Mapping Microbial Genomes through Time and Space in the Arctic Ocean

William Boulton

School of Computing Sciences

University of East Anglia

Student Number 100358520

For the degree of

Doctor of Philosophy

April 2026

This thesis is dedicated to my wife Izzi, and to my parents, John and Purabi

Acknowledgements

This work would not have been possible without the support of my supervisory team, Vincent Moulton, Thomas Mock, and Richard Leggett. I am also very grateful to Anthony Duncan and Andrew Toseland, and others in the Mock Lab for their help and advice, and to the members of the JGI, particularly Asaf Salamov, Igor Grigoriev, Frederick Schultz, Tanja Wokye and Sara Calhoun, all of whom contributed valuable advice, and provided computational resources from the NERSC facility and JGI.

Also I am thankful to all those who participated in the MOSAiC expedition, mentioned in the MOSAiC extended acknowledgement [1], and to Allison Fong, Katja Metfies, and Klaus Valentin, who kindly hosted my stay at AWI.

This work was supported by the Natural Environment Research Council and the ARIES Doctoral Training Partnership [grant number NE/S007334/1], and the US Department of Energy Joint Genome Institute (DoE JGI), who funded the sequencing project (The International Arctic Ice Drift Experiment MOSAiC: Seasonal changes of microbial communities across the Arctic Ocean, Proposal ID: 505419), and to international collaborators including the German Federal Ministry for Education and Research (BMBF) through financing the Alfred-Wegener-Institut Helmholtz Zentrum für Polar und Meeresforschung (AWI) and the Polarstern expedition PS122 (grant nos. MOSAiC20192020 and AWI_PS122_00).

Abstract

The Arctic Ocean acts as a leading indicator for the microbial response to the changing ocean climate, and hosts a wealth of microbial diversity, much of which is of unknown composition. Microbes play an important role as primary producers and base of the marine food web, in the carbon cycle, and in other biogeochemical cycles. They are also a source of novel genes, many of which are of interest to the biotechnology industry. It is therefore important to understand the effect of changing environmental conditions on these organisms and their genes. However, due to the extreme inaccessibility of the central Arctic, there is a gap in our understanding of microbial communities in the region.

To study these communities, we analyse datasets from the Multidisciplinary Drifting Observatory for Study of the Arctic Climate (MOSAiC) expedition. MOSAiC was the longest ever Arctic expedition, and the first to observe the Arctic throughout an annual cycle.

The aim of this thesis is to catalogue the Arctic microbial diversity captured by MOSAiC in the form of metagenomes, i.e. DNA from environmental samples. We focus on metagenome-assembled genomes (MAGs); putative semi-complete genomes derived from metagenomes computationally. MAGs are useful since most microbial species are too numerous, and too difficult to isolate, culture, and sequence individually. We develop a computational pipeline to recover prokaryotic and eukaryotic MAGs, and a new method to visualise eukaryotic MAGs and improve their quality.

We also quantify the microbial diversity recovered from MOSAiC across different environments, including a huge number of novel genera and species, and identify species co-occurrence patterns, functional profiles, and diversification of important gene families, with an emphasis on MAGs as operational taxonomic units. We found highly distinct sea ice and seawater communities, with temporal abundance profiles dictated by seasonal changes. In the final chapter we describe avenues for future research, using the MAG catalogue as a foundation to model and predict microbial communities in a changing Arctic climate.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ADAM Adaptive Moment Estimation. 249

alr additive log-ratio. 58

AMOC Atlantic meridional overturning circulation. 10

AMP antimicrobial peptide. 190

ANI average nucleotide identity. 135

ARG antimicrobial resistance gene. 190

BGC Biosynthetic Gene Cluster. 190

BLAST Basic Local Alignment Search Tool. 39, 41

BWA Burrows-Wheeler Aligner. 41

CAO Central Arctic Ocean. 1, 3, 16, 20, 21, 62, 101, 128, 191

CAS CRISPR-associated protein. 190

CDS coding domain sequence. 24, 33, 44, 45, 145

clr centred log-ratio. 58, 137, 138, 164, 177

CRISPR Clustered Regularly Interspersed Short Palindromic Repeat. 44, 190

CTD conductivity, temperature, density measurement. 22, 62, 87

DCM Deep Chlorophyll Maximum. 9, 10

DMSP dimethylsulfoniopropionate. 15, 25

DNA deoxyribonucleic acid. 23, 24

DUF domain of unknown function. 5, 85, 86, 90, 96, 145, 176, 177, 185

GO Gene Ontology. 20, 60

GOLD Genomes Online Database. 87, 247

GTDB Genome Taxonomy Database. 149, 153

HAVOC Ridges - HAVens for ice-associated flora and fauna in a seasonally ice-covered Arctic OCean. ix, 4, 22, 62, 63, 83, 129, 130, 138

HGT horizontal gene transfer. 85, 101

HMM Hidden Markov model. 41–43

HMSC Hierarchical Modelling of Species Communities. 60, 186, 193, 194

HNLC high-nutrient, low-chlorophyll. 9

HPC high-performance computing. 125, 194

IBP ice-binding protein. 85–87, 90–92, 94, 96, 99, 101, 187, 246

IMG/M Integrated Microbial Genomes & Microbiomes. ix, 4, 5, 32, 34, 36, 39, 45, 46, 62, 68, 71, 131, 133, 149, 250

JGI Joint Genome Institute. 4, 7, 32, 62, 81

JSDM joint species distribution model. 6, 60, 193

KEGG Kyoto Encyclopedia of Genes and Genomes. 60

LCA least common ancestor. 107

MAG metagenome-assembled genome. ix, x, 3–7, 33, 48–50, 62, 68, 71, 72, 74–76, 80, 81, 83, 84, 87, 90–92, 96, 101–104, 110–113, 116, 125–127, 131–133, 144, 149, 152, 176, 177, 183–185, 187–194

MAP Metagenome Annotation Pipeline. 4, 5, 32, 33, 35, 38, 71, 133, 136, 144, 149, 150

MCL Markov Clustering Algorithm. 133

MCMC Monte-Carlo Markov-Chain. 194

MDS multidimensional scaling. 55, 57

MMETSP Marine Microbial Eukaryote Transcriptome Sequencing Project. 192

MOSAiC Multidisciplinary Drifting Observatory for Study of the Arctic Climate. 1, 3–5, 20, 21, 23, 27, 29, 31, 32, 39, 62, 87, 102, 129, 187, 188, 190–192, 194, 195

mRNA messenger RNA. 23, 30, 31

NCBI National Centre for Biotechnology Information. 32, 101

NCLDV nucleocytoplasmic large DNA virus. 191

NERSC National Energy Research Scientific Computing Center. 4

NMF Non-negative Matrix Factorisation. 56

OTU operational taxonomic units. 29, 44, 59, 172, 194

PAM Point Accepted Mutation. 32

PAR photosynthetically active radiation. 192

PCA principal component analysis. 50, 54–57, 187

PCoA principal coordinate analysis. 50, 54, 55, 57, 194

PCR polymerase chain reaction. 26, 29, 66, 89

PSU Practical Salinity Units. 141

RNA ribonucleic acid. 23

RPKM reads per kilobase million. 49, 50, 90, 94

RPM reads per million. 49, 50, 75, 76, 90, 128, 135, 137, 149

SBL serine β-lactamase. 190

SCMG single copy marker gene. 47, 48, 188

SCNIC Sparse Co-occurrence Network Investigation for Compositional Data. 60

SOM self-organising map. 56

SP signal peptide. 86

SRA Short Read Archive. 87, 101

t-SNE t-distributed Stochastic Neighbour Embedding. 56, 57, 248, 249

TMD transmembrane domain. 86, 97

TNF tetanucleotide frequency. 46

TPM transcripts per million. 49

UMAP Uniform Manifold Approximation. 56, 57, 133, 187, 189, 249

VAE variational autoencoder. 50

VALENCE Visualising Autoencoder Latent-Space Embedding Network for Clustering Eukaryotes. 104, 105, 110

VAMB Variational Autoencoder for Metagenomic Binnning. 46, 50, 84, 108, 189

WGCNA Weighted Gene Correlation Network Analysis. 5, 59, 60, 128, 138, 186, 190, 194

Chapter 1

Introduction

1.1 Motivation

Microbes are the principal primary producers of the ocean ecosystem, and are responsible for approximately 50% of annual global carbon fixation [2]. In the Arctic Ocean, marine microbes support a rich food web, including zooplankton, polar cod, seabirds, and aquatic mammals [3]. In turn, the success, migration, or decline of these species affects fish stocks further south, including those in the North Sea and North Atlantic Ocean. Microbes also affect, and are affected by, biogeochemical cycles. The most topical of these is the carbon cycle, in which microbes are involved in the sequestration of carbon dioxide (CO_2), but are also hugely affected by CO_2 concentrations.

Even so, the Arctic is currently undergoing extreme changes, warming over 1°C per decade, approximately four times the average global rate [4]. Understanding how these microbes respond to rapid environmental changes is therefore critical, not only to predict effects throughout the food web, but also in forecasting stoichiometric changes in the ocean and atmosphere. This is challenging due to the logistical difficulties of gathering data in the inhospitable conditions of the Arctic. These problems are particularly pronounced during the long polar winter. For this reason, the Central Arctic Ocean has been dubbed a 'black hole' in terms of our understanding of the processes governing the region [5].

1.2 MOSAiC

To fill this gap in the knowledge base, the Multidisciplinary Drifting Observatory for Study of the Arctic Climate (MOSAiC) expedition set out to conduct the largest ever survey of the Arctic Ocean, both in terms of length and number of personnel [6]. This was a year-long expedition with the objective of filling the void of information missing on the Central Arctic Ocean [7].

More specifically, MOSAiC gathered over 200 researchers to investigate five areas of scientific focus in the Central Arctic Ocean (CAO); atmosphere, sea ice, oceanography, biogeochemistry, and ecosystems. The goal of the project was to improve our understanding

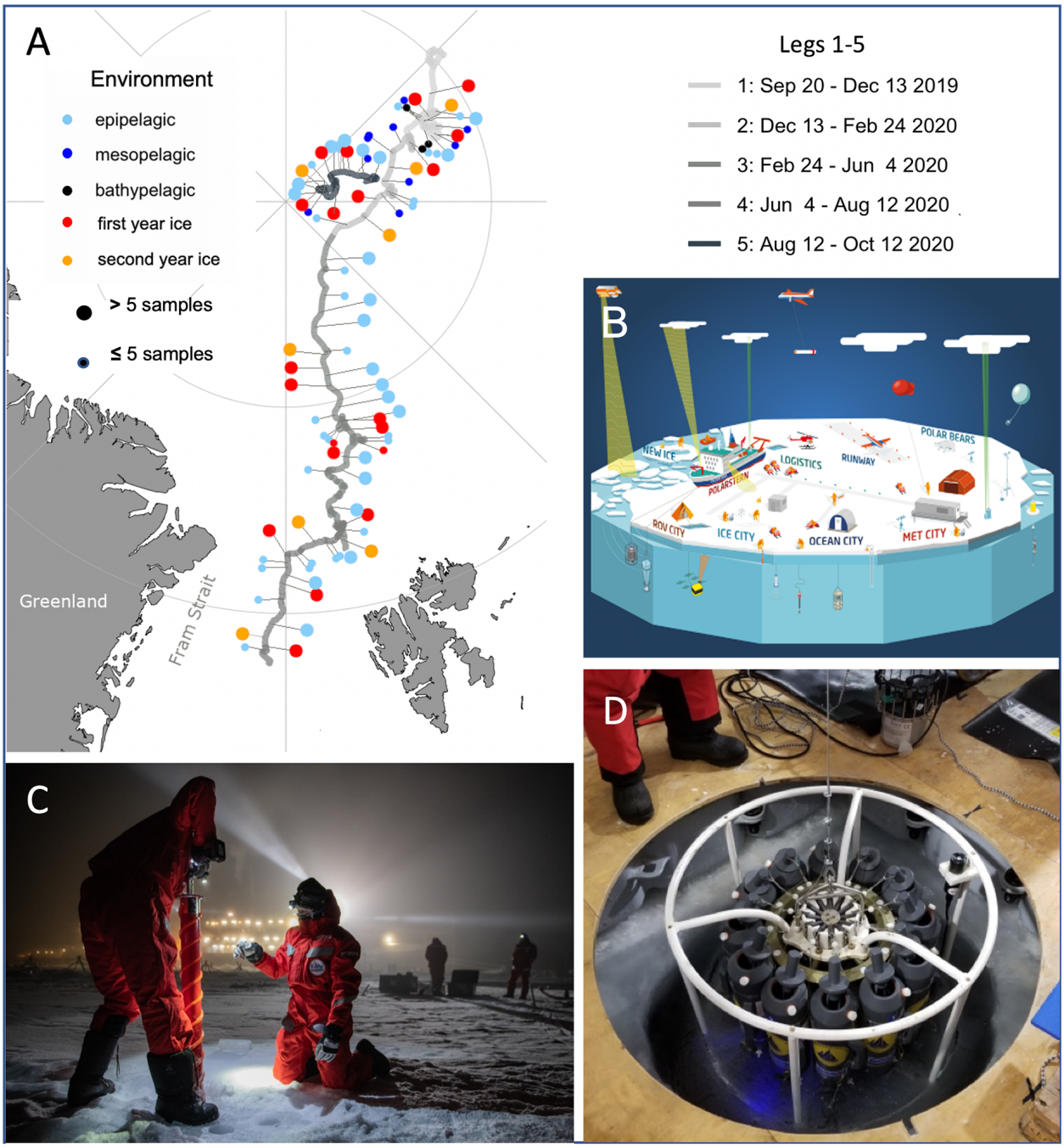


Figure 1.1: Elements of the MOSAiC expedition, from Mock *et al.* [8]. (A) Map of the drift route, with pins indicating the sampling site, and surface ocean sampling aggregated by week. (B) Infographic of the drifting ice camp representing diverse scientific groups and their "cities" on the surface of sea ice. (Credits Alfred-Wegener Institute.) (C) Drilling of sea-ice cores in the dark during leg 2 (Credits Esther Horvath.) (D) CTD (conductivity, temperature, and depth) rosette for measuring and sampling the sea water underneath sea ice as part of Ocean City. The CTD rosette is lowered through a hole (ca. 1.4 m in diameter) in the sea ice. (Credits Ying-Chih Fang.)

of the Arctic climate system, providing a robust dataset necessary for the region to be included within climate and Earth system models. This was a highly interdisciplinary project, though our focus was on the ecosystems component of the scientific programme, particularly the metagenomics data.

The expedition set sail from Tromsø (Norway) in September 2019 and returned to port over a year later in October 2020. Conducted aboard the research vessel Polarstern, MOSAiC acted as a drifting observatory; the movement of the vessel was dictated by the body of water supporting it, and it drifted with the same ice floe. Studies of this kind are called Lagrangian observations (as opposed to Eulerian observations, where a vessel would remain in a fixed spot while the water and ice mass move around it). Over the year-long expedition, MOSAiC collected over 1000 metagenomic and metatranscriptomic samples in a variety of habitats including sea ice, the surface ocean, and the deep ocean.

Given that these samples contain a significant number of microbes that are very difficult to culture but constitute a large proportion of the genetic diversity in the oceans, metagenomic analysis provides the most promising way to uncover the landscape of genes and their functions within these marine ecosystems. Despite being of huge ecological importance, the genomic and metabolic content of the ocean is not thoroughly understood, with roughly 40% of the genes found in other ocean surveys, such as the TARA oceans survey, being of unknown function [9].

The focus of this thesis is the investigation of metagenomic samples collected by MOSAiC. The fundamental questions in a metagenomic analysis are often very basic; which species are present in an ecosystem, and what functions do those species fulfil. A central aim is therefore to generate a comprehensive collection of the prokaryotic and eukaryotic genomic diversity of the metagenomes, in the form of metagenome-assembled genomes (MAGs). MAGs - putative partially-complete genomes of individuals recovered from a metagenome - are the fundamental unit of analysis in this thesis. Furthermore, we will investigate the drivers of microbial diversity (α and β diversity, covered in Chapter 2), and traits associated with abiotic filtering in Arctic habitats, such as sea ice, of which freezing-tolerance is one important example.

1.3 Structure of this Thesis

In the next two chapters, we will introduce the necessary ecological, biological, and bioinformatic context in order to progress with these questions. More specifically, in Chapter 2, we will provide an overview of the keystone microbial species present in the Central Arctic Ocean, including their functions within biogeochemical cycles and niches. We will then de-

scribe the MOSAiC drift in more detail, including information on the course of the drift, and the study design, in terms of sampling. Finally, we will introduce the necessary background in molecular biology, in particular the basics of DNA sequencing.

In Chapter 3 we will summarise the bioinformatics and data analysis techniques to be used in later chapters. Once these techniques have been defined in sufficient detail, we will be able to carry out analyses of metagenome-assembled genomes (MAGs). Much of our work builds on the bioinformatics analysis pipeline developed at the JGI called the Integrated Microbial Genomes & Microbiomes Metagenome Annotation Pipeline (IMG/M MAP), which we abbreviate as the IMG/M pipeline, or just the MAP. Sections 3.1 to 3.7.2 describe the steps involved in this pipeline as a particular example, while also describing some of the bioinformatics methods used more generally. We then review methods from numerical ecology used to study microbial diversity, such as α and β diversity indices.

The subsequent chapters contain the results from analysing different subsets of the data, culminating in an investigation of a large time-series of samples that span the full course of the MOSAiC drift.

Chapter 4 presents a description of MAGs recovered from two sets of samples that we received early on. The first was a set of pilot samples, taken during the Arctic winter. The second was a set of samples from ice ridges and sediment traps; these samples were collected under a MOSAiC subproject called Ridges - HAVens for ice-associated flora and fauna in a seasonally ice-covered Arctic OCean (HAVOC), led by Mats Granskog, and Oliver Muller, who provided input and review for details of the sampling process. One particular MAG, *Bacillariophyceae sp.* MOSAICH1_1, was of sufficient quality to be included in Phycocosm [10], an online resource for algal genomes, available to researchers worldwide. This work was conducted using the supercomputing resources available at National Energy Research Scientific Computing Center (NERSC), under the guidance and supervision of collaborators at the Joint Genome Institute (JGI), particularly Asaf Salamov, Igor Grigoriev, and Sara Calhoun - the annotations were performed myself but under the supervision of Sara Calhoun, and with review from Asaf Salamov and Igor Grigoriev. We also used results from the IMG/M MAP pipeline, which generated the assemblies, gene and Pfam annotations, and prokaryotic MAGs. The subsequent analysis was conducted by myself, in consultation with my supervisory team. Data from this work is published in Boulton *et al.* [11]. My contribution was in writing the text, generating the figures, preliminary data analysis and the eukaryotic coassembly (with guidance from Asaf Salamov and my supervisory team). Details on sampling and sample processing (DNA extraction) were provided by Oliver Muller. Some exploratory results from this pilot dataset were published in the overview paper Mock *et al.* [8], where my contribution was creation of the figures and preliminary data analysis.

In Chapter 5, we analyse a particular gene family as a case-study; genes containing the domain of unknown function DUF3494 ice-binding domain. These ice-binding proteins are known to be influential in Arctic ecosystems, and have some biotechnological relevance. We explore the diversity of ice-binding proteins within the pilot samples, making use of the MAGs generated in Chapter 4 to provide a genomic context for these genes. Work from this chapter was published in Winder *et al.* [12], where we analysed the distributions of this protein family across various different kinds of sample, and in different taxonomic groups of bacteria and archaea. This was a joint-first-author paper where I was responsible for the metagenomic analysis, description of the methods, and phylogenetic and diversity figures, in consultation with my supervisory team (the AlphaFold structures and domain architecture diagrams were by Johanna Winder). Similarly to Chapter 4, we made use of the Pfam annotations and prokaryotic MAGs generated by JGI's IMG/M MAP pipeline. Subsequent bioinformatics analysis was conducted by me, in consultation with my supervisory team and J. Winder, and the writing of the manuscript was a joint effort between myself and JW, again with guidance from my supervisory team. The structural modelling (Figure 5.1) was the work of JW alone, but is included with permission. We surveyed the gene families that appeared alongside the ice binding domain, and the different architectures that were present in genes containing the ice-binding domain.

Chapter 6 describes a pipeline I developed for visualising and binning eukaryotic MAGs, with the support of my supervisory team. Whereas for prokaryotes, the generation of MAGs is a routine and often highly automated process, metagenomic software are not typically optimised for finding eukaryotic MAGs, though more recently this is beginning to change. Nevertheless, there are only a relatively small number of eukaryotic genomes and medium to high quality MAGs in online databases. Eukaryotic MAGs are generated at a much lower frequency than prokaryotic MAGs, making manual curation of these practical and worthwhile.

In the final chapter of results, Chapter 7, we undertake a much larger analysis of over 300 samples, including all the samples previously analysed in Chapter 4, and utilising the pipeline from Chapter 6 to generate a catalogue of MAGs that represent the microbial diversity of the sea ice and ocean, across the whole year. My contribution was the diversity analysis, WGCNA and functional analyses, coassembly, generation and annotation of eukaryotic MAGs, as well as further prokaryotic binning, though we still used the IMG/M pipeline for the initial single-sample assembly, binning and prokaryotic annotation. These samples are structured as a time-series, covering the full extent of the drift period of MOSAiC. We analyse taxonomic and genomic diversity, and explore changes in time and differences between the environments sampled, particularly differences between water and ice.

In Chapter 8 we give an overview of our main results, and conclude with a summary of the prospective directions for future work, of which there are many. Catalogues of MAGs provide a foundational dataset, on top of which methods from genome-resolved metagenomics can be used to link established methods such as joint species distribution models (JSDMs) with gene-functional analyses and metabolic modelling. This can provide us with a greater understanding of how microbial species and their genes interact with each other, and with global biogeochemical systems.

Chapter 2

Biological Context

In this section, we will introduce some basic terminology and background information about the habitats we are studying and their ecology, including an overview of the major groups of microbes that are typically present in marine and sea ice samples. We will then go on to describe the standard operating procedures used by MOSAiC when collecting seawater and sea ice samples. From there, we will introduce the relevant terminology from molecular biology and genomics, so as to introduce the concept of MAGs, as well as describing a typical procedure used to sequence metagenomes, similar to that used at the Joint Genome Institute (JGI).

2.1 Microbial Ecology of the Global Ocean

Oceans cover roughly 70% of the Earth's surface, and host an enormous range of microbial communities [13]. These communities can be extremely varied; for example algal blooms might consist of overwhelmingly just a few species (e.g. blooms of the coccolithophore *Emiliania huxleyi*, which can be seen from space [14]), whereas samples from the deep ocean may be made up of more complex mixtures, often containing novel species [15] and extremophiles with enzymes that could be useful to the biotechnology industry [16]. Through the work of ocean surveys such as TARA Oceans, Global Ocean Sampling [17], [18], as well as numerous smaller sequencing projects, there has been some progress in characterising the microbial diversity of the oceans, for example in Tully *et al.* [19], even though a large proportion of this diversity still remains unknown. Two important components of marine microbial diversity are plankton (particularly phytoplankton) and algae. Neither of these terms are defined in a phylogenetically consistent manner - plankton (from the Greek planktos, meaning 'wandering') covers any marine organism that drifts with ocean currents; this covers organisms from all domains of life, including most marine microbes. Phytoplankton - 'plant-like' plankton - are photoautotrophic plankton, meaning they generate their own food from sunlight. Algae is an informal term referring to several groups of photosynthetic eukaryotes found in both freshwater and marine environments; these groups have distinct evolutionary histories and are not all closely related to one another. These two terms are often prefixed with a length

scale to further categorise unicellular groups by size, i.e. nano- and pico- plankton, and microalgae.

Microbial communities are important for both terrestrial biogeochemical processes and food webs, and for those in the ocean. However, a major difference between terrestrial and marine ecosystems is the extent of this importance. On land, primary production (i.e. the synthesis of organic molecules from inorganic material such as CO_2, water, and nitrogen) is driven by plants. In the ocean, essentially all primary production is driven by unicellular organisms [20]. Microorganisms dominate the ocean in terms of ecological importance and purely in terms of biomass [21], and although the biomass of the ocean comprises just 3% of total global biomass, it accounts for approximately 45% of global primary production [2].

Some species and groups of microbes in the ocean play a more significant role than others in the food web, and in biogeochemical cycles. In the next subsections, we will summarise several important groups of eukaryotic phytoplankton, as well as marine prokaryotic diversity, both generally and for the Arctic Ocean in particular. However, we will begin with a brief overview of physical oceanography, and introduce some guiding principles which help to explain general patterns of microbial abundance in the oceans.

2.1.1 Physical Oceanography Background

Oceans are stratified into pelagic layers, layers of the ocean which have distinct physical and biological properties. These layers can be based on biological properties, such as the concentration of chlorophyll, or on physical properties such as density, temperature, and salinity. Ocean layers are important because different layers tend not to mix with each other, and therefore nutrient concentrations can be limited in the upper ocean depending on the strength of the stratification, which in turn can limit microbial growth. Deeper ocean waters are not as depleted in nutrients and are therefore more nutrient-rich [22]. In general, water density increases with increasing salinity and with decreasing temperature. Where mixing between layers does occur, for example due to Ekman transport (a physical process of upwelling deeper water, caused by the interaction between the Coriolis force and shear wind forces on the surface ocean) in some coastal regions and at the equator, there is a corresponding increase in primary productivity [23].

Oceanographic Terminology: Ocean Layers

- **Upper Mixed Layer:** The upper layer of the ocean, homogenised by the effect of the wind, and wave-driven turbulence. Can be from a few metres to a few hundred metres in depth.
- **Deep Chlorophyll Maximum (DCM):** The layer at which the concentration of chlorophyll is highest. This is typically at a depth of between 20 and 50 m.
- **Pycnocline:** A boundary layer, defined by a rapid change in water density per change in depth, that separates warmer, less salty water above from denser, colder, saltier water below.
- **Epipelagic:** The surface layer of the ocean, 0 to 200 m. Also called the photic zone, this is where light levels are high enough to sustain photosynthesis above the rate of respiration.
- **Mesopelagic:** (Twilight zone.) Middle ocean layer, typically starting at a depth of 200 m and ending at 1000 m.
- **Bathypelagic and Abyssopelagic:** The bathypelagic zone is between 1000 and 4000 m, while the abyssopelagic zone extends beyond 4000 m.
- **Benthos:** The sea floor.

Pelagic water masses can be further categorised by their geography, topography, chemistry, and their dynamics. Open oceans are often nutrient-poor, and microbial growth is limited by the availability of basic nutrients, particularly nitrogen [24], [25]. The ratio of three basic nutrients, carbon, nitrogen, and phosphate, has been empirically found to be at proportions of 106:16:1 C:N:P in the surface ocean (the Redfield ratio [26], [27]), though with some variation [28]. Though these nutrients are often limiting for microbial growth, the Southern Ocean is exceptional as a high-nutrient, low-chlorophyll (HNLC) zone, due to the limitation of iron and other micronutrients (including zinc, cobalt) [24], [29], as is the northern Pacific Ocean. The Arctic Ocean is also exceptional, where the availability of light is a limiting factor, in addition to nitrate [30], [31]; this will be discussed in more detail in a later section. Ocean topography can influence these micronutrient concentrations - near continental shelves, the concentration of micronutrients is generally higher, and micronutrients such as iron are less likely to be limiting. In the context of the Arctic, iron has not received as much attention as a potentially limiting micronutrient; however, a 2013 study found iron, nitrogen, and light as potentially colimiting in the Beaufort Sea[32], and a 2020

analysis of the GEOTRACES programme found that in the Arctic summer, nitrogen and iron were limiting in the Fram Strait[33]. The other two components of the Redfield ratio, phosphate and carbon, are generally not deemed as limiting in the Arctic Ocean, though there may be a future decline in phosphate availability in the upper ocean due to climate change [34]. Due to the abundance of diatoms in the Arctic, silica has been found as a potentially bloom-terminating nutrient, at least during the spring [35].

The dynamics of oceans plays a large role in determining microbial communities and ocean biogeochemistry. These dynamics are determined by the atmosphere and by temperature, salinity, and density gradients, and are ultimately driven by energy from the sun. Surface ocean circulation is in large part controlled by the wind, which itself is driven by three pairs of atmospheric cells, called Hadley cells (near the equator), Ferrel cells (at mid-latitudes), and Polar cells. The force of the atmosphere on the surface ocean, interacting with the rotation of the Earth, cause surface ocean currents to rotate in large gyres, anticlockwise in the northern hemisphere and clockwise in the southern hemisphere. At the poles this effect is the strongest, with the Antarctic circumpolar current the strongest ocean current in the world, with mean transport of up to 150 million m^3/s [36]. (There is no equivalent current in the Arctic, since there is land at the corresponding latitude.) Deep ocean currents are much slower than those on the surface, and are driven by convection and changes in temperature and salinity. The thermohaline conveyor is a deep ocean circulation current spanning the globe, driven by changes in temperature and salinity, which transports energy and nutrients around the world [37]. This current, especially part called the Atlantic meridional overturning circulation (AMOC), appears to be slowing down due to the influence of climate change [38].

We can use levels of light, temperature, and nutrient concentrations to predict where microbial species of different morphology might occur, based on a few general empirical principles. In general, smaller cells with a faster cell cycle have a comparative advantage in warmer environments, an empirical observation called the temperature-size rule [40], [41]. Additionally, species diversity increases toward the equator, as the higher temperature is able to support a greater range of species [42]. Smaller cells also tend to cope better with nutrient limitation than larger cells [43], [44]. The availability of light is also a key factor in determining the distributions of microbial species; photoautotrophs are present in highest abundance at the DCM where light intensity is optimal for photosynthesis. These principles are backed up by trait-based modelling and ocean survey sampling [42], [45]. Beyond these very broad categorisations of microbial distributions, in the next sections we will look at a few important groups of organism present in the Global Ocean in more depth, and their functions within the ecosystem and in biogeochemical cycles.

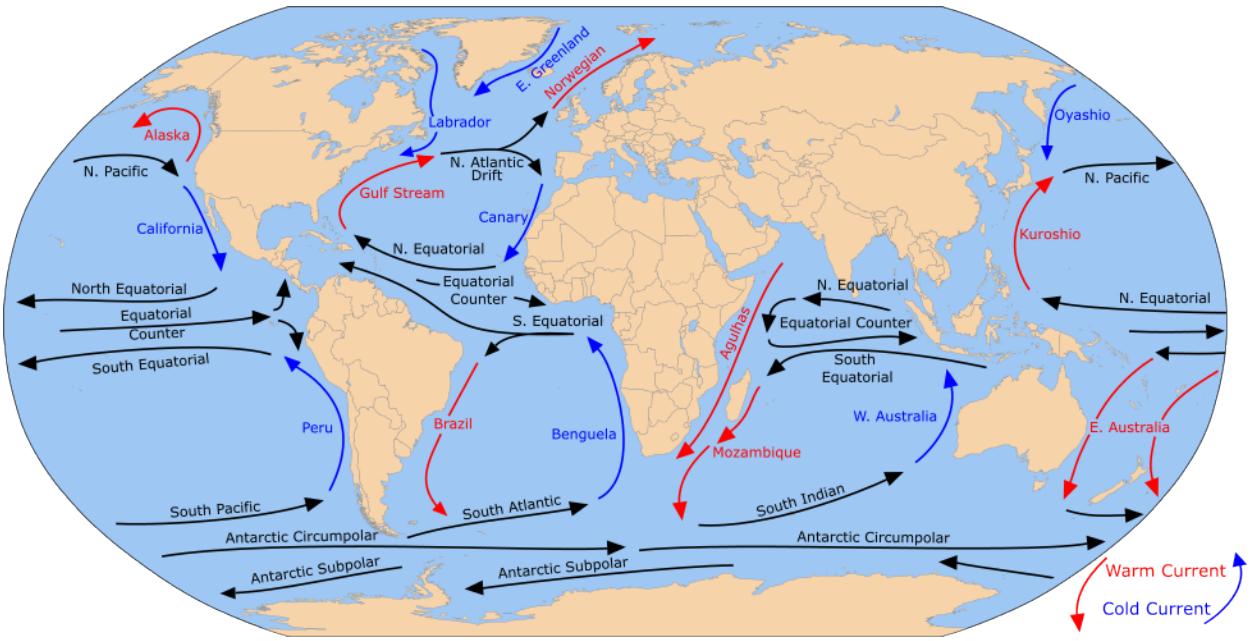


Figure 2.1: Map of Global Ocean surface circulation currents and gyres, adapted from Pidwirny [39].

2.1.2 Microbial Prokaryotes in the Global Ocean

The clades of bacteria and archaea are known collectively as prokaryotes. Prokaryotic cells have no nucleus, instead their DNA is generally mobile within the cell, most often as a large circular loop. Prokaryotic genomes and cells are usually much smaller than those of eukaryotes, and typically have a much shorter cell generation time, on the scale of minutes or hours rather than closer to a day. They also have smaller mobile genetic elements called plasmids, which can be exchanged between individuals (even of different species), allowing for one mechanism of horizontal gene transfer. Prokaryotic cells are on the order of 0.5 to 5 μm in length, with genomes of size typically less than 10 Mbp, and more commonly about 1 Mbp. Some prokaryotes are motile, propelled with a flagellum, a whip-like structure which drives them through the water. Bacterial cells have both a membrane and a cell wall containing a sugar crossed-linked polymer called peptidoglycan. Bacterial cells also contain ribosomes, and may contain other intracellular structures (e.g. carboxysomes and magnetosomes), but are otherwise simple in structure.

Two highly abundant phyla of bacteria in the Global Ocean are Pseudomonadota (previously called Proteobacteria) and Bacteroidota. These phyla are found in a wide range of environments, including the ocean and sea ice but also the human gut [46], [47]. Pseudomonadota alone constitute 15 to 50% of the prokaryotic diversity in some parts of the surface ocean [48], [49]; they are split into 5 groups (labelled with the Greek letters alpha to

epsilon), of which the Alpha- and Gammaproteobacteria are the most dominant. In surface oceans, groups such as SAR11, SAR86 and Roseobacter are abundant heterotrophs which rely of the availability of dissolved organic carbon to thrive. Roseobacter in particular (part of the Alphaproteobacteria) are found in all marine habitats, sometimes at an abundance of up to 25% [50], [51]. Several species have been found to have mutualistic connections with eukaryotic microalgae [52]. Within the Gammaproteobacteria, the two largest orders in the ocean are the Pseudomonadales and Enterobacterales; though these are ubiquitous, also found in soil, or host associated, and include some human pathogens.

Bacteroidota, similarly to Pseudomonadota, are also found in many environments including the human gut, in other animal hosts, but also in soil, and marine environments. Within the Bacteroidota, the Flavobacteria are one of the largest classes found in the marine environment, and play a role in carbon cycle, with large numbers of diversified carbohydrate-active enzymes [53]. Many Bacteroidota are opportunistic pathogens [54], and most are Gram-negative heterotrophs. A flavobacterial genus within polar environments are the *Polaribacter*, some species of which have a metabolism enriched in carbohydrate-active enzymes, and in proteorhodopsins [55].

The next six most diverse prokaryotic phyla (in terms of species richness) found in the Global Ocean based on the surveys Chen *et al.* [56] and Nishimura *et al.* [57], are (ordered alphabetically): Actinobacteriota, Chloroflexota, Cyanobacteriota, Planctomycetota, Thermoplasmatota, and Verrucomicrobiota.

Two of these particularly of note are Planctomycetota for their large genomes (up to 18 Mbp [56]), and the Cyanobacteriota (blue-green algae), due to their ability to photosynthesise, and their prevalence in warmer, surface oceans. Photosynthetic bacteria are an important group due to their role as primary producers and the Cyanobacteriota are a phylum responsible for most of this primary production, with a single genus, *Prochlorococcus*, accounting for up to 50% of chlorophyll-a in large parts of the oceans, though mainly nutrient-poor areas [58]. However, while globally significant, *Prochlorococcus* are not typically present in the polar oceans. Other bacterial phyla can photosynthesise, though not necessarily using the exact same biochemical pathway; Cyanobacteriota is the dominant prokaryotic phyla that carry out aerobic photosynthesis.

Archaea are a separate domain of life, only discovered in 1977 by Carl Woese and George Fox [59]. Morphologically, bacteria and archaea can appear similar, however evolutionarily, they are quite distant; archaea have distinct ribosomal RNA compared to bacteria, and a very different cellular membrane structure. Whilst they are the minority, archaea are still an important component of the microbial ecology of the ocean, and comprise approximately 30% of microbial cells some marine environments [60]; this is surprising given that archaea

were originally thought to be rare extremophiles, living in habitats such as hydrothermal vents. In fact, while some species of archaea are extremophiles (such as the genera *Thermococcus* and *Pyrococcus*, which live in temperatures up to 100 °C and some of which can withstand extremely high amounts of ionising radiation), they are quite ubiquitous, and are even found in the human microbiome. A large amount of archaeal diversity is within the ocean, and particularly within the phylum of Thermoplasmatota. Thermoplasmatota are the only archaeal phylum with comparable abundance to the major bacterial phyla in the surface ocean [57].

2.1.3 Microbial Eukaryotes in the Global Ocean

Eukaryotes are defined as organisms whose cells contain a nucleus. Eukaryotes may have begun diverging from archaea approximately 1000 to 2000 million years ago, and the complexity of eukaryotic cells has been in part due to several endosymbiosis events over the course of their evolution [61], [62]. In the oceans, phytoplankton are the main primary producers, with eukaryotic phytoplankton comprising a substantial proportion of this primary productivity.

Of the eukaryotic phytoplankton, there are some key groups which are highly represented in the oceans, including diatoms, dinoflagellates, and coccolithophores.

Diatoms are a class of unicellular microalgae, normally ranging between 20 to 200 μm in size. Diatoms are equivalently defined as the class of Bacillariophyceae, within the Stramenopiles. It is estimated that there are between 30,000 and 100,000 species of diatom [63], with efforts to sequence a fraction of these, for example in the 100 Diatom Genomes project [64]. Their genomes range in size between 27 Mbp (*Phaeodactylum tricornutum*) and 1.5 Gbp (*Thalassiosira tumida*) [65], though some of the more abundant model species such as *Thalassiosira pseudonana* and *Fragilariopsis cylindrus*, are toward the lower end of this scale, 34 and 61 Mbp respectively - diatom genomes over 1 Gbp are atypical. Diatoms have been called the 'jewels of the sea' due to their highly symmetrical body plans, and silica cell walls, called frustules, which are unique to diatoms and make them particularly important in the global silica cycle. Diatoms split into two evolutionarily distinct clades; pennate (rod-shaped) and centric (circularly symmetric, disc-shaped). In general, diatoms can be mixotrophic, though most species are photoautotrophs, with energy harvested through photosynthesis [66]. Collectively, they are responsible for approximately 45% of the primary production in the Global Ocean [67]. A few species have been studied extensively as model organisms, such as *Phaeodactylum tricornutum*, and *Thalassiosira pseudonana*. Diatom species can exhibit a 'bloom and bust' life-cycle, growing into enormous seasonal

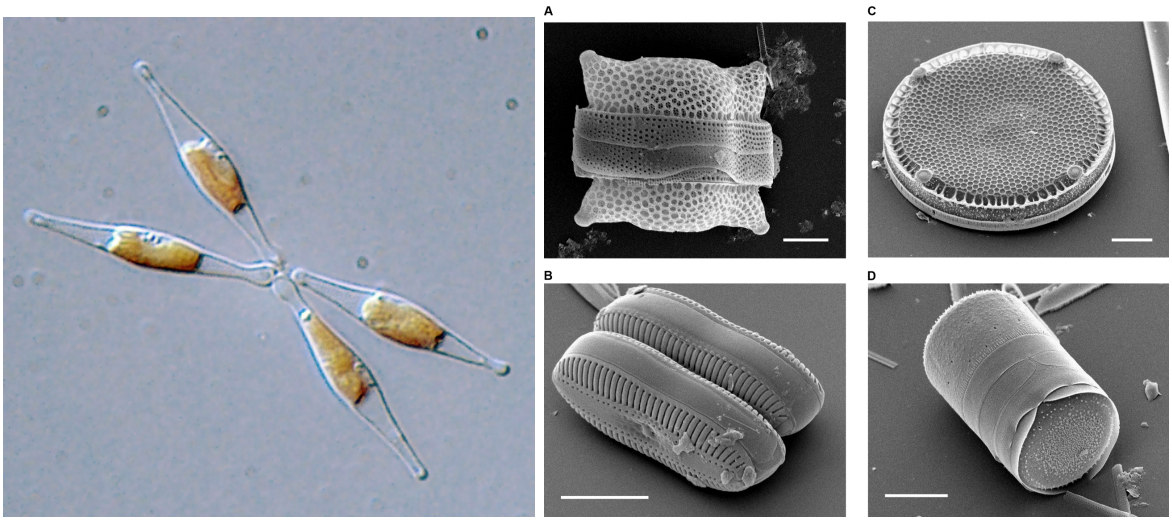


Figure 2.2: Left: Light microscopy image of *Phaeodactylum tricornutum*. Right: Electromicrograph of diatoms. These images highlight the pennate / centric split between major diatom clades. Images from Bradbury [72]

blooms. In the Arctic Ocean, psychrophilic ('cold-loving') diatoms are the main primary producers, with genera and species such as *Pseudo-Nitzschia* and *Fragilariopsis cylindrus* colonising sea ice, with a main bloom in spring and then later a smaller bloom in autumn [68].

Another key group of unicellular eukaryotes are the dinoflagellates, which range in size from 2 to 2000 μm, and are predominantly mixotrophic, though there are some heterotrophic species, including those that predate on diatoms. Some species exhibit bioluminescence, and others produce toxins, which can move through the food web, including to humans who eat affected shellfish. Dinoflagellates can create harmful algal blooms when in significant numbers, causing environmental and economic damage. They are typically more prevalent in warmer, coastal waters, though are present in polar oceans, with some species predating on polar diatoms [69]. Dinoflagellates have large, complex genomes and a complicated evolutionary history, having acquired plastids from green algae, haptophytes, and cryptomonads in several separate secondary (or tertiary) endosymbiotic events [70]. Their nuclear genome can be up to 240 Gbp, 80 times that of the human genome [71]. It is therefore virtually impossible to recover a complete dinoflagellate genome outside of a culture.

Coccolithophores, and haptophytes more generally, are a third group of unicellular photosynthetic eukaryotes which are again notable for rapidly producing extremely large blooms. Coccolithophores, which make up around 85% of known haptophyte species, are calcifying microalgae; they have calcium carbonate shells, which can act to sequester CO_2 if they fall to the sea bed. Fossilised coccoliths are responsible for producing chalk sedimentation. The

genome of *Emiliania huxleyi* , a model coccolithophore species commonly found in Atlantic coastal waters, is on the order of 100 to 130 Mbp in size [73].

A final group we cover are the Chlorophyta, a phylum including a large number of green algae. Chlorophyte algae are the closest algal relatives of land plants, and algae such as *Volvox carteri* within the chlorophytes provide insight into the origin of multicellularity [74], [75]. Chlorophyta include the Mamiellophyceae and in particular the genus *Micromonas*; polar species such as *Micromonas polaris*, which is the dominant picophytoplankton in the Arctic, especially during summer [76]. *Micromonas* genomes are on the order of 20 Mbp [77]. This genome is extremely small for a eukaryote, though other chlorophyte genomes are smaller still, such as *Ostreococcus tauri* [78], at 12.5 Mbp. Of course, being green algae, these species are all predominantly photoautotrophs, though interestingly there is some evidence that they may engage in mixotrophy under certain circumstances [79].

2.1.4 Biological Function

All of the above groups are significant in the Global Ocean due to their interaction with the biogeochemical systems of the Earth, and their species interactions with each other. They are all are affected by anthropogenic climate change, and in particular are sensitive to changes in CO_2 concentration and temperature. For example, formation of calcium carbonate shells of coccolithophores is negatively impacted by increasing CO_2 concentrations in the ocean and the corresponding decrease in pH. As well as the carbon cycle, they are all involved in other biogeochemical cycles. In the sulphur cycle; both eukaryotic and prokaryotic groups produce dimethylsulfoniopropionate (DMSP), which is transferred into the atmosphere as dimethyl sulphide (DMS), before progressing through the sulphur cycle. This compound is important as it contributes to the formation of cloud condensation nuclei, and has a modulating effect on temperature [80]. All groups are involved in cycling organic and inorganic carbon, through photosynthesis, metabolism, and other more specialised pathways such as generating biofilms and extracellular polymeric substances (mainly prokaryotes) [81] or coccoliths (haptophytes). A further key function is the fixation of nitrogen; N_2 is plentiful in the air as the relatively inert gas but only some bacterial groups can convert this to ammonia, and then to other biologically useful forms [82]. Diazotrophic (nitrogen fixing) marine bacterial groups from the phylum Cyanobacteriota have been studied in detail [83], however diazotrophy has also been found across a number of prokaryotic phyla including several subclades of Proteobacteria [84]. A related process, called anammox (anaerobic ammonia oxidation) also occurs; in this pathway, ammonium and nitrite are converted to nitrogen gas under anoxic conditions. Unlike nitrogen fixation, anammox is restricted to a specific clade within the Planctomycetota

[85], [86].

In addition to their roles within chemical cycles, different species fulfil different ecosystem functions, such as prokaryotic-algal associations within a 'phycosphere' [87], a sphere of influence near algal cells where mutualistic effects with microbes may take place - for example exchange of the vitamin B_{12}. Within prokaryotic communities, cell-to-cell communication takes place such as the coordination of bioluminescence within *Vibrio fischeri* [88]; this communication is known as quorum-sensing.

The proliferation of horizontally-transferred genes in prokaryotes means that assigning functions to specific groups is hard - however there are a few groups that have strong associations with particular biological functions. Cyanobacteria are strongly associated with aerobic photosynthesis, and the purple sulphur bacteria (a kind of Gammaproteobacteria) are associated with a form of anaerobic photosynthesis. In their model of prokaryotic phyla and functional traits, Finn *et al.* [89] demonstrated that there was an association between taxonomy and function; however the fact that their models were able to infer bacterial phyla from functional traits with just 80% accuracy highlights the complexity of this relationship.

2.2 Microbial Ecology of the Arctic Ocean

A brief oceanographic description of the Arctic Ocean is as a mediterranean basin extending from the North Pole down to a latitude of approximately 66° north, at the boundary of the Arctic Circle. As a mediterranean (meaning land-surrounded) basin, the largest two points of connection with the Global Ocean are the Bering Strait (a mere 45 m deep, 82 km wide connection to the Pacific Ocean, between Alaska and Russia), and the much larger Norwegian sea and Denmark strait, linking to the Atlantic Ocean. There are two basins within the Arctic, the Eurasian basin and the Canada basin, separated by the Gakkel ridge. The deepest points of the Arctic basins are approximately 4000 m in depth. The Arctic Ocean is the world's smallest ocean, covering just 14 million km^2. The key features of Arctic oceanography are covered in Figure 2.3.

The points from Section 2.1.1 relevant to the Global Ocean are also applicable to the Arctic, however there are other circumstances which single out the Central Arctic Ocean (CAO) as a unique environment, such as extreme physical conditions, including high winds, freezing temperatures, and huge seasonal variations. In the last decade, sea ice covered between approximately 5 and 14×10^6 km^2 through the course of the year [91] (though this figure is rapidly declining), and over the winter months the sun is permanently below the horizon. The typical depiction of the Arctic is therefore of a dark, inhospitable desert. However, at a microbial level, the Arctic is full of life.

Figure 2.3: Map of Arctic Ocean circulation and oceanography, adapted from Zeimusu [90].

Three defining features for microbial biology in the Arctic Ocean are then the low temperature, the seasonal darkness and light limitation, and sea ice cover. Sea ice, in particular, is an environment unique to the poles. As sea ice freezes, highly saline water is ejected, leading to the formation of porous ice filled with brine channels and brine. Many Arctic microbes have evolved metabolic pathways to help them survive in this environment; for example the diatom *Fragilariopsis cylindrus* produces ice-binding proteins, which inhibit the formation of ice crystals that would otherwise damage the cell [92]. These ice-binding proteins are extremely diversified in polar ecosystems [93]. Extracellular polymeric substances (chains of sugars and other organic molecules, generating a biofilm) are also expelled by prokaryotic microbes within sea ice [94], possibly providing protection from the extremely haline brine. In terms of prokaryotic community composition, sea ice is known to be particularly rich in Gammaproteobacteria. In Bowman *et al.* [95], the prokaryotic composition of multi-year sea ice was made up of 84% Gammaproteobacteria and Bacteroidetes (more specifically, Flavobacteria), from which the families Moraxellaceae and Flavobacteriaceae were the largest two families. In Bellas *et al.* [96], sea ice metagenomes from the Hudson Bay in northern Canada were particularly rich in Oceananospirillales, Alteromonadales, (both Gammaproteobacteria), and Flavobacterales (Bacteroidota). Genera from the Gammaproteobacteria and Bacteroidota are present at both poles, for example *Glaciecola*, *Paraglaciecola* (Gammaproteobacteria), and *Psychroserpens* and *Polaribacter* (Bacteroidota) [97], [98]. Sea ice will often contain a dense algal layer near its interface with the sea, which is easily recognisable as a brown 'stripe' running through an ice core - consisting mostly of diatom species. In their study of summer Arctic sea ice and seawater communities, Rapp *et al.* [97] found the diatom *Melosira arctica* populating the thinner ice floes, as well as *Nitzschia*, *Fragilariopsis*, *Cylindrotheca*, and the stramenopile protist *Labyrinthulomycetes*. They identified *Micromonas* as a eukaryotic generalist present in both sea ice, seawater, and sediment, alongside 6 other generalist eukaryotic genera and 3 prokaryotic genera.

The Arctic microbial community in the water column is highly distinct from that of the sea ice, and has been much better studied. The following authors [100]–[104] all measured the prokaryotic community composition in the Arctic Ocean water column, though nearer to continental shelves than MOSAiC, either monitoring populations in the Fram Strait, or at the sites sampled by the TARA Oceans expedition. In the Royo-Llonch study [105], the most abundant phyla were again Gammaproteobacteria and Bacteroidota, but there was a much greater range of prokaryotic phyla present compared to sea ice; Alphaproteobacteria, Thermoplasmatota, Chloroflexota, Actinobacteriota and Verrucomicrobiota were the next most abundant phyla present from that study. Qualitatively, this was closer to the phyla described within the Global Ocean, compared to the phyla in sea ice. The main difference

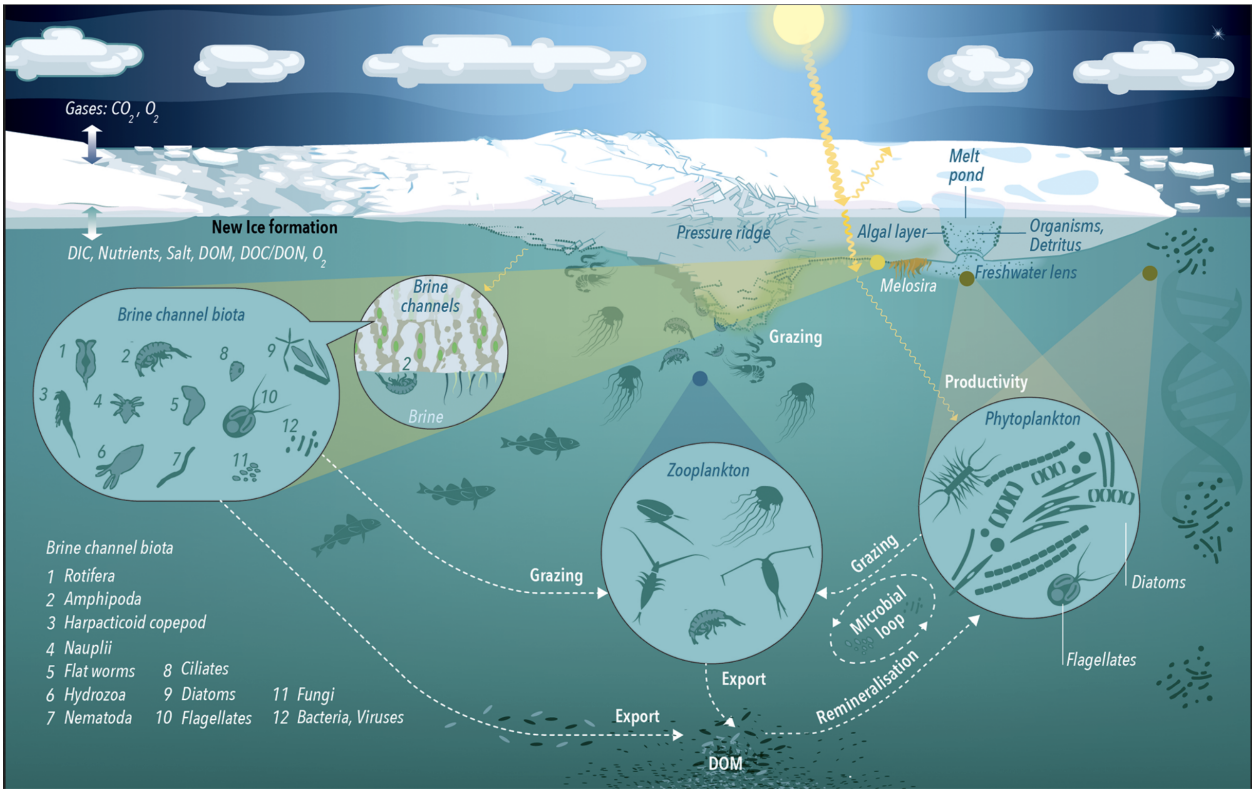


Figure 2.4: Conceptual diagram highlighting some major elements of the Arctic ecosystem, from Fong *et al.* [99]

with the Global Ocean was the relatively lower abundance of Cyanobacteriota, and the relative increase in Bacteroidota and Gammaproteobacteria in the Arctic. Pedrós-Alió *et al.* [106] speculated that in the Arctic, the ecological niche of Cyanobacteriota was filled by eukaryotic picophytoplankton such as *Micromonas*.

Cold adaptation is a second important feature of Arctic microbiology. In Duncan *et al.* [104], heat shock proteins were noted as the most abundant functional (GO) term associated with polar prokaryotes; these are known to be protective not just from heat but from temperature changes in general [107]. The optimal temperature for growth of polar diatoms such as *Fragilariopsis cylindrus* is in the range of 3 to 5 °C [108]. Hüner *et al.* [109], suggested that functional redundancy could in part explain the psychrophilic traits of polar algae, where a larger amount of gene duplication and regulation could allow algal species to tolerate lower temperatures (with the trade off of a slower overall growth rate).

Light limitation also plays a key role in shaping the microbial community and the biological functions present in the Arctic - the productive period of the Arctic Ocean begins immediately following the onset of light in March. Marine chlorophyll-a concentrations are high at the poles relative to the equator [110], and polar photosynthetic species seem primed to take advantage of light as soon as it becomes available [111]. In the water column, the chlorophyte genus *Micromonas* is known to dominate the algal community in the spring and summer (April to July) [112]. It is much less clear how photosynthetic species remain active through the polar night when the sun remains below the horizon (approximately October to February, though varying by location) [113]; though switching to heterotrophy has been suggested as a survival mechanism; see Berge *et al.* [114] for a review.

2.3 MOSAiC in Detail

The previous section provided an overview of the Arctic ecosystem in general; here, we examine the MOSAiC expedition specifically. Most of the information provided here are from the following overview articles [99], [115]–[117].

2.3.1 Geography of the MOSAiC Drift

The MOSAiC expedition was a Lagrangian drift survey, based around the icebreaker RV Polarstern [118]. The Polarstern began drifting on October 4th, 2019, starting in the Barents Sea and moving northward in leg 1 of the expedition (19th September to 15th December 2019), into the CAO, where it spent the majority of the drift (see Figure 2.5). By February 24th, 2020 (during leg 2), the vessel remained still in the CAO but was drifting south, toward

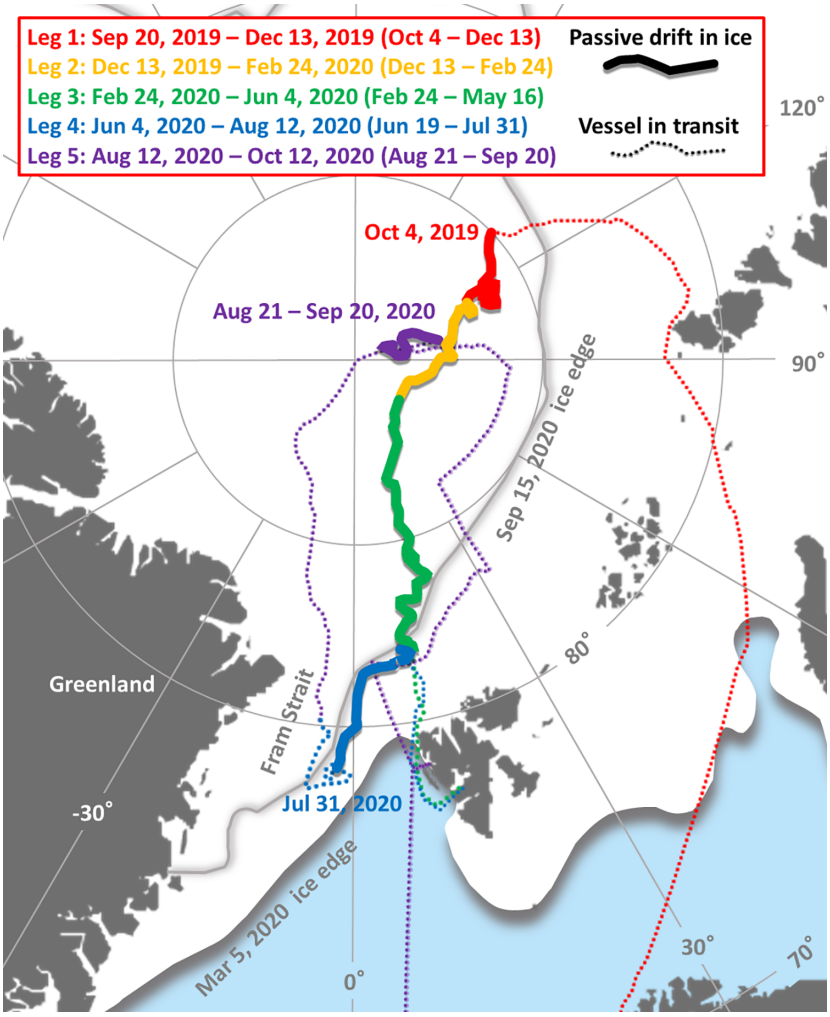


Figure 2.5: Route of the Polarstern during the MOSAiC expedition, from Rabe *et al.* [7]. The general course of the drift was at first toward the North Pole, then southward toward Svalbard and the Fram Strait, before the vessel returned to the CAO in August 2020, to complete the last leg of its drift.

the Fram Strait and the Greenland Sea. There was a break for resupply at Svalbard in May / June (between leg 3 and 4), after which the Polarstern returned to its previous position to continue its drift. Between July 31st and August 21st 2020 (start of leg 5), the vessel returned to the CAO, to complete the research expedition by September 20th 2020 and set a return to land. Most sampling efforts were therefore relatively uninterrupted, with the exception of the two breaks (May 16th to June 19th, July 31st to August 21st) just mentioned.

2.3.2 Study Design

Throughout the course of the year-long drift, the observatory was responsible for collecting a wide range of data, with work split across five teams. These including atmospheric measurements, physical oceanography, sea ice physics and snow, biogeochemistry, and ecosystems science [99], [115]–[117].

The ecosystems sampling effort collected approximately 1000 metagenomes and metatranscriptomes in total, with two strata of metagenomic sampling. The first was a core time series consisting of sea ice and water samples typically collected weekly, with two or three replicate samples taken in most cases. This core time series constituted the most consistent and relatively uninterrupted set of metagenomic samples. Ice was collected from two coring sites using an ice-corer (essentially a hollow metal drill bit, manually extruded into the ice). Seawater samples were collected with a conductivity, temperature, density measurement (CTD) rosette, a circular array of Niskin bottles with sensors attached. Figure 2.6 shows an example of each sampling method. Several Niskin bottles were often lowered and opened concurrently in one cast; where this was the case with our samples we will note that they are biological replicates.

Samples from the sea ice were taken from sections of ice cores; typically biological samples were taken from the meltwater of a 5-10 cm section of an ice core. The depth of these sections was a combination of opportunistic (choosing the depth containing a visible brown 'stripe' of algae) and systematic, i.e. sampling the layer at the base of the ice core closest to the sea-ice interface.

The second kind of sampling effort consisted of satellite projects, such as intense observation periods (e.g. samples collected throughout the course of 24 hours at highly frequent intervals), opportunistic sampling (e.g. of relatively transient features such as melt ponds), or systematic projects focussing on particular features, such as the project HAVOC (Ridges - HAVens for ice-associated flora and fauna in a seasonally ice-covered Arctic OCean), which studied ecosystems within ice ridges. Almost all metagenomic samples had some minimal metadata attached; either CTD parameters (for seawater) or sea ice physical properties, and often, nutrient concentrations, in particular concentrations of nitrate, nitrite, phosphate, and silicate.

Both seawater and sea ice samples were homogenised and filtered with a 0.22 μm mesh before DNA was extracted. This mesh size is extremely small, so essentially all cells (both prokaryotic and eukaryotic) were trapped in the filter. On the other hand, most viruses (except giant viruses) are smaller still, on the scale of tens of nanometres in diameter, and were therefore not captured. Sea ice was mixed with 0.22 μm-mesh-filtered seawater and

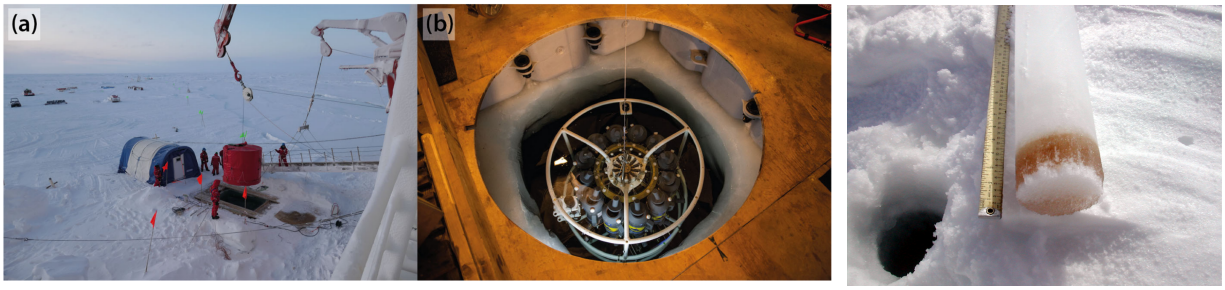


Figure 2.6: Left and middle: Images from the MOSAiC CTD sampling site, including a CTD rosette being lowered into the sea, adapted from [7]. Right: An example ice core, from [119]. The brown 'stripe' is from the sea-ice interface, and comprised of a dense algal layer.

melted prior to filtering. This was due to logistical constraints during the drift; the use of sterile seawater instead would have been preferable. However, the small mesh size should have minimised contamination from the seawater community. The filters were then frozen with liquid nitrogen and stored at -80 °C.

2.4 Molecular Biology and Sequencing

This section covers some basic preliminaries in molecular biology. The sequencing protocol used for the MOSAiC samples is described in Section 2.4.4.

2.4.1 DNA, RNA, and Protein

Deoxyribonucleic acid (DNA) is a double-stranded sugar-phosphate polymer, consisting of two helically interwoven deoxyribose-phosphate backbones running in antiparallel directions, cross-linked with hydrogen bonded base-pairs. In DNA there are 4 possible bases that occur in two pairs: adenosine (A) makes two hydrogen bonds with thymine (T), and guanine (G) makes 3 hydrogen bonds with cytosine (C). Sections of DNA called genes are transcribed into a single-strand ribose-phosphate polymer called messenger RNA (mRNA), by a DNA-dependent RNA polymerase. The RNA-polymerase will bind to a site at the start of the gene, called the promoter, and move along the template strand to produce the mRNA. Through the action of a fundamental organelle called the ribosome (a composite 'molecular machine' consisting of several protein subunits and functional RNAs), mRNA is then translated into a protein. This translation is mediated by transfer RNAs, which allows the mRNA to be read in units called codons, consisting of triples of nucleotides. In eukaryotic mRNA, this step is preceded by mRNA maturation; capping the $3'$ and $5'$ ends of the mRNA, and the removal of intervening sequences, called introns, between coding regions.

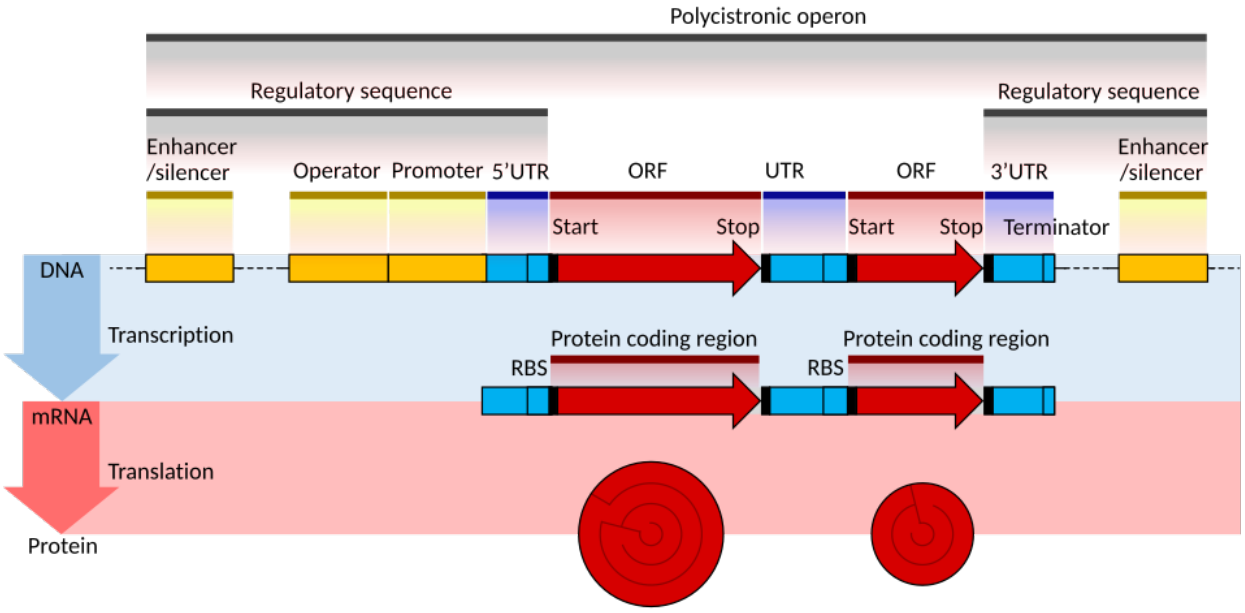


Figure 2.7: Architecture of a prokaryotic gene, from Shafee *et al.* [124].

2.4.2 Genes and Genomes

The majority of DNA is organised in genes and inter-genic regions; coding domain sequences (CDSs) lie within the genes, and it is these which are translated into proteins. Prokaryotic genomes are typically rich in genes, with a coding density of roughly 90% [120]. In prokaryotes, genes can be transcribed together if they lie within the same operon; a contiguous set of genes that are controlled by the same promoter. The prototypical example of prokaryotic gene regulation is the regulation of the Lac operon in *Escherichia coli* [121] - this operon produces the enzyme β-galactosidase, which is used to metabolise lactose. However, *E. coli* will only produce this enzyme in the presence of lactose; otherwise the gene is repressed.

Eukaryotes in contrast can have a large amount of non-coding DNA (over 99% in humans, for example [122]), much of which may be 'junk', though there is debate over how much of junk DNA has been mischaracterised and performs some useful role, and exactly what constitutes junk DNA [123]. Eukaryotic gene regulation is much more complex than in prokaryotes; sites on the DNA, such as an enhancer region, may control whether or not a gene is transcribed. Enhancer regions are just one of several mechanisms for gene regulation, alongside DNA methylation and other epigenetic factors, post-transcriptional regulation, and alternative splicing.

The totality of DNA in a cell is its genome; in prokaryotes this genome is usually in the form of a single circular loop of DNA as previously described (as well as smaller plasmids). In eukaryotes, the genome is structured as nuclear DNA, which sits within chromosomes, possi-

bly in a number of homologous sets (ploidy), as well as organellar DNA in the mitochondria and chloroplasts.

2.4.3 Biological Function Revisited

In Section 2.1.4 we discussed a few important biological functions carried out in the Global Ocean; here we revisit them from a biochemical perspective.

Many proteins coded for in a genome are enzymes, which catalyse specific biochemical reactions. Chained series of reactions form biochemical pathways and cycles, two of the most important of which are carbon and nitrogen fixation. Nitrogen fixation, the process of forming ammonia from N_2, is catalysed by the enzyme nitrogenase, which is encoded by Nif genes. Nitrogenases contain multiple protein subunits, but the presence of the gene for a particular subunit, NifH, if often used as a proxy; it is assumed that the presence of this gene implies the presence of a Nif operon, and therefore the ability to fix nitrogen. Certain nitrifying bacteria may then be able to convert ammonia to nitrates and nitrites. Carbon fixation is the process of converting inorganic carbon sources such as CO_2 to organic compounds. A key biochemical cycle for this process is the Calvin cycle (also called the light-independent reactions within photosynthesis), where the enzyme ribulose-1,5-bisphosphate carboxylase (RuBisCo) plays an important role. This enzyme catalyses a reaction between ribulose-1,5-bisphosphate and CO_2; RuBisCo is thought to be the most abundant enzyme in the world [125], and in the Global Ocean, algae use this pathway (called C_3-photosynthesis) for carbon fixation, augmented by a carbon-concentrating mechanism within an organellar subcompartment called the pyrenoid [126], [127]. In the sulphur cycle, DMSP demethylation is catalysed by enzymes from genes labelled DmdA to DmdD. Additionally, a core set of metabolic functions are present within almost all organisms, such as the glycolytic pathway and tricarboxylic acid (TCA) cycle; these pathways are important in the carbon cycle, both directly (since they interconvert various small organic compounds), and indirectly (by providing a chemical potential from which other reactions can be powered, including the synthesis of adenosine triphosphate, ATP).

2.4.4 Sequencing

Sequencing is the process of identifying the sequence of bases in a section of DNA. In 1977, Frederick Sanger invented the chain termination method of sequencing (Sanger sequencing), which went on to be the dominant form of DNA sequencing for almost 30 years [128]. Though Sanger sequencing is still used in some applications, it has been superseded by next generation sequencing (NGS) technology, such as Illumina sequencing by synthesis.

Illumina Sequencing By Synthesis

Illumina's method of paired-end sequencing by synthesis involves multiple steps of adding a single base at a time to fragments of DNA attached to a structure called a flow cell. The steps involved are described below, and in Figure 2.8.

1. **Add Adapters**: Short known sections of DNA are ligated to the $5'$ ends each DNA fragment. The flow cell is layered with complementary sections of DNA, so by Crick-Watson base pairing, the fragments attach to the flow cell.

2. **Create Mate-Pairs**: Another adapter is added to the $3'$ ends of the fragments, and a single polymerase chain reaction (PCR) step is performed to create complementary strands of each fragment. Adapters are added to these and they are attached to the flow-cell. Each fragment and its reverse complement are now attached to the flow cell.

3. **Bridge Amplification**: Polymer colonies (polonies) are created by using PCR to amplify each fragment of DNA into a localised section of identical single-stranded fragments.

4. **Single Synthesis Step**: The flow cell is then flooded with modified deoxy-oligonucleotide triphosphates (dNTPs), and DNA polymerase, which can bind to the Illumina adapter sequences. The modified dNTPs ensure that only a single step of polymerisation can take place; the modification is a coloured tag in the usual place where further polymerisation would occur. The coloured tag depends on the nucleotide base. The two-colour encoding uses red for A, green for T, yellow for C (combination of red and green), and no colour for G; other Illumina platforms use a 4-colour system. By taking a high resolution photograph of the flow cell and analysing the colours of each polony, we can infer the first base in each fragment.

5. **Tag Cleaving and Repeat**: The flow cell is washed of dNTPs, and the coloured tags of the dNTPs are cleaved chemically, freeing up the site for a further polymerisation step. The previous step can then be repeated, to sequence the next base in each fragment. For paired-end reads, both ends of the fragment are sequenced, and depending on the fragment size, there may be an unsequenced middle section.

The platform used at the JGI to sequence the MOSAiC pilot metagenomes is the Illumina

Novaseq S4, using 151 base pair paired-end reads. The library preparation method for the samples involved fragmenting the DNA to a length of roughly 250 base pairs, and the subsequent removal of any sequences beyond 300 bp, using a size-selection method such as Solid Phase Reversible Immobilization (SPRI). This ensured that paired-end reads would overlap. For low DNA samples (those below 1 ng / μL concentration), the samples could first be enriched by up to 15 cycles of PCR (though fewer than this was preferable, 9 was given as the default number [129]). Rivers [130] provides an overview of the JGI standard and minimal drafts; the typical numbers of read pairs for these being in the ranges 35 to 107 M, and 3 to 17 M, respectively.

The advantages of Illumina sequencing are its extremely high throughput, low error rates, and low sequencing cost per base. Up to 2 billion fragments can be sequenced in parallel in under 48 hours, for a cost of about $40 per Gb, and generating over 1 Tb of sequence data (according to Illumina marketing information [132]). However, there are some drawbacks. The quality of the read drops as the read length increases, as the sequenced bases within a polony slowly desynchronise due to random errors where a sequencing step is skipped. For this reason, reads are kept relatively short, usually 150 or 250 bp in most applications. Systematic errors such as homopolymer errors (errors introduced as or in stretches of identical bases) are more likely than in some other sequencing methods. Depending on the relative luminosity of the various tags used, tags in one polony can be masked by those nearby, and biases may also be introduced in the PCR step.

More recently, third generation sequencing methods have become more popular, such as the nanopore method from Oxford Nanopore Technologies, and SMRT sequencing from Pacific Biosciences. These methods currently have lower throughput, higher per-base cost, and higher error rates than Illumina sequencing, but can generate reads that are tens of kb long (reads using Oxford Nanopore have reached over 2 Mb long). We will not consider third generation sequencing in this thesis, since for the large-scale survey projects such as MOSAiC, it is still not a cost-effective option. However, future work involving samples recovered from the Arctic, particularly live cultures of mixed eukaryotic-prokaryotic communities, could potentially benefit greatly from this kind of sequencing.

A second recent technology used in microbial ecology such as in Doud *et al.* [133], and reviewed in Woyke *et al.* [134], is the application of single-cell genomes to microbial communities. Again, while we do not dwell on single-cell protocols, since these were not used in MOSAiC, we mention the method here for completeness, and because certain single-cell methodologies, particularly regarding the analysis and visualisation of large amounts of data, might be transferable to environmental metagenomics more generally.

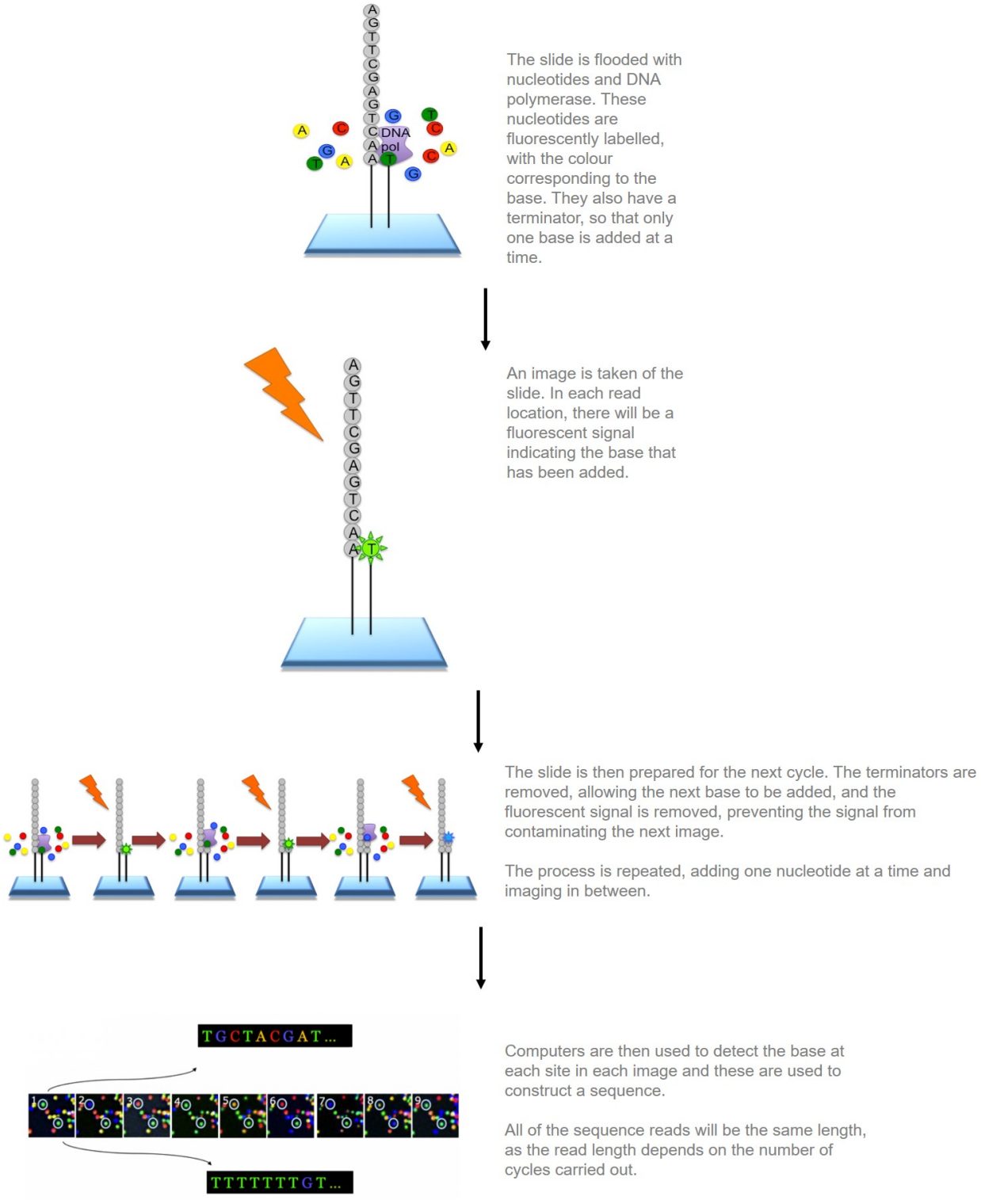


Figure 2.8: Overview of the steps 4 and 5 in Illumina Sequencing by Synthesis for a single fragment of DNA. From EMBL-EBI training material [131].

2.4.5 16S and 18S rRNA Gene Sequencing

The ribosome is such a fundamental organelle that it is found across all three domains of life, and its sequence is extremely strongly conserved. There exist, in particular, stretches of the ribosome that are variable, flanked by stretches which are so strongly conserved that they are near-universal. One such region is the small ribosomal subunit, part of which is called the 16S region in prokaryotes, or an 18S region, found in eukaryotes. This presents an opportunity for phylogenetically cataloguing species, since we can amplify the flanking regions using PCR, with universal primers, and the variable region acts as a kind of 'barcode'. This is the basis of 16S or 18S rRNA gene sequencing, and a large number of taxonomic profiling studies are based on this method, using 16S or 18S rRNA gene sequences as OTUs. Taxa can also be inferred either through tree building, or by searching for sequence similarity based on databases such as SILVA [135]. While the data used in this thesis are not from 16S rRNA gene sequencing, some work from MOSAiC are based on this data, for example Chamberlain *et al.* [136], and currently unpublished work from Muller and Metfies.

2.4.6 Metagenomics

Our main method for studying microbial communities is through metagenomics. Metagenomics is the study of the totality of genomic content in a sample, as opposed to genomics, which would usually seek to understand the genome of a single organism, often by sequencing a clonal culture. While there are other methods available for studying microbial communities (for example with flow cytometry and single-cell sequencing), metagenomics is advantageous in that it is relatively less expensive, requires less technical expertise (compared with flow cytometry, single-cell sequencing, and cell culture), and is less time-consuming. Metagenomics also provides insight into both species that are present in a sample and metabolic potential, something that is impossible with 16S and 18S rRNA gene sequencing, or other marker gene based approaches. This means that by studying metagenomes, we can get an idea of both which species are in a community, and what functions they could carry out.

A key problem when studying microbial communities is that only a small proportion of species in any given sample can be isolated and cultured [137]. Most of the time, researchers cannot grow cultures from individuals taken from a given species in a sample, because they do not know or cannot exactly replicate the conditions necessary for these fastidious individuals to thrive. Finding the right set of environmental conditions for each species would be prohibitively time-consuming and expensive.

Metagenomics provides a partial solution to this problem, by forgoing any kind of isolation and culture of species, and instead proceeding with untargeted sequencing of the total

DNA extracted from a sample. This essentially transforms the problem of isolating individuals in the sample from a biological problem to a computational one, since the extracted DNA contains short fragments of all the genomes of all the species in the sample together. Reassembling this 'genome spaghetti' is a major bioinformatics challenge (called metagenomic assembly), though algorithms exist that are capable of overcoming this problem with reasonable success.

2.4.7 Sequencing a Metagenome

In some respects, sequencing a metagenome is simpler than trying to sequence the genome of a single organism, since there is no need to try to culture a clonal colony. Instead, DNA is extracted from a sample, by lysing cells, removing any cellular debris (for example by centrifugation), isolating and washing the DNA (e.g. adding an alcohol, so the DNA forms a precipitate), and finally eluting the purified DNA in a solution, usually a Tris-EDTA buffer. This is then sequenced as described above. Cell lysis and DNA purification needs to be effective for a diverse range of taxa to avoid introducing biases, and contamination can also occur as small amounts of 'unusual' DNA can be indistinguishable from rare taxa [138]. Sequencing in this untargeted ('shotgun') approach avoids amplification biases; the sequenced DNA can be more representative of an unbiased selection of the total fragmented DNA than for example 16S or 18S rRNA gene sequencing based on PCR amplification of these ribosomal subunits using universal primers, and certainly more representative of a community than trying to isolate and culture individual species. Both processes of DNA extraction and (where applicable) PCR amplification can contribute to bias; in a 16S study of species composition of a known mock community of bacterial species, biases of up to 85% were observed for particular species [139], with biases from DNA extraction and PCR amplification both contributing significantly.

2.4.8 Sequencing Metatranscriptomes

Sequencing the (meta)transcriptome provides a complementary view of the metabolic potential of a sample. Sequencing a genome tells us all the genes encoded by proteins, but the transcriptome tells us which proteins are being produced, and therefore what the organism is actually doing. This is important because many genes are highly regulated, and may only be expressed under certain conditions. This kind of regulation is not observable through genome sequencing alone - to understand what metabolic processes are happening in a community at a particular time, we need to look at the metatranscriptome.

Sequencing a metatranscriptome requires purifying the total mRNA in a sample, and then

reverse-transcribing this into complementary DNA (cDNA), when it can then be sequenced as usual. This process can be challenging; the total RNA in a sample is usually overwhelmingly ribosomal and transfer RNA, and can swamp any signal in the small fraction of remaining mRNA (approximately 1 to 2% of the total) [140]. The purification process can be performed using a specific RNA-Seq library preparation method to enrich for mRNA within a total RNA sample, for example by screening for poly-A tails.

2.5 Discussion

This chapter has provided an overview of the microbial oceanography of the Global Ocean, and the Arctic Ocean more specifically. We have looked at the broad components of the MOSAiC expedition. We have also introduced the most important microbial groups and biochemical functions in the Global Ocean, and the main tool we will use in this thesis to study microbial communities in MOSAiC - metagenomic sequencing. In the next chapter we will describe the bioinformatics tools used to analyse these metagenomic data.

Chapter 3

Bioinformatics Overview

The use of algorithms to analyse biological data, such as phylogenetic relationships between species and genes, or the alignment of homologous protein sequences, emerged even before the invention of Sanger sequencing. These early bioinformatics methods were often computed by hand, and the computational methods pioneered by Dayhoff [141], such as the use of PAM matrices for amino-acid similarity, were some of the first applications of computers in biology. However, the cheap and ubiquitous availability of high throughput sequencing has fundamentally changed how biological data are analysed. Most large-scale analysis requires some sort of computing cluster, for example the supercomputing clusters like Perlmutter and Cori available to the JGI. Certain parts of the processing of high-throughput sequencing data are rote, and have been published as highly automated pipelines, for example [142], [143].

This chapter covers some of the general categories of tools used in bioinformatics, and specifically the bioinformatics pipeline used at the JGI to analyse data for Integrated Microbial Genomes & Microbiomes (IMG/M), which is called the IMG/M Metagenome Annotation Pipeline (MAP).

IMG/M is a repository for annotated genomic and metagenomic projects. The sequencing data for these projects are externally provided by researchers, generated by the JGI, or collected from other databases in the public domain, such as National Centre for Biotechnology Information (NCBI). As such, IMG/M provides an invaluable resource to researchers as a platform to share sequencing data and perform state-of-the-art analyses in a standardised way.

The metagenomes and metatranscriptomes sequenced as part of MOSAiC were all processed by the MAP, and our results make use of the results generated by the MAP, so it is useful to understand the key steps in this bioinformatics pipeline, and their limitations. We will describe the steps of the pipeline that are applicable to Illumina reads.

3.1 Summary of the MAP Pipeline

An Illumina sequencer produces reads in a format called FASTQ (described in Figure 3.1); each read is given a unique identifier, and the sequence of nucleotides is provided along with quality scores (encoded as ASCII characters) for each base that are calculated based on the probability that the base has been called incorrectly. Paired-end reads will produce two FASTQ files, with corresponding pairs in the same order in each file (the second file will be in the reverse orientation of the original DNA fragments). These data are then be run through the MAP [144]. MAP performs some core bioinformatics functions including assembly, read mapping, gene-finding, functional annotation, metagenome binning, taxonomic classification, and several other functions such as scanning for CRISPRs.

The raw read files constitute an enormous amount of data, most of which is not valuable without further processing. The MAP will perform this processing to generate more useful data, including quality controlled reads, assembled stretches of contiguous DNA (contigs), gene and functional annotations, and metagenome bins (see Summary Box 3.1 below). The rest of this chapter explains these steps in more detail.

IMG/M Pipeline Overview

- **Filtering and Quality Control**: Obvious errors and low quality reads are removed.
- **Assembly**: Overlapping reads from different fragments of identical stretches of DNA can be reassembled back into continuous pieces (contigs).
- **Binning**: Contigs can be clustered together into the same *metagenome bin*, based on their provenance. These bins represent fragmented genomes, also called metagenome-assembled genomes (MAGs).
- **Sequence annotation**: Contigs are annotated, identifying coding domain sequences (CDSs), CRISPRs, tRNA and rRNA genes. CDSs are then further annotated with their gene product, often by searching for similar sequences within gene databases.
- **Taxonomic Classification**: Contigs and MAGs are assigned taxonomic classifications.

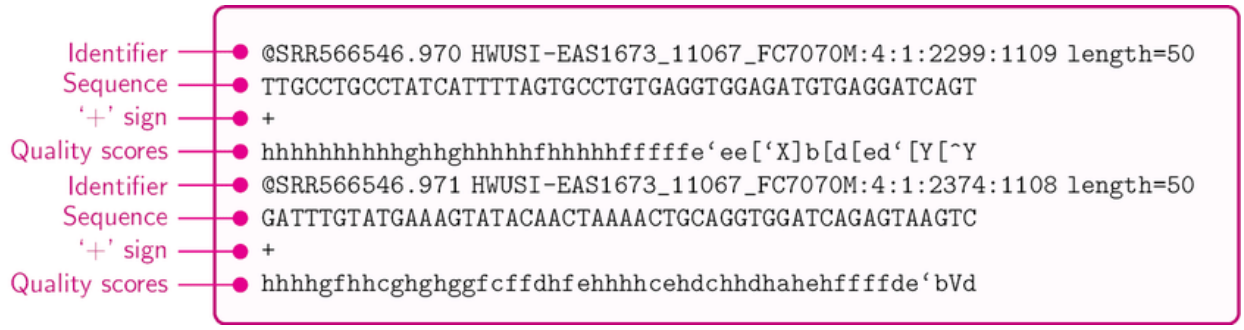


Figure 3.1: Annotated sample of a FASTQ file, from Hosseini *et al.* [145].

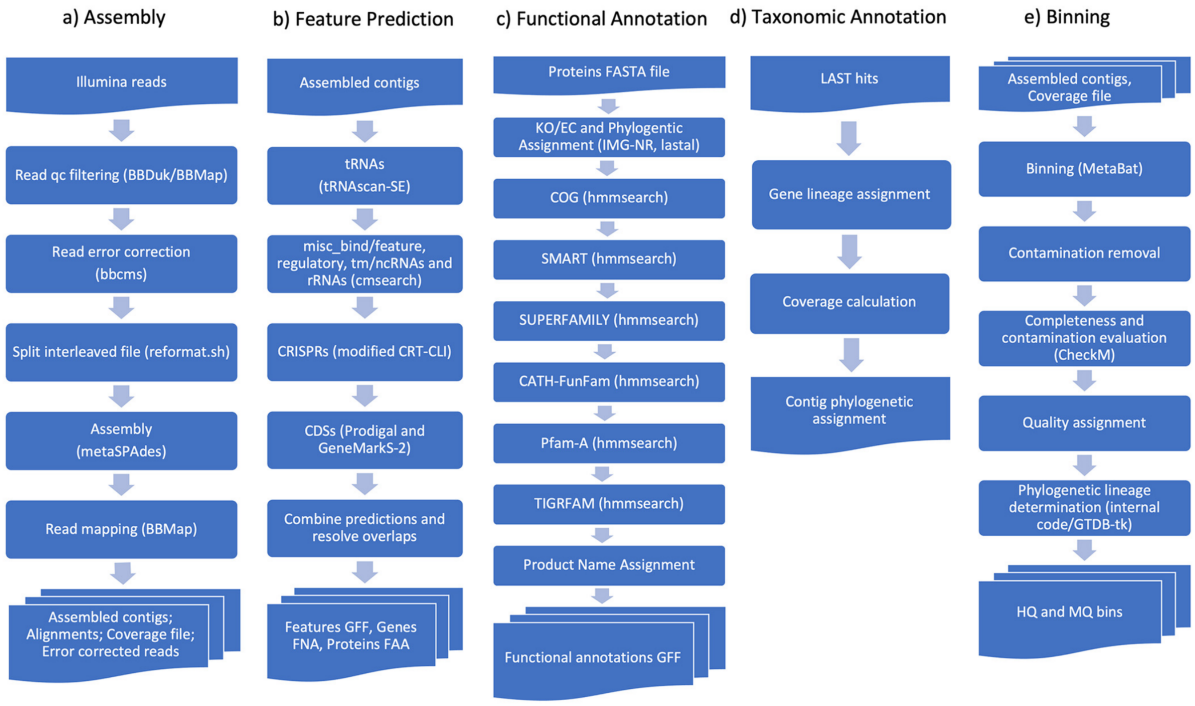


Figure 3.2: An overview of the IMG/M pipeline, figure from Clum *et al.* [144].

3.2 Read Preprocessing and Quality Control

The MAP performs a number of initial checks to remove low quality reads from any subsequent analysis. Additionally, reads from potential contaminants are removed and Illumina adapter sequences are trimmed from the ends of reads. These steps are performed by a suite of software called BBTools [146]. In order, the steps performed are:

1. **Trimming Illumina adapters**: The synthetic sections of DNA used in Illumina library preparation are removed from the ends of reads, and bases where the quality score drops to 15 or below are trimmed off. Homopolymer stretches of 5 G's or more trailing at the ends of reads are also removed - homopolymer errors can occur at the ends of Illumina reads and particularly with G, since G is indistinguishable from 'no signal' in the two-colour-chemistry versions of Illumina sequencing. These steps are performed by the tool BBDuk [146].

