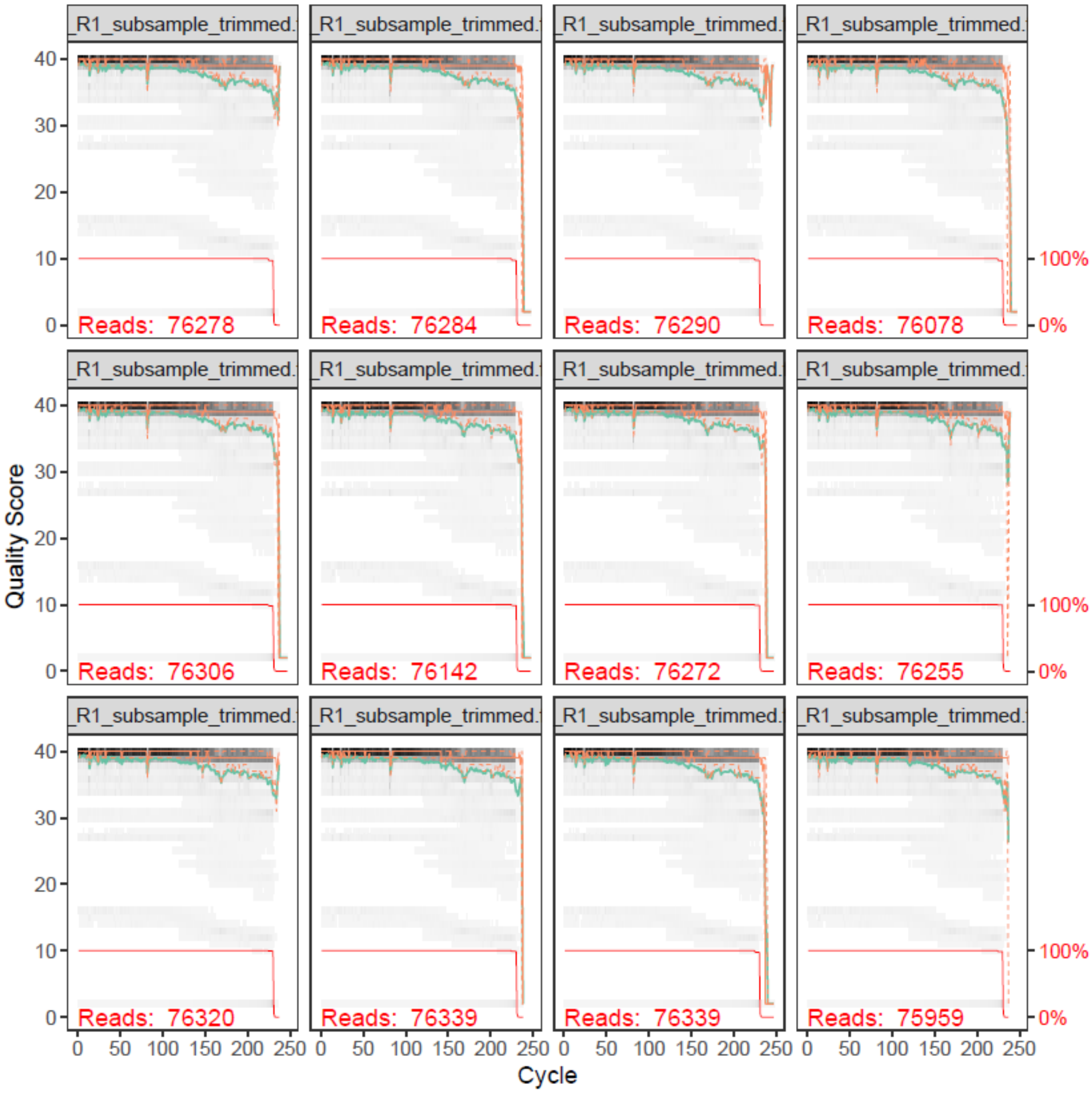


Figure A1 **Example output from the DADA2 plotQualityProfile() function for the forward reads from twelve samples. The gray-scale is a heatmap of the frequency of each quality score at each base position. The green lines represent the mean quality score at each position, the orange lines show the quartiles of the quality score distribution and the red lines shows the scaled proportion of reads that extend to at least that position.**

Appendix B.

P. globosa strains 16S rDNA sequence assembly methods

Forty-nine *P. globosa* strains were selected from bloom water samples using a micropipette under an inverted light microscope Nikon TS2 (Tokyo, Japan), washed and transferred to culture dishes, which were subsequently maintained at 20ºC under a 12:12 h light:dark cycle. DNA was extracted for whole-genome sequencing from each of the cultured strains using the HP Plant DNA kit (Omega, USA) according to the manufacturer’s instructions.The quality of the extracted DNA was monitored on 1% agarose gels and the DNA concentrations were measured using the Qubit® DNA Assay Kit with the Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). Sequencing libraries were generated using NEBNext DNA Library Prep Kit following the manufacturer's recommendations with indices added to each sample. Briefly, the genomic DNA was randomly fragmented to a size of 350 bp by shearing, DNA fragments were end polished, A-tailed and ligated with the NEBNext adapter for Illumina sequencing and further PCR enrichment was performed using generic P5 and P7 oligo adapters. The PCR products were purified (AMPure XP system) and the resulting libraries were analyzed for size distribution using the Agilent 2100 Bioanalyzer system and quantified using real-time PCR. Libraries were sequenced on the NovaSeq Illumina platform (Illumina, San Diego, CA, USA) using 2x150 paired-end reads at a depth of 35 million reads/sample (~100X coverage).

The 16S rRNA gene sequences were assembled using the whole-genome sequencing reads for each strain. The raw reads were filtered by 1) removing reads with >10% unidentified nucleotides (N), 2) removing reads with > 50% bases with Phred score < 5, 3) removing reads with > 10 nucleotides aligned to the adapter sequences and 4) removing PCR duplicates. The filtered reads were assembled using Platanus-allee v. 2.0.2 (Kajitani et al. 2019), ABySS v. 2.1.5 (Jackman et al. 2017) and SPAdes (Bankevich et al. 2012). Next, the chloroplast DNA (cpDNA) scaffolds were identified using BLAST (Camacho et al. 2009) against the NC_021637.1 *P. globosa* chloroplast genome. Finally, the 16S rDNA sequences were determined by two methods: 1) BLAST analysis of the NC_021637.1 16S rDNA sequence against the assembled cpDNA scaffolds for each strain or 2) alignment of the filtered reads for each strain to the NC_021637.1 16S rDNA sequence using BWA v. 0.7.17 (Li and Durbin 2009), extraction of the aligned reads using SAMtools v. 1.9 (Li et al. 2009) and assembly with SPAdes (Bankevich et al. 2012).