2. **Contaminant Removal**: Reads are mapped to contaminants, including the PhiX virus genome. (PhiX is an *Escherichia coli* phage, used in Illumina sequencing as a positive control.) They are then compared to genomes of common biological contaminants, including from human, cat, dog, and mouse genomes, matching at 93% identity or above. This step is performed by BBMap, a fast read alignment method. In Section 3.4.1 we will discuss sequence alignment in more detail.

3. **Remove ambiguous nucleotides**: Any non ACGT characters are replaced by N's, and reads with more than 5 N's are removed. Reads with an average quality score less than 18 are removed, as are reads less than 51 bases long.

3.3 Sequence Assembly

Sequence assembly is the process of assembling reads into longer continuous regions of DNA (contigs). Whereas short-read technologies such as Illumina sequencing by synthesis produce reads with a maximum length of around 300 base pairs, assembly algorithms can stitch together overlapping reads into contigs that can be several thousand or even a million base pairs long. Long contigs are much more informative for a researcher than just reads, as they may contain several genes (or large stretches of a genome). Contigs have a further advantage in that they do not contain the redundancy in information provided by large numbers of overlapping reads. When these information are required, reads can be mapped back to contigs, as will be described in Section 3.4.1. Sequence assembly is therefore often a critical

part of any metagenomics pipeline. We now briefly review the Overlap-Layout Consensus method of assembly, which is intuitively easier to understand, before moving on to a more detailed discussion of de Bruijn Graph sequence assemblers, which are far more commonly used (especially when dealing with large numbers of short reads), and look specifically at MetaSPAdes [147], [148], the assembler used in the IMG/M annotation pipeline.

3.3.1 Overlap-Layout Consensus Assembly

Historically, reads generated through Sanger sequencing were assembled together by identifying overlaps between reads and combining these together. Sanger sequencing was typically applied to clonal cultures, so that a consensus between overlapping reads could be built, removing the majority of sequencing errors. With a smaller number of relatively long reads produced by Sanger sequencing (500 to 1000 bp), assembling together overlapping reads in this fashion was relatively successful, and indeed the Celera assembler [149], which played an important role in the Human Genome Project, used this method.

From an abstract standpoint, the overlap-layout consensus method can be reformulated by building a graph where the vertices represent reads and the edges represent k-mer overlaps (i.e. the last k bases in one read equals the first k bases of the other). In this formulation of the problem, building an assembly is equivalent to finding a Hamiltonian path through the graph.

Assembly Graph Terminology

Graphs are abstract mathematical objects, consisting of a set of vertices, and edges connecting pairs of vertices. In the context of sequence assembly, graphs can be used to represent sets of reads. A summary of this terminology is below:

- **k-mer**: A string of k consecutive nucleotides, for some positive integer k.
- **Graph**: A mathematical object consisting of a set of vertices, connected by edges. Where edges are given an orientation, the graph is directed.
- **Overlap-Consensus Graph**: A directed graph where vertices represent reads. Two vertices A and B are connected by an edge A to B when the last k bases in A equal the first k bases of B.
- **De Bruijn Graph**: A directed graph where vertices represent k-mers, which are connected by an edge if the first $k-1$ bases of one k-mer equal the last $k-1$ bases of the other.
- **Hamiltonian Path**: A path through a graph is a sequence of vertices, where each vertex is connected to the next by an edge. A Hamiltonian path is a path which visits each vertex in the graph precisely once.
- **Eulerian Path**: A Eulerian path is a path between two vertices which passes through each edge precisely once.

Unfortunately, this approach does not scale as the number of reads increases. The main obstacle to this approach is that the Hamiltonian path problem is in the class of NP-complete problems, and there is no known algorithm to solve large problems in this class in a reasonable amount of time.

There are some situations where overlap-based methods are feasible. Third generation sequencing methods such as Oxford Nanopore or PacBio can generate single reads of on the orders of tens or hundreds of kbp in length. With a smaller number of long reads, heuristic methods can make the Hamiltonian path problem more tractable, and this can be favourable compared to the loss of information when splitting long reads into k-mers, as is done in De Bruijn graph methods.

3.3.2 De Bruijn Graph Methods and MetaSPAdes

Since the widespread adoption of next generation sequencing methods, producing upwards of 1 million reads per run, overlap-layout-consensus methods were no longer a practical solution for assembling this enormous number of relatively short reads (150 to 250 base pairs). The solution was to reframe the problem, using a construction called a De Bruijn Graph. To construct the De Bruijn Graph of an assembly, reads are first split into k-mers of a fixed length. Choosing this length affects the performance of later assembly steps, however there is not a methodical way of choosing an optimal k; in the MAP, the values of k chosen are 55, 77, 99, 127, and the resulting contigs from each run are aggregated. Each vertex in the De Bruijn graph represents a k-mer, and two vertices are connected by an edge if the last $(k-1)$ bases in one k-mer are identical to the first $k-1$ bases in the other; i.e. if the two k-mers completely agree except in the first and last positions. One technicality is that vertices need to be counted with repetition - i.e. each vertex must also have a positive integer attached to it denoting how many times that k-mer was seen in the set of reads. Finding an Eulerian path through this graph is equivalent to creating a single contig, so by repeatedly finding Eulerian paths through the graph, eventually an assembly can be created. Whereas the Hamiltonian path problem was NP-complete, there is a simple algorithm for finding Eulerian paths in a graph.

MetaSPAdes [150] is the assembly algorithm used in the MAP pipeline. The method used in MetaSPAdes follows procedure described above, however there are some extra technicalities. The first is that the software must keep track of gaps and insert sizes between paired-end reads; we will say a little more on this in the next subsection. A second technicality is in the disconnection of 'bubbles' in the k-mer graph, which represent different variants of otherwise highly similar sequences, possibly representing a sequencing error but also possibly representing a nucleotide variation between two otherwise highly similar strains sharing an almost identical genome. Most genome assembly programs would remove the less occurring branch of a bulge; MetaSPAdes retains the information about the coverage of variants, but removes the bulge from the graph to simplify the De Bruijn graph. MetaSPAdes also uses heuristics based on the coverage information to try to resolve longer repeats (appearing as loops in the De Bruijn Graph); this is done in a module called exSPAnder [148].

3.3.3 Scaffolds

When the order of several contigs is known, and the lengths of the gaps between each contig has already been determined (or estimated), these contigs are grouped together with the correct number of interleaving gap characters to form a structure known as a scaffold.

For the purposes of this thesis, we will not dwell on scaffolding, for two reasons. The first is that this is usually taken care of by assembly software automatically. Secondly, scaffold creation is not relevant for the MOSAiC samples, since the sequencing libraries (described in Chapter 2) were constructed as overlapping paired ends, and reads with long stretches of gap characters are therefore filtered out in the initial IMG/M quality control stages. We will simply note that the creation of scaffolds can often be part of assembly pipelines. For the purpose of this thesis, scaffolds and contigs are used semi-interchangeably, and we will note when the distinction between the two is relevant.

3.3.4 Assembly Quality and Statistics

A simple metric used to gauge the quality of an assembly is the N_{50}. N_{50} is defined as the largest L such that the set $S = \{c \in C : \text{len}(c) \geq L\}$ covers 50% of the assembly; i.e. 50% of the assembly sequence is contained in contigs of length N_{50} or greater. Here, C denotes the set of all contigs, and $\text{len}(c)$ is the length of contig c. This measure is generalised to N_X for any X% cover of the assembly. A larger N_{50} is indicative of longer contigs and is used as one metric to gauge assembly quality.

Often, when only scaffolds more than a certain threshold (e.g. 500 bp) are retained, the percentage of reads mapped to the assembly is provided as another summary statistic, as is average coverage. Coverage depth (the average number of times a base in a contig is covered by reads) is an important statistic for gauging assembly quality; when sequencing a genome it is often advised to sequence to a depth of 30x coverage at least. Since a metagenome is a mixture of multiple species of different relative abundances, rare species might be covered at a relatively lower depth than the more abundant species, which makes them harder to assemble and more likely to contain sequencing errors.

3.4 Sequence Search and Sequence Similarity

Two fundamental operations in bioinformatics are searching for a target sequence in a large database of genomes and sequences (sequence searching), and comparing sequences to assess how similar they are (sequence similarity). Finding similar DNA sequences to a sequence of interest is often the first step in understanding the provenance of that sequence, its function, and how it might be evolutionarily related to other sequences (their homology). The most influential search tool is Basic Local Alignment Search Tool (BLAST) [151].

Similar sequences are often compared to one another, either to try to find a possible evolutionary relationship between them (phylogenetics), or as the first step of understanding

their function. The first step in this comparison is a sequence alignment - matching bases or amino acids from one sequence with the other, and where necessary adding gap characters.

3.4.1 Sequence Alignment

Sequence alignments fall into two categories, global alignments and local alignments. Local alignments involving matching subsections of longer sequences to one another, and ignoring the irrelevant non-aligned sections, whereas in a global alignment, the full length of a set of sequences are matched with one another. Global alignments are further split into multiple sequence alignments (finding an alignment between a set of ≥ 3 sequences), and pairwise alignments (aligning just 2 sequences). In all cases, the aim is to optimise an alignment score, based on a scoring matrix representing a cost for mismatching different symbols (either amino-acids or nucleotides).

Different algorithms are appropriate for each situation. For pairwise alignments, there are exact dynamic programming algorithms; the Needleman-Wunsch algorithm [152] can be applied to global alignments, and the Smith-Waterman algorithm [153] for local alignments. These two algorithms are conceptually very similar. Both are dynamic programming algorithms, which find an optimal pairwise local or global alignment in $O(NM)$ time (for sequences of length N and M). They are based on the additive properties of the alignment scores that they are trying to optimise, and work by recursively iterating through a table of all possible sub-alignments of the given alignment problem. Although there are still new optimisations for these algorithms on certain hardware [154], these pairwise alignment problems are essentially completely solved, and the alignments they generate are optimal.

On the other hand, generating a multiple sequence alignment (MSA) is a much harder problem that is far from solved except in special cases. The analogue of the Needleman-Wunsch algorithm for multiple sequences does produce an optimal alignment, but in $O(N^K)$ time, for K sequences of length of order N - which even for moderate K and N is computationally intractable. Heuristic methods are therefore used, and many different algorithms (e.g. based on hierarchical pairwise alignments such as MUSCLE, and ClustalΩ [155], [156], or Fourier transforms as in MAFFT [157]) are used to generate good approximate MSAs, but there is no known general method to produce an optimal MSA in a reasonable amount of time. Although MSAs are extremely important when considering phylogenetic classifications and sequence homology, we will not go into this topic in any further depth.

3.4.2 Sequence Homology: BLAST and HMMer

The tools of local and global alignment can be used to quickly compare pairs of sequences, but modern research consists of scanning new sequences against databases containing potentially millions of target sequences. Doing this quickly requires a completely different approach. Basic Local Alignment Search Tool (BLAST) is a workhorse of bioinformatics, and is based on the principles of exact k-mer matching of seed sequences, followed by extension of these matches using a variant of the Smith-Waterman algorithm. Finding exact matches, even in a large database, can be performed using hash tables, which only require a constant-time look-up to search for a particular query sequence. Depending on the target database, there might be a large number of false positive matches between k-mers in the query and target sequences; hence the need for extension (i.e. local alignment) between the query and matching sequences. The quality of this alignment, represented by the alignment score, can also be transformed to represent a metric for the likelihood that such a match appeared through chance (an E-value) once the size of the database has been taken into account. A different score, the bit-score, is essentially just a scaled alignment score, taking into account various statistical properties of the scoring matrix.

The formulae for these scores are:

$$\text{bit-score} = \frac{\lambda S - \ln K}{\ln 2}$$

$$E = m \times n \times 2^{-\text{bit-score}}$$

with E the E-value, S the raw alignment score, m and n the lengths of the query sequence and database respectively, and K and λ empirically derived constants dependent on the particular scoring matrix used.

The search carried out by BLAST is an example of mapping; a routine bioinformatics problem involving both sequence search and local alignment. In mapping, reads are matched to potentially homologous 'hits' in a database (e.g. an indexed assembly) and then locally aligned to each hit, with only the best scoring hit or hits after alignment retained. There are a large number of different strategies for read alignment (Alser *et al.* [158] provides a review), but most are based on one of two methods; either the hashing and extension method described above, or alignment based on the Burrows-Wheeler transform, used in several popular aligners such as Bowtie and Burrows-Wheeler Aligner (BWA) [159], [160].

Hidden Markov models (HMMs) (Summary Box 3.4.2) are another ubiquitous tool used to compare sequences, and the most commonly used HMM based tool is HMMer [161]. HMMs are state-based graph models, which rely on the concepts of using observed states to

infer a hidden sequence of underlying states. HMMs apply to sequential non-independent data, of which nucleotide and amino-acid sequences are key practical examples.

Hidden Markov Model Terminology

Hidden Markov models (HMMs) are state-based graph models used to identify a sequence of transitioning underlying or hidden states from an observed sequences. Some of the terminology and key algorithms used when discussing HMMs are below:

- **Markov Chain**: A memory-less probabilistic model used to model discrete time-series or sequential data. They are defined by a discrete set of states and a transition matrix between those states, determining the likelihood of moving from one state to the next.
- **Transition Matrix**: A matrix defining the probabilities of moving from one state to another in a Markov chain.
- **Hidden States**: A discrete set representing unknown or as-of-yet unidentified states which influence the observed sequence of variables. Together with a transition matrix, these states define the hidden part of a Hidden Markov model (HMM).
- **Emission Probabilities**: The conditional probabilities of observing state j given that the system is in hidden state i. A set of hidden states, along with emission probabilities mapping to a set of observable states, constitute a Hidden Markov model (HMM).
- **Viterbi Algorithm**: A dynamic programming algorithm used to identify the most likely sequence of hidden states, given an HMM and an observed sequence [162].
- **Baum-Welch Algorithm**: A dynamic programming algorithm used to train optimal transition and emission probabilities in an HMM [163].

In the context of gene-finding, the observed states of a HMM could correspond to the observed sequence of nucleotides, and the underlying hidden states are whether or not the nucleotide is part of a coding sequence or not. (In this example, there might only appear to be two hidden states, however the actual HMMs used by gene-finding programs, for example GeneMark [164], are generally not this simple.)

For both BLAST and HMMer, query sequences need to be compared to references in

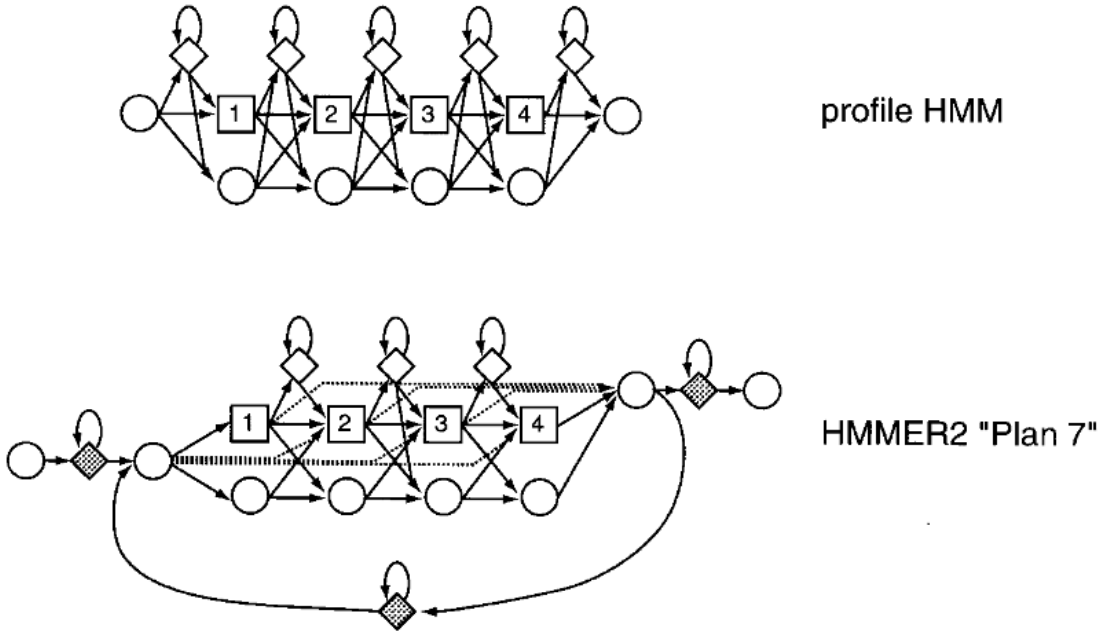


Figure 3.3: Examples of HMM state space architectures, from Eddy [161]. The diamond and circularly shaped states represent insertions and deletions respectively.

sequences databases. Sequence databases such as the NCBI NT and NR nucleotide and amino acid databases, as well as reference genome databases like the NCBI RefSeq [165], MMETSP (for eukaryotic transcripts) [166], MarRef and MarDB databases [167], are useful to catalogue known prokaryotic diversity, both in general, and for marine habitats in particular. For Hidden Markov models, protein databases are used to make profile HMMs, where a multiple protein alignment is used to derive a model of the form shown in Figure 3.3. One example of such is the Pfam database [168]. However, these databases have biases, for example over-representing well-studied environments such as the human microbiome, and with a bias towards easily cultivable species.

3.5 Sequence Annotation

Having covered the prerequisites of sequence search and sequence homology, we return to the MAP pipeline, which uses these tools to locate and annotate genes. The process of finding coding sequences (CDSs) is typically performed by the tools Prodigal and GeneMark [169], [170], which use hidden Markov models to identify open reading frames in contigs, as well as obvious markers such as start codons and the absence of stop codons in the middle of the

reading frame. MAP then uses a consensus from these tools to identify CDSs.

Gene annotation is again done using HMMs, this time with the tool HMMer, and using profiles built from the KO, TIGRFAM, COG, SMART, SUPERFAMILY and Pfam-A databases [168], [171]–[175]. These databases differ in their scope and granularity: Pfam-A catalogues protein domain families built from curated seed alignments; COG (Cluster of Orthologous Genes) groups proteins into orthologous clusters with a focus on comparative genomics; TIGRFAM provides manually curated HMM profiles for specific (mostly prokaryotic) protein families with known well-defined biological roles; KO (KEGG Orthology) maps genes to metabolic pathways and biological systems; SMART specialises in signalling and extracellular protein domains; and SUPERFAMILY classifies proteins into structural superfamilies based on known three-dimensional structures. Together, these databases provide complementary layers of functional annotation. Profile HMMs build on regular HMMs by using multiple sequence alignments from similar genes to build a profile of that gene, using the Baum-Welch expectation maximisation algorithm to infer maximum-likelihood transition and emission probabilities for the sequence. This can then be used to scan query sequences for regions fitting that profile, using the Viterbi algorithm.

Other feature predictions than tagging CDSs are done similarly, for example tRNAs are identified with tRNAScan [176]. While this tool is based on covariance models rather than HMMs, these are still state-based graph models, and again can use variants of the Baum-Welch and Viterbi algorithms to optimise its weights, and estimate most likely hidden states. Ribosomal RNA genes are annotated using the tool INFERNAL and profiles from the Rfam database, and CRISPRs are annotated with the tool CRT-CLI [177]–[179].

3.6 Phylogenetics and Taxonomic Classification

When comparing sequences to try to identify those with a recent common ancestor, there are two general methods available: phylogenetics, and taxonomic classification. Two major aims of phylogenetics are to either generate a plausible tree of related operational taxonomic unitss (OTUs) based on estimated evolutionary distances, or place a new sequence on an existing such tree. The field of phylogenetics is incredibly rich, and there is not space here to go into detail about either the phylogenetic placement problem or tree-building, but we will very briefly summarise the principles used in maximum-likelihood tree-inference algorithms such as FastTree or IQTree [180], [181]. These methods estimate a maximum-likelihood tree by stochastically searching through tree space until they settle on a (locally) optimum tree. Optimality is defined through maximising a likelihood score, based on a forward model of mutation rates (such as a general time-reversible model) and a molecular

clock hypothesis. These models generally assume that the mutations on each site follow independent transitions, obeying the Markov property.

Taxonomic classification is a related but easier problem than tree building or phylogenetic placement. In this problem, the aim is to assign as specific a taxon as possible to a given sequence. Placing a sequence in a sufficiently refined tree will automatically resolve its taxonomy. However there are several methods of assigning taxonomy to sequences at the level of reads, genes (and contigs), or whole genomes, and most of these avoid needing to solve the computationally harder phylogenetic placement problem. At the simplest level, some of these use tools we have already discussed do this; binning based on k-mer frequencies can be used, as can BLAST searches of genes to find matches to known species in sequence databases. Placement of genes on an existing tree is also possible and computationally simpler than generating a tree from scratch; this method is employed by pplacer [182].

3.6.1 Taxonomic Classification in the IMG/M pipeline

To assign a taxonomy to contigs, MAP uses a tool called LAST [183] to assign a taxonomy to each CDS, based on the best hits to the NR database [184]. Based on all the CDSs on a contig, MAP will then try to take a consensus, down to the lowest taxonomic rank at which over 50% the CDSs on the contig agree on a taxonomy. Although LAST has slightly worse type 1 and 2 error characteristics than BLAST (and other similar methods such as mmseq2 [185]), it scales well with extremely large numbers of sequences.

While the IMG/M pipeline does not assign a taxonomy to reads, this can be done; Kraken [186] attempts to do this in a reference-free way (i.e. without trying to BLAST sequences against databases such as the non-redundant protein database NR) by assigning reads to a taxon based on its k-mer frequencies, utilising its own database of k-mer frequencies per taxon.

3.7 Metagenome Bins and MAGs

After reads have been assembled into contigs, the next stage is to further group together contigs which can be identified as coming from the same species, or individual. These groups are called metagenome 'bins', each bin being a putative fragment of a whole genome. Bins which pass a set of quality selection criteria can be referred to as MAGs, or metagenome-assembled-genomes.

3.7.1 Binning

There have been two general approaches to binning; either reference-based, or reference-free methods. Reference based methods, such as using a combination of the software DIAMOND and MEGAN [187] focus on aligning coding regions to known protein databases, and in general create new bins similar to those in the reference database. We will focus instead on the unsupervised method used by MetaBAT2 [188], since this is the software used in the IMG/M pipeline, and is more suitable for metagenomes with potentially many novel species. However, it is worth noting that at a high level, most unsupervised methods use essentially the same two step to generate bins:

1. Calculate a metric on the set of contigs (usually derived from 4-mer distributions, also called tetranucleotide frequencies, and from coverage distributions).

2. Use a clustering algorithm to group the contigs into bins.

The reason why metrics derived from these criteria (tetranucleotide frequencies and coverage) work at all is that different species tend to use amino acids in varying proportions, and even when they use the same amino acids, have different preferences for synonymous codons [189]. Coverage is also a reasonable metric to use in binning; on average one might expect that each contig would be sequenced at a depth proportional to the abundance of the parent genome (though this does not account for repeating sequences, or sequences shared between species).

MetaBAT2 only varies slightly from the above two steps in that when species might be present across multiple metagenomic samples, it can use a hybrid metric derived from both tetranucleotide frequencies and coverage depth across samples (as measured by mapped reads across all samples). The clustering algorithm in MetaBAT2 is graph-based, and does not require any kind of dimensionality reduction in the set of contigs. Instead, it is based on a form of label propagation algorithm. Other tools such as VizBin [190] use dimensionality reduction to provide a visual aid for the generated metagenome bins.

Some more recent binning algorithms use machine learning methods cluster contigs; these are usually based on the same input data of tetranucleotide frequencies and coverage information, but use a neural network or other classifier-based approach to generate bins. The binners VAMB, Comebin, and SemiBin, all make use of a latent space approach to improve binning [191]–[193]. VAMB, for example, uses a variational autoencoder architecture, where the high-dimensional tetanucleotide frequency (TNF) and coverage data are 'squeezed' through a lower dimensional latent embedding; this helps with clustering, which is done by an iterative medoid clustering algorithm.

Although in theory, these methods are agnostic about binning prokaryotic or eukaryotic contigs, in practice binning algorithms are less effective for generating eukaryotic MAGs. This is down to the nature of eukaryotic genomes, which are usually much larger than those of prokaryotes (can be a number of Gbp rather than a few Mbp), have a more complex gene structure (containing introns and exons), and have more repetitive and low entropy regions of DNA, which are much harder to assemble and bin. To overcome these problems, eukaryotic binning is often more manual and time intensive, and a first step is to identify eukaryotic contigs, using a program such as EukRep or Tiara, which use support vector machines and deep neural networks respectively to classify contigs from different domains [194], [195].

3.7.2 Binning Quality

Once metagenome bins have been created, they must be assessed for quality. The MIMAG standards [196] lists a minimum set of criteria to determine what constitutes a low, medium, or high quality MAG. These criteria are based on single copy marker genes (SCMGs); genes which appear exactly once in a lineage or clade. Given a set of SCMGs, which needs to be selected suitably, bin completeness is then defined as the percentage of SCMGs that are present, and bin contamination is defined as the percentage of SCMGs that are duplicated, counting repetitions. The definitions of low quality, medium quality, and high quality MAGs are then given simply as thresholds on these two criteria. High quality MAGs additionally require the presence of genes for the 16S, 23S, and 5S ribosomal subunits, and at least 18 tRNA genes.

Given a metagenome bin, selecting the right set of SCMGs to perform this analysis fairly is not always straightforward. For prokaryotes, the CheckM software is routinely used to estimate bin quality. For eukaryotes, EukCC, or BUSCO, are used instead. In all three cases, the software operates in roughly the same way. In CheckM, the metagenome bin is placed on a precomputed genome-tree of high-quality genomes, decorated with lineage specific SCMGs (within a lineage, a gene is determined as a SCMG if it appears as single-copy in $\geq$97% of the genomes). This initial placement is done by scanning for an initial set of 43 marker genes in the bin using HMMer, placing these in corresponding gene-trees using pplacer, and taking a consensus. Based on this initial lineage designation, completeness and contamination are then estimated based on the percentage of SCMGs present and duplicated for that lineage.

A similar marker-gene driven approach is also used to assign a taxonomy to MAGs. The MAP uses the tool GTDB-Tk [197], which uses a consensus of marker genes identified in the MAG and placed on a reference tree using pplacer [182], followed by using the criteria of relative evolutionary divergence and average nucleotide identity to establish taxonomy,

possibly down to the species level.

3.7.3 Limitations of MAGs

MAG analyses are popular and can be extremely powerful since they allow for the synthesis of species and functional information. However, as with many methods, they have some limitations. MAGs capture a biased sample of the diversity within a metagenome, in particular skewed towards the fraction of a community that is assembled well [198]. This bias is made worse by the fact that most MAG analyses will only retain the medium and high quality bins. This trades off having a quality-controlled set of MAGs, for missing rarer or hard-to-bin species. Furthermore, many binning algorithms specifically target only the prokaryotic fraction of the community. The binning algorithms in MaxBin, SemiBin, and ComeBin [193], [199], [200] all use sets of prokaryotic SCMGs to refine the bins they generate. Not only does this mean that these methods may specifically avoid eukaryotic and viral bins, but also that bins are both generated and assessed using the same criteria - the numbers of SCMGs - which may lead to a form of overfitting or search bias.

An alternative to a MAG-based approach is to perform functional analyses directly on assembled contigs, without binning. In this approach, open reading frames (ORFs) are predicted on all contigs, and functional domains (e.g. Pfams, KEGG annotations) are assigned to individual genes. This avoids the quality constraints and potential biases of the binning step, capturing a greater fraction of the community — including rarer or harder-to-bin species. However, the contig-based approach lacks taxonomic resolution at the genome level; while contig-level taxonomy can be estimated (e.g. via LAST or DIAMOND BLAST), it is difficult to assign a gene's functional context to a specific organism. MAG-based approaches sacrifice breadth of coverage in exchange for genome-resolved information, linking function to taxonomy in a way that contig-based analyses cannot.

3.8 Tools from Numerical Ecology

Although there are often model and keystone species within certain ecosystems, such as *Fragilariopsis cylindrus* in the Arctic, microbial ecologists are more often interested in larger communities of microbes, especially in metagenomics where the untargeted approach to sequencing is a definite advantage. However, this presents a new challenge; microbial community composition data is often extremely high-dimensional, with the total number of species present within a sample (i.e. the species richness) possibly on the scale of 10^3 to 10^6 [201]–[203], depending on the environment being sampled. In a marine context, species richness ranges widely but can be in the thousands within a single sample of seawater; for example [204], where species richness was estimated as high as 1,200. We therefore require tools to be able to analyse these data in a principled manner. Fortunately, there are a number of tools from mathematical ecology and data science which can help us explore high-dimensional datasets such as community compositions and gene abundances, without cherry-picking factors that we believe should have some importance. Additionally, there are several methods of calculating abundance; reads per million (RPM), reads per kilobase million (RPKM), and transcripts per million (TPM) being the three most common. We will begin with an outline of the general principles that are applied to community composition data, and then examine some of the methods most commonly used to analyse ecological diversity, in increasing order of complexity.

3.8.1 Measures of Abundance for Community Composition Data

The relative abundance of any particular gene, contig or MAG can be measured based on the number of reads mapped to the feature in question. However, as the total number of reads sequenced within each metagenome is not biologically informative (based in part on budgetary constraints, for example), it is necessary to normalise these counts appropriately to compare the abundances of these features across samples. The three most common ways of doing this normalisation are RPM, RPKM, and TPM. Reads per million (RPM) is the simplest form of normalisation, which simply scales the read counts per million reads sequenced per sample. Reads per kilobase million (RPKM) further scales this value to account for the length in kilobases of the feature (either gene, contig, or MAG). TPM normalises by length first, and then converts that to a per-million score; this measure if most often used in RNA-seq experiments. Whether to use RPM, TPM or RPKM depends on the situation and the information to be conveyed. RPKM values can be viewed as a proxy for particle count (normalised per million reads); RPM is a more faithful representation of the relative

amounts of DNA present. In this thesis, we use RPKM in situations where particle count is more relevant, e.g. when comparing gene abundances (such as in Chapter 5). When comparing MAGs, especially when comparing across domains of life (eukaryotes vs prokaryotes vs viruses) where genome size and biomass per particle vary by many orders of magnitude, we use RPM as a measure of relative abundance. This is because particle count is not wholly informative when considering eukaryotic and prokaryotic primary production; the number of eukaryotic cells is dwarfed by the number of prokaryotic cells (and both these numbers are again dwarfed by the number of virus particles), however, the eukaryotic contribution to biogeochemical cycling and primary production is certainly not negligible. In this situation, RPM is an imperfect but more meaningful measure of abundance than RPKM.

3.8.2 Principles of Dimensionality Reduction

Dimensionality reduction is the process of taking some high-dimensional input data, and reducing it to an output that can be more simply analysed or visualised. In the extreme, we can reduce the information in a dataset to a single dimension, i.e. a number, such as a mean, diversity index, or count value. These single-number statistics (also called indices) are simple but can be extremely informative, as they succinctly quantify an important feature of the data. A key family of indices in ecology are those that measure α diversity, which we cover in section 3.8.3.

A more sophisticated approach is to reduce high-dimensional data to two or three dimensions, which can be visualised in a scatter plot. We have already come across some examples of dimensionality reduction; in Section 3.7.1 we noted that the proportions of 4-mers in contigs occupies a 135 dimensional space, and that the program VAMB uses a VAE to reduce this complexity into a lower dimensional latent space where the contigs are clustered to produces bins and MAGs. Other examples are abundance tables, where a large number of species might have different abundances at different frequencies across many samples; standard methods such as PCA and PCoA (see Section 3.8.5) allow us to plot these data in two dimensions.

There are two central concepts in dimensionality reduction, that of a distance (or metric), and of an embedding. A distance is used to quantify the difference between each pair of input data. Formally, a distance is a function d which takes any two pairs of inputs $\boldsymbol{x}$ and $\boldsymbol{y}$ and generates a number, satisfying the following criteria:

1. Zero distance to self: $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ if and only if $\boldsymbol{x} = \boldsymbol{y}$

2. Positivity: $d(\boldsymbol{x}, \boldsymbol{y}) > 0$ for all distinct points $\boldsymbol{x}$ and $\boldsymbol{y}$

3. Symmetry: $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$ for all pairs of points $\boldsymbol{x}$ and $\boldsymbol{y}$

4. Triangle inequality: $d(\boldsymbol{x}, \boldsymbol{y}) + d(\boldsymbol{y}, \boldsymbol{z}) \geq d(\boldsymbol{x}, \boldsymbol{z})$ for all triples of points $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$

Examples of distance functions include: the Euclidean distance, the Jaccard distance, and the Aitchison distance. We will develop some of these further in Section 3.8.4.

The triangle inequality is often fairly trivially satisfied, however there are a few cases (such as for the measure defined by Bray and Curtis [205]) where this does not necessarily hold. In those cases, the measure is called a dissimilarity rather than a distance. When we have a dataset X, along with a distance function d, we can call X a metric space. Instead of considering a dataset X with data points $\boldsymbol{x}_i$ with distance function d, it is sometimes more convenient to consider a distance matrix D_{ij} indexed by the data points instead; i.e. the square table defining distances between each pair of points.

Building on distance and dissimilarity is the concept of embedding. The idea behind an embedding is to represent points in a high-dimensional metric space X, by points in a lower dimensional space Y, so that the distances between points in X are the same (or as close as possible) to the distances in Y. Normally, Y will just be points in a two dimensional plane, so that we can then visualise them on a scatter plot.

Typically, there is no way to perfectly embed points from a high-dimensional (e.g. hundreds of dimensions) space down to two dimensions without warping or distorting some of the distances. Different methods and metrics have therefore been developed, which accentuate certain features of the 'true' distances in the embedded space, at the expense of others.

The above principles of dimensionality reduction are fairly universal, i.e. they can apply to any sort of high-dimensional data. In the next sections we will apply these principles specifically to the task of analysing ecological diversity. We will begin with the simplest measures, α diversity indices, and moving on to more complex measures of between-sample diversity, i.e. β diversity.

3.8.3 Alpha Diversity Indices

The simplest way to describe a dataset is with a single number. For an ecological dataset of species, the numerical count of the number of distinct species present is the species richness. However, this fails to take into account variations in abundance, i.e. the differing numbers or proportions of each species (evenness). There are various different indices of α diversity, which give a measure of the amount of diversity present within a sample; in general, the main thing that distinguishes different diversity indices is how much they weigh (or ignore) relatively rare species compared to the most prevalent. A few diversity indices are listed below.

The most commonly used is the Shannon index, which borrows the formula for entropy from information theory and thermodynamics, applied to the species distribution. The Shannon index sits at a theoretical midpoint between two extremes of diversity measurement. It does not give abundant species an additional weighting compared to rare species, but neither does it treat both common and rare species equally (which richness does).

Alpha Diversity Indices

Alpha (α) diversity is a single numerical measure of the diversity present at a site (within-sample diversity). This quantity is measured by a diversity index; for a set of species labelled from 1 to S, with count of the ith species as n_i, a few of these indices are as follows:

- **Species richness**: This is the number S, the total number of species observed in the sample.
- **Shannon index**: $H = -\sum_{i=1}^{S} p_i \ln p_i$, where $p_i = \frac{n_i}{N}$, is the proportion of the ith species observed, $N = \sum_{i=1}^{S} n_i$ is the total number of individuals.
- **Simpson index**: $\lambda = \sum_{i=1}^{S} p_i^2$, again with N and p_i defined as above.
- **Inverse Simpson index**: ${}^2D = \frac{1}{\lambda}$, the reciprocal of the Simpson index.
- **Hill numbers**: ${}^qD = (\sum_{i=1}^{S} p_i^q)^{\frac{1}{(q-1)}}$, a generalised formula for diversity indices with parameter q. When $q = 1$, this is equivalent to the Shannon index, for $q = 2$, it is equivalent to the (inverse) Simpson index. Smaller values of q put less weight on the most abundant species. For $q = 0$, it is equivalent to (inverse) richness.

3.8.4 Beta Diversity Indices

Beta (β) diversity describes between-sample diversity. For each sample, we have a vector consisting of counts (or relative abundances, in the case of metagenomic data) of each species - these are community composition vectors. Our aim is to measure distances or dissimilarities between the community composition vectors, so that we can compare diversity across samples. There are several dissimilarities and distances that are used in metagenomics. We will discuss five of these, beginning with the simplest distance function as a pedagogical example, Euclidean distance, before moving on to four of the most commonly used in metagenomics: Bray-Curtis dissimilarity, the Jaccard index, UniFrac distance, and Aitchison distance.

Formulae for Common Beta Diversity Indices

For two samples, X and Y, with community composition (in terms of relative abundance) of the ith species given by x_i and y_i respectively, assembled into community composition vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, we have the following formulae for β diversity distances or dissimilarities between them:

- **Euclidean distance**: $||\boldsymbol{x} - \boldsymbol{y}||_2$. This is the straight-line distance between two points in space. In two and three dimensions, it is the length of the line connecting the two points.
- **Bray-Curtis dissimilarity**: $1 - \frac{2\sum_{i=1}^{S} \min(x_i, y_i)}{\sum_{i=1}^{S}(x_i + y_i)}$
- **Jaccard distance**: $1 - \frac{|X \cap Y|}{|X \cup Y|}$, the sizes of the intersection over the union of the two sets of species - discounting abundances.
- **Weighted UniFrac distance**: Weighted UniFrac distance takes into account phylogenetic placement, and requires that the species are placed onto a phylogenetic tree.
- **Aitchison distance**: $||\mathrm{clr}(\boldsymbol{x}) - \mathrm{clr}(\boldsymbol{y})||_2$, where clr is the centred log-ratio, i.e. $\log(\boldsymbol{x}) - \overline{\log(\boldsymbol{x})}$. This is, equivalently, the log of the vector $\boldsymbol{x}$, once the components of $\boldsymbol{x}$ have been scaled by their geometric mean.

Euclidean distance is just the straight-line distance between the two points. The problem with using it as a metric when comparing community compositions is that it tends to overemphasise the importance of the most abundant species, and it is possible to end up with counter-intuitive results such as the Orlóci paradox [206], where two samples with no species in common might be closer together than a third sample sharing all species with the first. Bray-Curtis dissimilarity is a measure which tries to overcome this, by scaling the contributions from each species by the total number of individuals across both samples. Bray-Curtis dissimilarity is often used in community ecology, even though it is not a true metric, as differences do not obey the triangle inequality.

Two other metrics commonly used are the Jaccard distance and UniFrac distance. The Jaccard distance is only a metric on the boolean presence-absence datasets (i.e. the species richness information), and does not take into account abundance information. UniFrac distances are weighted by phylogenetic information, and computes distance taking into account branch lengths on a phylogenetic tree. This offers quite a rich metric, though phylogenetic

distances between different domains of life are a topic of active research (for example see [207]), making it tricky to build an agreed-upon species tree for mixed communities.

3.8.5 PCA and PCoA

Principal component analysis (PCA) and principal coordinate analysis (PCoA) represent two common ways of embedding dissimilarities and distances into 2 or 3 dimensions. Beyond being simply embeddings, they are examples of ordinations, meaning that the components of the embedded space are themselves physically meaningful. We treat each of these in turn.

PCA

Given an abundance matrix $\boldsymbol{M}$, where the rows are samples and the columns are species (so component M_{ij} is the abundance of species j in sample i) there is a mathematical theorem called the spectral theorem which guarantees that we can decompose $\boldsymbol{M}$ into a form

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and $\boldsymbol{\Lambda}$ is a diagonal matrix, consisting of decreasing non-negative entries called eigenvalues. This special decomposition is unique and is called the singular value decomposition; when we apply it to our dissimilarity matrix, we get an ordered set of vectors as the rows of $\boldsymbol{U}$, called principal components. We can keep the first few principal components, and use this as a dimensionality reduction tool. The references [208], [209] provide some good introductions to PCA.

PCA has a few theoretically optimal properties. The data are projected into a subspace where:

1. The components are uncorrelated.
2. The original data are a linear combination of the full set of components.
3. Components are ordered in decreasing order of how much variance they 'explain' in the original dataset, this value is measured by the singular value corresponding to that component.
4. (Ekhart-Young-Mirsky theorem) The differences in Euclidean distances between the original data and the transformed data are minimised in the following way:

$$\min_{\boldsymbol{y}} ||D_{ij} - ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2||_F$$

where the matrix norm $||.||_F$ is the Frobenius norm, D_{ij} is the Euclidean distance matrix of the high-dimensional data, for indices i and j running over all pairs of points, and $\boldsymbol{y}$'s the transformed data.

In low numbers of dimensions, PCA corresponds to making rotations and reflections of the data points so that the transformed data lie with most of the variation in the data lies parallel to the axes. One assumption made is that the data have zero mean, and it is sometimes useful to 'non-dimensionalise' the data by also dividing by the standard deviation. These transformations are easy to apply, but can have an impact on interpretability - for instance normalising count data in this way could make values harder to compare.

PCoA

PCoA attempts to generalise PCA for situations where other metrics than Euclidean are used in the high-dimensional space. This is useful because as we have already mentioned in Section 3.8.4, Euclidean distance is not always an ideal choice of index for β diversity in community composition data. PCoA is also called classical multidimensional scaling (MDS), for reasons that will become clear in the next section.

The idea behind PCoA is to try to retain the theoretically desirable properties of PCA, using information directly from the matrix of distances between points. We can apply a singular value decomposition directly to this distance matrix $\boldsymbol{D}$, with components D_{ij}. Concretely, PCoA minimises a strain function

$$S = ||B_{ij} - ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2||_F,$$

where the matrix $B_{ij} = D_{ij} \odot D_{ij} - D_{ii} - D_{jj}$, so depends on the distance matrix in a straightforward way. This strain function is minimised by taking the first few columns in the singular value decomposition of B as components for the embedded vectors $\boldsymbol{y}_i$. In the special case where the matrix D_{ij} corresponds to matrix of Euclidean distances, this is the same thing as PCA, but the method will work equally for other metrics, or for dissimilarities.

Other Methods for Dimensionality Reduction

There are a number of other popular methods of dimensionality reduction and ordination. Classical multidimensional scaling (i.e. PCoA) can be further generalised to forms of metric and non-metric MDS [210]. In these formulations, the assumptions about the distance or dissimilarity matrix used are weakened still further.

There are also several other methods which aim to decompose high-dimensional data into linear components, which drop some of the nice theoretical properties of PCA but add other useful constraints.

One example is sparse PCA [211], where PCA is reformulated as an optimisation problem and an ℓ_1 penalty is added - the effect of this is to enforce sparsity in the principal component vectors. This can be useful since without sparsity, component vectors are generally a linear combination of all inputs, which could be an unmanageably large number. Depending on the size of the ℓ_1 penalty added, many of these contributions, except the most important, are set to 0. A different use of the ℓ_1 norm is in ℓ_1-PCA, which is more robust to outliers [212]. Other decompositions, such as independent component analysis (ICA) [213], drop the condition of zero correlation between components in favour of generating components that are independent, under a model of additive mixing of the various sources.

A final decomposition method, applicable to count data, is Non-negative Matrix Factorisation (NMF), where non-negative data such as count or relative abundance data are factorised into a set of components and loadings, retaining the constraint that all the data remain non-negative. This can aid interpretability, especially when negative data (such as count values) have no physical meaning [214].

3.8.6 UMAP and t-SNE

The methods up until now have mostly been linear decomposition algorithms, meaning that for data $\boldsymbol{x}_i$ with components in matrix form X_{ik}, we generate matrices of coefficients α_{ij} and (principal) components PC_{jk} such that

$$X_{ik} = \sum_j \alpha_{ij} PC_{jk}$$

However, data are not always built of linear components in this manner. Two more recent methods in dimensionality reduction algorithms are the use of t-distributed Stochastic Neighbour Embedding (t-SNE) [215] and Uniform Manifold Approximation (UMAP) [216]. While these are not the only non-linear methods available (there are others such as self-organising maps (SOMs), and Isomap [217], [218]), they are the most ubiquitous, especially in machine learning and data science, and have recently gained huge popularity for their ability to handle large datasets. UMAP in particular has become a routinely used in single-cell sequencing analysis, as part of the Scanpy package [219]. Both have good runtime characteristics, and produce visually appealing plots. UMAP is a network-based method, where the high-dimensional data are converted into a weighted k-nearest-neighbours graph, for a

user-defined value of k. In this formulation, each point is connected by an edge to its k closest neighbours, and a weight is assigned to each edge based on their proximity. This graph is then laid out in an embedded space (usually two-dimensional), using an optimisation algorithm to minimise a quantity called cross-entropy. The t-SNE algorithm is conceptually similar but slightly simpler. In t-SNE, points are laid out in the embedded space so as to again minimise a cross-entropy, but this time based not on a weighted adjacency matrix but on probability distributions. These distributions are derived from the probabilities of each point picking one of its neighbours at random, based on a Gaussian distribution (in the high-dimensional space) or a Student's t-distribution (in the embedded space). A fuller treatment of both UMAP and t-SNE are given in Chapter 6. A key point of both algorithms is that they do not attempt to preserve all pairs of distances between points, instead, the aim is to preserve distances locally between points and their neighbours.

There are several limitations to using either of these non-linear methods to visualise data. The first is that these are not ordinations, but are simply embeddings, where the coordinates of the points in the visualisation space no longer have a physical meaning that could in principle relate to units in the original data. The second problem is that although locally, distances are preserved relatively well, the global scale of points tend to be distorted. A third problem is that each algorithm has some key parameters which greatly affect the clustering of the plots, either perplexity for t-SNE, or the number of neighbours, for UMAP. Care must be taken to ensure that what look like meaningful clusters are not in fact just artifacts from this choice of parameters. Finally, these algorithms are stochastic - they depend on random initial conditions, and different initial conditions can produce quite different plots.

Given this extensive list of issues, it is reasonable to wonder if there is any benefit at all to using either of these methods. With the choices for parametrisation, and stochastic nature of the algorithms, it is certainly possible to cherry-pick one's favourite looking plot and over-interpret the tea-leaf style patterns that emerge.

There are however two important advantages that these methods have over ordinations such as PCA, PCoA, and MDS: (1) they perform well on large datasets, and (2) many interesting datasets are non-linear. While MDS can in principle be used as an ordination method on non-linear data, in practice, the embeddings produced tend to look meaninglessly messy for large datasets (over 10^5 points), if the algorithm can even finish at all. PCA and PCoA do not suffer from poor runtime characteristics, but are only applicable to linear datasets; if there are multiple interesting clusters of points that cannot be parametrised by two or three variables, PCA is unlikely to be able to show much finer structure within the data.

3.8.7 Compositional Data Analysis

The previous sections have given an overview of some general techniques from data science and community ecology, but there are certain features of metagenomic data which make analysis more difficult. The most important of these is that metagenomic data is compositional. In a metagenome, the total number of reads is somewhat arbitrary, and is a function of an estimated desired depth of coverage, and a limited budget of time and money for running the sequencing machine. Within a sample, the raw number of bases itself is therefore not very illuminating, instead, only ratios such as the fraction of the total read count, are guaranteed to be meaningful. This can greatly confound an analysis. Spurious correlations can arise in compositional datasets between pairs of otherwise uncorrelated variables, purely because as fractions of the whole sample, they share a common denominator. In a limiting case, any set of samples made from just two species will always appear have a negative correlation (or possibly 0) between them, because with only two species present, an increase in the fraction of one can only result in a corresponding decrease in the other. The problem persists in less extreme cases; for example a large increase in abundance of one species will introduce a corresponding drop in all other species, making them appear correlated, when there may in fact be no relationship between them. There does not yet seem to be a standard method of compositional data analysis that is accepted amongst all microbiome researchers, though there is an increasing awareness that taking into account compositional effects is important [220], [221].

There are several proposed methods to analyse compositional data, the simplest of which are a family of log-ratio transformations which convert compositional data to a new set of coordinates. These transformations include the centred log-ratio (clr), isometric log-ratio (ilr), and additive log-ratio (alr) transformations. For the clr and alr, these transform the original data by application of functions $x \mapsto \log x - \log g(x)$, or $x \mapsto \log x - \log x_D$ respectively. Here, x is an untransformed vector of components, $g(x)$ is the geometric mean of the components of x, x_D is some particular component of x. The ilr is more difficult to describe succinctly; the ith component of the transformed vector is given by $\sqrt{\frac{i}{i+1}} \log \frac{g(x_{(i)})}{x_{i+1}}$, where $x_{(i)}$ represents the vector x truncated to the first i components. While the isometric log-ratio has some favourable properties (such as preserving distances and angles), the resulting components are not easy to interpret. The alr is simple to interpret, but requires a choice of special component against which all others are scaled against. The clr was first used to analyse geophysical data [222], though has become popular in metagenomics [223]–[225]. All three methods require from having to add a pseudocount value (a small positive offset) if there are zeros in the untransformed data, since $\log 0$ is undefined.

Other methods for analysing compositional data involve applying corrections to correlations between OTUs, or using a different measure rather than Pearson correlation. The python package SparCC [226] attempts to estimate the true (i.e. without compositional bias) correlations between OTUs, assuming that on average, most correlations should be small.

A complementary approach to dealing with the compositional nature of metagenomic data is the use of quantitative methods that generate absolute rather than relative counts. Quantitative PCR (qPCR) targeting specific marker genes (e.g. the 16S rRNA gene for total bacterial load, or functional genes for particular processes) can provide absolute cell or gene copy counts per unit volume of sample. By combining qPCR-derived absolute abundances with metagenomic relative abundances, it becomes possible to approximate absolute concentrations for particular taxa or functions, thereby overcoming some of the compositional artefacts described above. However, this requires careful primer design for broadly conserved targets, and introduces its own biases from PCR amplification efficiency.

3.8.8 Correlation Networks and Clustering Algorithms

Associations between genes, or species, measured as correlations, or through other metrics such as proportionality, have a natural interpretation in network theory. In this formulation, genes (for example, though species or other units of interest could be valid) are represented as vertices in a network, with two vertices linked by an edge with a weight representing the strength of the association between the two vertices. If associations are measured through the Pearson correlation coefficient r, then the weighted adjacency matrix of the network is simply the matrix of correlation coefficients.

There are often more genes (or OTUs) than there are samples, and far more pairs. This leads to several problems. The first is that there are too many genes to be able to analyse them all, and some kind of summary information is required. A second related problem is to do with multiple testing corrections. In essence, there are so many genes or pairs of genes that some correlations are likely to appear high by chance. Applying a multiple testing correction in this situation is possible, but can reduce the power of the test. This is before one even considers problems stemming from compositional biases discussed in the previous section.

We will look at Weighted Gene Correlation Network Analysis (WGCNA), a foundational method in network analysis, as well as some more recent developments since.

Weighted Gene Correlation Network Analysis

Weighted Gene Correlation Network Analysis (WGCNA), first introduced in Langfelder *et al.* [227], is a method to find subsets of genes (modules) within RNA-expression data, correlated with one another and with exogenous variables. WGCNA has been extremely influential, not just in RNA-seq studies but also metagenomics more widely, and has been applied to microarray, 16S amplicon and proteomic datasets [228]–[230]. WGCNA assumes a scale free network topology, popularised by Barabási *et al.* [231] - the gene network is assumed to have a power law degree distribution. The elements of a correlation matrix (some other similarity matrix can be used) are raised to a power to form the adjacency matrix of the weighted network. Modules are then formed based on a hierarchical clustering of the nodes, and eigengenes (linear combinations of a set of genes, proportional to the first principal component of that set) are built from their respective modules. Since there are fewer modules, these can be tested for statistical significance against exogenous traits, with fewer issues due to multiple testing corrections. Modules can also be tested for gene enrichment, for example based on Gene Ontology (GO) terms, or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

Modern Methods in Biological and Ecological Network Analysis

Since WGCNA, many gene network analysis tools have been published, either building on top of WGCNA (such as Lemoine *et al.* [232]), or dedicated to a particular part of network analysis such as finding modules of co-expression [233], [234]. In Jiang *et al.* [235] and Kong *et al.* [236], WGCNA was used in comparison to other machine learning classification methods to find groups of genes in biomedical datasets - rheumatoid arthritis and ischaemic heart disease respectively. The fact that in all these studies, WGCNA was used as the benchmark against which other methods were compared, shows how influential it has been for analysing gene expression and gene correlation datasets.

While WGCNA has also been used in species co-occurrence network analysis (treating the species abundances as units, instead of genes), some assumptions made by WGCNA may not be applicable, particularly the assumption of a scale free network topology. The package Sparse Co-occurrence Network Investigation for Compositional Data (SCNIC) uses the sparse correlation matrix built by the SparCC to generate a co-occurrence network, with modules of species generated using a threshold of pairwise correlations within the module [237]. These species networks start to overlap with joint species distribution models (JSDMs) more generally, where a lot of theory has been developed, but the complexity also increases. Methods such as Mefisto [238], SpiecEasi [239] and HMSC [240], which model species or

gene networks, but also include more complex models of the covariates, potentially including modelling the temporal or geographical structure of the data.

3.9 Discussion

This chapter has contained a survey of the methods used in bioinformatics for sequence assembly, annotation, and binning, and the methods used for analysing microbial diversity, species correlations, and gene and species networks. Of the dimensionality reduction methods described here, PCA was used for exploratory analysis of Pfam compositions in Chapters 4; t-SNE was used to visualise functional similarity of prokaryotic MAGs, and Pfams (Chapters 4 and 7); and UMAP was central to the eukaryotic MAG refinement pipeline described in Chapter 6. The choice of each was determined by the complexity (number of dimensions) of the data being analysed. PCA was appropriate to visualise difference between a small number of samples, when the number of features to visualise was large, either t-SNE or UMAP were more appropriate. The choice of UMAP over t-SNE for the contig visualisation pipeline (Chapter 6) was motivated primarily by UMAP's superior runtime on large datasets and its use in the VAMB latent space, where it integrates naturally with the variational autoencoder's output.

Chapter 4

Generating MAGs

The full set of the MOSAiC metagenomic and metatranscriptomic data comprises a year-long time series of over 1000 samples, from diverse habitats including sea-ice ridges, under-ice sediment traps, and all ocean layers, from the epipelagic to bathypelagic. Sequencing and processing this collection of metagenomes was a relatively large effort; the IMG/M pipeline is still processing some samples at the time of writing. In this chapter, we describe the bioinformatics pipelines that we initially used to generate prokaryotic and eukaryotic MAGs from the first 73 samples we received, both from a set of pilot samples, and as part of a MOSAiC sub-project, called Ridges - HAVens for ice-associated flora and fauna in a seasonally ice-covered Arctic OCean (HAVOC). We present a catalogue of MAGs recovered from these samples, including 2407 prokaryotic and 56 eukaryotic MAGs, as well as annotations of a near complete eukaryotic MAG, using an annotation pipeline developed by the JGI for the online resource Phycocosm.

We also present results based on analysis of MAGs from each of the two sets of samples; these were published in Boulton *et al.* [11] and Mock *et al.* [8]. Finally, we discuss the performance of the various assembly and binning strategies that we utilised. The final set of MAGs we generated can be used to benchmark microbial biodiversity in the Central Arctic Ocean, compare individual strains across space and time, and to study changes in Arctic microbial communities from the winter to summer, at a genomic level.

4.1 Background and Summary

The metagenomes we study in this chapter consist of two sets of samples from sub-projects which were received early on in our investigation; a set of 15 samples sequenced as part of a pilot sequencing project and collected during the Arctic winter (pilot samples), and a set of 58 samples from the HAVOC project, associated with sea-ice ridges (HAVOC samples). The pilot samples were collected from the first-year and multi-year MOSAiC ice-coring sites, and the CTD, during the Arctic winter. Seawater samples cover depths ranging from the epipelagic to bathypelagic, whilst the sea ice samples are from both first-year and multi-year ice, from two coring sites on the ice flow. The HAVOC samples were taken as part of

a satellite project, collecting samples from ice-ridges, under-ice water, and from sediment traps beneath sea ice ridges and level ice [241], throughout the course of the drift.

Sea ice ridges are characteristic features covering 25 to 45% of the Arctic sea ice area [242]. Ridges are formed by pressure from drifting ice. When ice floes are forced together, they break up and are pushed up and down to form a sail above and a keel below the surface water level. The keel consists of ice blocks separated by voids, described as macroporosity, making up approximately 15 to 30% of the volume [243], [244]. The voids may be empty (in the sail) or filled with liquid or frozen seawater or meltwater (in the keel). While the bottom of level sea ice is known as an important habitat for Arctic marine biodiversity and activity, much less is known of the life within ridge keel voids which constitute unique habitats and biological hotspots in the Arctic Ocean [245]–[247]. As ridges are logistically harder to navigate and take samples from, they are relatively understudied compared to level ice. The aim of the HAVOC project was to better understand how ice ridges act as a refuge for microbial biodiversity and activity, and how food web and biogeochemical processes at the ice-ocean interface and the underlying water column differ between ridges and level sea ice [241].

Advances in metagenomic sequencing, assembly and binning, have generated a wealth of MAG datasets, even for Arctic environments such as in Royo-Llonch *et al.* [105]. However, challenges associated with the assembly and binning of eukaryotes have meant that these generally focus on prokaryotic MAGs. Two exceptions are Duncan *et al.* [104] and Delmont *et al.* [248], which generated 21 and 25 eukaryotic medium and high quality MAGs from Arctic metagenomes respectively (i.e. above 50% completeness). The data presented here includes both prokaryotic and eukaryotic MAGs, using coassembly to improve the coverage of eukaryotes. These MAGs can be used to gain insights into microbial diversity and the metabolic potential of microbiomes during the Arctic winter, and the role of ice ridges in maintaining the microbial biodiversity of the Arctic Ocean.

4.2 Methods

4.2.1 Sampling

Of the 73 samples presented here, 15 were collected as part of a larger time-series, during leg 2 of the expedition (between 13 January 2020 and 7 February 2020) as described in Winder *et al.* [12] and in Chapter 5, and sequenced as part of a pilot sequencing project (hereon called pilot samples). Of these samples, 8 were from pelagic layers and the remaining 7 from sea ice. The pelagic pilot samples were collected using a CTD rosette, on three different days.

Sequenced pilot samples from a sampling event on February 6th 2020 consist of 2 co-located samples (i.e. replicates, from the same CTD cast and sampled at the same depth, but from different Niskin bottles) taken from a depth of 20 m, and one sample from a depth of 202 m. One sample was collected on February 7th 2020, at a depth of 4082 m. Further, 2 replicates from a depth of 50 m sampled on January 16th 2020 were sequenced. Additionally, for each of the two biological replicates, a third technical replicate was generated by pooling remaining material from the two replicate samples. These data are summarised in Supplementary Table 1 of [249], with the two samples generated through pooling identified with a sample identifier suffix 'pool' (column E).

The remaining 7 pilot samples from sea ice were taken from the first-year and second-year MOSAiC ice-coring sites on the floe, as described in Nicolaus *et al.* [116] and Fong *et al.* [99]. First- and second-year sea ice chemical and physical properties are available in Oggier *et al.* [250], [251], Lei *et al.* [252], and properties for snow in Macfarlane *et al.* [253]. Generally, 2-4 cores were collected on a weekly basis, cut in 10 cm sections, except the bottom where two 5 cm thick section were cut, and pooled per section to allow for enough biomass in the DNA samples. For the pilot samples, each ice metagenome always represents a pool of three individual cores collected after each other in the same location (adjacent within 40 cm). Of these, five were collected from different 5-10 cm thick sections of cores from the same coring site on February 3rd 2020 from first-year ice; three from the upper part (20-50 cm from the top), one from the middle section (70-80 cm from the top), and one from the bottom-most section (122 to 127 cm from the top) of the sea ice, i.e. at the sea-ice interface. The remaining two samples were second-year ice, also from the bottom most 5 cm section, i.e. from the sea–ice interface, 1.23 to 1.28 m and 1.43 to 1.48 m from the top, collected on January 13th and 27th 2020.

The 58 metagenomes from the HAVOC project were collected during legs 2, 3 and 4 (collection dates between 22nd January 2020 and 26th July 2020), either: from sediment traps, directly below an ice ridge (7 samples) or level ice (10 samples) at depths of 5, 15 and 50 m, in the water column at 20 m (2 samples), from under ice water below an ice ridge (2 samples) and level ice (3 samples), from seawater taken from voids in the ice ridge (10 samples), or from a 10 cm ice core section at different depths of the ice ridge, as either ridge bottom ice (3 samples), top of void ice (5 samples), bottom of void ice (6 samples), refrozen void ice (4 samples) and ice samples at irregular depths (6 samples). Samples from the same location, depth, and time (Supplementary Table 1 in Boulton [249]) are considered replicates, with samples taken from a total of 24 distinct locations and depths. The number of pooled core sections for each sample, and section thickness, are recorded in Supplementary Table 1 [249]. Figure 4.1 summarises the locations of the samples.

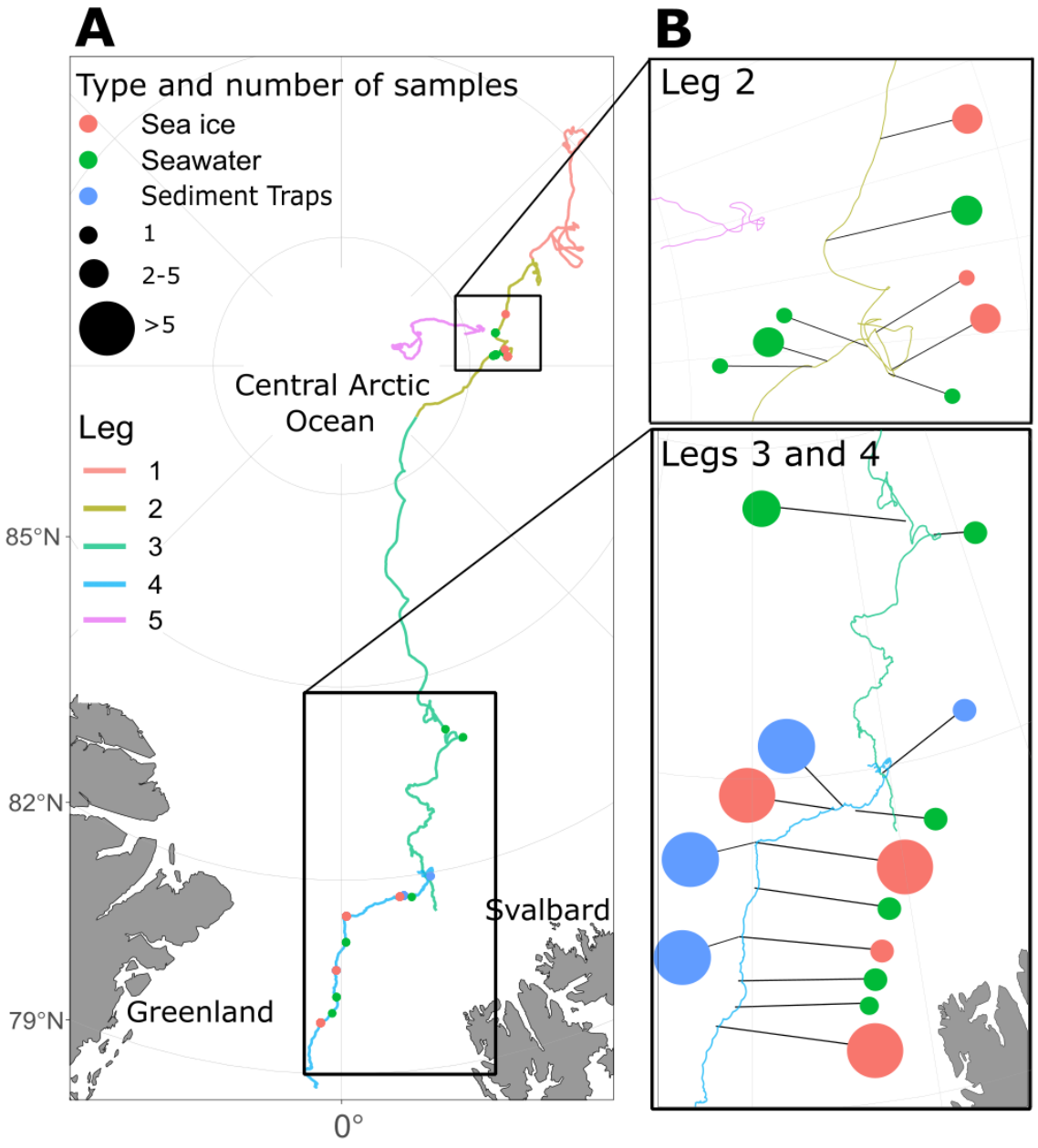


Figure 4.1: Map showing the locations of the samples. Panel A shows the overall course of the drift, and locations of the samples. Panel B shows more detailed locations, zoomed within the boxes marked in panel A. Samples are from leg 2 (15th Dec. 2019 – 3rd March 2020), leg 3 (3rd March – 6th June 2020), and leg 4 (6th June – 12th August 2020), with the drift route generally moving southward from the Central Arctic Ocean. Often, multiple (replicate) samples are co-located, either from the same CTD cast, or as different layers within a single ice core. In panel B the number of co-located samples is represented by the size of the marker.

For both the HAVOC and pilot sea ice samples, ice cores were sectioned in the field, transferred to sterile plastic bags, and brought back to the Polarstern. On board, 50 ml 0.22 μm filtered sea water was added per cm sea ice, and the sea ice samples melted within 24 to 36 hours in the dark at around 17 to 22 °C. The use of filtered seawater was made necessary due to logistical constraints during the drift; this is a possible source of contamination for the sea ice samples. For both sea ice and pelagic samples, water was filtered through a Sterivex 0.22 μm filter or when volumes were $<$ 500 mL (HAVOC sediment trap samples and three HAVOC ice samples) onto 0.22 μm Durapore filters. The filters were immediately flash frozen in liquid nitrogen and stored at -80 °C on board the Polarstern, and subsequently shipped to either the Alfred Wegener Institute (pilot samples), or the University of Bergen (HAVOC samples), at a temperature of -80 °C.

4.2.2 DNA Extraction, Purification, and Sequencing

Following shipping, DNA from Sterivex filters was extracted using the Qiagen PowerWater DNA kit, following the QIAGEN DNeasy Power Water SOP version 1 for the ice and water samples, and the QIAGEN Dneasy Power Soil SOP version 1 (QIAGEN N.V., Hilden, Germany) used for the DNA extraction from Durapore filters. This work was carried out by collaborators at AWI, particularly Sarah Lena Eggers. Plates were shipped to the Joint Genome Institute (CA, USA) under dry ice, and sequenced using either the Illumina low or regular concentration protocol, with between 0 and 15 rounds of PCR applied to samples. The sequencing project was managed by Kerrie Barry at the JGI.

For the regular protocol, the DNA was sheared to 300 bp using the Covaris LE220-Plus and size selected with SPRI using TotalPure NGS beads (Omega Bio-tek, Norcross, GA, USA). The fragments were treated with end-repair, A-tailing and the ligation of Illumina compatible adapters (IDT, Inc, Gladesville, Australia) using the KAPA-HyperPrep kit (KAPA Biosystems, Wilmington, MA, USA). The prepared libraries were quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The sequencing of the flowcell was performed with the Illumina NovaSeq sequencer using NovaSeq XP V1.5 reagent kits, S4 flowcell, following a 2 x 151 indexed run recipe. For the low input protocol (less than 10 ng of DNA, and less than 1 ng per μL concentration), the procedure was the same, except that the sample was first enriched using between 5 and 15 cycles of PCR. In all but 5 cases, the number of PCR cycles was restricted to 5 rounds. Supplementary Table 3 in Boulton [249] outlines the library preparation steps and sequencing protocols used for each of the samples.

4.2.3 Genome Assembly and Binning

Samples were first assembled individually using the JGI MAP pipeline [254], with prokaryotic bins recovered on a per-sample basis. In brief, samples were filtered for quality with BBDuk, error corrected using BBCMS, assembled using SPAdes [147], and reads mapped back to contigs using BBMap (version 38.86) [146]. Binning was performed using MetaBAT2 (version 2.15.1) [188] and assessed for quality using CheckM (version 1.1.3) [255]. Software and pipeline versions used are listed in Supplementary Table 3 [249]. To extract further eukaryotic bins, we used a coassembly method. We used a custom filtering pipeline to identify the eukaryotic fraction of reads from each sample, before pooling these reads for coassembly and binning. To extract the eukaryotic fraction of reads, we used MMSeqs2 (version 01889*) [185] with the default parameters (length cut-off of 500 base pairs) to taxonomically identify contigs that had already been assembled using a per-sample assembly method, using a combination of MMETSP [256] and NCBI NR [257] as a reference database. Contigs identified at the domain level as anything other than Bacteria, Archaea, or viruses were retained, leaving a list of putatively eukaryotic contigs. Contigs identified as belonging to already existing prokaryotic bins were removed from this list. Next, the quality-filtered reads were mapped to this subset of contigs with BBMap (version 3.17). These reads were pooled, depending on whether they were from the HAVOC or pilot dataset, and assembled with MetaHipMer (version 2.1.0) [258] for the HAVOC samples, or SPAdes (version 3.14.0) with the metaspades.py --only-assembler option for the pilot samples. The choice of assembler differed between the two datasets for logistical reasons: the HAVOC coassembly was performed on the NERSC supercomputing infrastructure using 500 CPU nodes (256 GB RAM each), where MetaHipMer was available and optimised for large-scale coassembly. The pilot coassembly was performed on a high-memory node at the UEA (1 TB RAM), for which MetaHipMer was not available; SPAdes was used instead, albeit with a substantially longer runtime. Samples with no pre-existing metagenome bins from their single assembly were excluded. Finally, the pooled reads from each dataset were mapped to their respective coassemblies, and then the new contigs were binned using MetaBAT2 (version 2.15), and checked for quality with EukCC (version 2.1.1, database version 1.1) [259]. Bins of over 90% completeness and less than 5% contamination, and with at least 18 tRNA genes, and with 23S, 16S and 5S rRNA genes, were designated as high-quality MAGs, those with above 50% completeness and less than 10% contamination were designated medium-quality, and, for the eukaryotic MAGs, those above 30% completeness and less than 10% contamination were retained and designated as low quality, as per Alexander *et al.* [260]. Figure 4.2 shows a schematic diagram of the bioinformatics pipeline, and 4.3 provides an overview of the completeness and contamination of

the generated MAGs.

4.2.4 Functional and Taxonomic Annotation

MAGs recovered from single-sample assemblies were annotated using the IMG/M annotation pipeline (versions ranging between 5.0.23 and 5.1.11), using Genemark (version 1.05) and Prodigal (version 2.6.3) [169], [170] for gene calling, and HMMer (version 3.1b2) [261] to combine analyses from the COG, Pfam (version 0.34) [168], TIGRFAM (version 15.0) [172], Cath-Funfam (version 4.1.0) [262], SuperFamily (version 1.75) [175], and SMART (access date 01_06_2016) [174] databases. CRISPRs, and tRNAs were identified with CRT (version 1.8.2) [179] and tRNAscan-SE (version 2.0.4) [263] respectively. Prokaryotic MAGs were taxonomically placed with GTDB-Tk (version 2.4.0, database release 220) [197]. GO terms were included based on the Pfam2GO mapping provided by Interpro [264], [265]. To identify genes within the coassembled eukaryotic MAGs within the coassemblies, we used MetaEuk (version f32e8*) [266] with the --easy-annotate option, using a custom database of the combined Phycocosm [10], MMETSP, and UniRef [267] databases, with UniRef clustered at a 50% identity level. These were combined with genes identified through Genemark-ES (version 4.71, gmes_pepal.pl –ES), with genes from MetaEuk given priority and retained if overlapping with genes from Genemark-ES. Pfam (version 35.0), PANTHER (version 17.0), SMART (version 9.0), NCBIfam (version 12.0) and SuperFamily (version 1.75) domains were then annotated using InterProScan (version 5.63) [268]. Eukaryotic MAGs were placed on a phylogenetic tree (Figure 4), using a set of 100 concatenated BUSCO genes (BUSCO version 5.1.1 odb_eukaryota_10 gene set, aligned using MUSCLE version 3.8.1551) [155], [269], alongside a set of 140 eukaryotic reference genomes from Phycocosm and NCBI RefSeq [270]. A maximum-likelihood tree was generated using FastTree (version 2.1.11) [180].

4.3 Results from the Pilot Samples

We analysed the pilot samples in terms of the quality of their assemblies, as well as their taxonomic profiles. We compared the taxonomic profiles of the prokaryotic fraction of the samples with the fraction that could be mapped to MAGs. We also compared the functional profiles of the prokaryotic MAGs from the pilot samples. We found highly distinct prokaryotic communities between each of the four sample types, and especially between ice and water. This was also true when we looked at functional composition, at the level of Pfams. When we looked at functional profiles of MAGs, we found that functional similarity was largely dependent on prokaryotic phylum. We observed generalist and specialist clades of

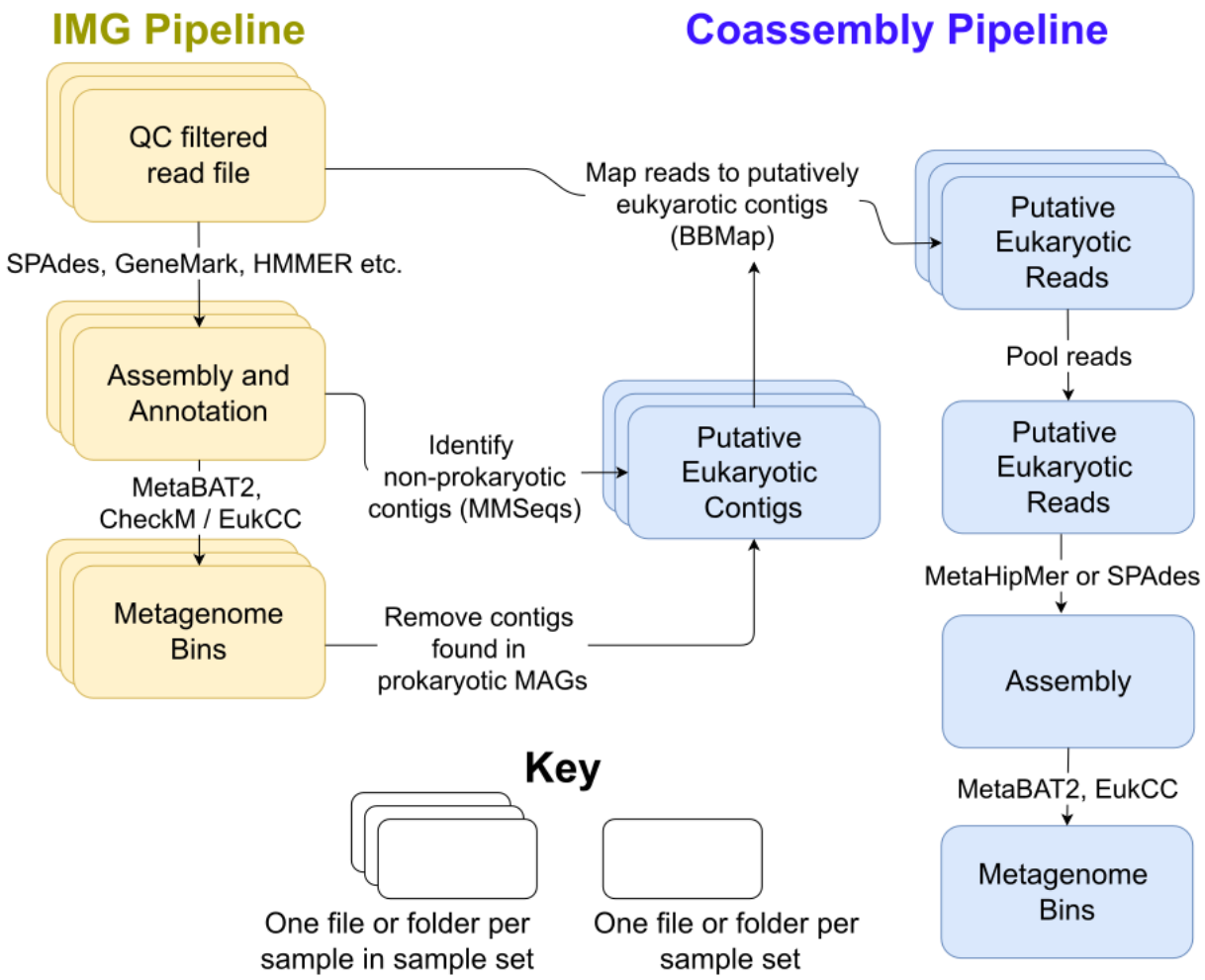


Figure 4.2: A summary of the IMG metagenome annotation pipeline, and the coassembly pipeline used for the two sample sets; either the pilot samples or the HAVOC samples. Coloured boxes indicate intermediate folders or files, either one per sample in the case of the stacked boxes, or one for each sample set, in the case of the coassemblies. Arrows indicate which files are inputs and outputs for other processes.

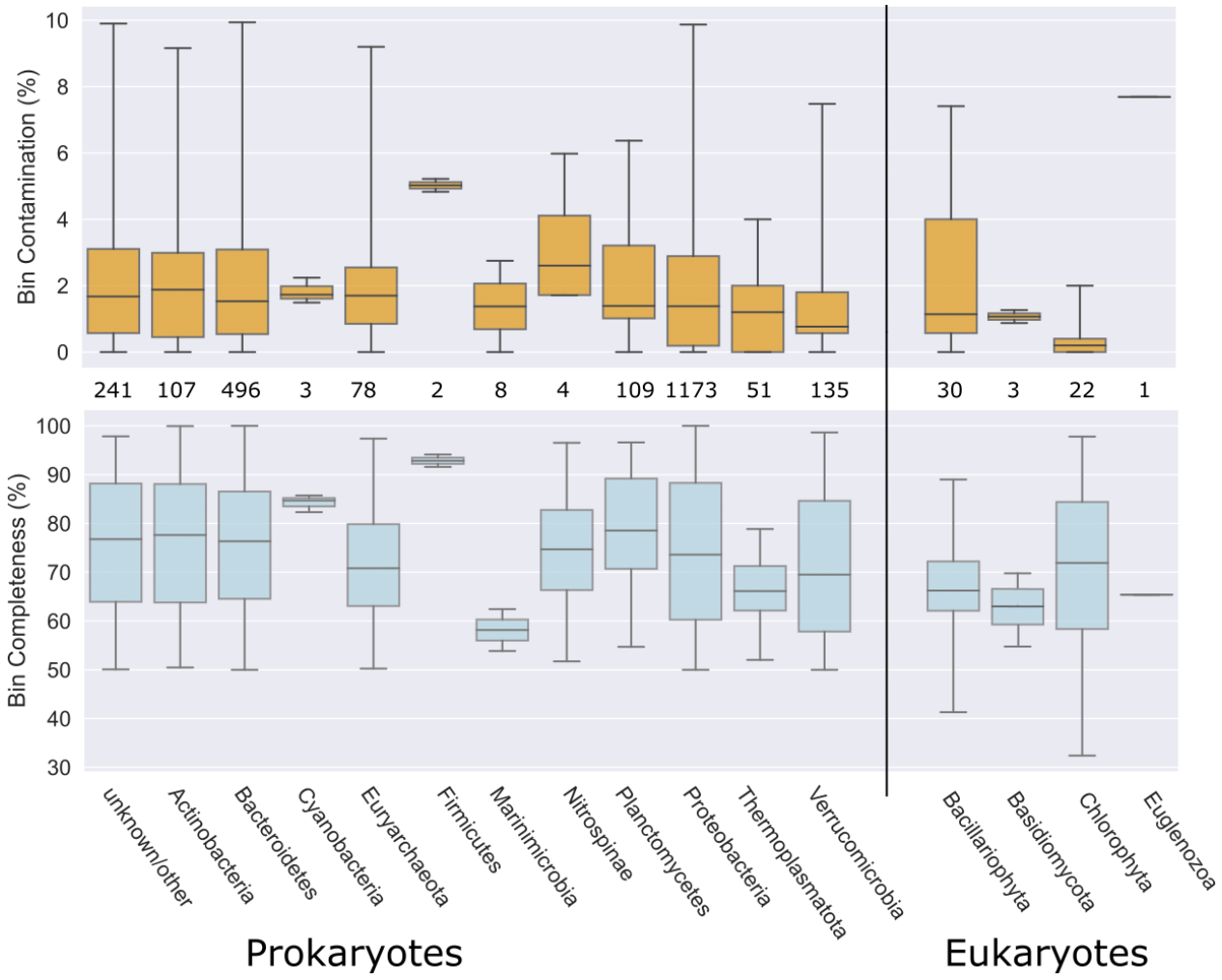


Figure 4.3: Completeness and contamination of the 2463 MAGs recovered across the 73 samples; 2407 prokaryotic and 56 eukaryotic MAGs. In each panel, a vertical line separates the eukaryotic and prokaryotic MAGs. The number of MAGs per taxon is shown between the two boxplots.

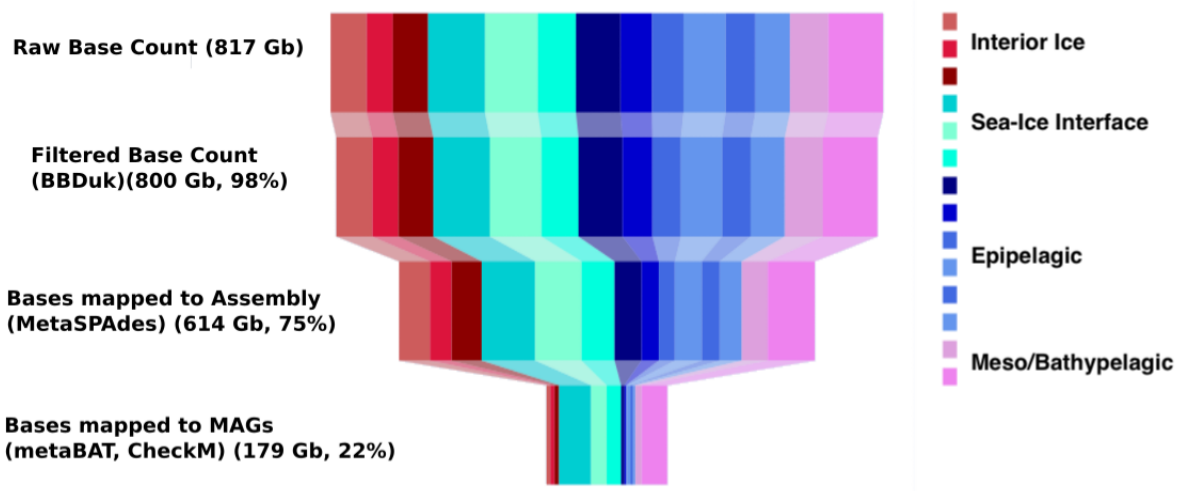


Figure 4.4: Funnel diagram showing the proportion of bases retained from quality filtering, assembly mapping, and binning. Note the diagram only contains 14 samples - the IMG/M binning pipeline failed for one sample (interior ice 2).

MAGs, including some clades of MAGs which were all functionally similar but were identified in all environments, whereas others could only be recovered from a particular ecotype.

4.3.1 Assembly and Mapping Statistics

There were a total of 817 Gbp sequenced from the 15 pilot samples, with a mean of 59.1 Gbp per sample (s.d. 13.9). This resulted in 2.53 M contigs per sample (s.d. 0.76 M) based on the single assemblies, of which the contig N_{50} from seawater was in general shorter than the contig N_{50} from sea ice (875 bases, s.d. 105 in seawater, compared to 1130 bases, s.d. 336 in ice). The prokaryotic fraction of these assemblies were 91% in total, ranging between 75% and 98% of the samples. The remainder was almost all eukaryotic, with just 2.5% viral (between 1.0% and 4.8%). Eukaryotes were more abundant in the sea ice than in water, with a mean abundance of 12.5% in sea ice (s.d. 7.6%) and 2.3% in seawater (s.d. 0.7%). This difference was significant at the $p < 0.05$ level (Welch's t-test).

The MAP pipeline generated 702 MAGs of medium quality and 48 high quality MAGs, defined as above 90% completeness, less than 5% contamination, as well as the presence of at least 18 tRNA genes and 3 rRNA genes. MAG binning initially failed for one out of 15 of the samples (interior ice 2); therefore in some of the later analyses we excluded this sample. The IMG/M pipeline did not generate any eukaryotic MAGs from the pilot samples; we later generated eukaryotic MAGs through coassembly as described in Section 4.2.3. Overall, 22% of bases from reads were mapped back to MAGs (179 out of 817 Gb, Figure 4.4). The ma-

jority of sequenced data — including reads from rarer species, poorly assembled organisms, and taxonomic groups resistant to binning — remains unaccounted for in a MAG-centred analysis. An alternative approach that sidesteps binning entirely is to perform functional analyses at the contig or gene level, capturing a broader fraction of the community. Contig-based taxonomic annotations are less accurate than MAG-based taxonomies, however. A comparison of MAG-based and contig-based Pfam profiles is described in the following section. The reads mapped back to MAGs were distributed unevenly; a smaller proportion of bases from the interior ice and epipelagic environments were mapped to MAGs (12% and 6% respectively) compared to the sea-ice interface and meso/bathypelagic samples (41% and 34%). From here on we concentrate only on the prokaryotic fraction of the assemblies, and the 750 prokaryotic MAGs recovered from the pilot samples.

Sample Label	Raw Base Count (Gbp)	Raw Read Count (10^6)	Filtered Base Count (Gbp)	Filtered Read Count (10^6)	Number of Contigs (10^6)	Contig N_{50} (Bases)	Reads Aligned to Assembly (10^6)	Reads Aligned to Assembly (%)
sea ice interface 1	84.5	560.1	83.4	556.6	1.26	941	526.9	94.7
epipelagic 1	66.3	439.6	64.7	433.3	3.53	961	265.2	61.2
epipelagic 2	46.6	308.9	43.9	295.0	2.72	883	170.4	57.8
epipelagic 3	46.9	310.7	41.7	280.5	2.46	915	159.2	56.8
sea ice interface 2	78.6	520.7	76.2	509.1	2.26	779	457.5	89.9
interior ice 1	53.7	356.0	53.0	354.7	2.15	1227	304.6	85.9
interior ice 2	64.5	427.2	63.6	425.0	2.41	1605	372.4	87.6
interior ice 3	38.9	258.2	38.3	256.4	1.53	1442	213.7	83.4
interior ice 4	52.3	346.6	51.5	344.4	2.04	1208	295.4	85.8
sea ice interface 3	56.7	376.0	54.5	364.3	1.66	707	325.0	89.2
epipelagic 4	62.8	416.2	60.5	406.4	3.48	784	26.3	64.9
epipelagic 5	44.0	291.5	41.9	281.7	2.46	853	178.6	63.4
epipelagic 6	51.4	340.8	50.0	335.4	3.05	773	215.0	64.1
meso/bathypelagic 1	56.9	377.1	55.8	374.2	3.82	766	255.5	68.3
meso/bathypelagic 2	81.8	542.3	80.6	539.0	3.09	1067	464.4	86.2

Table 4.1: Pilot sample assembly statistics.

4.3.2 Taxonomic and Functional Profiles of Prokaryotes

Across all 15 samples, the most common prokaryotic phyla (Figure 4.5a) were the Gammaproteobacteria (42% relative abundance), Alphaproteobacteria (20%), and Bacteroidota (7.7%), followed by the Verrucomicrobiota (5.3%), Actinomycetota (3.6%), Delta/Epsilon Proteobacteria (3.6%), Planctomycetota (3.4%), Chloroflexota (2.8%), Thaumarchaeota (1.8%), and 86 other prokaryotic phyla (including unclassified, some subphyla, candidate phyla, and miscellaneous clades), each of less than 1% abundance, and combined abundance of 8.2%.

The five most common orders were the Alteromonadales (10.5%, Gammaproteobacteria), Cellvibrionales (10.1%, Gammaproteobacteria), Rhodobacterales (9.8%, Alphaproteobacteria), Oceanospirillales (5.6%, Gammaproteobacteria) and Flavobacteriales (4.0%, Bacteroidota), followed by other bacterial phyla listed above, though they could not be identified at the order level. In the ice, the 5 orders previously named were the most abundant orders, and together made up 65% of prokaryotic abundance, whereas in water, the Gamma-, Delta-, Epsilon- and Alphaproteobacteria, Verrucomicrobiota, and Chloroflexota (orders all unidentified) were the most abundant orders, and in total made up 32% of the relative abundance in seawater.

Within just the proportion of contigs that were binned into MAGs (Figure 4.5b), the most abundant phyla were the Gamma- and Alphaproteobacteria (38% and 18% abundance of the fraction of reads mapped to MAGs), followed by Actinomycetota (10%), Bacteroidota (9%), Verrucomicrobiota (8%), and Planctomycetota (4%). The most abundant orders were the Pseudomonadales (26%, Gammaproteobacteria), Rhodobacterales (10%, Alphaproteobacteria), Acidimicrobiales (8.7%, Actinomycetota), Flavobacteriales (7.7%, Bacteroidota), and Enterobacterales (4.4%, Gammaproteobacteria).

Both the contig-based (Figure 4.5a) and MAG-based (Figure 4.5b) taxonomic profiles can introduce sources of bias. The MAG-based approach provides a genome-resolved view but is biased towards abundant, well-assembled species that meet quality criteria (>50% completeness, <10% contamination), under-representing rare taxa. The contig-based approach captures a larger proportion of the metagenome, including contigs from rare or fragmented genomes that may not meet quality thresholds for binning. However, taxonomic assignments at the contig level carry greater uncertainty, and the use of a taxonomic database (in this case, NCBI nr) introduces a searchlight bias, where some of our results are reflective of the makeup of the taxonomic database rather than the samples themselves.

PCA was applied to the matrix of log-scaled Pfam abundances in each of the samples, (Figure 4.6) with abundance measured in RPM, with a pseudocount of 10^{-6} to avoid taking the log of zero. There was a separation between ice and water samples based on the ordina-

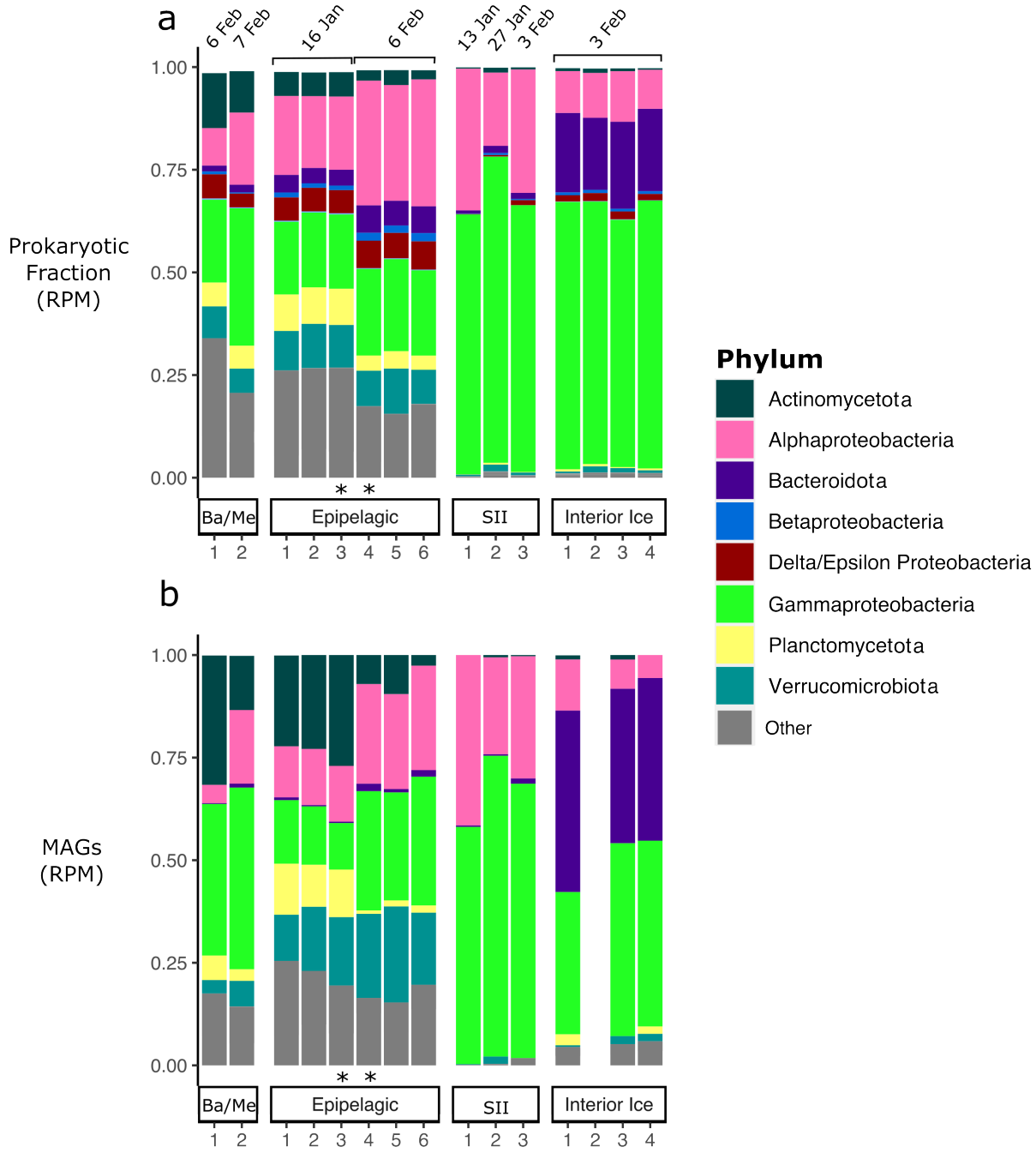


Figure 4.5: Relative abundances of the prokaryotic phyla present, either **a** determined using the RPM abundances of contigs within the prokaryotic fraction of the assembly, or **b** based only on the RPM abundances mapped to MAGs. Ba/Me; bathy-/mesopelagic samples, SII; sea-ice interface. The stars indicate samples generated through pooling. Interior ice 2 is the sample where MAG binning initially failed.

tion of Pfams, with the first principal component clearly splitting these two sample types. Furthermore, the sample subtypes (interior ice, sea-ice interface, epipelagic, meso/bathypelagic) also clustered together closely, with an average Euclidean distance of 20 within each group in the PCA plot, compared to 199 between groups. A permANOVA test statistically verified that there were significant differences between these four groups in terms of their Pfam composition (pseudo-F = 179.75, $p = 0.001$).

To investigate functional differences between ice and water metagenomes, we analysed overall and differential abundances of Pfams within the prokaryotic fraction of the assembled contigs in the pilot samples. We shortlisted Pfams with a total abundance (RPM) over 10, this included 4452 out of 8012 Pfams. The five Pfams most abundant overall were: ABC transporter, response regulator receiver domain, histidine kinase-/DNA gyrase B-/HSP90-like ATPase, binding-protein-dependent transport system inner membrane component, enoyl-(acyl carrier protein) reductase (pfam00005, pfam00072, pfam02518, pfam00528, pfam13561). These are Pfams known to be involved in basic cellular processes such as cellular transport and signal transduction, and were present in over 98% of the MAGs. The five Pfams that were the most comparatively more abundant in ice compared to water based on two-fold change (Figure 4.7), were: DUF4842, propionate catabolism activator, DUF3622, DUF3494 ice-binding, YggL 50S ribosome-binding protein (pfam16130, pfam06506, pfam12286, pfam11999, pfam04320). The five Pfams that were the most comparatively more abundant in water than in ice were: archaebacterial flagellin, DUF63, signal-peptide peptidase/presenilin aspartyl protease, D-aminopeptidase, HTH DNA binding domain (pfam01917, pfam01889, pfam06550, pfam04951, pfam04967). In several cases these Pfams are directly linked to taxonomy; archaebacterial flagellin, DUF63, signal-peptide peptidase/presenilin aspartyl protease and HTH DNA binding domain are all linked to Archaea. DUF4842 is present in a larger number of bacterial phyla, but particularly prevalent within Bacteroidota and Gammaproteobacteria, and the YggL ribosome-binding protein is predominantly found within Gammaproteobacteria.

We generated t-SNE plots (Figure 4.8), using functional similarity in terms of counts of Pfams, to ordinate MAGs. MAGs of the same phylum clustered together, indicating that taxonomic similarity is a strong indicator of functional similarity, at least at the level of phylum. In contrast, several groups of MAGs clustered closely together but were recovered from distinct environments. several other groups seem to be pelagic or sympagic specialists, these are highlighted in Figure 4.8 panel b. Archaea were functionally separate compared to other prokaryotes, and were also recovered only from pelagic samples. In contrast, clades of Alpha- and Gammaproteobacteria (circled in panel b) were functionally closer to one another, but were recovered from all sampled environments, including the bathypelagic.

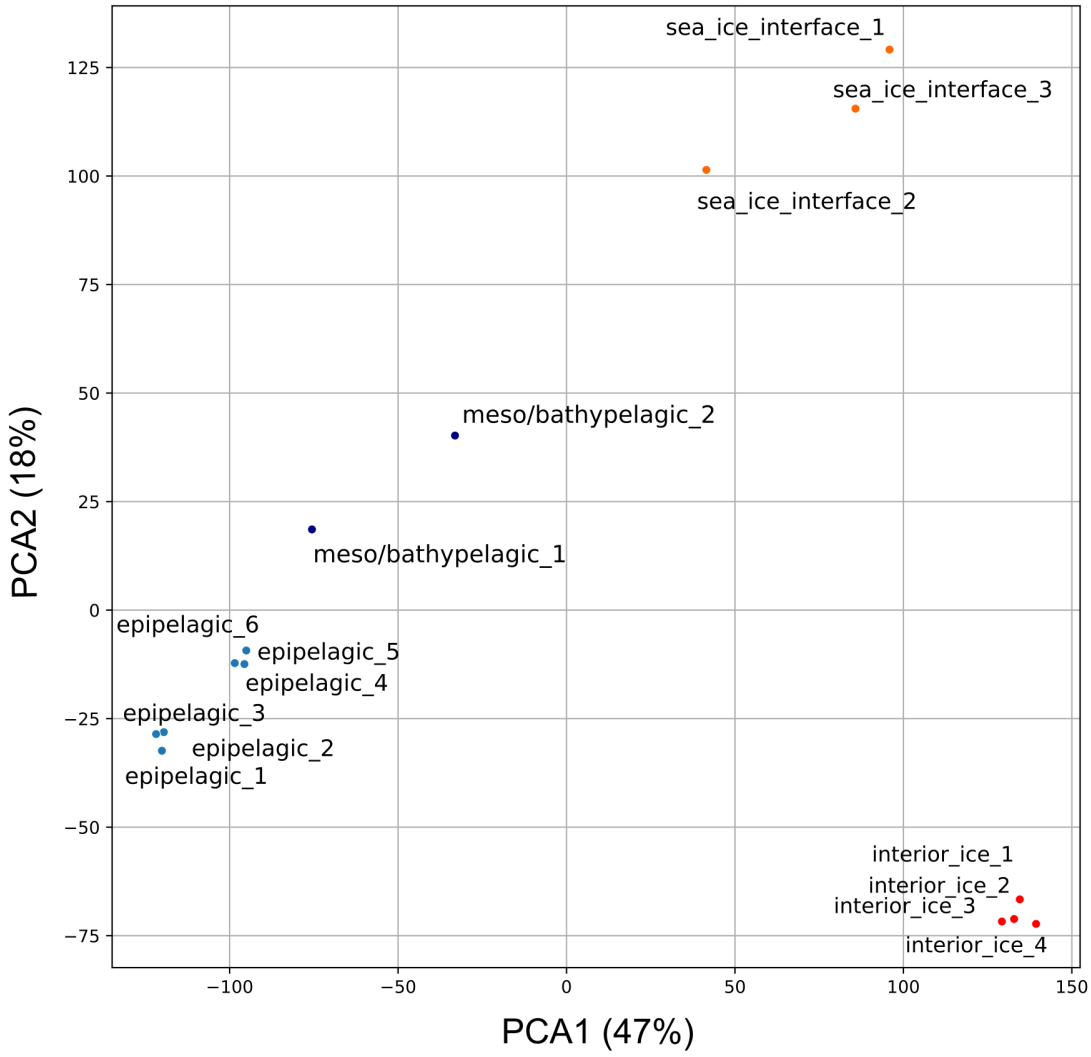


Figure 4.6: PCA plot of Pfam abundances ($\log_{10}$-RPM with a pseudocount of 10^{-6}) within the prokaryotic fraction of the pilot samples. The axes are labelled alongside the percentage of variance within the Pfam abundance matrix that is explained by the corresponding principal component. The first principal coordinate divides the water and ice samples, this coordinate explains 47% of the total variance.

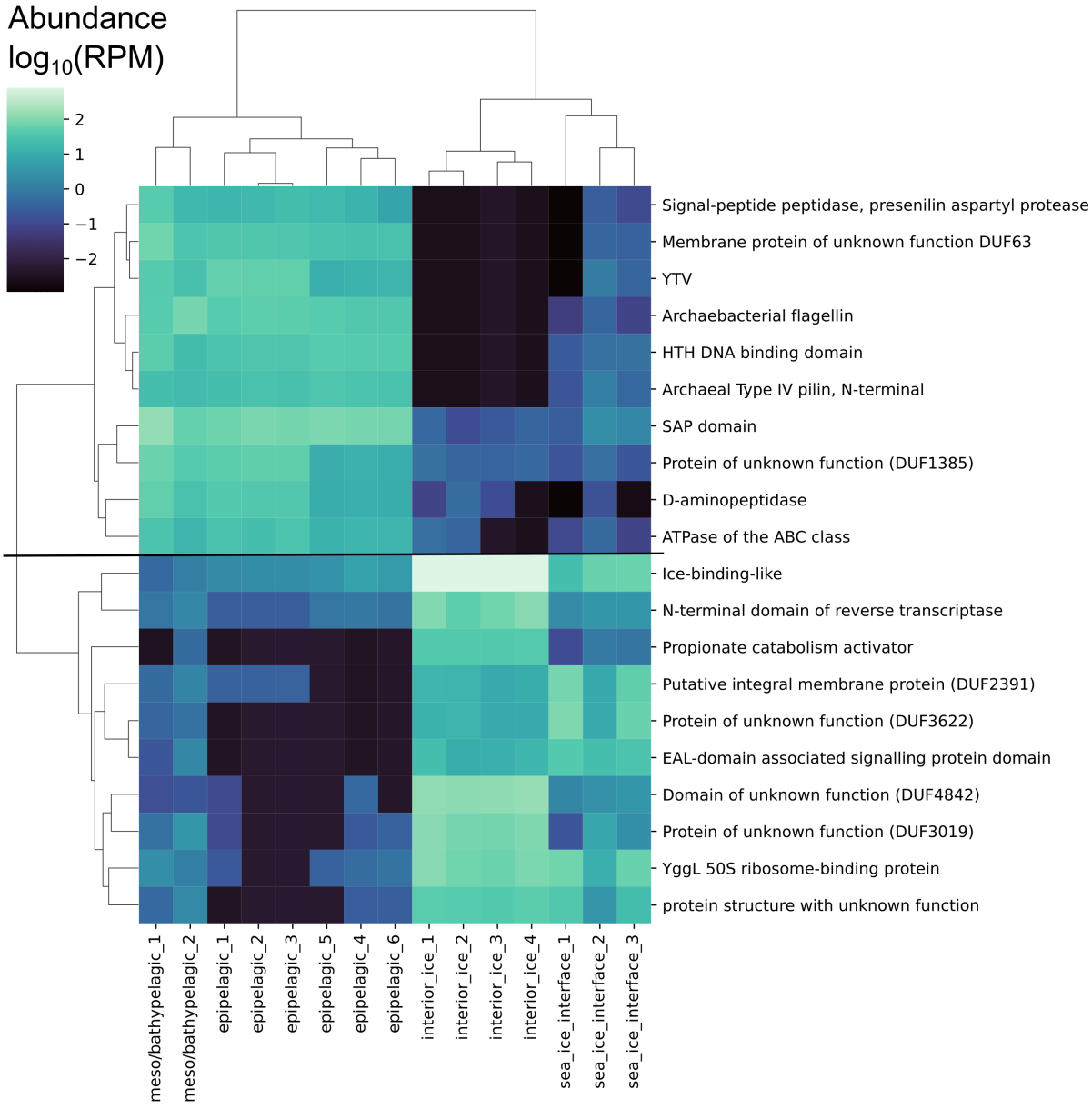


Figure 4.7: Clustered heatmap showing the 10 most differentially abundant Pfams within water (top half) and ice (bottom half) when compared all assembled contigs. Note the log-scale. Only the DUF3494 ice-binding-like domain had an RPM abundance above 100 in any sample.

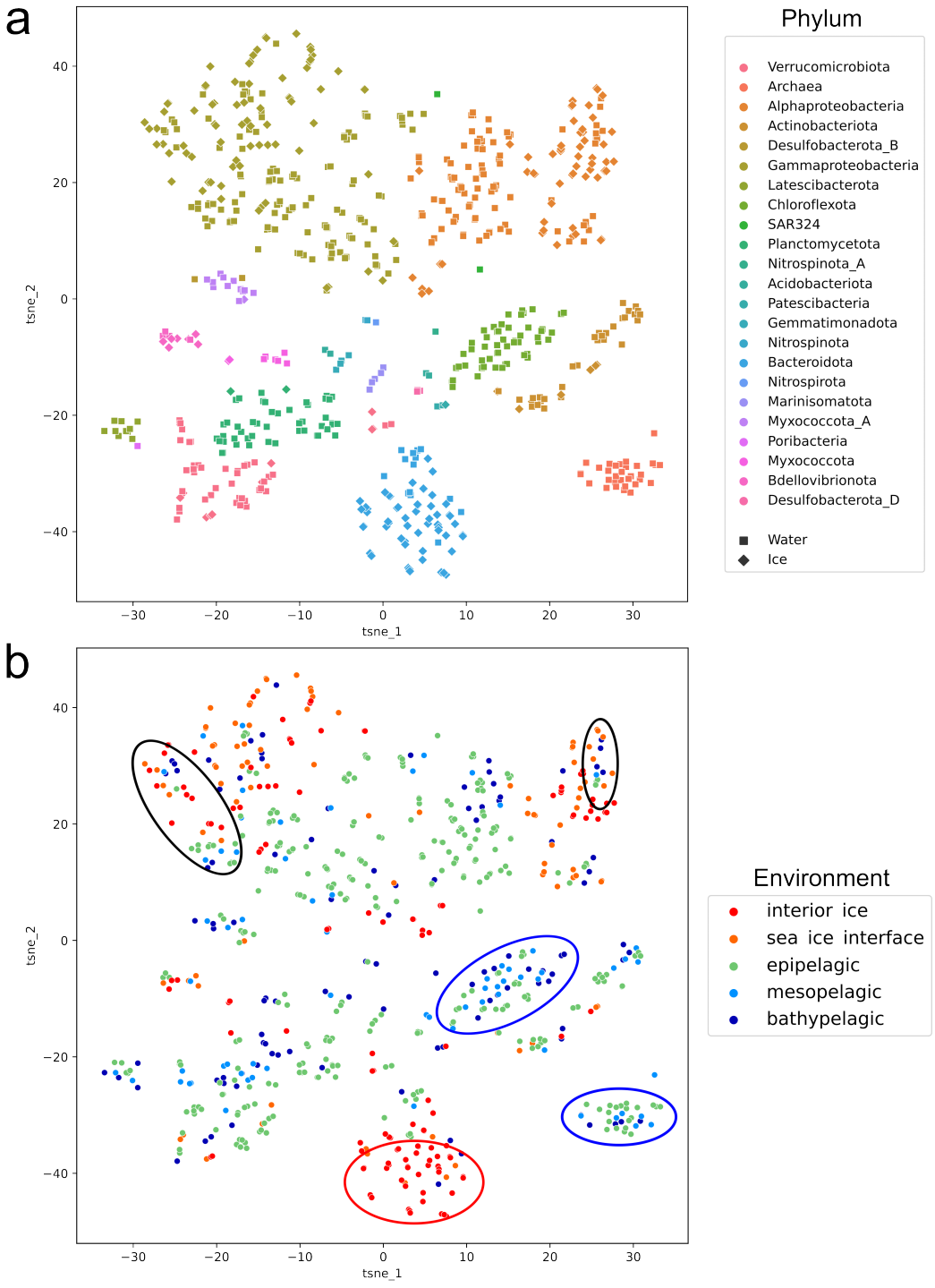


Figure 4.8: t-SNE plot showing prokaryotic MAGs, based on their metabolic profiles (Pfams). Mags are coloured by phylum in **a**, by environment in **b**. Some putative specialists are circled; Bacteroidota (red) for ice, archaea (bottom right, blue) and Chloroflexota (middle, blue) for water. Two putative generalist clades of Gamma- and Alphaproteobacteria are circled in black (top left/right, respectively).

Only the Bacteroidota appeared to have clades which were sympagic specialists; the other groups with MAGs recovered from ice also had functionally similar MAGs recovered from other environments.

4.4 Results from Eukaryotic Coassemblies

4.4.1 Coassembly Statistics

There were a total of 4,092,885 contigs generated from the HAVOC coassembly, with an N_{50} of 1092, and a minimum contig size of 500 bp. The overall size of the assembly was 4.33 Gbp, and the longest contig was 322 kbp. Within the pilot coassembly, there were 821,549 contigs with an N_{50} of 1952 and a minimum size of 500 bp. The total assembly size was 1.26 Gbp, and the longest contig was 467 kbp in length.

4.4.2 Eukaryotic MAGs

We generated a total of 56 eukaryotic MAGs; 27 Stramenopiles, 19 Chlorophyta, 3 Sporidiobolaceae, 1 Kinetoplastida (Figure 4.9), and 6 which were lacking enough marker genes to be placed on a phylogenetic tree, but which included 3 Chlorophyta and 3 Bacillariophyta based on the most abundant taxonomy of their contigs. There were 13 MAGs generated through coassembly, those remaining were from single-assembly. The single-assembled MAGs were predominantly from just two clades. Sixteen *Micromonas* MAGs were all extremely closely related to one another, and to a reference genome *Micromonas sp.* AD1. Thirteen MAGs were closely related to the *Fragilariopsis cylindrus* reference genome, while three were nested between *Pseudo-nitzschia multiseries* and *Nitzschia potreida*. The most complete MAGs were from the genus *Micromonas*; the highest quality MAG had a completeness of 98%.

4.4.3 Case-Study of a High-Quality MAG

From our eukaryotic coassembly, we flagged bin havoc.90 from the HAVOC dataset as having good contamination and completeness scores, measured by BUSCO (version 5.1.1, odb_eukaryota_10 gene set), with completeness and contamination of 73% and 0.4% respectively. This MAG, *Bacillariophyceae sp.* MOSAICH1_1, was annotated using the JGI Phycocosm pipeline; in brief, scaffolds were masked for repeats with RepeatMasker [271], and genes were called using a combination of Genemark-ES (version 4), GeneWise 51 (version 1), and fgenesh 52 (fgenesh1_pg), with the best fitting model for each locus picked to form a set of filtered gene models. Genes were functionally annotated for signal peptides, transmembrane domains, and assigned functional descriptions including assignment of Pfam, GO, KOG and KEGG terms, and genes were formed into gene clusters using the MCL algorithm [272]. This genome had a length of 32.07 Mbp, contained in 2,963 scaffolds, with

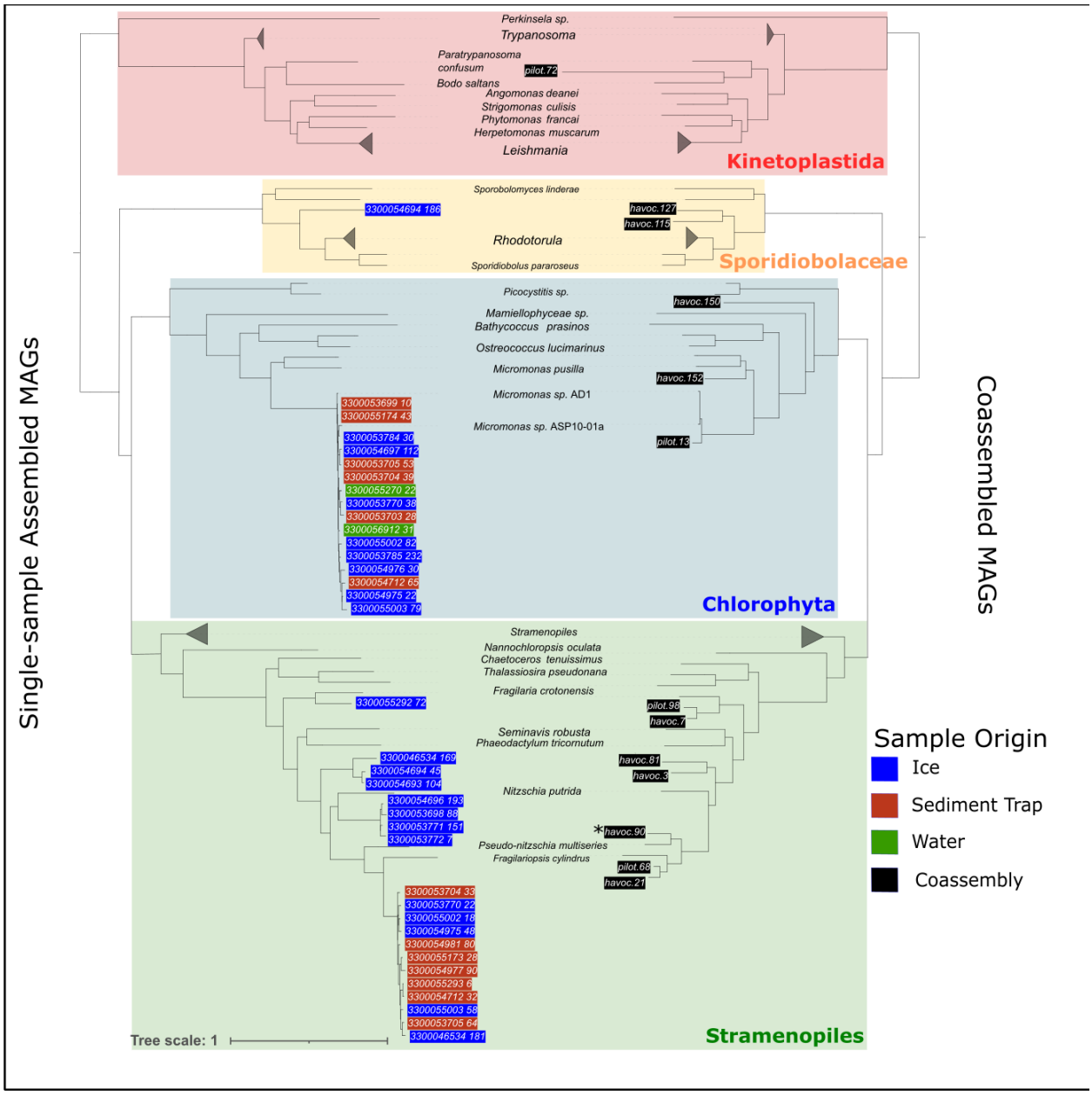


Figure 4.9: Phylogenetic trees showing MAGs from singly assembled samples (left) and MAGs from coassemblies (right). Reference genomes, common to both trees, have their leaves aligned and are labelled at the centre of the tree, with collapsed clades represented in the tree by a wedge. Some leaves of reference genomes are unlabelled for legibility, where they share their genus with their closest relative in the tree. MAGs labelled are shown on the tips of each tree. For each MAG, the colour of the background indicates the type of the sample (or indeterminate type, in the case of the coassemblies). The coassembled MAG havoc.90, marked with an asterisk *, had good completeness and contamination scores and was subsequently renamed *Bacillariophyceae sp.* MOSAICH1_1.

13,169 genes called among the set of filtered gene models. The most closely related isolate genome in Phycocosm was the diatom *Pseudo-nitzschia multiseries* CLN-47, with an average nucleotide identity of 76%, estimated with FastANI (version 1.33) [273].

4.5 Discussion

Prokaryotic MAGs in the Pilot Samples

We found a much higher diversity of prokaryotic MAGs within water than in ice, and that ice was dominated by species from the phyla Bacteroidota, Alphaproteobacteria, and Gammaproteobacteria. This is consistent with 16S rRNA gene sequencing studies of Arctic sea ice, such as Bowman *et al.* [95], though we found a smaller proportion of Cyanobacteriota than that study; this is likely due to the fact that our samples were collected during the polar night. We also found a much higher proportion of Bacteroidota in interior ice than had been reported in that study, with over 50% of abundance made up of Bacteroidota in the interior ice samples. This compares with approximately 20% in Bowman *et al.* [95], though a key methodological difference was how ice cores were processed, with the MOSAiC expedition using sections of ice cores to explore changes in microbial community with vertical space.

By making use of MAGs, we were able to gain an understanding of the genomic context of IBPs, as well as an overview of how metabolic and functional potential affected species niches. The t-SNE ordination showed a few groups of MAGs which had similar metabolic profiles (labelled by black ovals in Figure 4.8), but which appeared in multiple different environments. These might be generalists, and it would be interesting to compare the metabolic profiles of MAGs from these clades that were recovered from both ice and water (particularly at the strain level), however we leave this for future work. In contrast, some clades, such as Archaea, and Chloroflexota, were only present in the pelagic samples.

Coassembly and Eukaryotic MAGs

Coassembly is a resource intensive operation, and coassembly on a terabase-scale has only recently become possible, due to the JGI's assembler MetaHipMer [274], run on their supercomputing infrastructure. For most other institutions, this amount of computing resource is still out of reach. Our pilot coassembly made use of the assembler MetaSPAdes, using a high-memory compute node at the UEA (almost 1TB of RAM), while the HAVOC coassembly used 500 CPU nodes, each with 256 GB of RAM, at NERSC. While the HAVOC coassembly completed in under a few hours, the pilot coassembly ran for over 10 days, using a highly limited computing resource at the UEA (the high-memory nodes). Both of these were some-

what impractical to scale, and in later chapters, we used MEGAHIT instead [275], which is known to have lower resource requirements, as well as an approach known as 'bin-splitting' by the authors of VAMB, where multiple assemblies are concatenated, bins are generated, and then split into their sample-specific sub-bins [276]. We found, as expected from Hofmeyr *et al.* [258], that coassembly lost strain-specific resolution of MAGs, though this was compensated by an increase in the average quality of the MAGs recovered (with the exception of *Micromonas* MAGs), as well as a higher diversity overall in the MAGs recovered.

The eukaryotic MAGs themselves were dominated by two genera; *Micromonas* and *Fragilariopsis* (specifically, *F. cylindrus*). Although most MAGs of these genera were recovered from the sea ice, or from the sinking sediment traps, *Micromonas* were found to be much more relatively abundant in the water compared to other eukaryotic species. In Chapter 7 we explore this in more detail. While the number of eukaryotic MAGs is tiny compared to prokaryotic MAGs, certain species such as *Micromonas polaris* and *Fragilariopsis cylindrus* can have an outsized impact on biogeochemical cycles and food webs in the Arctic ocean [277], [278], so generating new genomes of these species from the environment, both at a strain resolved level and as a consensus-type coassembled genome, will be valuable to better understand how these species may adapt to a changing Arctic climate.

4.6 Data Records

All MAGs are available through Figshare [279], at NCBI BioProject PRJNA1160706 [280], and replicated in the GOLD database [281], Study ID 505419. Annotations of *Bacillariophyceae sp.* MOSAICH1_1 are also available from Phycocosm, and via Figshare, as indicated above. Individual read files for samples are stored in the NCBI SRA, with BioSample, BioProject, and SRP accessions listed in Supplementary Table 1 in [249], and as citations [282]–[354].

4.7 Code Availability

The custom pipelines used for eukaryotic MAG binning and annotation are available at https://github.com/willboulton/mosaic-pilot-havoc-mags.

Chapter 5

An Exploration of Ice-Binding Proteins

This chapter is a case study into one particular family of proteins; ice-binding proteins (IBPs) containing the domain of unknown function (DUF) 3494. This name DUF3494 is a historical artifact from the Pfam naming system - the domain was categorised as a DUF and only later structurally and biochemically characterised to show that it was ice-binding. IBPs are produced by many psychrophilic organisms, such as the diatom *Fragilariopsis cylindrus* and the sea ice bacterium *Colwellia sp.* [355], [356]. These proteins inhibit ice crystal formation and depress the freezing point of water [357], [358], aiding survival in cold environments such as sea ice by preventing or limiting cellular damage. Fusion proteins (such as an ice-binding protein, attached to a longer mobile peptide) have also been proposed as acting like tethers to the ice by some bacteria [359]. Despite having diverse functions, the underlying principle is the same; IBPs adhere to the surface of ice, and this makes it energetically less favourable for further ice to form. In IBP literature, this is called ice-recrystallisation inhibition, while the depression of the freezing point (a consequence of the Gibbs-Thomson effect) is thermal hysteresis.

The most common family of IBP are those containing the DUF3494 domain - a protein domain of approximately 200 amino-acids in length [356]. DUF3494 IBPs (hereon just IBPs) are known to be particularly diversified in polar species, particularly those inhabiting sea ice. In Dorrell *et al.* [93], enrichment p-values for Arctic versus non-Arctic species were 20 orders of magnitude smaller for IBPs compared to any other protein family. While this chapter focuses on the DUF3494 family, it should be noted that several other structurally distinct IBP families have been characterised. In the Vance *et al.* [358] review, at least 10 other classes of IBPs with distinct architectures were identified in polar organisms; mostly polar fishes, plants and insects. At least some of these seem to have convergently evolved an ice-binding function and have limited structural homology to the DUF3494 domain, or each other, though they often share a flat face that binds ice.

IBPs are important biotechnologically, and are known to be transferred between species of all domains of life through horizontal gene transfer (HGT) [360]. However there have not been any metagenomic studies surveying prokaryotic IBPs within the natural environment, (there are numerous studies of particular isolates from algae, fungi, and sea ice bacteria,

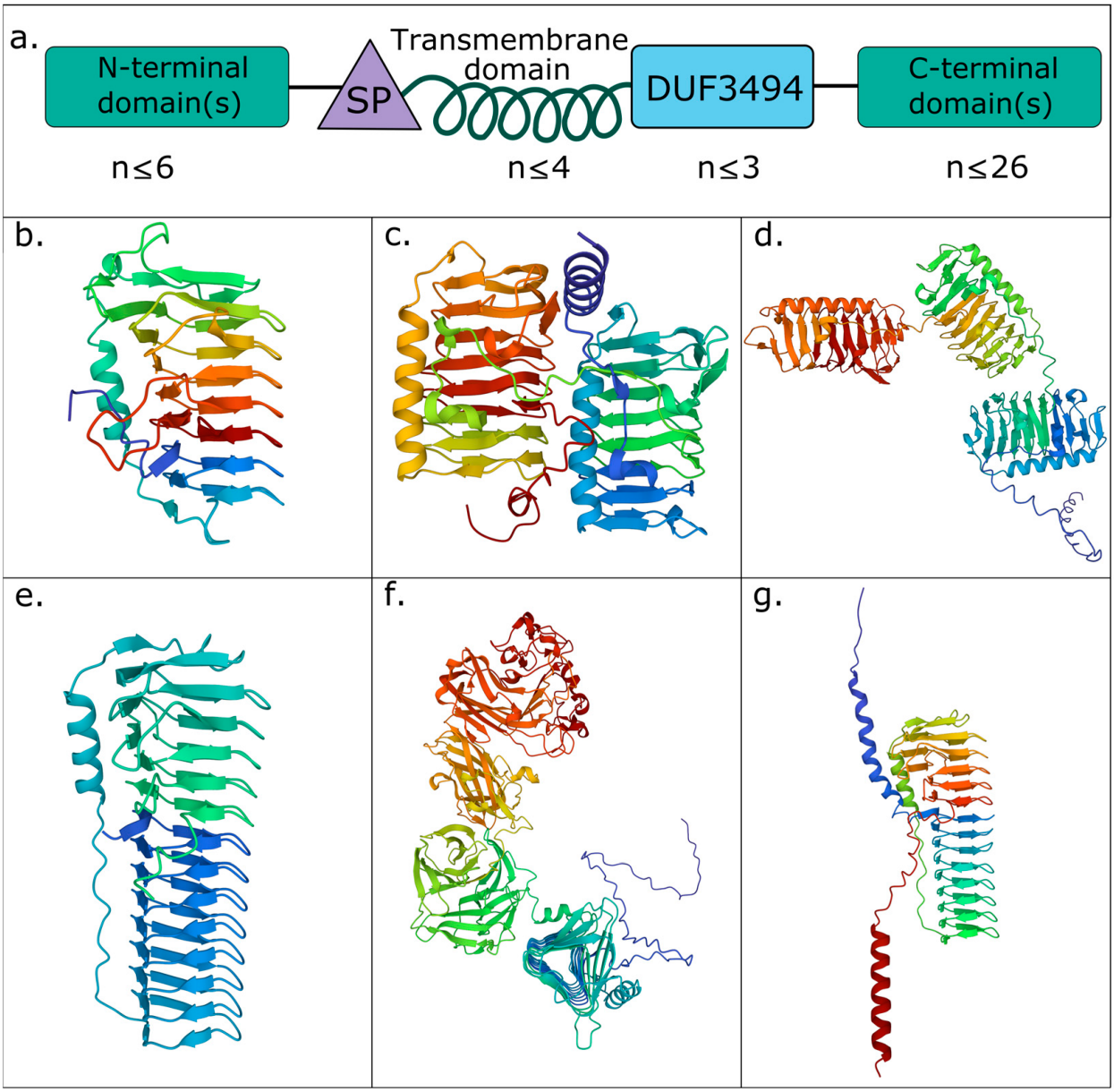


Figure 5.1: Examples of ice-binding protein structures, from Winder *et al.* [12]. The DUF3494 domain is the cylinder-shaped domain (β-solenoid) common to all 6 panels. Figure **a.** shows a concept diagram of the proteins surveyed; these proteins contained potentially up to six other domains toward the N-terminus of the protein, a signal peptide (SP), up to six transmembrane domains (TMDs), up to three copies of the DUF3494 domain, and up to 26 copies of other domains towards the C-terminus. Figure **b.** shows a protein comprised of a single IBP domain. Figures **c.**, **d.** show proteins comprised of 2 and 3 fused DUF3494 domains respectively. The vast majority of IBPs we surveyed were of the forms **b.** to **d.**. Figures **e.** to **g.** show proteins with more diverse domains fused to the DUF3494 domain; the DUF3494 domains in these proteins are also unusually elongated. The colour denotes the residue position, violet to red from the N- to C-terminus respectively.

see [355], [356], [359]–[363]). In this chapter, we survey the prokaryotic diversity of IBPs in the 15 MOSAiC pilot samples, using MAGs to provide a genomic context. We find diverse protein structures and gene architectures, which may have been due to domain shuffling, as well as evidence of horizontal gene transfer. We found ice-binding proteins (IBPs) to be one of the most differentially abundant protein domains when comparing sea ice and seawater, affirming their importance in Arctic species, particularly those living within sea ice.

5.1 Background and Summary

5.1.1 Sample Collection

We studied the 15 pilot samples, described in Chapter 4, (collection dates between 13 January 2020 and 7 February 2020). These samples were collected during the Arctic winter, from pelagic layers, with seawater collected via sampling from a CTD rosette, and from sea ice layers. The sample volumes used can be found in Table 5.1, which summarises the descriptions from Chapter 4. A more detailed map of the sample locations can be found in Figure 5.2, which defines the different subtypes that we used to categorise the samples (interior ice, sea-ice interface, epipelagic, meso/bathypelagic). Associated metadata are in Table 5.1 and Appendix A.2, which also provides the IDs of the relevant GOLD databases and SRA accession numbers.

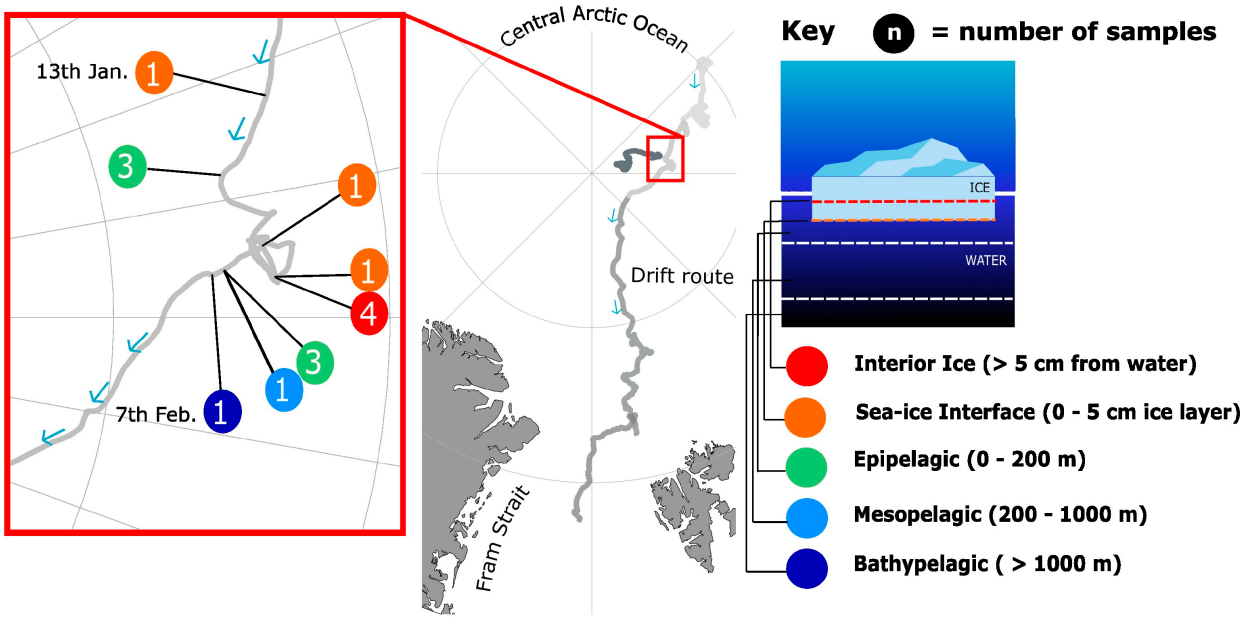


Figure 5.2: Drift route of the MOSAiC expedition, and with just the 15 pilot samples highlighted, from [12]. The red box shows the drift route of the RV Polarstern between the 13th January and the 7th February 2020. Co-occurring samples (either from the same CTD rosette, or neighbouring ice cores) are shown, with the number of co-located samples from the same environment circled. The schematic diagram on the right describes the environment of each of the samples.

Sample Label	Collection Date	Depth (m)	Habitat	Latitude	Longitude	Pooled?	Illumina Protocol	Filter Volume (μL)
sea ice interface 1	13/01/2020	0.00 - 0.05	Sea Ice	87.323	107.442		Low input	1220
epipelagic 1	16/01/2020	51	Seawater	87.551	102.085		Regular	8500
epipelagic 2	16/01/2020	51	Seawater	87.551	102.085		Regular	8500
epipelagic 3	16/01/2020	51	Seawater	87.551	102.085	Yes, from epipelagic 1 and 2	Regular	n/a
sea ice interface 2	27/01/2020	0.00 - 0.05	Sea Ice	87.445	95.670		Low input	1230
interior ice 1	03/02/2020	0.30 - 0.40	Sea Ice	87.412	93.215		regular	2920
interior ice 2	03/02/2020	0.05 - 0.30	Sea Ice	87.412	93.215		Regular	2150
interior ice 3	03/02/2020	0.50 - 0.60	Sea Ice	87.412	93.215		Regular	1930
interior ice 4	03/02/2020	0.40 - 0.50	Sea Ice	87.412	93.215		Regular	1180
sea ice interface 3	03/02/2020	0.00 - 0.05	Sea Ice	87.412	93.215		Low input	1500
epipelagic 4	06/02/2020	20	Seawater	87.595	94.084	Yes, from epipelagic 5 and 6	Regular	n/a
epipelagic 5	06/02/2020	20	Seawater	87.595	94.084		Regular	5500
epipelagic 6	06/02/2020	20	Seawater	87.595	94.084		Regular	8500
meso/bathypelagic 1	06/02/2020	202	Seawater	87.595	94.084		Low input	9500
meso/bathypelagic 2	07/02/2020	4082	Seawater	87.636	93.749		Low input	6000

Table 5.1: Sample location, processing and sequencing data. Depth represents the distance from the sea-ice interface (sea level). For these samples, the Illumina Low Input protocol differed from the Regular protocol only with the addition of 5 cycles of PCR.

5.2 Methods

MAGs and metagenomes were processed according to the method described in Chapter 4. The subsequent analysis of the IBPs used all genes that were annotated with the PF11999 Pfam domain, a HMMer profile developed by the Pfam team which categorises the DUF3494 domain, and using the model specific e-value cutoff developed for that HMM profile. We used the taxonomic annotations of contigs from MMSeqs2 (Section 4.2.3) to identify the prokaryotic fraction of each sample. The abundance of the PF11999 was measured using reads per kilobase million (RPKM), based on read mapping from BBMap (version 38.79) [146]. We used the Phobius web server [364] to further annotate transmembrane domains and signal peptides. The genomic context of each IBP gene within MAGs was explored using a custom script that we developed, to search for the closest 5 genes and Pfams adjacent to each IBP.

Barplots of community composition were generated used the phyloseq and ggplot2 packages in R (versions 1.40.0, 3.4.0, 4.3.3 respectively) [365], based on reads per million (RPM) abundances.

To determine how the phylogenetic relationships between IBPs varied depending on the domain architecture, environment and taxonomic assignments, we produced gene trees of the most environmentally abundant gene architectures, as well as gene trees of IBPs across all domain architectures. The alignments of the amino acid sequences of HMMER hits to the DUF3494 domain were produced using MUSCLE (version 2.0.4) [155], and low quality columns of the alignment were removed using TrimAl (version 1.2) [366]. The trees were generated with FastTree (version 2.1.1) [180], using the default parameters, and visualised using interactive tree of life (IToL; version 6.6) [367]. We repeated this method for IBPs within MAGs. Gene trees with fewer than 60 leaves, or with multi-copy DUF3494 domain architectures, were rooted at their midpoint. For the remaining trees, we rooted the trees using an outgroup of 130 IBPs from the dinoflagellate *Polarella glacialis* [368] (accessions in Appendix A.1).

5.3 Results

5.3.1 Prokaryotic IBP Functional and Taxonomic Profiles

Taxonomic Profile

We recovered a total of 3869 prokaryotic IBP genes from the 15 samples (Figure 5.3a), containing 4446 DUF3494 domains. Of these domains, 3581 were from the interior ice, 797 from the sea-ice interface, 60 from the epipelagic and 8 from the meso/bathypelagic. Of all IBPs, 85.7% could be assigned an order-level taxonomy, comprising 60 bacterial and 5 archaeal orders. The most common orders were from the Bacteroidota (Flavobacteriales, 1936 IBPs, 50.8%) and Gammaproteobacteria (Alteromonadales, 893 IBPs, 23.4%). There were very few IBPs assigned to archaeal orders, since archaea were far more abundant within water than ice. (No archaeal MAGs were recovered from ice, whereas 33 were recovered from water; 20 from the epipelagic and 13 from the meso- and bathypelagic.) Of IBPs from archaeal orders, the most common were 14 were from Methanomicrobiales (0.36%, phylum Euryarchaeota), and 4 from Candidatus Poseidonales (0.11%, Thermoplasmatota).

There were 199 IBPs encoded by 79 MAGs (Figure 5.3b), of which, 89 IBPs were found within 29 MAGs of the order Flavobacteriales, 19 were encoded by just 3 Acidimicrobiales MAGs (phylum Actinomycetota), and 18 were encoded by 8 Enterobacterales MAGs (phylum Gammaproteobacteria). The remaining 73 IBPs were from 8 other orders of bacteria, or from bacterial MAGs unidentified at the order level. Most MAGs encoding IBPs were recovered from the interior ice samples (67/79 MAGs), followed by the sea–ice interface (8/79 MAGs); just 3 and 1 MAGs containing IBPs were recovered from the epipelagic and meso/bathypelagic, respectively.

IBPs were one of the most differentially abundant Pfams, when comparing their enrichment level in ice to that in water (Figure 5.4). The IBPs had either a much higher abundance than other Pfams (222 total RPM abundance), or, had a much high two-fold change in abundance between ice and water (174 two-fold change). Other Pfams with a comparable or higher two-fold change were DUF4842, propionate catabolism activator, DUF3622 (two-fold change 701, 374, 229) with abundances of 35, 11, 14 total RPM. The Pfams with a higher abundance, and largest two-fold change were EAL domain, LytTr DNA-binding domain, methyl-accepting chemotaxis protein (1668, 321, 764 total RPM abundance) with the much lower two-fold changes of 9.5, 9.4, 9.2 respectively.

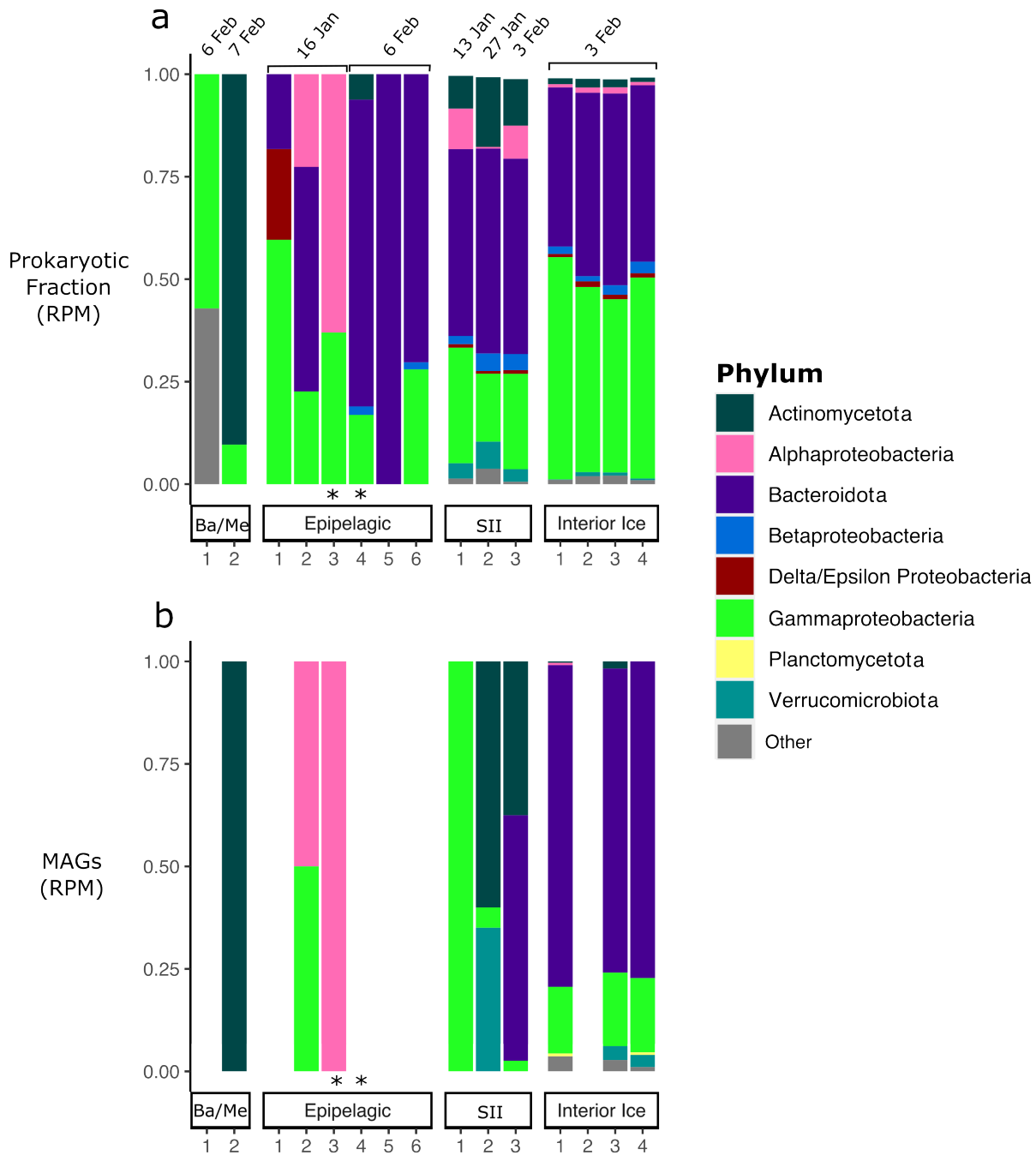


Figure 5.3: Relative abundance of contigs containing an IBP, either from the whole metagenomes (**a**), or just the fraction within MAGs (**b**). In the pelagic layers there are far fewer IBPs overall, but those present have a higher proportion of Alphaproteobacteria (pink) within Epipelagic samples 2 and 3. Otherwise, Gammaproteobacteria (light green) and Bacteroidota (purple) dominate.

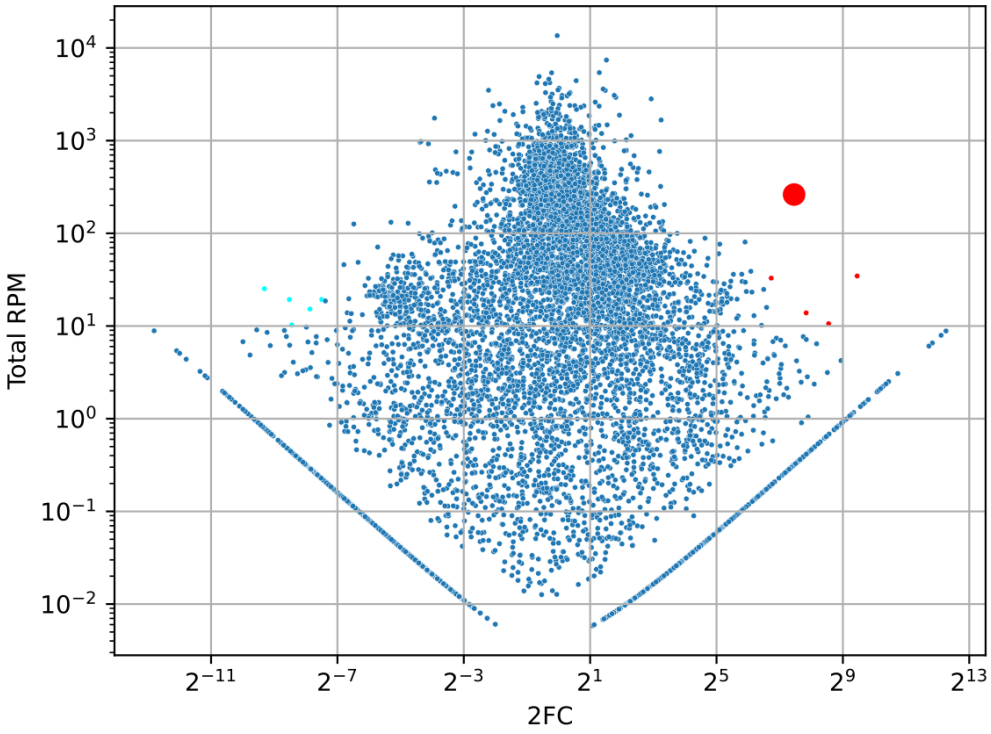


Figure 5.4: Two-fold change of Pfam abundance in ice and water, and Pfam total abundance, log-scaled axes. The five Pfams with the highest and lowest two-fold changes are highlighted in red, and light blue, respectively. The DUF3494 Pfam (pfam11999) is represented by the large red marker. Pfams lying on diagonal lines (all with a total RPM of less than 10) represent cases where there are no sequence representatives in ice or water respectively; the two-fold change in those cases with 0 in the numerator or denominator is ill-defined.

Gene Architectures

IBPs had a wide variety of different gene architectures (Table 5.2), but these were predominantly one of just a few kinds; 82% of IBPs abundance was either single domain (henceforth sdIBPs, containing just a single DUF3494 domain), or double domain (henceforth ddIBPs, consisting of two DUF3494 domains linked by a peptide of usually between 15 to 40 amino acids). This was the case based either on prevalence (86% IBPs were either single or double domain) or on abundance (82% of the total IBP RPKM). The remainder of IBPs genes did however display a broad range of different architectures; there were 116 different gene architectures in total, where gene architecture was categorised by the order of the Pfams within that gene. Of these, 101 appeared in fewer than 10 IBPs, and 74 of those appeared just once or twice. Only 0.4% of IBPs contained 3 DUF3494 domains, the remainder all contained either one or two; either alone (in the case of the sdIBPs and ddIBPs), or in conjunction with other Pfams. This is roughly consistent with the gene architectures found on Interpro (accessed June 2025), where by far the most abundant gene architecture were the sdIBPs, followed by 284 other IBP gene architectures. Out of all IBPs, 52% contained a signal peptide and 11% contained at least one transmembrane domain, however, within the sdIBPs, these values were 40%, and 9.4% respectively, whereas for the ddIBPs, they were 58% and 4.0%.

Excluding the single and double domain IBPs, the most abundant gene architectures were those involving a β-barrel bacterial immunoglobulin-like fold (DUF4842, pfam16130). This gene architecture had a total abundance of 5.1%, and appeared in 2.4% of the IBPs. DUF4842 was the most differentially abundant Pfam, in terms of two-fold change between ice and water (and with a total abundance of over 10 RPM). The most prevalent other IBP architectures were those with Pfams involving cell adhesion and exopolysaccharides. Architectures of this type together constituted 5.8% of the relative abundance, and 4.9% of all IBPs found. Some examples of this type, notable for their length, were IBPs with two DUF3494 domains and up to 26 C-terminal thrombospondin type-3 repeats (pfam2412), and single and double domain IBPs containing up to 7 C- or N-terminal bacterial immunoglobulin-like (BIg) domains (pfam13205). IBPs containing an N-terminal PEP-CTERM motif (pfam07589), involved in a protein sorting / exopolysaccaride function were the most abundant architecture of this type (2.6% abundance, 2.4% of IBPs); 65% of IBPs with this architecture had at least one transmembrane domain (60/92), well above the average. Of the IBPs that had an adhesion function, 1% contained only a transmembrane domain, 15% contained only a signal peptide, and 33% contained both.

Domain Architecture	Protein Family	Broader Function	Abundance (RPKM)	Fraction of IBPs (%)	Fraction with TMD (%)	Fraction with SP (%)
PF11999	DUF3494	sdIBP	4983.46	74.59	9.36	39.92
PF11999_PF11999	DUF3494	ddIBP	1660.68	11.76	3.96	57.8
PF11999_PF16130	DUF4842	β-barrel Ig-fold	413.49	2.22	19.77	37.21
PF11999_PF07589	PEP C-term motif	Sorting / Exopolysaccharides	209.68	2.38	65.22	53.26
PF11999_PF11999_PF11999	DUF3494	Triple domain IBP	93.11	0.44	5.88	41.18
PF11999_PF11999_PF01345	DUF11	Cell wall related	63.74	0.59	8.7	73.91
PF11999_PF02010	REJ domain	Membrane associated	58.44	0.36	0.0	35.71
PF04519_PF11999	Polymer-forming cytoskeletal	Cytoskeleton	47.68	0.34	0.0	76.92
PF11999_PF11999 + PF13517_PF13517_PF07593	FG-GAP-like repeat, ASPIC and UnbV	Cell adhesion	44.35	0.16	0.0	100.0
PF11999_PF11999 + PF13517_PF13517_PF13517 + PF07593	FG-GAP-like repeat, ASPIC and UnbV	Cell adhesion	25.93	0.1	0.0	100.0
PF11999_PF03797	Autotransporter β-domain	Secretion	24.84	0.23	0.0	66.67
PF13205_PF13205_PF11999	BIg-like domain	Tethering	24.45	0.59	0.0	69.57
PF11999_PF11999 + PF02412_PF02412_PF02412 + PF02412_PF02412	Thrombospondin type 3 repeat	Cell adhesion	21.36	0.13	0.0	80.0
PF11999_PF01391	Collagen triple helix repeat	Cell adhesion	20.53	0.31	0.0	66.67

Table 5.2: Table of the most abundant IBP domain architectures, and the proportions of those with signal peptides (SP) and transmembrane domains (TMD), and abundances. We grouped Pfams into broader categories based on the information provided by the Interpro database.

Other broad classes of Pfam functions within the diverse gene architectures (i.e. not sdIBPs or ddIBPs) were DUF domains (6.53% abundance; 3.15% of IBPs), calcium binding proteins (0.78%; 1.16%) and trafficking/secretion-related proteins (1.15%; 0.70%). Within the MAGs, the most prevalent domain architectures after sdIBPs (112/199 IBPs) and ddIBPs (41/199) were the single domain IBP and DUF4842 architecture (9/199), the single domain IBP and PEP C-term motif (6/199) and the triple domain IBP (3/199).

Genomic Context of IBPs within MAGs

We used the MAGs to explore which Pfams were present in genes upstream and downstream of IBPs, as well as the distribution of IBPs within MAGs. Out of 79 MAGs containing IBPs, 46 contained more than one IBP, and two contained 9 IBPs (these were from the orders Acidimicrobiales and Flavobacteriales). In MAGs with multiple IBPs, these IBPs were frequently found in the same contig, immediately upstream or downstream of one another. Furthermore, these IBPs often had identical domain architectures, e.g., double domains. The most frequent domain architectures found downstream of IBPs in MAGs were single domain IBPs (6.12%) and double domain IBPs (4.76%). Following this, the five most abundant downstream domain architectures contained small solute membrane transport proteins (MFS; pfam07690; 3.40%), bacterial 2-component systems containing a DNA binding domain and a response regulator receiver domain (pfam04397_pfam00072; 2.72%), an antioxidant enzyme (AhpC/TSA family; pfam00578; 2.04%), DNA topoisomerase (pfam01131_pfam01751; 2.04%) and a phosphodiesterase (pfam01663; 2.04%).

5.3.2 Prokaryotic Ice-Binding Protein Phylogenetics

We generated phylogenetic gene trees of IBPs, based on alignments of just the DUF3494 domain (Figures 5.5, 5.6). We found that structurally diverse prokaryotic IBPs are phylogenetically widely distributed, and that the presence of signal peptides was distributed across the tree, somewhat evenly. A large clade of mostly monophyletic Bacteroidota IBPs seemed to contain the majority of ddIBPs; these contained a higher proportion of signal peptides (58%) compared to the background rate (52%), and a lower proportion of transmembrane domains (4.0%). Of the 455 ddIBPs, 298 were within Flavobacterales, 81 within other Bacteroidota of unknown order. Just 26 were within Gammaproteobacteria (most frequently Alteromonadales), 18 within Actinomycetota (most frequently Acidimicrobiales), and 16 within all other phyla combined. There were two clearly distinct clades within the Bacteroidota, with one clade closer in the phylogenetic tree to a clade containing almost all ddIBPs from non-Bacteroidota phyla.

A second mostly monophyletic clade of Bacteroidota IBPs was enriched with double and triple domain IBPs, this clade contained the majority of the triple domain IBPs. IBPs of this kind were otherwise found within Gammaproteobacteria and Actinomycetota, though overall these were few in number; there were just 17 triple domain IBPs in total and 9 were found within Bacteroidota (4 Flavobacteriales, 3 Cytophagales, 2 unidentified order), 4 within Actinomycetota (Micrococcales) and 4 within Gammaproteobacteria (2 Cellvibrionales, 1 Thiotrichales, and 1 unidentified order). Of the triple domain IBPs, 41% contained a signal peptide and 5.9% at least one TMD.

Other IBPs with diverse gene architectures were similarly formed several small subclades dispersed widely across the phylogenetic tree. IBPs containing the DUF4842 domain formed four separate closely related subclades within Gammaproteobacteria (predominantly, of the order Alteromonadales), accounting for 62 out of 86 IBPs with this gene architecture. The remainder were within Bacteroidota (either Flavobacteriales, or an unidentified order), within a monophyletic clade. IBPs containing a Pep C-term motif were found within at least 4 distinct clades, all attributed to Gammaproteobacteria (Alteromonadales), as well as a single clade of Verrucomicrobiota (Verrucomicrobiales). These two phyla represented 86 and 7 IBPs with this architecture, respectively. The DUF4842 and Pep C-term motif gene architectures had a higher prevalence of TMDs than the background rate; 20% and 65% of genes with these architectures had at least one TMD, respectively.

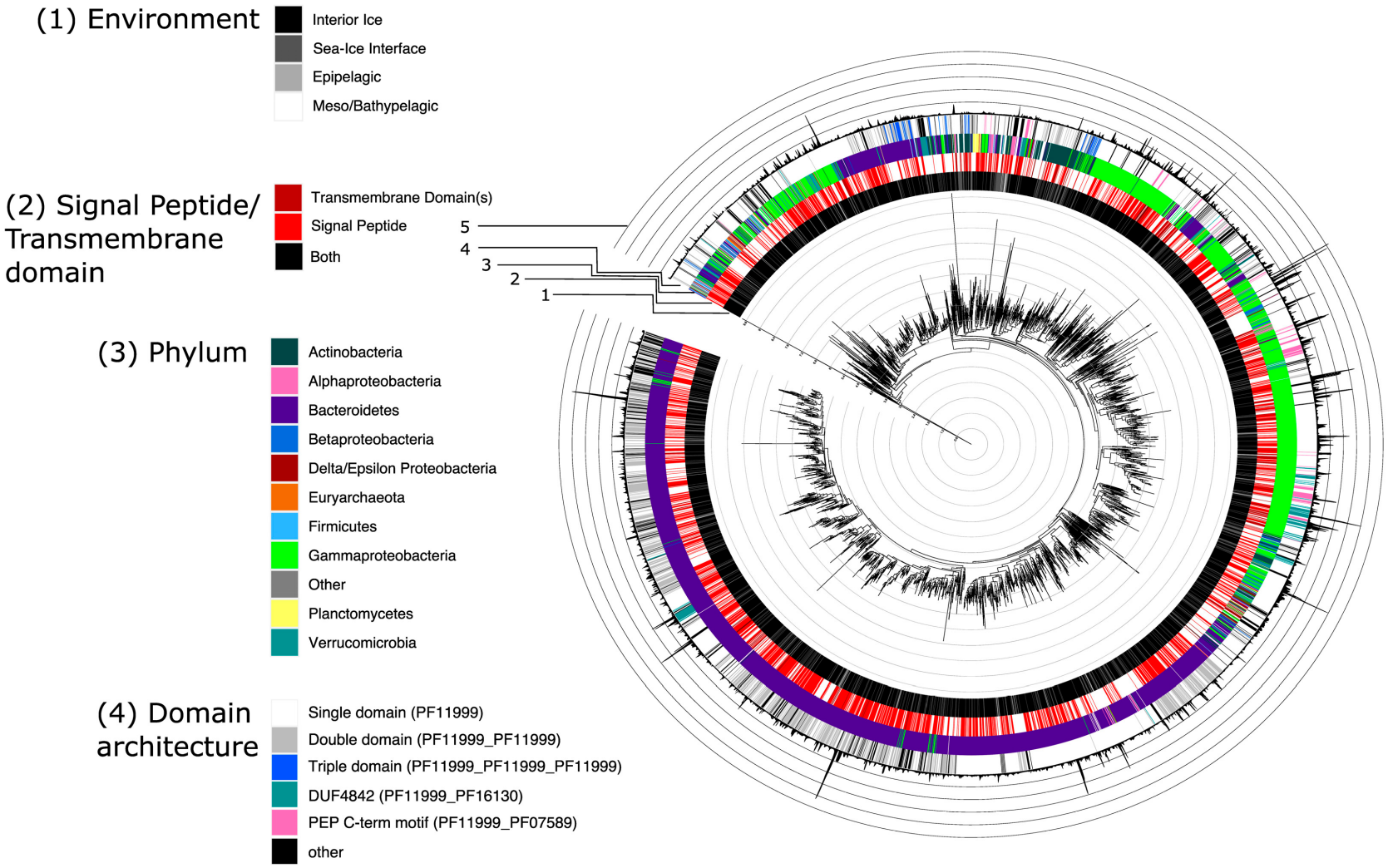


Figure 5.5: Phylogenetic gene tree showing all DUF3494 domains and rings marking (moving inwards); abundance in RPKM multiples of 10, gene architecture, phylum, signal peptide / TMD, and environment.

5.4 Discussion

Although prokaryotic IBPs were dominated in terms of abundance by just two gene architectures, we found a large amount of structural diversity within the genes surveyed. IBPs containing immunoglobulin-like domains, and domains with functions involving adhesion were abundant. Similarly, IBPs were found within a wide range of diverse taxa and environments, though again dominated by a relatively small number of taxonomic orders within the sea ice interior. We found that the likeliest genomic context for IBPs was other IBPs, and found that the majority of IBP-encoding MAGs contained more than one IBP, and in some cases up to nine. The prevalence of multiple IBPs within single MAGs may reflect several non-mutually exclusive mechanisms. First, gene duplication and subsequent functional divergence could allow a bacterium to produce IBPs with distinct ice affinities (e.g. targeting different crystal faces) or acting at different temperatures. Second, domain shuffling, as evidenced by the variety of gene architectures observed, could produce IBPs with different subcellular localisations or secondary functions (e.g. adhesion, exopolysaccaride biosynthesis). Third, some IBPs even within the same MAG appear phylogenetically diverse, consistent with horizontal gene transfer of multiple IBP copies from different donors. Having multiple IBPs may therefore confer a selective advantage by broadening the range of ice crystal structures that can be inhibited, or by allowing differential regulation under varying environmental conditions. Additionally, bacteria use operons as a means of co-regulating proteins [369], and it may be the case that adjacent IBPs are within the same operon and are duplicated as a means of regulating overall IBP expression relative to other genes within the operon, though this is something we leave to test in future work.

Just over 50% of IBPs did not contain a signal peptide, and so could have putatively had some intracellular function. Although there are other mechanisms for extracellular secretion (some of which were found within the gene architectures of the IBPs surveyed), it seems plausible that a significant fraction of IBPs have an intracellular function, potentially preventing the formation of intracellular ice. Conversely, a higher proportion of the ddIBPs were found encoded by Bacteroidota and contained a signal peptide; Bacteroidota MAGs were putatively ice-specialists and it is possible that the ddIBPs play an important role within the sea-ice community, and in exopolymeric substances within sea ice [370].

We found that IBPs clustered taxonomically, but sometimes there was a complicated relationship between gene phylogeny and taxonomy; many clades of IBPs were polyphyletic and there were at least two highly distinct clades of IBPs that were predominantly encoded within Bacteroidota, and were present even within the same Flavobacteria MAG. There was also no straightforward relationship between gene architecture and phylogeny, although

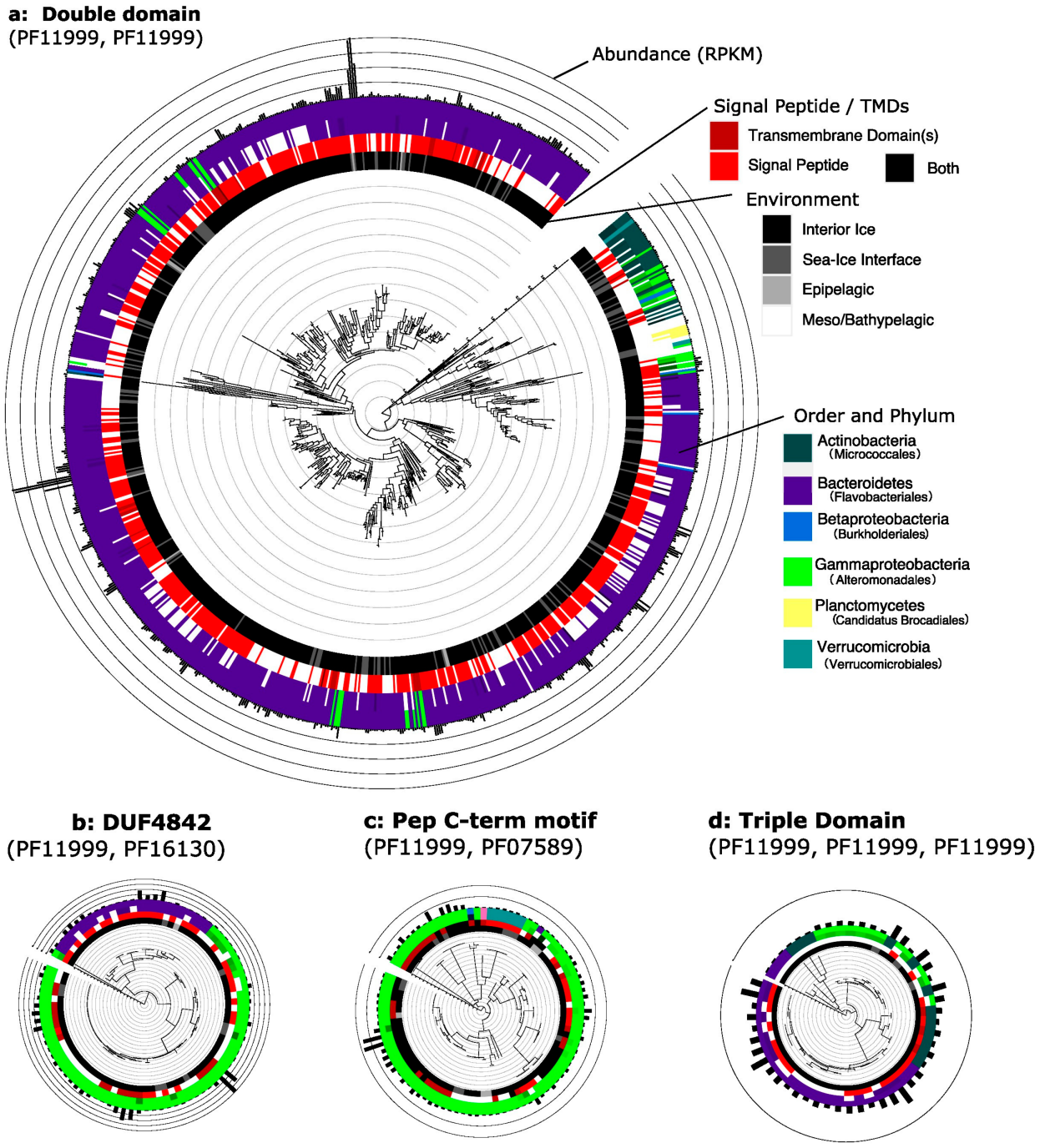


Figure 5.6: Gene trees of the DUF3494 domains with the most abundant gene architectures, excluding the sdIBP. From outside to centre, the rings represent: environmental abundance, taxonomy (phylum, then order), transmembrane domain / signal peptide presence, and sampled environment.

there were some links, such as between the ddIBPs and Bacteroidota, and the Pep C-term motif (an exopolysaccaride/sorting function) and Gammaproteobacteria. We found several were flanked by transcriptases, indicating a possible mechanism for gene duplication. Some bacteria are believed to have acquired their IBPs through HGT [371], however, our results suggest that in addition to this, instances of gene duplication and domain shuffling may be positively selected for, if the resulting proteins fulfil more specific roles within (or adjacent to) sea ice. Our results provide the first survey of IBPs in the CAO, and highlight the utility of MAGs as a resource for genome-resolved studies of genetic diversity within environmental samples.

5.5 Data Availability

Sequence data are available for download from the NCBI SRA, accession numbers list in Appendix A.2, and Appendix A. Code for the analysis of IBPs and MAG and Pfam diversity are available on Github at https://github.com/willboulton/mosaic-pfam, as well as in https://github.com/willboulton/mosaic-ibps. The Winder *et al.* [12] paper is appended to the end of this thesis, after the appendices.

Chapter 6

Refining Eukaryotic MAGs with UMAP Visualisations

In this chapter we present a pipeline for recovering and visualising eukaryotic contigs, with the aim of refining eukaryotic MAGs from metagenome bins. We benchmark our method against popular binning tools (MetaBAT2, SemiBin2, and VAMB [276], [372], [373]) in a straightforward way, focussing on the quantity and quality of eukaryotic MAGs generated, and on resource usage (e.g. computing time / memory). We evaluate our method on different subsets of metagenomes from MOSAiC. We also qualitatively compare our visualisation method against another method that is highly representative of the other bin-visualisation methods available.

6.1 Introduction

The generation of prokaryotic MAGs from metagenomes has made the new field of genome-resolved metagenomics possible, contributing to our understanding of microbial species in a range of habitats including the human microbiome, surface ocean, soil, as well as extreme environments such as hydrothermal vents [103], [274], [374], [375]. More recently, some progress has been made on recovering eukaryotic MAGs, such as in [104], [248], [260]. Several bioinformatics pipelines have also incorporated tools for eukaryotic MAG generation, including automatic quality control checks using EukCC and BUSCO [259], [269]. However, the recovery of eukaryotic MAGs is still a challenging task, due to the more repetitive and complex eukaryotic genome, and greater average genome length [376], [377]. While recent surveys have recovered prokaryotic MAGs from marine environments in the tens of thousands [56], [378], the collection of eukaryotic MAGs is much sparser; there are on the order of a few thousand medium quality eukaryotic MAGs recovered from the ocean in all eukaryotic binning studies combined, with no more than 500 recovered within a single study, though this number has more recently been increasing at a faster rate [379].

Publication	Dataset	Number of Samples	Method Used	Number of Euk. MAGs (Quality Filtered)
West *et al.* [194]	Various public metagenomes	17	EukRep	20
Olm *et al.* [380]	Infant gut	1198	EukRep	14
Duncan *et al.* [104]	Tara Oceans Arctic/Atlantic**	11	EukRep + Custom binning	21
Delmont *et al.* [248]	Tara Oceans**	939	Coassembly + Anvi'o	324*
Alexander *et al.* [260]	Tara Oceans**	441	EukHeist	485*
Xu *et al.* [381]	South China Sea	24	Coassembly + EukRep	28
Tagirdzhanova *et al.* [382]	Lichen metagenomes	456	Binning + EukCC	326
Espinoza *et al.* [383]	Biofilm, marine	7	VEBA / VEBA2	8
Rocha *et al.* [384]	Various public metagenomes	574	MuDoGeR	5
Michoud *et al.* [385]	Glacier-fed Streams	156	Custom binning	42
Seong *et al.* [386]	Human microbiome	167	Coassembly, ACR	36
Peng *et al.* [387]	Goat gut	43	MetaBAT2	18*
Saraiva *et al.* [388]	Multiple Terrestrial	6000	EukRep	121*
Boulton *et al.* [11]	MOSAiC Pilot and HAVOC	73	Coassembly + Tiara	56*

Table 6.1: Numbers of eukaryotic MAGs recovered from metagenomic datasets. Most methods use a strategy of first identifying eukaryotic contigs (commonly using EukRep [194]) followed by binning. Often, similar samples were coassembled. The 'Method Used' column covers all of the major bioinformatics pipelines with a eukaryotic binning component. *These studies used 30% completeness as a quality cut-off when reporting numbers of MAGs. **These studies included size-fractionated samples enriched with eukaryotic phytoplankton.

In this chapter, we set out a method for the generation, visualisation and manual refinement of eukaryotic MAGs. Our method is called Visualising Autoencoder Latent-Space Embedding Network for Clustering Eukaryotes (VALENCE). Other data visualisation methods exist for examining metagenomic data, particularly bins; VizBin [190] and Anvi'o [389] being the two most popular and highly cited. Two more recently published tools, BinaRena and BusyBee Web [390], [391], function as a more feature-rich and easily accessible successors to VizBin.

Tool Name and Publication	Tool Application	Notes
EukRep [194]	Contig Classifier	Used as first stage in multiple euk. pipelines
Tiara [195]	Contig Classifier	Similar principle to EukRep
Eukfinder [392]	Contig and Read Classifier	
EukHeist [260]	Bioinformatics Pipeline	Uses coassembly, EukRep, MetaBAT2
VEBA / VEBA2 [142], [383]	Bioinformatics Pipeline	Uses Tiara, MetaBAT2, CONCOCT for euk. binning
EukCC [259]	Euk. MAG Assessment	Also has capability for MAG refinement
BUSCO [269]	Euk. MAG Assessment	
Anvi'o [389]	Contig Visualisation and Refinement	Complex platform combining supervised binning and conventional binning
VizBin [190]	Contig Visualisation and Refinement	t-SNE on k-mers
BusyBee Web [390]	Contig Visualisation and Refinement	Platform for supervised binning
BinaRena [391]	Contig Visualisation and Refinement	Supervised binning only

Table 6.2: Metagenomic pipelines and software with a component for eukaryotic MAG generation, refinement, or visualisation.

In terms of pipelines and tools for recovering eukaryotic MAGs (see Table 6.2 for an overview), most bioinformatics pipelines are based on contig domain-classifiers such as Eukrep or Tiara [195], [393]. These pipelines, namely EukHeist and VEBA2 [260], [383], first identify a non-prokaryotic component of contigs using either of these classifiers, and then run a binning method such as MetaBAT2 or CONCOCT [372], [394] on the reduced set of

contigs. Bins are then evaluated for quality with BUSCO or EukCC [259], [269].

6.2 Methods

6.2.1 Summary of the Pipeline

VALENCE consists of two modules; a Snakemake pipeline for taxonomic annotation of contigs, coassembly, binning, and the generation of the latent space of coordinates, and a Jupyter Notebook (an interactive web server) [395] for interacting with the visualisation of the latent space, converting highlighted bins into FASTA files. The pipeline assumes that each set of reads has been quality filtered, assembled by single-sample assembly, that these reads have then been mapped to the assemblies, and that gene-calling has been run on each assembly individually.

There are five overall stages to the pipeline, each of which will be described in the next subsections:

- The non-prokaryotic fraction (i.e. the fraction annotated as either viral, eukaryotic, or unknown) of each metagenome is identified using both MMSeqs and Tiara to taxonomically annotate contigs; additionally the user can provide their own set of annotations. Taxonomic annotations are then aggregated using a voting algorithm, and in cases of inconsistent annotations, the least trusted method is removed (as decided by the user) until the remaining methods come to a majority consensus.
- Reads are mapped to contigs using Strobealign [396].
- Reads from the non-prokaryotic fraction are coassembled using MEGAHIT [275]. Additionally, all non-prokaryotic contigs are collated into one file, for multi-binning.
- Eukaryotic bins are generated using both multi-binning and coassembly, with bins generated using MetaBAT2 and VAMB. These bins are then checked for quality with EukCC.
- UMAP is then applied to the latent space of the VAMB output, to produce a two-dimensional visualisation of the contigs, and MCL clustering is applied to the nearest-neighbour graph generated by UMAP, to produce an additional set of bins. These bins can then be further refined manually by interacting with the UMAP visualisations in a Jupyter Notebook.

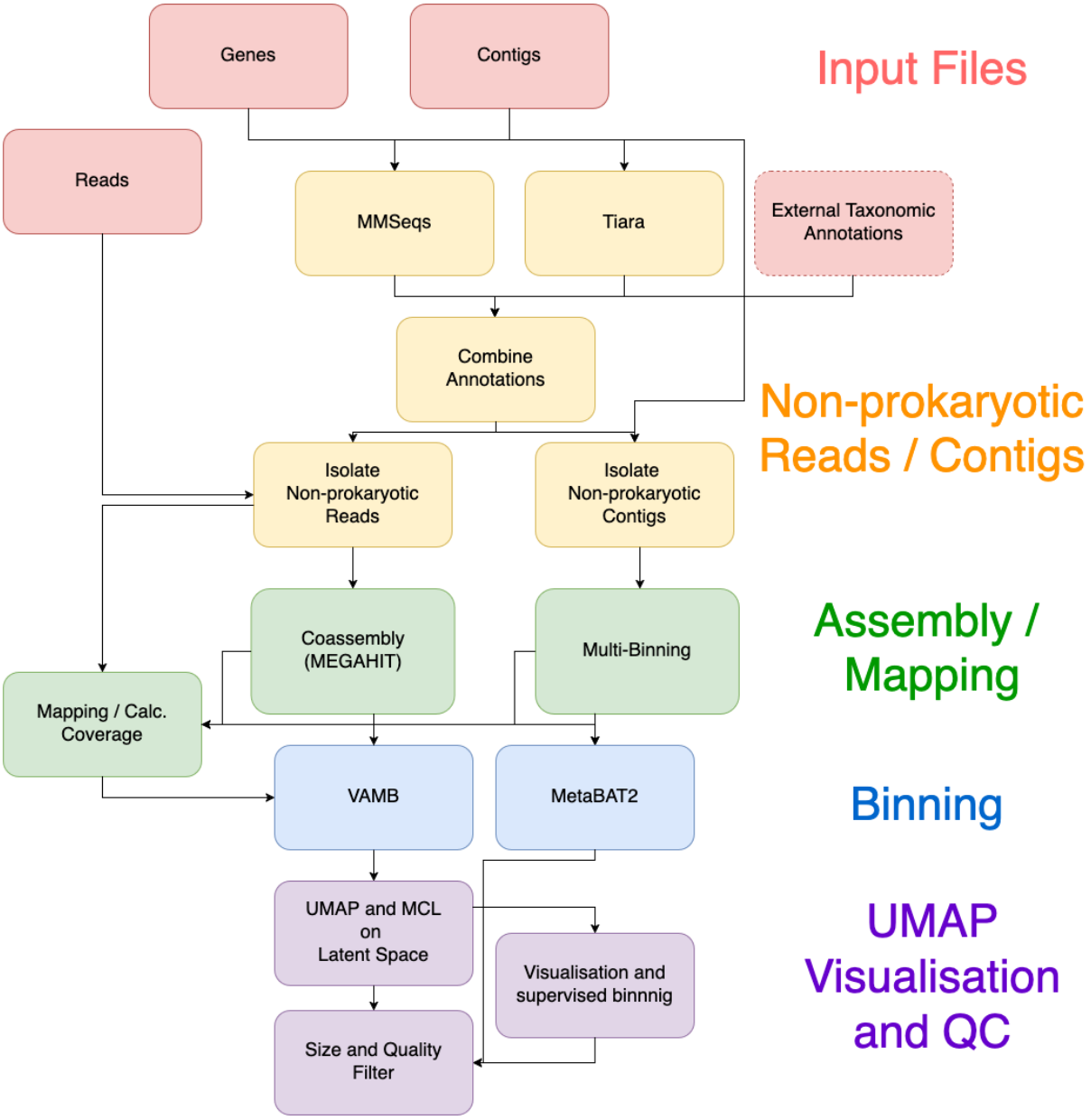


Figure 6.1: Stages of the VALENCE pipeline. The colours of the stages indicate a grouping: red for input files, yellow identifying non-prokaryotic data, green coassembly and read mapping, blue for binning, and violet for the UMAP visualisations and MCL clustering. A dotted line around a stage indicates that it is optional. Arrows indicate a dependency between stages. The user can optionally choose to run only coassembly or multi-binning stages, and to not run MetaBAT2 for binning or MMSeqs for taxonomic annotation.

Isolation of Non-Prokaryotic Reads and Taxonomy Estimation

To identify a non-prokaryotic fraction of reads within each sample, we estimate the taxonomy of each contig of single assemblies (supplied by the user), and pool together all reads that map to contigs of eukaryotic, viral, or unknown taxa. We discard unmapped reads; while this may reduce the quality of the final assemblies, it drastically reduces time and memory usage during assembly. Reads are then mapped to contigs using Strobealign (version 0.13.0) and sorted with Samtools. (version 1.16.1).

Discarding unmapped reads is not without limitations and other bioinformatics pipelines specialised for classifying eukaryotes (e.g. EukFinder [392]) pool unmapped reads, before applying a read classifier such as Centrifuge or Kraken to identify a eukaryotic component [397], [398]. However, these tools often rely on huge databases, normally several hundred gigabytes in size, and the reads they add are more likely to represent lower abundance organisms that are too rare to assemble, from genomic regions that are too repetitive, or from novel taxa with no close relative in the assembled contigs. Retaining and pooling unmapped reads (which in the two datasets described later in this chapter accounted for 11% and 14% of reads) from all samples before a further assembly step could, in principle, increase the representation of rare taxa in the coassembly, though this would substantially increase computational cost and may not greatly improve MAG recovery, since very rare taxa may still not reach sufficient coverage for assembly.

Contig taxonomy is estimated with the following methods:

- A custom MMSeqs2 (version ed4c55*) least common ancestor (LCA) method, whereby taxonomy is estimated for all amino-acid sequences (supplied by the user) on each contig, and then for each contig a consensus taxonomy is taken using a least common ancestor method, where we chose the most specific taxon supported by over 50% of annotations (excluding annotations at the rank of domain, or unknown). This avoids using the command 'MMSeqs taxonomy', which consumes a very large amount of time and memory per sample [399]. When running MMSeqs, we queried amino-acid sequences against a database combining UniRef50, Phycocosm, and MMETSP [10], [256], [267].

- The program Tiara (version 1.0.3) [195], which estimates contig taxonomy at the domain level based on a deep-learning classifier of k-mer composition ($k = 6$).

Additionally, we allow the user to supply their own contig taxonomy estimates via an external file. To generate a single taxonomic estimate at the domain level, we use the domain chosen by the majority of methods, and in cases of a tie, remove estimates one by one (in an

order supplied by the user; by default Tiara first, then MMSeqs, then external user-supplied annotations) until the tie is broken.

Read Mapping and Coassembly

For each sample, quality-filtered reads are mapped to contigs from their respective single-assembly using Strobealign, and then the resulting bam files sorted with Samtools using the command samtools sort. Paired reads that map to non-prokaryotic contigs are extracted using samtools view and samtools fastq, using the --regions-file parameter, and the flag -F 12. All successfully mapped reads are then concatenated and passed to MEGAHIT (version 1.2.9) using the flags --min-contig length 500, --presets meta-large, and with a default of 400 GB of memory. Additionally, all non-prokaryotic contigs from the single-assemblies are concatenated into a single file.

Binning and Abundance Estimation

Eukaryotic MAGs were generated using multiple binning algorithms, and a combination of both multi-binning and coassembly (see Figure 6.2 for a definition of these terms). We use abundance estimates of reads mapped to the non-prokaryotic coassembly to generate differential coverage of contigs within the coassembly, and further, we use differential coverage of reads across the concatenated set of non-prokaryotic contigs – a method termed multi-binning by the authors of VAMB. These abundance estimations are again performed using Strobealign. These differential coverages are produced from the alignment files using the script jgi_summarise_bam_contig_depths, which comes packaged with MetaBAT2.

Binning is then performed by VAMB (version 4.1.3) and MetaBAT2 (version 2:2.15). VAMB additionally generates a file which provides the coordinates of its latent representation of contigs.

UMAP ordination of the VAMB Latent Space

We generated further bins using a custom visualisation script. This uses the embedding space provided by VAMB as a feature set, which was then mapped into a two dimensional space using UMAP (UMAP-Learn, version 0.5.6). UMAP was run with default parameters (n_neighbours=15, Euclidean metric) on the VAMB latent space coordinates. The number of nearest neighbours (n_neighbours=15) determines the balance between local and global structure preserved in the embedding; smaller values preserve finer local structure, while larger values better preserve global topology. The default value of 15 was used as a starting

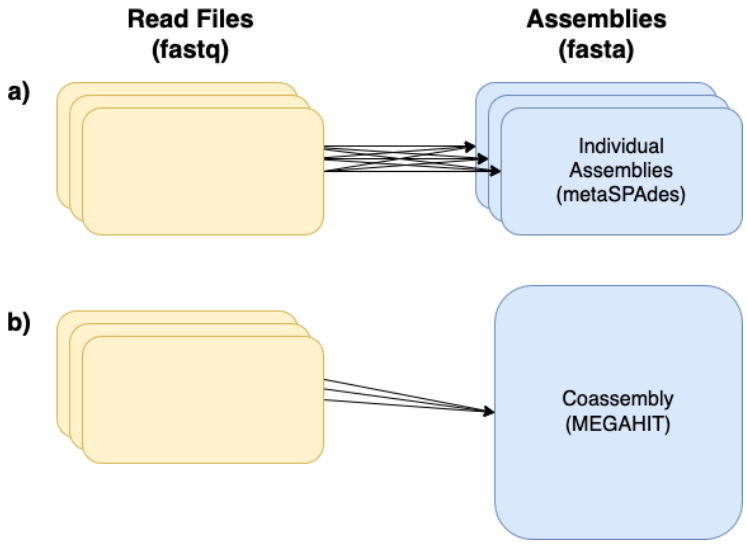


Figure 6.2: A schematic diagram of multi-binning and binning from a coassembly. Multi-binning (a) involves an all-versus-all mapping from reads to contigs. In (b), one coassembly is generated from a pool of all reads, and then each set of reads is mapped to the coassembly. In both cases, the differential coverage of reads across samples provides data that is utilised in binning algorithms.

point and found to produce visually interpretable clusters. UMAP also generates a k-nearest-neighbours graph based on the Euclidean distances in the latent space. Mathematical details of the UMAP algorithm are provided in Appendix B.1. The two dimensional representation of contigs can then be visually inspected for clusters by the user, and the k-nearest-neighbours graph is used as input for the Markov Clustering Algorithm (MCL, version 14-137). Clusters are extracted into bins, using SeqKit (version 2.8.2) to extract sequences into FASTA files based on their contig IDs, using the command seqkit grep. The user can use a Jupyter Notebook (an interactive web server, running Jupyter-core version 4.9.1) [395] to select contigs that are subsequently extracted into bins.

Deduplication and Quality Estimation

We filter bins using SeqKit to retain only those with a size of at least 5 Mbp, and use EukCC (version 2.1.1, database 1.2) to estimate completeness and contamination of bins, retaining only those with above 30% completeness and less than 15% contamination.

To ensure that MAGs do not overlap in their set of contigs, we developed a custom script, which identified contigs shared between MAGs. Contigs shared between two MAGs most likely arise either from the same organism being binned twice (redundancy due to overlapping reads mapped across coassembly and multi-binning strategies), or from genomic regions shared between closely related strains (e.g. conserved core genes). In either case, having

the same contig in two MAGs would artificially inflate their estimated completeness and create ambiguity in abundance estimation (reads would be double-counted). We therefore resolve overlaps conservatively: when the overlap is large (>100 shared contigs), we treat the two MAGs as likely representing the same genome and retain only the more complete representative. When the overlap is smaller, the shared contigs may derive from genomic regions conserved between related but distinct species; in this case we assign contigs to the more complete MAG and rerun quality estimation on the trimmed MAG.

6.2.2 Validation

We applied VALENCE to two datasets and compared its performance against other metagenomic tools; BinaRena (visualisation) and MetaBAT2, VAMB, and SemiBin2 (binning).

We ran our evaluation using real metagenomic from the MOSAiC expedition. We ran VALENCE on two sets of samples, grouped together based on their provenance (water or ice) and the month of year; MOSAiC_Ice_April ($N = 9$ metagenomes) and MOSAiC_Water_June ($N = 18$ metagenomes). These two datasets were chosen as they contained the largest numbers of eukaryotic MAGs from ice and water respectively. The sample IDs of metagenomes in each dataset are provided in Appendix B.2. We compared these bins with results from MetaBAT2, SemiBin2 (version 2.1.0), and VAMB, run in single-sample modes for each of the samples within each set, as well as results from coassembly and multi-binning, and manual bin refinement. Bins were generated using a minimum contig length of 2000 for MetaBAT2 and VAMB, and 4000 for SemiBin2. The higher contig length cut-off for SemiBin2 was necessary due to memory constraints; the memory usage for SemiBin2 scales quadratically with the number of contigs, so without aggressively filtering out smaller contigs the memory usage would have been multiple terabytes per sample, which was beyond the limit of the hardware available at the UEA (512 GB Ice Lake Intel Xeon Platinum 8358 nodes). This is the most likely cause of the poorer performance of SemiBin2 compared to the other binners. Bins were again assessed for quality using EukCC, and only bins above 30% completeness and below 15% contamination were retained. In addition to the deduplication method performed by VALENCE, we ran dRep (version 2.0.1) using an ANI threshold of 99% to identify clusters of highly similar MAGs.

Finally, we ran scripts provided by BinaRena (https://github.com/qiyunlab/binarena commit 279d358), using the k-mer compositions of contigs as an input, for $k = 5$, and qualitatively compared the contig visualisations for the MOSAiC datasets. We used a t-SNE visualisation with a perplexity of 30 on k-mers with $k = 5$, since all of BinaRena, BusyBee Web, and VizBin use t-SNE as a dimensionality reduction method by default,

though BinaRena and BusyBee can optionally support UMAP embeddings. We compared visualisations based on the coassemblies generated by VALENCE, filtering out contigs of size smaller than 1500 base pairs, since this produced the most like-for-like comparison; the BinaRena visualisation would otherwise be overwhelmed by an extremely large number of prokaryotic contigs.

6.3 Results

6.3.1 Case Study of Eukaryotic MAGs Recovered from MOSAiC

We compared binning results of the VALENCE pipeline against single-sample binning using MetaBAT2, VAMB, and SemiBin2, and also compared which binning strategy (coassembly, multi-binning, or single-sample binning) performed best in terms of recovering eukaryotic MAGs with the highest completeness, and least contamination. From the MOSAiC_Water_June and MOSAiC_Ice_April datasets, we recovered a total 116 and 91 eukaryotic MAGs respectively, before removing duplicates. From the dataset MOSAiC_Water_June, 71 eukaryotic bins were generated through single-sample binning (i.e., not using the VALENCE pipeline), 39 through multi-binning, and 6 through coassembly. In terms of the binning methods used, 44 (eukaryotic) bins were generated through manual refinement (i.e. interactively refining bins generated via MCL clustering from VALENCE using a Jupyter Notebook), 46 generated using MetaBAT2, 16 from SemiBin2, and 10 from VAMB. From MOSAiC_Ice_April, 59 bins were generated through multi-binning, 14 through coassembly, and 18 through the single-sample binning. The different binning methods generated 46 bins from MetaBAT2, 32 through manual refinement, 10 from SemiBin2, and 3 from VAMB.

From the MOSAiC_Water_June dataset (Figures 6.3-6.6), single-sample binning produced a larger number of smaller, less complete bins, though with a slightly lower mean level of contamination. The bins produced by this method were predominantly those from taxa with smaller average genome sizes; mostly Bacillariaceae and *Micromonas* (roughly 60 Mbp and 20 Mbp respectively). Of the 70 single-sample-binned eukaryotic MAGs that could be assigned a taxonomy, 44 were *Micromonas*, 20 Bacillariaceae, 4 were other Chlorophyta, and 2 were Haptophyta. We deduplicated the MAGs, based on the heuristic of counting shared contigs described in Section 6.2, as well as using a 99% ANI similarly threshold, and from each cluster of similar MAGs, retained only the MAG with the highest completeness, and contamination lower than 5%. This generated 33 MAGs, of which 26 were from multi-binning, 4 from coassembly, and 3 from single-sample binning. Coassembly using the VALENCE pipeline was the only method that managed to generate a haptophyte MAG at above 50%

completeness (2 MAGs; 58% / 9% and 53% / 6% completeness / contamination respectively). The recovery of rarer taxa (e.g. Haptophyta, Fungi, Metazoa) exclusively through coassembly and multi-binning is consistent with the hypothesis that these taxa, being at low abundance in any single sample, provide insufficient read depth for single-sample assembly and binning. Pooling reads across samples in coassembly effectively increases the read depth for rare taxa, improving the contiguity of the assembly; the longer contigs that are generated are subsequently easier to bin. A recently developed binning algorithm, Bin Chicken [400], takes advantage of this, by iteratively binning contigs, and pooling the remaining reads.

The majority of the deduplicated eukaryotic bins (29 out of 33) were generated using manual bin refinement. Within each taxon, manual refinement increased the completeness of MAGs by up to 42% (mean of 10.1%, s.d. 14.5% increase), though it did also increase the contamination, by a mean of 0.8% (s.d. 1.0%). Highly similar MAGs are generated through two effects, one is an artifact from the computational pipeline and the other is biological. The computational artifact is from each sample being present both within a coassembly, and a single-assembly; in some cases a MAG generated from the coassembly is essentially just representative of a single dominant sample. In the cases of difficult-to-bin taxa (e.g. Haptophyta), coassembly generates slightly higher quality bins, and increasing the completeness cut-off to 50% can remove some of the 'duplicate' bins. In other cases such as with common and abundant taxa such as *Micromonas*, the multiple clusters highlight that there are many strains of the same species in the samples, and persist when the completeness cut-off is raised to 50%. The 99% ANI threshold is higher than the species boundary used in GTDB taxonomy, so within-cluster MAGs can be considered the same species by this criterion.

Within the MOSAiC_Ice_April dataset (Figures 6.7-6.10), single-sample binning produced bins with a lower mean contamination, though once again only produced bins within the clades of Bacillariaceae and *Micromonas*. In these clades, completeness was between 32% and 18% lower than coassembly and multi-binning, though contamination was also lower, by 2.5% and 0.3% respectively. Two fungal MAGs were generated through coassembly with VALENCE (completeness 31.4 and 32.0%, contamination 0.4%, for both MAGs), and two metazoans, most likely copepods, were generated through multi-binning (completeness 31% and 59%, contamination 5.6% and 12.2%). Once deduplicated, and the most complete MAG chosen from each cluster, there were 69 representative MAGs, of which 50 were generated through multi-binning, 13 through coassembly, and 6 through single-sample binning. In terms of the binning method, 35 were generated by MetaBAT2, 26 by manual refinement, 6 by SemiBin2, and 2 by VAMB. Within the Phaeodactylaceae, manually refined MAGs had 1.3% higher completeness than MAGs assembled by MetaBAT2 through multi-binning,

though also had 2.5% higher contamination. Within the Bacillariaceae, the most complete MAG was a 95.2% complete, 3.1% contaminated MAG generated through coassembly and refined manually, improving on VAMB by 0.8% in terms of completeness, though with 0.3% higher contamination.

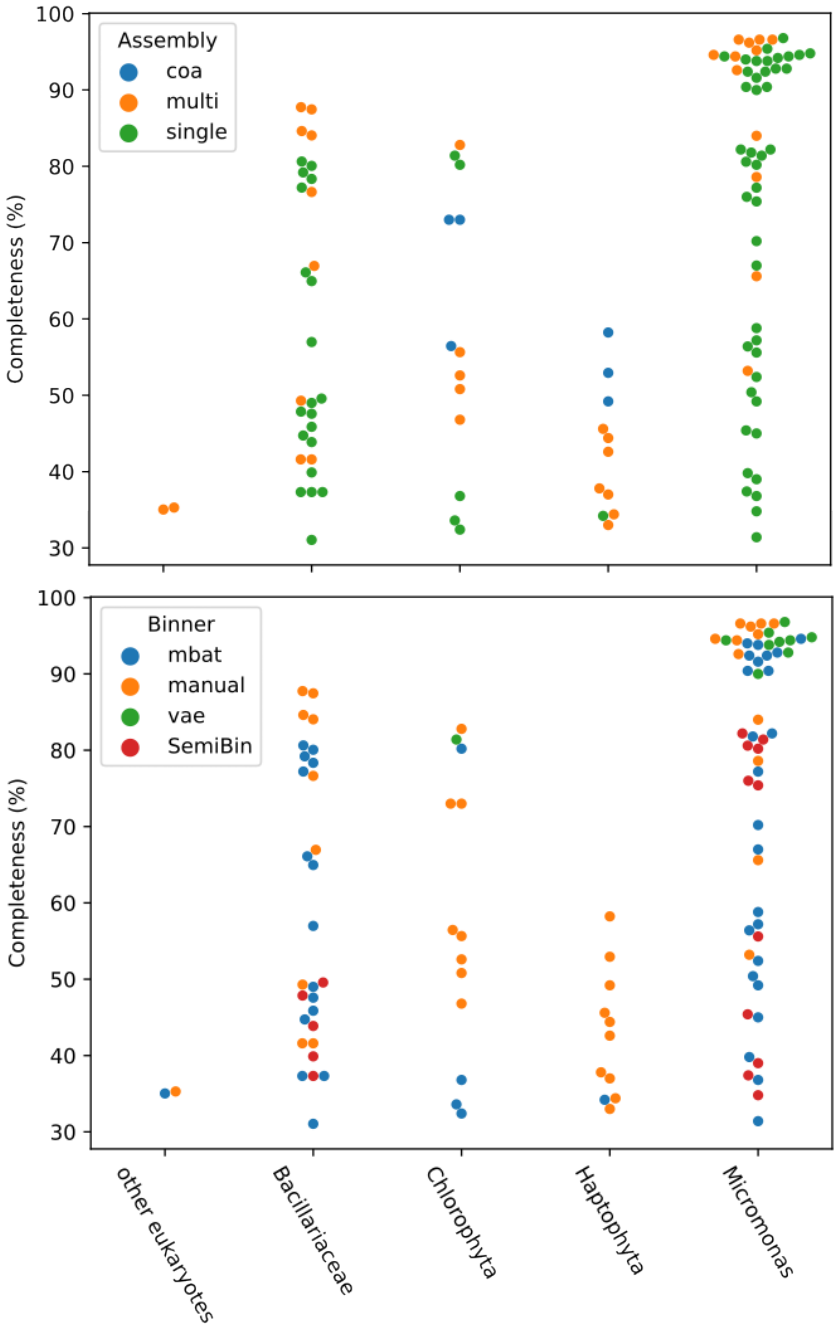


Figure 6.3: Completeness of all MAGs (before deduplication) in the MOSAiC_Water_June dataset, for different taxa. Chlorophyta denotes all non-*Micromonas* chlorophytes. Top: MAGs are coloured based on the assembly method used; coassembly, multi-binning, single-sample binning. Bottom: MAGs are coloured based on the binning method used to generate them: mbat, MetaBAT; vae, VAMB.

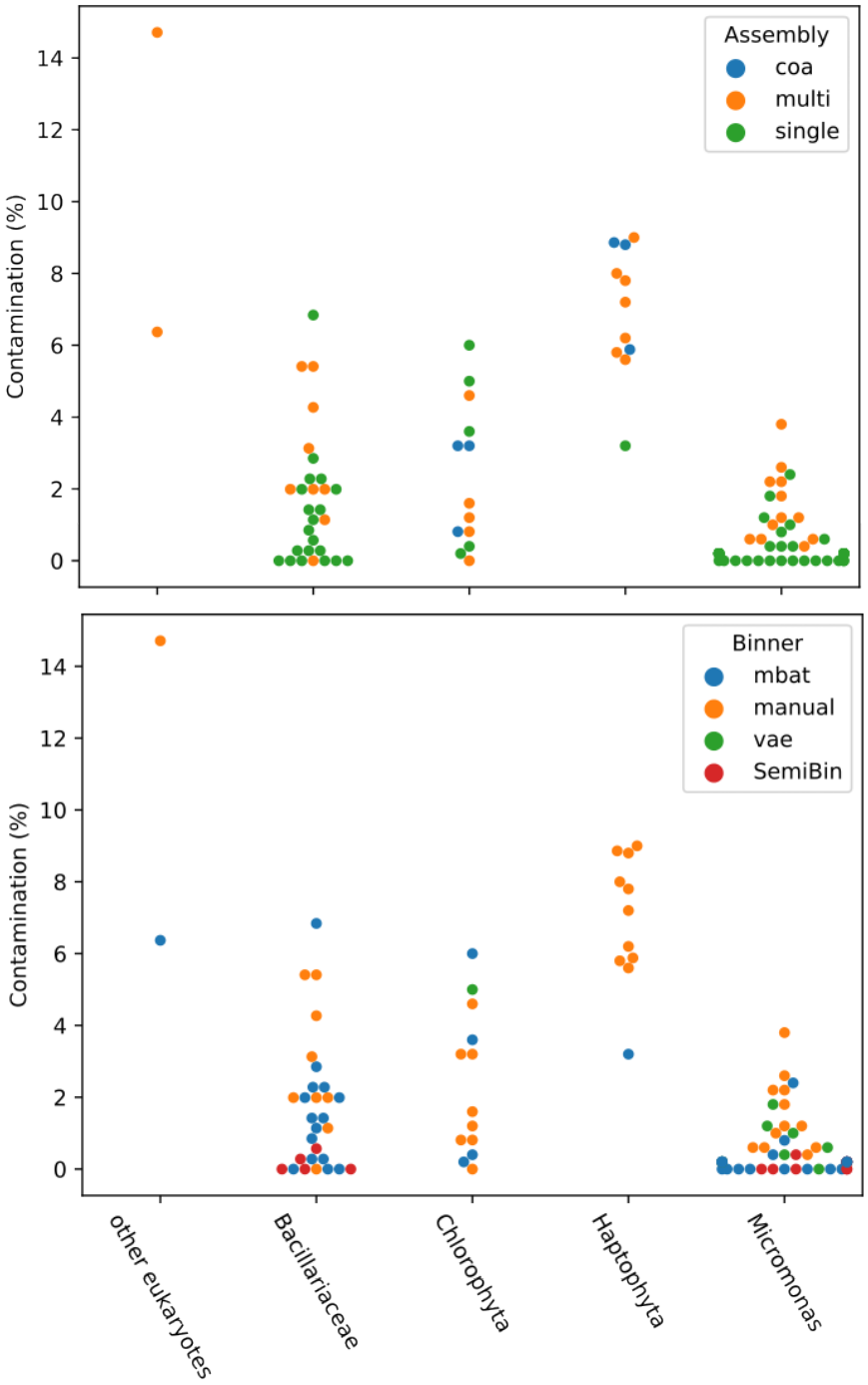


Figure 6.4: Contamination of all MAGs (before deduplication) in the MOSAiC_Water_June dataset, using the same colours and formatting as Figure 6.3. Single-sample binned MAGs tended to have a lower contamination, though at the cost of a lower completeness (see Figure 6.3), and were generally only able to recover Bacillariaceae and Chlorophyta MAGs. Manually binned MAGs tended to have a slightly higher completeness, but also higher contamination.

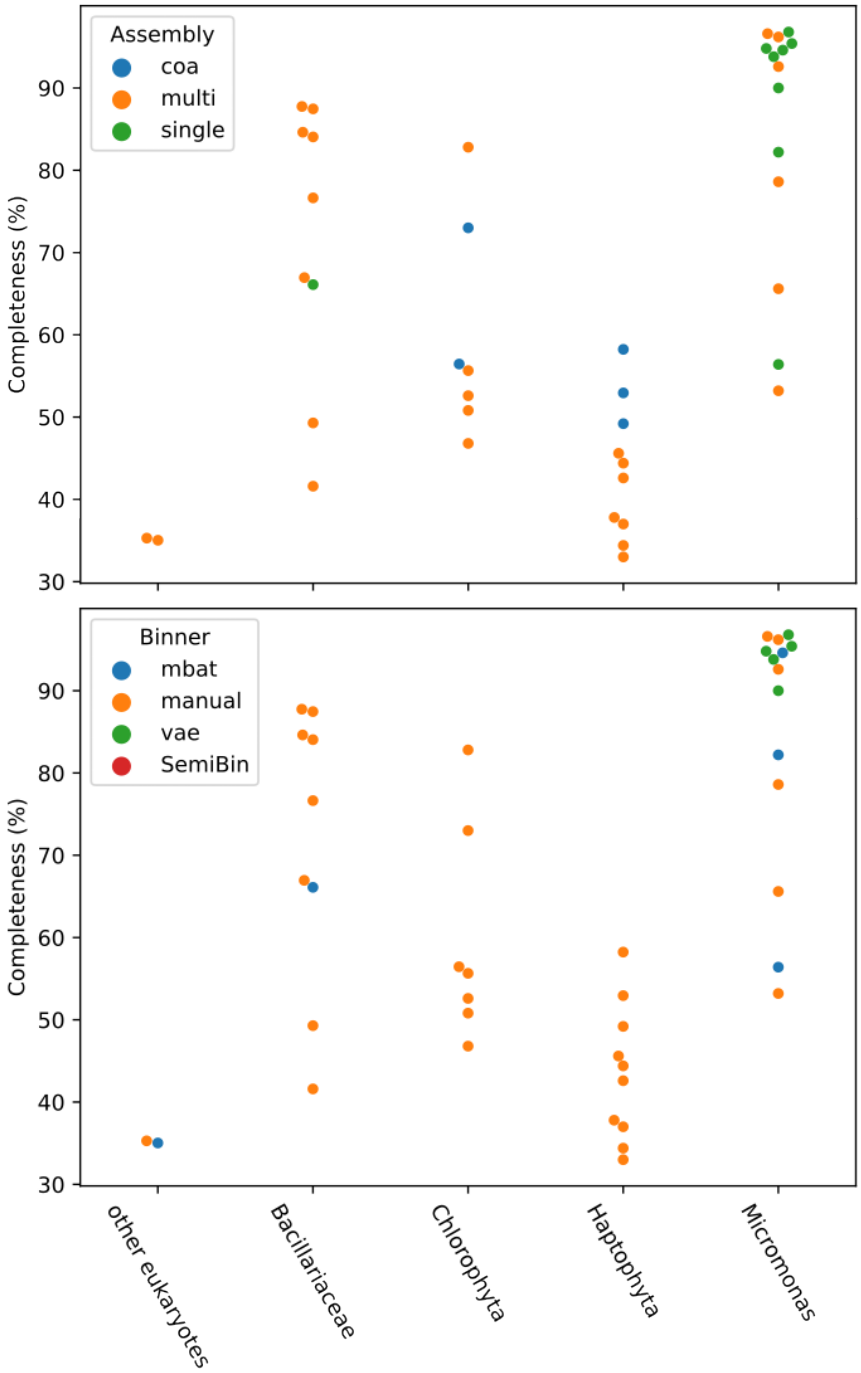


Figure 6.5: Completeness of all MAGs (after deduplication at 99% ANI) in the MOSAiC_Water_June dataset, using the same colours and formatting as Figure 6.3. For each 99% ANI cluster, most often the most complete MAG was a manually-refined bin produced through multi-binning.

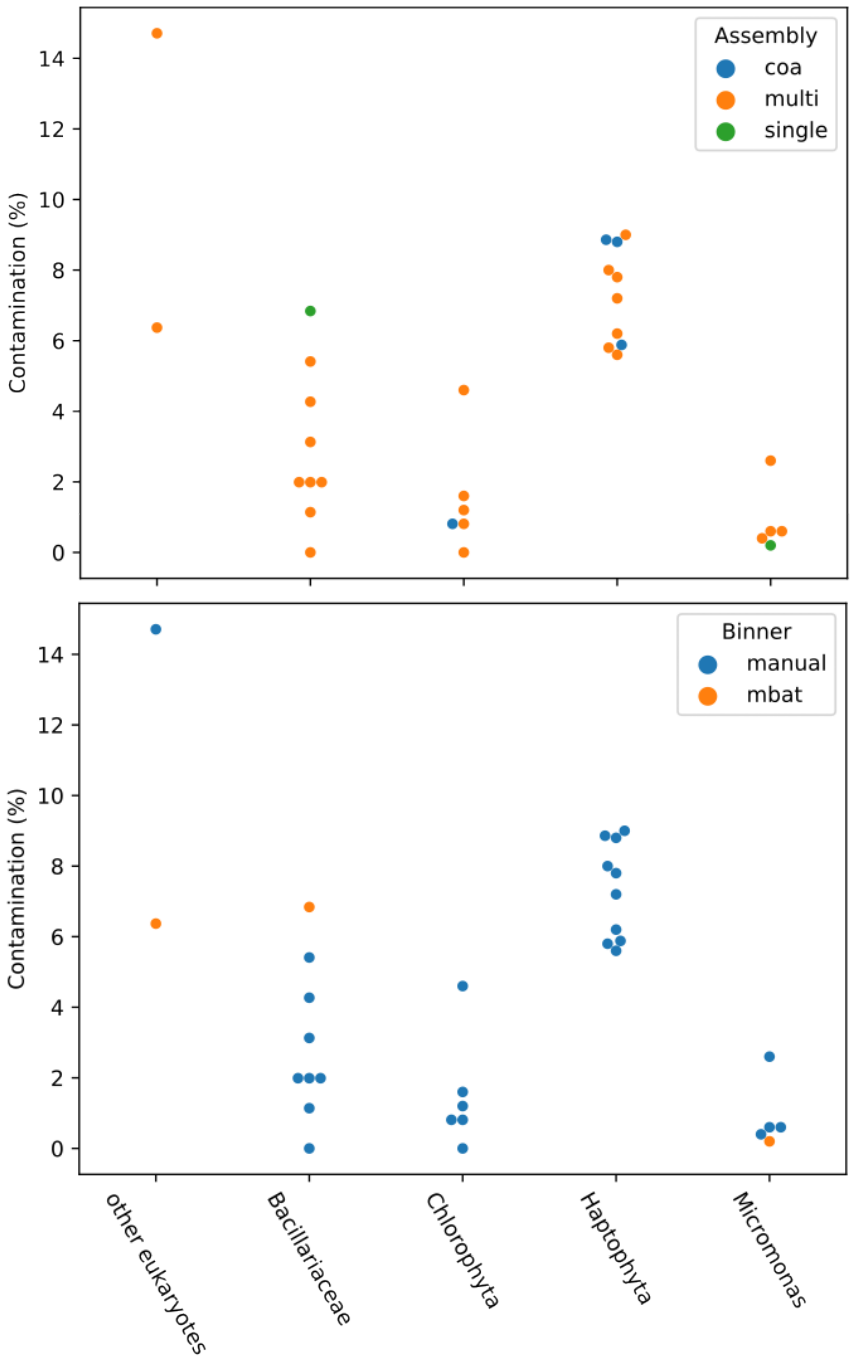


Figure 6.6: Contamination of all MAGs (after deduplication at 99% ANI) in the MO-SAiC_Water_June dataset, using the same colours and formatting as Figure 6.3. Haptophytes generally had a higher amount of contamination, perhaps linked to their larger genome size compared to the other clades. Micromonas MAGs had the lowest mean contamination.

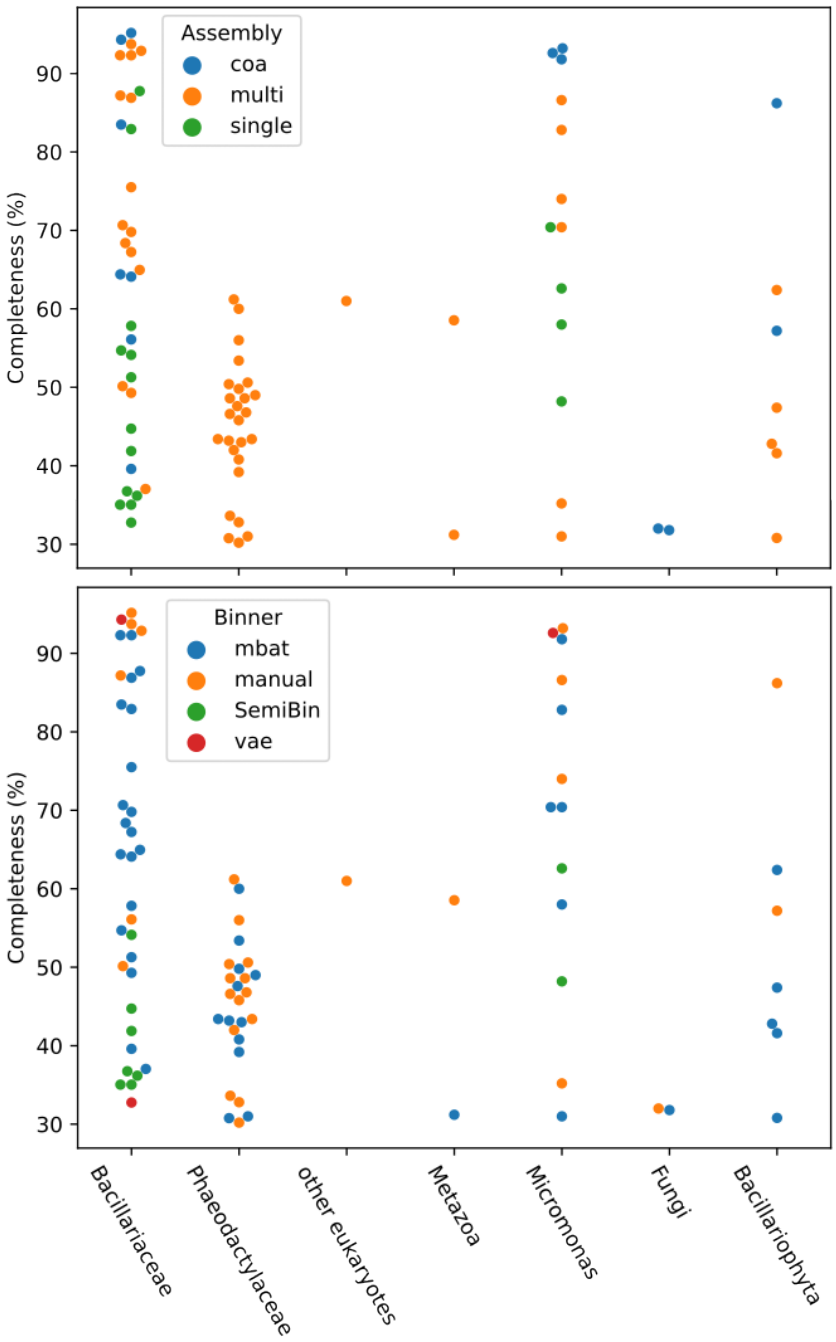


Figure 6.7: Completeness of all MAGs (before deduplication at 99% ANI) in the MOSAiC_Ice_April dataset, using the same colours and formatting as Figure 6.3. Bacillariophyta refers to all MAGs which could not identified more specifically as Bacillariaceae or Phaeodactylaceae. Multi-binning was the only method which produced any bins from the Phaeodactylaceae, while coassembly was the only method to produce Fungi bins. Single-sample binning (not using VALENCE) only produced *Micromonas* and Bacillariaceae MAGs.

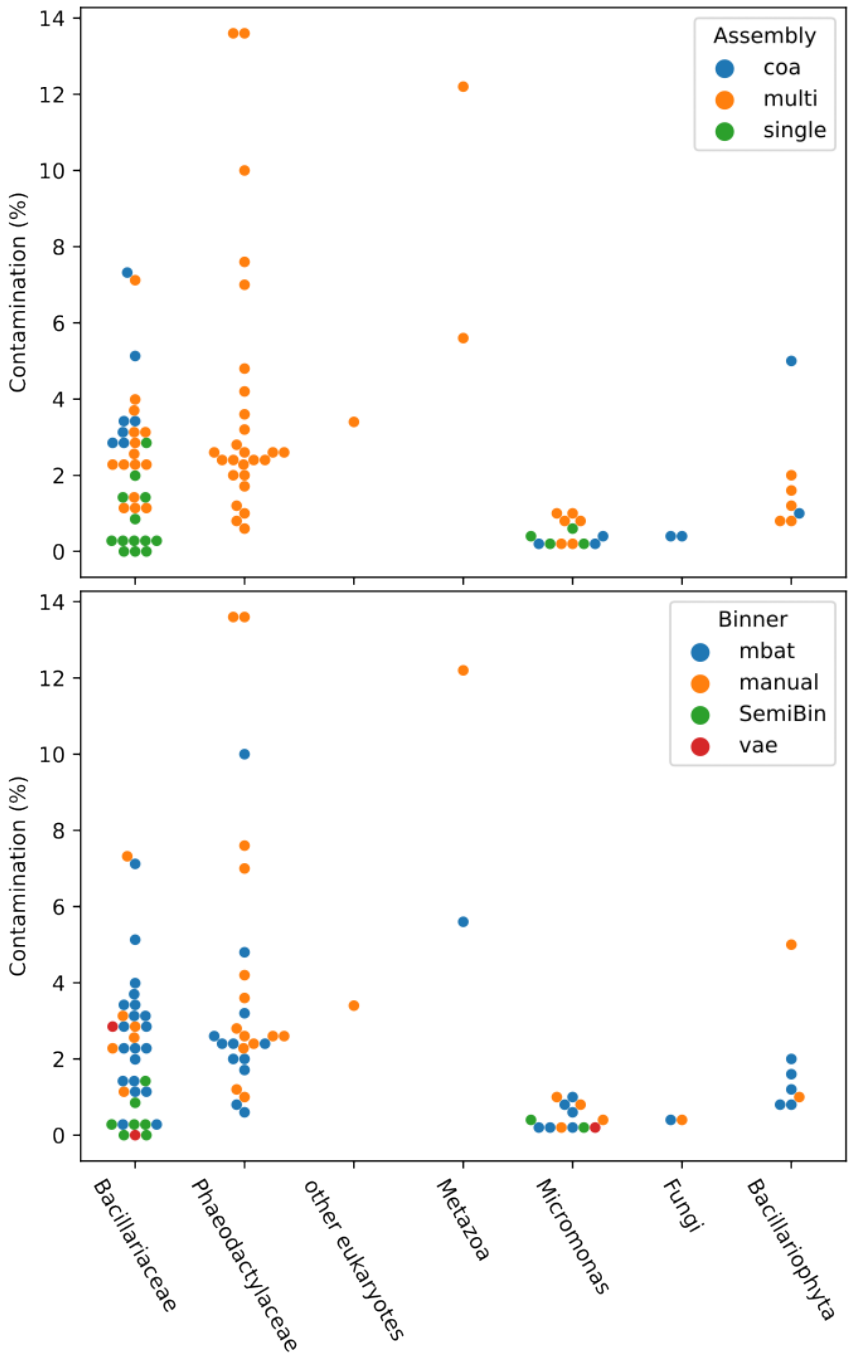


Figure 6.8: Contamination of all MAGs (before deduplication at 99% ANI) in the MOSAiC_Ice_April dataset, using the same colours and formatting as Figure 6.3. Once again, single-sample binning produced bins with the lowest mean contamination, though only within the *Micromonas* and Bacillariaceae, and these MAGs were of lower completeness.

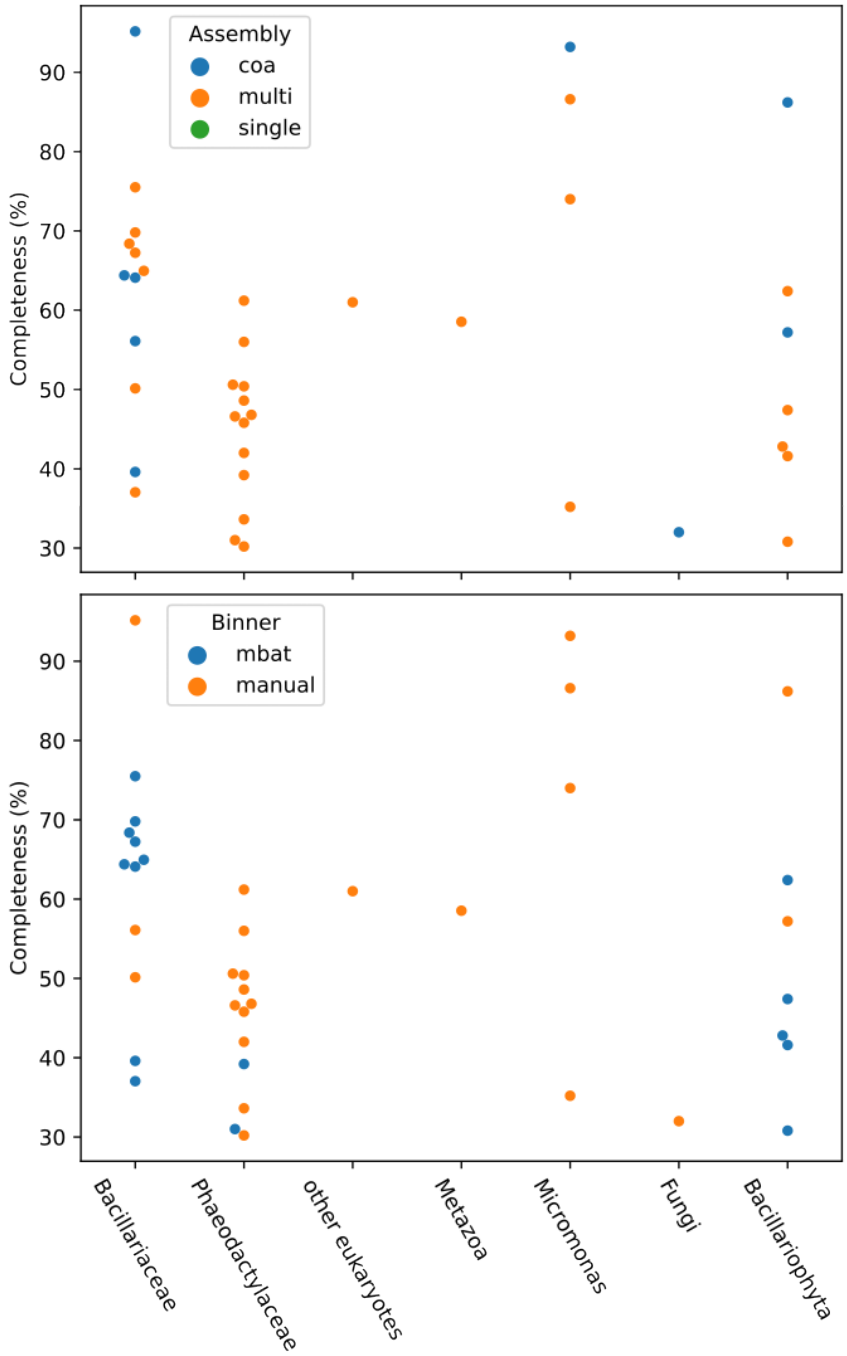


Figure 6.9: Completeness of all MAGs (after deduplication at 99% ANI) in the MOSAiC_Ice_April dataset, using the same colours and formatting as Figure 6.3. Once MAGs were deduplicated, the most complete MAG from each ANI cluster was retained - in all cases these were coassembled or multi-binned MAGs, either manually refined or binned using MetaBAT2.

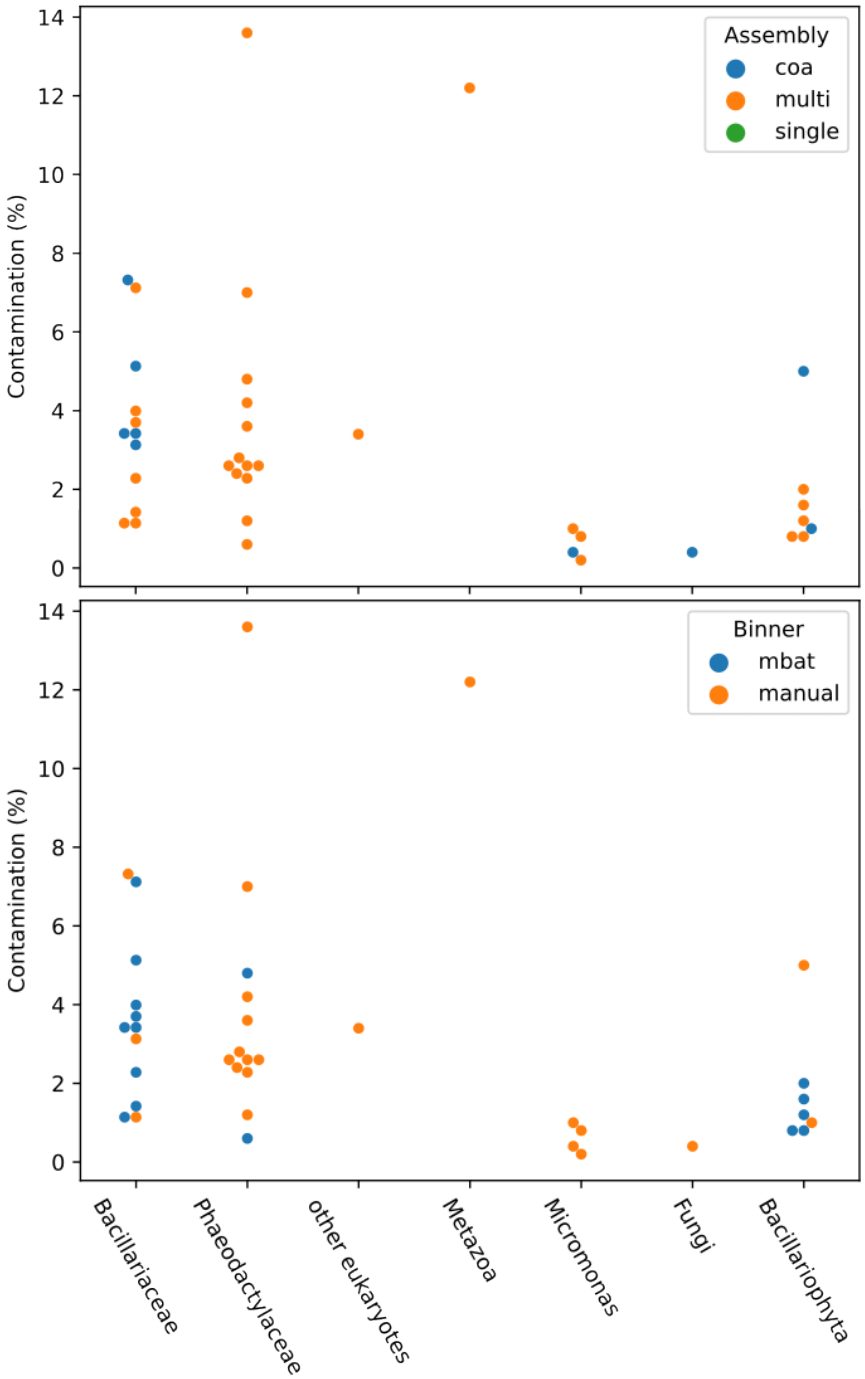


Figure 6.10: Contamination of all MAGs (after deduplication at 99% ANI) in the MOSAiC_Ice_April dataset, using the same colours and formatting as Figure 6.3.

6.3.2 Qualitative Comparisons of MOSAiC Data Visualisations

Figures 6.11 and 6.12 show comparisons of the UMAP plots generated by VALENCE alongside the visualisations provided by BinaRena, using t-SNE (the method used in the BinaRena publication [391], as well as the default option in VizBin [401]). Plots within BinaRena, where UMAP and PCA are applied directly to the k-mer compositions (using the scripts packaged with BinaRena), are provided in Appendix B.3. The boxes on the UMAP plots from VALENCE in Figures 6.11 and 6.12 highlight how manual bin refinement is performed; the user must select a rectangular area within the plot, which can then optionally be intersected with pre-existing bins, either from MetaBAT2, VAMB, or MCL clustering. Qualitatively, contigs of different taxa seem to group together using the VALENCE visualisation, whereas it was harder to identify as many groups within BinaRena, though with UMAP (Appendix B.3), there was a greater qualitative separation between contigs with different taxonomy. This may be because the latent-space method used in the VALENCE pipeline uses both contig coverage information and k-mer composition to generate a visualisation, whereas BinaRena only uses k-mer composition.

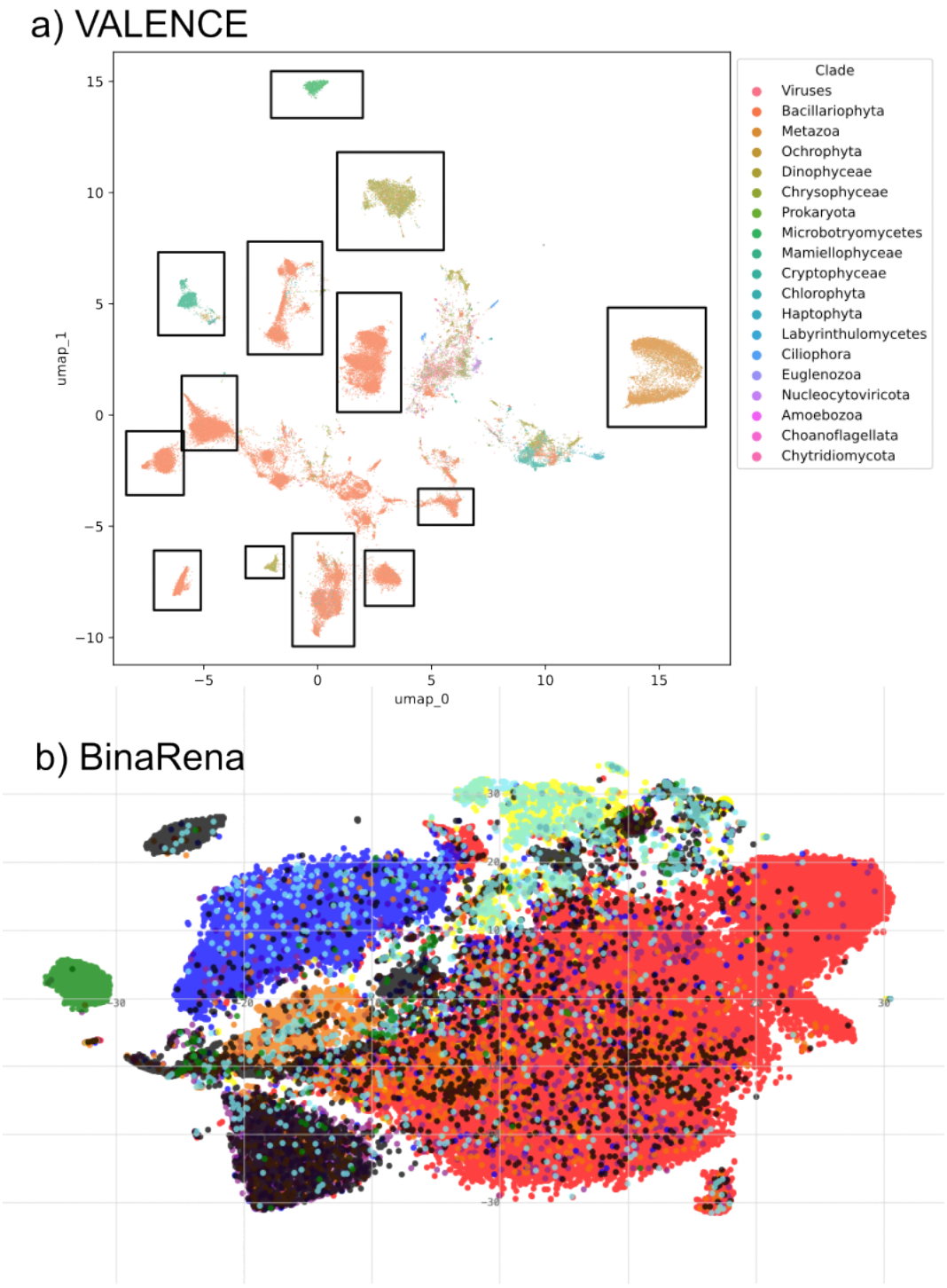


Figure 6.11: Visualisation from a) VALENCE and b) BinaRena, for the MOSAiC Ice April dataset. BinaRena does not provide a colour legend; contigs are coloured using the same taxonomic categories as in a), but with a different colour scheme. The boxes in a) exemplify how manual bin refinement is performed in VALENCE - each box describes contigs which are extracted into a separate bin.

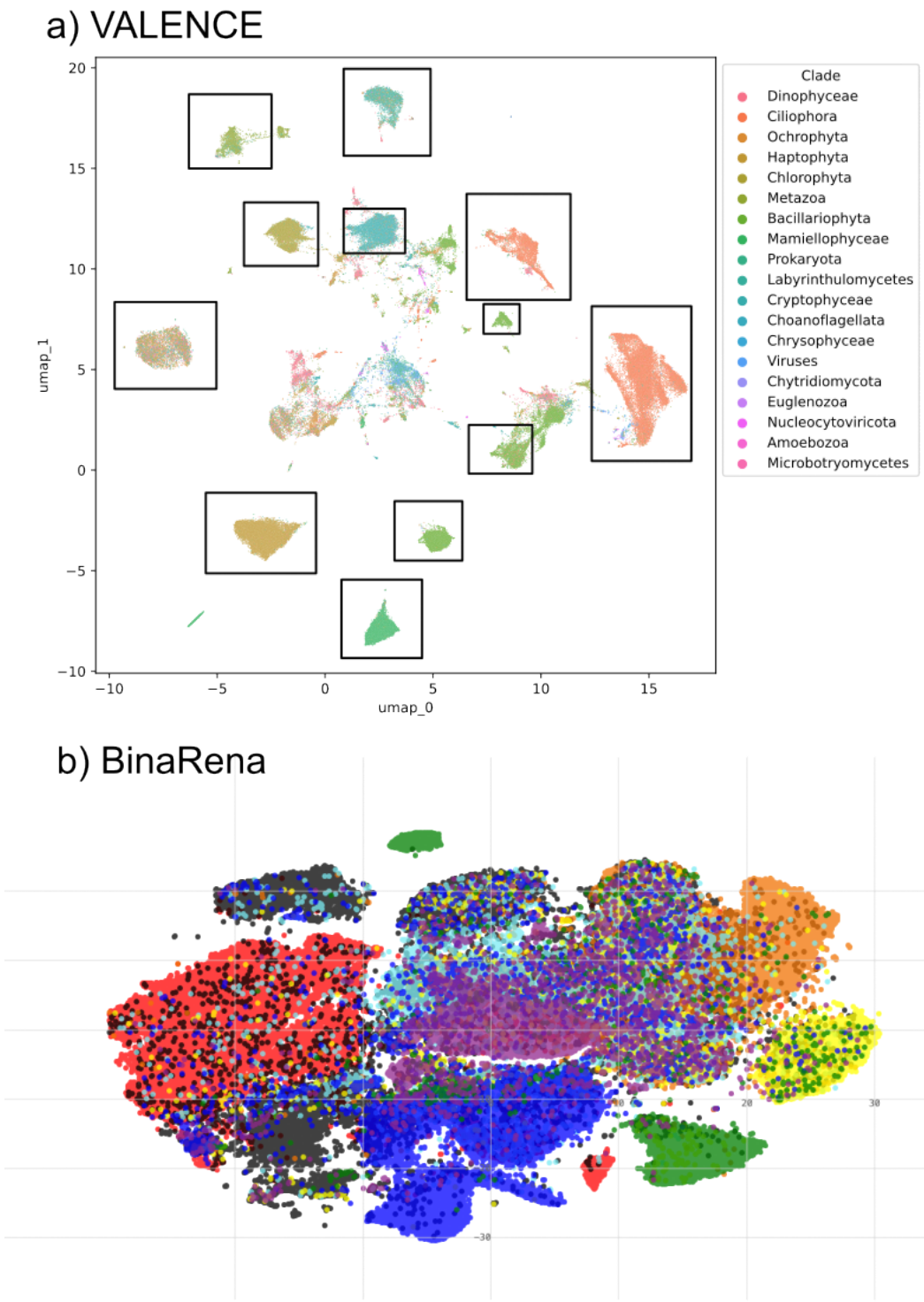


Figure 6.12: VALENCE a) and BinaRena b) visualisations for the MOSAiC Water June dataset.

6.3.3 Time and Memory Usage

We measured the time and memory usage of VALENCE on the MOSAiC_Water_June dataset, which was carried out on the university high-performance computing (HPC) system, using the commands seff and sacct packaged with SLURM (version 19.05.3-2). The most time- and memory-intensive single step was the metagenomic assembly using Megahit, which required 400 GB of RAM, and took 18 hours on 48 CPU threads, however this needed to be done just once for the whole coassembly. Read mappings with Strobealign were less memory-intensive but were performed much more frequently; there were 54 separate mappings required, though most were on the non-prokaryotic fraction of the reads. These required a peak memory of 250 GB, and ran on 8 CPU cores in 8 hours. The MMSeqs sequence searches also required 240 GB of memory each, and were performed once per sample. These stages were the bottlenecks in terms of computational throughput. Other hardware requirements were the MMSeqs database, requiring 38 GB of disc space, and the generation of intermediate bam files, requiring 144 GB of disc space.

6.4 Discussion

Eukaryotic MAG recovery presents several challenges beyond those encountered in prokaryotic binning. First, eukaryotic genomes are typically orders of magnitude larger than prokaryotic genomes (often 10 - 1000 times larger), meaning more reads are needed for adequate coverage. Second, eukaryotic genomes are highly repetitive: transposable elements, tandem repeats, and multi-copy gene families create ambiguous kmers in the de Bruijn graph, leading to fragmented assemblies. Third, intron-exon structure means that many coding sequences are interrupted, complicating gene prediction and functional annotation. Fourth, most binning algorithms were designed and benchmarked on prokaryotic data, and their scoring functions (often based on prokaryotic single-copy marker genes) do not directly translate to eukaryotes. Fifth, eukaryotic taxonomy is less well characterised in databases such as GTDB, and many marine eukaryotes lack close reference genomes, making validation of bin completeness and contamination difficult.

Coassembly and multi-binning improve the quality of eukaryotic MAGs generated, compared to single-sample binning, though at quite a significant extra cost in terms of time and computational burden. However, this can be important when searching for rarer taxa or larger genomes. In our case-study of MOSAiC metagenomes, coassembly and multi-binning were the only methods that were able to recover MAGs from the clades of Haptophyta, Fungi and Metazoa. It is notable that for the more abundant clades (e.g. *Micromonas*,

Bacillariophyceae) coassembly was not strictly superior to single-assembly for generating more complete MAGs. There are at least two possible reasons for this; firstly, for taxa found in high abundance across many samples, the sequencing depth from single-assembly may have been good enough, and provided lots of independent opportunities for generating a good MAG, and secondly, it is possible that with too much intra-species variation in a coassembly, the contigs generated are much shorter and so are difficult to bin (or below the length threshold that binning programs tolerate, usually two kilobases).

Since single-sample assembly, binning and annotation are routine and expected steps in any metagenomic analysis, pipelines that leverage this information without duplicating work will be useful for broadening the scope of MAG studies in an economical manner. This is a harder design problem because it requires handling a larger number of input configurations compared to simply processing raw read files one-by-one. However, software such as VEBA2, that are built in a modular way, are a useful step towards solving this challenge by allowing the user more flexibility when running the bioinformatics pipeline. Similarly, VALENCE provides versatility by allowing multiple configuration options such as using user-supplied taxonomic annotations, or only running various portions of the pipeline, at the user's discretion.

Beyond eukaryotic binning pipelines, the visualisation of metagenomic 'big data' (i.e. millions of contigs) is an important and often under-utilised tool available to bioinformaticians. Most tools tend to be tuned to optimise a handful of metrics; in the case of MAG binning, often completeness, contamination, and the number of medium or high quality MAGs. While these metrics are important, and cannot be replaced by a human eyeball (which has its own biases), interpreting UMAP ordinations generated through binning latent-spaces provides a secondary sense-check. This can be important when using bioinformatics tools slightly outside their intended scope, e.g. when using binning algorithms that have been validated and optimised for recovering prokaryotic MAGs to eukaryotic datasets. They may also have a place in hypothesis generation. For example, in a different MOSAiC dataset (MOSAiC_Water_April, Figure B.4 in Appendix B.3), we found the contigs of a diatom MAG, clustered closely together with contigs from a dinoflagellate, suggestion they may have had a correlated abundance profile across samples. On closer inspection, the closest taxonomic hits of the dinoflagellate were to *Kryptoperidinium foliaceum*, a 'dinotom', i.e. a dinoflagellate species with an endosymbiotic relationship with a diatom [402]. We leave an investigation of this as further work; the identification of symbioses in metagenomes is an area of active research, with new methods such as Symclatron [403], a symbiont classifier, in development.

More speculatively, data visualisation may also be beneficial for public engagement, since identifying bins within a UMAP plot can be done without any prior technical bioinformatics

or biological experience. This is particularly true for the tool BinaRena, which requires only a web-browser to operate. While in our analysis the default visualisations created by BinaRena were qualitatively not always particularly illuminating, there is no reason in principle why the program could not accept the ordinations produced by VALENCE. One simple improvement to the VALENCE pipeline could then be to generate a file correctly formatted for BinaRena, which has a very user-friendly interface and could then be displayed at science fairs, where members of the public can try their hand at binning MAGs.

Eukaryotic MAG recovery does not yet have quite as mature an ecosystem of bioinformatics tools as for prokaryotic MAGs; as this field continues to develop we aspire to update the VALENCE pipeline with the latest methods. Improvements in long-read sequencing technology, and latent-space binning algorithms (e.g. COMEBin [200]) mean that recovering eukaryotic MAGs from metagenomes is becoming more achievable, and may soon become a standard part of any metagenomic pipeline, especially when applied to marine datasets that may contain an abundance of microeukaryotes. Future developers should keep non-prokaryotic genomes in mind when creating tools to analyse metagenomic data in a more holistic manner.

6.5 Data Availability

The VALENCE pipeline is available for download and installation from Github at https://github.com/willboulton/valence.

Chapter 7

Network Analysis of MAG Diversity

7.1 Background and Summary

The previous chapters have focussed on smaller-scale analyses of genes and species within MAGs from the MOSAiC expedition. This chapter will explore a longer time-series of samples, covering a time period from November 2019 to September 2020, and geographically ranging from the Central Arctic Ocean (CAO), drifting south roughly parallel with the Gakkel ridge, to the Fram Strait, between Greenland and Svalbard. We will explore prokaryotic and eukaryotic diversity through a MAG analysis, and identify networks of correlated genes and species, relating these to environmental variables, via a Weighted Gene Correlation Network Analysis (WGCNA) [404].

The MAG analysis comprises 4 steps, which are fully described in Section 7.2.2:

1. **Binning and Annotation**: Generate a catalogue of MAGs, with taxonomic and functional annotations for each MAG.

2. **Abundance Estimation**: Map reads from each sample to the MAG catalogue, generating an abundance table, normalised as reads per million (RPM).

3. **Diversity Analysis**: Identify correlations between environmental variables and MAG abundances, using PCoA plots.

4. **WGCNA**: Apply WGCNA to identify functional modules and gene correlation networks that are associated with time, geography, and changing environmental parameters.

7.2 Methods

7.2.1 Sample Description

This chapter will synthesise data from two sets of metagenomes from MOSAiC:

1. The 73 metagenomes analysed in Chapter 4.
2. A further 241 metagenomes, with sample collection dates spanning from November 2019 to September 2020, from the main MOSAiC time-series.

Figure 7.1 outlines the locations, dates, and environments (ice, water, or sinking particles from sediment traps) from where the metagenomes were collected by the MOSAiC team, and particularly team ECO, lead by Allison Fong [1], [99]. Of the 314 total samples, 186 are from sea ice, 111 from seawater, and a further 17 are from sinking particles caught in sediment traps (sediment trap samples from the HAVOC subproject, previously described). Table 7.1 provides a summary of the physical parameters for various sets of metagenome samples. Note that some samples, particularly the sinking particles, are missing some of these physical parameters, such as nutrient concentrations.

To identify each metagenomic sample with physical parameter values, we cross-referenced each metagenome with other datasets that measured CTD and sea-ice parameters listed in Table 7.1. Each sample taken by MOSAiC was assigned an event ID, based on either its CTD cast number (sampling water) or ice coring effort (sampling ice). Samples were further differentiated by their depth layer, typically from 0 to 200 m for the water column (with a focus on the surface 50 m), or 10 cm depth layer increments within ice cores. Metagenomic samples were matched with samples measuring physical parameters based on their event IDs and depth layer, using the closest available depth layer. Data from these measurements were downloaded from open-access datasets [405]–[409]. Typically, the combination of event ID and depth layer uniquely defined a single sample for each property (temperature, salinity, oxygen concentration, nutrient concentrations, chlorophyll-a). Where there were multiple samples per property that matched with a metagenomic sample, a parameter value was generated by taking the mean average. Where there was no sampling matching with a metagenome, the parameter was estimated by taking an average of samples from the same day, or if none were available, from the same week.

7.2.2 DNA Extraction, Library Preparation, and Bioinformatics

Metagenome sample collection and library preparation followed the same protocol as described in Chapter 4 for the 15 pilot samples. We then developed and applied a more

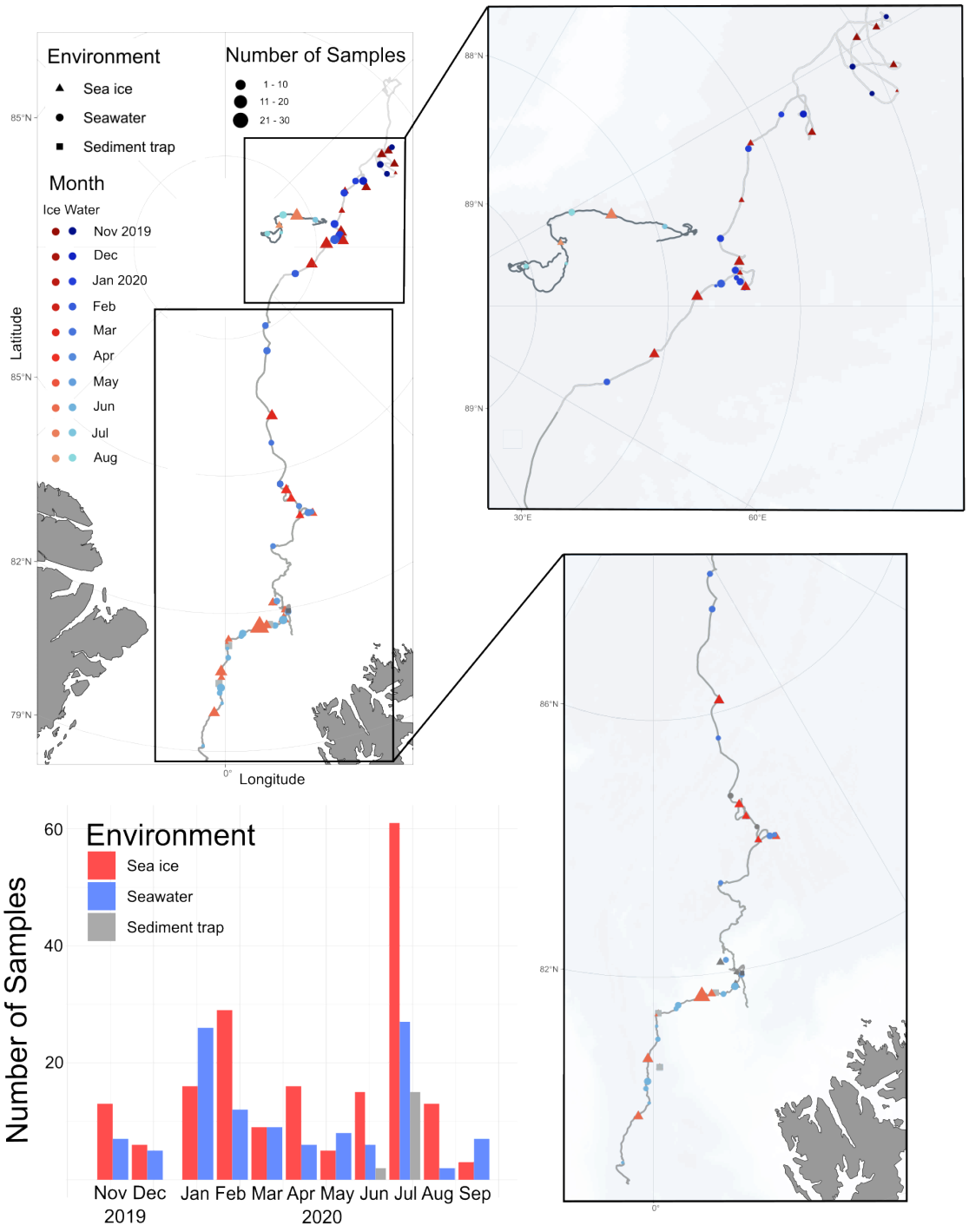


Figure 7.1: Geographic and temporal distribution of the 314 metagenomic samples. The shading indicates time progress along the drift, blue circles represent seawater, red triangles sea ice, grey squares are the sediment trap samples. The excess of ice samples collected in July are in part due to HAVOC, which collected a large number of ridge ice samples in this month.

Physical parameter	Ice	Seawater	Sediment traps
sample collection site	available for all samples		
collection date	available for all samples		
latitude	available for all samples		
longitude	available for all samples		
depth	available for all samples		
salinity	available for all samples		
temperature	available for all samples		
dissolved oxygen	missing	available	available
total concentration of C	missing	available	available
total potential of C	missing	available	available
attenuation	missing	available	available
phosphate	available	available	missing
nitrate	available	available	missing
nitrite	available	available	missing
silicate	available	available	missing
ammonium	available	available	missing
chlorophyll-a	missing	available	missing
first / multi-year ice	available	NA	NA
ice depth	available	NA	NA
snow depth	available	NA	NA
volume of brine	available	NA	NA
volume of gas	available	NA	NA
δ deuterium	available	NA	NA
δ 180 H_20	available	NA	NA

Table 7.1: Availability of various environmental measurements across different environments.

complex bioinformatics pipeline, which we now describe in more detail.

Bioinformatics Overview

The aim of the bioinformatics pipeline was to generate as comprehensive a set of MAGs as possible, within the constraints of the computing resources available on the university cluster. We did not reprocess the 73 samples described in Chapter 4, instead we used the MAGs described in that section.

The remaining 241 new samples were processed according to a custom pipeline, based on the work from Chapter 6. First, each sample was processed individually by the IMG/M pipeline, version 5.1, to generate prokaryotic MAGs. We applied additional binning, and annotation, described in Section 7.2.2, to improve the quality of the prokaryotic MAGs produced, and so that the Pfam annotations of all MAGs were consistent. Next, we split

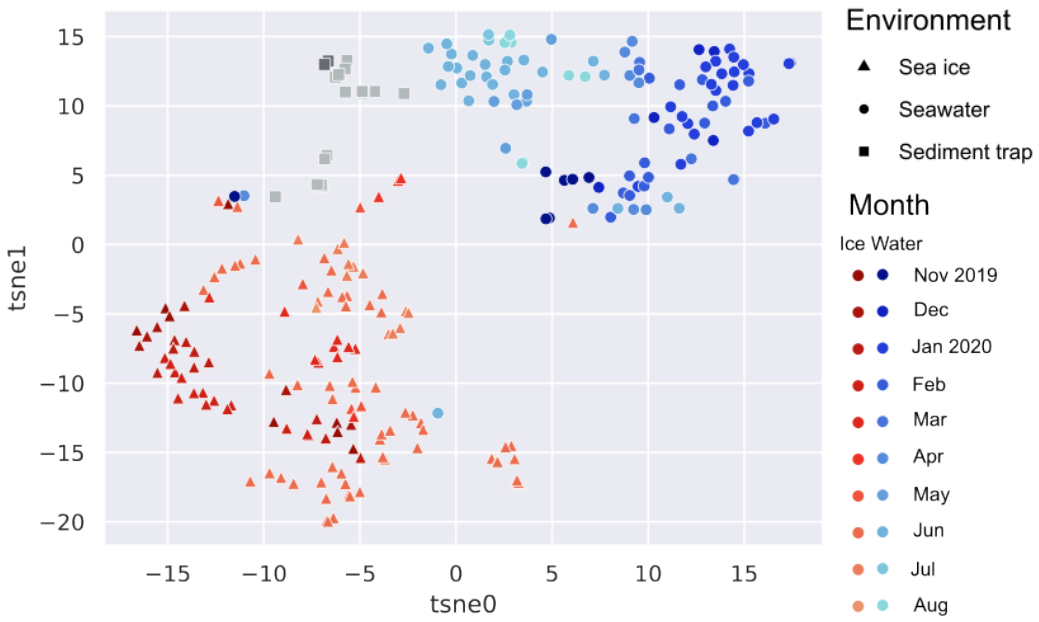


Figure 7.2: t-SNE ordination of metagenomic samples, based on their Pfam richness (counts). Points are metagenomic samples, coloured by environment and collection month, with functionally similar samples clustered closer together. Samples were aggregated into batches based on their environment and collection month, these batches were then co-assembled. Based on this ordination, aggregating samples based on environment and month seemed to group functionally similar samples (based on Pfam richness) relatively well.

the samples into several batches, and for each batch we applied the coassembly and binning pipeline described in Chapter 6. Since we wanted to retain strain diversity within eukaryotic MAGs, we chose not to generate a small number of large coassemblies, but instead generate a larger number of smaller coassemblies, with samples stratified by time and space. Samples were grouped into batches of those from the same environment (ice or water) and into date ranges, typically within the same month. Batches had a size of up to 16 samples; larger batches required too much time and memory during assembly. A t-SNE visualisation, Figure 7.2, based on the Pfam abundances within each sample, was used to visually inspect whether or not samples within a batch were clustered together; our aim was to coassembly similar samples (i.e. group them into the same batch), and similarity at the level of Pfams was a simple proxy to determine that our groupings were indeed sensible. In total, we applied the pipeline from Chapter 6 to 33 batches of samples, with an average of 7.3 samples in each batch. We then annotated any resulting eukaryotic MAGs, retaining those with over 30% completeness and below 15% contamination.

An overview of the pipeline is described in Figure 7.3, which is similar to Figure 4.2

in Chapter 4, but includes annotation stages, and the more complex binning process. The stages from the IMG/M pipeline, and the pipeline from Chapter 6, are highlighted. The remaining stages are described in further detail below.

Initial Assembly and Annotation with IMG/M MAP

All new samples were processed using the IMG/M MAP (version 5.1). This pipeline is extremely similar to the one described in Chapter 4, though [410] presents a change-log of differences between both this version, and the version used the annotate the pilot set of 15 metagenomes. Table 7.2.2 summarises the changes in software and database versions used to process the pilot samples, HAVOC samples, and the 241 new metagenomes.

Although most changes in software versions correspond to minor updates, changes in the database versions had an impact on some preliminary analysis. In particular, Pfam databases version 30.0 and 34.0 were incomparable, with version 34.0 including an additional 1980 protein families. Since our analyses relied heavily on Pfam annotations, we re-annotated the 15 pilot metagenomes that used Pfam database 30.0 to include Pfams from database version 34.0.

To generate a more complete set of prokaryotic MAGs, we aggregated the results of 3 binning algorithms (MetaBAT2, VAMB, SemiBin2, versions (2.2.15), (4.1.3) and (2.1.0), respectively), as well as the bins generated by the MAP pipeline, using DASTool (version 1.1.6) to generate a single set of MAGs for each single-assembly, with no shared sets of contigs. Completeness and contamination of the prokaryotic bins was assessed with CheckM2 (version 1.0.2).

Cross-Domain MAG Catalogue and Abundance Estimation

Contigs in the non-prokaryotic fraction of each batch were processed using the pipeline described in Chapter 6 to generate eukaryotic MAGs. We used the default settings of the eukaryotic pipeline; MCL clustering was run, using an inflation parameter of 1.1, to generate bins from the UMAP graph, and further bins were generated by visually inspecting the UMAP scatterplots of each batch.

Additionally, we ran genomad (version 1.8.0) with default parameters (version 4.1.0) to identify viral contigs. We used the abundance estimates across all samples to bin the viral contigs, using both MetaBAT2 and VAMB. We retained only the 5000 largest bins, to avoid our MAG catalogue becoming swamped with an extremely large number of single-contig viruses.

	Pilot Sample Single Assembly	Pilot Sample Non-Prok. Coassembly	HAVOC Sample Single Assembly	HAVOC Sample Non-Prok. Coassembly	These Data Single Assembly	These Data Non-Prok. Coassembly
Chapter	Chapter 4	Chapter 4	Chapter 4	Chapter 4	Chapter 7	Chapter 7
IMG/M Pipeline Version	5.0	NA	5.1	NA	5.1	NA
Read QC	BBDuk (v. 38)	NA	BBDuk (v. 38)	NA	BBDuk (v. 38)	NA
Assembly	metaSPAdes (3.14.1)	metaSPAdes (3.14.0)	metaSPAdes (3.15.2)	Metahipmer (2.1.0)	metaSPAdes ()	MEGAHIT (1.2.9)
Contig Taxonomy	NA	MMSeqs2 (v. 01889*)	NA	MMSeqs2 (v. 01889*)	NA	Consensus method[3]
Gene Calling	Prodigal (2.6.3), Genemark (1.05)	MetaEuk (f32e8*), Genemark-ES (4.71)[2]	Prodigal (2.6.3), Genemark (1.05)	MetaEuk (f32e8*), Genemark-ES (4.71)[2]	Prodigal (2.6.3), Genemark (1.25)	MetaEuk (f32e8*), BRAKER3
Annotation	HMMER (3.1b2)	InterProScan (5.63)	HMMER (3.1b2)	InterProScan (5.63)	HMMER (3.3)	HMMER (3.3)
Pfam Database	Pfam-A (v. 30)[2]	Pfam-A (v. 34)	Pfam-A (v. 34)	Pfam-A (v. 34)	Pfam-A (v. 34)	Pfam-A (v. 34)
Binning	MetaBAT2 (2.12.1)	MetaBAT2 (2.15)	MetaBAT2 (2.15)	MetaBAT2 (2.15)	Consensus method[3]	Consensus method[3]
Binning QC	CheckM (1.0.12)[2]	EukCC (2.1.1), database 1.1[2]	CheckM (1.1)[2]	EukCC (2.1.1), database 1.1[2]	CheckM2 (1.0.2)	EukCC (2.1.1), database 2.0
MAG Taxonomy	GTDB-Tk (0.2.2), database release 95[2]	Tree, Contigs[1]	GTDB-Tk (2.4.0), database release 220	Tree, Contigs[1]	GTDB-Tk (2.4.0), database release 220	Tree, Contigs[1]

Table 7.2: Summaries of the pipeline versions, database versions, and methods used, for each set of samples. [1]Taxonomy estimated based on position in a eukaryotic tree, and on most common taxon of contigs. [2]The updated version of this tool was re-run for work in this chapter. [3]The consensus methods are described later in this section.

The resulting catalogue of MAGs consisted of all prokaryotic MAGs (generated by the IMG/M pipeline for all samples, and by the multiple binning programs described above for the 241 new samples, and with redundancy removed using DASTool), and all eukaryotic MAGs, both from previous chapters, and this chapter (once again with redundancy removed, i.e. no bins with overlapping sets of contigs).

To estimate the abundance of these MAGs, we generated a single sequence database from this catalogue of MAGs and mapped reads from each metagenome to the database using Strobealign (0.13.0), using --aemb mode. We multiplied these values by the read lengths to produce a table of bases mapped per contig per sample, which we converted into reads per million (RPM). Additionally, we mapped a set of 135 metatranscriptomes, collected from the surface ocean (top 200 m pelagic layer) throughout the course of the drift, to the same catalogue, again using Strobealign. Although mRNA expression was not the focus of this study, these data might be useful for further work.

Gene Identification and Functional Annotation

Genes and functional annotations provided by the MAP were used for prokaryotic and viral MAGs; this was the same as in Chapter 4, with updated software versions listed in Table 7.2.2.

Eukaryotic annotations were performed using MetaEuk (version f32e8*) using the --easy-annotate option, and the BRAKER3 pipeline, in --ES mode (i.e. using Genemark-ES to generate a set of genes, and then Augustus to train from these predictions). We used a much shorter --min_contig_length of 500 compared to the default of 20000, since these metagenomes were often highly fragmented. Just as for the prokaryotes, we identified tRNAs and rRNAs with tRNAScan-SE (version 2.0.12) and barrnap (version 0.9.0) respectively, with barrnap in euk mode. Pfam annotations were added using the script Pfam_scan.py, which relies on HMMer (version 3.3.0) using hmmsearch, and e-value cut-offs supplied by the Pfam team.

Annotation results were combined using the AGAT package (version 1.0.0) with the script agat_sp_combine_results.pl, with rRNAs taking first priority, then tRNAs, then BRAKER3 predictions, then Metaeuk. To identify GO terms, we used the Pfam2Go mapping [411], maintained by the Interpro team.

Taxonomy Estimation and Phylogenetic Placement

Eukaryotic and prokaryotic MAGs were deduplicated at the 99% average nucleotide identity (ANI) level using dRep (version 2.0.1), this generated 2867 clusters from 9987 prokaryotic MAGs, and 239 clusters from 354 eukaryotic MAGs. Prokaryotic taxonomy was estimated

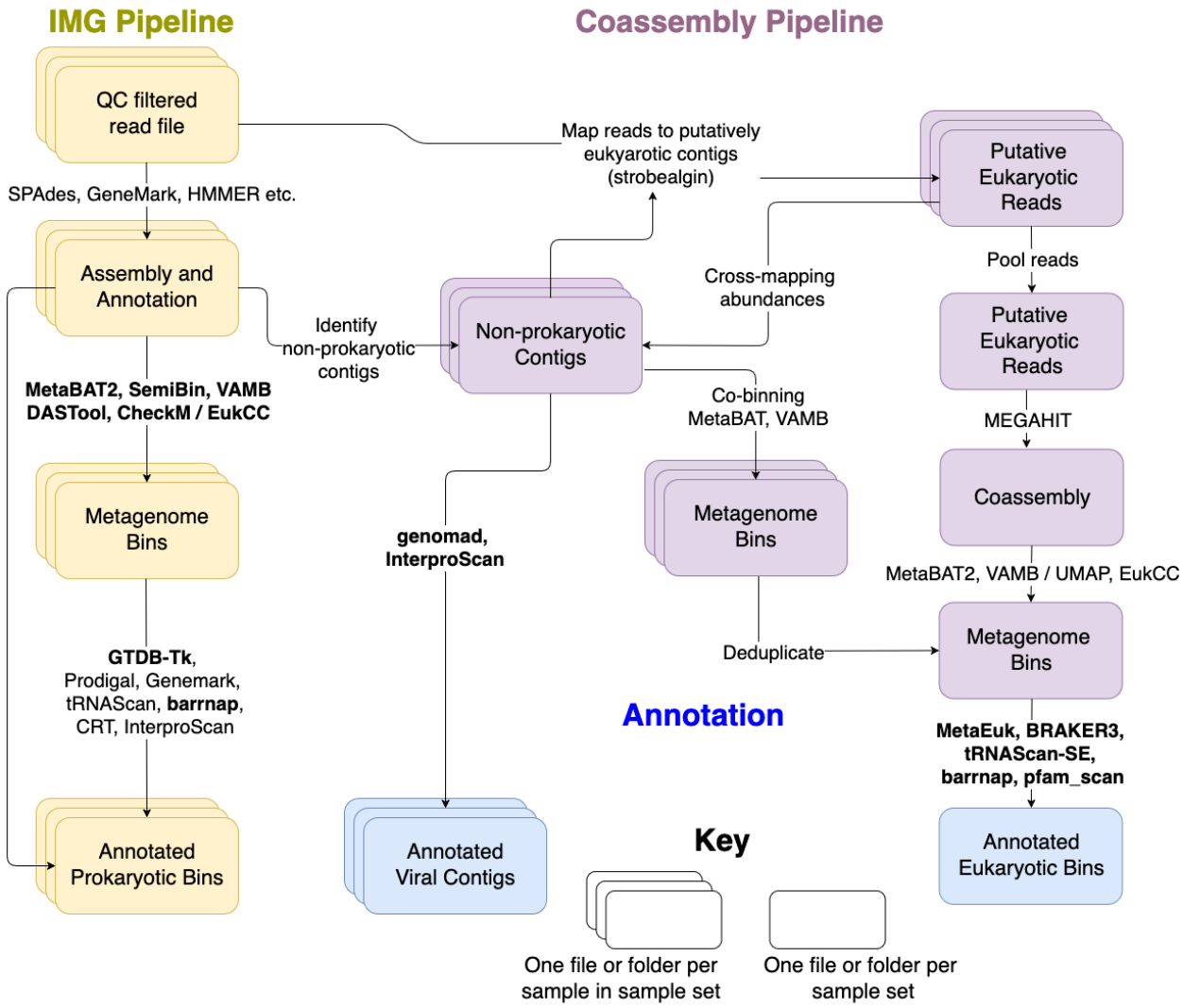


Figure 7.3: Assembly, binning and annotation pipeline for each batch of the 241 newly processed samples. In yellow, the steps of the MAP. These annotations are used for prokaryotic bins. In purple, multi-binning and coassembly steps are used to generate eukaryotic MAGs (Chapter 6). Viral contigs are extracted with genomad. Eukaryotic contigs are re-annotated using MetaEuk and the BRAKER3 pipeline, in --ES mode. The blue boxes, and the programs highlighted in bold, were run in addition to the IMG/M and Coassembly pipelines, and are described in this chapter, Section 7.2.2.

with GTDB-Tk (version 2.4.0, database 220). For the eukaryotes, we developed a custom script to aggregate the taxonomies obtained from the MMSeqs2 LCA method when classifying non-prokaryotic contigs.

Species trees of prokaryotes were generated through GTDB-Tk classify workflow, which uses pplacer to place species on a GTDB reference tree, using a selection of 53 archaeal or 120 bacterial marker genes [412], aligned to their respective HMM models, concatenated, and trimmed to approximately 5000 amino acids. GTDB-Tk uses a predefined mask to trim a large concatenated alignment of approximately 40,000 amino acids down to 5000, as generating a tree from this is more computationally tractable, and furthermore the sites chosen are determined to be well-aligned (few gaps) but also phylogenetically informative.

For eukaryotic MAGs, we followed a similar methodology to Duncan *et al.* [104], also used in Chapter 4. We selected a set of 298 reference genomes from NCBI RefSeq and JGI Phycocosm and Mycocosm; accessions listed in Appendix C.2. To generate a concatenated alignment of marker genes, we used BUSCO (version 5.4.3) with the otdb_eukaryota_10 gene set, and with a threshold of 30% for the number of marker genes present across 90% of MAGs and reference genomes. This identified a set of 111 marker genes which we individually aligned using MUSCLE (version 2.1.1), and concatenated the amino acid alignments. We then used trimAl (version 1.2) to remove low quality sites in the alignment; this reduced the alignment from approximately 35,000 sites to 11,112. To generate the tree, we used FastTree (version 2.1.11) to get a maximum likelihood tree using a Jones-Taylor-Thorton substitution model, with 20 rate categories. We rooted the tree using the Bacillariophyta as an outgroup. Trees were visualised with the Interactive Tree of Life [367].

Species Network Analysis

A species correlation network was generated using the SparCC package (version 1.0.0) [226], this was used as input to generate network modules, using the python package SCNIC (version 0.6.6) [237]. This generated modules by using a shared minimum distance method; MAGs were clustered applying complete linkage hierarchical clustering to correlation coefficients, and modules generated by identifying fully-connected clusters of MAGs using a threshold correlation value of 0.35). To visualise the species connectivity graph, we used a force-directed graph layout method from the iGraph package (version 0.9.9).

Sample α and β diversity were measured using the Shannon index, and Aitchison distances, respectively. Aitchison distance is simply Euclidean distance between clr-transformed data; in this case RPM abundances. PCoA plots were generated with the python package scikit-bio (version 0.5.6).

Gene Network Analysis

We applied a WGCNA analysis to the Pfam abundances within MAGs, using the mean coverage of each contig as a proxy for the coverage of Pfams within that contig, accounting for multiplicity. These data were clr-transformed before being input into the WGCNA analysis. We used a signed network with mid-weight bi-correlation to generate a linkage between species, raised to a power of 15. To identify GO term enrichment within WGCNA modules we ran a GO enrichment analyses, using a binomial test (scikit-learn, version 1.2.0) to test for the likelihood that a GO term was disproportionately prevalent within a particular WGCNA module.

7.3 Results I: Overview and Quality Control

7.3.1 Summary of Sample Physical Parameters

Of the 314 metagenomes, 186 were from sea ice, 111 from seawater, and 17 from sediment caught in sediment traps beneath ice ridges. Samples were collected during each month of the cruise between November 2019 and September 2020, with a mean of 16.9 sea ice samples collected per month (standard deviation 16.2), and 10.5 (s.d. 8.3) and 1.5 (s.d. 4.5) collected from seawater and sediment traps respectively. Sediment trap samples were only collected during the months of June and July. Sea ice and seawater samples were collected during each month, with a maximum of 88 samples in July (61 sea ice, 27 seawater), and a minimum of 11 in December (6 sea ice, 5 seawater).

Seawater samples from CTD casts were typically from the epipelagic ocean layer. Of the 111 samples, 43 were from the upper 10 m ocean layer, a further 45 from between 10 and 100 m, and 11 from between 100 and 200 m. 10 seawater samples were from voids in ice ridges, part of the HAVOC project (see Chapter 4). The remaining 2 samples were from depths of 202 and 4082 meters.

Sea ice samples were from both first-year ice (78 samples) and multi-year ice (82 samples), with the remaining 6 from ice of an unknown age. There were 64 samples were from the bottom 5 cm layer of ice core sections, 41 were from the top 10 cm layer of a core section, and 78 from a middle layer (neither the bottom 5 cm of the ice core, at the sea-ice interface, nor the top 10 cm of an ice core). Three samples were from a miscellaneous depth layer - either a refrozen layer (2 samples) or rafted ice (1 sample). Twenty samples were ridge ice, part of the HAVOC project, as described in Chapter 4.

The sediment trap samples have already been described in Chapter 4; these collected sinking particles from beneath ice ridges, and resided at a depths of 5 m (5 samples), 15 m

(6 samples) and 50 m (6 samples).

Figure 7.4a shows a PCA plot of the physical parameters associated with the ice and water samples, with directions corresponding to each of these parameters superimposed. Sediment trap samples were excluded, since these were missing most nutrient concentration data relevant for the PCA. The single bathypelagic ocean sample was also excluded as an outlier. Parameters that were not common to both ice and water samples (e.g. chlorophyll concentration, density) were also excluded. The first principal component, explaining 35% of the total variance in the data, almost completely separates the ice and water samples. The second principal component is aligned with the direction of changing time, and geography (since the course of the drift was generally from north to south, and in a direction of decreasing latitude, these variables were all correlated).

Figure 7.4b,c also show correlation plots of the various nutrient concentrations (phosphate, silicate, nitrate, nitrite, and ammonium), and of temperature versus salinity. Ammonium concentration was negatively correlated with all 6 other parameters (a Pearson correlation r value between -0.49 and -0.10). All other pairs were positively correlated, except nitrate and nitrite, which had an r value of -0.10. Phosphate, silicate, nitrate, and salinity were more strongly positively correlated, with the lowest r between any pair of 0.49 between nitrate and silicate, and the highest of 0.88 between phosphate and salinity. All other pairs are weakly positively correlated, with r between 0.14 and 0.38. A table of all correlations is provided in Appendix C.4.

Figure 7.5 show time series of temperature and salinity, both for ice and water, as well as sea ice thickness, both in first-year ice and multi-year ice. Salinity and temperature are important variables in physical oceanography since they dictate ocean stratification and mixing layers (see Section 2.1.1). Both are also important biologically; temperature influences metabolic rates, while salinity influences osmotic effects and is an important proxy for overall nutrient concentrations, particularly phosphate and nitrate (Table C.4 shows correlations between salinity and other nutrient concentrations).

We concentrated only on the upper 100 m ocean layer, stratifying the seawater samples into two groups of surface ocean (less than 25 m deep) and a deeper 25 - 100 m layer. For the ice cores, we used temperature and salinity measurements from the bottom sections of the cores, i.e. the 5 cm ice core section interfacing with the seawater. Sea ice thickness increased at an approximately constant rate from November to April, from 0.5 m to 1.6 m in first-year ice, and 0.8 m to 2.2 m in multi-year ice. From April to July, it remained roughly stable, before decreasing from July until September. First-year ice core measurements ceased in mid July with the disappearance of the coring site. From May to September, surface ocean temperature increased from -1.85 °C to -1.58 °C peaking at -1.52 °C. During this period,

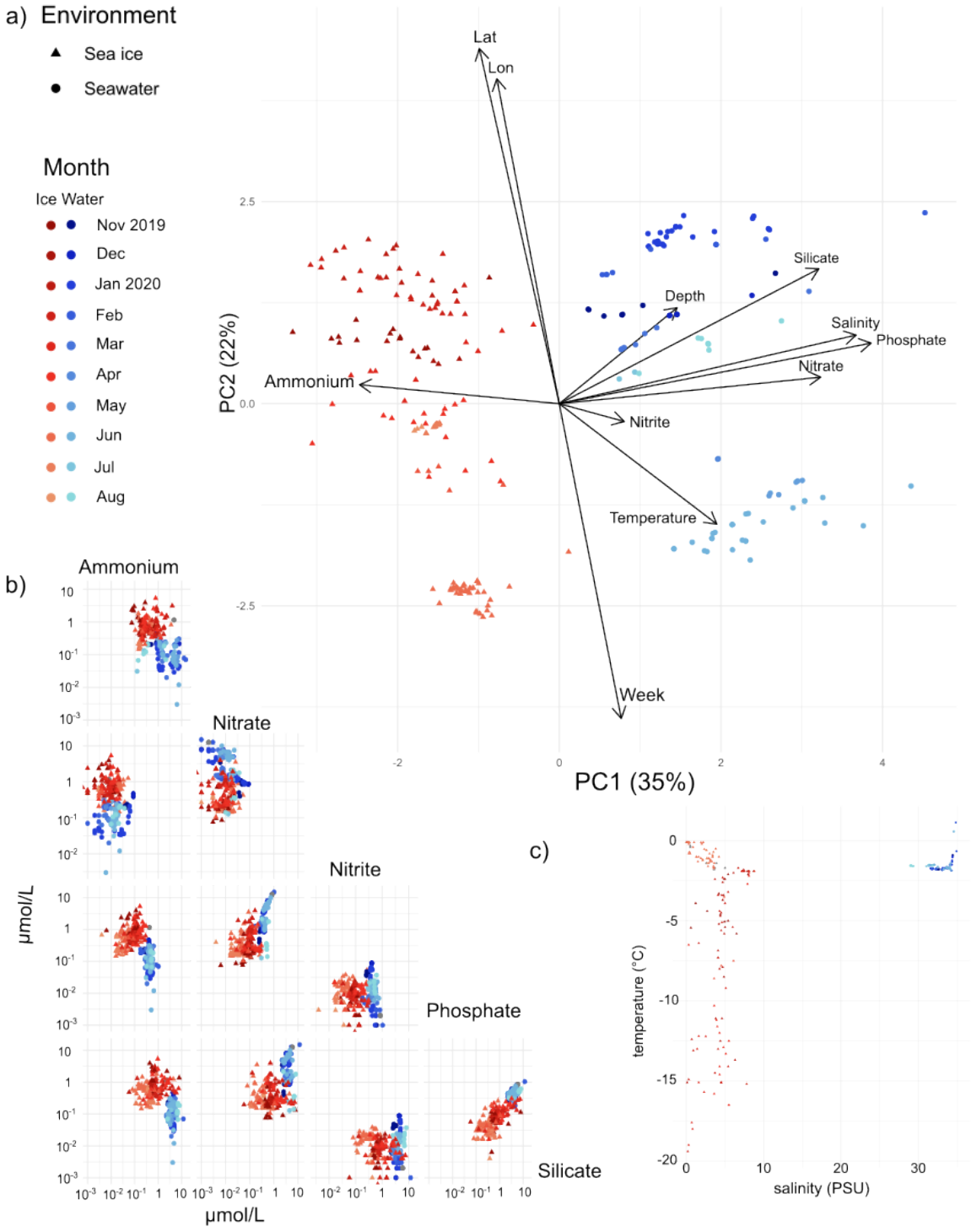


Figure 7.4: a) PCA plot of samples, using the shared physical metadata of the samples. b,c) Correlation scatterplots of the nutrient concentrations, and of temperature against salinity. Note that panel b) uses a log scale. Sediment trap samples are omitted, as they did not have any nutrient metadata.

surface ocean salinity decreased from 34 to 29 Practical Salinity Units (PSU). Some of the most rapid decreases in salinity occurred between July and September, concurrently with the largest overall reductions in sea ice thickness. In the upper 25 m ocean there is a negative correlation between temperature and salinity ($r = -0.58$), in the 25 - 100 m layer, this is reversed and there is a positive correlation ($r = 0.47$). In ice, there are two distinct regions in the temperature - salinity plot, separated by time. From November to May, temperature and salinity were positively correlated ($r = 0.59$), from June to September, the correlation changed to negative ($r = 0.61$).

7.3.2 Assembly and Annotation Statistics

There were a total of 148 billion quality-filtered reads across the set of samples, with a cumulative size of 22.1 Tbp sequenced, and an average of 69.4 Gbp (± 23.4 Gbp) per sample. The maximum and minimum numbers of bases sequenced were 198.1 and 10.9 Gbp respectively. The SPAdes single-sample assemblies resulted in a total of 535 million contigs, with a total length of 399 Gbp.

	MEGAHIT mean (s.d.) across coassemblies	SPAdes mean (s.d.) across single-assemblies
Average Length (bp)	817 (284)	995 (89.3)
Max Length (Mbp)	0.07 (0.07)	0.12 (0.05)
Min Length (bp)	500	200
Num. Seqs. (M)	1.68 (1.14)	1.11 (0.60)
Sum Length (Gbp)	1.26 (0.70)	1.10 (0.60)
N50 (bp)	1012 (434)	1018 (137)
L50 (i.e. N50_num)	11500 (3660)	8740 (2180)
GC(%)	46.7 (2.7)	48 (2.23)

Table 7.3: Means and standard deviations for summary statistics of the two sets of assemblies, single sample for SPAdes, and coassemblies for MEGAHIT. The minimum contig length for each assembler has no standard deviation across the samples / batches - in all cases the shortest contig was always produced as a default minimum value set by the assembler.

The pooling of non-prokaryotic reads from batches of similar samples, and their subsequent coassembly, using MEGAHIT, generated a total of 36 million contigs from 284 million reads (42.6 Gbp). Table 7.3 provides overall summary statistics for the assembly qualities of coassembled batches, and the single-assemblies. SPAdes generated a much larger number of long contigs (e.g. over 1 Mbp) compared to MEGAHIT, with a maximum contig length of 3.4 Mbp (SPAdes) compared to 272 kbp (MEGAHIT), and 72000 contigs over 50 kbp

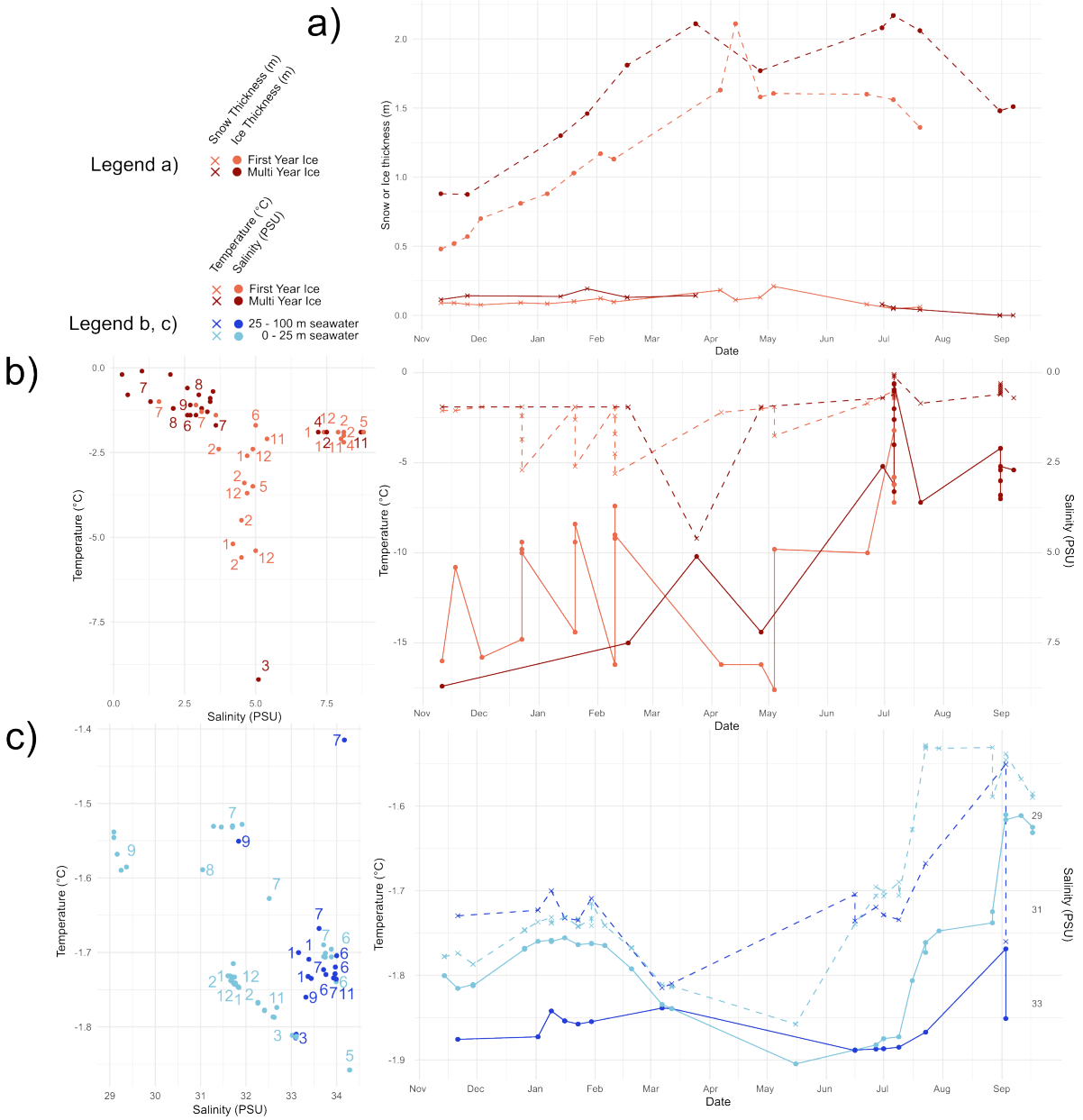


Figure 7.5: a) Sea ice and snow thickness in multi-year and first-year ice. By late July, no first-year ice remained. b,c) Left diagrams show temperature-salinity plots, the numbers represent the month of the closest data point, i.e. 11 for November, 1 for January, 2 February, etc.. Right, time series of temperature (dashed line, crosses) and salinity (solid line, dots). The salinity scale in the time series is inverted for legibility.

compared to just 875. However, there were also a large number of poorly assembled single-assemblies, mostly within sea-ice samples. This somewhat skewed some of the statistics in Table 7.3. Both assemblers had similar N50 statistics, with an average N50 of 1018 compared to 1012 from SPAdes and MEGAHIT respectively. Assembly statistics of the 33 MEGAHIT coassemblies are listed in C.1.

7.3.3 Binning Quality

There were a total of 9987 prokaryotic MAGs generated, of which 8476 had a completeness above 50% and contamination below 10%. Of those, 2340 had a completeness above 90% and contamination below 5%. Mean completeness and contamination of the prokaryotic MAGs were 75% and 2.3% respectively. 4665 MAGs were from sea ice, 4933 from seawater, and 247 from sediment traps. An average of 25.6 prokaryotic MAGs were recovered per sample in sea ice (s.d. 17.0), compared to 43.6 (s.d. 33.4) in seawater and 14.5 (s.d. 9.3) in sediment traps. Table 7.4 outlines the average numbers of prokaryotic and eukaryotic MAGs recovered per sample per month in each of the three environments. Once deduplicated using dRep at a 99% ANI similarity, there were 2867 resulting species-level clusters.

	Sea ice	Seawater	Sediment trap
Nov *	30.3 (16.1)	9.0 (6.8)	N/A
Dec	18.8 (12.5)	41.4 (24.2)	N/A
Jan **	27.9 (13.8)	64.8 (35.3)	N/A
Feb **	35.9 (13.1)	71.8 (29.9)	N/A
Mar *	20.6 (8.6)	52.0 (39.6)	N/A
Apr *	17.5 (12.9)	57.7 (36.7)	N/A
May *	15.0 (11.9)	45.0 (21.4)	N/A
Jun	13.2 (5.3)	15.3 (8.3)	25.5 (7.8)
Jul	29.8 (20.7)	23.4 (16.7)	13.1 (8.7)
Aug	11.9 (7.1)	26.0 (5.7)	N/A
Sep	17.3 (4.0)	31.4 (29.6)	N/A

Table 7.4: Means (and standard deviations) of the numbers of MAGs recovered per sample, per month, for each environment, across all 317 samples. Asterisks indicate statistically significant difference between ice and water (Welch's t-test, * $p < 0.05$; ** $p < 0.001$). Numbers of MAGs from sediment traps showed no significant difference to either the other environments in June, but were different to both ($p < 0.05$) in July.

The mean genome size was 2.58 Mbp (s.d. 1.39 Mbp), with a mean of 2500 CDSs per MAG (s.d. 1250), and coding density of 0.90 (s.d. 0.03). The mean number of contigs per prokaryotic MAG was 233 (s.d. 176); 100 MAGs had 10 contigs or fewer and 4 were composed of just a single contig. Contigs had an average N50 of 5.3 kbp. Contigs in MAGs

are inherently longer on average than contigs in the assemblies; most binning algorithms require a minimum contig length of 2 kbp, which is already above the average N50 of the assemblies. The largest prokaryotic genomes were from the phylum Planctomycetota (and family Pirellulaceae); 3 MAGs were greater than 10 Mbp in length, and out of the 23 prokaryotic MAGs greater than 9 Mbp, 16 of them were Planctomycetota. This is similar to the findings from Chen *et al.* [56], which found several large prokaryotic genomes from the phylum Planctomycetota. Cyanobacteriota had the shortest genomes on average; with a mean of 350 kbp (s.d. 651 kbp) compared to an overall prokaryotic genome size of 2.5 Mbp (s.d. 1.0 Mbp).

For eukaryotes, we used the ensemble of methods described in Section 6.2.1 to generate a total of 789 MAGs; this reduced to a set of 354 eukaryotic MAGs once we removed MAGs sharing sets of contigs, and retained only those that were more complete, and had completeness above 30% and contamination below 15%. This reduced set had a mean completeness of 60.3% (s.d. 21.2%) and contamination of 3.44% (s.d. 3.52%). Of these, 215 had completeness above 50% and contamination below 10%. The mean eukaryotic contamination percentage (3.4%) was higher than the mean prokaryotic contamination level (2.3%) - though still relatively low. This may have been down to the more stringent binning process applied to eukaryotic MAGs, involving both read filtering and a manual inspection step, the slightly different methodology when calculating contamination (prokaryotes use CheckM2, EukCC was used for eukaryotes), and the lower taxonomic diversity of eukaryotes compared to prokaryotes (and greater genomic differences between taxa, for example with less horizontal gene transfer between species) making it less likely for binning algorithms to conflate different eukaryotic species.

The ensemble of different binners and assembly pipelines performed variably when generating eukaryotic MAGs. 59 MAGs were generated by the MAP, 36 were generated from single-sample binning. Using multi-binning, 76 were generated using MetaBAT, and 111 from manual refinement of MCL clusters using the VAMB/UMAP visualisation method described in Chapter 6. The coassemblies produced a further 32 MAGs using MetaBAT, and 55 MAGs derived from VAMB/UMAP visualisations, either directly using MCL, or by manual refinement of MCL clusters using UMAP visualisations. These MAGs had an average length of 26 Mbp (s.d. 15 Mbp); the largest was a 130 Mbp copepod genome with completeness and contamination of 49% and 7.3%.

When deduplicated at a 99% ANI level with dRep, and the most complete representative for each ANI cluster chosen, this produced a set of 239 non-redundant MAGs, and 49 primary clusters (i.e. coarser clusters generated by dRep; MAGs with at least 90% MASH ANI similarity). This non-redundant set had mean completeness and contamination of 55% (s.d.

16%) and 3.9% (s.d. 3.5%) respectively. The reason why these average completeness and contamination values dropped in the non-redundant compared to the redundant set, was that more abundant strains generally had higher completeness. Removing redundant MAGs therefore tended to remove those with better completeness and contamination. For example, from a cluster of 48 *Micromonas* MAGs, 30 had a completeness above 90%, of which only a single MAG (completeness 99.8%) was retained.

Eukaryotic MAGs were fragmented, with a mean N50 of 5.6 kbp in the non-redundant set, and an average of 6900 contigs per MAG (s.d. 4700). They had an average of 12 rRNAs (s.d. 35), 65 tRNAs (s.d. 106), 54000 CDSs (s.d. 43000), and 14000 Pfam annotations (s.d. 5500). Of these Pfam annotations, only 0.7% were DUFs. This is possibly an artifact of the gene-calling methods employed; eukaryotic gene-calling is notoriously harder than prokaryotic gene-calling due to the presence of introns. Consequently, the most effective eukaryotic gene calling methods for metagenomes (we used a combination of mmseqs and BRAKER) use eukaryotic gene databases to map putative genes to known amino-acid sequences. These databases are by their nature comprised of better-described sequences, and are possibly easier to annotate than de-novo gene predictions in prokaryotes. A secondary effect is the bias in the eukaryotic MAG dataset, which is comprised of overwhelmingly better characterised clades, e.g. *Micromonas* and *Fragilariopsis*, which are both model organisms and have relatively well-annotated reference genomes.

The total size of the MAG catalogue came to 37 Gbp; of this, 26 Gbp was prokaryotic, 9 Gbp eukaryotic, and 2 Gbp viral. Contigs in MAGs were longer than the average; there were only 2.2 million contigs in MAGs (of 535 million total), yet these accounted for approximately 7.5% of the total assembly size. On average, 57% of reads mapped to the MAG catalogue (s.d. 19%), though this ranged greatly, between 96% and 17% for the best and worst-mapping samples. This mapped fraction was lower in the water (44%, s.d. 8%) compared to the sea ice (67%, s.d. 19%), although both were higher than for the sediment traps (37%, s.d. 12%). For all cases, these differences were statistically significant ($p < 0.01$). Figure 7.7 shows how this mapped fraction varied across the samples. Figure 7.6 shows the completeness and contamination of all MAGs, grouped by their taxonomy.

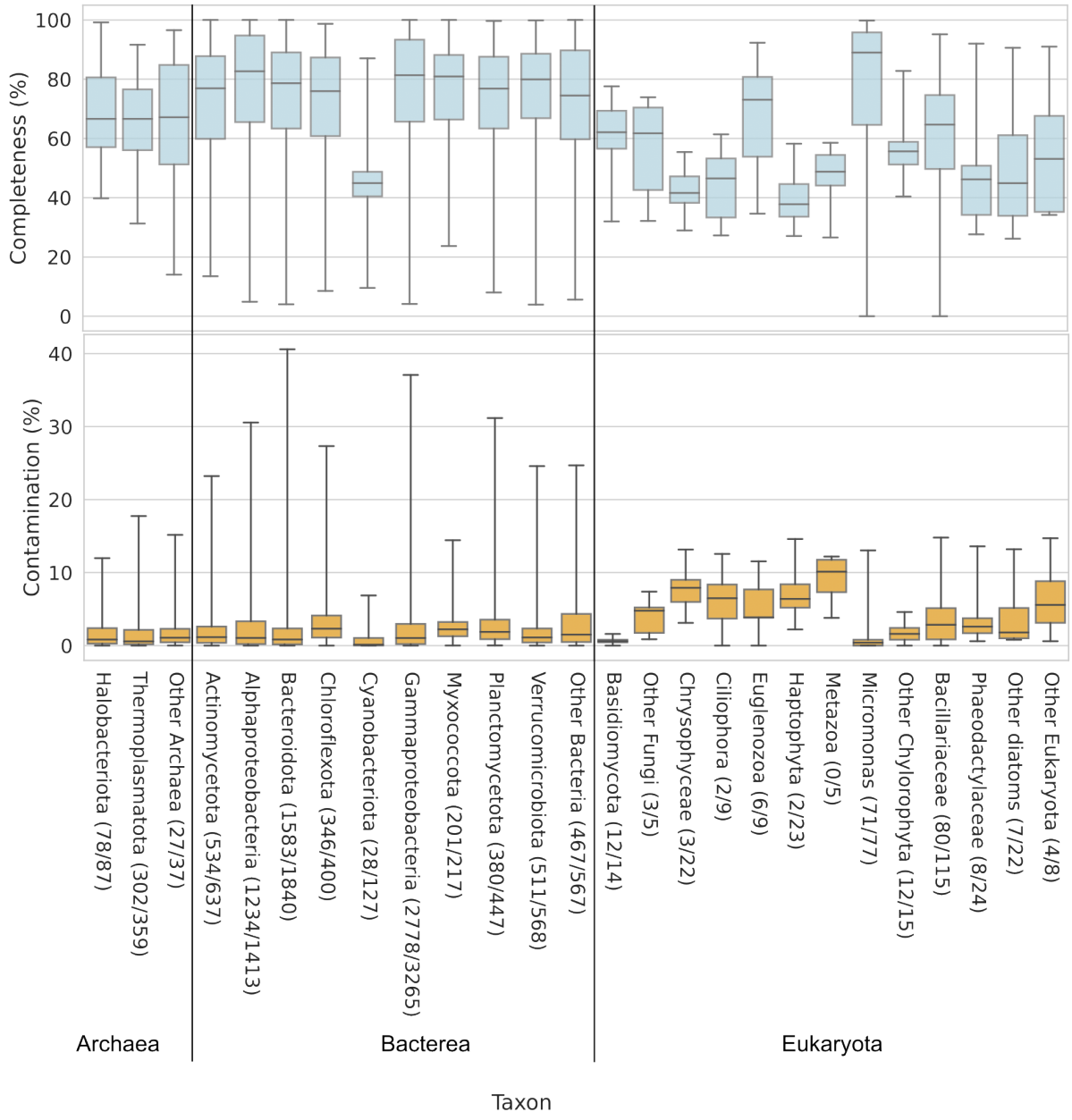


Figure 7.6: Completeness and contamination box plots of all prokaryotic and eukaryotic MAGs. Bracketed fractions denote the number of medium quality MAGs (completeness $\geq$ 50%, contamination $\leq$ 10%) out of the total number of MAGs. Taxa are mostly at the phylum level, though some Eukaryotic clades have been split, e.g. Chlorophyta, diatoms.

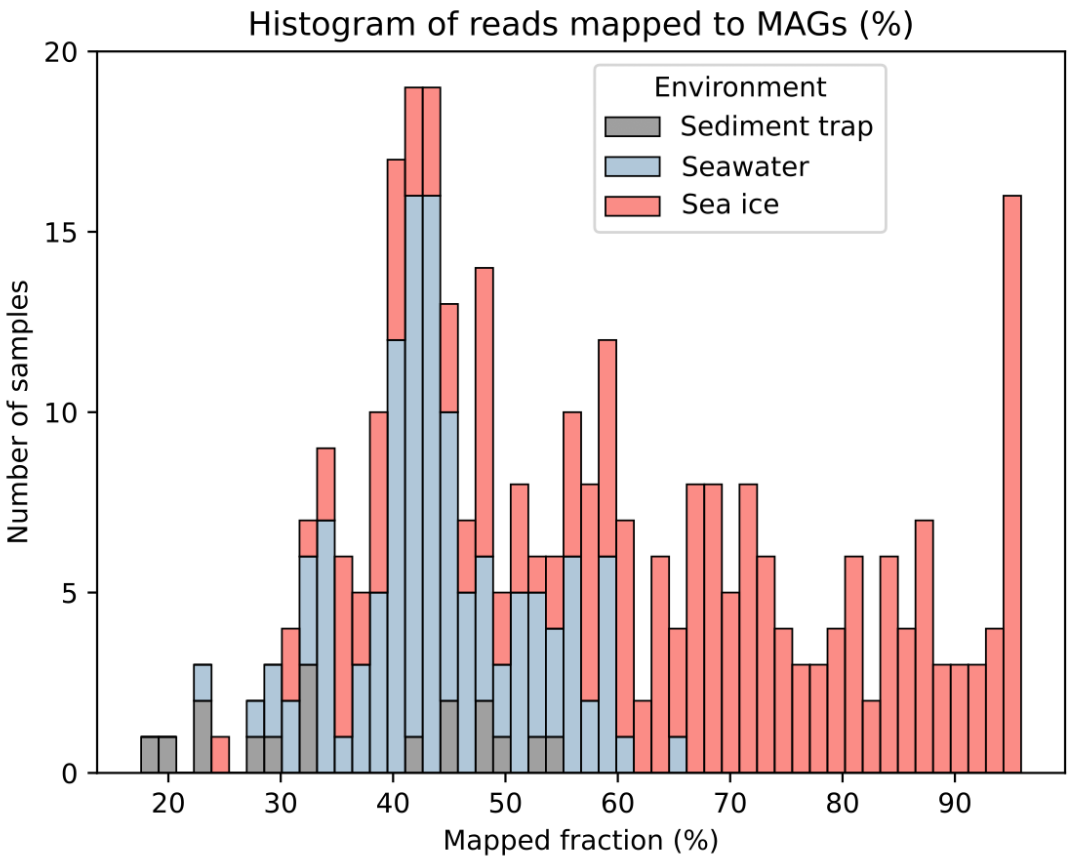


Figure 7.7: Proportion of reads mapped to the MAG catalogue.

7.4 Results II: Species

7.4.1 Taxonomy and Phylogenetics

Overview

Bacteria were the dominant domain, within both ice and water. In the ice metagenomes, 82.5% of genes were bacterial, with 14.6% eukaryotic, 2.5% viral and just 0.4% archaeal. In seawater, these proportions were 86%, 7.5%, 2.8% and 3.7% respectively. Proportions were computed by the IMG/M pipeline based on numbers of BLAST hits (at least 30% identity). With the exception of 7 sediment trap samples, which were dominated by archaea, and 2 sea ice samples, where Eukaryota were the most abundant domain, Bacteria accounted for more than 50% of gene abundance in every sample. The dominant groups were the Gammaproteobacteria, followed by the Bacteroidetes and Alphaproteobacteria. These 3 groups alone accounted for an average of 44% of genes in the metagenomes, and made up 62% of the MAG catalogue. These groups are known to be abundant in the Global Ocean [167]. Within the medium and high quality bacterial MAGs, 1205 had no species level GTDB-Tk annotation, 542 no genus level annotation, and 70 had no family level annotation. Bacteria were more abundant in the winter than in the summer, and more abundant in the seawater than the sea ice. Since abundances are computed only in relative terms, these values are more indicative of the relative lack of eukaryotes in winter rather than changes in absolute abundance (e.g. cell counts).

Eukaryotes were much more abundant in ice than in water, and more abundant again in summer than in winter. The two most prevalent classes of eukaryotes were the Mamiellophyceae, in particular the genus *Micromonas*, and the Bacillariophyceae (diatoms). There were 75 and 139 MAGs from these two classes respectively, of 354 eukaryotic MAGs total.

Archaea accounted only for 3% of the genes in the metagenomes, and just only 485 MAGs were archaeal out of 9987 prokaryotic MAGs. However, archaea were extremely abundant within certain sediment trap samples, where in some case, species of Halobacteria (phylum Haloarchaeota) constituted between 35% - 65% of the genes in the metagenome. This could represent them filling a niche, though it is hard to interpret this without further context around the sediment trap samples including more details about their processing, and the nutrient concentrations within the traps. Other archaeal clades detected by the IMG/M pipeline were the candidate class Poseidoniia (formerly Marine Group II archaea, known to be abundant in pelagic samples [413]), and the class Nitrososphaeria (both within the phylum Thermoplasmatota), which were present at levels under 2% in the seawater, and 0.1% or less in the sea ice. These 3 classes accounted for almost all the archaeal MAGs, with 359 being

Poseidoniia, 87 Halobacteria, and 13 Nitrososphaeria. Binning also generated 18 archaeal MAGs of the class Nanosalinia, and 7 MAGs of other classes. 23 MAGs were unidentified by GTDB-Tk at the genus level.

Amongst viruses, the population was mostly represented by two groups, Caudoviricites, and Megaviricites. These represented approximately 2% of genes across ice and water. Megaviricites include the clade of Phycodnaviridae; algae-infecting giant viruses, which was the most frequent family within the genomad annotations.

We first compared the taxonomic distributions of the whole metagenomes (based on the numbers of BLAST hits in genes, generated by the IMG/M MAP) with those based on abundances of MAGs, using RPM. This was to evaluate the extent to which MAGs were representative of the full metagenome, including both binned and unbinned contigs. Figure 7.8 shows the Pearson correlation coefficients between the two methods, for taxonomic ranks down to the family level, including only the taxa shared between both abundance estimation methods. (The IMG/M pipeline used slightly different taxonomic nomenclature compared to GTDB lineages, which confounded some of the analysis for a few of the rarer taxa.) At the domain level there is a very high (mean of 0.93, s.d. 0.15) correlation between the gene-abundances and MAG-abundances. Correlation generally decreased as the specificity of the taxonomic rank increased, however at the rank of family, the correlation was still 0.67 (s.d. 0.2).

This analysis, combined with the fact that the proportion of reads mapped to the MAG catalogue was relatively high (57%), suggests that using MAGs to measure abundance is a valid approach in this case, though there are limitations in this methodology, as described in Section 3.7.3. Calculating abundance through gene counts of BLAST hits also has some bias associated with the choice of database - ultimately it is necessary to pick some reasonable methodology and stick with it.

From hereon, we focus on abundances within the MAGs, and all metrics based on abundance are calculated using RPM mapped to the MAG catalogue, unless specified otherwise.

Prokaryotes

Figure 7.9 shows phylogenetic species trees of prokaryotic MAGs, with several major phyla and the provenance of MAGs indicated. For the bacterial tree, only species level representatives (i.e. 99% ANI clusters) and displayed.

The largest prokaryotic phylum was the Gammaproteobacteria; 3265 MAGs were from this clade, of which 2778 were at least medium quality. Within the Gammaproteobacteria, there were 704 species-level clusters. The largest order was the Pseudomonadales, with 1672 MAGs and 321 species-level clusters. The next two largest orders were the Enterobacterales

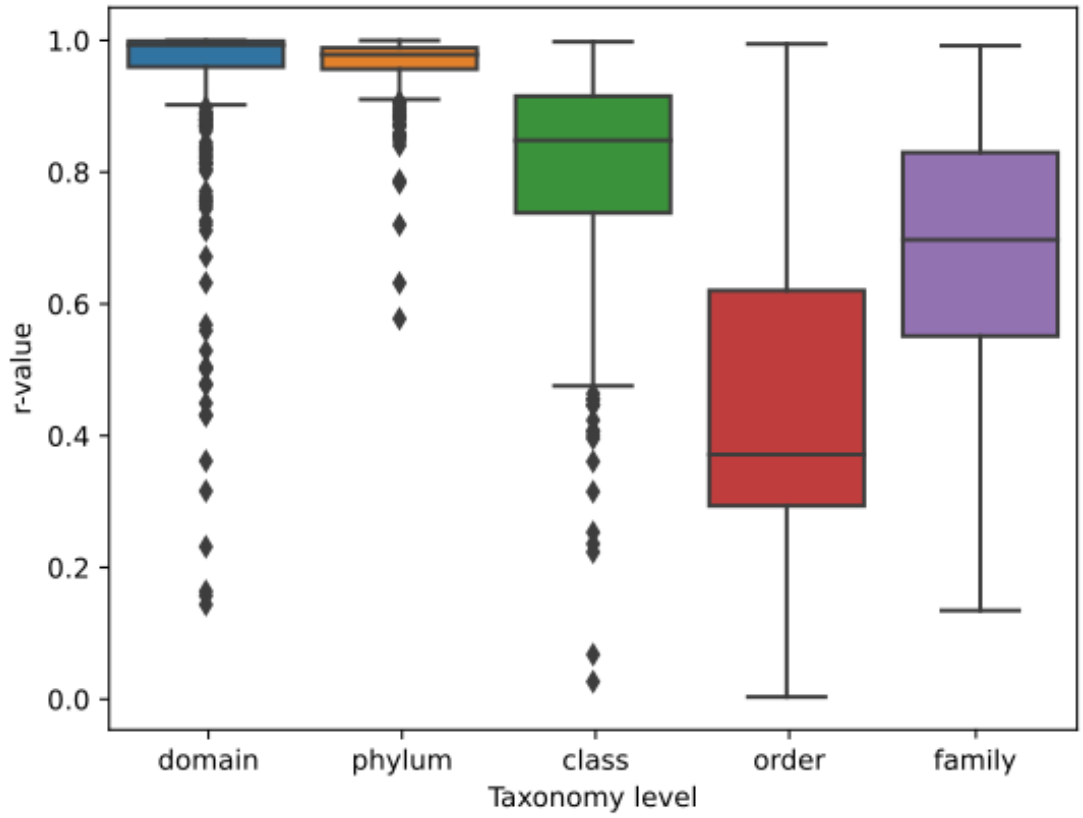


Figure 7.8: Correlation coefficients between the proportions of taxonomic labels per samples identified within genes through IMG/M MAP, and taxonomic abundances of MAGs (RPM). A high correlation indicates that the two methods agree well on the proportions within a sample, for that specific taxonomic level. As the taxonomy becomes more specific, the correlation generally decreases as there are more distinct categories. Additionally, taxonomic nomenclature diverges between NCBI and GTDB, which confounds some of the analysis.

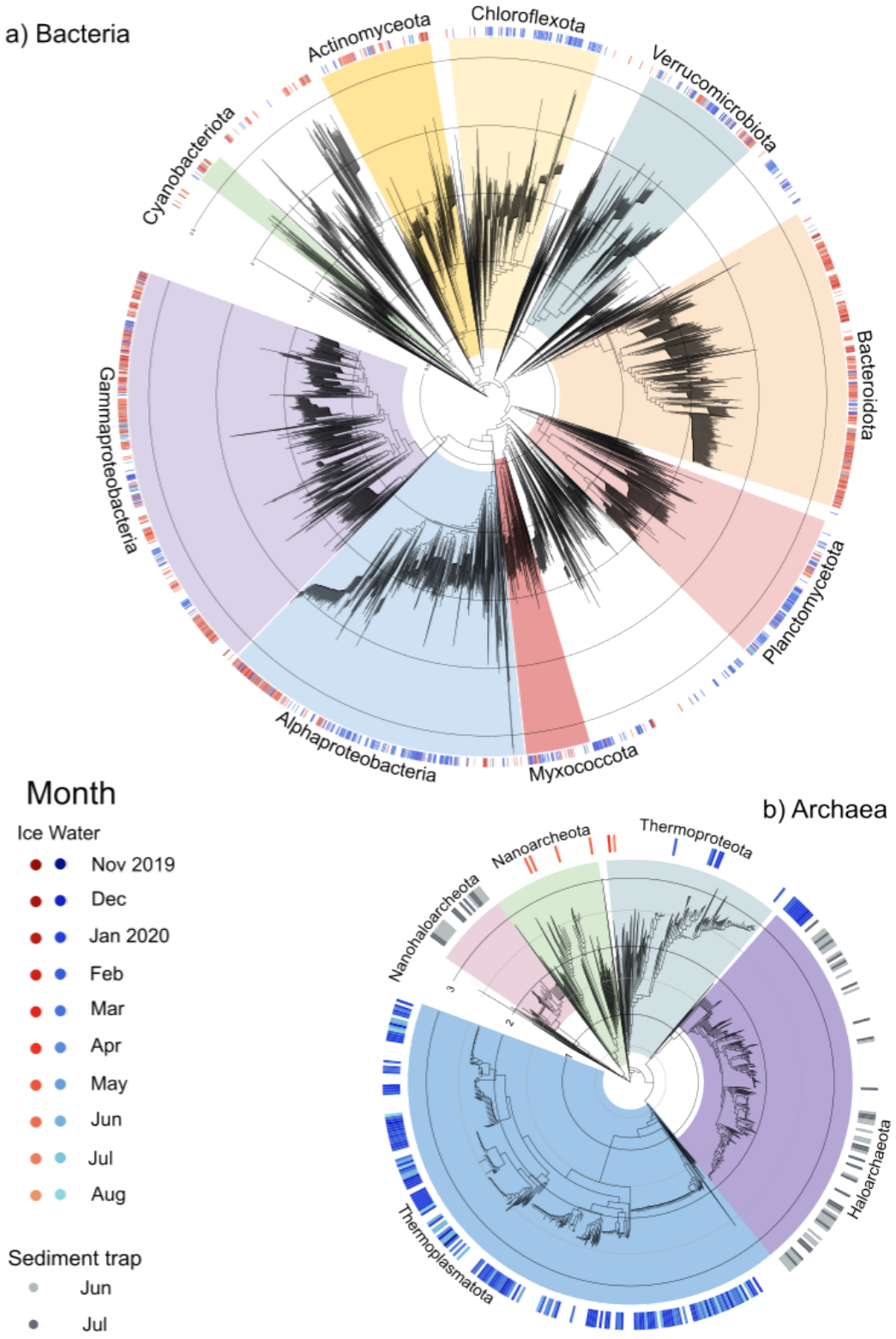


Figure 7.9: Species trees of bacteria and archaea, with GTDB genomes as references. Several significant clades are highlighted. The environment from which MAGs were recovered is indicated by the outside coloured strip. No colour indicates a reference genome.

(600 MAGs) and Burkholderiales (417 MAGs). Most prevalent within these two taxa were several known psychrophilic genera, for example *Glaciecola* and *Paraglaciecola* (95 and 78 MAGs respectively), which were first found in ice cores, see Bowman *et al.* [98].

Within the second largest phylum, Bacteroidota, there were 1840 MAGs and 477 species-level clusters. The largest family by far was the Flavobacteriaceae (770 MAGs), with Flavobacterium and Polaribacter the two largest genera within this family. The third largest phylum, Alphaproteobacteria, had 1413 MAGs and 438 species-level clusters. The largest of these species level clusters consisted of 69 MAGs which were identified as the species *Sulfitobacter sp. CW3*. These are known DMSP (dimethylsulfoniopropionate) degraders; this compound is known to be produced by eukaryotic algae, so when DMSP data from MOSAiC becomes available it would be interesting to correlate these taxa with DMSP concentrations, as well as other organic sulphur compounds (DMSO, DMS, both other DMSP derivatives) and DMSP-degrading related genes such as *DmdA* and *DmdB*. Less prevalent phyla, with more than 100 MAGs, are listed in the table below, and a full list of numbers of MAGs of all phyla is provided in Appendices C.5 and C.6.

Phylum	Number of MQ MAGs	Number of MAGs
Actinomycetota	534	637
Alphaproteobacteria	1234	1413
Bacteroidota	1583	1840
Chloroflexota	346	400
Cyanobacteriota	28	127
Gammaproteobacteria	2778	3265
Myxococcota	201	217
Planctomycetota	380	447
Thermoplasmatota	302	359
Verrucomicrobiota	511	568
Other Bacteria	472	573

Table 7.5: Numbers of MAGs (medium quality or above, and total) per phylum, in the most prevalent prokaryotic phyla. Other bacteria represents all remaining phyla, including (in descending numbers): Patescibacteria, Gemmatimonadota, Bdellovibrionota, Marinisomatota, Acidobacteriota, Latescibacterota, SAR324, Nanohaloarchaeota, Desulfobacterota, Nitrospinota, and 21 other phyla each with 10 MAGs or fewer.

Several other less prevalent bacterial phyla were noteworthy, either for their novelty, or abundance within specific environments. We recovered 127 Cyanobacteriota bins, all but three of which were from unknown families; however most had too low completeness to be considered medium quality MAGs. Only 28 were medium completeness or above. Cyanobacteriota are typically not abundant within the Arctic; in the Royo-Llonch MAG

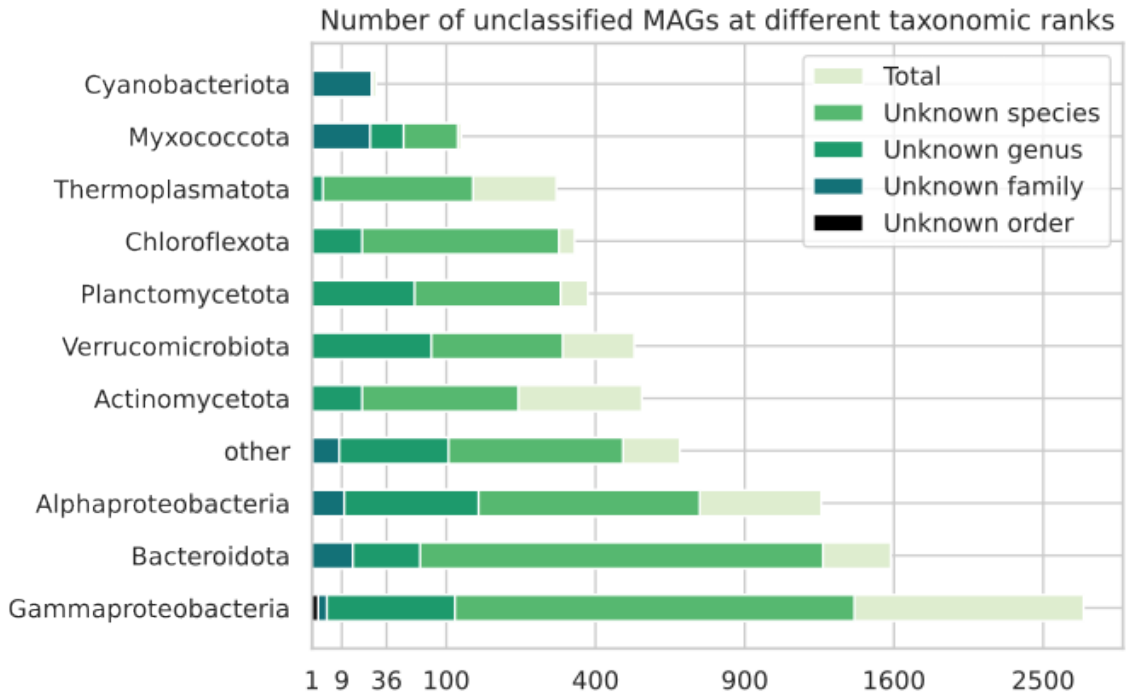


Figure 7.10: Prokaryotic MAGs unclassified by the GTDB, representing new species and potential novelty at higher taxonomic ranks. Only medium quality MAGs are listed. The x-axis is square-root scaled.

compendium of Arctic MAGs [105], only 2 of 530 MAGs were from Cyanobacteriota. Though there were only 217 Myxococcota (previously called Deltaproteobacteria), 26 of these MAGs had no taxonomic annotation beyond the order level, and the annotations that exist were generally of other MAGs such as UBA786, a MAG recovered from a hydrothermal vent. Figure 7.10 shows the numbers of medium quality MAGs unclassified by the GTDB at different taxonomic ranks. These represent novel taxa, compared to the Genome Taxonomy Database.

Archaeal MAGs were conspicuously prevalent with the sediment trap samples. Several large clades of Haloarchaeota and Nanohaloarchaeota were the dominant MAGs recovered from sediment traps (95 out of 261 MAGs). Conversely, just a single clade of 10 closely related Haloarchaeota were recovered from all the other samples combined. Within the water samples, Thermoplasmatota were the most abundant archaeal phylum, and accounted for 359 of the 485 archaeal MAGs. This was the only archaeal phylum that produced a significant number of MAGs, similar to other major bacterial clades that were present within the water column such as Verrucomicrobia (568 MAGs), Chloroflexota (400 MAGs)

and Planctomycetota (447 MAGs).

Eukaryotes

There were two dominant groups within the eukaryotes; *Micromonas* and Bacillariophyta (Figure 7.12). These two groups are known to be highly abundant in the Arctic Ocean [112], [414]. Both groups also have slightly smaller genomes relative to many other eukaryotes, often smaller than 100 Mbp. They are therefore more successfully assembled and binned into medium quality MAGs than other eukaryotes; this can be a source of sampling bias within MAG studies. To alleviate this problem we retained low quality eukaryotic bins (those above 30% completeness, less than 15% contamination, but below 50% completeness or above 10% contamination) in the MAG catalogue; these low quality bins were harder to place phylogenetically due to a lack of marker genes. We will distinguish between low and medium quality eukaryotic MAGs in our subsequent analysis. For low quality MAGs, and MAGs on long branches in the phylogenetic tree, we assessed their taxonomy based on their placement within the tree, on the assessment of their clade from EukCC, and from mmseqs hits to our combined MMETSP, Phycocosm, and UniRef database. For more complete genomes, close in the tree to references, we used ANI to refine their taxonomy.

The largest clade was the Bacillariophyta, with 159 MAGs (95 MQ, medium quality), and 125 species level clusters. Within these, 113 were from the family of pennate diatoms Bacillariaceae (80 MQ), 24 from the family Phaeodactylaceae (8 MQ), and 22 either unclassified at the family level or from other families (7 MQ). The largest of the species level clusters comprised just 9 and 8 MQ MAGs respectively; 9 from the algal genus *Cylindrotheca*, and 8 from the genus *Pseudo-nitzschia*. All other species level clusters of Bacillariophyta were of a size at most 2, and 111 were singletons; they did not share more than 99% ANI with any other MAG. Most (45) of the Bacillariaceae were unidentified beyond the family level, 37 diatoms were from the genus *Fragilariopsis*, 24 from the genus *Phaeodactylum*, 17 from *Cylindrotheca*, 10 from *Pseudo-Nitzschia*, the remainder from other genera (see table 7.6). The taxonomy of contigs was variable; in 23 MAGs, over 10% of contigs were given a taxonomic annotation inconsistent with Bacillariophyta. However, the numbers of contigs that were annotated as prokaryotic was always extremely low, in most cases fewer than 10 contigs.

The second largest clade by far was the genus *Micromonas*, comprising 75 MAGs (71 MQ), but just 29 species level clusters. This reduction in diversity was due to one enormous species level cluster of 47 MQ MAGs; the remainder were all singletons that did not share 99% ANI with another MAG, however, they were close to one another on the phylogenetic tree. The closest reference genome was the MAG *Micromonas sp.* AD1, described in [104].

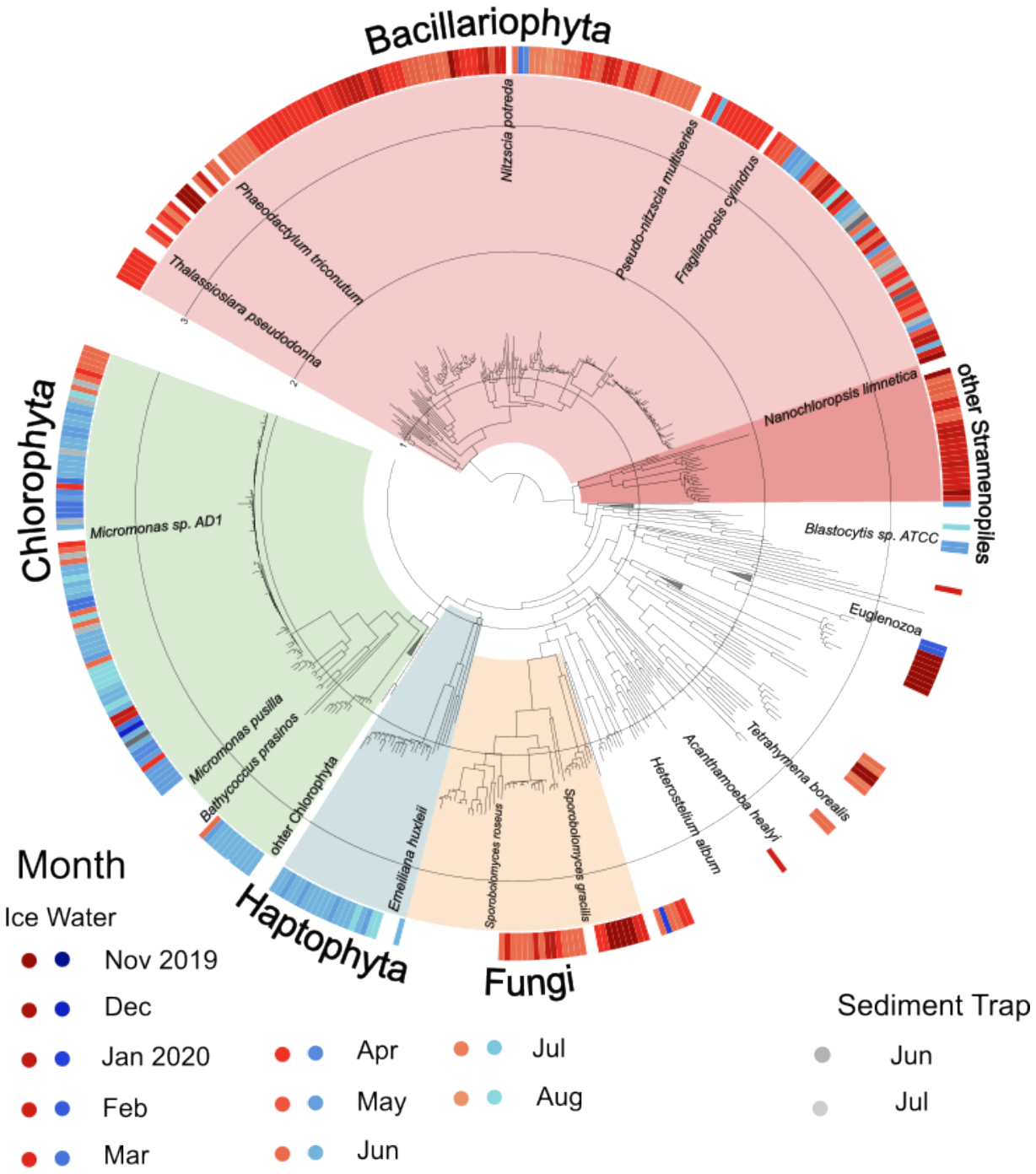


Figure 7.11: Species trees of Eukaryotic MAGs. Several major clades are highlighted. The coloured strip represents the provenance of the MAG, blue shades for water, red for ice, grey for sediment trap, and white represents reference genomes from NCBI RefSeq or Phycocosm. Some reference genomes are labelled, though most labels are omitted for legibility, and some clades of reference genomes are collapsed. The tree is rooted with Bacillariophyta as an outgroup.

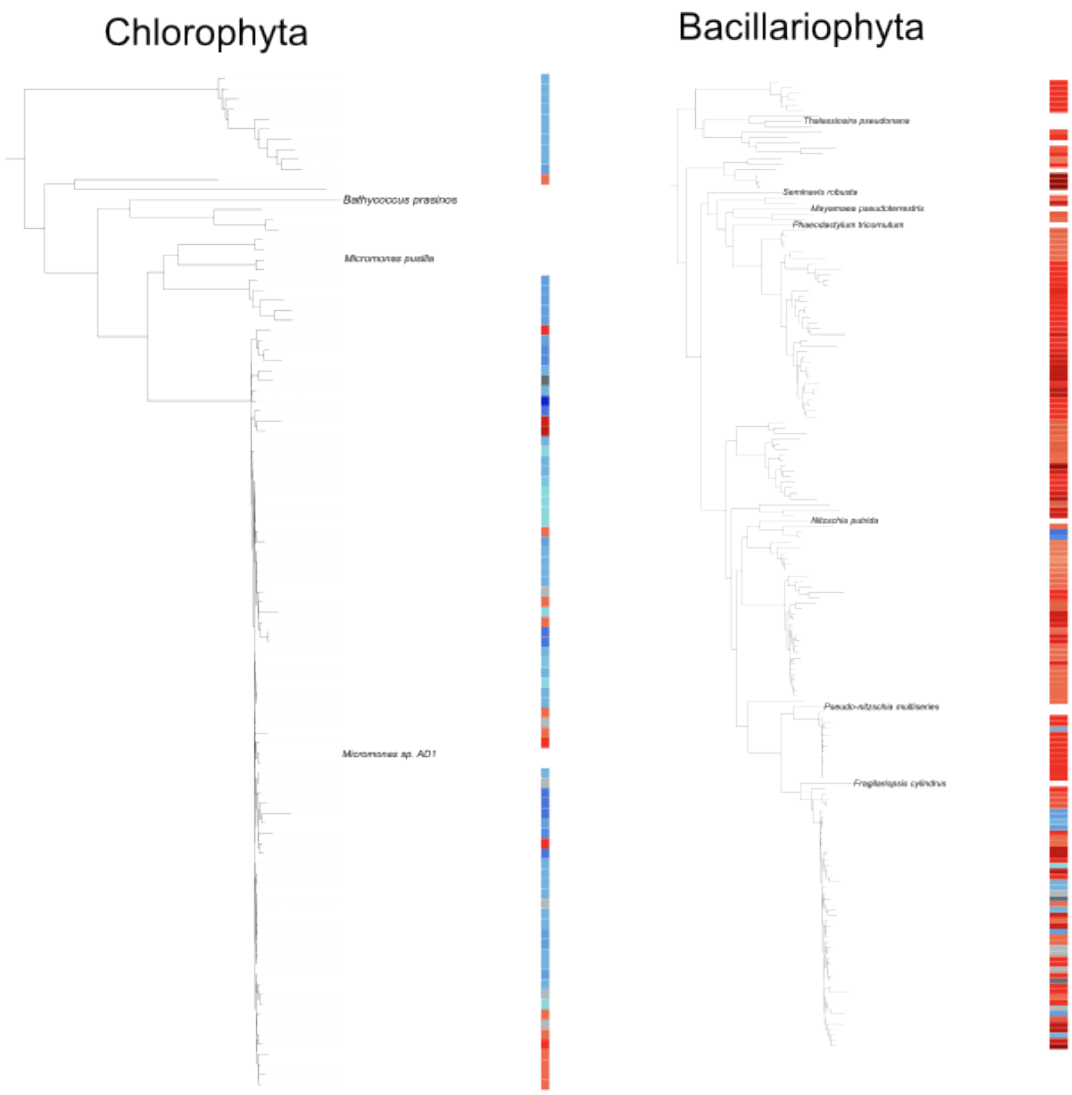


Figure 7.12: Subtrees of the phylogenetic species tree for the two largest groups of MAGs; Bacillariophyta and Chlorophyta. The legend for the coloured strip is the same as in Figure 7.11.

Taxon	Number of MQ MAGs	Number of MAGs	Number of Species level clusters
Bacillariaceae	37	45	36
Fragilariopsis	20	37	33
Phaeodactylum	8	24	21
Cylindrotheca	9	17	13
Pseudo-nitzschia	10	10	3
Attheya	2	6	6
Nitzschia	4	4	4
Chaetoceros	1	4	4
Naviculales	2	2	2
Fistulifera	0	2	2
Synedropsis	1	2	2
Amphora	1	2	2
Bacillariophycidae	0	1	1
Bacillariophyta	0	1	1
Bacillariophyceae	0	1	1
Seminavis	0	1	1

Table 7.6: Numbers of MQ MAGs, all MAGs, and species level clusters for the Bacillariophyta. Taxonomy is based on the most common contig annotation from mmseqs.

Micromonas genomes were the most complete eukaryotic genomes recovered, with 35 MAGs above 90% completeness, and less than 5% contaminated. Other evidence from taxonomic annotations of contigs (using mmseqs) suggested that these MAGs were generally of good quality; there were just 10 medium quality *Micromonas* MAGs where over 1% of contigs were inconsistent with the genus *Micromonas*. In 38 MAGs, there were fewer than 10 contigs with an inconsistent taxonomic annotation. High quality eukaryotic MAGs were normally uncommon (both in this study and in other studies, e.g. [104], [248]), *Micromonas* were an exception due to their small genome size, abundance within Arctic pelagic environments, and the availability of at least one reference genome.

From Figures 7.11 and 7.12 it is clear that Bacillariophyta were mainly recovered from the ice (as well as other Stramenopiles and Fungi), whereas Chlorophyta and Haptophyta are predominantly recovered from the water column. There were exceptions to this pattern - several *Micromonas* were recovered from ice and some *Fragilariopsis* were recovered from water, most often within the summer months (indicated by the lighter shading in the figures). We will explore this further in the next section.

Other Chlorophyta included 6 MAGs from the genus *Pyramimonas*. Within these MAGs, only approximately 88% of contigs were annotated consistently with *Pyramimonas*, though all MAGs were assessed as medium quality by EukCC. These MAGs were all within one

Taxon	Number of MQ MAGs	Number of MAGs	Number of Species level clusters
Stramenopiles	3	23	23
Haptophyta	2	23	21
Fungi	15	20	12
Chlorophyta	12	15	14
Alveolata	3	12	11
Euglenozoa	6	9	8
Metazoa	0	5	5
other	1	2	2
Choanoflagellata	1	1	1
Apusomonadida	0	1	1
Discosea	0	1	1
Streptophyta	1	1	1
Bacillariophyta	95	159	125
Micromonas	71	75	29

Table 7.7: Numbers of MQ MAGs, all MAGs, and species level clusters for all eukaryotic phyla. Bacillariophyta and *Micromonas* are counted separately below the double line.

clade, on a relatively long branch in the phylogenetic tree.

Other clusters included two groups of closely related MAGs from the fungal genus *Sporidobolmyces*, and two groups of Euglenazoa. While all the fungal MAGs were recovered from ice, the two distinct Euglenazoa clades were from the ice and water respectively. There were a large number (23) of Stramenopile MAGs which were identified as Chrysophyceae through taxonomy of their contigs, but which did not have any close relative in the tree. The closest relative was the microalga *Nannochloropsis limnetica*. These MAGs were mostly of poor quality, only 3 were medium quality, and some were on very long branches. All were recovered from the ice. There was a similar case with the Haptophytes; of 23 MAGs only 2 were medium quality and their taxonomy could not be ascertained either through the phylogenetic tree or by taxonomy of their contigs. The closest relative in the tree was *Emiliania huxleyi*, though they were separated by a long branch. The taxonomy of the contigs was also not very specific or consistent. In contrast to the Stramenopiles, all the Haptophyte genomes were recovered from water rather than ice, between April and August.

7.4.2 Abundance and Diversity

Alpha diversity, measured via the Shannon index from RPM MAG abundance (Figure 7.13), was higher in the seawater samples compared to the sea ice, with an average Shannon index of 10.3 in water (s.d. 0.8) and 8.9 in ice (s.d. 0.9), this difference was significant ($p < 0.001$).

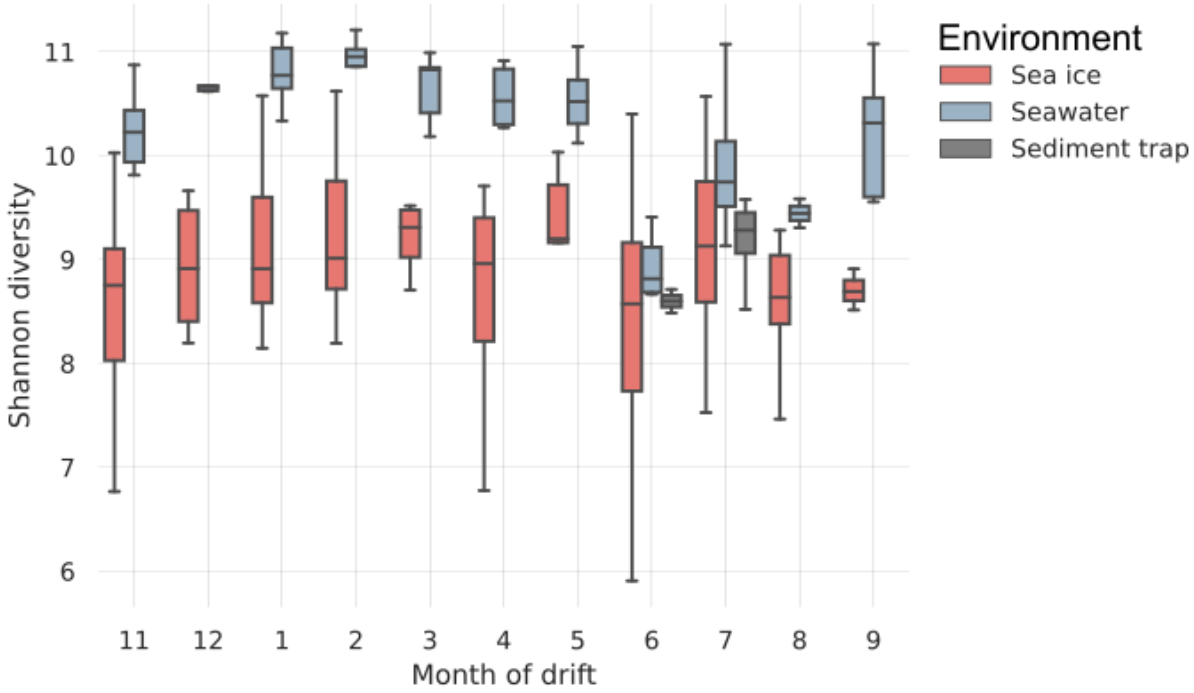


Figure 7.13: Alpha diversity (Shannon index) aggregated per month of the drift.

Alpha diversity in the sediment traps was 9.1 (s.d. 0.4). Alpha diversity remained higher in the water than ice in each month; this difference was significant at the 5% level for each month except June. This roughly followed the same pattern as the richness results (see Table 7.4), where more MAGs were recovered per sample from seawater than sea ice in each month except November and June. A large amount of the drop in diversity in June in the water samples could be explained by a sharp increase in eukaryotic abundance in that month; when eukaryotes were removed from the analysis, diversity dropped by much less between May and June in water, with an average reduction in the Shannon index of 1.9 when eukaryotes were included, and 1.0 without. In these months, a small number of eukaryotic species (particularly Haptophytes and Chlorophytes), drastically increased in number and so reduced overall diversity.

MAG abundance and richness were correlated; more abundant phyla were the same as the phyla with large numbers of MAGs, and the environment from which a MAG was recovered from was a good predictor of the environment where that MAG was most abundant. Prokaryotes, specifically bacteria, were more abundant than eukaryotes, with an average abundance (RPM) of 71% compared to 17% (archaea 2%, viruses 8% and the remainder unclassified). The 10 most abundant prokaryotic phyla were the same as those listed in

table 7.5.

We classified MAGs as being present or absent within a sample using a simple threshold, where a MAG was considered present if its mean depth of coverage was at least 0.5. Under this criterion, only 4 MAGs were considered to be absent within all samples, and no MAGs were considered present in all samples. 1215 MAGs were obligate pelagic species (i.e. absent from sea ice and sediment traps), while 2158 were obligate sympagic, and 104 were present only in sediment traps. There were just 84 MAGs which were present in over 50% of both water and ice samples, of these, 33 were medium quality prokaryotes: 13 Gammaproteobacteria, 8 Bacteroidota, 3 Thermoproteota (Archaea), 2 Alphaproteobacteria, 2 Cyanobacteriota, and one of each of Patescibacteria, Verrucomicrobiota, and the SAR324 group. 9 were eukaryotic: 2 Chrysophyceae, 2 Bacillariophyta, 1 *Micromonas*, 1 Apusozoa, 1 Haptophyta, 2 uncharacterised eukaryotes (all of low quality). Some of these results are hard to explain; the Thermoproteota, Patescibacteria, SAR324, and Cyanobacteria were relatively rare MAGs, yet specific MAGs from these phyla were apparently present (even if in low abundance) in over half of all samples, regardless of sample type.

Environmental differences

Gammaproteobacteria were the most abundant group, both in ice and water, though they were particularly abundant in the ice, especially in winter. The mean abundance (total RPM) in ice was 47% (s.d. 23%, compared to 22% (s.d. 9%) in water samples. From November to February, Gammaproteobacteria accounted for over 75% of total abundance in more than half of ice samples. In contrast, the abundance of Gammaproteobacteria in all water samples was at most 30% over the same period, though this still meant it was the single most abundant prokaryotic phylum during that time. Bacteroidota were similarly much more abundant in ice (16%, s.d. 13%) rather than water (5%, s.d. 5%), however the seasonal pattern of the Bacteroidota was very different to the Gammaproteobacteria, with a peaks of abundance in March (mean 23%, s.d. 10%) and June and July (mean 24%, s.d. 15%) in the ice. The only other prokaryotic phyla with an overall abundance in ice over 1% were the Alphaproteobacteria (7.5%, s.d. 6.7%) and the Actinomycetota (2.5%, s.d. 2.8%), though these phyla were more abundant in water (11.8% s.d. 5.5% and 4.3% s.d. 3.8%, respectively).

Interestingly, we found the only other prokaryotic group with an overall abundance above 0.5%, and which was more abundant in ice rather than water, was the Cyanobacteriota, with 0.6% abundance in ice (s.d. 1.2%) and 0.2% abundance in water (s.d. 0.3%). Though these abundances were small, it was noteworthy the abundance of Cyanobacteriota within sea ice has been unclear; Cyanobacteriota have been associated with melt ponds and near coastal

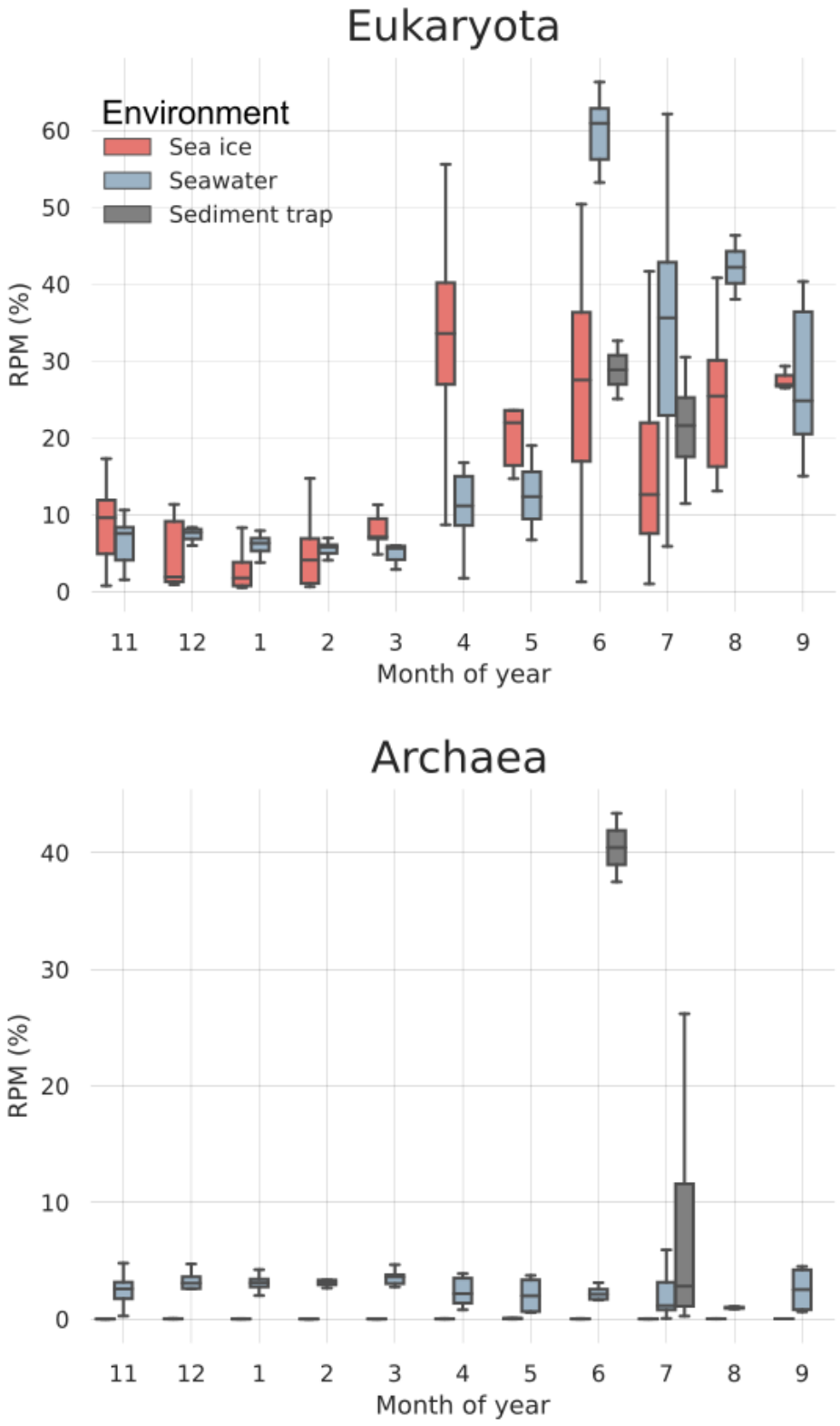


Figure 7.14: RPM abundances of eukaryotes and archaea. Virus abundances are in Appendix C.5. The remainder in each month are bacteria.

Arctic regions, but not necessarily within the ice [415], [416]. In one sea ice sample, the abundance of Cyanobacteriota was as high as 7.7%.

All significant prokaryotic phyla (those listed in table 7.5) other than the Gammaproteobacteria, Bacteroidota, and Cyanobacteriota were more abundant in water rather than ice. In all cases this difference was statistically significant ($p < 0.01$). This included the Verrucomicrobiota, Chloroflexota, Planctomycetota, and Myxococcota. For these phyla, 84% of their MAGs were recovered from water, and their combined abundance within the ice was 1.3% compared to 15.4% in water. 30% of these phyla were obligate pelagic species; 10% were obligate sympagic. For the Chloroflexota in particular, these proportions were 44% pelagic and 0.5% sympagic. Table 7.8 shows abundances and counts of MAGs from the different environments.

There was one set of deep ocean MAGs, consisting primarily of Actinomycetota, which were particularly associated with the single bathypelagic sample (4042 m depth). These MAGs also had higher abundance in other mesopelagic samples, especially those from 100 m depth or more. Some of these MAGs were present also in ice, particularly in the winter, and were correlated with higher nitrate and ammonium concentrations.

Within the Eukaryotes, the Haptophyta and Chlorophyta were the only two clades more abundant within water than ice. Haptophytes were present at an abundance of up to 25% in the water, Chlorophytes (specifically *Micromonas*) were present at up to 40%. However, only 1 eukaryotic MAG (a Haptophyte) was absent from all sea ice samples. In contrast, 98 eukaryotic MAGs were absent from the water, including 88% of the Phaeodactylaceae and all but 1 of the Fungi.

Viruses were omnipresent within the samples at high abundances; we do not study viruses in depth but merely note that their abundances in ice and water were 6.8% and 10.1% respectively, making them the third most abundant group in each environment. Since RPM abundance is not normalised using genome size, the number of virus particles must be extremely high, at least an order of magnitude higher than any phylum. This is consistent with previously observed viral abundance in the oceans more generally [417]. The two largest viral classes were the Caudoviricetes and Megaviricetes.

All abundance distributions were highly skewed - for every single phylum the median abundance was less than the mean abundance, with the mean often being inflated by certain samples with a particularly high abundance of one phylum. However, using median statistics rather than the mean did not meaningfully change any of the analysis.

	MAG richness (counts)			MAG abundance (mean % RPM)		
	Ice	Water	S.T.	Ice	Water	S.T.
Actinomycetota	249	383	2	2.45	4.34	0.42
Alphaproteobacteria	596	778	23	7.52	11.76	3.7
Bacteroidota	1494	255	37	16.35	5.35	18.75
Chloroflexota	2	398	0	0.05	2.78	0.07
Cyanobacteriota	86	41	0	0.62	0.21	0.76
Gammaproteobacteria	1817	1322	79	47.17	22.3	28.62
Halobacteriota	0	10	77	0.0	0.01	11.29
Myxococcota	53	76	0	0.21	0.74	0.04
Planctomycetota	67	376	0	0.26	3.49	0.23
Thermoplasmatota	0	358	1	0.03	2.7	0.24
Verrucomicrobiota	120	433	10	0.75	6.36	1.21
other prokaryotes	176	417	18	2.07	10.7	4.04
Bacillariaceae	92	13	7	4.76	1.69	4.9
Phaeodactylaceae	24	0	0	1.31	0.12	0.25
other Bacillariophyta	22	0	0	0.98	0.37	0.56
Chrysophyceae	22	0	0	3.3	1.51	2.81
Ciliophora	9	0	0	0.36	0.05	0.05
Euglenozoa	7	2	0	0.16	0.06	0.05
Fungi	19	0	0	0.26	0.02	0.05
Haptophyta	0	23	0	0.93	4.15	3.41
Metazoa	4	1	0	1.92	0.79	0.59
Micromonas	16	51	7	0.51	7.17	6.55
other Chlorophyta	1	14	0	0.49	1.41	2.18
other eukaryotes	6	4	0	0.67	0.7	0.61

Table 7.8: Numbers of MAGs, and their mean abundance, in each environment, and for each major clade. S.T.; sediment trap. Prokaryotes are listed alphabetically above the double line, then eukaryotes, with diatom families first, then other eukaryotic clades, and finally *Micromonas* along with other Chlorophyta. In every single instance, there was a significant difference between abundance in water and ice (Welch's t-test, $p < 0.05$).

Seasonal differences

The two main factors for changes in abundance and diversity were from eukaryotic blooms, most likely driven by the return of light (beginning in February, but with a detectable increase in chlorophyll-a only from late March / April) [111], [418], [419] which corresponded with an increase in eukaryotic species in the sea ice, and a summer bloom starting from June which corresponded to the start of the melt season [420]. The increase in eukaryotic abundance in June and July in the water was reflected by corresponding decreases in alpha diversity

in those months - since a smaller number of eukaryotic species tended to dominate a large fraction of each sample during these bloom seasons.

Several prokaryotic phyla showed a distinct annual growth pattern in the water (figures C.5 of Alphaproteobacteria, Chloroflexota, Myxococcota, Planctomycetota, and Verrucomicrobiota), with increasing abundance from November to February, followed by a dip and then rise in abundance until May, and then a sudden large decline from June until September. Some of the sharp drop in abundance might be explained by the corresponding increase in eukaryotic phyla that occurred in June in the water column, particularly of Haptophyta and *Micromonas*, since some drop in abundance from May to June was observed for other prokaryotic phyla as well (Alpha- and Gammaproteobacteria). A second large group of prokaryotic pelagic specialists was the archaeal phylum of Thermoplasmatota, which were not present in the ice, and had a consistent abundance in the water column of 2.7% (s.d. 1.5%). This group was different from other pelagic phyla in that it had a more consistent abundance throughout the year.

Figure 7.15 shows a heatmap of the most abundant prokaryotic and eukaryotic clades, with samples ordered by environment and by time, grouped by week of the drift. Seawater samples from below 100 m depth were excluded as they produced outliers which made clustering results less clear. Clusters were generated by hierarchical agglomerative clustering with the linkage function from Scipy (nearest point algorithm), using a cosine similarity metric on the clr-transformed RPM abundances. The first 3 rows of the heatmap (Micromonas, other Chlorophyta, Haptophyta) show a pattern of low abundance in the ice and in the water before June, followed by a bloom from June onwards which either dropped in the case of the Haptophyta and other Chlorophyta, or increased until August for the *Micromonas*. Verrucomicrobiota and Actinomycetota had very similar abundance patterns in the water, with a sustained increase before June, followed by a sudden drop following the eukaryotic bloom from June. In the ice, levels of Verrucomicrobiota were consistently low, whereas in contrast Actinomycetota consisted of various subclades, some of which were relatively more abundant in ice, and others water (though never both).

Bacteroidota, viruses, and Alpha and Gammaproteobacteria were ubiquitous throughout the year, though Bacteroidota had seasonal peaks in the ice (March, and June) and peaks in the water in July and November. Gammaproteobacteria were overwhelmingly abundant in the sea ice in winter (over 80% abundance in some samples). Their abundances in ice and water throughout the year tracked each other approximately, reducing from a maximum in December (85% in ice, 24% in water) to June (17% ice, 7% water), before increasing in July (33% ice, 38% water), and then remaining somewhat stable until the end of the drift.

The abundance pattern for Alphaproteobacteria was different, with a peak abundance in

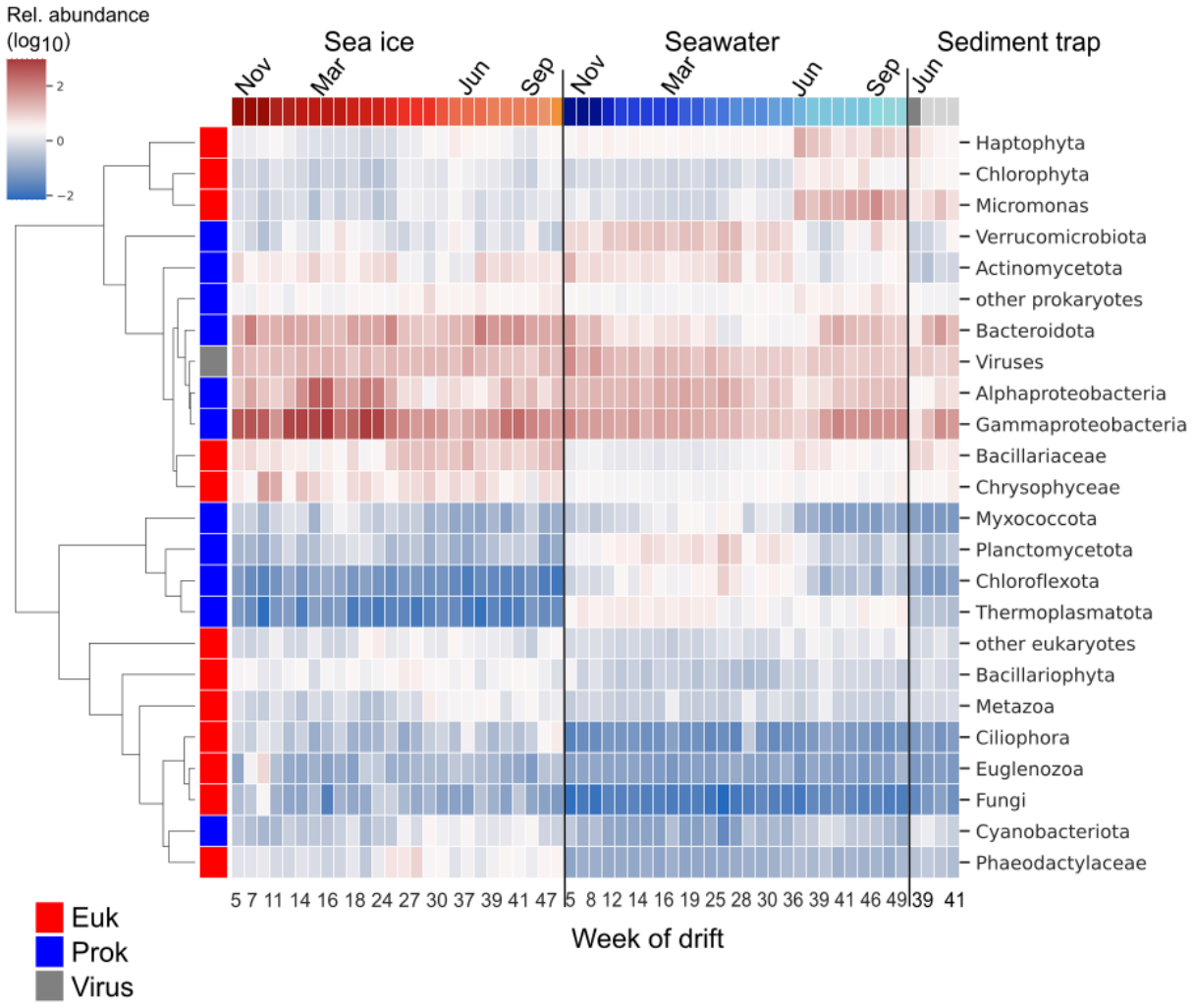


Figure 7.15: Heatmap of relative abundance (log10 RPM) of prokaryotic and eukaryotic clades, and total viruses, per week of drift. Samples are grouped by environment, and ordered by time.

the water column of 19% in February, and two peaks within the ice, once in January (19%) and then again in September (13%).

The Haptophyta were all closely related to one another on the tree and all exhibited a similar abundance profile; a large peak in abundance in June in the water (up to 25% abundant), but quickly declining back to below 5% by July, and a sustained abundance in water of 2% throughout the year. They had a much lower abundance in ice, where there maximum abundance at any time was just 2%, in June. Within the Chlorophytes, the *Micromonas* exhibited low abundance from November to June, followed by a large increase in abundance from June to September. The remaining Chlorophytes followed a third distinct abundance pattern, with a marked (but small) increase in abundance in both the ice and water from April, and then a larger bloom in June, which was sustained until July before reducing back to around 1% in August and September.

The two main diatom families present, Bacillariaceae and Phaeodactylaceae, both peaked in abundance in ice in April; for the Bacillariaceae this peak was sustained until the end of the drift, whereas for the Phaeodactylaceae, the peak abundance was in April, and then dropped in May and June, rising again until the end of the drift. Whereas the abundance of the Phaeodactylaceae was low in the water, throughout the year, there was an increase in abundance of Bacillariaceae in water from April to June, followed by a decrease until September. Other eukaryotic clades were much less abundant in either ice or water. The Metazoa were present in ice and water from April onward, though more-so in ice.

Abundance patterns for each taxon, grouped by month, and for each environment, are provided in the Appendix C.5.

Beta diversity

The plot of beta diversity (Figure 7.17) shows the ice and water samples split into two distinct groups. Beta diversity was calculated using Aitchison distances, i.e. Euclidian distances between clr-transformed abundances. All but 1 of the water samples had their first principal component greater than 0; all but 20 of the ice samples had their first principal component less than 0, and only 1 had this component above 50. These two outlying samples, i.e. the water sample in the ice cluster, and vice versa, were sample HAVOC10 (a seawater sample, from a gap in arctic sea ice ridge) and sample 446018iDtx20x30Pxxx (a melted ice core top). These two samples were fairly visible as outliers a few other plots. However, there were only 13 and 1 MAGs recovered from these two samples respectively, meaning that they barely affected any conclusion of the overall MAG analyses.

The second principal component had a strong negative correlation with time (using week of the drift, $r = -0.73$). These two components explained 48% of the variation in the

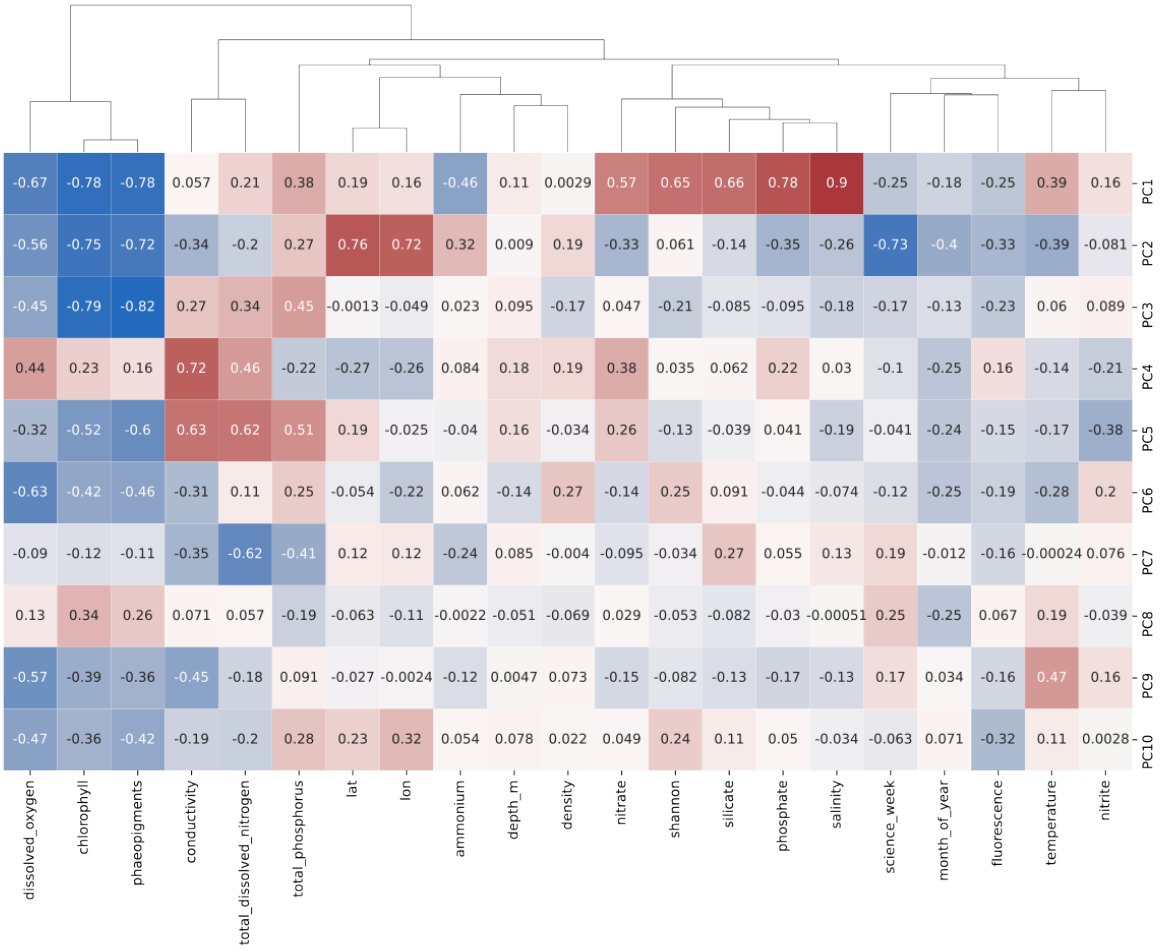


Figure 7.16: Pearson correlation coefficients between the first 10 PCoA components and the environmental parameters of the samples; the colour indicates the strength of the correlation with red positive and blue negative.

samples; the top 10 components together explained 71% of the total variation. Figure 7.16 shows a heatmap of Pearson correlations coefficients of the first 10 PCoA components, and the environmental parameters. Projections of the MAG abundances onto these two directions also showed a clear distinction between those recovered from ice and water, based on the first principal component; though otherwise the components were not particularly enlightening (see Appendix Figure C.1).

More informative was a UMAP plot of these data. Figure 7.18 shows a UMAP plot of the clr-transformed MAG abundances; MAGs represented by points, with MAGs having similar abundance profiles placed closer together. The UMAP plot clearly shows ice and water clusters, separated by the first component (umap_0). Only 125 MAGs recovered from sea ice had this component greater than 1, out of 4751 MAGs; conversely just 560 MAGs recovered

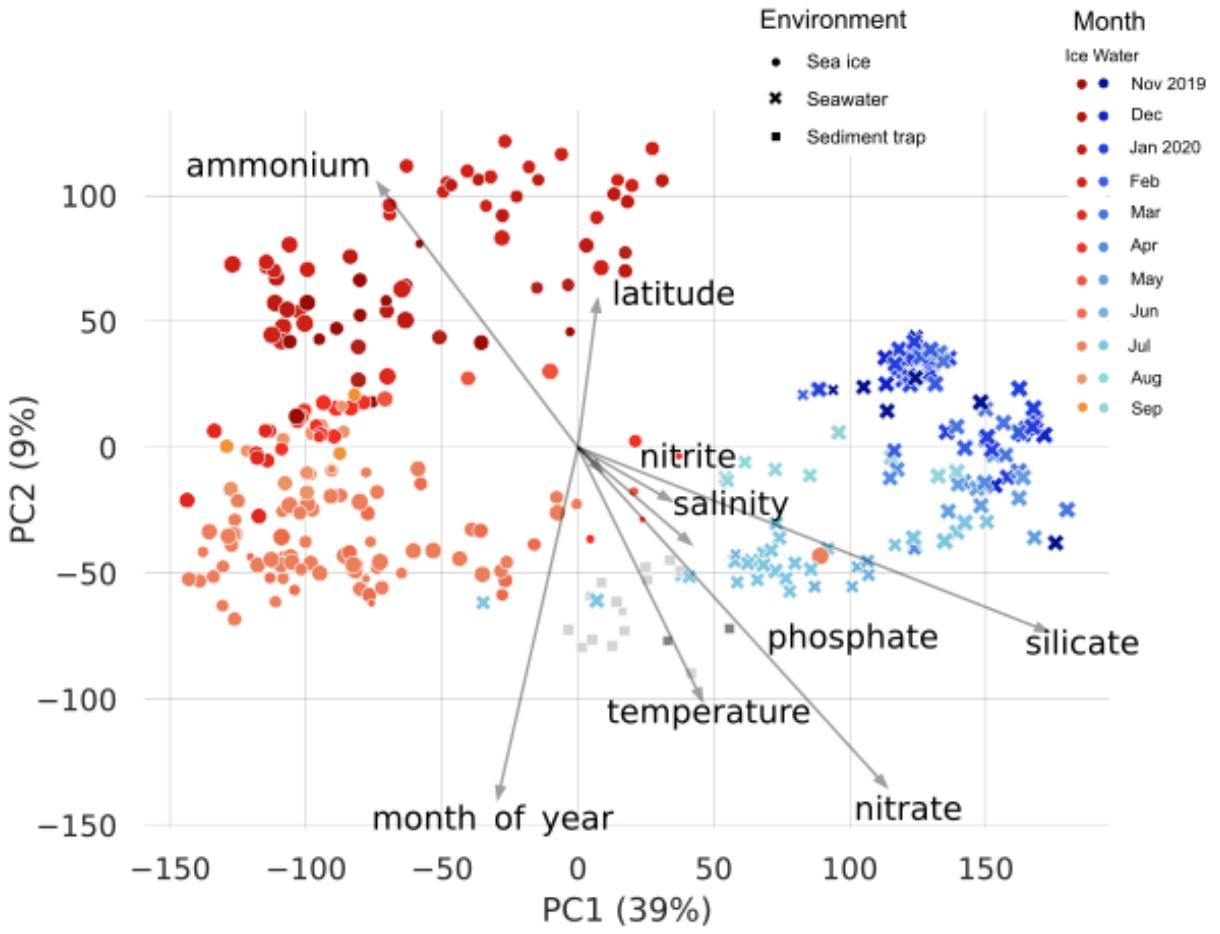


Figure 7.17: Sample beta diversity (Aitchison distance). Blue shaded crosses denote seawater samples, red circles denote sea ice, grey squares are sediment traps. Shading denotes progress in time (lighter is later in the drift).

from seawater out of 5580 were in the ice cluster (umap_0 < 1). Within the main two ice and water clusters, several subclusters had slightly different abundance profiles. Several manually annotated subclusters were marked from the main UMAP, and their abundance profiles are provided in the Appendix (C.5). Often, subclusters had a linear topology, for example clusters 1, 4, 6, 7, 8 and 9 in Figure 7.18, where points were on straight lines. The interpretation of this is that the abundances of these MAGs varies in just 1 dimension. In fact, within clusters 1, 4, 6, 7, 8 and 9, MAGs all had very similar abundance profiles across the drift, but scaled (in terms of total RPM) by differing amounts. Figure 7.19 shows the same data, coloured by taxonomy instead of provenance. Table 7.9 gives a brief overview of each one of these manually annotated clusters. There were only a few visible clusters of water-recovered MAGs in the main ice cluster and vice versa. One group of light blue MAGs within the main ice cluster consisted of 29 Cyanobacteriota, and 25 Bacteroidota, and 12 MAGs from 3 other prokaryotic phyla. Most of these MAGs were from samples annotated as chlorophyll-maximum layer seawater, from between June and September, and several were from water samples taken from gaps in ridge ice from the HAVOC project.

The main group of water-recovered MAGs seemed to have a principal axis, running from approximately the coordinate $(5, 5)$ down to the coordinate $(7.5, -4)$. This axis corresponded to seasonal distribution; MAGs near the top of the plot had decreasing abundance until June in water, followed by a massive increase in abundance from June to September (again, in water). Conversely, those near the bottom were more abundant during the winter, with peak abundance in February, and were much less abundant between July and September. These abundance profiles are provided in the Appendix, C.5.

7.4.3 Species Correlation Network Analysis

We generated a table of the correlation coefficients between each pair of MAGs in our catalogue (i.e. all prokaryotic and eukaryotic MAGs, both from this chapter and previous chapters). Species networks were generated by applying a threshold to the correlation coefficients between MAG abundances, and by the python packages SCNIC and SparCC. SCNIC also generated network modules, based on hierarchical clustering of the correlation matrices. SCNIC generated 215 modules, each of which consisted of clusters of MAGs which all pairwise positive correlations greater than 0.35.

The network modules were, in general, consistent with the results of the UMAP plot; in addition to the network modules, we also classified MAGs depending on whether they were in the ice module (UMAP0 < 1) or water (UMAP0 > 1). There were 33 SCNIC modules (out of 216) which had MAGs in both the water and ice UMAP modules, but in all these

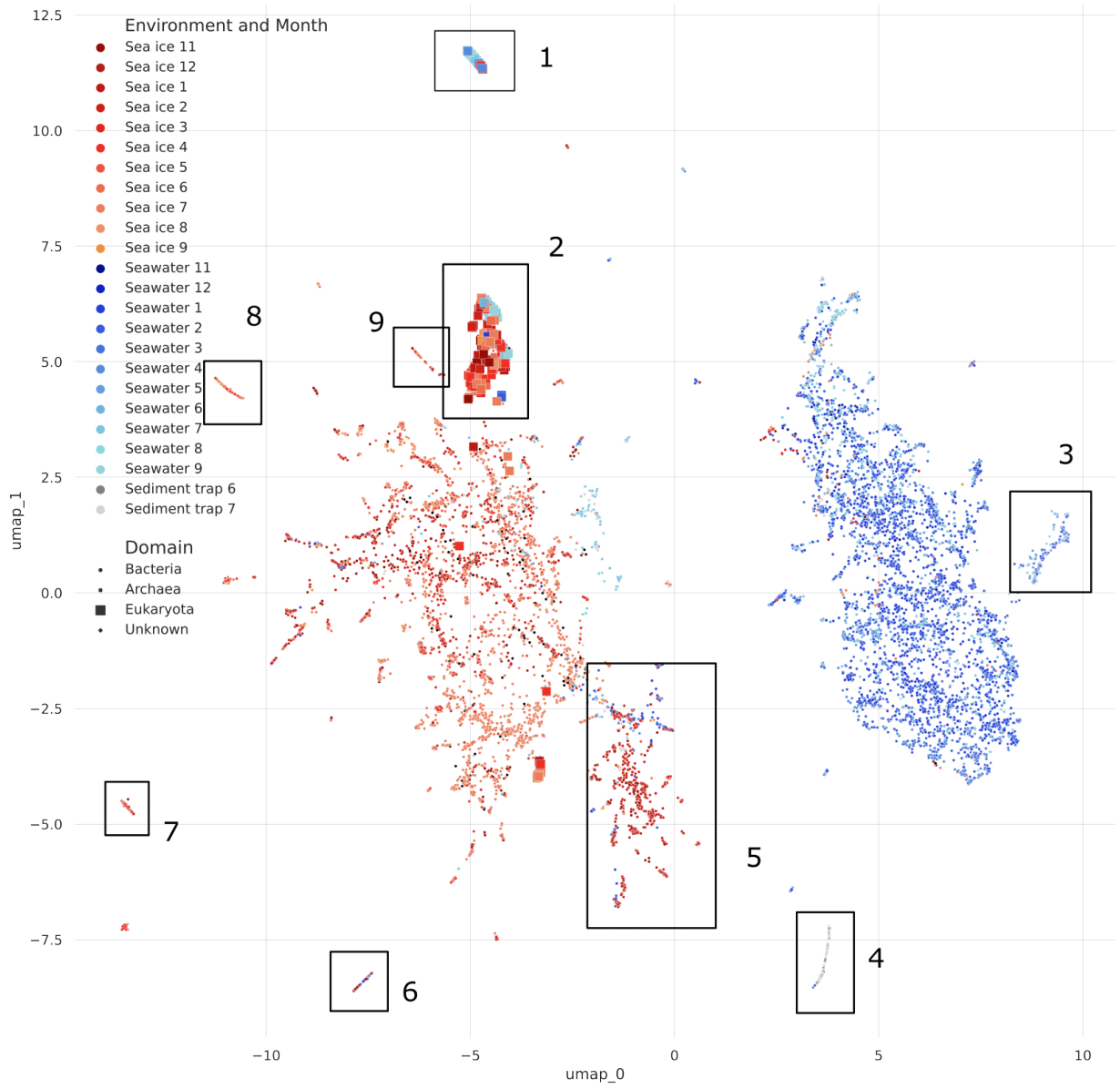


Figure 7.18: UMAP plot of MAGs based on Aitchison distances in RPM abundances. Points are ordinated using the k-nearest neighbours data ($k = 15$), based on abundance profile similarity - similar to a species correlation network. MAGs are coloured by the environment and month from which they were recovered. Eukaryotes are indicated by the larger square markers. Several subclusters (identified and annotated manually) are marked 1-9.

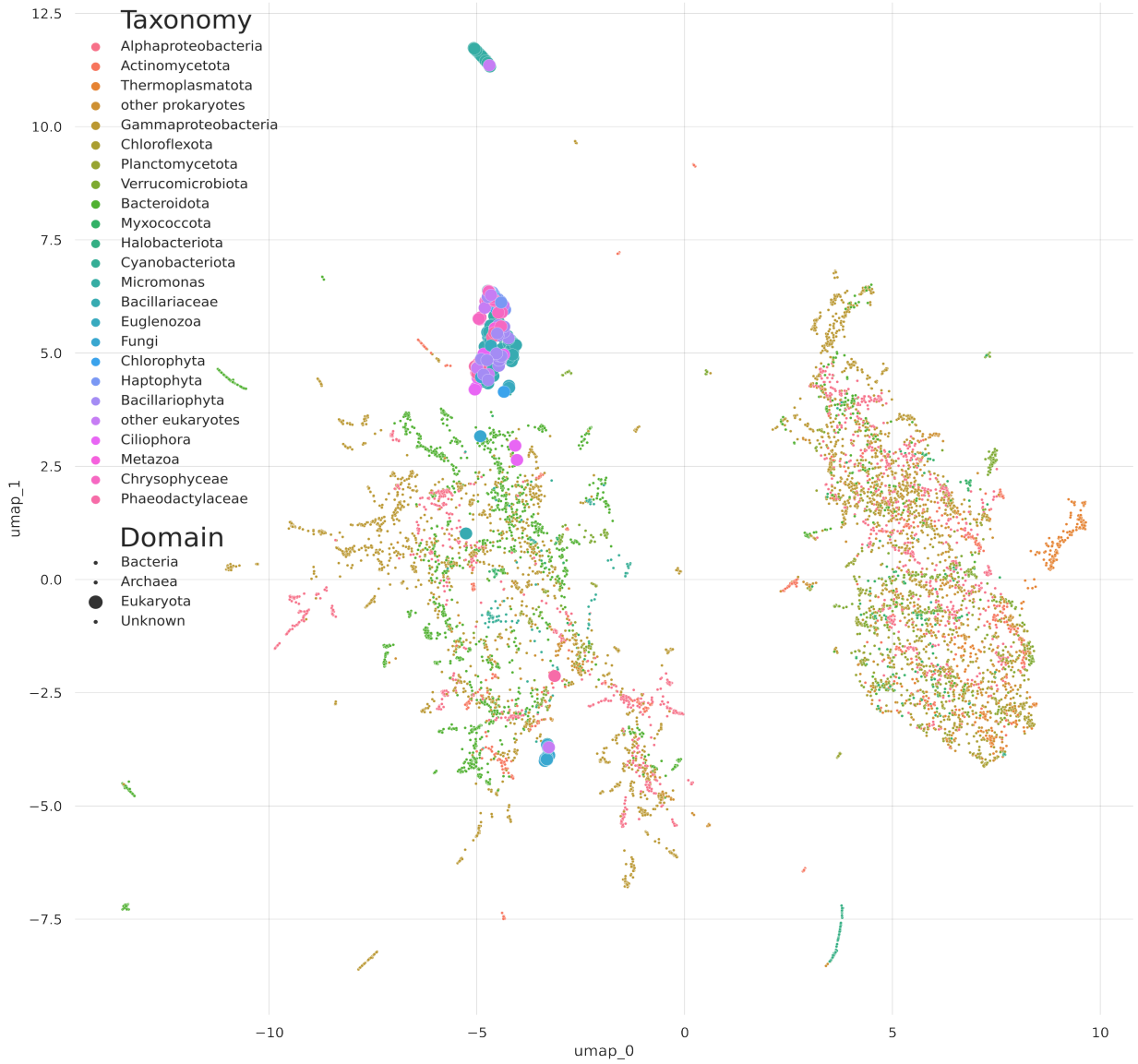


Figure 7.19: UMAP plot with identical coordinates to 7.18, coloured by MAG taxonomy (phylum). The larger markers are eukaryotes.

Cluster id	Number of MAGs	Main taxa present	Comments
1	72	*Micromonas*	1 MAG other than *Micromonas*
2	276	Bacillariaceae (111) Haptophyta (23) Phaeodactylaceae (23)	diatoms, haptophytes, and non-*Micromonas* chlorophytes
3	174	Thermoplasmatota	
4	104	Halobacteria and 27 other prokaryotes	sediment trap MAGs
5	729	Alpha- and Gamma-proteobacteria (574)	mainly Jan/Feb ice and water
6	65	Gammaproteobacteria	Mostly 1 species; *Alcanivorax sp947258575* (63)
7	86	Bacteroidota	Mostly 1 species; from family Spirosomataceae (83)
8	71	Bacteroidota	All 1 species; *Neolewinella sp026421165*
9	48	Actinomycetota and Gammaproteobacteria	Mostly 1 species; *Aquiluna sp913057795*

Table 7.9: Descriptions of the MAG clusters labelled in 7.18.

cases, the modules were all small, with a maximum of 30 MAGs, and the most common case was for a single MAG to be on the opposite side of the ice/water split compared to the rest of the MAGs in that module.

We used the SCNIC modules, and the UMAP ice/water modules, to identify two subsets of MAGs; firstly, OTUs which appeared both within the ice cluster and the water cluster (i.e. generalist OTUs), and secondly, MAGs with a cross-domain association (i.e. MAGs of different domains within the same SCNIC module).

Generalist Clades

To be considered as putatively a generalist, we required there to be more than one MAG within the ice cluster, and more than one in the water cluster. This condition was rarely met by MAGs within the same 99% ANI species cluster; we found 5 examples where this was the case, out of 3075 of these OTUs.

There were just three examples of groups of MAGs with the same GTDB species (meaning that they had 95% ANI similarity with one another, as well as the GTDB-Tk species reference genome), fulfilling the above definition of generalist. These were: *Neolewinella sp026421165* and *Croceibacter atlanticus* (Bacteroidota), and *Oleibacter sp002733645* (Gammaproteobac-

teria).

However, at the level of genus, we identified 38 prokaryotic genera that fulfilled these criteria as generalists. Only a handful of these were notable for having relatively similar numbers of MAGs within both the ice and water clusters; for most, the number of MAGs within one environment outweighed the other by a factor of more than ten. The two largest genera of these, in terms of richness, had 99 and 60 MAGs respectively; these were Pseudohongiellaceae *UBA9145* (a metagenome-derived genus), and *Halioglobus*. Both were Gammaproteobacteria. Other generalist genera were within the phyla Bacteroidota (3 genera; 85 ice-specialist; 10 water-specialist), Actinomycetota (1 genus; 5 ice; 28 water), Verrucomicrobiota (1 genus; 3 ice; 9 water), Myxococcota (1 genus; 8 ice; 2 water), Bdellovibrionota (2 genera; 50 ice; 6 water), other Gammaproteobacteria (8 genera; 120 ice; 27 water).

At higher taxonomic ranks, there were larger proportions of generalists, and it was more notable for a phylum to be specialist rather than generalist. The prokaryotic specialist phyla were already mentioned in Section 7.4.2 and Table 7.8; these were the Chloroflexota and Thermoplasmatota (both pelagic specialists).

Eukaryote-prokaryote Associations

We identified putative inter-domain interactions based on the following criteria:

- different domains - we only considered pairs of MAGs from different domains
- same SCNIC modules - we only considered pairs within the same network modules
- correlation - we further required there to be a high correlation coefficient between the two prospectively interacting MAGs (Pearson $r^2 > 0.8$).

Based on the above criteria, we identified 13 SCNIC modules where it was possible for cross-kingdom interactions to occur, and of those, we found 80 eukaryote-prokaryote pairs that had a correlation r^2 above 0.8, within 5 network modules. This included 27 eukaryotes and 29 prokaryotes. The eukaryotes were primarily diatoms, 13 Bacillariaceae (7 unclassified beyond family, 1 was *Fragilariopsis* and 5 *Cylindrotheca*), 10 Phaeodactylaceae (*Phaeodactylum*), 1 Chrysophyceae, 1 Bacillariophyta, 1 *Micromonas*, and 1 Fungi.

The prokaryotes included 17 Bacteroidota, 4 Cyanobacteriota, 4 Gammaproteobacteria, 2 Alphaproteobacteria, 1 Verrucomicrobiota and 1 unclassified prokaryote. Some of these prokaryotes were low quality MAGs; once they were removed, only 16 Bacteroidota, 2 Gammaproteobacteria, 1 Cyanobacteriota, 1 Alphaproteobacteria and 1 Verrucomicrobiota remained. The strongest cross-domain associations ($r^2 > 0.85$) were between an unknown

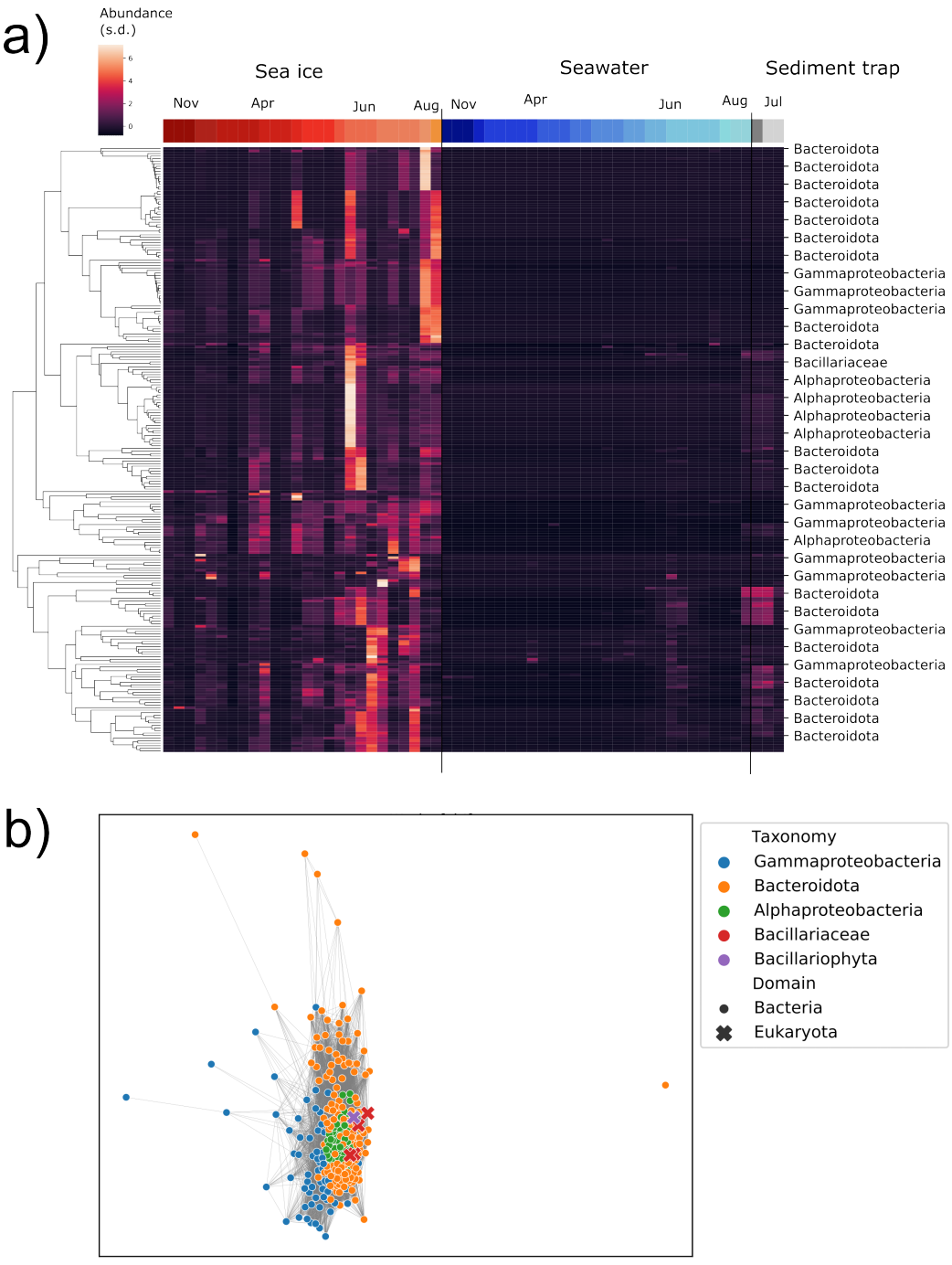


Figure 7.20: Species network module 10. **a)** Heatmap showing the abundance profiles of MAGs within the network module. Rows are scaled in terms of their standard deviation. **b)** Species correlation network module. The larger markers indicate eukaryotes, smaller markers, prokaryotes. A line indicates a correlation coefficient r above 0.5.

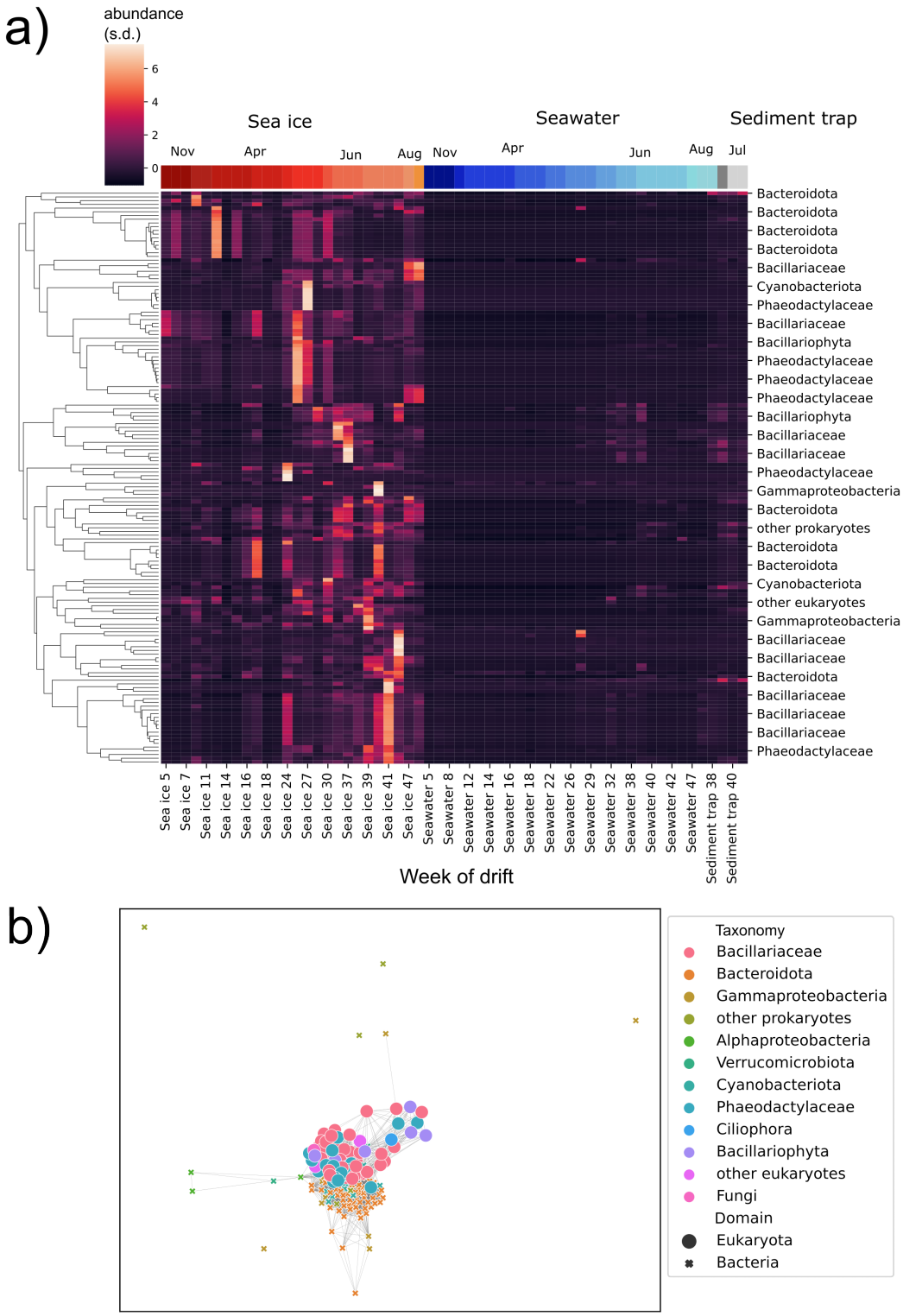


Figure 7.21: Species network module 17.

Cyanobacteriota MAG (with order PCC 6307, the same as *Synechococcus sp. PCC 6307*) and a *Phaeodactylum* MAG, and a Verrucomicrobia MAG of genus *Lentimonas*, which was associated with the diatom *Cylindrotheca*.

Figures 7.20 and 7.21 show the two largest network modules containing these eukaryote-prokaryote associations, as well as heatmaps of the species abundance patterns within each network module. The remaining three network modules containing cross-domain species associations are in Appendix C.5.1.

7.5 Results III: Gene Network Modules

7.5.1 Overview of Genes and Functional Annotations

A total of 762 million genes were predicted in the single-sample assemblies, with a maximum of 8.3 million genes predicted in one sample, and a minimum of 31000. There were a total of 24.9 million genes predicted within prokaryotic MAG, and 19.0 million within eukaryotic MAGs, with a mean of 2495 within prokaryotes (s.d. 1250) and 53605 within eukaryotes (s.d. 42840). There was a correlation of $r = 0.60$ between MAG completeness and the number of CDSs called, and $r = 0.99$ between genome size and number of CDSs.

7.5.2 DUFs and Hypothetical Proteins

Within the annotations of the single assemblies, 34% of CDSs had at least 1 Pfam domain, while 63% had no annotation (annotated with 'hypothetical protein'). There was a significantly higher percentage of unannotated proteins within ice than within water (Welch's t-test, $p = 0.0007$); in ice 63% (s.d. 13%) of proteins were annotated 'hypothetical protein' compared to 58% (s.d. 9%) in water. Of the 318 million Pfam annotations, 4.4% were domains of unknown function, varying between 3.1% and 5.9% across the samples. (We counted Pfams where 'DUF' was in the accession and 'unknown function' in the description - this ruled out a few DUF domains where a function is known, for example DUF3494, ice-binding proteins.)

Within MAGs, there was once again a higher proportion of Pfams were DUFs within sea ice MAGs than within water; 5.1% (s.d. 0.01%) compared to 4.7% (s.d. 2.7%) for MAGs recovered from seawater. The overall average percentage of Pfams annotated as DUF was 4.9%. The difference between ice and water is partly driven by differences in taxonomic composition between the two environments; the Bacteroidota, which dominated many sea ice samples, had more DUF domains than the background rate, and furthermore, had more DUF

domains in the ice subcommunity (6.2%) than water (5.2%). This was similarly the case with Gammaproteobacteria, the other large group dominating the ice; ice-Gammaproteobacteria had a percentage of 4.8% DUFs compared to 4.2% for seawater-Gammaproteobacteria. Overall, phyla with the highest proportions of DUF domains were the Verrucomicrobiota (9.4%, s.d. 3.3%), Planctomycetota (7.3%, s.d. 4.3%), Bacteroidota (6.1%, s.d. 0.8%) and Halobacteriota (5.6%, s.d. 0.7%). In the cases of the Verrucomicrobiota and Planctomycetota, there was a correlation with genome size; ($r = 0.49$ and $r = 0.71$, respectively). These two phyla both tended to have large genomes. The number of proteins within prokaryotic MAGs annotated as just 'hypothetical protein' was correlated with the number of DUFs ($r = 0.51$) and genome size ($r = 0.43$); this was highest in the Myxococcota (mean 37.5%, s.d. 8.8%), Planctomycetota (37.2%, 6.0%), Verrucomicrobiota (36.3%, 6.6%), and then Bacteroidota (31.0%, 5.5%).

7.5.3 WGCNA Modules

Our WGCNA analysis generated 21 modules, based on the clr-transformed Pfam abundances within MAGs. The largest module (blue module) contained 5145 Pfams, the smallest (dark orange) contained 55. The dark red, dark green, dark turquoise, dark grey, orange and dark orange modules each contained 100 Pfams or fewer. The blue, midnightblue, yellow, and green-yellow modules each contained over 1000 Pfams. We clustered the environmental factors, and WGCNA modules, and computed the Pearson correlation coefficients between each factor and module eigengene. The midnightblue, cyan, magenta, and pink modules all had positive correlations with ice (treated as a binary factor); these r values were 0.78, 0.7, 0.48, and 0.39 respectively. The yellow, royal blue, orange, and black modules had the largest negative correlations with ice; r of and -0.85, -0.76, -0.63 and -0.44. These same modules had positive correlations (all above 0.4) with nitrate, silicate, phosphate, and salinity; however these factors are orders of magnitude different between ice and water, so the correlations here most likely reflect this. The modules most strongly correlated with seasonality were the blue, magenta, light green, dark green and green-yellow modules. The blue and magenta modules had positive correlations with the week of drift (r above 0.45) while the remaining three had negative correlations with this variable ($r \leq -0.44$). The correlations were reversed for some other variables associated negatively with the progress in time of the drift, e.g. latitude and longitude, since the drift on the whole progressed southwards, and westerly.

All modules except for the royal blue and midnight blue modules showed a correlation above 0.2, or below -0.2, with both chlorophyll and phaeopigments (alga-associated non-chlorophyll pigments, produced by chlorophyll degradation).

There were strong links between WGCNA module abundance and the abundance of different clades. The midnightblue and pink modules had strongest correlations ($r \geq 0.80$) with Gammaproteobacteria. The blue module was positively correlated with all eukaryotic phyla, and negatively correlated with most prokaryotic phyla that were generally more abundant within water than ice; Alphaproteobacteria, Actinomycetota, Chloroflexota, Thermoplasmatota, Planctomycetota, Myxococcota, Verrucomicrobiota (from here on; pelagic-associated prokaryotes). The pelagic-associated prokaryotes were all pairwise positively correlated with the royalblue, yellow, orange, black, greenyellow and lightgreen modules, with just a single exception, and were pairwise negatively associated with the midnightblue, pink, grey, cyan and magenta modules.

Diatom clades and the Ciliophora were positively correlated with the grey, cyan, magenta, darkturquoise and blue modules, the other eukaryotic clades associated with ice (Fungi, Euglenozoa, Chysophyceae) were more strongly correlated with the grey module ($r \geq 0.39$). These modules were positively correlated with chlorophyll and phaeopigment concentrations. The other eukaryotic clades, more strongly correlated with water (*Micromonas*, other Chlorophyta, Haptophyta, Metazoa, and other eukaryotes) were on the other hand positively correlated only with the royalblue and blue modules. Cyanobacteriota were also associated with these modules. Overall, there were 6 large functionally correlated supergroups of taxa; pelagic-associated prokaryotes, ice-associated non-diatom eukaryotes, diatoms (plus the Ciliophora), water-associated eukaryotes (plus Cyanobacteriota), the Gammaproteobacteria, and the Bacteroidota. These last two constituted the vast majority of ice-associated prokaryotes, though both were also found in pelagic environments. Viruses formed a seventh distinct functional pattern, positively correlated with the same modules as the pelagic-associated prokaryotes, and the dark turquoise and blue modules that were otherwise associated with eukaryotes. This may be due to at least two separate clades of viruses, infecting diatoms, and separately, other eukaryotic microalgae.

As well as correlations between clade abundance and modules eigengenes, we also looked at the relative abundance of each module within each clade (Appendix, Figure C.2). This showed that some phyla had a high relative abundance of Pfams within particular modules: for Cyanobacteriota, the orange (82%) module, for viruses, dark turquoise (67%), for Fungi, grey (13%), for Metazoa, blue (34%), for Bacteroidota, magenta (14%), for Alphaproteobacteria, lightgreen (10%) and for Actinomycetota, lightyellow (8%). In each case, the enrichment within that module for that phylum was much larger relative to the enrichment for all other phyla.

For each module, we looked at the GO terms which were enriched, using a binomial test, and used a p-value of 0.001 to test for significance. The blue module had a significantly

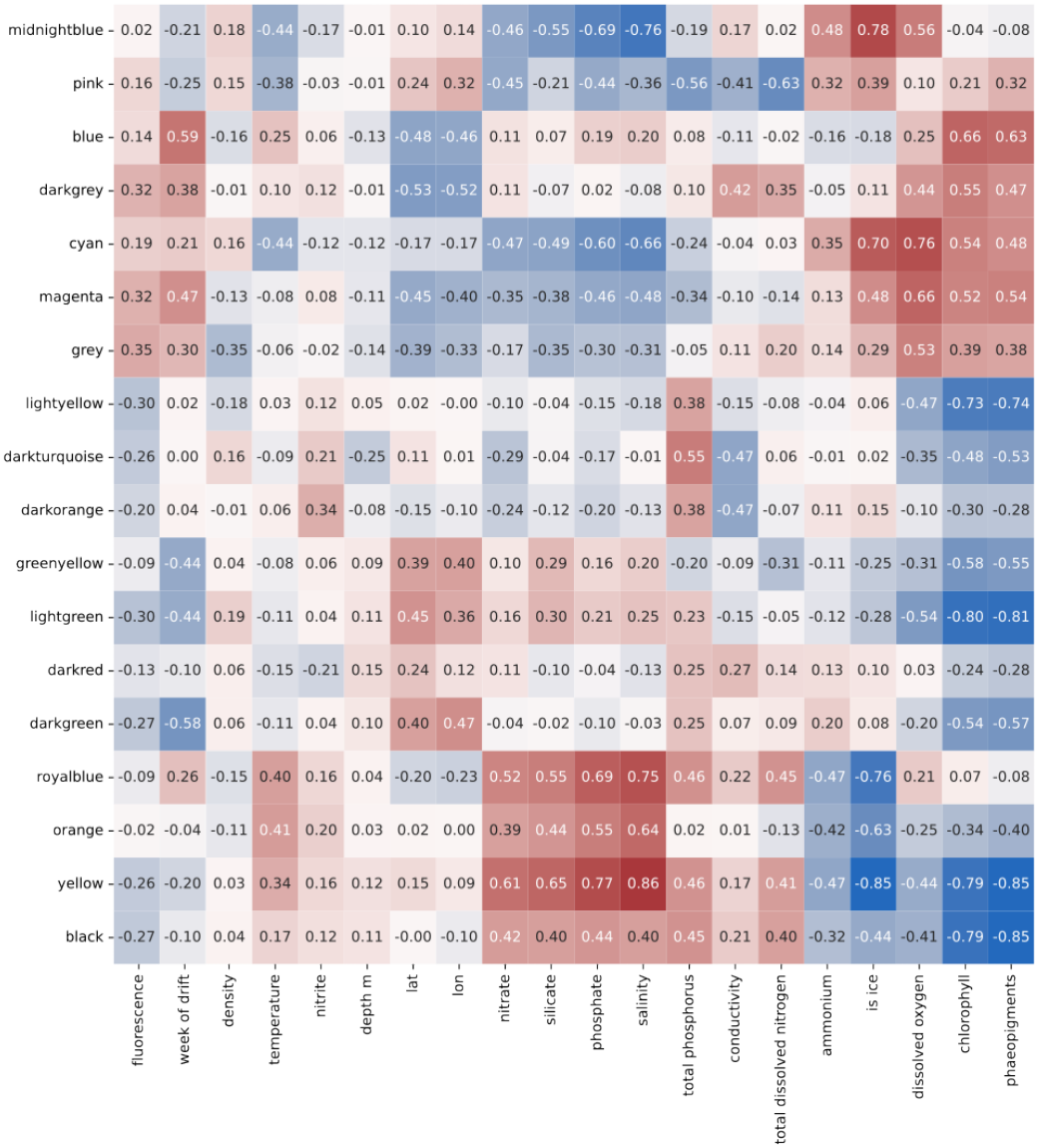


Figure 7.22: Pearson correlation coefficients between WGCNA module eigengenes and the sample physical parameters. Parameters and WGCNA modules have been hierarchically clustered, but the dendrograms are omitted for legibility.

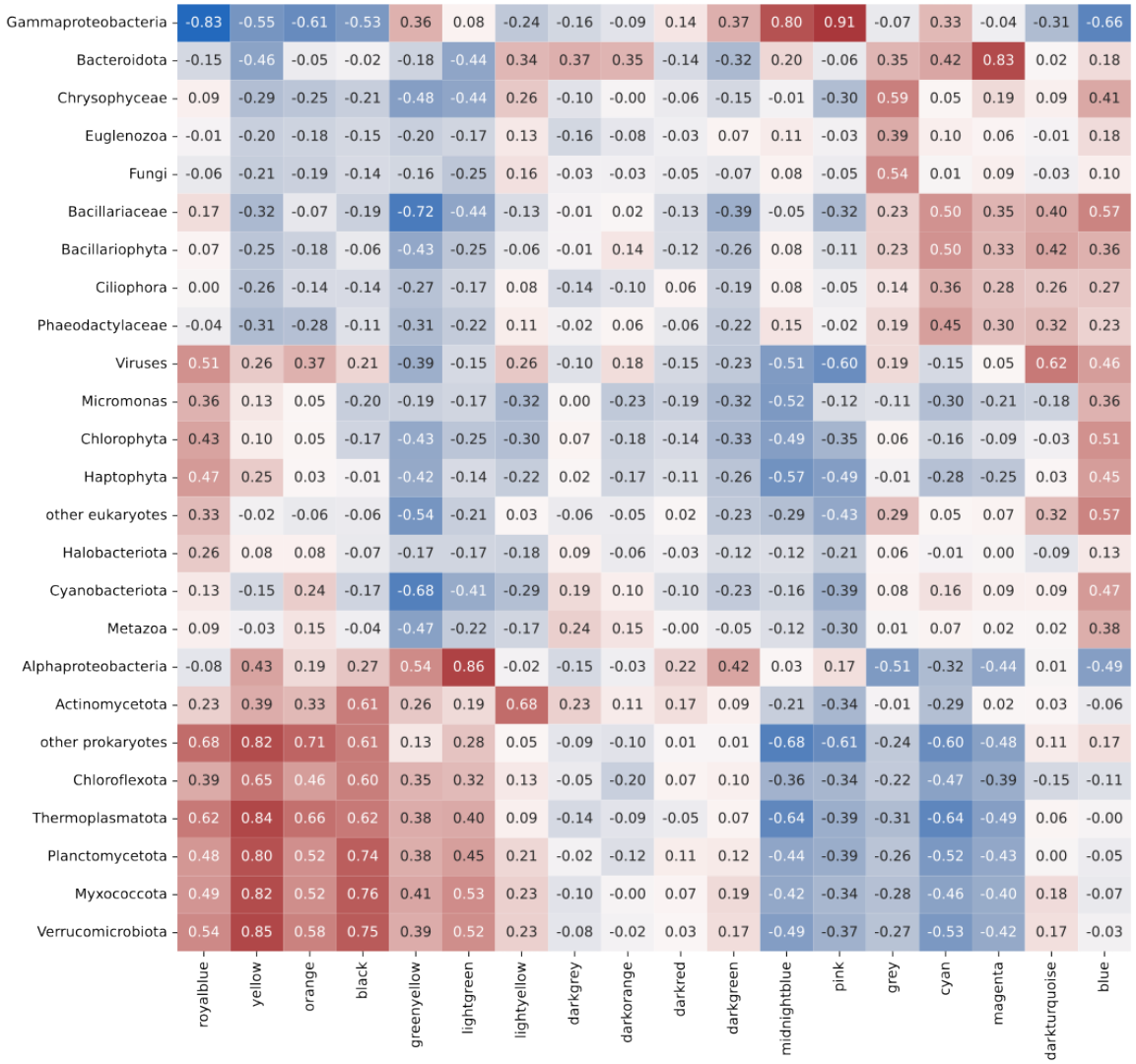


Figure 7.23: Pearson correlation coefficients between WGCNA module eigengenes and the abundance of different clades.

higher proportion of photosynthesis GO terms, and terms for vesicle mediated transport. The midnight-blue module, which was correlated with Gammaproteobacteria, had enriched terms for transposase activity, signal transduction, and chemotaxis. The second module linked with Gammaproteobacteria (pink), contained terms for type II secretion systems, and sodium ion transport. The magenta module, strongly correlated with Bacteroidota, contained terms related to sugar hydrolases and carbohydrate metabolism, spore germination, but also several other terms including bioluminescence (though with $p = 0.00337$). The royal blue and orange modules were both highly enriched with ribosome-related terms. The orange module almost solely contained terms for rRNA and translation. This module was abundant within all phyla except for the viruses, and was particularly abundant within the Cyanobacteriota. The light-green module, associated with Alphaproteobacteria, was enriched with terms for cobalamin synthesis, and aromatic compound metabolic processes. For the remaining modules, the top five terms with the smallest p-values (less than 10^{-3}) are listed in Table 7.5.3.

Module	GO Term	p-value
blue	protein binding	0.0
	nucleus	0.0
	photosynthesis	0.0
	intracellular protein transport	2e-05
	vesicle-mediated transport	2e-05
midnightblue	transposase activity	1e-05
	phosphorelay signal transduction system	3e-05
	chemotaxis	6e-05
	transposition, DNA-mediated	7e-05
	endonuclease activity	0.00017
yellow	oxidoreductase activity	0.0
	NADH dehydrogenase (ubiquinone) activity	7e-05
	IMP biosynthetic process	0.00072
	one-carbon metabolic process	0.00072
greenyellow	ATP binding	0.0
	nucleotide binding	0.0
	aminoacyl-tRNA ligase activity	0.0
	catalytic activity	1e-05
	tRNA aminoacylation for protein translation	2e-05
magenta	hydrolase activity, hydrolysing O-glycosyl compounds	0.0

	carbohydrate metabolic process	4e-05
	spore germination	0.00022
black	modification-dependent protein catabolic process	0.00052
	phosphoenolpyruvate carboxykinase activity	0.00052
	proteasomal protein catabolic process	0.00052
pink	protein secretion by the type II secretion system	2e-05
	type II protein secretion system complex	3e-05
	sodium ion transport	0.00031
	oxidoreductase activity	0.00087
royalblue	ribosome	0.0
	translation	0.0
	DNA-templated transcription	0.0
	structural constituent of ribosome	0.0
	DNA-directed 5'-3' RNA polymerase activity	0.0
cyan	ribonuclease activity	0.00063
grey	DASH complex	0.0
	attachment of spindle microtubules to kinetochore	0.0
	proton motive force-driven ATP synthesis	4e-05
	mitotic spindle	5e-05
	spindle microtubule	0.00019
lightgreen	cellular aromatic compound metabolic process	0.00045
	cobalamin biosynthetic process	0.001
lightyellow	phosphoenolpyruvate-dependent sugar phosphotransferase system	0.0
	protein-N(PI)-phosphohistidine-sugar phosphotransferase activity	0.0
darkred	cell wall macromolecule biosynthetic process	2e-05
	glycosyltransferase activity	0.00043
darkgreen	cell adhesion	0.0008
darkgrey	phycobilisome	1e-05
	chloride ion binding	0.0009
	light absorption	0.0009
orange	translation	0.0
	ribosome	0.0

	structural constituent of ribosome	0.0
	DNA-directed 5'-3' RNA polymerase activity	0.0
	DNA-templated transcription	6e-05

We searched for the enriched GO terms within MAGs, limiting to only those with a mean prevalence of at least 10 search term hits within any phylum. As expected, there was a set of core terms, such as ribosome, membrane, and translation, which were universally present. Photosynthesis terms were present only in eukaryotes and the Cyanobacteriota. Only transposase activity and DNA-mediated transposition could be described as uniquely prokaryotic, excluding Metazoa. Protein binding was much more prevalent within the eukaryotes than prokaryotes, as was (unsurprisingly) terms related to the nucleus, organelles, or vesicles. Carbohydrate metabolism terms were prevalent within both prokaryotes and eukaryotes, but moreso in eukaryotes, and in particular, the term 'hydrolysing O-glycosyl compounds' was prevalent in all phyla but particularly the Haptophytes. This is linked with the breakdown of cellulose and starch. This potential capacity for heterotrophy might be an adaptation for overwintering, or, in the case of the Haptophyta, which bloomed during the melt season, it might be used to take advantage of particulate organic carbon released from melting ice. Most clades, except for the Archaea, Cyanobacteriota, and two eukaryotic phyla, had an abundance of phosphorelay signal transduction systems, and in most cases, this was more prevalent within ice than water. The term 'cell adhesion' was found in all eukaryotic clades except *Micromonas*, Chrysophyceae, and Fungi, but was also found, enriched within the prokaryotic ice-associated clades (Bacteroidota and Gammaproteobacteria), and was most abundant in Myxococcota. In ice, cell adhesion could be important to remain attached within brine channels, and in water, to the underside of the sea ice. Myxococcota were enriched with multiple GO terms listed in Figure 7.24, such as sodium ion transport and ice-binding, particularly those species associated with ice.

Figure 7.24: GO terms and GO term enrichments when comparing ice and water. The size of the marker denotes the mean prevalence of that GO term within the clade (counts). The marker colour denotes enrichment comparing ice to water; red indicates a higher enrichment within ice, blue for water, grey indicates a ratio of 0.5 between ice and water. The GO terms chosen are those enriched within WGCNA modules with a p-value less than 10^{-3}, and also a mean prevalence within MAGs of at least 10 in any one phylum.

7.6 Discussion

Our analysis of Arctic microbiomes has uncovered a large number of novel species and genes, and substantially improved the coverage of MAG catalogues for the Arctic, especially for Arctic eukaryotes. While MAG analyses are not as sensitive at detecting species as 16S or 18S rRNA gene sequencing, the availability of genomic information is key when attempting to go beyond a purely correlative study. These expanded sets of genes may also be invaluable for data mining, where we just scratched the surface when exploring the data available, and would be particularly useful when searching for cold-adapted homologues to existing enzymes. MAGs do not perfectly capture microbial diversity, and the somewhat arbitrary quality constraints on completeness and contamination make recovering some important components of the community virtually impossible (e.g. dinoflagellate genomes, up to 200 Gbp in size, can probably never be captured through short-read metagenomics, even though we found that this clade did constitute a reasonable fraction of eukaryotic contigs). However, this study has demonstrated that MAGs are a valuable resource for understanding both taxonomy and function simultaneously, and that this can be done for both the prokaryotic and eukaryotic community. We found examples of large prokaryotic genomes, especially in the phylum Planctomycetota (three were over 10 Mbp), similar to [56], and found that these larger genomes had a higher proportion of DUFs and hypothetical proteins. This is consistent with [421], and we could hypothesise that smaller more streamlined genomes may have retained only the 'core' proteins that are better characterised.

While we were able to make putative inter-kingdom connections between species, these were left as somewhat speculative and based on a correlation network analysis. There were several observed strong correlations between diatoms and Bacteroidota MAGs within the same network modules. In the 'phycosphere' (the zone of microbial influence around algal cells), Bacteroidota are abundant heterotrophs that degrade algal-derived polysaccharides and other organic matter. The seasonal co-occurrence of diatom and Bacteroidota abundances in the MOSAiC dataset (both peaking during the spring and summer bloom periods in ice in the co-occurrence network modules we studied) is consistent with this ecological relationship. An analysis of the Bacteroidota MAGs co-occurring with diatom MAGs for carbohydrate-active enzymes targeting chrysolaminarin (a diatom-specific carbohydrate storage molecule) would provide stronger functional evidence for metabolic coupling between these groups. Additionally, Cyanobacteriota MAGs were disproportionately over-represented in our cross-domain correlation networks. These might be worth further investigation since cyanobacteria are thought not to be particularly present in sea ice, and their reduced genome size could be an indication of symbiosis. Cyanobacteriota and Metazoa were closest together

in terms of species co-occurrence, and had very similar correlations with WGCNA modules. Analysis of complementary metabolic pathways within pairs of putatively symbiotic MAGs could provide stronger evidence for species associations. We also omitted an investigation into viruses - though these may play an important role, particularly in regard to species interactions and co-evolution.

Our WGCNA and GO term enrichment analysis suggested some terms which could be used as a basis for trait-based models, such as HMSC; alternatively, the WGCNA modules themselves could be considered as traits. This narrows down the number of variables that could conceivably be used as traits, such as photosynthesis, chemotaxis, spore germination, cobalamin biosynthesis, cell adhesion, and carbohydrate metabolic processes. Other important processes such as ice-binding and nitrogen fixation were not enriched in one module in particular but could also be added as traits.

Chapter 8

Discussion and Future Work

8.1 Overview

In this thesis we have developed and applied bioinformatics pipelines, and used the resulting MAG data to analyse the microbial diversity captured by MOSAiC, including a case-study of one particular gene family.

In Chapter 4 we generated catalogues of MAGs from diverse Arctic environments, and used data from the pilot study and the HAVOC sub-project to begin exploring prokaryotic species. We found several distinct clades of prokaryotic pelagic and sympagic specialists in the Arctic winter, providing some evidence of environmental filtering of those groups. Other groups seemed more generalist; clades of Gamma- and Alphaproteobacteria were recovered across all sample types, including the deep ocean. We also used these first data to begin generating eukaryotic MAGs, testing various assembly and binning strategies. This generated a large set of eukaryotic MAGs, mostly diatoms and *Micromonas*, but also encompassing several fungi, and one euglenozoan. One of these, *Bacillariophyceae sp.* MOSAiCH1_1, was of sufficiently high quality to be included within the algal resource Phycocosm.

Chapter 5 provided a case-study into one influential family of proteins, namely IBPs. We found diverse protein and gene architectures, and mechanisms for generating this diversity, as well as evidence of a complex phylogenetic history, suggesting potentially a large number of horizontally transferred IBPs between prokaryotic species.

Although we successfully recovered large numbers of eukaryotic MAGs in Chapter 4, we realised that repeating the same methods again on a much larger set of samples would run into logistical difficulties. We therefore developed a pipeline that was capable of processing a much larger number of samples, and used a novel visualisation method to augment the automated binning algorithms. UMAP and t-SNE have been a hugely influential tool in machine learning, but has only recently become popular within biological and ecological sciences, and even then, often restricted to particular subdomains such as single-cell sequencing. In this thesis we hope to have shown the utility of these data visualisation methods in microbial ecology, even if using them involves certain trade-offs in terms of being less directly interpretable, compared to classical dimensionality reduction such as PCA.

In Chapter 7 we demonstrated the utility of our new pipeline by generating almost 240 eukaryotic MAGs, as well as nearly 10,000 prokaryotic MAGs. Using this catalogue of MAGs, we were able to analyse species α and β diversity across the MOSAiC drift. We found a large number of novel taxa, up to the level of family, but also hundreds of novel orders and species of prokaryotes. We saw clear shifts in the community composition due to the return of the light, and the melt season, and identified the clades associated with each change, as well as species associated with the ice, water and sediment traps. Interestingly, we found cyanobacteria species in the sea ice, which until now had rarely been reported, and found these same species within otherwise eukaryotic species co-occurrence networks. Eukaryotes seemed to drive changes in the community composition, with the most abrupt increases in abundance due to the eukaryotic summer bloom during the melt season.

8.2 Future Work

There are multiple different routes that could be chosen to extend the work from this thesis; below, we discuss the most important avenues to pursue, as well as some low-hanging fruit for future development.

8.2.1 Improving Binning of Eukaryotic MAGs

While generating eukaryotic MAGs in Chapters 4 and 6, we found that multiple binning programs were optimised for generating prokaryotic MAGs over eukaryotes. This was most pronounced in programs which used hits to a set of SCMGs to recluster and iteratively refine binning results, such as MaxBin or COMEBin [200], [422]. Furthermore, consensus programs such as DASTool [423] used the same set of prokaryotic marker genes to generate an optimised set of bins, given an overlapping set of candidates. Using the same set of marker genes to both search for MAGs and then evaluate their quality seems dubious; ignoring this issue, it certainly biases against generating eukaryotic MAGs. If using a marker gene based approach both for binning and quality estimation is seen as acceptable, then an easy modification to most programs would be to allow the use of a eukaryotic marker gene set, such as the OrthoDB10 Eukaryota genes used in BUSCO [424]. Our own implementation of a modified version of DASTool, which used a smaller set of protist marker genes developed by Tom Delmont [425], can be found on github, and provides a proof of concept for this idea. Another extremely simple method is to run parameter sweeps of each binning tool, and provide the sets of input parameters which work best to produce both eukaryotic and prokaryotic MAGs. This is not currently done and most binning benchmarks focus just on

prokaryotic MAGs.

8.2.2 Autoencoders, UMAP and t-SNE for MAG Visualisation

In Chapter 6, we found that a combination of the variational autoencoder in VAMB and UMAP, could be used to visualise contigs and then classify them into bins. Similarly, the t-SNE plots in Chapter 4 were used as a way of surveying the functional diversity of all prokaryotic MAG at once. However, at present a limitation of these methods is that unless two datasets contain a large common core of identical points, the two corresponding visualisations will not be comparable. This means that each new analysis must combine all previously visualised units (either MAGs, or contigs) into a single dataset, and then re-run the method from scratch. It is desirable for the method to be more consistent, so that new data can be compared with old data in a like-for-like manner. A similar problem is apparent in single-cell studies. In single-cell sequencing, UMAP plots are quite routinely produced to visualise transcription data within 'Cell Atlases' [426]; but currently, no two Cell Atlases are directly comparable, even if the underlying data are themselves comparable. This problem can be solved with a method called Parametric UMAP [427], which uses a neural network to learn the relationship between the data and an embedding, and is one avenue for further investigation.

8.2.3 Eukaryotic Pangenomics

The dataset of Arctic eukaryotic MAGs that we generated constitutes by far the largest compendium of Arctic genomes to date. These data will be useful for comparative genomics, such as a comparison of Arctic and Antarctic strains, or studying the phylogenomic diversity of globally distributed genera such as *Pseudo-Nitzschia*. A *Micromonas* pangenome is currently being generated by Alan Kuo at the JGI, and similar methods might be usefully applied to the *Fragilariopsis* and *Pseudo-Nitzschia* genomes that we recovered. This might also be useful for improving species reference genomes; the *F. cylindrus* reference genome CCMP1102 is known to have some irregularities in its genome (e.g. triploidy), not representative of individuals of that species more widely. Whereas *F. cylindrus* is found only at the poles, *Pseudo-nitzschia* species are distributed globally, and generating a pangenome in this case, incorporating MAGs from Alexander *et al.* [260] and Delmont *et al.* [248], may more clearly identify different ecotypes, contributing to our understanding of cold adaptation in diatoms.

8.2.4 Further Work Analysing MOSAiC Metagenomes

Data Mining Novel Genes

One possible application for the MOSAiC data is to use it as a source for novel genes and gene clusters. Biosynthetic Gene Clusters (BGCs) are an important class of genes for bioprospecting; BGCs are used to synthesise secondary metabolites, including many antimicrobial peptides. These gene clusters can be searched for, using the program AntiSMASH [428], grouped using BiG-SLiCE [429], and deep learning tools used to predict which secondary metabolites are indeed antimicrobials [430], as was done in Chen *et al.* [56]. In that study, over 600 novel BGCs were found to contain antimicrobial peptides from a total of over 60000, from pelagic environments. However, few if any samples from that study included sea ice. Studies of the Arctic have focussed either on particular isolates (e.g. [431]), or have been much smaller in scale (e.g. [432]), but still found numerous novel BGCs, highlighting the potential for Arctic metagenomes in bioprospecting. In Paoli *et al.* [433], polar regions were associated with a higher abundance of BGCs.

Additionally, the MOSAiC data can be used to find homologues of existing genes. An initial investigation, led by Yao Xiong (current PhD student at UEA), into serine β-lactamases (an antimicrobial resistance gene, or ARG), revealed SBLs in almost every single prokaryotic MAG in our catalogue. Certain modules in our WGCNA analysis, associated with the sea-ice interface, also exhibited a statistically significant enrichment of toxin-related Pfams. The enrichment of ARGs within these MAGs suggests that there could also be an enrichment of antimicrobials; this is an area that seems worth exploring further, as existing antibiotics are dwindling due to selective pressures from their overuse in a clinical and agricultural setting [434], [435].

Beyond potentially novel antimicrobial peptides, another direction for study is data-mining for cold adapted enzymes, with CRISPR-associated proteins (CASs) being one example. We found very few CAS enzymes in our dataset, and CRISPRs have been found to be less abundant in cold-adapted prokaryotes compared to thermophiles, and those living in warmer water [56]. However, those that are there may still be interesting and biotechnologically useful due to potential cold-adaptation.

Arctic Biogeochemistry

In addition to data-mining novel genes, the MOSAiC dataset can also be used to better understand the biogeochemical processes that are catalysed by well-studied biochemical pathways. While some of the MOSAiC biogeochemistry data has yet to be released (such as

DMSP concentrations), we have data on several components of the nitrogen cycle for both sea-ice and seawater. These data can be linked with the abundances of key nitrogen-related genes, both within the MAGs, and all contigs in general. In particular, the presence and abundance of the *nifH* gene (encoding the nitrogenase iron protein, a marker for diazotrophy) can be used to estimate the potential for nitrogen fixation across ice and water communities, and nitrification and anammox maker genes encoding ammonia monooxygenase (*amoA*) and hydrazine synthase (*hzsA*, a marker for anammox) can reveal environmental and temporal patterns of nitrification and anaerobic nitrogen removal in the Arctic. Some of these genes are relatively rare and are possibly found in only a handful of MAGs - in those cases, studying the contigs overall for these genes may be more fruitful, with limited MAG data used where available to improve the links between gene trees and taxonomy. When measurements of DMSP and DMS in the sea-ice and seawater are eventually released, correlating these with the abundances of DMSP-biosynthesis / DMSP-degrading genes (e.g. *dsyB* and *dmdA*, respectively), and with clades known to be DMSP producers and degraders, could provide some insight into the sulphur cycle in the CAO.

A second biogeochemically important question in the Arctic ocean is the mechanism by which phytoplankton quickly bloom in the spring after a prolonged period of overwintering. Our MAG collection has a large number of photosynthesising eukaryotic algae, and correlating their abundance with overall chlorophyll-a levels, with key marker genes (*psaA* for photosystem A, *rbcL* for RuBisCo) and transcripts, and with the light field data, may provide a partial answer as to how these photoautotrophic algae remain alive and 'primed' for photosynthesis over the long polar winter.

Viruses

Viruses are known to be extremely important within pelagic ecosystems, responsible for a 'viral shunt' which may cause an overturning of up to 25% of the microbial community per day [436]. However, our analysis of viruses was quite limited; we retained some viral contigs in our MAG catalogue, but we did not measure their diversity and they were excluded from any analysis of species diversity. Even so, we found that the abundance of viruses ranked third, compared to eukaryotic and prokaryotic phyla. Current PhD student Derui Song is analysing nucleocytoplasmic large DNA virus (NCLDV) MAGs (a supergroup of the 'giant viruses'), many of which infect algae, such as the Phycodnaviridae [437]. Preliminary results based on his work and from the JGI have already found a large diversity of these viruses in the MOSAiC data, and our own more limited binning effort found that Phycodnaviridae were the most abundant family of viruses in the metagenomes we studied.

Transcriptomics

MOSAiC collected a set of 130 metatranscriptomes from the surface ocean, which we mapped to our MAG catalogue in Chapter 7. However, we did not have time to analyse these data; we focussed only on metagenomes. Our preliminary work, mapping the MOSAiC transcripts to MAGs, should help refine the taxonomy of some of these transcripts, and furthermore can provide insight into which genes are over- or under-expressed at which times of year, at least in the surface ocean. This could help understand which genes are functionally important (and at which times), and so identify a smaller number of influential genes in the Arctic, which can then be studied in more detail, for example by incorporating them into trait-based models within a MAG analysis. Furthermore, a database combining a non-redundant set of the MOSAiC transcripts with MMETSP would immediately be useful for annotating eukaryotic MAGs (e.g. with MetaEuk [266]), especially when gene-calling Arctic eukaryotic genomes.

Light and Melt

Though we discussed major static environmental factors affecting gene and species distributions (depths and types of ice, and pelagic layers), we did not analyse the two major dynamical effects in the Arctic in much depth; light and melt. Their effects were obvious when investigating the temporal changes in species distributions. The onset of light was from March / April, and coincided with large increases in photosynthetic diatoms, in the ice. By contrast, the melt season was between June to September, and was accompanied by rapid increases in Chlorophyta and Haptophyta. The analysis of MOSAiC data was spread across several teams of researchers, and a unified PAR data product (synthesising data from the various available light sensors from MOSAiC) has not yet been published. It was unfortunate that some of the PAR meters on the CTD were reportedly broken, and seemed to produce somewhat meaningless results. This was a substantial limitation of the MOSAiC sample metadata. Light levels were captured by the MOSAiC teams, but in other formats; two PAR datasets are OptiCALs [438]–[440] and the LightHarp [419], [441]. In Hoppe *et al.* [111], the authors utilised these two data sources, and incorporated a model of radiative transfer for snow depth, which was applicable to biological samples from all depths, and in ice and water. To quantify how changes in light affected the microbial community, it would make sense to reproduce this methodology for the MOSAiC metagenome samples. Niels Fuchs (University of Hamburg) and Bonnie Light (University of Washington) are two points of contact within MOSAiC who are producing a data product of this kind. The melt season is less problematic to identify since new meltwater can be inferred through changes

in salinity and temperature. We did not incorporate samples from melt ponds (they were not available until after we had finished analysing the largest cohort of samples); this was an important yet short-lived environment, and it would be interesting to compare the community composition of melt ponds to sea ice and to surface ocean before, during, and after the periods of ice-melt.

Modelling Time-Dependent Network Evolution

In our analysis we constructed a single species co-occurrence network, and through this we were able to identify parts of the network which seemed associated with seasonality, particularly in the water column. We also found strong correlations between various phyla and time, and correlations between the time and some of the principal covariates in our analysis of β diversity. We clearly saw the effects of seasonality, and of the time-series structure of our data, however future modelling work will need to ensure that the time-dependence of the data is accounted for. An overview of several methods, building on the tools in Section 3.8 is given in [442] and [443], and more modern methods, such as Gaussian process modelling, are implemented in Mefisto [238]. Time-series analysis is also addressed within the HMSC framework (next section) [444].

8.2.5 Hierarchical Modelling of Species Communities (HMSC)

While we were able to identify covariates in our MAG analysis, for example examining the strength of the correlations between eukaryotic subpopulations and the melt season, we did not build a model relating each abiotic factor to the community composition, or incorporating species co-occurrence, and genomic content. Models that do incorporate these factors together are known as trait-based joint species distribution models (JSDMs). Trait-based JSDMs of phytoplankton often use cell size as a 'master trait' [445]–[448]; cell size is linked to genome size [449], [450], but focussing on this one parameter risks missing all the functional traits (e.g. nitrogen fixation, ice-binding, photosynthesis) which feedback into biogeochemical cycles or have an interesting community function. We looked for alternatives where other traits derived from genomic data could be included.

One promising method is HMSC [451], a statistical method from community ecology which tries to infer biotic and abiotic filtering in communities of species. HMSC uses species associations, environmental covariates, as well as information from the phylogeny and species traits to build a hierarchical Bayesian model of species-to-species and environment-to-species interactions. Sampling units can be a time-series, or vary spatially, or both. The inputs for this model are a phylogenetic tree, a matrix of species traits, environmental measurements

per sampling unit, as well as species abundances per sampling unit. From these data, the model will try to infer species niches; combinations of environmental factors measuring the influence on that species. Species interactions are modelled by looking at different species' covariance after controlling for the environmental covariates. Current PhD student Yao Xiong is currently experimenting with this modelling framework, using the data from Chapters 4 and 7. This method, which uses Monte-Carlo Markov-Chain (MCMC) to estimate model parameters given a set of inputs, can be computationally expensive as the size of the input data grows. However, adaptations of the method, allowing it to run in an HPC environment, have already been developed [452]. The aim of HMSC is to understand which factors affect the microbial diversity in both ice and ocean, and relate those to features of the data such as species traits, in a more causative manner than WGCNA, which simply identifies groups of correlated variables such as species, traits, genes, or abiotic factors. This may uncover niches beyond simply ice and water (potentially related to seasonal changes), and identify signs of environmental filtering, species co-occurrence, trait response, and niche conservatism within phylogenetic clades.

8.3 Concluding Remarks

The metagenomic data from MOSAiC has provided a benchmark for future Arctic exploration. The rapidly decreasing cost of next-generation sequencing has allowed metagenomic surveys on the scale of MOSAiC to be possible, but has also highlighted how much more there is to understand about this environment at a microbial level. As third-generation sequencing continues to become more cost-efficient, it will improve our understanding of microbial genomes, especially as these technologies open up the potential for telomere-to-telomere sequencing of eukaryotes. In the meantime, MAGs still play a major role in uncovering the unculturable majority of microbial life, and exploring the diversity of microbial ecosystems.

One key challenge in microbial ecology is understanding how to best use MAGs to link 'omics data to community composition. The standard tools from numerical ecology rely only on an OTU abundance table; PCoA plots, α and β diversity statistics, and species networks are all derived from this. MAGs bring a huge wealth of extra genomic information, and the trait-based methods that integrate genes, species, and abiotic factors together into a single model are correspondingly much more complex. For the Arctic in particular, understanding the linkage between sea, ice, and the communities within them, is necessary to bring the predictive quality of climate and ecological models of the region to parity with other much better studied parts of the Earth System.

The Arctic is already seeing drastic ecological changes due to the anthropogenic climate

change, and Arctic marine systems are exhibiting an 'Atlantification' effect [100], [102]. Further changes to the climate will undoubtedly have effects on the structure and function of Arctic microbial ecosystems, and hence on the biogeochemical cycles they contribute to and the food web supported by them. By benchmarking Arctic ecosystems, through the first large-scale analysis of MOSAiC sequencing data, we have provided an important starting point to understand how Arctic microbial communities will shift in the coming critical decades for climate change.

References

[1] U. Nixdorf et al., *MOSAiC Extended Acknowledgement*, Sep. 2021.

[2] C. B. Field, M. J. Behrenfeld, and J. T. Randerson, "Primary production of the biosphere: Integrating terrestrial and oceanic components," *Science*, vol. 281, no. 5374, pp. 237–240, Jul. 1998, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.281.5374.237.

[3] G. Darnis, D. Robert, C. Pomerleau, H. Link, P. Archambault, R. J. Nelson, M. Geoffroy, J.- Tremblay, C. Lovejoy, S. H. Ferguson, B. P. V. Hunt, and L. Fortier, "Current state and trends in Canadian Arctic marine ecosystems: II. Heterotrophic food web, pelagic-benthic coupling, and biodiversity," *Climatic Change*, vol. 115, no. 1, pp. 179–205, Nov. 2012, ISSN: 1573-1480. DOI: 10.1007/s10584-012-0483-8.

[4] G. Mcbean, G. Alekseev, D. Chen, E. Førland, J. Fyfe, P. Groisman, R. King, R. Melling, R. Vose, and P. Whitfield, "Arctic climate: Past and present," *In: Arctic Climate Impact Assessment*, Jan. 2005.

[5] The Alfred Wegener Institute, *The mosaic expedition website*, https://mosaic-expedition.org/, Accessed: 2026, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, 2022.

[6] The Alfred Wegener Institute, *Mosaic in numbers*, https://mosaic-expedition.org/expedition/mosaic-in-numbers/, Accessed: 2026, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, 2022.

[7] B. Rabe, C. Heuzé, J. Regnery, Y. Aksenov, J. Allerholt, M. Athanase, Y. Bai, C. Basque, D. Bauch, T. M. Baumann, D. Chen, S. T. Cole, L. Craw, A. Davies, E. Damm, K. Dethloff, D. V. Divine, F. Doglioni, F. Ebert, Y.-C. Fang, I. Fer, A. A. Fong, R. Gradinger, M. A. Granskog, R. Graupner, C. Haas, H. He, Y. He, M. Hoppmann, M. Janout, D. Kadko, T. Kanzow, S. Karam, Y. Kawaguchi, Z. Koenig, B. Kong, R. A. Krishfield, T. Krumpen, D. Kuhlmey, I. Kuznetsov, M. Lan, G. Laukert, R. Lei, T. Li, S. Torres-Valdés, L. Lin, L. Lin, H. Liu, N. Liu, B. Loose, X. Ma, R. McKay, M. Mallet, R. D. C. Mallett, W. Maslowski, C. Mertens, V. Mohrholz, M. Muilwijk, M. Nicolaus, J. K. O'Brien, D. Perovich, J. Ren, M. Rex, N. Ribeiro, A. Rinke, J. Schaffer, I. Schuffenhauer, K. Schulz, M. D. Shupe, W. Shaw, V. Sokolov, A. Sommerfeld, G. Spreen, T. Stanton, M. Stephens, J. Su, N. Sukhikh, A. Sundfjord, K. Thomisch, S. Tippenhauer, J. M. Toole, M. Vredenborg, M. Walter, H. Wang, L. Wang, Y. Wang, M. Wendisch, J. Zhao, M. Zhou, and J. Zhu, "Overview of the MOSAiC expedition: Physical oceanography," *Elementa: Science of the Anthropocene*, vol. 10, no. 1, Feb. 2022, ISSN: 2325-1026. DOI: 10.1525/elementa.2021.00062.

[8] T. Mock, W. Boulton, J.-P. Balmonte, K. Barry, S. Bertilsson, J. Bowman, M. Buck, G. Bratbak, E. J. Chamberlain, M. Cunliffe, J. Creamean, O. Ebenhöh, S. L. Eggers, A. A. Fong, J. Gardner, R. Gradinger, M. A. Granskog, C. Havermans, T. Hill, C. J. M. Hoppe, K. Korte, A. Larsen, O. Müller, A. Nicolaus, E. Oldenburg, O. Popa, S. Rogge, H. Schäfer, K. Shoemaker, P. Snoeijs-Leijonmalm, A. Torstensson, K. Valentin, A. Vader, K. Barry, I.-M. A. Chen, A. Clum, A. Copeland, C. Daum, E. Eloe-Fadrosh, B. Foster, B. Foster, I. V. Grigoriev, M. Huntemann, N. Ivanova, A. Kuo, N. C. Kyrpides, S. Mukherjee, K. Palaniappan, T. B. K. Reddy, A. Salamov, S. Roux, N. Varghese, T. Woyke, D. Wu, R. M. Leggett, V. Moulton, and K. Metfies, "Multiomics in the central Arctic Ocean for benchmarking biodiversity change," en, *PLOS Biology*, vol. 20, no. 10, e3001835, Oct. 2022, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3001835.

[9] S. Sunagawa, L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. d'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, n. null, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, E. Boss, C. Bowler, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. Sieracki, and D. Velayoudon, "Structure and function of the global ocean microbiome," *Science*, vol. 348, no. 6237, p. 1 261 359, 2015. DOI: 10.1126/science.1261359.

[10] I. V. Grigoriev, R. D. Hayes, S. Calhoun, B. Kamel, A. Wang, S. Ahrendt, S. Dusheyko, R. Nikitin, S. Mondo, A. Salamov, I. Shabalov, and A. Kuo, "PhycoCosm, a comparative algal genomics resource," *Nucleic Acids Research*, vol. 49, no. D1, pp. D1004–D1011, Jan. 2021, ISSN: 0305-1048. DOI: 10.1093/nar/gkaa898.

[11] W. Boulton, A. Salamov, I. V. Grigoriev, S. Calhoun, K. LaButti, R. Riley, K. Barry, A. A. Fong, C. J. M. Hoppe, K. Metfies, K. Oetjen, S. L. Eggers, O. Müller, J. Gardner, M. A. Granskog, A. Torstensson, M. Oggier, A. Larsen, G. Bratbak, A. Toseland, R. M. Leggett, V. Moulton, and T. Mock, "Metagenome-assembled-genomes recovered from the Arctic drift expedition MOSAiC," en, *Scientific Data*, vol. 12, no. 1, p. 204, Feb. 2025, ISSN: 2052-4463. DOI: 10.1038/s41597-025-04525-8.

[12] J. C. Winder, W. Boulton, A. Salamov, S. L. Eggers, K. Metfies, V. Moulton, and T. Mock, "Genetic and Structural Diversity of Prokaryotic Ice-Binding Proteins from the Central Arctic Ocean," en, *Genes*, vol. 14, no. 2, p. 363, Feb. 2023, ISSN: 2073-4425. DOI: 10.3390/genes14020363.

[13] M. Ferrer, C. Méndez-García, R. Bargiela, J. Chow, S. Alonso, A. García-Moyano, G. E. K. Bjerga, I. H. Steen, T. Schwabe, C. Blom, J. Vester, A. Weckbecker, P. Shahgaldian, C. C. C. R. de Carvalho, R. Meskys, G. Zanaroli, F. O. Glöckner, A.

Fernández-Guerra, S. Thambisetty, F. de la Calle, O. V. Golyshina, M. M. Yakimov, K.-E. Jaeger, A. F. Yakunin, W. R. Streit, O. McMeel, J.-B. Calewaert, N. Tonné, and P. N. Golyshin, "Decoding the ocean's microbiological secrets for marine enzyme biodiscovery," *FEMS Microbiology Letters*, vol. 366, no. 1, fny285, Dec. 2018, ISSN: 0378-1097. DOI: 10.1093/femsle/fny285.

[14] C. Garcia-Soto, E. Fernández, R. D. Pingree, and D. S. Harbour, "Evolution and structure of a shelf coccolithophore bloom in the Western English Channel," *Journal of Plankton Research*, vol. 17, no. 11, pp. 2011–2036, Nov. 1995, ISSN: 0142-7873. DOI: 10.1093/plankt/17.11.2011.

[15] M. L. Sogin, H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, "Microbial diversity in the deep sea and the underexplored “rare biosphere”," *Proceedings of the National Academy of Sciences*, vol. 103, no. 32, pp. 12 115–12 120, 2006. DOI: 10.1073/pnas.0605127103.

[16] G. Z. L. Dalmaso, D. Ferreira, and A. B. Vermelho, "Marine extremophiles: A source of hydrolases for biotechnological applications," eng, *Marine drugs*, vol. 13, no. 4, pp. 1925–1965, 2015, ISSN: 1660-3397. DOI: 10.3390/md13041925.

[17] Tara Oceans Coordinators, S. Sunagawa, S. G. Acinas, P. Bork, C. Bowler, D. Eveillard, G. Gorsky, L. Guidi, D. Iudicone, E. Karsenti, F. Lombard, H. Ogata, S. Pesant, M. B. Sullivan, P. Wincker, and C. de Vargas, "Tara Oceans: Towards global ocean ecosystems biology," en, *Nature Reviews Microbiology*, vol. 18, no. 8, pp. 428–445, Aug. 2020, ISSN: 1740-1526, 1740-1534. DOI: 10.1038/s41579-020-0364-5.

[18] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y.-H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter, "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific," en, *PLOS Biology*, vol. 5, no. 3, e77, Mar. 2007, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.0050077.

[19] B. J. Tully, E. D. Graham, and J. F. Heidelberg, "The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans," *Scientific Data*, vol. 5, no. 1, p. 170 203, 2018, ISSN: 2052-4463. DOI: 10.1038/sdata.2017.203.

[20] D. L. Kirchman, "Microbial primary production and phototrophy," in *Processes in Microbial Ecology*, D. L. Kirchman, Ed., Oxford University Press, Jul. 2018, p. 0, ISBN: 978-0-19-878940-6. DOI: 10.1093/oso/9780198789406.003.0006.

[21] Y. M. Bar-On and R. Milo, "The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet," *Cell*, vol. 179, no. 7, pp. 1451–1454, Dec. 2019, ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.11.018.

[22] R. Chester, "Nutrients, organic carbon and the carbon cycle in sea water," en, in *Marine Geochemistry*, R. Chester, Ed., Dordrecht: Springer Netherlands, 1990, pp. 272–320, ISBN: 978-94-010-9488-7. DOI: 10.1007/978-94-010-9488-7_9.

[23] P. Webb, "Introduction to oceanography," in *Introduction to Oceanography*, R. C. Press, Ed., Montreal: Rebus Community, 2023, ch. 7, pp. 147–163.

[24] T. J. Browning and C. M. Moore, "Global analysis of ocean phytoplankton nutrient limitation reveals high prevalence of co-limitation," en, *Nature Communications*, vol. 14, no. 1, p. 5014, Aug. 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-40774-0.

[25] P. M. Vitousek and R. W. Howarth, "Nitrogen Limitation on Land and in the Sea: How Can It Occur?" *Biogeochemistry*, vol. 13, no. 2, pp. 87–115, 1991, ISSN: 0168-2563.

[26] "Eighty years of Redfield," en, *Nature Geoscience*, vol. 7, no. 12, pp. 849–849, Dec. 2014, ISSN: 1752-0908. DOI: 10.1038/ngeo2319.

[27] A. C. Redfield, "On the proportions of organic derivatives in sea water and their relation to the composition of plankton," *James Johnstone Memorial Volume*, vol. James Johnstone Memorial Volume, D. R. J., Ed., pp. 176–192, 1934.

[28] T. Takahashi, W. S. Broecker, and S. Langer, "Redfield ratio based on chemical data from isopycnal surfaces," en, *Journal of Geophysical Research: Oceans*, vol. 90, no. C4, pp. 6907–6924, 1985, ISSN: 2156-2202. DOI: 10.1029/JC090iC04p06907.

[29] H. J. W. de Baar, A. G. J. Buma, R. F. Nolting, G. C. Cadée, G. Jacques, and P. J. Tréguer, "On iron limitation of the Southern Ocean: Experimental observations in the Weddell and Scotia Seas," *Marine Ecology Progress Series*, vol. 65, no. 2, pp. 105–122, 1990, ISSN: 0171-8630.

[30] L. W. Juranek, "Changing Biogeochemistry of the Arctic Ocean: Surface Nutrient and CO2 Cycling in a Warming, Melting North," *Oceanography*, vol. 35, no. 3-4, pp. 144–155, May 2022. DOI: 10.5670/oceanog.2022.120.

[31] A. Randelhoff, J. Holding, M. Janout, M. K. Sejr, M. Babin, J.- Tremblay, and M. B. Alkire, "Pan-Arctic Ocean Primary Production Constrained by Turbulent Nitrate Fluxes," English, *Frontiers in Marine Science*, vol. 7, Mar. 2020, ISSN: 2296-7745. DOI: 10.3389/fmars.2020.00150.

[32] R. L. Taylor, D. M. Semeniuk, C. D. Payne, J. Zhou, J.-E. Tremblay, J. T. Cullen, and M. T. Maldonado, "Colimitation by light, nitrate, and iron in the beaufort sea in late summer," *Journal of Geophysical Research: Oceans*, vol. 118, no. 7, pp. 3260–3277, 2013. DOI: https://doi.org/10.1002/jgrc.20244. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/jgrc.20244. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/jgrc.20244.

[33] S. Krisch, T. J. Browning, M. Graeve, K.-U. Ludwichowski, P. Lodeiro, M. J. Hopwood, S. Roig, J.-C. Yong, T. Kanzow, and E. P. Achterberg, "The influence of arctic fe and atlantic fixed n on summertime primary production in fram strait, north greenland sea," *Scientific Reports*, vol. 10, no. 1, p. 15 230, 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-72100-9. [Online]. Available: https://doi.org/10.1038/s41598-020-72100-9.

[34] S. D. Gerace, J. Yu, J. K. Moore, and A. C. Martiny, "Observed declines in upper ocean phosphate-to-nitrate availability," *Proceedings of the National Academy of Sciences*, vol. 122, no. 6, e2411835122, 2025. DOI: 10.1073/pnas.2411835122. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2411835122. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2411835122.

[35] J. W. Krause, I. K. Schulz, K. A. Rowe, W. Dobbins, M. H. S. Winding, M. K. Sejr, C. M. Duarte, and S. Agusti, "Silicic acid limitation drives bloom termination and potential carbon sequestration in an arctic bloom," *Scientific Reports*, vol. 9, no. 1, p. 8149, 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-44587-4. [Online]. Available: https://doi.org/10.1038/s41598-019-44587-4.

[36] S. R. Rintoul, C. W. Hughes, and D. Olbers, "Chapter 4.6 The antarctic circumpolar current system," in *International Geophysics*, ser. Ocean Circulation and Climate, G. Siedler, J. Church, and J. Gould, Eds., vol. 77, Academic Press, Jan. 2001, pp. 271–XXXVI. DOI: 10.1016/S0074-6142(01)80124-8.

[37] J. R. Toggweiler and R. M. Key, "Thermohaline Circulation," in *Encyclopedia of Ocean Sciences*, J. H. Steele, Ed., Oxford: Academic Press, Jan. 2001, pp. 2941–2947, ISBN: 978-0-12-227430-5. DOI: 10.1006/rwos.2001.0111.

[38] P. Ditlevsen and S. Ditlevsen, "Warning of a forthcoming collapse of the Atlantic meridional overturning circulation," en, *Nature Communications*, vol. 14, no. 1, p. 4254, Jul. 2023, ISSN: 2041-1723. DOI: 10.1038/s41467-023-39810-w.

[39] M. Pidwirny, *Physicalgeography.net*, https://commons.wikimedia.org/wiki/File:Corrientes-oceanicas-en.svg, Accessed: 2025-05-13, 2007.

[40] J. Forster and A. G. Hirst, "The temperature-size rule emerges from ontogenetic differences between growth and development rates," en, *Functional Ecology*, vol. 26,

no. 2, pp. 483–492, 2012, ISSN: 1365-2435. DOI: 10.1111/j.1365-2435.2011.01958.x.

[41] R. Richard, J. A. Lugsanay, and S.-P. Huang, "The temperature-size rule in the context of Dynamic Energy Budget theory," *Ecological Modelling*, vol. 493, p. 110 761, Jul. 2024, ISSN: 0304-3800. DOI: 10.1016/j.ecolmodel.2024.110761.

[42] F. M. Ibarbalz, N. Henry, M. C. Brandão, S. Martini, G. Busseni, H. Byrne, L. P. Coelho, H. Endo, J. M. Gasol, A. C. Gregory, F. Mahé, J. Rigonato, M. Royo-Llonch, G. Salazar, I. Sanz-Sáez, E. Scalco, D. Soviadan, A. A. Zayed, A. Zingone, K. Labadie, J. Ferland, C. Marec, S. Kandels, M. Picheral, C. Dimier, J. Poulain, S. Pisarev, M. Carmichael, S. Pesant, S. G. Acinas, M. Babin, P. Bork, E. Boss, C. Bowler, G. Cochrane, C. de Vargas, M. Follows, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels, L. Karp-Boss, E. Karsenti, F. Not, H. Ogata, S. Pesant, N. Poulton, J. Raes, C. Sardet, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, P. Wincker, M. Babin, E. Boss, D. Iudicone, O. Jaillon, S. G. Acinas, H. Ogata, E. Pelletier, L. Stemmann, M. B. Sullivan, S. Sunagawa, L. Bopp, C. de Vargas, L. Karp-Boss, P. Wincker, F. Lombard, C. Bowler, and L. Zinger, "Global Trends in Marine Plankton Diversity across Kingdoms of Life," *Cell*, vol. 179, no. 5, 1084–1097.e21, Nov. 2019, ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.10.008.

[43] K. J. Flynn, D. O. F. Skibinski, and C. Lindemann, "Effects of growth rate, cell size, motion, and elemental stoichiometry on nutrient transport kinetics," en, *PLOS Computational Biology*, vol. 14, no. 4, e1006118, Apr. 2018, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006118.

[44] K. H. Peter and U. Sommer, "Phytoplankton Cell Size Reduction in Response to Warming Mediated by Nutrient Limitation," en, *PLOS ONE*, vol. 8, no. 9, e71528, Sep. 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0071528.

[45] P. Brun, M. R. Payne, and T. Kiørboe, "Trait biogeography of marine copepods – an analysis across scales," en, *Ecology Letters*, vol. 19, no. 12, pp. 1403–1413, 2016, ISSN: 1461-0248. DOI: 10.1111/ele.12688.

[46] Z. Zhou, P. Q. Tran, K. Kieft, and K. Anantharaman, "Genome diversification in globally distributed novel marine Proteobacteria is linked to environmental adaptation," *The ISME Journal*, vol. 14, no. 8, pp. 2060–2077, Aug. 2020, ISSN: 1751-7362. DOI: 10.1038/s41396-020-0669-4.

[47] P. H. Bradley and K. S. Pollard, "Proteobacteria explain significant functional variability in the human gut microbiome," *Microbiome*, vol. 5, no. 1, p. 36, Mar. 2017, ISSN: 2049-2618. DOI: 10.1186/s40168-017-0244-z.

[48] Y. Wang, H. Lin, R. Huang, and W. Zhai, "Exploring the plankton bacteria diversity and distribution patterns in the surface water of northwest pacific ocean by

metagenomic methods," English, *Frontiers in Marine Science*, vol. 10, Apr. 2023, ISSN: 2296-7745. DOI: 10.3389/fmars.2023.1177401.

[49] A. Cordone, G. D'Errico, M. Magliulo, F. Bolinesi, M. Selci, M. Basili, R. de Marco, M. Saggiomo, P. Rivaro, D. Giovannelli, and O. Mangoni, "Bacterioplankton Diversity and Distribution in Relation to Phytoplankton Community Structure in the Ross Sea Surface Waters," English, *Frontiers in Microbiology*, vol. 13, Jan. 2022, ISSN: 1664-302X. DOI: 10.3389/fmicb.2022.722900.

[50] J. M. González and M. A. Moran, "Numerical dominance of a group of marine bacteria in the alpha-subclass of the class Proteobacteria in coastal seawater," *Applied and Environmental Microbiology*, vol. 63, no. 11, pp. 4237–4242, Nov. 1997. DOI: 10.1128/aem.63.11.4237-4242.1997.

[51] M. A. Moran, R. Belas, M. A. Schell, J. M. González, F. Sun, S. Sun, B. J. Binder, J. Edmonds, W. Ye, B. Orcutt, E. C. Howard, C. Meile, W. Palefsky, A. Goesmann, Q. Ren, I. Paulsen, L. E. Ulrich, L. S. Thompson, E. Saunders, and A. Buchan, "Ecological Genomics of Marine Roseobacters," *Applied and Environmental Microbiology*, vol. 73, no. 14, pp. 4559–4569, Jul. 2007. DOI: 10.1128/AEM.02580-06.

[52] R. Beiralas, N. Ozer, and E. Segev, "Abundant Sulfitobacter marine bacteria protect Emiliania huxleyi algae from pathogenic bacteria," en, *ISME Communications*, vol. 3, no. 1, pp. 1–10, Sep. 2023, ISSN: 2730-6151. DOI: 10.1038/s43705-023-00311-y.

[53] A. Gavriilidou, J. Gutleben, D. Versluis, F. Forgiarini, M. W. J. van Passel, C. J. Ingham, H. Smidt, and D. Sipkema, "Comparative genomic analysis of Flavobacteriaceae: Insights into carbohydrate metabolism, gliding motility and secondary metabolite biosynthesis," *BMC Genomics*, vol. 21, no. 1, p. 569, Aug. 2020, ISSN: 1471-2164. DOI: 10.1186/s12864-020-06971-7.

[54] J. Hudson and S. Egan, "Opportunistic diseases in marine eukaryotes: Could Bacteroidota be the next threat to ocean life?" en, *Environmental Microbiology*, vol. 24, no. 10, pp. 4505–4518, 2022, ISSN: 1462-2920. DOI: 10.1111/1462-2920.16094.

[55] J. M. González, B. Fernández-Gómez, A. Fernàndez-Guerra, L. Gómez-Consarnau, O. Sánchez, M. Coll-Lladó, J. del Campo, L. Escudero, R. Rodríguez-Martínez, L. Alonso-Sáez, M. Latasa, I. Paulsen, O. Nedashkovskaya, I. Lekunberri, J. Pinhassi, and C. Pedrós-Alió, "Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 25, pp. 8724–8729, Jun. 2008, ISSN: 0027-8424. DOI: 10.1073/pnas.0712027105.

[56] J. Chen, Y. Jia, Y. Sun, K. Liu, C. Zhou, C. Liu, D. Li, G. Liu, C. Zhang, T. Yang, L. Huang, Y. Zhuang, D. Wang, D. Xu, Q. Zhong, Y. Guo, A. Li, I. Seim, L. Jiang, L. Wang, S. M. Y. Lee, Y. Liu, D. Wang, G. Zhang, S. Liu, X. Wei, Z. Yue, S. Zheng,

X. Shen, S. Wang, C. Qi, J. Chen, C. Ye, F. Zhao, J. Wang, J. Fan, B. Li, J. Sun, X. Jia, Z. Xia, H. Zhang, J. Liu, Y. Zheng, X. Liu, J. Wang, H. Yang, K. Kristiansen, X. Xu, T. Mock, S. Li, W. Zhang, and G. Fan, "Global marine microbial diversity and its potential in bioprospecting," en, *Nature*, vol. 633, no. 8029, pp. 371–379, Sep. 2024, ISSN: 1476-4687. DOI: 10.1038/s41586-024-07891-2.

[57] Y. Nishimura and S. Yoshizawa, "The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments," en, *Scientific Data*, vol. 9, no. 1, p. 305, Jun. 2022, ISSN: 2052-4463. DOI: 10.1038/s41597-022-01392-5.

[58] F. Partensky, W. R. Hess, and D. Vaulot, "¡i¿prochlorococcus¡/i¿, a marine photosynthetic prokaryote of global significance," *Microbiology and Molecular Biology Reviews*, vol. 63, no. 1, pp. 106–127, 1999. DOI: 10.1128/MMBR.63.1.106-127.1999.

[59] F. G. Woese C., *Phylogenetic structure of the prokaryotic domain: The primary kingdoms*, en, 1977. DOI: 10.1073/pnas.74.11.5088.

[60] E. F. DeLong, "Exploring Marine Planktonic Archaea: Then and Now," *Frontiers in Microbiology*, vol. 11, 2021, ISSN: 1664-302X.

[61] L. Eme, A. Spang, J. Lombard, C. W. Stairs, and T. J. G. Ettema, "Archaea and the origin of eukaryotes," en, *Nature Reviews Microbiology*, vol. 15, no. 12, pp. 711–723, Dec. 2017, ISSN: 1740-1526, 1740-1534. DOI: 10.1038/nrmicro.2017.133.

[62] S. B. Hedges, H. Chen, S. Kumar, D. Y.-C. Wang, A. S. Thompson, and H. Watanabe, "A genomic timescale for the origin of eukaryotes," en, *BMC Evolutionary Biology*, p. 10, 2001.

[63] D. G. Mann and P. Vanormelingen, "An Inordinate Fondness? The Number, Distributions, and Origins of Diatom Species," en, *Journal of Eukaryotic Microbiology*, vol. 60, no. 4, pp. 414–420, Jul. 2013, ISSN: 10665234. DOI: 10.1111/jeu.12047.

[64] D. J. G. I. Thomas Mock, *100 diatom genomes project*, https://archive.jgi.doe.gov/csp-2021-100-diatom-genomes/, 2021.

[65] W. R. Roberts, A. M. Siepielski, and A. J. Alverson, "Diatom abundance in the polar oceans is predicted by genome size," en, *PLOS Biology*, vol. 22, no. 8, e3002733, Aug. 2024, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3002733.

[66] V. Villanova and C. Spetea, "Mixotrophy in diatoms: Molecular mechanism and industrial potential," en, *Physiologia Plantarum*, vol. 173, no. 2, pp. 603–611, 2021, ISSN: 1399-3054. DOI: 10.1111/ppl.13471.

[67] D. G. Mann, "The species concept in diatoms," *Phycologia*, vol. 38, no. 6, pp. 437–495, 1999. DOI: 10.2216/i0031-8884-38-6-437.1.

[68] H. M. Kauko, L. M. Olsen, P. Duarte, I. Peeken, M. A. Granskog, G. Johnsen, M. Fernández-Méndez, A. K. Pavlov, C. J. Mundy, and P. Assmy, "Algal Colonization of Young Arctic Sea Ice in Spring," en, *Frontiers in Marine Science*, vol. 5, p. 199, Jun. 2018, ISSN: 2296-7745. DOI: 10.3389/fmars.2018.00199.

[69] F. J. R. Taylor, M. Hoppenrath, and J. F. Saldarriaga, "Dinoflagellate diversity and distribution," en, *Biodiversity and Conservation*, vol. 17, no. 2, pp. 407–418, Feb. 2008, ISSN: 0960-3115, 1572-9710. DOI: 10.1007/s10531-007-9258-3.

[70] C. F. Delwiche, "CHAPTER 10 - The Origin and Evolution of Dinoflagellates," in *Evolution of Primary Producers in the Sea*, P. G. Falkowski and A. H. Knoll, Eds., Burlington: Academic Press, Jan. 2007, pp. 191–205, ISBN: 978-0-12-370518-1. DOI: 10.1016/B978-012370518-1/50011-4.

[71] H. Wang, P. Wu, L. Xiong, H.-S. Kim, J. H. Kim, and J.-S. Ki, "Nuclear genome of dinoflagellates: Size variation and insights into evolutionary mechanisms," eng, *European Journal of Protistology*, vol. 93, p. 126 061, Apr. 2024, ISSN: 1618-0429. DOI: 10.1016/j.ejop.2024.126061.

[72] J. Bradbury, "Nature's nanotechnologists: Unveiling the secrets of diatoms," *PLOS Biology*, vol. 2, no. 10, null, Oct. 2004. DOI: 10.1371/journal.pbio.0020306.

[73] B. A. Read, J. Kegel, M. J. Klute, A. Kuo, S. C. Lefebvre, F. Maumus, C. Mayer, J. Miller, A. Monier, A. Salamov, J. Young, M. Aguilar, J.-M. Claverie, S. Frickenhaus, K. Gonzalez, E. K. Herman, Y.-C. Lin, J. Napier, H. Ogata, A. F. Sarno, J. Shmutz, D. Schroeder, C. de Vargas, F. Verret, P. von Dassow, K. Valentin, Y. Van de Peer, G. Wheeler, J. B. Dacks, C. F. Delwiche, S. T. Dyhrman, G. Glöckner, U. John, T. Richards, A. Z. Worden, X. Zhang, and I. V. Grigoriev, "Pan genome of the phytoplankton Emiliania underpins its global distribution," en, *Nature*, vol. 499, no. 7457, pp. 209–213, Jul. 2013, ISSN: 1476-4687. DOI: 10.1038/nature12221.

[74] J. G. Umen, "Green Algae and the Origins of Multicellularity in the Plant Kingdom," *Cold Spring Harbor Perspectives in Biology*, vol. 6, no. 11, a016170, Nov. 2014, ISSN: 1943-0264. DOI: 10.1101/cshperspect.a016170.

[75] M. D. HERRON, "Origins of multicellular complexity: Volvox and the volvocine algae," *Molecular ecology*, vol. 25, no. 6, pp. 1213–1223, Mar. 2016, ISSN: 0962-1083. DOI: 10.1111/mec.13551.

[76] V. Jimenez, J. A. Burns, F. Le Gall, F. Not, and D. Vaulot, "No evidence of Phagomixotropy in Micromonas polaris (Mamiellophyceae), the Dominant Picophytoplank-

ton Species in the Arctic," eng, *Journal of Phycology*, vol. 57, no. 2, pp. 435–446, Apr. 2021, ISSN: 1529-8817. DOI: 10.1111/jpy.13125.

[77] M. J. van Baren, C. Bachy, E. N. Reistetter, S. O. Purvine, J. Grimwood, S. Sudek, H. Yu, C. Poirier, T. J. Deerinck, A. Kuo, I. V. Grigoriev, C.-H. Wong, R. D. Smith, S. J. Callister, C.-L. Wei, J. Schmutz, and A. Z. Worden, "Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants," *BMC Genomics*, vol. 17, no. 1, p. 267, Mar. 2016, ISSN: 1471-2164. DOI: 10.1186/s12864-016-2585-6.

[78] E. Derelle, C. Ferraz, S. Rombauts, P. Rouzé, A. Z. Worden, S. Robbens, F. Partensky, S. Degroeve, S. Echeynié, R. Cooke, Y. Saeys, J. Wuyts, K. Jabbari, C. Bowler, O. Panaud, B. Piégu, S. G. Ball, J.-P. Ral, F.-Y. Bouget, G. Piganeau, B. De Baets, A. Picard, M. Delseny, J. Demaille, Y. Van de Peer, and H. Moreau, "Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features," *Proceedings of the National Academy of Sciences*, vol. 103, no. 31, pp. 11 647–11 652, Aug. 2006. DOI: 10.1073/pnas.0604795103.

[79] R. Anderson, S. Charvet, and P. J. Hansen, "Mixotrophy in Chlorophytes and Haptophytes—Effect of Irradiance, Macronutrient, Micronutrient and Vitamin Limitation," English, *Frontiers in Microbiology*, vol. 9, Jul. 2018, ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.01704.

[80] R. Ghahreman, W. Gong, M. Galí, A.-L. Norman, S. R. Beagley, A. Akingunola, Q. Zheng, A. Lupu, M. Lizotte, M. Levasseur, and W. R. Leaitch, "Dimethyl sulfide and its role in aerosol formation and growth in the Arctic summer – a modelling study," English, *Atmospheric Chemistry and Physics*, vol. 19, no. 23, pp. 14 455–14 476, Nov. 2019, ISSN: 1680-7316. DOI: 10.5194/acp-19-14455-2019.

[81] A. W. Decho and T. Gutierrez, "Microbial Extracellular Polymeric Substances (EPSs) in Ocean Systems," *Frontiers in Microbiology*, vol. 8, p. 922, May 2017, ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.00922.

[82] D. A. Hutchins and D. G. Capone, "The marine nitrogen cycle: New developments and global change," en, *Nature Reviews Microbiology*, vol. 20, no. 7, pp. 401–414, Jul. 2022, ISSN: 1740-1534. DOI: 10.1038/s41579-022-00687-z.

[83] L. I. Falcón, F. Cipriano, A. Y. Chistoserdov, and E. J. Carpenter, "Diversity of Diazotrophic Unicellular Cyanobacteria in the Tropical North Atlantic Ocean," *Applied and Environmental Microbiology*, vol. 68, no. 11, pp. 5760–5764, Nov. 2002, ISSN: 0099-2240. DOI: 10.1128/AEM.68.11.5760-5764.2002.

[84] T. O. Delmont, J. J. Pierella Karlusich, I. Veseli, J. Fuessel, A. M. Eren, R. A. Foster, C. Bowler, P. Wincker, and E. Pelletier, "Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering

most of the sunlit ocean," *The ISME Journal*, vol. 16, no. 4, pp. 927–936, Apr. 2022, ISSN: 1751-7362. DOI: 10.1038/s41396-021-01135-1.

[85] X.-R. Yang, H. Li, J.-Q. Su, and G.-W. Zhou, "Anammox Bacteria Are Potentially Involved in Anaerobic Ammonium Oxidation Coupled to Iron(III) Reduction in the Wastewater Treatment System," English, *Frontiers in Microbiology*, vol. 12, Sep. 2021, ISSN: 1664-302X. DOI: 10.3389/fmicb.2021.717249.

[86] J. G. Kuenen, "Anammox bacteria: From discovery to application," en, *Nature Reviews Microbiology*, vol. 6, no. 4, pp. 320–326, Apr. 2008, ISSN: 1740-1534. DOI: 10.1038/nrmicro1857.

[87] J. R. Seymour, S. A. Amin, J.-B. Raina, and R. Stocker, "Zooming in on the phycosphere: The ecological interface for phytoplankton–bacteria relationships," en, *Nature Microbiology*, vol. 2, no. 7, pp. 1–12, May 2017, ISSN: 2058-5276. DOI: 10.1038/nmicrobiol.2017.65.

[88] J. W. Hastings and E. P. Greenberg, "Quorum Sensing: The Explanation of a Curious Phenomenon Reveals a Common Characteristic of Bacteria," *Journal of Bacteriology*, vol. 181, no. 9, pp. 2667–2668, May 1999. DOI: 10.1128/jb.181.9.2667-2668.1999.

[89] D. R. Finn, B. Bergk-Pinto, C. Hazard, G. W. Nicol, C. C. Tebbe, and T. M. Vogel, "Functional trait relationships demonstrate life strategies in terrestrial prokaryotes," *FEMS Microbiology Ecology*, vol. 97, no. 5, fiab068, May 2021, ISSN: 0168-6496. DOI: 10.1093/femsec/fiab068.

[90] Zeimusu, *Arctic ocean currents*, https://commons.wikimedia.org/wiki/File:Arctic_Ocean_circulation_map.svg, Accessed: 2025-05-22, 2012.

[91] B. Hwang, Y. Aksenov, E. Blockley, M. Tsamados, T. Brown, J. Landy, D. Stevens, and J. Wilkinson, "Impacts of climate change on Arctic sea ice," en, *MCCIP Science Review 2020*, 20 pages, 2020. DOI: 10.14465/2020.ARC10.ICE.

[92] M. Bayer-Giraldi, I. Weikusat, H. Besir, and G. Dieckmann, "Characterization of an antifreeze protein from the polar diatom Fragilariopsis cylindrus and its relevance in sea ice," en, *Cryobiology*, vol. 63, no. 3, pp. 210–219, Aug. 2011.

[93] R. G. Dorrell, A. Kuo, Z. Füssy, E. H. Richardson, A. Salamov, N. Zarevski, N. J. Freyria, F. M. Ibarbalz, J. Jenkins, J. J. Pierella Karlusich, A. Stecca Steindorff, R. E. Edgar, L. Handley, K. Lail, A. Lipzen, V. Lombard, J. McFarlane, C. Nef, A. M. Novák Vanclová, Y. Peng, C. Plott, M. Potvin, F. R. J. Vieira, K. Barry, C. De Vargas, B. Henrissat, E. Pelletier, J. Schmutz, P. Wincker, J. B. Dacks, C. Bowler, I. V. Grigoriev, and C. Lovejoy, "Convergent evolution and horizontal gene transfer in Arctic Ocean microalgae," en, *Life Science Alliance*, vol. 6, no. 3, e202201833, Mar. 2023, ISSN: 2575-1077. DOI: 10.26508/lsa.202201833.

[94] C Krembs, H Eicken, K Junge, and J. W Deming, "High concentrations of exopolymeric substances in Arctic winter sea ice: Implications for the polar ocean carbon cycle and cryoprotection of diatoms," *Deep Sea Research Part I: Oceanographic Research Papers*, vol. 49, no. 12, pp. 2163–2181, Dec. 2002, ISSN: 0967-0637. DOI: 10.1016/S0967-0637(02)00122-X.

[95] J. S. Bowman, S. Rasmussen, N. Blom, J. W. Deming, S. Rysgaard, and T. Sicheritz-Ponten, "Microbial community structure of Arctic multiyear sea ice and surface seawater by 454 sequencing of the 16S RNA gene," en, *The ISME Journal*, vol. 6, no. 1, pp. 11–20, Jan. 2012, ISSN: 1751-7370. DOI: 10.1038/ismej.2011.76.

[96] C. M. Bellas, K. Campbell, M. Tranter, and P. Sánchez-Baracaldo, "Nitrogen and sulfur metabolisms encoded in prokaryotic communities associated with sea ice algae," *ISME Communications*, vol. 3, no. 1, p. 131, Dec. 2023, ISSN: 2730-6151. DOI: 10.1038/s43705-023-00337-2.

[97] J. Z. Rapp, M. Fernández-Méndez, C. Bienhold, and A. Boetius, "Effects of Ice-Algal Aggregate Export on the Connectivity of Bacterial Communities in the Central Arctic Ocean," *Frontiers in Microbiology*, vol. 9, p. 1035, May 2018, ISSN: 1664-302X. DOI: 10.3389/fmicb.2018.01035.

[98] J. P. Bowman, S. A. McCammon, J. L. Brown, and T. A. McMeekin, "Glaciecola punicea gen. nov., sp. nov. and Glaciecola pallidula gen. nov., sp. nov.: Psychrophilic bacteria from Antarctic sea-ice habitats," *International Journal of Systematic and Evolutionary Microbiology*, vol. 48, no. 4, pp. 1213–1222, 1998, ISSN: 1466-5034. DOI: 10.1099/00207713-48-4-1213.

[99] A. A. Fong, C. J. M. Hoppe, N. Aberle, C. J. Ashjian, P. Assmy, Y. Bai, D. C. E. Bakker, J. P. Balmonte, K. R. Barry, S. Bertilsson, W. Boulton, J. Bowman, D. Bozzato, G. Bratbak, M. Buck, R. G. Campbell, G. Castellani, E. J. Chamberlain, J. Chen, M. Chierici, A. Cornils, J. M. Creamean, E. Damm, K. Dethloff, E. S. Droste, O. Ebenhöh, S. L. Eggers, A. Engel, H. Flores, A. Fransson, S. Frickenhaus, J. Gardner, C. E. Gelfman, M. A. Granskog, M. Graeve, C. Havermans, C. Heuzé, N. Hildebrandt, T. C. J. Hill, M. Hoppema, A. Immerz, H. Jin, B. P. Koch, X. Kong, A. Kraberg, M. Lan, B. A. Lange, A. Larsen, B. Lebreton, E. Leu, B. Loose, W. Maslowski, C. Mavis, K. Metfies, T. Mock, O. Müller, M. Nicolaus, B. Niehoff, D. Nomura, E.-M. Nöthig, M. Oggier, E. Oldenburg, L. M. Olsen, I. Peeken, D. K. Perovich, O. Popa, B. Rabe, J. Ren, M. Rex, A. Rinke, S. Rokitta, B. Rost, S. Sakinan, E. Salganik, F. L. Schaafsma, H. Schäfer, K. Schmidt, K. M. Shoemaker, M. D. Shupe, P. Snoeijs-Leijonmalm, J. Stefels, A. Svenson, R. Tao, S. Torres-Valdés, A. Torstensson, A. Toseland, A. Ulfsbo, M. A. Van Leeuwe, M. Vortkamp, A. L. Webb, Y. Zhuang, and R. R. Gradinger, "Overview of the MOSAiC expedition: Ecosystem," *Elementa: Science of the Anthropocene*, vol. 12, no. 1, p. 00 135, Aug. 2024, ISSN: 2325-1026. DOI: 10.1525/elementa.2023.00135.

[100] T. Priest, W.-J. von Appen, E. Oldenburg, O. Popa, S. Torres-Valdés, C. Bienhold, K. Metfies, W. Boulton, T. Mock, B. M. Fuchs, R. Amann, A. Boetius, and M. Wietz, "Atlantic water influx and sea-ice cover drive taxonomic and functional shifts in Arctic marine bacterial communities," en, *The ISME Journal*, vol. 17, no. 10, pp. 1612–1625, Oct. 2023, ISSN: 1751-7370. DOI: 10.1038/s41396-023-01461-6.

[101] T. Priest, E. Oldenburg, O. Popa, B. Dede, K. Metfies, W.-J. von Appen, S. Torres-Valdés, C. Bienhold, B. M. Fuchs, R. Amann, A. Boetius, and M. Wietz, "Seasonal recurrence and modular assembly of an Arctic pelagic marine microbiome," en, *Nature Communications*, vol. 16, no. 1, p. 1326, Feb. 2025, ISSN: 2041-1723. DOI: 10.1038/s41467-025-56203-3.

[102] E. Oldenburg, O. Popa, M. Wietz, W.-J. von Appen, S. Torres-Valdes, C. Bienhold, O. Ebenhöh, and K. Metfies, "Sea-ice melt determines seasonal phytoplankton dynamics and delimits the habitat of temperate Atlantic taxa as the Arctic Ocean atlantifies," *ISME Communications*, vol. 4, no. 1, ycae027, Jan. 2024, ISSN: 2730-6151. DOI: 10.1093/ismeco/ycae027.

[103] M. Royo-Llonch, P. Sánchez, C. Ruiz-González, G. Salazar, C. Pedrós-Alió, M. Sebastián, K. Labadie, L. Paoli, F. M. Ibarbalz, L. Zinger, B. Churcheward, S. Chaffron, D. Eveillard, E. Karsenti, S. Sunagawa, P. Wincker, L. Karp-Boss, C. Bowler, and S. G. Acinas, "Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean," en, *Nature Microbiology*, vol. 6, no. 12, pp. 1561–1574, Dec. 2021, ISSN: 2058-5276. DOI: 10.1038/s41564-021-00979-9.

[104] A. Duncan, K. Barry, C. Daum, E. Eloe-Fadrosh, S. Roux, K. Schmidt, S. G. Tringe, K. U. Valentin, N. Varghese, A. Salamov, I. V. Grigoriev, R. M. Leggett, V. Moulton, and T. Mock, "Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans," en, *Microbiome*, vol. 10, no. 1, p. 67, Dec. 2022, ISSN: 2049-2618. DOI: 10.1186/s40168-022-01254-7.

[105] M. Royo-Llonch, P. Sánchez, C. Ruiz-González, G. Salazar, C. Pedrós-Alió, M. Sebastián, K. Labadie, L. Paoli, F. M. Ibarbalz, L. Zinger, B. Churcheward, S. Chaffron, D. Eveillard, E. Karsenti, S. Sunagawa, P. Wincker, L. Karp-Boss, C. Bowler, and S. G. Acinas, "Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean," en, *Nature Microbiology*, vol. 6, no. 12, Dec. 2021, ISSN: 2058-5276. DOI: 10.1038/s41564-021-00979-9.

[106] C. Pedrós-Alió, M. Potvin, and C. Lovejoy, "Diversity of planktonic microorganisms in the Arctic Ocean," *Progress in Oceanography*, Overarching perspectives of contemporary and future ecosystems in the Arctic Ocean, vol. 139, pp. 233–243, Dec. 2015, ISSN: 0079-6611. DOI: 10.1016/j.pocean.2015.07.009.

[107] L. T. Gill, J. R. Kennedy, and K. E. Marshall, "Proteostasis in ice: The role of heat shock proteins and ubiquitin in the freeze tolerance of the intertidal mussel, Mytilus

trossulus," eng, *Journal of Comparative Physiology. B, Biochemical, Systemic, and Environmental Physiology*, vol. 193, no. 2, pp. 155–169, Mar. 2023, ISSN: 1432-136X. DOI: 10.1007/s00360-022-01473-2.

[108] M. Fiala and L. Oriol, "Light-temperature interactions on the growth of Antarctic diatoms," en, *Polar Biology*, vol. 10, no. 8, pp. 629–636, Oct. 1990, ISSN: 1432-2056. DOI: 10.1007/BF00239374.

[109] N. P. A. Hüner, D. R. Smith, M. Cvetkovska, X. Zhang, A. G. Ivanov, B. Szyszka-Mroz, I. Kalra, and R. Morgan-Kiss, "Photosynthetic adaptation to polar life: Energy balance, photoprotection and genetic redundancy," *Journal of Plant Physiology*, vol. 268, p. 153 557, Jan. 2022, ISSN: 0176-1617. DOI: 10.1016/j.jplph.2021.153557.

[110] S. Yasunaka, T. Ono, K. Sasaoka, and K. Sato, "Global distribution and variability of subsurface chlorophyll *a* concentrations," English, *Ocean Science*, vol. 18, no. 1, pp. 255–268, Feb. 2022, ISSN: 1812-0784. DOI: 10.5194/os-18-255-2022.

[111] C. J. M. Hoppe, N. Fuchs, D. Notz, P. Anderson, P. Assmy, J. Berge, G. Bratbak, G. Guillou, A. Kraberg, A. Larsen, B. Lebreton, E. Leu, M. Lucassen, O. Müller, L. Oziel, B. Rost, B. Schartmüller, A. Torstensson, and J. Wloka, "Photosynthetic light requirement near the theoretical minimum detected in Arctic microalgae," en, *Nature Communications*, vol. 15, no. 1, p. 7385, Sep. 2024, ISSN: 2041-1723. DOI: 10.1038/s41467-024-51636-8.

[112] F. Not, R. Massana, M. Latasa, D. Marie, C. Colson, W. Eikrem, C. Pedrós-Alió, D. Vaulot, and N. Simon, "Late summer community composition and abundance of photosynthetic picoeukaryotes in Norwegian and Barents Seas," en, *Limnology and Oceanography*, vol. 50, no. 5, pp. 1677–1686, 2005, ISSN: 1939-5590. DOI: 10.4319/lo.2005.50.5.1677.

[113] A. Vader, M. Marquardt, A. R. Meshram, and T. M. Gabrielsen, "Key Arctic phototrophs are widespread in the polar night," en, *Polar Biology*, vol. 38, no. 1, pp. 13–21, Jan. 2015, ISSN: 1432-2056. DOI: 10.1007/s00300-014-1570-2.

[114] J. Berge, P. E. Renaud, G. Darnis, F. Cottier, K. Last, T. M. Gabrielsen, G. Johnsen, L. Seuthe, J. M. Weslawski, E. Leu, M. Moline, J. Nahrgang, J. E. Søreide, Varpe, O. J. Lønne, M. Daase, and S. Falk-Petersen, "In the dark: A review of ecosystem processes during the Arctic polar night," *Progress in Oceanography*, Overarching perspectives of contemporary and future ecosystems in the Arctic Ocean, vol. 139, pp. 258–271, Dec. 2015, ISSN: 0079-6611. DOI: 10.1016/j.pocean.2015.08.005.

[115] B. Rabe, C. Heuzé, J. Regnery, Y. Aksenov, J. Allerholt, M. Athanase, Y. Bai, C. Basque, D. Bauch, T. M. Baumann, D. Chen, S. T. Cole, L. Craw, A. Davies, E. Damm, K. Dethloff, D. V. Divine, F. Doglioni, F. Ebert, Y.-C. Fang, I. Fer, A. A. Fong, R. Gradinger, M. A. Granskog, R. Graupner, C. Haas, H. He, Y. He, M. Hoppmann,

M. Janout, D. Kadko, T. Kanzow, S. Karam, Y. Kawaguchi, Z. Koenig, B. Kong, R. A. Krishfield, T. Krumpen, D. Kuhlmey, I. Kuznetsov, M. Lan, G. Laukert, R. Lei, T. Li, S. Torres-Valdés, L. Lin, L. Lin, H. Liu, N. Liu, B. Loose, X. Ma, R. McKay, M. Mallet, R. D. C. Mallett, W. Maslowski, C. Mertens, V. Mohrholz, M. Muilwijk, M. Nicolaus, J. K. O'Brien, D. Perovich, J. Ren, M. Rex, N. Ribeiro, A. Rinke, J. Schaffer, I. Schuffenhauer, K. Schulz, M. D. Shupe, W. Shaw, V. Sokolov, A. Sommerfeld, G. Spreen, T. Stanton, M. Stephens, J. Su, N. Sukhikh, A. Sundfjord, K. Thomisch, S. Tippenhauer, J. M. Toole, M. Vredenborg, M. Walter, H. Wang, L. Wang, Y. Wang, M. Wendisch, J. Zhao, M. Zhou, and J. Zhu, "Overview of the MOSAiC expedition: Physical oceanography," en, *Elementa: Science of the Anthropocene*, vol. 10, no. 1, p. 00 062, Feb. 2022, ISSN: 2325-1026. DOI: 10.1525/elementa.2021.00062.

[116] M. Nicolaus, D. K. Perovich, G. Spreen, M. A. Granskog, L. von Albedyll, M. Angelopoulos, P. Anhaus, S. Arndt, H. J. Belter, V. Bessonov, G. Birnbaum, J. Brauchle, R. Calmer, E. Cardellach, B. Cheng, D. Clemens-Sewall, R. Dadic, E. Damm, G. de Boer, O. Demir, K. Dethloff, D. V. Divine, A. A. Fong, S. Fons, M. M. Frey, N. Fuchs, C. Gabarró, S. Gerland, H. F. Goessling, R. Gradinger, J. Haapala, C. Haas, J. Hamilton, H.-R. Hannula, S. Hendricks, A. Herber, C. Heuzé, M. Hoppmann, K. V. Høyland, M. Huntemann, J. K. Hutchings, B. Hwang, P. Itkin, H.-W. Jacobi, M. Jaggi, A. Jutila, L. Kaleschke, C. Katlein, N. Kolabutin, D. Krampe, S. S. Kristensen, T. Krumpen, N. Kurtz, A. Lampert, B. A. Lange, R. Lei, B. Light, F. Linhardt, G. E. Liston, B. Loose, A. R. Macfarlane, M. Mahmud, I. O. Matero, S. Maus, A. Morgenstern, R. Naderpour, V. Nandan, A. Niubom, M. Oggier, N. Oppelt, F. Pätzold, C. Perron, T. Petrovsky, R. Pirazzini, C. Polashenski, B. Rabe, I. A. Raphael, J. Regnery, M. Rex, R. Ricker, K. Riemann-Campe, A. Rinke, J. Rohde, E. Salganik, R. K. Scharien, M. Schiller, M. Schneebeli, M. Semmling, E. Shimanchuk, M. D. Shupe, M. M. Smith, V. Smolyanitsky, V. Sokolov, T. Stanton, J. Stroeve, L. Thielke, A. Timofeeva, R. T. Tonboe, A. Tavri, M. Tsamados, D. N. Wagner, D. Watkins, M. Webster, and M. Wendisch, "Overview of the MOSAiC expedition: Snow and sea ice," *Elementa: Science of the Anthropocene*, vol. 10, no. 1, p. 000 046, Feb. 2022, ISSN: 2325-1026. DOI: 10.1525/elementa.2021.000046.

[117] M. D. Shupe, M. Rex, B. Blomquist, P. O. G. Persson, J. Schmale, T. Uttal, and D. Althausen, "Overview of the MOSAiC expedition: Atmosphere," *Elementa: Science of the Anthropocene*, vol. 10, no. 1, p. 00 060, Feb. 2022, ISSN: 2325-1026. DOI: 10.1525/elementa.2021.00060. (visited on 05/03/2025).

[118] R. Knust, "Polar Research and Supply Vessel POLARSTERN Operated by the Alfred-Wegener-Institute," en, *Journal of large-scale research facilities JLSRF*, vol. 3, A119–A119, Oct. 2017, ISSN: 2364-091X. DOI: 10.17815/jlsrf-3-163.

[119] J. D. Christopher Krembs, *What do we know about organisms that thrive in arctic sea ice?* https://www.pmel.noaa.gov/arctic-zone/essay_krembsdeming.html, 2022.

[120] A. Rodríguez-Gijón, M. Buck, A. F. Andersson, D. Izabel-Shen, F. J. A. Nascimento, and S. L. Garcia, "Linking prokaryotic genome size variation to metabolic potential and environment," *ISME Communications*, vol. 3, p. 25, Mar. 2023, ISSN: 2730-6151. DOI: 10.1038/s43705-023-00231-x.

[121] F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of Molecular Biology*, vol. 3, no. 3, pp. 318–356, 1961, ISSN: 0022-2836. DOI: https://doi.org/10.1016/S0022-2836(61)80072-7.

[122] A. Piovesan, M. Caracausi, F. Antonaros, M. C. Pelleri, and L. Vitale, "GeneBase 1.1: A tool to summarize data from NCBI Gene datasets and its application to an update of human gene statistics," *Database*, vol. 2016, baw153, Jan. 2016, ISSN: 1758-0463. DOI: 10.1093/database/baw153.

[123] N. J. Fagundes, R. Bisso-Machado, P. I. Figueiredo, M. Varal, and A. L. Zani, "What We Talk About When We Talk About "Junk DNA"," *Genome Biology and Evolution*, vol. 14, no. 5, evac055, May 2022, ISSN: 1759-6653. DOI: 10.1093/gbe/evac055.

[124] T. Shafee and R. Lowe, "Eukaryotic and prokaryotic gene structure," *WikiJournal of Medicine*, vol. 4, no. 1, p. 1, Apr. 2015, ISSN: 20024436. DOI: 10.15347/wjm/2017.002. (visited on 05/23/2025).

[125] Y. M. Bar-On and R. Milo, "The global mass and average rate of rubisco," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4738–4743, Mar. 2019. DOI: 10.1073/pnas.1816654116.

[126] M. R. Badger, T. J. Andrews, S. M. Whitney, M. Ludwig, D. C. Yellowlees, W Leggat, and G. D. Price, "The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO2-concentrating mechanisms in algae," *Canadian Journal of Botany*, vol. 76, no. 6, pp. 1052–1071, Jun. 1998, ISSN: 0008-4026. DOI: 10.1139/b98-074.

[127] L. C. M. Mackinder, M. T. Meyer, T. Mettler-Altmann, V. K. Chen, M. C. Mitchell, O. Caspari, E. S. Freeman Rosenzweig, L. Pallesen, G. Reeves, A. Itakura, R. Roth, F. Sommer, S. Geimer, T. Mühlhaus, M. Schroda, U. Goodenough, M. Stitt, H. Griffiths, and M. C. Jonikas, "A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle," *Proceedings of the National Academy of Sciences*, vol. 113, no. 21, pp. 5958–5963, May 2016. DOI: 10.1073/pnas.1522866113.

[128] F. Sanger, S. Nicklen, and A. R. Coulson, "DNA sequencing with chain-terminating inhibitors," en, *Proceedings of the National Academy of Sciences*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.74.12.5463.

[129] E. Eloe-Fadrosh, *Metagenome analysis of low-biomass samples: A jgi user facility perspective*, https://sma.nasa.gov/docs/default-source/sma-disciplines-and-programs/planetary-protection/emiley-eloe_nasa-planetary-protection-workshop---eloe-fadrosh-jgi.pdf?sfvrsn=f5c8d7f8_0, 2024.

[130] A. Rivers, "A quick reference guide to metagenome sequencing at JGI," Jun. 2018. DOI: 10.6084/m9.figshare.6653519.v1. [Online]. Available: https://figshare.com/articles/preprint/A_quick_reference_guide_to_metagenome_sequencing_at_JGI/6653519.

[131] EMBL-EBI, *Illumuna sequencing technology workflow*, https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/second-generation-sequencing/illumina-sequencing, Accessed: 2025-05-23, 2025.

[132] Competition and Markets Authority, *Anticipated acquisition by Illumina, Inc. of Pacific Biosciences of California, Inc. Provisional findings report*, 2019.

[133] D. F. R. Doud, R. M. Bowers, F. Schulz, M. De Raad, K. Deng, A. Tarver, E. Glasgow, K. Vander Meulen, B. Fox, S. Deutsch, Y. Yoshikuni, T. Northen, B. P. Hedlund, S. W. Singer, N. Ivanova, and T. Woyke, "Function-driven single-cell genomics uncovers cellulose-degrading bacteria from the rare biosphere," en, *The ISME Journal*, vol. 14, no. 3, pp. 659–675, Mar. 2020, ISSN: 1751-7370. DOI: 10.1038/s41396-019-0557-y.

[134] T. Woyke, D. F. R. Doud, and F. Schulz, "The trajectory of microbial single-cell sequencing," en, *Nature Methods*, vol. 14, no. 11, pp. 1045–1054, Nov. 2017, ISSN: 1548-7105. DOI: 10.1038/nmeth.4469.

[135] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner, "The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools," *Nucleic Acids Research*, vol. 41, no. D1, pp. D590–D596, Jan. 2013, ISSN: 0305-1048. DOI: 10.1093/nar/gks1219.

[136] E. J. Chamberlain, J. P. Balmonte, A. Torstensson, A. A. Fong, P. Snoeijs-Leijonmalm, and J. S. Bowman, "Impacts of sea ice melting procedures on measurements of microbial community structure," *Elementa: Science of the Anthropocene*, vol. 10, no. 1, p. 00 017, Dec. 2022, ISSN: 2325-1026. DOI: 10.1525/elementa.2022.00017.

[137] M. S. Rappé and S. J. Giovannoni, "The uncultured microbial majority," eng, *Annual Review of Microbiology*, vol. 57, pp. 369–394, 2003, ISSN: 0066-4227. DOI: 10.1146/annurev.micro.57.030502.090759.

[138] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," en, *Nature Biotechnology*, vol. 35, no. 9, pp. 833–844, Sep. 2017, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3935.

[139] Vaginal Microbiome Consortium (additional members), J. P. Brooks, D. J. Edwards, M. D. Harwich, M. C. Rivera, J. M. Fettweis, M. G. Serrano, R. A. Reris, N. U. Sheth, B. Huang, P. Girerd, J. F. Strauss, K. K. Jefferson, and G. A. Buck, "The truth about metagenomics: Quantifying and counteracting bias in 16S rRNA studies," en, *BMC Microbiology*, vol. 15, no. 1, p. 66, Dec. 2015, ISSN: 1471-2180. DOI: 10.1186/s12866-015-0351-6.

[140] A. Conesa, P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szcześniak, D. J. Gaffney, L. L. Elo, X. Zhang, and A. Mortazavi, "A survey of best practices for RNA-seq data analysis," en, *Genome Biology*, vol. 17, no. 1, p. 13, Dec. 2016, ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8.

[141] M. Dayhoff, R. Schwartz, and B. Orcutt, "A model of evolutionary change in proteins," *Atlas of Protein Sequence and Structure*, vol. 5, "345–352", 1978.

[142] J. L. Espinoza and C. L. Dupont, "VEBA: A modular end-to-end suite for in silico recovery, clustering, and analysis of prokaryotic, microeukaryotic, and viral genomes from metagenomes," *BMC Bioinformatics*, vol. 23, no. 1, p. 419, Oct. 2022, ISSN: 1471-2105. DOI: 10.1186/s12859-022-04973-8.

[143] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki, L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell, "NCBI prokaryotic genome annotation pipeline," *Nucleic Acids Research*, vol. 44, no. 14, pp. 6614–6624, Aug. 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkw569.

[144] A. Clum, M. Huntemann, B. Bushnell, B. Foster, B. Foster, S. Roux, P. P. Hajek, N. Varghese, S. Mukherjee, T. B. K. Reddy, C. Daum, Y. Yoshinaga, R. O'Malley, R. Seshadri, N. C. Kyrpides, E. A. Eloe-Fadrosh, I.-M. A. Chen, A. Copeland, and N. N. Ivanova, "DOE JGI Metagenome Workflow," *mSystems*, vol. 6, no. 3, e00804–20, ISSN: 2379-5077. DOI: 10.1128/mSystems.00804-20.

[145] M. Hosseini, D. Pratas, and A. Pinho, "A Survey on Data Compression Methods for Biological Sequences," en, *Information-an International Interdisciplinary Journal*, vol. 7, no. 4, p. 56, Oct. 2016, ISSN: 2078-2489. DOI: 10.3390/info7040056.

[146] B. Bushnell, *"BBMap: A Fast, Accurate, Splice-Aware Aligner"*, 2014.

[147] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," en, *Journal of Computational Biology*, vol. 19, no. 5, pp. 455–477, May 2012, ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2012.0021.

[148] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaSPAdes: A new versatile metagenomic assembler," en, *Genome Research*, vol. 27, no. 5, pp. 824–834, May 2017, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.213959.116.

[149] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, "A whole-genome assembly of drosophila," *Science*, vol. 287, no. 5461, pp. 2196–2204, 2000. DOI: 10.1126/science.287.5461.2196.

[150] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaSPAdes: A new versatile metagenomic assembler," en, *Genome Research*, vol. 27, no. 5, pp. 824–834, May 2017, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.213959.116.

[151] S. F. AltschuP, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," en, p. 8, 1990.

[152] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, Mar. 1970, ISSN: 0022-2836. DOI: 10.1016/0022-2836(70)90057-4.

[153] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, Mar. 1981, ISSN: 0022-2836. DOI: 10.1016/0022-2836(81)90087-5.

[154] M. Farrar, "Striped Smith-Waterman speeds database searches six times over other SIMD implementations," en, *Bioinformatics (Oxford, England)*, vol. 23, no. 2, pp. 156–161, Jan. 2007, ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btl582.

[155] R. C. Edgar, "MUSCLE: A multiple sequence alignment method with reduced time and space complexity," *BMC Bioinformatics*, vol. 5, no. 1, p. 113, Aug. 2004, ISSN: 1471-2105. DOI: 10.1186/1471-2105-5-113.

[156] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega," *Molecular Systems Biology*, vol. 7, no. 1, p. 539, Jan. 2011, ISSN: 1744-4292. DOI: 10.1038/msb.2011.75.

[157] K. Katoh, "MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform," en, *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, ISSN: 13624962. DOI: 10.1093/nar/gkf436.

[158] M. Alser, J. Rotman, D. Deshpande, K. Taraszka, H. Shi, P. I. Baykal, H. T. Yang, V. Xue, S. Knyazev, B. D. Singer, B. Balliu, D. Koslicki, P. Skums, A. Zelikovsky, C. Alkan, O. Mutlu, and S. Mangul, "Technology dictates algorithms: Recent developments in read alignment," en, *Genome Biology*, vol. 22, no. 1, pp. 1–34, Dec. 2021, ISSN: 1474-760X. DOI: 10.1186/s13059-021-02443-7.

[159] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, Jul. 2009, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp324.

[160] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, no. 3, R25, Mar. 2009, ISSN: 1474-760X. DOI: 10.1186/gb-2009-10-3-r25.

[161] S. R. Eddy, "Profile hidden Markov models," en, *Bioinformatics (Oxford, England)*, vol. 14, no. 9, pp. 755–763, Oct. 1998, ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/14.9.755.

[162] G. David Forney Jr, "The Viterbi Algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, Mar. 1973.

[163] L. E. Baum, "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes," en,

[164] J. Besemer and M. Borodovsky, "GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses," en, *Nucleic Acids Research*, vol. 33, no. Web Server, W451–W454, Jul. 2005, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gki487.

[165] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins," en, *Nucleic Acids Research*, vol. 35, no. Database, pp. D61–D65, Jan. 2007, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkl842.

[166] P. J. Keeling, F. Burki, H. M. Wilcox, B. Allam, E. E. Allen, L. A. Amaral-Zettler, E. V. Armbrust, J. M. Archibald, A. K. Bharti, C. J. Bell, B. Beszteri, K. D. Bidle, C. T. Cameron, L. Campbell, D. A. Caron, R. A. Cattolico, J. L. Collier, K. Coyne, S. K. Davy, P. Deschamps, S. T. Dyhrman, B. Edvardsen, R. D. Gates, C. J. Gobler, S. J. Greenwood, S. M. Guida, J. L. Jacobi, K. S. Jakobsen, E. R. James, B. Jenkins, U. John, M. D. Johnson, A. R. Juhl, A. Kamp, L. A. Katz, R. Kiene, A. Kudryavtsev, B. S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A. M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M. A. Moran, S. Murray, G. Nadathur, S. Nagai, P. B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M. C. Posewitz, K. Rengefors, G. Romano, M. E. Rumpho, T. Rynearson, K. B. Schilling, D. C. Schroeder, A. G. B. Simpson, C. H. Slamovits, D. R. Smith, G. J. Smith, S. R. Smith, H. M. Sosik, P. Stief, E. Theriot, S. N. Twary, P. E.

Umale, D. Vaulot, B. Wawrik, G. L. Wheeler, W. H. Wilson, Y. Xu, A. Zingone, and A. Z. Worden, "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing," en, *PLoS Biology*, vol. 12, no. 6, R. G. Roberts, Ed., e1001889, Jun. 2014, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001889.

[167] T. Klemetsen, I. A. Raknes, J. Fu, A. Agafonov, S. V. Balasundaram, G. Tartari, E. Robertsen, and N. P. Willassen, "The MAR databases: Development and implementation of databases specific for marine metagenomics," en, *Nucleic Acids Research*, vol. 46, no. D1, pp. D692–D699, Jan. 2018, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkx1036.

[168] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn, and A. Bateman, "Pfam: The protein families database in 2021," *Nucleic Acids Research*, vol. 49, no. D1, pp. D412–D419, Jan. 2021, ISSN: 0305-1048. DOI: 10.1093/nar/gkaa913.

[169] D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser, "Prodigal: Prokaryotic gene recognition and translation initiation site identification," *BMC Bioinformatics*, vol. 11, no. 1, p. 119, Mar. 2010, ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-119.

[170] A. V. Lukashin and M. Borodovsky, "GeneMark.hmm: New solutions for gene finding," *Nucleic Acids Research*, vol. 26, no. 4, pp. 1107–1115, Feb. 1998, ISSN: 0305-1048. DOI: 10.1093/nar/26.4.1107.

[171] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, Jan. 2016, ISSN: 0305-1048. DOI: 10.1093/nar/gkv1070.

[172] D. H. Haft, B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White, "TIGRFAMs: A protein family resource for the functional identification of proteins," eng, *Nucleic Acids Research*, vol. 29, no. 1, pp. 41–43, Jan. 2001, ISSN: 1362-4962. DOI: 10.1093/nar/29.1.41.

[173] M. Y. Galperin, R. Vera Alvarez, S. Karamycheva, K. S. Makarova, Y. Wolf, D. Landsman, and E. V. Koonin, "COG database update 2024," *Nucleic Acids Research*, vol. 53, no. D1, pp. D356–D363, Jan. 2025, ISSN: 1362-4962. DOI: 10.1093/nar/gkae983.

[174] I. Letunic, T. Doerks, and P. Bork, "SMART 6: Recent updates and new developments," eng, *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D229–232, Jan. 2009, ISSN: 1362-4962. DOI: 10.1093/nar/gkn808.

[175] D. Wilson, R. Pethica, Y. Zhou, C. Talbot, C. Vogel, M. Madera, C. Chothia, and J. Gough, "SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny," eng, *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D380–386, Jan. 2009, ISSN: 1362-4962. DOI: 10.1093/nar/gkn762.

[176] P. P. Chan and T. M. Lowe, "tRNAscan-SE: Searching for tRNA genes in genomic sequences," eng, *Methods in molecular biology (Clifton, N.J.)*, vol. 1962, pp. 1–14, 2019, ISSN: 1940-6029. DOI: 10.1007/978-1-4939-9173-0_1.

[177] E. P. Nawrocki and S. R. Eddy, "Infernal 1.1: 100-fold faster RNA homology searches," *Bioinformatics*, vol. 29, no. 22, pp. 2933–2935, Nov. 2013, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt509.

[178] N. Ontiveros-Palacios, E. Cooke, E. Nawrocki, S. Triebel, M. Marz, E. Rivas, S. Griffiths-Jones, A. Petrov, A. Bateman, and B. Sweeney, "Rfam 15: RNA families database in 2025," *Nucleic Acids Research*, vol. 53, no. D1, pp. D258–D267, Jan. 2025, ISSN: 1362-4962. DOI: 10.1093/nar/gkae1023.

[179] C. Bland, T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, and P. Hugenholtz, "CRISPR Recognition Tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats," *BMC Bioinformatics*, vol. 8, no. 1, p. 209, Jun. 2007, ISSN: 1471-2105. DOI: 10.1186/1471-2105-8-209.

[180] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments," en, *PLOS ONE*, vol. 5, no. 3, e9490, Mar. 2010, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0009490.

[181] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear, "IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era," *Molecular Biology and Evolution*, vol. 37, no. 5, pp. 1530–1534, May 2020, ISSN: 0737-4038. DOI: 10.1093/molbev/msaa015.

[182] F. A. Matsen, R. B. Kodner, and E. V. Armbrust, "Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree," en, *BMC Bioinformatics*, vol. 11, no. 1, p. 538, Dec. 2010, ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-538.

[183] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, "Adaptive seeds tame genomic sequence comparison," en, *Genome Research*, vol. 21, no. 3, pp. 487–493, Jan. 2011, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.113985.110.

[184] E. W. Sayers, J. Beck, E. E. Bolton, J. R. Brister, J. Chan, R. Connor, M. Feldgarden, A. M. Fine, K. Funk, J. Hoffman, S. Kannan, C. Kelly, W. Klimke, S. Kim, S. Lathrop, A. Marchler-Bauer, T. D. Murphy, C. O'Sullivan, E. Schmieder, Y. Skripchenko, A. Stine, F. Thibaud-Nissen, J. Wang, J. Ye, E. Zellers, V. A. Schneider, and K. D.

Pruitt, "Database resources of the National Center for Biotechnology Information in 2025," eng, *Nucleic Acids Research*, vol. 53, no. D1, pp. D20–D29, Jan. 2025, ISSN: 1362-4962. DOI: 10.1093/nar/gkae979.

[185] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," en, *Nature Biotechnology*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3988.

[186] D. E. Wood and S. L. Salzberg, "Kraken: Ultrafast metagenomic sequence classification using exact alignments," en, *Genome Biology*, vol. 15, no. 3, R46, 2014, ISSN: 1465-6906. DOI: 10.1186/gb-2014-15-3-r46.

[187] C. Bağcı, S. Patz, and D. H. Huson, "DIAMOND+MEGAN: Fast and easy taxonomic and functional analysis of short and long microbiome sequences," *Current Protocols*, vol. 1, no. 3, e59, 2021. DOI: https://doi.org/10.1002/cpz1.59.

[188] D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang, "MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, e7359, Jul. 2019. DOI: 10.7717/peerj.7359.

[189] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glöckner, "Application of tetranucleotide frequencies for the assignment of genomic fragments," *Environmental Microbiology*, vol. 6, no. 9, pp. 938–947, 2004. DOI: https://doi.org/10.1111/j.1462-2920.2004.00624.x.

[190] C. C. Laczny, T. Sternal, V. Plugaru, P. Gawron, A. Atashpendar, H. H. Margossian, S. Coronado, L. v. der Maaten, N. Vlassis, and P. Wilmes, "VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data," en, *Microbiome*, vol. 3, no. 1, p. 1, Dec. 2015, ISSN: 2049-2618. DOI: 10.1186/s40168-014-0066-1.

[191] J. N. Nissen, J. Johansen, R. L. Allesøe, C. K. Sønderby, J. J. A. Armenteros, C. H. Grønbech, L. J. Jensen, H. B. Nielsen, T. N. Petersen, O. Winther, and S. Rasmussen, "Improved metagenome binning and assembly using deep variational autoencoders," en, *Nature Biotechnology*, vol. 39, no. 5, pp. 555–560, May 2021, ISSN: 1546-1696. DOI: 10.1038/s41587-020-00777-4.

[192] Z. Wang, R. You, H. Han, W. Liu, F. Sun, and S. Zhu, "Effective binning of metagenomic contigs using contrastive multi-view representation learning," en, *Nature Communications*, vol. 15, no. 1, p. 585, Jan. 2024, ISSN: 2041-1723. DOI: 10.1038/s41467-023-44290-z.

[193] S. Pan, C. Zhu, X.-M. Zhao, and L. P. Coelho, "A deep siamese neural network improves metagenome-assembled genomes in microbiome datasets across different en-

vironments," en, *Nature Communications*, vol. 13, no. 1, p. 2326, Apr. 2022, ISSN: 2041-1723. DOI: 10.1038/s41467-022-29843-y.

[194] P. T. West, A. J. Probst, I. V. Grigoriev, B. C. Thomas, and J. F. Banfield, "Genome-reconstruction for eukaryotes from complex natural microbial communities," en, *Genome Research*, vol. 28, no. 4, pp. 569–580, Apr. 2018, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.228429.117.

[195] M. Karlicki, S. Antonowicz, and A. Karnkowska, "Tiara: Deep learning-based classification system for eukaryotic sequences," *Bioinformatics*, vol. 38, no. 2, pp. 344–350, Jan. 2022, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab672.

[196] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. Murat Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke, "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea," en, *Nature Biotechnology*, vol. 35, no. 8, pp. 725–731, Aug. 2017, ISSN: 1546-1696. DOI: 10.1038/nbt.3893.

[197] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "GTDB-tk: A toolkit to classify genomes with the genome taxonomy database," *Bioinformatics (Oxford, England)*, vol. 36, no. 6, pp. 1925–1927, Nov. 2019, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz848.

[198] W. C. Nelson, B. J. Tully, and J. M. Mobberley, "Biases in genome reconstruction from metagenomic data," *PeerJ*, vol. 8, e10119, Oct. 2020, ISSN: 2167-8359. DOI: 10.7717/peerj.10119.

[199] Y.-W. Wu, B. A. Simmons, and S. W. Singer, "MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets," *Bioinformatics*, vol. 32, no. 4, pp. 605–607, Feb. 2016, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv638.

[200] Z. Wang, R. You, H. Han, W. Liu, F. Sun, and S. Zhu, "Effective binning of metagenomic contigs using contrastive multi-view representation learning," en, *Nature Communications*, vol. 15, no. 1, p. 585, Jan. 2024, ISSN: 2041-1723. DOI: 10.1038/s41467-023-44290-z.

[201] L. F. W. Roesch, R. R. Fulthorpe, A. Riva, G. Casella, A. K. M. Hadwin, A. D. Kent, S. H. Daroub, F. A. O. Camargo, W. G. Farmerie, and E. W. Triplett, "Pyrosequencing enumerates and contrasts soil microbial diversity," *The ISME Journal*, vol. 1, no. 4, pp. 283–290, Aug. 2007, ISSN: 1751-7362. DOI: 10.1038/ismej.2007.53.

[202] V. Torsvik, L. Øvreås, and T. F. Thingstad, "Prokaryotic Diversity–Magnitude, Dynamics, and Controlling Factors," *Science*, vol. 296, no. 5570, pp. 1064–1066, May 2002. DOI: 10.1126/science.1071698.

[203] J. Gans, M. Wolinsky, and J. Dunbar, "Computational Improvements Reveal Great Bacterial Diversity and High Metal Toxicity in Soil," *Science*, vol. 309, no. 5739, pp. 1387–1390, Aug. 2005. DOI: 10.1126/science.1112665.

[204] P. D. Schloss and J. Handelsman, "Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness," *Applied and Environmental Microbiology*, vol. 71, no. 3, pp. 1501–1506, Mar. 2005. DOI: 10.1128/AEM.71.3.1501-1506.2005.

[205] C. Ricotta and J. Podani, "On some properties of the Bray-Curtis dissimilarity and their ecological meaning," *Ecological Complexity*, vol. 31, pp. 201–205, Sep. 2017, ISSN: 1476-945X. DOI: 10.1016/j.ecocom.2017.07.003.

[206] L. Orlóci, "Resemblance Functions," en, in *Multivariate Analysis in Vegetation Research*, L. Orlóci, Ed., Dordrecht: Springer Netherlands, 1975, pp. 24–62, ISBN: 978-94-017-5608-2. DOI: 10.1007/978-94-017-5608-2_2.

[207] L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, and J. F. Banfield, "A new view of the tree of life," en, *Nature Microbiology*, vol. 1, no. 5, pp. 1–6, Apr. 2016, ISSN: 2058-5276. DOI: 10.1038/nmicrobiol.2016.48.

[208] J. Shlens, "A tutorial on principal component analysis," *CoRR*, vol. abs/1404.1100, 2014. arXiv: 1404.1100. [Online]. Available: http://arxiv.org/abs/1404.1100.

[209] M. Richardson, *Principal Component Analysis*, en, http://www.dsc.ufcg.edu.br/~hmg/disciplinas/posgraduacao/rn-copin-2014.3/material/SignalProcPCA.pdf, 2009.

[210] *Modern Multidimensional Scaling* (Springer Series in Statistics), en. New York, NY: Springer, 2005, ISBN: 978-0-387-25150-9. DOI: 10.1007/0-387-28981-X.

[211] H. Zou, T. Hastie, and R. Tibshirani, "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006, ISSN: 1061-8600.

[212] N. Tsagkarakis, P. P. Markopoulos, G. Sklivanitis, and D. A. Pados, "L1-Norm Principal-Component Analysis of Complex Data," en, *IEEE Transactions on Signal Processing*, vol. 66, no. 12, pp. 3256–3267, Jun. 2018, ISSN: 1053-587X, 1941-0476. DOI: 10.1109/TSP.2018.2821641.

[213] A. Tharwat, "Independent component analysis: An introduction," *Applied Computing and Informatics*, vol. 17, no. 2, pp. 222–249, Jan. 2021, ISSN: 2634-1964, 2210-8327. DOI: 10.1016/j.aci.2018.08.006.

[214] F. W. Townes and B. E. Engelhardt, "Nonnegative spatial factorization applied to spatial genomics," en, *Nature Methods*, vol. 20, no. 2, pp. 229–238, Feb. 2023, ISSN: 1548-7105. DOI: 10.1038/s41592-022-01687-w.

[215] L. Van der Maaten and G. Hinton, "Visualising Data using t-SNE," en, *Journal of Machine Learning Research*, no. 9, pp. 2579–2605, 2008.

[216] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, en, Sep. 2020.

[217] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, Twenty-fifth Anniversay Commemorative Issue, vol. 37, pp. 52–65, Jan. 2013, ISSN: 0893-6080. DOI: 10.1016/j.neunet.2012.09.018.

[218] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," en, *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.290.5500.2319.

[219] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: Large-scale single-cell gene expression data analysis," *Genome Biology*, vol. 19, no. 1, p. 15, Feb. 2018, ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0.

[220] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome Datasets Are Compositional: And This Is Not Optional," English, *Frontiers in Microbiology*, vol. 8, Nov. 2017, ISSN: 1664-302X. DOI: 10.3389/fmicb.2017.02224.

[221] M. L. Calle, M. Pujolassos, and A. Susin, "Coda4microbiome: Compositional data analysis for microbiome cross-sectional and longitudinal studies," *BMC Bioinformatics*, vol. 24, no. 1, p. 82, Mar. 2023, ISSN: 1471-2105. DOI: 10.1186/s12859-023-05205-3.

[222] J. Aitchison, "The statistical analysis of compositional data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 44, no. 2, pp. 139–177, 1982, ISSN: 00359246. [Online]. Available: http://www.jstor.org/stable/2345821 (visited on 05/23/2025).

[223] H. Lin and S. D. Peddada, "Analysis of microbial compositions: A review of normalization and differential abundance analysis," en, *npj Biofilms and Microbiomes*, vol. 6, no. 1, pp. 1–13, Dec. 2020, ISSN: 2055-5008. DOI: 10.1038/s41522-020-00160-w.

[224] Z. Sun, S. Huang, M. Zhang, Q. Zhu, N. Haiminen, A. P. Carrieri, Y. Vázquez-Baeza, L. Parida, H.-C. Kim, R. Knight, and Y.-Y. Liu, "Challenges in Benchmarking Metagenomic Profilers," *Nature methods*, vol. 18, no. 6, pp. 618–626, Jun. 2021, ISSN: 1548-7091. DOI: 10.1038/s41592-021-01141-3.

[225] M. S. Matchado, M. Lauber, S. Reitmeier, T. Kacprowski, J. Baumbach, D. Haller, and M. List, "Network analysis methods for studying microbial communities: A mini review," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 2687–2698, Jan. 2021, ISSN: 2001-0370. DOI: 10.1016/j.csbj.2021.05.001.

[226] J. Friedman and E. J. Alm, "Inferring Correlation Networks from Genomic Survey Data," en, *PLoS Computational Biology*, vol. 8, no. 9, C. Von Mering, Ed., e1002687, Sep. 2012, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002687.

[227] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, Dec. 2008, ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559.

[228] M. Farhadian, S. A. Rafat, B. Panahi, and C. Mayack, "Weighted gene co-expression network analysis identifies modules and functionally enriched pathways in the lactation process," en, *Scientific Reports*, vol. 11, no. 1, p. 2367, Jan. 2021, ISSN: 2045-2322. DOI: 10.1038/s41598-021-81888-z.

[229] B. D. Jameson, S. A. Murdock, Q. Ji, C. J. Stevens, D. S. Grundle, and S. Kim Juniper, "Network analysis of 16S rRNA sequences suggests microbial keystone taxa contribute to marine N2O cycling," en, *Communications Biology*, vol. 6, no. 1, pp. 1–14, Feb. 2023, ISSN: 2399-3642. DOI: 10.1038/s42003-023-04597-5.

[230] G. Pei, L. Chen, and W. Zhang, "Chapter Nine - WGCNA Application to Proteomic and Metabolomic Data Analysis," in *Methods in Enzymology*, ser. Proteomics in Biology, Part A, A. K. Shukla, Ed., vol. 585, Academic Press, Jan. 2017, pp. 135–158. DOI: 10.1016/bs.mie.2016.09.016.

[231] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999. DOI: 10.1126/science.286.5439.509.

[232] G. G. Lemoine, M.-P. Scott-Boyer, B. Ambroise, O. Périn, and A. Droit, "GWENA: Gene co-expression networks analysis and extended modules characterization in a single Bioconductor package," *BMC Bioinformatics*, vol. 22, no. 1, p. 267, May 2021, ISSN: 1471-2105. DOI: 10.1186/s12859-021-04179-4.

[233] S. Roy, S. Lagree, Z. Hou, J. A. Thomson, R. Stewart, and A. P. Gasch, "Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks," en, *PLOS Computational Biology*, vol. 9, no. 10, e1003252, Oct. 2013, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003252.

[234] W. Saelens, R. Cannoodt, and Y. Saeys, "A comprehensive evaluation of module detection methods for gene expression data," en, *Nature Communications*, vol. 9, no. 1, p. 1090, Mar. 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-018-03424-4.

[235] F. Jiang, H. Zhou, and H. Shen, "Identification of Critical Biomarkers and Immune Infiltration in Rheumatoid Arthritis Based on WGCNA and LASSO Algorithm," eng, *Frontiers in Immunology*, vol. 13, p. 925 695, 2022, ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.925695.

[236] X. Kong, H. Sun, K. Wei, L. Meng, X. Lv, C. Liu, F. Lin, and X. Gu, "WGCNA combined with machine learning algorithms for analyzing key genes and immune cell infiltration in heart failure due to ischemic cardiomyopathy," *Frontiers in Cardiovascular Medicine*, vol. 10, p. 1 058 834, Mar. 2023, ISSN: 2297-055X. DOI: 10.3389/fcvm.2023.1058834.

[237] M. Shaffer, K. Thurimella, J. D. Sterrett, and C. A. Lozupone, "SCNIC: Sparse correlation network investigation for compositional data," en, *Molecular Ecology Resources*, vol. 23, no. 1, pp. 312–325, 2023, ISSN: 1755-0998. DOI: 10.1111/1755-0998.13704.

[238] B. Velten, J. M. Braunger, R. Argelaguet, D. Arnol, J. Wirbel, D. Bredikhin, G. Zeller, and O. Stegle, "Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO," en, *Nature Methods*, Jan. 2022, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-021-01343-9.

[239] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, "Sparse and Compositionally Robust Inference of Microbial Ecological Networks," en, *PLOS Computational Biology*, vol. 11, no. 5, e1004226, May 2015, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004226.

[240] G. Tikhonov, O. H. Opedal, N. Abrego, A. Lehikoinen, M. M. J. de Jonge, J. Oksanen, and O. Ovaskainen, "Joint species distribution modelling with the r-package Hmsc," en, *Methods in Ecology and Evolution*, vol. 11, no. 3, pp. 442–447, 2020, ISSN: 2041-210X. DOI: 10.1111/2041-210X.13345.

[241] M. Granskog and O. Muller, "A peek beneath the surface of arctic sea ice," *EU Research*, vol. 37, pp. 38–39, 2024, HAVOC project contribution. [Online]. Available: https://issuu.com/euresearcher/docs/havoc_eur37_h_res.

[242] S. Mårtensson, H. E. M. Meier, P. Pemberton, and J. Haapala, "Ridged sea ice characteristics in the Arctic from a coupled multicategory sea ice model," en, *Jour-*

nal of Geophysical Research: Oceans, vol. 117, no. C8, 2012, ISSN: 2156-2202. DOI: 10.1029/2010JC006936.

[243] G. W. Timco and R. P. Burden, "An analysis of the shapes of sea ice ridges," *Cold Regions Science and Technology*, vol. 25, no. 1, pp. 65–77, Jan. 1997, ISSN: 0165-232X. DOI: 10.1016/S0165-232X(96)00017-1.

[244] L. Strub-Klein and D. Sudom, "A comprehensive analysis of the morphology of first-year sea ice ridges," *Cold Regions Science and Technology*, vol. 82, pp. 94–109, Oct. 2012, ISSN: 0165-232X. DOI: 10.1016/j.coldregions.2012.05.014.

[245] R. Gradinger, B. Bluhm, and K. Iken, "Arctic sea-ice ridges—Safe heavens for sea-ice fauna during periods of extreme ice melt?" *Deep Sea Research Part II: Topical Studies in Oceanography*, Observations and Exploration of the Arctic's Canada Basin and the Chukchi Sea: the Hidden Ocean and RUSALCA Expeditions, vol. 57, no. 1, pp. 86–95, Jan. 2010, ISSN: 0967-0645. DOI: 10.1016/j.dsr2.2009.08.008.

[246] E. E. Syvertsen, "Ice algae in the Barents Sea: Types of assemblages, origin, fate and role in the ice-edge phytoplankton bloom," en, *Polar Research*, vol. 10, no. 1, pp. 277–288, Jan. 1991, ISSN: 1751-8369. DOI: 10.3402/polar.v10i1.6746.

[247] M. Fernández-Méndez, L. M. Olsen, H. M. Kauko, A. Meyer, A. Rösel, I. Merkouriadi, C. J. Mundy, J. K. Ehn, A. M. Johansson, P. M. Wagner, Ervik, B. K. Sorrell, P. Duarte, A. Wold, H. Hop, and P. Assmy, "Algal Hot Spots in a Changing Arctic Ocean: Sea-Ice Ridges and the Snow-Ice Interface," *Frontiers in Marine Science*, vol. 5, 2018, ISSN: 2296-7745.

[248] T. O. Delmont, M. Gaia, D. D. Hinsinger, P. Frémont, C. Vanni, A. Fernandez-Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. Da Silva, M. Wessner, B. Noel, J.-M. Aury, S. Sunagawa, S. G. Acinas, P. Bork, E. Karsenti, C. Bowler, C. Sardet, L. Stemmann, C. de Vargas, P. Wincker, M. Lescot, M. Babin, G. Gorsky, N. Grimsley, L. Guidi, P. Hingamp, O. Jaillon, S. Kandels, D. Iudicone, H. Ogata, S. Pesant, M. B. Sullivan, F. Not, K.-B. Lee, E. Boss, G. Cochrane, M. Follows, N. Poulton, J. Raes, M. Sieracki, S. Speich, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, and O. Jaillon, "Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean," *Cell Genomics*, vol. 2, no. 5, p. 100 123, May 2022, ISSN: 2666-979X. DOI: 10.1016/j.xgen.2022.100123.

[249] W. Boulton, "Metagenome-assembled-genomes recovered from the Arctic drift expedition MOSAiC: spreadsheet of MAGs, sample metadata, and sequencing information," May 2024. DOI: 10.6084/m9.figshare.25765707.v1. [Online]. Available: https://figshare.com/articles/dataset/Metagenome-assembled-genomes_recovered_from_the_Arctic_drift_expedition_MOSAiC_spreadsheet_of_MAGs_sample_metadata_and_sequencing_information/25765707.

[250] M. Oggier, E. Salganik, L. M. Whitmore, A. A. Fong, C. J. M. Hoppe, R. Rember, K. V. Høyland, D. V. Divine, R. Gradinger, S. W. Fons, K. Abrahamsson, A. M. Aguilar-Islas, M. Angelopoulos, S. Arndt, J. P. Balmonte, D. Bozzato, J. S. Bowman, G. Castellani, E. Chamberlain, J. Creamean, A. DAngelo, E. Damm, A. Dumitrascu, S. L. Eggers, J. Gardner, L. Grosfeld, J. Haapala, A. Immerz, N. Kolabutin, B. A. Lange, R. Lei, C. M. Marsay, S. Maus, O. Müller, L. M. Olsen, A. Nuibom, J. Ren, A. Rinke, I. Sheikin, E. Shimanchuk, P. Snoeijs-Leijonmalm, S. Spahic, J. Stefels, S. Torres-Valdés, A. Torstensson, A. Ulfsbo, J. Verdugo, M. Vortkamp, L. Wang, M. Webster, L. Wischnewski, and M. A. Granskog, *First-year sea-ice salinity, temperature, density, oxygen and hydrogen isotope composition from the main coring site (MCS-FYI) during MOSAiC legs 1 to 4 in 2019/2020*, 2023. DOI: 10.1594/PANGAEA.956732.

[251] M. Oggier, E. Salganik, L. M. Whitmore, A. A. Fong, C. J. M. Hoppe, R. Rember, K. V. Høyland, R. Gradinger, D. V. Divine, S. W. Fons, K. Abrahamsson, A. M. Aguilar-Islas, M. Angelopoulos, S. Arndt, J. P. Balmonte, D. Bozzato, J. S. Bowman, G. Castellani, E. Chamberlain, J. Creamean, A. DAngelo, E. Damm, A. Dumitrascu, L. Eggers, J. Gardner, L. Grosfeld, J. Haapala, A. Immerz, N. Kolabutin, B. A. Lange, R. Lei, C. M. Marsay, S. Maus, L. M. Olsen, O. Müller, A. Nuibom, J. Ren, A. Rinke, I. Sheikin, E. Shimanchuk, P. Snoeijs-Leijonmalm, S. Spahic, J. Stefels, S. Torres-Valdés, A. Torstensson, A. Ulfsbo, J. Verdugo, M. Vortkamp, L. Wang, M. Webster, L. Wischnewski, and M. A. Granskog, *Second-year sea-ice salinity, temperature, density, oxygen and hydrogen isotope composition from the main coring site (MCS-SYI) during MOSAiC legs 1 to 4 in 2019/2020*, 2023. DOI: 10.1594/PANGAEA.959830.

[252] R. Lei, B. Cheng, M. Hoppmann, F. Zhang, G. Zuo, J. K. Hutchings, L. Lin, M. Lan, H. Wang, J. Regnery, T. Krumpen, J. Haapala, B. Rabe, D. K. Perovich, and M. Nicolaus, "Seasonality and timing of sea ice mass balance and heat fluxes in the Arctic transpolar drift during 2019–2020," *Elementa: Science of the Anthropocene*, vol. 10, no. 1, p. 000 089, Jul. 2022, ISSN: 2325-1026. DOI: 10.1525/elementa.2021.000089.

[253] A. R. Macfarlane, M. Schneebeli, R. Dadic, A. Tavri, A. Immerz, C. Polashenski, D. Krampe, D. Clemens-Sewall, D. N. Wagner, D. K. Perovich, H. Henna-Reetta, I. Raphael, I. Matero, J. Regnery, M. M. Smith, M. Nicolaus, M. Jaggi, M. Oggier, M. A. Webster, M. Lehning, N. Kolabutin, P. Itkin, R. Naderpour, R. Pirazzini, S. Hämmerle, S. Arndt, and S. Fons, "A Database of Snow on Sea Ice in the Central Arctic Collected during the MOSAiC expedition," en, *Scientific Data*, vol. 10, no. 1, p. 398, Jun. 2023, ISSN: 2052-4463. DOI: 10.1038/s41597-023-02273-1.

[254] A. Clum, M. Huntemann, B. Bushnell, B. Foster, B. Foster, S. Roux, P. P. Hajek, N. Varghese, S. Mukherjee, T. B. K. Reddy, C. Daum, Y. Yoshinaga, R. O'Malley, R. Seshadri, N. C. Kyrpides, E. A. Eloe-Fadrosh, I.-M. A. Chen, A. Copeland, and N. N. Ivanova, "DOE JGI Metagenome Workflow," *mSystems*, vol. 6, no. 3, e00804–20, ISSN: 2379-5077. DOI: 10.1128/mSystems.00804-20.

[255] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," en, *Genome Research*, vol. 25, no. 7, pp. 1043–1055, Jan. 2015, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.186072.114.

[256] P. J. Keeling, F. Burki, H. M. Wilcox, B. Allam, E. E. Allen, L. A. Amaral-Zettler, E. V. Armbrust, J. M. Archibald, A. K. Bharti, C. J. Bell, B. Beszteri, K. D. Bidle, C. T. Cameron, L. Campbell, D. A. Caron, R. A. Cattolico, J. L. Collier, K. Coyne, S. K. Davy, P. Deschamps, S. T. Dyhrman, B. Edvardsen, R. D. Gates, C. J. Gobler, S. J. Greenwood, S. M. Guida, J. L. Jacobi, K. S. Jakobsen, E. R. James, B. Jenkins, U. John, M. D. Johnson, A. R. Juhl, A. Kamp, L. A. Katz, R. Kiene, A. Kudryavtsev, B. S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A. M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M. A. Moran, S. Murray, G. Nadathur, S. Nagai, P. B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M. C. Posewitz, K. Rengefors, G. Romano, M. E. Rumpho, T. Rynearson, K. B. Schilling, D. C. Schroeder, A. G. B. Simpson, C. H. Slamovits, D. R. Smith, G. J. Smith, S. R. Smith, H. M. Sosik, P. Stief, E. Theriot, S. N. Twary, P. E. Umale, D. Vaulot, B. Wawrik, G. L. Wheeler, W. H. Wilson, Y. Xu, A. Zingone, and A. Z. Worden, "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing," en, *PLoS Biology*, vol. 12, no. 6, R. G. Roberts, Ed., e1001889, Jun. 2014, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1001889.

[257] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation," eng, *Nucleic Acids Research*, vol. 44, no. D1, pp. D733–745, Jan. 2016, ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

[258] S. Hofmeyr, R. Egan, E. Georganas, A. C. Copeland, R. Riley, A. Clum, E. Eloe-Fadrosh, S. Roux, E. Goltsman, A. Buluç, D. Rokhsar, L. Oliker, and K. Yelick, "Terabase-scale metagenome coassembly with MetaHipMer," en, *Scientific Reports*, vol. 10, no. 1, p. 10 689, Jul. 2020, ISSN: 2045-2322. DOI: 10.1038/s41598-020-67416-5.

[259] P. Saary, A. L. Mitchell, and R. D. Finn, "Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC," en, *Genome Biology*,

vol. 21, no. 1, p. 244, Dec. 2020, ISSN: 1474-760X. DOI: 10.1186/s13059-020-02155-4.

[260] H. Alexander, S. K. Hu, A. I. Krinos, M. Pachiadaki, B. J. Tully, C. J. Neely, and T. Reiter, "Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton," *mBio*, vol. 14, no. 6, e01676–23, Nov. 2023. DOI: 10.1128/mbio.01676-23.

[261] J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, and M. Punta, "Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions," *Nucleic Acids Research*, vol. 41, no. 12, e121, Jul. 2013, ISSN: 0305-1048. DOI: 10.1093/nar/gkt263.

[262] I. Sillitoe, A. L. Cuff, B. H. Dessailly, N. L. Dawson, N. Furnham, D. Lee, J. G. Lees, T. E. Lewis, R. A. Studer, R. Rentzsch, C. Yeats, J. M. Thornton, and C. A. Orengo, "New functional families (FunFams) in CATH to improve the mapping of conserved functional sites to 3D structures," *Nucleic Acids Research*, vol. 41, no. Database issue, pp. D490–D498, Jan. 2013, ISSN: 0305-1048. DOI: 10.1093/nar/gks1211.

[263] P. P. Chan and T. M. Lowe, "tRNAscan-SE: Searching for tRNA genes in genomic sequences," *Methods in molecular biology (Clifton, N.J.)*, vol. 1962, pp. 1–14, 2019, ISSN: 1064-3745. DOI: 10.1007/978-1-4939-9173-0_1.

[264] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: Tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000, ISSN: 1061-4036. DOI: 10.1038/75556.

[265] D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. V. Auken, J. Ramsey, and D. A. Siegele, "The Gene Ontology knowledgebase in 2023," en,

[266] E. Levy Karin, M. Mirdita, and J. Söding, "MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics," en, *Microbiome*, vol. 8, no. 1, p. 48, Dec. 2020, ISSN: 2049-2618. DOI: 10.1186/s40168-020-00808-x.

[267] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, and C. H. Wu, "UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches," *Bioinformatics*, vol. 31, no. 6, pp. 926–932, Mar. 2015, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu739.

[268] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez, "InterProScan: Protein domains identifier," *Nucleic Acids Research*, vol. 33,

no. Web Server issue, W116–W120, Jul. 2005, ISSN: 0305-1048. DOI: 10.1093/nar/gki442.

[269] M. Manni, M. R. Berkeley, M. Seppey, F. A. Simão, and E. M. Zdobnov, "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes," *Molecular Biology and Evolution*, vol. 38, no. 10, pp. 4647–4654, Oct. 2021, ISSN: 1537-1719. DOI: 10.1093/molbev/msab199.

[270] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins," en, *Nucleic Acids Research*, vol. 35, no. Database, pp. D61–D65, Jan. 2007, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkl842.

[271] M. Tarailo-Graovac and N. Chen, "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences," en, *Current Protocols in Bioinformatics*, vol. 25, no. 1, pp. 4.10.1–4.10.14, 2009, ISSN: 1934-340X. DOI: 10.1002/0471250953.bi0410s25.

[272] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, Apr. 2002, ISSN: 0305-1048.

[273] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, "High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries," en, *Nature Communications*, vol. 9, no. 1, p. 5114, Dec. 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-018-07641-9.

[274] R. Riley, R. M. Bowers, A. P. Camargo, A. Campbell, R. Egan, E. A. Eloe-Fadrosh, B. Foster, S. Hofmeyr, M. Huntemann, M. Kellom, J. A. Kimbrel, L. Oliker, K. Yelick, J. Pett-Ridge, A. Salamov, N. J. Varghese, and A. Clum, "Terabase-Scale Coassembly of a Tropical Soil Microbiome," *Microbiology Spectrum*, vol. 11, no. 4, e00200–23, Jun. 2023. DOI: 10.1128/spectrum.00200-23.

[275] D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam, "MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph," en, *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, May 2015, ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btv033.

[276] J. N. Nissen, J. Johansen, R. L. Allesøe, C. K. Sønderby, J. J. A. Armenteros, C. H. Grønbech, L. J. Jensen, H. B. Nielsen, T. N. Petersen, O. Winther, and S. Rasmussen, "Improved metagenome binning and assembly using deep variational autoencoders," en, *Nature Biotechnology*, vol. 39, no. 5, pp. 555–560, May 2021, ISSN: 1546-1696. DOI: 10.1038/s41587-020-00777-4.

[277] A. Otte, J. C. Winder, L. Deng, J. Schmutz, J. Jenkins, I. V. Grigoriev, A. Hopes, and T. Mock, "The diatom Fragilariopsis cylindrus: A model alga to understand cold-adapted life," en, *Journal of Phycology*, vol. 59, no. 2, pp. 301–306, 2023, ISSN: 1529-8817. DOI: 10.1111/jpy.13325.

[278] Z. M. McKie-Krisberg and R. W. Sanders, "Phagotrophy by the picoeukaryotic green alga Micromonas: Implications for Arctic Oceans," *The ISME Journal*, vol. 8, no. 10, pp. 1953–1961, Oct. 2014, ISSN: 1751-7362. DOI: 10.1038/ismej.2014.16.

[279] W. Boulton, A. Salamov, I. V. Grigoriev, S. Calhoun, K. LaButti, R. Riley, K. Barry, A. A. Fong, M. Hoppe, K. Metfies, K. Oetjen, L. Eggers, O. Müller, J. Gardner, M. A. Granskog, A. Torstensson, M. Oggier, A. Larsen, G. Bratbak, A. Toseland, R. M. Leggett, V. Moulton, and T. Mock, *Metagenome-assembled-genomes recovered from the Arctic drift expedition MOSAiC*. 2024. DOI: https://doi.org/10.6084/m9.figshare.27879576.

[280] *NCBI BioProject*, 2024. DOI: http://identifiers.org/bioproject:PRJNA1160706.

[281] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, J. Sundaramurthi, J. Lee, M. Kandimalla, I.-M. A. Chen, N. C. Kyrpides, and T. B. K. Reddy, "Genomes On-Line Database (GOLD) v.8: Overview and updates," *Nucleic Acids Research*, vol. 49, no. D1, pp. D723–D733, Jan. 2021, ISSN: 0305-1048. DOI: 10.1093/nar/gkaa983.

[282] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497145.

[283] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497146.

[284] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497147.

[285] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497148.

[286] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497149.

[287] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497150.

[288] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497151.

[289] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497152.

[290] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497153.

[291] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497154.

[292] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497155.

[293] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497157.

[294] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497160.

[295] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497161.

[296] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497165.

[297] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497166.

[298] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497167.

[299] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497168.

[300] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497174.

[301] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497178.

[302] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497181.

[303] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497183.

[304] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497184.

[305] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497185.

[306] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497186.

[307] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497187.

[308] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497189.

[309] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497190.

[310] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497191.

[311] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497192.

[312] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497193.

[313] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497194.

[314] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497195.

[315] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497196.

[316] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497197.

[317] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497198.

[318] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497199.

[319] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497200.

[320] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497201.

[321] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497202.

[322] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497203.

[323] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497204.

[324] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497205.

[325] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497207.

[326] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497209.

[327] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497213.

[328] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497214.

[329] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497216.

[330] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP497220.

[331] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506343.

[332] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506344.

[333] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506347.

[334] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506348.

[335] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506350.

[336] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506351.

[337] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506352.

[338] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506353.

[339] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506355.

[340] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506356.

[341] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506357.

[342] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506359.

[343] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506366.

[344] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506368.

[345] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506369.

[346] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506371.

[347] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506373.

[348] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506374.

[349] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506375.

[350] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506376.

[351] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506379.

[352] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506382.

[353] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506389.

[354] *NCBI sequence read archive*, 2024. DOI: http://identifiers.org/insdc.sra:SRP506779.

[355] J. A. Raymond, C. Fritsen, and K. Shen, "An ice-binding protein from an Antarctic sea ice bacterium," *FEMS Microbiology Ecology*, vol. 61, no. 2, pp. 214–221, Aug. 2007, ISSN: 0168-6496. DOI: 10.1111/j.1574-6941.2007.00345.x.

[356] A. Krell, B. Bánk, D. Gerhard, G. Gernot, V. Klaus, and T. Mock, "A new class of ice-binding proteins discovered in a salt-stress-induced cDNA library of the psychrophilic diatom Fragilariopsis cylindrus (Bacillariophyceae)," *European Journal of Phycology*, vol. 43, no. 4, pp. 423–433, Nov. 2008, ISSN: 0967-0262. DOI: 10.1080/09670260802348615.

[357] Y. Yeh and R. E. Feeney, "Antifreeze Proteins: Structures and Mechanisms of Function," *Chemical Reviews*, vol. 96, no. 2, pp. 601–618, Jan. 1996, ISSN: 0009-2665. DOI: 10.1021/cr950260c.

[358] T. D. R. Vance, M. Bayer-Giraldi, P. L. Davies, and M. Mangiagalli, "Ice-binding proteins and the 'domain of unknown function' 3494 family," en, *The FEBS Journal*, vol. 286, no. 5, pp. 855–873, 2019, ISSN: 1742-4658. DOI: 10.1111/febs.14764.

[359] T. D. Vance, L. A. Graham, and P. L. Davies, "An ice-binding and tandem beta-sandwich domain-containing protein in Shewanella frigidimarina is a potential new type of ice adhesin," en, *The FEBS Journal*, vol. 285, no. 8, pp. 1511–1527, 2018, ISSN: 1742-4658. DOI: 10.1111/febs.14424.

[360] J. A. Raymond and D. Remias, "Ice-Binding Proteins in a Chrysophycean Snow Alga: Acquisition of an Essential Gene by Horizontal Gene Transfer," English, *Frontiers in Microbiology*, vol. 10, Nov. 2019, ISSN: 1664-302X. DOI: 10.3389/fmicb.2019.02697.

[361] L. Procházková, D. Remias, L. Nedbalová, and J. A. Raymond, "A DUF3494 ice-binding protein with a root cap domain in a streptophyte glacier ice alga," English, *Frontiers in Plant Science*, vol. 14, Jan. 2024, ISSN: 1664-462X. DOI: 10.3389/fpls.2023.1306511.

[362] J. A. Raymond, "Variations on a theme: Non-canonical DUF3494 ice-binding proteins," en, *Extremophiles*, vol. 29, no. 1, p. 8, Dec. 2024, ISSN: 1433-4909. DOI: 10.1007/s00792-024-01374-y.

[363] H. Kondo, Y. Hanada, H. Sugimoto, T. Hoshino, C. P. Garnham, P. L. Davies, and S. Tsuda, "Ice-binding site of snow mold fungus antifreeze protein deviates from structural regularity and high conservation," *Proceedings of the National Academy of Sciences*, vol. 109, no. 24, pp. 9360–9365, Jun. 2012. DOI: 10.1073/pnas.1121607109.

[364] L. Käll, A. Krogh, and E. L. Sonnhammer, "Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server," *Nucleic Acids Research*, vol. 35, no. suppl_2, W429–W432, Jul. 2007, ISSN: 0305-1048. DOI: 10.1093/nar/gkm256.

[365] P. J. McMurdie and S. Holmes, "Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data," en, *PLOS ONE*, vol. 8, no. 4, e61217, Apr. 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0061217.

[366] Fiksen, M. J. Follows, and D. L. Aksnes, "Trait-based models of nutrient uptake in microbes extend the Michaelis-Menten framework," en, *Limnology and Oceanography*, vol. 58, no. 1, pp. 193–202, Jan. 2013, ISSN: 0024-3590, 1939-5590. DOI: 10.4319/lo.2013.58.1.0193.

[367] I. Letunic and P. Bork, "Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool," *Nucleic Acids Research*, vol. 52, no. W1, W78–W82, Jul. 2024, ISSN: 0305-1048. DOI: 10.1093/nar/gkae268.

[368] T. G. Stephens, R. A. González-Pech, Y. Cheng, A. R. Mohamed, D. W. Burt, D. Bhattacharya, M. A. Ragan, and C. X. Chan, "Genomes of the dinoflagellate Polarella glacialis encode tandemly repeated single-exon genes with adaptive functions," *BMC Biology*, vol. 18, no. 1, p. 56, May 2020, ISSN: 1741-7007. DOI: 10.1186/s12915-020-00782-8.

[369] R. Overbeek, M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev, "The use of gene clusters to infer functional coupling," eng, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2896–2901, Mar. 1999, ISSN: 0027-8424. DOI: 10.1073/pnas.96.6.2896.

[370] C. Krembs and J. W. Deming, "The Role of Exopolymers in Microbial Adaptation to Sea Ice," en, in *Psychrophiles: from Biodiversity to Biotechnology*, R. Margesin,

F. Schinner, J.-C. Marx, and C. Gerday, Eds., Berlin, Heidelberg: Springer, 2008, pp. 247–264, ISBN: 978-3-540-74335-4. DOI: 10.1007/978-3-540-74335-4_15.

[371] U. Sorhannus, "Evolution of antifreeze protein genes in the diatom genus fragilariopsis: Evidence for horizontal gene transfer, gene duplication and episodic diversifying selection," eng, *Evolutionary Bioinformatics Online*, vol. 7, pp. 279–289, 2011, ISSN: 1176-9343. DOI: 10.4137/EBO.S8321.

[372] D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang, "MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies," *PeerJ*, vol. 7, e7359, Jul. 2019, ISSN: 2167-8359. DOI: 10.7717/peerj.7359.

[373] S. Pan, X.-M. Zhao, and L. P. Coelho, "SemiBin2: Self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing," *Bioinformatics*, vol. 39, no. Supplement_1, pp. i21–i29, Jun. 2023, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad209.

[374] S. Zeng, D. Patangia, A. Almeida, Z. Zhou, D. Mu, R. Paul Ross, C. Stanton, and S. Wang, "A compendium of 32,277 metagenome-assembled genomes and over 80 million genes from the early-life human gut microbiome," en, *Nature Communications*, vol. 13, no. 1, p. 5139, Sep. 2022, ISSN: 2041-1723. DOI: 10.1038/s41467-022-32805-z.

[375] R. E. Anderson, J. Reveillaud, E. Reddington, T. O. Delmont, A. M. Eren, J. M. McDermott, J. S. Seewald, and J. A. Huber, "Genomic variation in microbial populations inhabiting the marine subseafloor at deep-sea hydrothermal vents," en, *Nature Communications*, vol. 8, no. 1, p. 1114, Oct. 2017, ISSN: 2041-1723. DOI: 10.1038/s41467-017-01228-6.

[376] M. A. Biscotti, E. Olmo, and J. S. P. Heslop-Harrison, "Repetitive DNA in eukaryotic genomes," en, *Chromosome Research*, vol. 23, no. 3, pp. 415–420, Sep. 2015, ISSN: 1573-6849. DOI: 10.1007/s10577-015-9499-z.

[377] T. R. Gregory, "Synergy between sequence and size in Large-scale genomics," en, *Nature Reviews Genetics*, vol. 6, no. 9, pp. 699–708, Sep. 2005, ISSN: 1471-0064. DOI: 10.1038/nrg1674.

[378] Y. Nishimura and S. Yoshizawa, "The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments," en, *Scientific Data*, vol. 9, no. 1, p. 305, Jun. 2022, ISSN: 2052-4463. DOI: 10.1038/s41597-022-01392-5.

[379] R. Massana and D. López-Escardó, "Metagenome assembled genomes are for eukaryotes too," *Cell Genomics*, vol. 2, no. 5, p. 100 130, May 2022, ISSN: 2666-979X. DOI: 10.1016/j.xgen.2022.100130.

[380] M. R. Olm, P. T. West, B. Brooks, B. A. Firek, R. Baker, M. J. Morowitz, and J. F. Banfield, "Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms," *Microbiome*, vol. 7, no. 1, p. 26, Feb. 2019, ISSN: 2049-2618. DOI: 10.1186/s40168-019-0638-1.

[381] W. Xu, Y. Xu, R. Sun, E. Rey Redondo, K. K. Leung, S. H. Wan, J. Li, and C. C. M. Yung, "Revealing the intricate temporal dynamics and adaptive responses of prokaryotic and eukaryotic microbes in the coastal South China Sea," *Science of The Total Environment*, vol. 952, p. 176 019, Nov. 2024, ISSN: 0048-9697. DOI: 10.1016/j.scitotenv.2024.176019.

[382] G. Tagirdzhanova, P. Saary, E. S. Cameron, C. C. G. Allen, A. I. Garber, D. D. Escandón, A. T. Cook, S. Goyette, V. T. Nogerius, A. Passo, H. Mayrhofer, H. Holien, T. Tønsberg, L. Y. Stein, R. D. Finn, and T. Spribille, "Microbial occurrence and symbiont detection in a global sample of lichen metagenomes," en, *PLOS Biology*, vol. 22, no. 11, e3002862, Nov. 2024, ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3002862.

[383] J. L. Espinoza, A. Phillips, M. Prentice, G. Tan, P. Kamath, K. G. Lloyd, and C. Dupont, "Unveiling the microbial realm with VEBA 2.0: A modular bioinformatics suite for end-to-end genome-resolved prokaryotic, (micro)eukaryotic and viral multi-omics from either short- or long-read sequencing," *Nucleic Acids Research*, vol. 52, no. 14, e63, Aug. 2024, ISSN: 0305-1048. DOI: 10.1093/nar/gkae528.

[384] U. Rocha, J. Coelho Kasmanas, R. Kallies, J. P. Saraiva, R. B. Toscan, P. Štefanič, M. F. Bicalho, F. Borim Correa, M. N. Baştürk, E. Fousekis, L. M. Viana Barbosa, J. Plewka, A. J. Probst, P. Baldrian, P. F. Stadler, and T. C. Clue, "MuDoGeR: Multi-Domain Genome recovery from metagenomes made easy," *Molecular Ecology Resources*, vol. 24, no. 2, e13904, Feb. 2024, ISSN: 1755-098X. DOI: 10.1111/1755-0998.13904.

[385] G. Michoud, H. Peter, S. B. Busi, M. Bourquin, T. J. Kohler, A. Geers, L. Ezzat, and T. J. Battin, "Mapping the metagenomic diversity of the multi-kingdom glacier-fed stream microbiome," en, *Nature Microbiology*, vol. 10, no. 1, pp. 217–230, Jan. 2025, ISSN: 2058-5276. DOI: 10.1038/s41564-024-01874-9.

[386] H. J. Seong, J. J. Kim, and W. J. Sul, "ACR: Metagenome-assembled prokaryotic and eukaryotic genome refinement tool," *Briefings in Bioinformatics*, vol. 24, no. 6, bbad381, Nov. 2023, ISSN: 1477-4054. DOI: 10.1093/bib/bbad381.

[387] X. Peng, S. E. Wilken, T. S. Lankiewicz, S. P. Gilmore, J. L. Brown, J. K. Henske, C. L. Swift, A. Salamov, K. Barry, I. V. Grigoriev, M. K. Theodorou, D. L. Valentine, and M. A. O'Malley, "Genomic and functional analyses of fungal and bacterial consortia that enable lignocellulose breakdown in goat gut microbiomes," en,

Nature Microbiology, vol. 6, no. 4, pp. 499–511, Apr. 2021, ISSN: 2058-5276. DOI: 10.1038/s41564-020-00861-0.

[388] J. P. Saraiva, A. Bartholomäus, R. B. Toscan, P. Baldrian, and U. Nunes da Rocha, "Recovery of 197 eukaryotic bins reveals major challenges for eukaryote genome reconstruction from terrestrial metagenomes," en, *Molecular Ecology Resources*, vol. 23, no. 5, pp. 1066–1076, 2023, ISSN: 1755-0998. DOI: 10.1111/1755-0998.13776.

[389] A. M. Eren, C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont, "Anvi'o: An advanced analysis and visualization platform for 'omics data," *PeerJ*, vol. 3, e1319, Oct. 2015, ISSN: 2167-8359. DOI: 10.7717/peerj.1319.

[390] G. P. Schmartz, P. Hirsch, J. Amand, J. Dastbaz, T. Fehlmann, F. Kern, R. Müller, and A. Keller, "BusyBee Web: Towards comprehensive and differential composition-based metagenomic binning," *Nucleic Acids Research*, vol. 50, no. W1, W132–W137, Jul. 2022, ISSN: 0305-1048. DOI: 10.1093/nar/gkac298.

[391] M. J. Pavia, A. Chede, Z. Wu, H. Cadillo-Quiroz, and Q. Zhu, "BinaRena: A dedicated interactive platform for human-guided exploration and binning of metagenomes," *Microbiome*, vol. 11, no. 1, p. 186, Aug. 2023, ISSN: 2049-2618. DOI: 10.1186/s40168-023-01625-8.

[392] D. Zhao, D. E. Salas-Leiva, S. K. Williams, K. A. Dunn, J. D. Shao, and A. J. Roger, "Eukfinder: A pipeline to retrieve microbial eukaryote genome sequences from metagenomic data," *mBio*, vol. 0, no. 0, e00699–25, Apr. 2025. DOI: 10.1128/mbio.00699-25.

[393] P. T. West, A. J. Probst, I. V. Grigoriev, B. C. Thomas, and J. F. Banfield, "Genome-reconstruction for eukaryotes from complex natural microbial communities," en, *Genome Research*, vol. 28, no. 4, pp. 569–580, Apr. 2018, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.228429.117.

[394] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince, "Binning metagenomic contigs by coverage and composition," en, *Nature Methods*, vol. 11, no. 11, pp. 1144–1146, Nov. 2014, ISSN: 1548-7105. DOI: 10.1038/nmeth.3103.

[395] B. E. Granger and F. Pérez, "Jupyter: Thinking and Storytelling With Code and Data," *Computing in Science & Engineering*, vol. 23, no. 2, pp. 7–14, Mar. 2021, ISSN: 1558-366X. DOI: 10.1109/MCSE.2021.3059263.

[396] K. Sahlin, "Strobealign: Flexible seed size enables ultra-fast and accurate read alignment," *Genome Biology*, vol. 23, no. 1, p. 260, Dec. 2022, ISSN: 1474-760X. DOI: 10.1186/s13059-022-02831-7.

[397] D. E. Wood and S. L. Salzberg, "Kraken: Ultrafast metagenomic sequence classification using exact alignments," en, *Genome Biology*, vol. 15, no. 3, R46, 2014, ISSN: 1465-6906. DOI: 10.1186/gb-2014-15-3-r46.

[398] D. Kim, L. Song, F. P. Breitwieser, and S. L. Salzberg, "Centrifuge: Rapid and sensitive classification of metagenomic sequences," *Genome Research*, vol. 26, no. 12, pp. 1721–1729, Dec. 2016, ISSN: 1088-9051. DOI: 10.1101/gr.210641.116.

[399] M. Steinegger and J. Söding, "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," en, *Nature Biotechnology*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3988.

[400] S. T. N. Aroney, R. J. P. Newell, G. W. Tyson, and B. J. Woodcroft, "Bin chicken: Targeted metagenomic coassembly for the efficient recovery of novel genomes," *Nature Methods*, vol. 22, no. 12, pp. 2516–2524, 2025, ISSN: 1548-7105. DOI: 10.1038/s41592-025-02901-1. [Online]. Available: https://doi.org/10.1038/s41592-025-02901-1.

[401] C. C. Laczny, T. Sternal, V. Plugaru, P. Gawron, A. Atashpendar, H. H. Margossian, S. Coronado, L. v. der Maaten, N. Vlassis, and P. Wilmes, "VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data," en, *Microbiome*, vol. 3, no. 1, p. 1, Dec. 2015, ISSN: 2049-2618. DOI: 10.1186/s40168-014-0066-1.

[402] B. Imanian, J.-F. Pombert, and P. J. Keeling, "The Complete Plastid Genomes of the Two 'Dinotoms' Durinskia baltica and Kryptoperidinium foliaceum," en, *PLOS ONE*, vol. 5, no. 5, e10711, May 2010, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0010711.

[403] J. C. Villada, Y. M. Vasquez, G. Szabo, E. Whittaker-Walker, M. F. Romero, S. Qin, N. Varghese, E. A. Eloe-Fadrosh, N. C. Kyrpides, S. d. Consortium, A. Visel, T. Woyke, and F. Schulz, *A genomic catalog of Earth's bacterial and archaeal symbionts*, en, May 2025. DOI: 10.1101/2025.05.29.656868.

[404] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," en, *BMC Bioinformatics*, vol. 9, no. 1, p. 559, Dec. 2008, ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559.

[405] S. Tippenhauer, M. Vredenborg, C. Heuzé, A. Ulfsbo, B. Rabe, M. A. Granskog, J. Allerholt, J. P. Balmonte, R. G. Campbell, G. Castellani, E. Chamberlain, J. Creamean, A. DAngelo, U. Dietrich, E. S. Droste, L. Eggers, Y.-C. Fang, A. A. Fong, J. Gardner, R. Graupner, J. Grosse, H. He, N. Hildebrandt, C. J. M. Hoppe, M. Hoppmann, T. Kanzow, S. Karam, Z. Koenig, B. Kong, D. Kuhlmey, I. Kuznetsov, M. Lan, H. Liu, M. Mallet, V. Mohrholz, M. Muilwijk, O. Müller, L. M. Olsen, R. Rember, J. Ren, S. Sakinan, J. Schaffer, K. Schmidt, I. Schuffenhauer, K. Schulz, K. Shoemaker,

S. Spahic, N. Sukhikh, A. Svenson, S. Torres-Valdés, A. Torstensson, L. Wischnewski, and Y. Zhuang, *Physical oceanography water bottle samples based on Ocean City CTD during POLARSTERN cruise PS122*, 2023. DOI: 10.1594/PANGAEA.959966.

[406] S. Torres-Valdés, R. Rember, L. Heitmann, K.-U. Ludwichowski, A. Ulfsbo, A. A. Fong, C. J. M. Hoppe, I. Kuznetsov, E. Damm, M. Graeve, U. Dietrich, E. Chamberlain, E. S. Droste, J. Creamean, J. Gardner, O. Müller, J. P. Balmonte, and B. Rost, *Dissolved nutrients data from the PS122 MOSAiC expedition carried out at the AWI nutrient facility*, 2024. DOI: 10.1594/PANGAEA.966217.

[407] C. J. M. Hoppe, J. Creamean, J. S. Bowman, L. Heitmann, E. Chamberlain, T. Brenneis, A. Terbrüggen, and M. Vortkamp, *Year-round discrete underway water column Chlorophyll a concentrations from the central Arctic*, 2023. DOI: 10.1594/PANGAEA.962597.

[408] M. Oggier, E. Salganik, L. M. Whitmore, A. A. Fong, C. J. M. Hoppe, R. Rember, K. V. Høyland, R. Gradinger, D. V. Divine, S. W. Fons, K. Abrahamsson, A. M. Aguilar-Islas, M. Angelopoulos, S. Arndt, J. P. Balmonte, D. Bozzato, J. S. Bowman, G. Castellani, E. Chamberlain, J. Creamean, A. DAngelo, E. Damm, U. Dietrich, E. S. Droste, A. Dumitrascu, S. L. Eggers, J. Gardner, L. Grosfeld, J. Haapala, L. Heitmann, A. Immerz, N. Kolabutin, B. A. Lange, R. Lei, C. M. Marsay, S. Maus, L. M. Olsen, O. Müller, A. Nuibom, J. Ren, A. Rinke, K. Schmidt, I. Sheikin, E. Shimanchuk, P. Snoeijs-Leijonmalm, S. Spahic, J. Stefels, S. Torres-Valdés, A. Torstensson, A. Ulfsbo, J. Verdugo, M. Vortkamp, L. Wang, M. Webster, and M. A. Granskog, *Second-year sea-ice salinity, temperature, density, nutrient, oxygen and hydrogen isotope composition from the main coring site (MCS-SYI) during MOSAiC legs 1 to 4 in 2019/2020, version 2*, 2025. DOI: 10.1594/PANGAEA.974764.

[409] E. Salganik, L. M. Whitmore, D. Bauch, E. Chamberlain, U. Dietrich, E. S. Droste, A. A. Fong, L. Heitmann, M. Nicolaus, N. Kolabutin, Y. Li, K.-U. Ludwichowski, A. Marent, M. Mellat, H. Meyer, D. Nomura, K. Schmidt, E. Shimanchuk, L. Thielke, S. Torres-Valdés, A. L. Webb, M. Weiner, and M. A. Granskog, *Sea-ice salinity, temperature, density, nutrient, oxygen and hydrogen isotope composition from the coring sites during MOSAiC leg 5 in August-September 2020*, 2024. DOI: 10.1594/PANGAEA.971266.

[410] JGI, *Img/m annotation version change log*, https://sites.google.com/lbl.gov/imghelp/img-annotation-pipeline-v-5-x/change-log, 2025.

[411] A. Mitchell, H.-Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Thomas, and R. D. Finn, "The InterPro protein families database: The classification resource

after 15 years," en, *Nucleic Acids Research*, vol. 43, no. D1, pp. D213–D221, Jan. 2015, ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gku1243.

[412] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz, "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life," en, *Nature Biotechnology*, vol. 36, no. 10, pp. 996–1004, Nov. 2018, ISSN: 1546-1696. DOI: 10.1038/nbt.4229.

[413] C. Rinke, F. Rubino, L. F. Messer, N. Youssef, D. H. Parks, M. Chuvochina, M. Brown, T. Jeffries, G. W. Tyson, J. R. Seymour, and P. Hugenholtz, "A phylogenomic and ecological analysis of the globally abundant marine group ii archaea (ca. poseidoniales ord. nov.)," *The ISME Journal*, vol. 13, no. 3, pp. 663–675, 2019, ISSN: 1751-7362. DOI: 10.1038/s41396-018-0282-y. eprint: https://academic.oup.com/ismej/article-pdf/13/3/663/55551449/41396_2018_article_282.pdf. [Online]. Available: https://doi.org/10.1038/s41396-018-0282-y.

[414] H. Hodal, S. Falk-Petersen, H. Hop, S. Kristiansen, and M. Reigstad, "Spring bloom dynamics in Kongsfjorden, Svalbard: Nutrients, phytoplankton, protozoans and primary production," en, *Polar Biology*, vol. 35, no. 2, pp. 191–203, Feb. 2012, ISSN: 1432-2056. DOI: 10.1007/s00300-011-1053-7.

[415] E. Y. Koh, R. O. M. Cowie, A. M. Simpson, R. O'Toole, and K. G. Ryan, "The origin of cyanobacteria in Antarctic sea ice: Marine or freshwater?" en, *Environmental Microbiology Reports*, vol. 4, no. 5, pp. 479–483, 2012, ISSN: 1758-2229. DOI: 10.1111/j.1758-2229.2012.00346.x.

[416] T. Mock and D. N. Thomas, "Recent advances in sea-ice microbiology," en, *Environmental Microbiology*, vol. 7, no. 5, pp. 605–619, 2005, ISSN: 1462-2920. DOI: 10.1111/j.1462-2920.2005.00781.x.

[417] Bergh, K. Y. BØrsheim, G. Bratbak, and M. Heldal, "High abundance of viruses found in aquatic environments," en, *Nature*, vol. 340, no. 6233, pp. 467–468, Aug. 1989, ISSN: 1476-4687. DOI: 10.1038/340467a0.

[418] J. M. Nielsen, M. F. Sigler, L. B. Eisner, J. T. Watson, L. A. Rogers, S. W. Bell, N. Pelland, C. W. Mordy, W. Cheng, K. Kivva, S. Osborne, and P. Stabeno, "Spring phytoplankton bloom phenology during recent climate warming on the Bering Sea shelf," *Progress in Oceanography*, vol. 220, p. 103 176, Jan. 2024, ISSN: 0079-6611. DOI: 10.1016/j.pocean.2023.103176.

[419] N. Fuchs, P. Anhaus, M. Hoppmann, T. Kagel, C. Katlein, R. Reese, L. Riemenschneider, R. Tao, R. Winkelmann, and D. Notz, "In-ice light measurements during the MOSAiC expedition," en, *Scientific Data*, vol. 11, no. 1, p. 702, Jun. 2024, ISSN: 2052-4463. DOI: 10.1038/s41597-024-03472-0.

[420] I. A. Raphael, D. K. Perovich, C. M. Polashenski, D. Clemens-Sewall, P. Itkin, R. Lei, M. Nicolaus, J. Regnery, M. M. Smith, M. Webster, and M. Jaggi, "Sea ice mass balance during the MOSAiC drift experiment: Results from manual ice and snow thickness gauges," *Elementa: Science of the Anthropocene*, vol. 12, no. 1, p. 00 040, Jul. 2024, ISSN: 2325-1026. DOI: 10.1525/elementa.2023.00040.

[421] K. T. Konstantinidis and J. M. Tiedje, "Trends between gene content and genome size in prokaryotic species with larger genomes," *Proceedings of the National Academy of Sciences*, vol. 101, no. 9, pp. 3160–3165, Mar. 2004. DOI: 10.1073/pnas.0308653100.

[422] Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer, "MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm," *Microbiome*, vol. 2, no. 1, p. 26, Aug. 2014, ISSN: 2049-2618. DOI: 10.1186/2049-2618-2-26.

[423] C. M. K. Sieber, A. J. Probst, A. Sharrar, B. C. Thomas, M. Hess, S. G. Tringe, and J. F. Banfield, "Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy," en, *Nature Microbiology*, vol. 3, no. 7, pp. 836–843, Jul. 2018, ISSN: 2058-5276. DOI: 10.1038/s41564-018-0171-1.

[424] E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. Simão, and E. M. Zdobnov, "OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs," *Nucleic Acids Research*, vol. 47, no. D1, pp. D807–D811, Jan. 2019, ISSN: 0305-1048. DOI: 10.1093/nar/gky1053.

[425] T. O. Delmont, *Assessing the completion of eukaryotic bins with anvi'o*, https://merenlab.org/2018/05/05/eukaryotic-single-copy-core-genes, 2018.

[426] L. Pan, P. Parini, R. Tremmel, J. Loscalzo, V. M. Lauschke, B. A. Maron, P. Paci, I. Ernberg, N. S. Tan, Z. Liao, W. Yin, S. Rengarajan, X. Li, and The SCA Consortium, "Single Cell Atlas: A single-cell multi-omics human cell encyclopedia," *Genome Biology*, vol. 25, no. 1, p. 104, Apr. 2024, ISSN: 1474-760X. DOI: 10.1186/s13059-024-03246-2.

[427] T. Sainburg, L. McInnes, and T. Q. Gentner, *Parametric UMAP embeddings for representation and semi-supervised learning*, Aug. 2021. DOI: 10.48550/arXiv.2009.12981.

[428] K. Blin, S. Shaw, H. E. Augustijn, Z. L. Reitz, F. Biermann, M. Alanjary, A. Fetter, B. R. Terlouw, W. W. Metcalf, E. J. N. Helfrich, G. P. van Wezel, M. H. Medema, and T. Weber, "antiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation," *Nucleic Acids Research*, vol. 51, no. W1, W46–W50, Jul. 2023, ISSN: 0305-1048. DOI: 10.1093/nar/gkad344.

[429] S. A. Kautsar, J. J. J. van der Hooft, D. de Ridder, and M. H. Medema, "BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters," *GigaScience*, vol. 10, no. 1, giaa154, Jan. 2021, ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa154.

[430] Y. Ma, Z. Guo, B. Xia, Y. Zhang, X. Liu, Y. Yu, N. Tang, X. Tong, M. Wang, X. Ye, J. Feng, Y. Chen, and J. Wang, "Identification of antimicrobial peptides from the human gut microbiome using deep learning," en, *Nature Biotechnology*, vol. 40, no. 6, pp. 921–931, Jun. 2022, ISSN: 1546-1696. DOI: 10.1038/s41587-022-01226-0.

[431] S. Sonjak, J. C. Frisvad, and N. Gunde-Cimerman, "Comparison of secondary metabolite production by Penicillium crustosum strains, isolated from Arctic and other various ecological niches," *FEMS Microbiology Ecology*, vol. 53, no. 1, pp. 51–60, Jun. 2005, ISSN: 0168-6496. DOI: 10.1016/j.femsec.2004.10.014.

[432] A. Rego, A. Fernandez-Guerra, P. Duarte, P. Assmy, P. N. Leão, and C. Magalhães, "Secondary metabolite biosynthetic diversity in Arctic Ocean metagenomes," *Microbial Genomics*, vol. 7, no. 12, p. 000 731, 2021, ISSN: 2057-5858. DOI: 10.1099/mgen.0.000731.

[433] L. Paoli, H.-J. Ruscheweyh, C. C. Forneris, F. Hubrich, S. Kautsar, A. Bhushan, A. Lotti, Q. Clayssen, G. Salazar, A. Milanese, C. I. Carlström, C. Papadopoulou, D. Gehrig, M. Karasikov, H. Mustafa, M. Larralde, L. M. Carroll, P. Sánchez, A. A. Zayed, D. R. Cronin, S. G. Acinas, P. Bork, C. Bowler, T. O. Delmont, J. M. Gasol, A. D. Gossert, A. Kahles, M. B. Sullivan, P. Wincker, G. Zeller, S. L. Robinson, J. Piel, and S. Sunagawa, "Biosynthetic potential of the global ocean microbiome," en, *Nature*, vol. 607, no. 7917, pp. 111–118, Jul. 2022, ISSN: 1476-4687. DOI: 10.1038/s41586-022-04862-3.

[434] C. Llor and L. Bjerrum, "Antimicrobial resistance: Risk associated with antibiotic overuse and initiatives to reduce the problem," *Therapeutic Advances in Drug Safety*, vol. 5, no. 6, pp. 229–241, Dec. 2014, ISSN: 2042-0986. DOI: 10.1177/2042098614554919.

[435] C. Manyi-Loh, S. Mamphweli, E. Meyer, and A. Okoh, "Antibiotic Use in Agriculture and Its Consequential Resistance in Environmental Sources: Potential Public Health Implications," *Molecules : A Journal of Synthetic Chemistry and Natural Product Chemistry*, vol. 23, no. 4, p. 795, Mar. 2018, ISSN: 1420-3049. DOI: 10.3390/molecules23040795.

[436] S. W. Wilhelm and C. A. Suttle, "Viruses and Nutrient Cycles in the Sea: Viruses play critical roles in the structure and function of aquatic food webs," *BioScience*, vol. 49, no. 10, pp. 781–788, Oct. 1999, ISSN: 0006-3568. DOI: 10.2307/1313569.

[437] D. D. Dunigan, L. A. Fitzgerald, and J. L. Van Etten, "Phycodnaviruses: A peek at genetic diversity," *Virus Research*, Comparative Genomics and Evolution of Complex

Viruses, vol. 117, no. 1, pp. 119–132, Apr. 2006, ISSN: 0168-1702. DOI: 10.1016/j.virusres.2006.01.024.

[438] P. Anderson, J. Berge, M. A. Granskog, D. V. Divine, C. Katlein, P. Itkin, I. Raphael, G. Johnsen, D. Vogedes, T. Kopec, A. Zolich, M. Geoffroy, F. Cottier, and P. R. De La Torre, *Upwelling and downwelling visible radiation measurements of the autonomous ice-tethered OptiCAL gg buoy deployed during MOSAiC in the Monster Bay area*, 2023. DOI: 10.1594/PANGAEA.954849.

[439] P. Anderson, J. Berge, M. A. Granskog, D. V. Divine, C. Katlein, P. Itkin, I. Raphael, G. Johnsen, D. Vogedes, T. Kopec, A. Zolich, M. Geoffroy, F. Cottier, and P. R. De La Torre, *Upwelling and downwelling visible radiation measurements of the autonomous ice-tethered OptiCAL hh buoy deployed during MOSAiC in the Dark site Second Year Ice*, 2023. DOI: 10.1594/PANGAEA.955045.

[440] P. Anderson, J. Berge, M. A. Granskog, D. V. Divine, C. Katlein, P. Itkin, I. Raphael, G. Johnsen, D. Vogedes, T. Kopec, A. Zolich, M. Geoffroy, F. Cottier, and P. R. De La Torre, *Upwelling and downwelling visible radiation measurements of the autonomous ice-tethered OptiCAL ee buoy deployed during MOSAiC in the Sea Ice Ridge Observatory area*, 2023. DOI: 10.1594/PANGAEA.928495.

[441] N. Fuchs, T. Kagel, C. Katlein, R. Lei, M. Nicolaus, D. Notz, and L. Riemenschneider, *Measurements of in-ice light profiles and optical properties of Arctic sea ice on MOSAiC 2020: Unified NetCDF data from the lightharp and lightchain instruments*, 2024. DOI: 10.1594/PANGAEA.963743.

[442] A. R. Coenen, S. K. Hu, E. Luo, D. Muratore, and J. S. Weitz, "A Primer for Microbiome Time-Series Analysis," English, *Frontiers in Genetics*, vol. 11, Apr. 2020, ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00310.

[443] K. Faust, L. Lahti, D. Gonze, W. M. de Vos, and J. Raes, "Metagenomics meets time series analysis: Unraveling microbial community dynamics," *Current Opinion in Microbiology*, Environmental microbiology • Extremophiles, vol. 25, pp. 56–66, Jun. 2015, ISSN: 1369-5274. DOI: 10.1016/j.mib.2015.04.004.

[444] O. Ovaskainen, G. Tikhonov, D. Dunson, V. Grøtan, S. Engen, B.-E. Sæther, and N. Abrego, "How are species interactions structured in species-rich communities? A new method for analysing time-series data," *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, no. 1855, p. 20 170 768, May 2017. DOI: 10.1098/rspb.2017.0768.

[445] G. Weithoff and B. E. Beisner, "Measures and Approaches in Trait-Based Phytoplankton Community Ecology – From Freshwater to Marine Ecosystems," English, *Frontiers in Marine Science*, vol. 6, Feb. 2019, ISSN: 2296-7745. DOI: 10.3389/fmars.2019.00040.

[446] A. L. N. Guislain, J. C. Nejstgaard, J. Köhler, E. Sperfeld, U. Mischke, B. Skjelbred, H.-P. Grossart, A. Lyche Solheim, M. O. Gessner, and S. A. Berger, "Cell size explains shift in phytoplankton community structure following storm-induced changes in light and nutrients," eng, *Ecology*, vol. 106, no. 3, e70043, Mar. 2025, ISSN: 1939-9170. DOI: 10.1002/ecy.70043.

[447] B. Matthiessen, G. S. I. Hattich, S. Pulina, T. Hansen, T. B. H. Reusch, and J. Hamer, "Phytoplankton mean cell size and total biomass increase with nutrients are driven by both species composition and evolution of plasticity," en, *Oikos*, vol. 2025, no. 1, e10910, 2025, ISSN: 1600-0706. DOI: 10.1111/oik.10910.

[448] N. M. Levine, M. A. Doblin, and S. Collins, "Reframing trait trade-offs in marine microbes," en, *Communications Earth & Environment*, vol. 5, no. 1, p. 219, Apr. 2024, ISSN: 2662-4435. DOI: 10.1038/s43247-024-01381-z.

[449] D. Čertnerová, P. Škaloud, I. Jadrná, and M. Čertner, "Large Genomes Are Associated With Greater Cell Size and Ecological Shift Towards More Nitrogen-Rich and Higher-Latitude Environments in Microalgae of the Genus Synura," *The Journal of Eukaryotic Microbiology*, vol. 72, no. 4, e70026, 2025, ISSN: 1066-5234. DOI: 10.1111/jeu.70026.

[450] J. M. Beaulieu, I. J. Leitch, S. Patel, A. Pendharkar, and C. A. Knight, "Genome size is a strong predictor of cell size and stomatal density in angiosperms," en, *New Phytologist*, vol. 179, no. 4, pp. 975–986, 2008, ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.2008.02528.x.

[451] O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego, "How to make more out of community data? A conceptual framework and its implementation as models and software," en, *Ecology Letters*, vol. 20, no. 5, pp. 561–576, 2017, ISSN: 1461-0248. DOI: 10.1111/ele.12757.

[452] A. U. Rahman, G. Tikhonov, J. Oksanen, T. Rossi, and O. Ovaskainen, "Accelerating joint species distribution modelling with Hmsc-HPC by GPU porting," en, *PLOS Computational Biology*, vol. 20, no. 9, e1011914, Sep. 2024, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011914.

[453] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, Jan. 2017. DOI: 10.48550/arXiv.1412.6980.

Chapter A

Appendix A

A.1 Accession Numbers of *Polarella glacialis* IBPs

CAE8679677.1	CAE8631646.1	CAE8583686.1	CAE8583687.1
CAE8636836.1	CAE8582104.1	CAE8583689.1	CAE8636834.1
CAE8582102.1	CAE8582101.1	CAE8631645.1	CAE8582103.1
CAE8679681.1	CAE8729160.1	CAE8729152.1	CAE8618267.1
CAE8741406.1	CAE8583685.1	CAE8637704.1	CAE8618265.1
CAE8634595.1	CAE8639355.1	CAE8637703.1	CAE8729150.1
CAE8581014.1	CAE8741407.1	CAE8618268.1	CAE8639358.1
CAE8688622.1	CAE8639354.1	CAE8639356.1	CAE8636651.1
CAE8628465.1	CAE8588114.1	CAE8595931.1	CAE8636649.1
CAE8671768.1	CAE8671762.1	CAE8671764.1	CAE8636642.1
CAE8636644.1	CAE8640504.1	CAE8696859.1	CAE8625316.1
CAE8636647.1	CAE8604292.1	CAE8607075.1	CAE8668303.1
CAE8643832.1	CAE8652165.1	CAE8601668.1	CAE8671765.1
CAE8720042.1	CAE8717284.1	CAE8611419.1	CAE8622661.1
CAE8677484.1	CAE8640451.1	CAE8647134.1	CAE8635871.1
CAE8622663.1	CAE8581017.1	CAE8635872.1	CAE8594356.1
CAE8594357.1	CAE8647038.1	CAE8641817.1	CAE8647135.1
CAE8585364.1	CAE8654565.1	CAE8650135.1	CAE8602834.1
CAE8743901.1	CAE8706293.1	CAE8729157.1	CAE8591189.1
CAE8649369.1	CAE8675450.1	CAE8634922.1	CAE8700145.1
CAE8623192.1	CAE8647136.1	CAE8616734.1	CAE8631824.1
CAE8627979.1	CAE8720727.1	CAE8688779.1	CAE8616733.1
CAE8706292.1	CAE8616732.1	CAE8687224.1	CAE8624346.1
CAE8625315.1	CAE8743391.1	CAE8723718.1	CAE8624936.1
CAE8720725.1	CAE8743390.1	CAE8702088.1	CAE8588202.1
CAE8588206.1	CAE8588203.1	CAE8703796.1	CAE8671446.1

Table A.1: List of *Polarella glacialis* IBP accessions from the NCBI SRA, BioProject PRJEB33539.

Sample Label	MOSAiC Sample Identifier	GOLD ID	JGI ID	IMG/M ID	NCBI SRA Accession	Citation
sea-ice interface 1	PS122_totDNA_309	Gp0561256	1290821	3300045789	SRX24021761	[317]
epipelagic 1	PS122_totDNA_328	Gp0561257	1290823	3300046532	SRX24021756	[312]
epipelagic 2	PS122_totDNA_327	Gp0561266	1292144	3300046450	SRX24021754	[310]
epipelagic 3	PS122_totDNA_327_328_pool	Gp0561269	1292150	3300047669	SRX24021764	[319]
sea-ice interface 2	PS122_totDNA_371	Gp0561258	1290827	3300045790	SRX24021758	[314]
interior ice 1	PS122_totDNA_414	Gp0561263	1290837	3300049783	SRX24021733	[307]
interior ice 2	PS122_totDNA_411	Gp0561260	1290831	3300046534	SRX24021731	[305]
interior ice 3	PS122_totDNA_412	Gp0561261	1290833	3300047666	SRX24021752	[308]
interior ice 4	PS122_totDNA_413	Gp0561262	1290835	3300047667	SRX24021732	[306]
sea-ice interface 3	PS122_totDNA_405	Gp0561259	1290829	3300046449	SRX24021755	[311]
epipelagic 4	PS122_totDNA_424_425_pool	Gp0561270	1292152	3300046103	SRX24021767	[322]
epipelagic 5	PS122_totDNA_424	Gp0561267	1292146	3300046451	SRX24021759	[315]
epipelagic 6	PS122_totDNA_425	Gp0561268	1292148	3300047668	SRX24021765	[320]
meso/bathypelagic 1	PS122_totDNA_418	Gp0561264	1290839	3300046467	SRX24021763	[318]
meso/bathypelagic 2	PS122_totDNA_432	Gp0561265	1290841	3300045738	SRX24021766	[321]

Table A.2: Accession numbers and identifiers of the pilot metagenomic samples; Label used in Chapter 5, ID used by MOSAiC, ID in the GOLD database, JGI analysis project ID, IMG/M taxon ID, NCBI SRA ID, and citation to an identifiers.org reference.

Chapter B

Appendix B

The next two subsections provide some mathematical background for UMAP and t-SNE, two commonly used non-linear dimensionality reduction methods used in machine learning.

B.1 UMAP and t-SNE Mathematical Details

t-SNE

The first method, t-SNE, uses a relatively straightforward algorithm to embed points from a high-dimensional dataset (using a Euclidean metric) to 2 or 3D. Given N data points $\boldsymbol{x}_i$, we want to find points in a low-dimensional embedding, $\boldsymbol{y}_i$, $(i = 1...N)$.

First a set of probability distributions are computed using the formula:

$$p_{j|i} = \frac{\exp\left(||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(||\boldsymbol{x}_i - \boldsymbol{x}_k||^2/2\sigma_i^2\right)}$$

These represent the conditional probabilities of $\boldsymbol{x}_i$ picking $\boldsymbol{x}_j$ as its nearest neighbour, if a neighbour was randomly picked according to a Gaussian distribution placed at the point $\boldsymbol{x}_i$. The variance σ_i^2 is chosen using a user-defined parameter called the perplexity, P, so that for each data point $\boldsymbol{x}_i$,

$$P = 2^{H_i}, \text{ where } H_i = -\sum_k p_{k|i}\log_2 p_{k|i}$$

is the Shannon entropy. For each $\boldsymbol{x}_i$, a value of σ_i^2 can be found by binary search to ensure that this holds. The $p_{j|i}$ are then symmetrised to form a dissimilarity matrix with components $p_{ij} = \frac{1}{2N}(p_{i|j} + p_{j|i})$.

Points in the low dimensional space $\boldsymbol{y}_i$ are then initialised randomly. We then compute a similar set of probability distributions

$$q_{ij} = \frac{(||1 + \boldsymbol{y}_i - \boldsymbol{y}_j||)^{-1}}{\sum_{k\neq l}(||1 + \boldsymbol{y}_k - \boldsymbol{y}_l||)^{-1}}$$

which has a similar form to that of $p_{i|j}$, except the Gaussian distribution has been replaced by a Student's t-distribution with 1 degree of freedom, and the normalisation is done over all pairs of points rather than conditional for each point.

Finally we iteratively update the points $\boldsymbol{y}_i$ so as to minimise a cost function called the cross-entropy,

$$\sum_{i \neq j} p_{ij} \log q_{ij}$$

This iteration can be done by an algorithm called gradient descent, where the points are updated in a direction proportionally to how much they decrease the cost function. Optimisation algorithms such as Adaptive Moment Estimation (ADAM) [453] will carry out such an iteration scheme, either until the points no longer move (to within some tolerance), or until some set number of iterations has been reached.

UMAP

Uniform Manifold Approximation (UMAP) is a theoretically more involved algorithm than t-SNE, though in practice there are a similar number of corresponding steps in the algorithm. Whereas the mapping from t-SNE is based on concepts from information theory, UMAP builds an embedding using concepts from graph theory, and in particular, the first step of the UMAP algorithm is to build a weighted nearest-neighbours graph. This connects points to their n nearest neighbours, with n as an input parameter. Similar to t-SNE, the weightings for the n nearest neighbours of point $\boldsymbol{x}_i$ are the given by a scaled exponential

$$w(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{d(\boldsymbol{x}_i, \boldsymbol{x}_j) - \rho_i}{\sigma_i}\right)$$

where ρ_i is the distance from $\boldsymbol{x}_i$ to its closest neighbour and σ_i is a deviation, chosen such that the total weight of $\boldsymbol{x}_i$'s neighbours is $\log_2(n)$.

Just as in t-SNE, the (weighted) adjacency matrix M is symmetrised, this time using the formula $M + M^T - M \odot M^T$, where $\odot$ is the pointwise product. However, for the low-dimensional embedding, points are not initialised randomly but instead embedded using the non-zero eigenvectors of a particular graph Laplacian matrix L, given by $L = D^{1/2}(D - M)D^{1/2}$; D being the diagonal matrix recording the degrees of each vertex. Other differences to t-SNE, besides the initialisation, are the choice of optimisation algorithm, the cost function to be minimised, and the corresponding dissimilarity measure associated with pairs of points in the embedded space. UMAP uses an optimisation algorithm called simulated annealing to move the embedded points along a gradient, where the cost function being minimised is

again a cross-entropy, this time formulated as a sum over the edges of the weighted graph

$$C = \sum_{i,j} (w(\boldsymbol{x}_i, \boldsymbol{x}_j) \log(\Phi(\boldsymbol{y}_i, \boldsymbol{y}_j)) + (1 - w(\boldsymbol{x}_i, \boldsymbol{x}_j)) \log(1 - \Phi(\boldsymbol{y}_i, \boldsymbol{y}_j))$$

.

The corresponding distance in the embedded space is the function Φ;

$$\Phi(\boldsymbol{y}_i, \boldsymbol{y}_j) = (1 + a||\boldsymbol{y}_i - \boldsymbol{y}_j||_2^{2b})^{-1}$$

for hyperparameters a and b. This choice of distance in the embedded space is essentially a heuristic choice, made similarly to the choice of a t-distribution in t-SNE, so as to avoid points crowding.

B.2 IMG Taxon IDs of MOSAiC Samples

The IMG Taxon IDs of the samples from the datasets MOSAiC_Water_June and MOSAiC_Ice_April are the following:

- **MOSAiC_Water_June**: 3300061190, 3300061567, 3300061558, 3300060030, 3300060847, 3300060893, 3300061740, 3300071295, 3300061542, 3300060858, 3300060716, 3300069123, 3300057219, 3300057218, 3300056912
- **MOSAiC_Ice_April**: 3300061187, 3300074301, 3300074315, 3300074331, 3300074487, 3300077200, 3300078168, 3300074336, 3300075611, 3300075631, 3300077802, 3300078170

These can be found by querying the IMG/M website https://img.jgi.doe.gov/.

B.3 BinaRena UMAP and PCA Plots

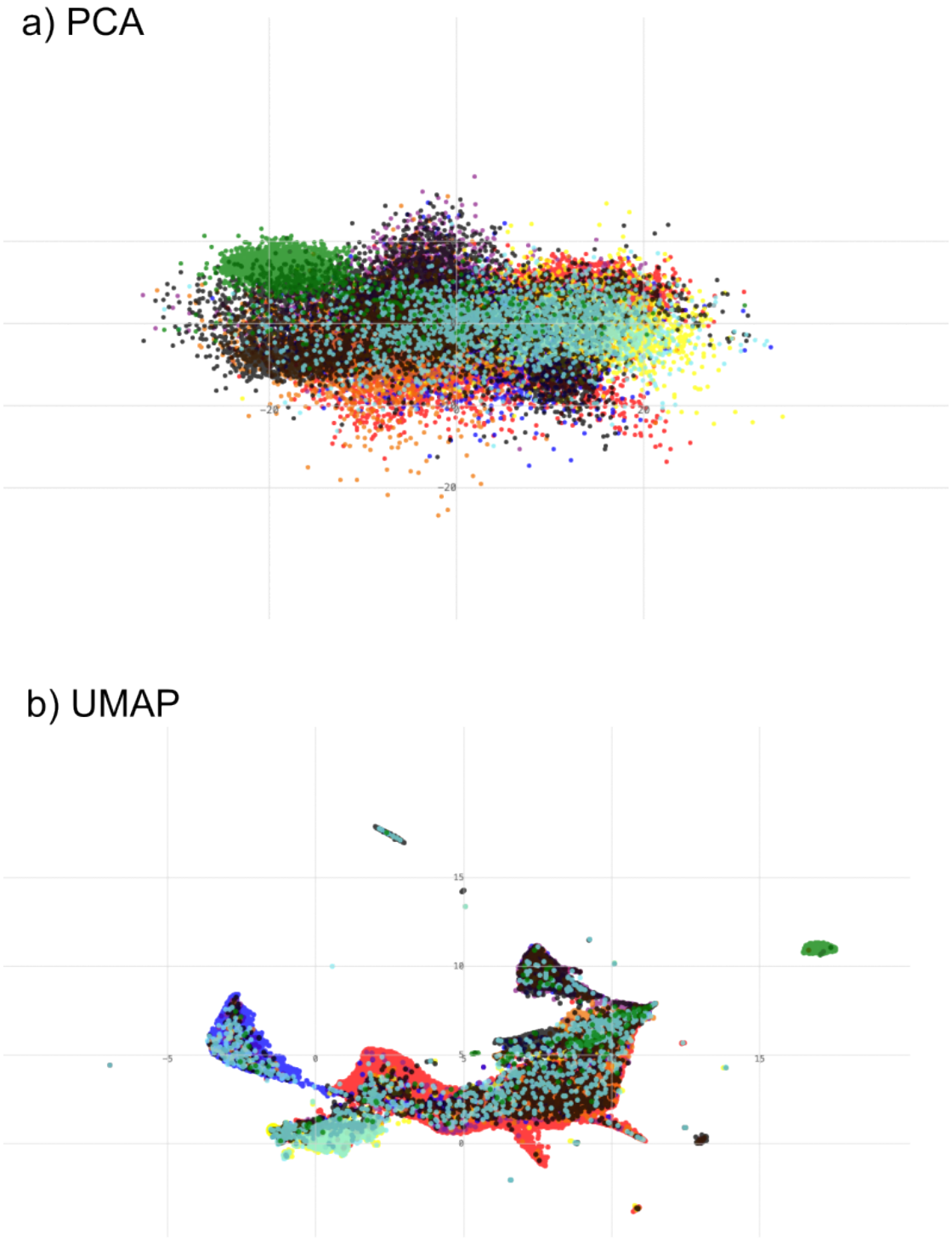


Figure B.1: BinaRena PCA and UMAP plots for the MOSAiC Ice April dataset.

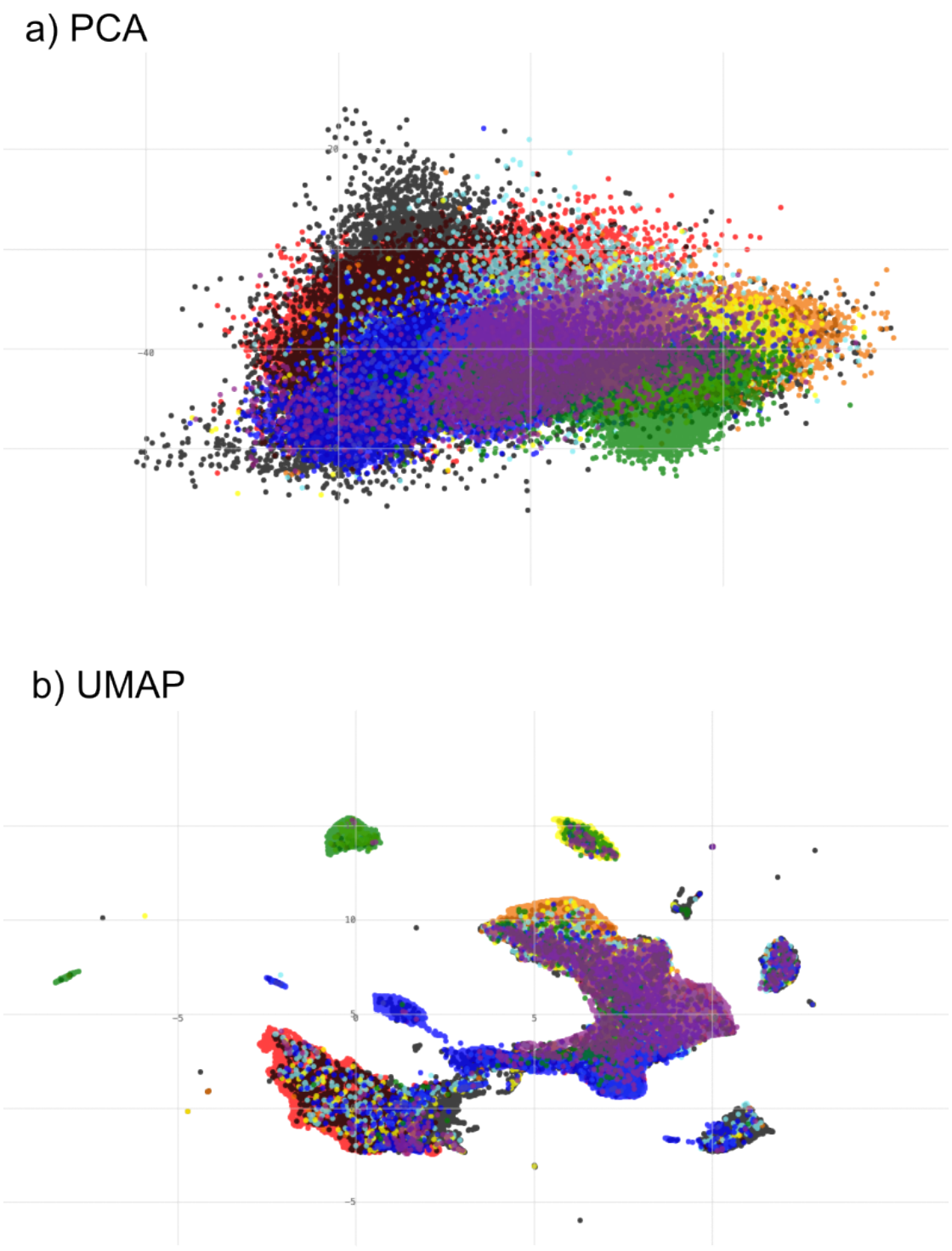


Figure B.2: BinaRena PCA and UMAP plots for the MOSAiC Water June.

B.4 VALENCE Plot for MOSAiC Water April

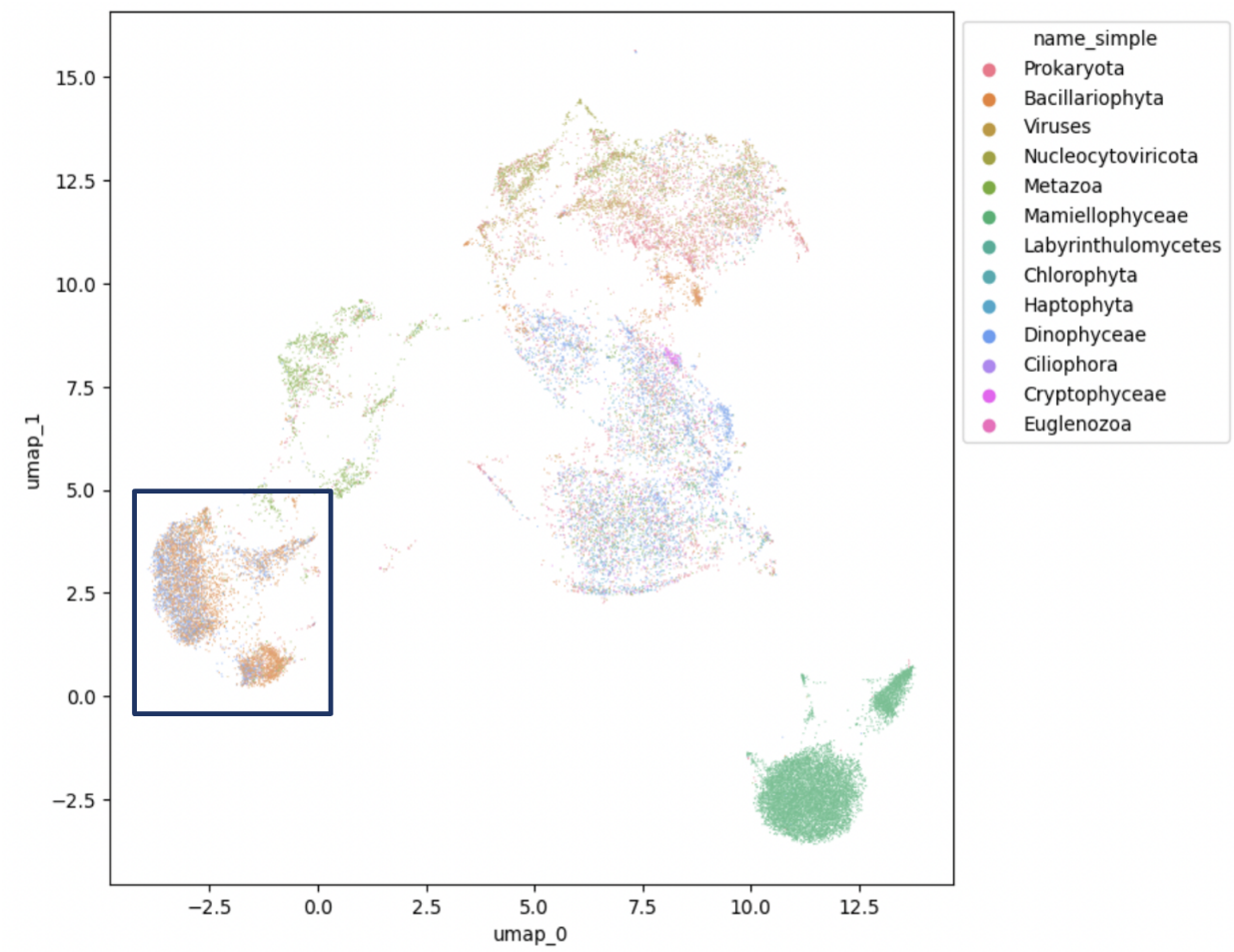


Figure B.3: VALENCE Plot for a MOSAiC Water April Dataset, where we noticed one Bacillariophyceae MAG coinciding with a large cluster of Dinophyceae contigs (marked). On inspection, the closest taxonomic hits were to *Attheya septentrionalis* and *Kryptoperidinium foliaceum*, species with a potential endosymbiotic relationship.

Chapter C

Appendix C

C.1 Coassembly Statistics

Batch Name	Num. Seqs	Sum Len.	Min Len.	Avg. Len.	Max. Len.	Q1	Q2	Q3	N50	N50_num	GC(%)
E01_1	786373	696596354	500	885.8	114902	575.0	688.0	925.0	855	6651	47.75
E01_2	1862786	1757231245	500	943.3	110251	584.0	718.0	1007.0	946	8909	46.69
E02_1	1076666	989386686	500	918.9	103663	585.0	711.0	979.0	908	7256	50.59
E02_2	834335	720984610	500	864.1	105715	570.0	683.0	916.0	839	6233	47.57
E03_1	1033131	934254664	500	904.3	83701	575.0	699.0	964.0	894	6997	49.42
E04_1	830462	728805984	500	877.6	58061	570.0	684.0	929.0	856	6554	53.31
E04_2	1297342	1201434886	500	926.1	74349	569.0	687.0	955.0	910	9147	51.38
E04_3	730082	649445626	500	889.6	111740	578.0	687.0	909.0	846	6973	49.6
E06_1	2954203	3019901512	500	1022.2	184206	581.0	711.0	1012.0	1030	14822	48.94
E07_1	1275797	1264470934	500	991.1	94742	581.0	710.0	999.0	987	10308	48.27
E07_2	1238881	1209304386	500	976.1	100583	575.0	696.0	973.0	963	10032	50.74
E09_1	1068118	1135689011	500	1063.3	117417	588.0	746.0	1121.0	1136	9097	54.53
E11_1	223304	195731460	500	876.5	81866	566.0	674.0	919.0	850	4250	46.13
E12_1	1813976	1662149658	500	916.3	105694	579.0	706.0	977.0	908	8604	48.84
I01_1	946442	1037292055	500	1096.0	113251	595.0	755.0	1117.0	1160	10476	46.89
I01_3	391006	424602907	500	1085.9	98381	585.0	727.0	1059.0	1130	8243	46.28
I02_2	771209	876178154	500	1136.1	153644	598.0	762.0	1136.0	1219	10740	46.33
I02_3	397755	462139020	500	1161.9	74297	599.0	775.0	1198.0	1301	7861	46.64
I02_4	714844	790560468	500	1105.9	74426	598.0	769.0	1163.0	1198	8920	47.02
I03_1	513274	497773471	500	969.8	105721	582.0	712.0	998.0	966	6934	46.99
I04_1	1886883	1903191730	500	1008.6	272219	582.0	723.0	1047.0	1035	11236	47.03
I04_2	1525226	1590065384	500	1042.5	201465	592.0	750.0	1101.0	1092	9942	46.27
I06_1	625946	576841054	500	921.6	98089	571.0	695.0	987.0	924	5751	48.37

I06_2	1060172	1135575658	500	1071.1	112736	588.0	739.0	1083.0	1119	10744	47.16
I06_3	1062019	994917667	500	936.8	130157	576.0	700.0	975.0	926	7667	46.93
I07_2	1534363	1428962540	500	931.3	102787	581.0	705.0	972.0	916	9254	47.55
I07_3	1959584	2057103851	500	1049.8	226926	595.0	752.0	1109.0	1103	10815	47.74
I07_7	429580	405330793	500	943.6	71850	575.0	696.0	968.0	929	6465	43.33
I07_8	1982348	2057419704	500	1037.9	192103	587.0	733.0	1058.0	1066	12447	45.15
I08_1	1715333	1780898699	500	1038.2	93271	588.0	741.0	1098.0	1093	10077	48.48
I11_1	929848	955032246	500	1027.1	98080	596.0	760.0	1113.0	1079	8059	47.43
I11_2	674928	811077223	500	1201.7	96661	599.0	776.0	1209.0	1368	10528	48.47
I12_1	484055	487872270	500	1007.9	65149	591.0	741.0	1075.0	1043	6470	46.21

Table C.1: Contig statistics of the MEGAHIT assemblies. Q1 to Q3 are length quartiles.

C.2 Accessions for Eukaryotic Reference Genomes

Phycocosm and Mycocosm Accessions

Chloso1228_1	Sceobl2630_1	Edade1	Chleu1
TetrdesSNI2_1	MicrAD1_1	Ostque1	Botrbrau1_1
Chloso1230_1	Rhodia1_1	Rhoto1	RhoCCFEE5036_1
Trivag1	Chaten1	Picre1	Psemu1
Picsp_1	Rhoba1_1	Ostta1115_2	Rapsub1_1
Chlin1	Auxpr25_1	Nitput1	Gonpec1
Astpho2	Maypse1	Astgub1	SceoblDOE13_1
Chrzof1	Cosub3	Monneg1	Sporo1
Rhomuc1	DesarB2533_2	Tetrobl72_1	Sobl393_1
ChlreiCC4532_1	RhodotJ31_1	Flerot1_1	SymretSp1
Nandes2526_1	SceoblEN4_1	Rhota1	Semro1
Scesp_1	Rhoto_IFO0880_4	Chlsc1	Spopa1
Nandes2437_1	Auxeprot1	Spogra1_1	Chloso1602_1
Pico_ML_1	Chlre5_6	SymretAf1	Mosaich1_1
Dunsal1_1	Spoli1	Cyccr2_2	Rhoto_IFO0559_1
Thaoce1	ChlNC64A_1	Ostta4221_3	Volca2_1
Batpra1	Heli50920_1	SymretSw1	Rhoglu1
Ulvmu1_1	Sceobl393_2	Trybru1	Monmin1
Tetso1	Fracy1	ChloA99_1	Dicre1
SymretSc1	Rhosp1	Giaint1	MicpuN3v2
Sceob152z_1	Thaps3		

Table C.2: Accessions and genome IDs from the Mycocosm and Phycocosm web portals, used in the species tree in Chapter 7.

NCBI Genbank Accessions

GCA Accession
GCA_900092255.1_ASM90009225v1
GCA_000231825.2_Tetra_elliot_V2
GCA_003568905.1_KSI-1_01
GCA_000691245.1_Tgr_V1
GCA_003297045.1_SymC_ver_1.0
GCA_003671325.1_LVH60
GCA_900243725.1_Aphanomyces_stellatus_v1
GCA_001586965.3_ASM158696v3
GCA_001600495.1_JCM_30514_assembly_v001
GCA_000497125.1_SSK3.0
GCA_000733215.1_ASM73321v1
GCA_000004695.1_dicty_2.7
GCA_002087225.1_Tth_v3
GCA_000220395.1_JCVI-IMG1-V.1
GCA_000318465.2_MPF4_v2.0
GCA_000002725.2_ASM272v2
GCA_000190715.1_v1.0
GCA_002811675.1_ASM281167v1
GCA_000499745.1_EMH001
GCA_000260095.1_Tetra_borealis_V1
GCA_000006355.1_ASM635v1
GCA_002024145.1_C_fragrantissima_v5
GCA_006510595.1_ASM651059v1
GCA_001447515.1_ASM144751v1
GCA_003719475.1_ASM371947v1
GCA_001186125.1_Spha_arctica_JP610_V1
GCA_900002385.1_PY17X01
GCA_003255715.1_ASM325571v1
GCA_002814315.1_ASM281431v1
GCA_000387425.2_pir_scaffolds_v1
GCA_000387445.2_pag1_scaffolds_v1

GCA_003287315.1_Pcac_10300_v1
GCA_000151265.1_Micromonas_pusilla_CCMP1545_v2.0
GCA_000002875.2_ASM287v2
GCA_001029375.1_Pythium_insidiosum_1.0
GCA_000257125.1_ENU1_v1
GCA_900240875.1_crithidia-expoeki.GDC.2015.v1
GCA_900088475.1_hypho_2016
GCA_900092275.1_ASM90009227v1
GCA_008037345.1_MG_SEMC4_Ver1.0
GCA_000963415.1_ASM96341v1
GCA_001724245.1_ASM172424v1
GCA_007859695.1_ASM785969v1
GCA_000963455.1_ASM96345v1
GCA_000387505.2_par_scaffolds_v1
GCA_000686205.4_P.fr2.0
GCA_002812785.1_ASM281278v1
GCA_000482105.1_Caca_1.0
GCA_001584585.1_ASM158458v1
GCA_000444285.2_Leishmania_aethiopica-L147-2.0.3
GCA_005317125.1_ASM531712v1
GCA_003573635.1_ASM357363v1
GCA_001880345.1_ASM188034v1
GCA_900128565.1_TOSAG23-6
GCA_001625125.1_ASM162512v1
GCA_003664525.1_ASM366452v1
GCA_000482205.1_Hmus_1.0
GCA_002087855.2_ASM208785v2
GCA_000281045.1_Sap_diclina_VS20_V1
GCA_000092065.1_ASM9206v1
GCA_000441995.1_Leishmania_turanica_LEM423-1.0.2
GCA_003612995.1_ASM361299v1
GCA_000165395.1_ASM16539v1
GCA_001273305.2_ASM127330v2
GCA_004115355.1_Halite_2017MT

GCA_003324165.1_Nlova_1.1
GCA_000691945.2_ASM69194v2
GCA_000818905.1_ASM81890v1
GCA_000499725.1_EBH001
GCA_900108755.1_sob1
GCA_001299535.1_ASM129953v1
GCA_000331125.1_PhytSerpensv01
GCA_001235845.1_ASM123584v1
GCA_002921335.1_ASM292133v1
GCA_000331325.2_Crithidia_fasciculata-14.0
GCA_000002595.2_v3.0
GCA_001766655.1_ASM176665v1
GCA_003693705.1_ASM369370v1
GCA_000982615.1_AKI_PRJEB1539_v1
GCA_000252605.1_ASM25260v1
GCA_000826245.1_Acanthamoeba_astronyxis
GCA_004335715.1_ASM433571v1
GCA_000715435.1_caudatum_43c3d_assembly_v1
GCA_001839685.1_ASM183968v1
GCA_900092265.1_ASM90009226v1
GCA_000482125.1_Sgal_1.0
GCA_002179805.1_ALPT14_1.0
GCA_000277465.1_ASM27746v1
GCA_000826945.1_ASM82694v1
GCA_000209065.1_ASM20906v1
GCA_000410755.2_Leishmania_enrietti_LEM3045-1.0.2
GCA_003843895.1_ASM384389v1
GCA_004335835.1_ASM433583v1
GCA_000826425.1_Acanthamoeba_lugdunensis
GCA_001680005.1_ASM168000v1
GCA_000325885.1_Phytophthora_capsici_LT1534_v11.0
GCA_000733375.1_ASM73337v1
GCA_000143455.1_v1.0
GCA_001430725.1_ASM143072v1

GCA_000247585.2_PP_INRA-310_V2
GCA_000372725.1_Emiliana_huxleyi_CCMP1516_main_genome_assembly_v1.0
GCA_000210295.1_ASM21029v1
GCA_000818945.1_ASM81894v1
GCA_001662385.1_Cas_assembly01
GCA_900097035.1_PBLACG01
GCA_000443025.1_Leishmania_gerbilii_LEM452-1.0.2
GCA_001643675.1_Mono14B
GCA_006782975.1_ASM678297v1
GCA_001460835.1_BSAL
GCA_004764695.1_ASM476469v1
GCA_000333855.2_Endotrypanum_monterogeii-LV88-1.0.3
GCA_003568945.1_KIPB_1.0
GCA_001314365.1_MP94-48v2
GCA_000981925.2_Ld_v2
GCA_001653735.1_Ucit_macronuclear_v.1.0
GCA_001659865.1_Angomonas_deanei_v1.0
GCA_000410715.1_Leishmania_tropica_L590-2.0.2
GCA_000165425.1_ASM16542v1
GCA_003613005.1_ASM361300v1
GCA_001922765.1_Pythium_periplocum_1.0
GCA_002216565.1_ASM221656v1
GCA_004335685.1_ASM433568v1
GCA_000520075.1_Apha_asta_APO3_V1
GCA_000410695.2_Leishmania_arabica_LEM1108-1.0.3
GCA_000482145.1_Scul_1.0
GCA_001655075.1_ASM165507v1
GCA_000963465.1_ASM96346v1
GCA_004335455.1_ASM433545v1
GCA_004335735.1_ASM433573v1
GCA_001614225.1_ASM161422v1
GCA_002335675.1_C.eustigma_genome_v1.0
GCA_000482185.1_Ades_1.0
GCA_001593455.1_CryBaiTAMU-09Q1-1.0

GCA_900538255.1_Ulvmu_WT_fa
GCA_000388065.2_Font_alba_ATCC_38817_V2
GCA_000090985.2_ASM9098v2
GCA_004337835.1_UKCU2.v0
GCA_002286825.1_Lag_gig_ARSEF373_v1.0
GCA_004335635.1_ASM433563v1
GCA_000203815.1_DFas_2.0
GCA_001598975.1_PK2152_assembly
GCA_001457755.2_Trypanosoma_equiperdum_OVI_V2
GCA_000524495.1_Plas_inui_San_Antonio_1_V1
GCA_000006405.1_JCVI_PMG_1.0
GCA_000439335.1_Phyto_alni_1.0
GCA_000981445.1_Bbig001
GCA_000687305.2_P.frag2.0
GCA_900240985.1_crithidia-bombi.GDC.2013.v1
GCA_004335645.1_ASM433564v1
GCA_000002765.2_ASM276v2
GCA_000755165.1_ASM75516v1
GCA_002288995.1_ASM228899v1
GCA_000147415.1_v_1.0
GCA_000826305.1_Acanthamoeba_healyi
GCA_001403675.1_Leishmania_peruviana_PAB-4377_V1
GCA_002287245.1_ASM228724v1
GCA_004138255.1_ASM413825v1
GCA_003665715.1_OU_Pico_1.0
GCA_003130725.1_ASM313072v1
GCA_000482165.1_Sonc_1.0
GCA_003640625.1_ASM364062v1
GCA_003833335.1_Pldbra_eH_r1
GCA_003676415.1_INRA_Pmur_1
GCA_006503475.1_ASM650347v1
GCA_002247145.1_Plurivora_assembly_v1.fn
GCA_001651215.1_ASM165121v1
GCA_000826265.1_Acanthamoeba_culbertsoni_genome_assembly

GCA_000234665.4_ASM23466v4
GCA_001314345.1_NZFS3378v2
GCA_000512085.1_Reti_assembly1.0
GCA_005966545.1_ASM596654v1
GCA_000635995.1_ASM63599v1
GCA_001606155.1_ASM160615v1
GCA_004335555.1_ASM433555v1
GCA_000582765.1_AKH_PRJEB1535_v1
GCA_003664395.1_CDC_Llain_216-34_v1
GCA_002245815.2_ASM224581v2
GCA_002081555.1_ASM208155v1
GCA_002151225.1_ASM215122v1
GCA_001712635.2_PfChile5v2.0
GCA_003719485.1_ASM371948v1
GCA_005223375.1_ASM522337v1
GCA_000227135.2_ASM22713v2
GCA_000149755.2_P.sojae_V3.0
GCA_000409445.2_Leishmania_MAR_LEM2494-1.0.3
GCA_000002845.2_ASM284v2
GCA_000165365.1_ASM16536v1
GCA_003024175.1_ASM302417v1
GCA_003203535.1_Rsub_1.0
GCA_002751075.1_ASM275107v1
GCA_002794665.1_JCM_9641_assembly_v001
GCA_001179505.1_Vbrassicaformis
GCA_006384855.1_TSEL_PacBio_SMRT
GCA_001293395.1_ASM129339v1
GCA_000004825.1_PolPal_Dec2009
GCA_001273295.2_ASM127329v2
GCA_004337795.1_ABER_CVIA_1.0

Table C.3: Genbank accessions for eukaryotes used in the species tree in Chapter 7.

C.3 Correlations Between Physical Parameters

	temperature	phosphate	salinity	ammonium	nitrate	nitrite	silicate
temperature	1.00	-	-	-	-	-	-
phosphate	0.27	1.00	-	-	-	-	-
salinity	0.38	0.88	1.00	-	-	-	-
ammonium	-0.33	-0.41	-0.49	1.00	-	-	-
nitrate	0.25	0.79	0.63	-0.31	1.00	-	-
nitrite	0.21	0.15	0.28	-0.10	-0.14	1.00	-
silicate	0.14	0.82	0.75	-0.38	0.46	0.21	1.00

Table C.4: Pearson correlation coefficients between the sample physical parameters from Figure 7.4.

C.4 MAGs Per Phylum

Domain	Phylum	MAGs	ANI Clusters	GTDB Lineages
Archaea	other Archaea	37	17	16
Archaea	Halobacteriota	87	47	23
Archaea	Thermoplasmatota	359	102	30
Bacteria	Acidobacteriota	59	36	14
Bacteria	Actinomycetota	637	181	62
Bacteria	Alphaproteobacteria	1413	438	194
Bacteria	Bacteroidota	1842	479	157
Bacteria	Bdellovibrionota	63	8	6
Bacteria	Campylobacterota	6	6	2
Bacteria	Chlamydiota	4	2	2
Bacteria	Chloroflexota	400	149	63
Bacteria	Cyanobacteriota	127	22	4
Bacteria	Deinococcota	1	1	1
Bacteria	Dependentiae	1	1	1
Bacteria	Desulfobacterota	5	4	4
Bacteria	Desulfobacterota_B	17	6	4
Bacteria	Desulfobacterota_D	19	11	2
Bacteria	Desulfobacterota_E	1	1	1
Bacteria	Electryoneota	2	1	1
Bacteria	Fusobacteriota	1	1	1
Bacteria	Gammaproteobacteria	3265	704	263
Bacteria	Gemmatimonadota	74	39	15
Bacteria	Hydrogenedentota	1	1	1
Bacteria	Krumholzibacteriota	4	1	1
Bacteria	Latescibacterota	58	18	11
Bacteria	Margulisbacteria	3	3	2
Bacteria	Marinisomatota	78	21	12
Bacteria	Myxococcota	217	101	38
Bacteria	Nitrospinota	15	10	5
Bacteria	Nitrospirota	1	1	1
Bacteria	Omnitrophota	5	4	2
Bacteria	Patescibacteria	81	49	18
Bacteria	Planctomycetota	447	204	73
Bacteria	SAR324	58	33	7
Bacteria	Sumerlaeota	1	1	1
Bacteria	UBA8248	7	4	3
Bacteria	Verrucomicrobiota	568	181	52

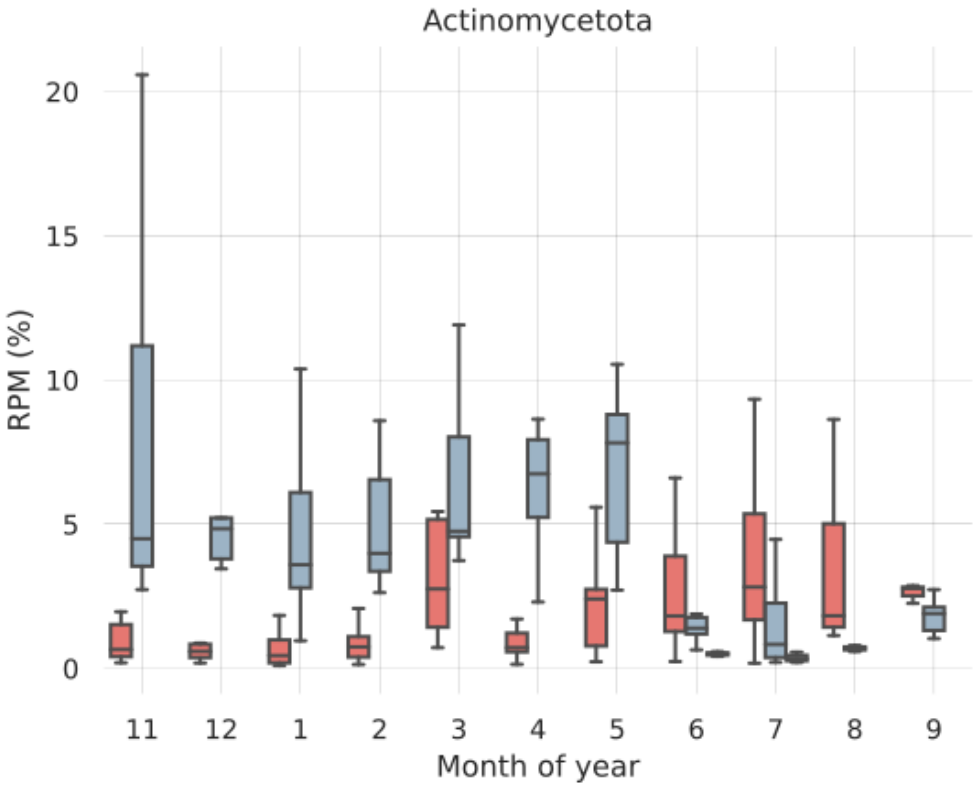
Table C.5: Numbers of prokaryotic MAGs per phylum, including the number of MAGs, number of 99% ANI clusters, and number of distinct GTDB lineages (species, or above).

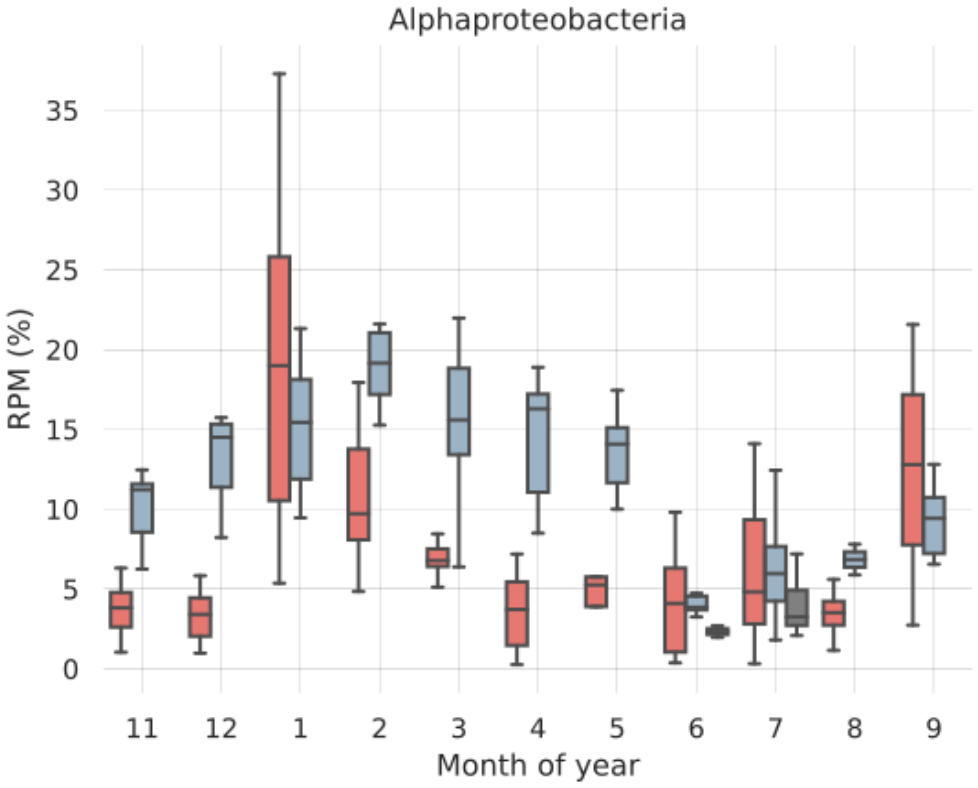
Domain	Phylum	MAGs	ANI Clusters	Lineages
Eukaryota	Bacillariaceae	113	84	5
Eukaryota	Bacillariophyta	22	20	10
Eukaryota	Chlorophyta	15	14	3
Eukaryota	Chrysophyceae	22	22	4
Eukaryota	Ciliophora	9	8	3
Eukaryota	Euglenozoa	9	8	2
Eukaryota	Fungi	20	12	3
Eukaryota	Haptophyta	23	21	2
Eukaryota	Metazoa	5	5	1
Eukaryota	Micromonas	75	27	2
Eukaryota	Phaeodactylaceae	24	21	1
Eukaryota	other eukaryotes	10	10	8
Viruses	Caudoviricetes	2359	1	22
Viruses	Mimiviridae	217	1	1
Viruses	Nucleocytoviricota	109	1	8
Viruses	Phycodnaviridae	1335	1	1
Viruses	Preplasmiviricota	46	1	4
Viruses	Riboviria	44	1	28
Viruses	Viruses	502	1	12

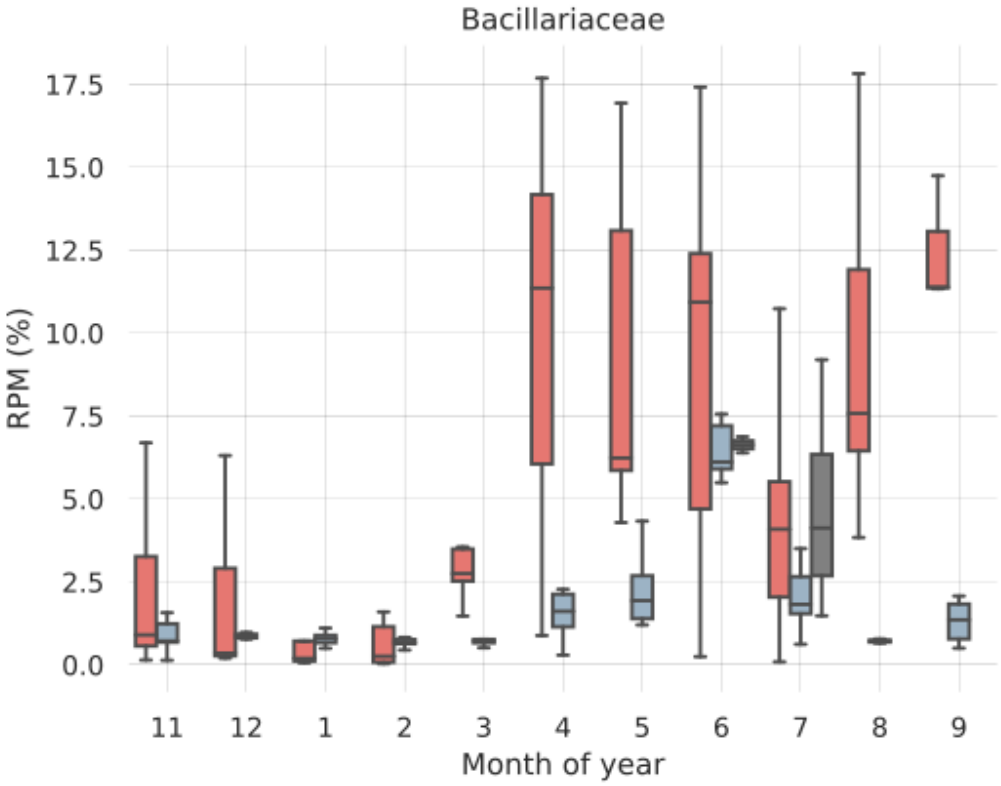
Table C.6: Numbers of eukaryotic and viral MAGs per phylum, including the number of MAGs, number of 99% ANI clusters, and number of distinct lineages (species, or above), based on the taxonomy of contigs and position within the eukaryotic species tree, or annotation from genomad. Lineage is typically no more specific than genus level, and more often family.

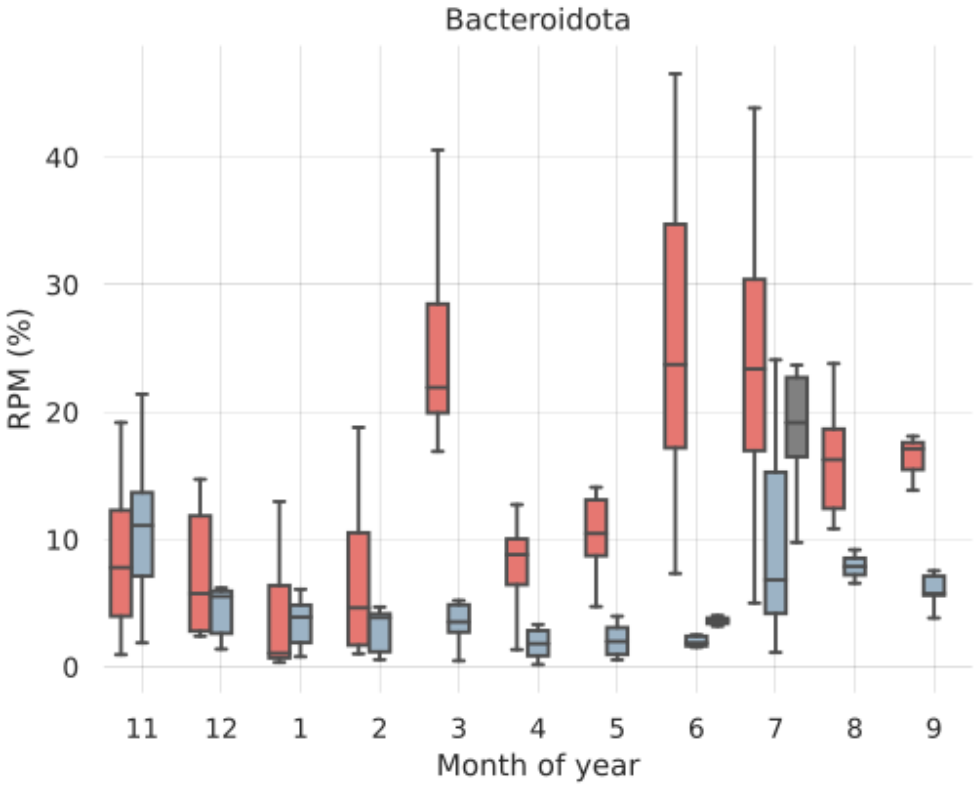
C.5 Abundance Plots

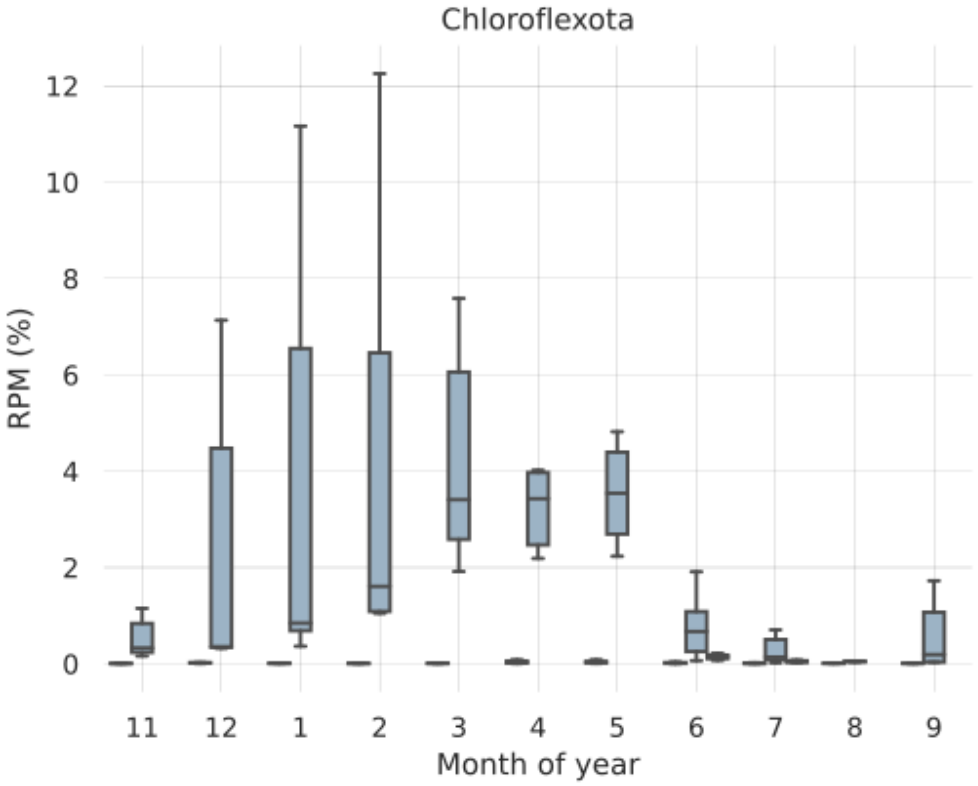
Abundances of major prokaryotic and eukaryotic clades (and viruses), per month of the drift. In all plots, ice samples are coloured red, water blue, and sediment trap samples grey.

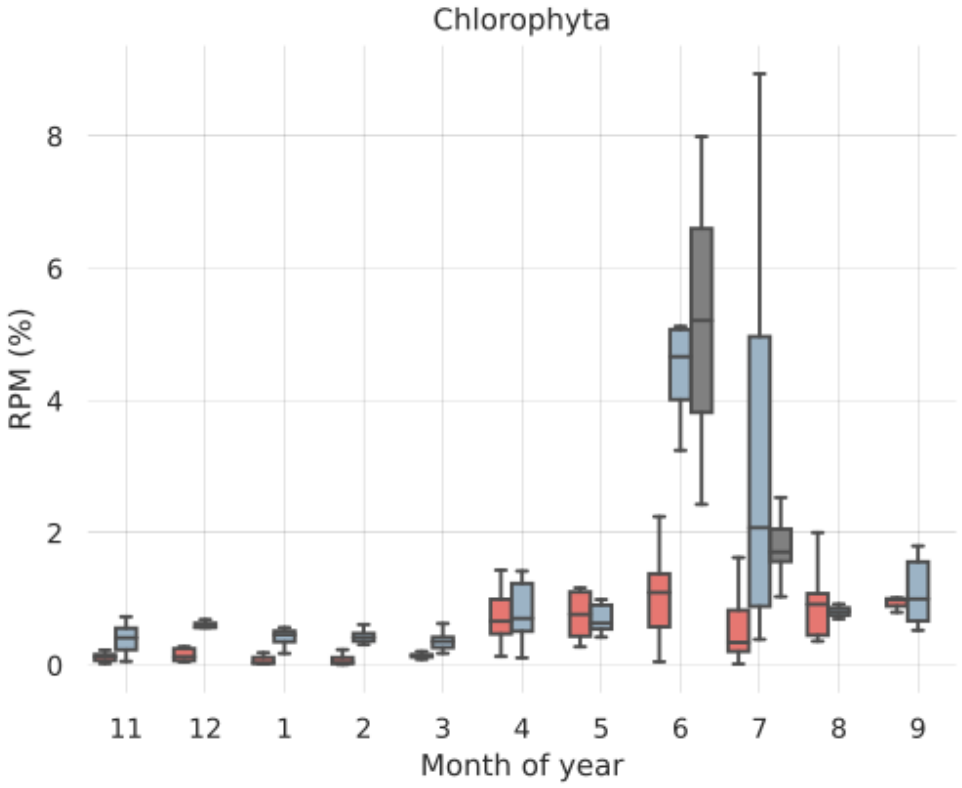

Actinomycetota
RPM (%)
0
5
10
15
20
11
12
1
2
3
4
5
6
7
8
9
Month of year

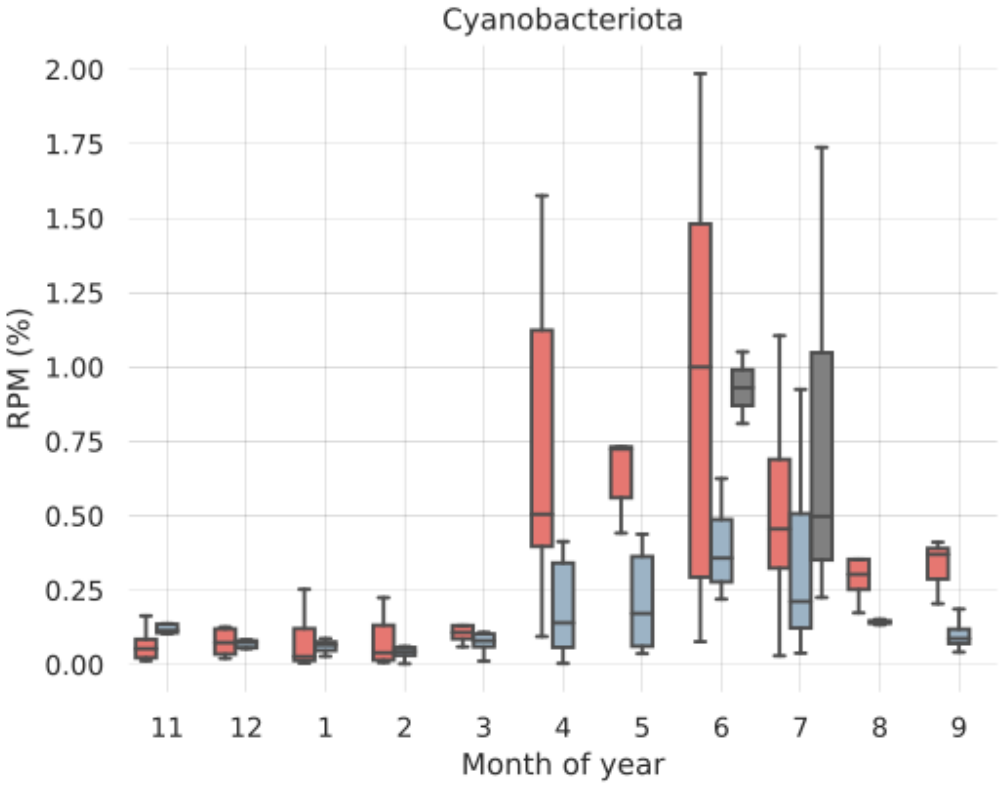

Alphaproteobacteria
RPM (%)
0
5
10
15
20
25
30
35
11
12
1
2
3
4
5
6
7
8
9
Month of year

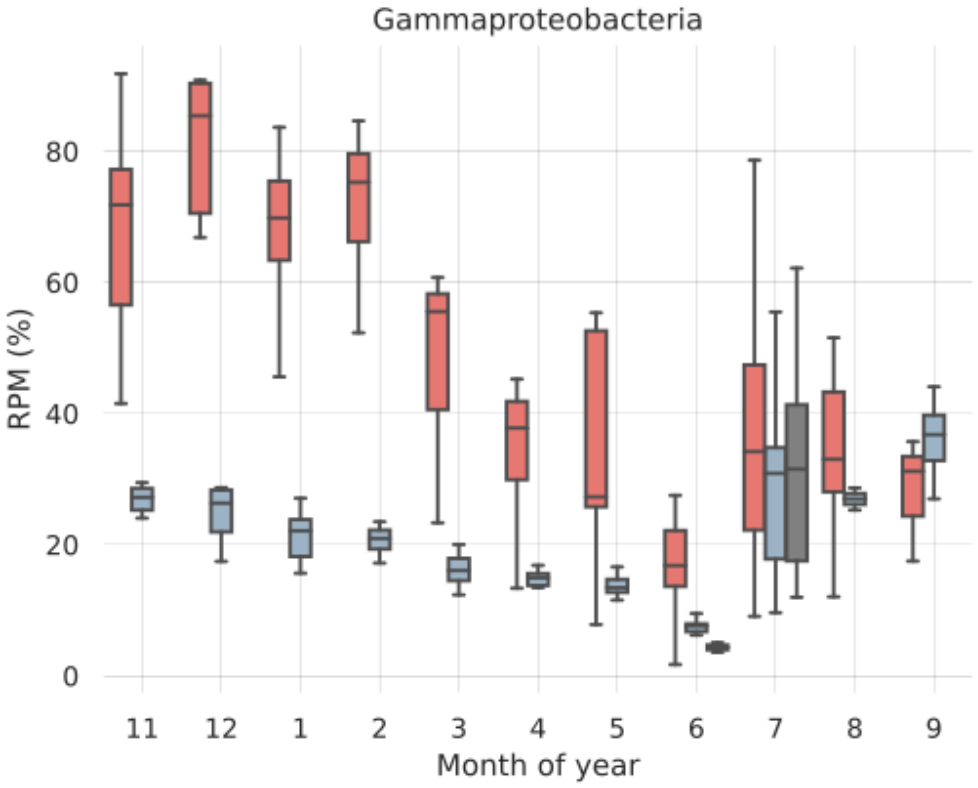

Bacillariaceae
RPM (%)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
11
12
1
2
3
4
5
6
7
8
9
Month of year

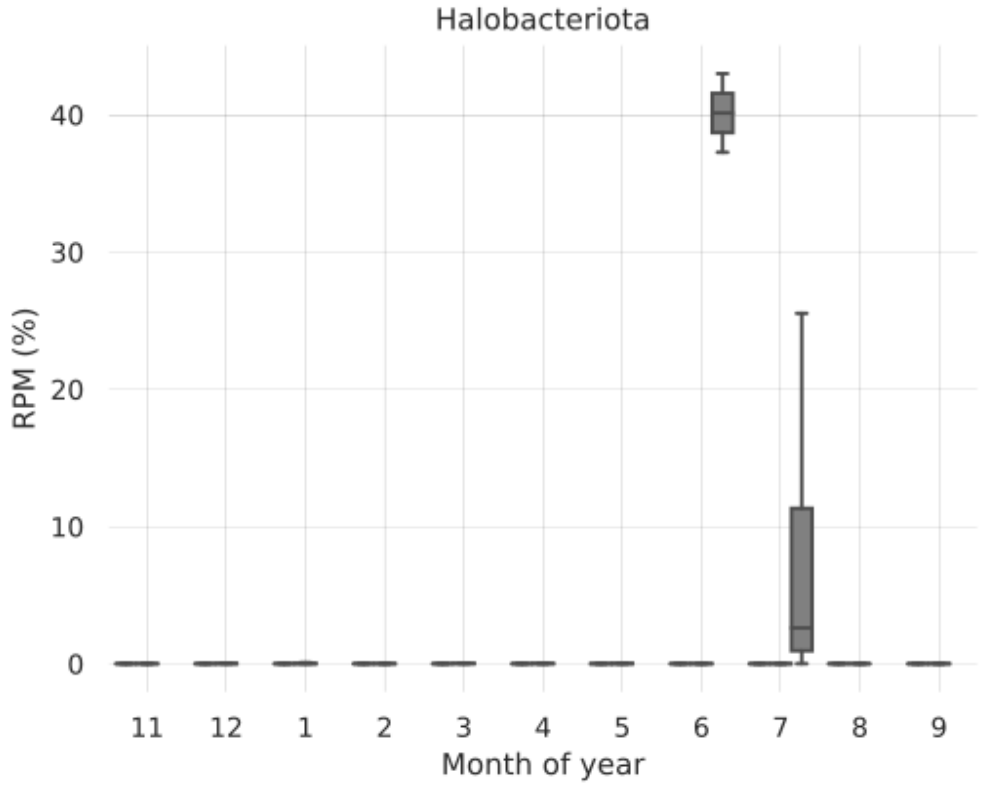

Bacteroidota
RPM (%)
40
30
20
10
0
11
12
1
2
3
4
5
6
7
8
9
Month of year

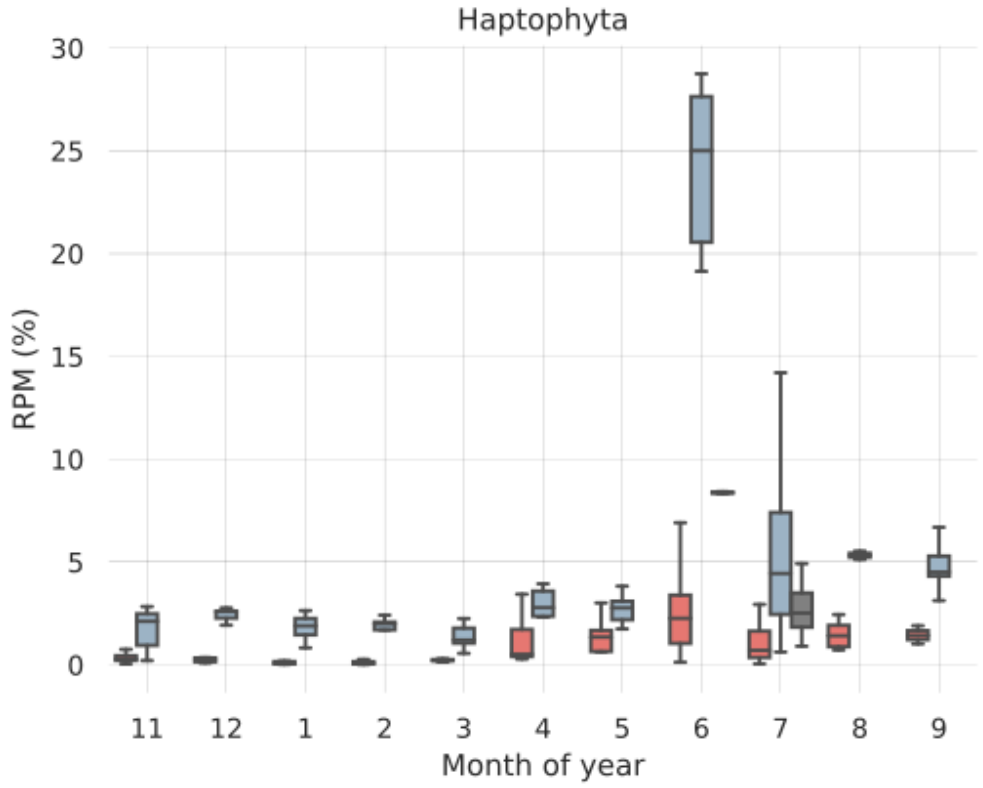

Chloroflexota
RPM (%)
0
2
4
6
8
10
12
11
12
1
2
3
4
5
6
7
8
9
Month of year

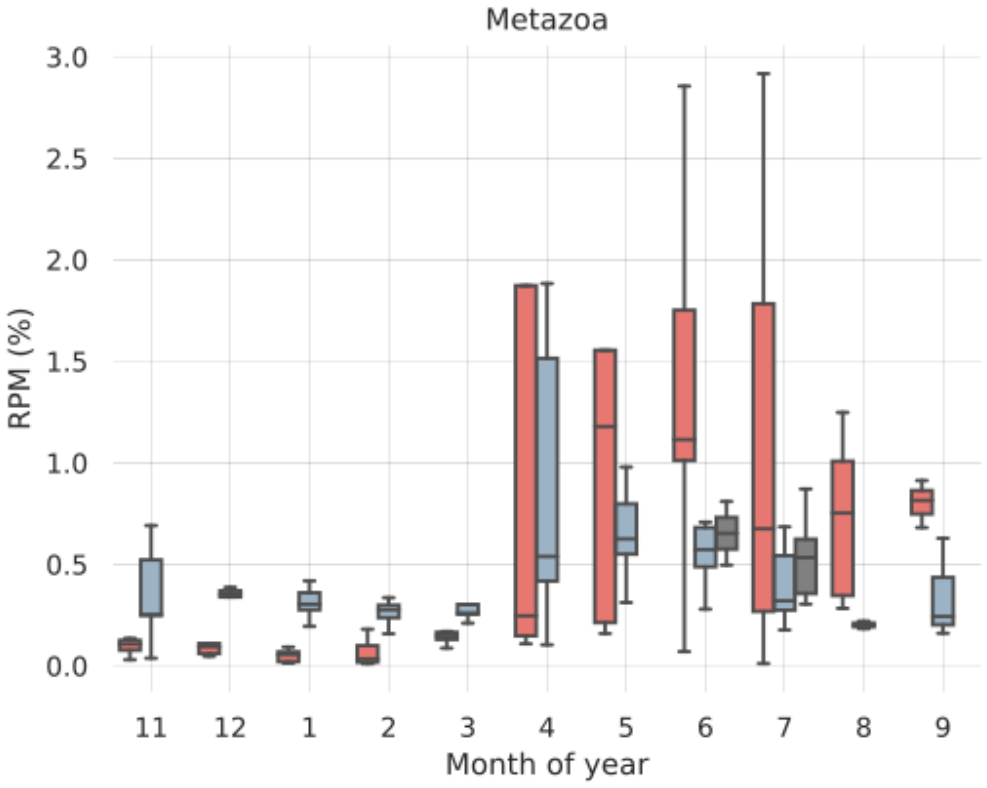

Chlorophyta
RPM (%)
0
2
4
6
8
11
12
1
2
3
4
5
6
7
8
9
Month of year

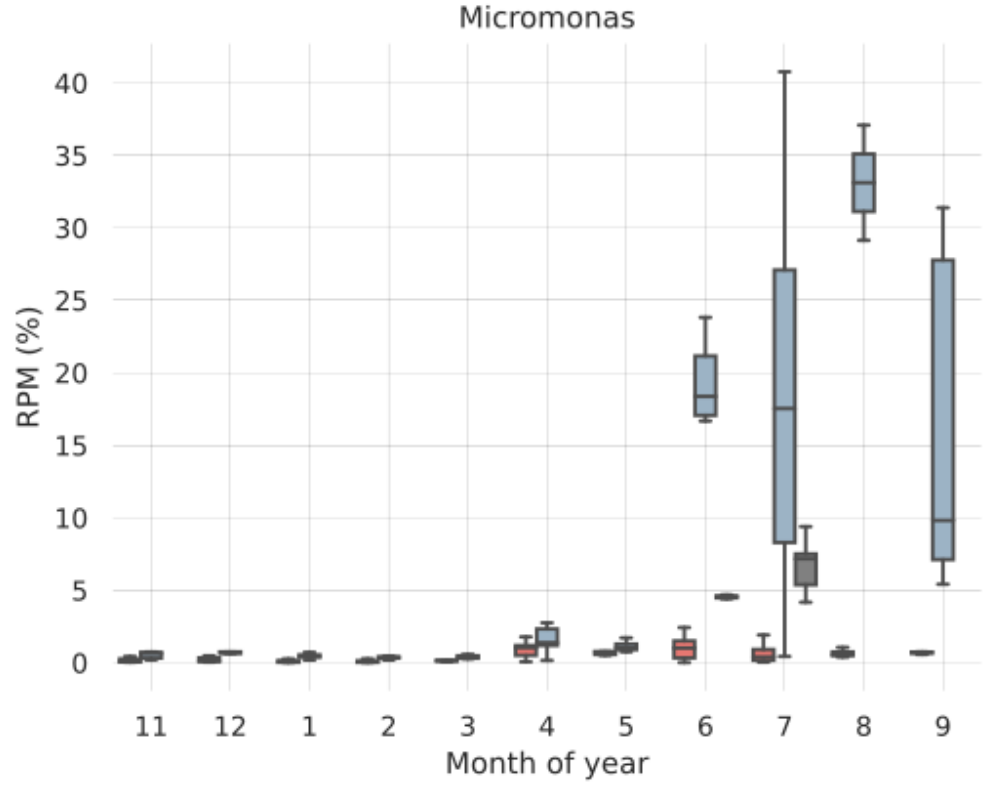

Cyanobacteriota
RPM (%)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
11
12
1
2
3
4
5
6
7
8
9
Month of year

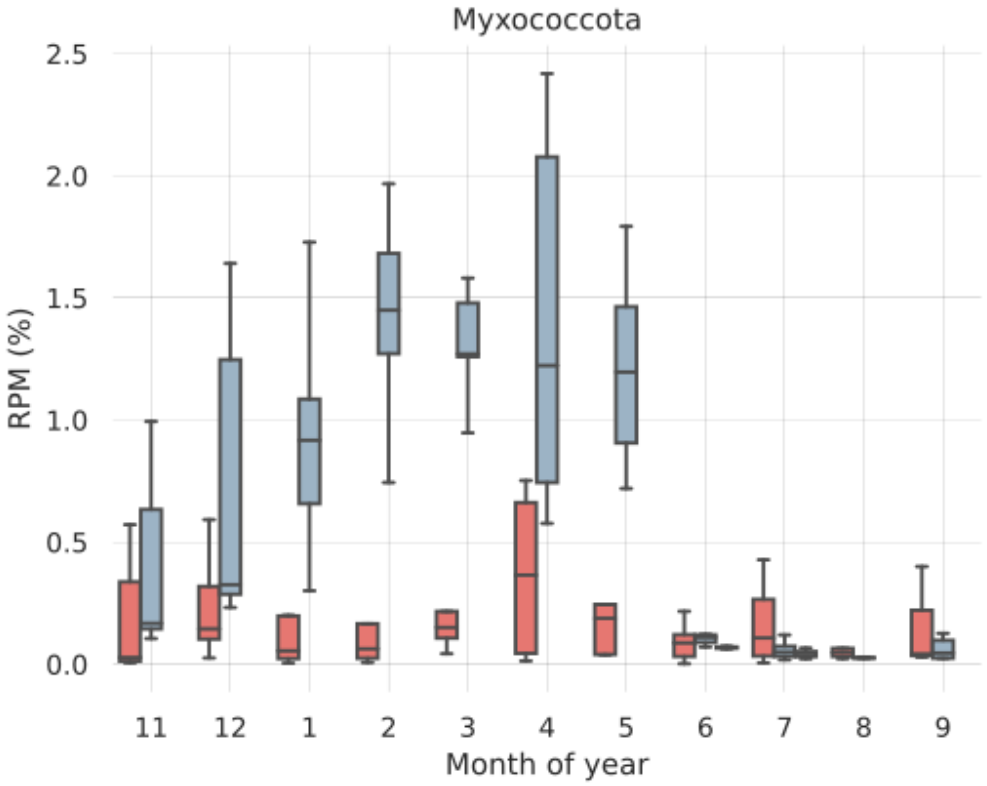

Gammaproteobacteria
RPM (%)
80
60
40
20
0
11
12
1
2
3
4
5
6
7
8
9
Month of year

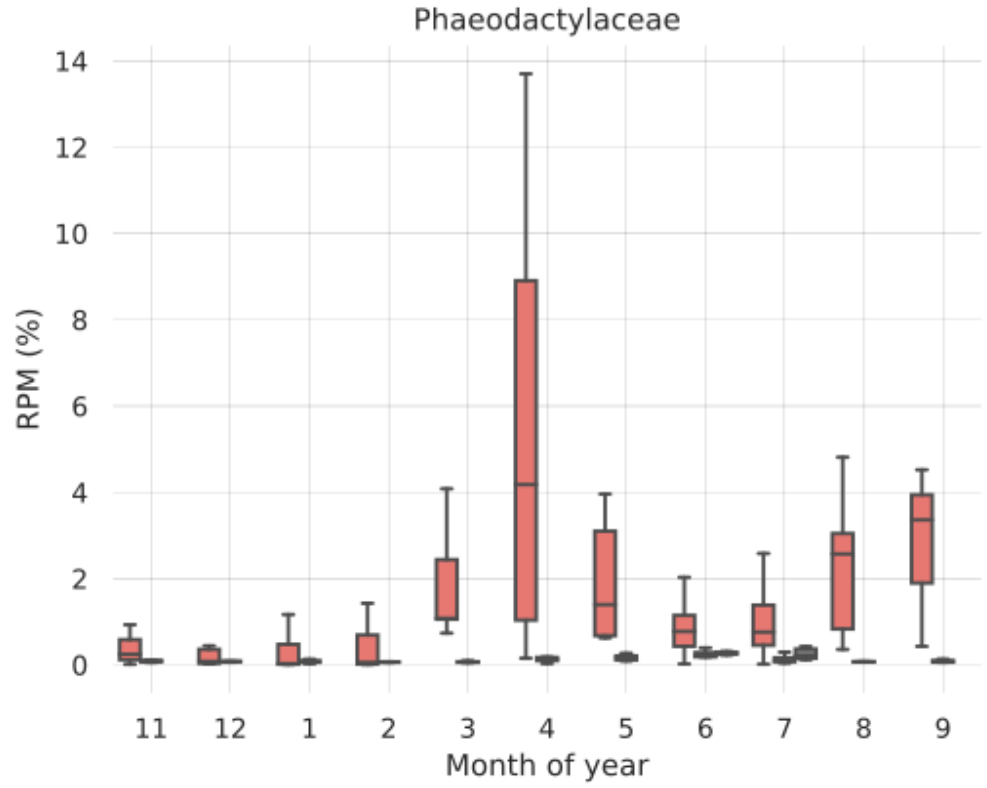

Halobacteriota
RPM (%)
0
10
20
30
40
11
12
1
2
3
4
5
6
7
8
9
Month of year

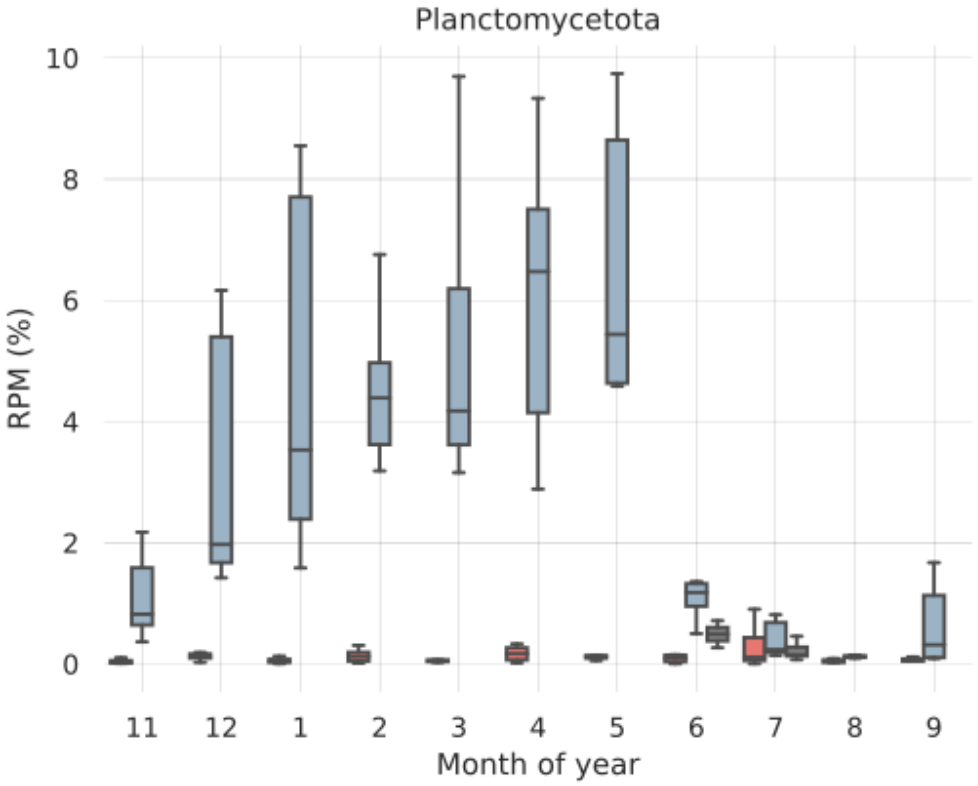

Haptophyta
RPM (%)
0
5
10
15
20
25
30
11
12
1
2
3
4
5
6
7
8
9
Month of year

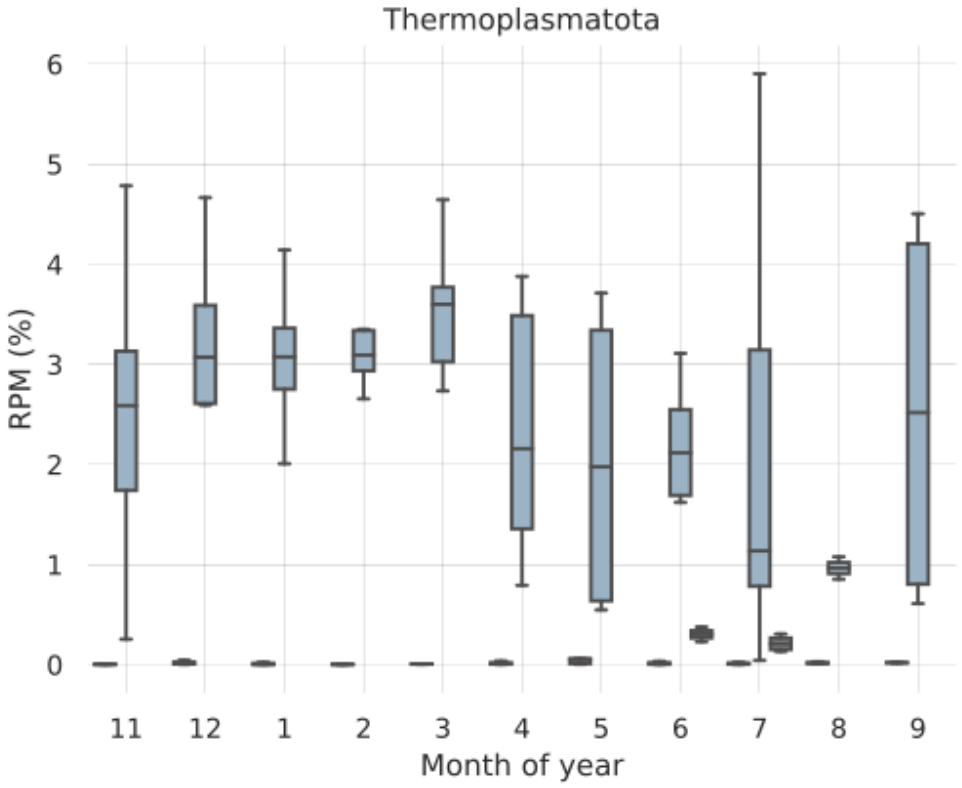

Metazoa
RPM (%)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
11
12
1
2
3
4
5
6
7
8
9
Month of year

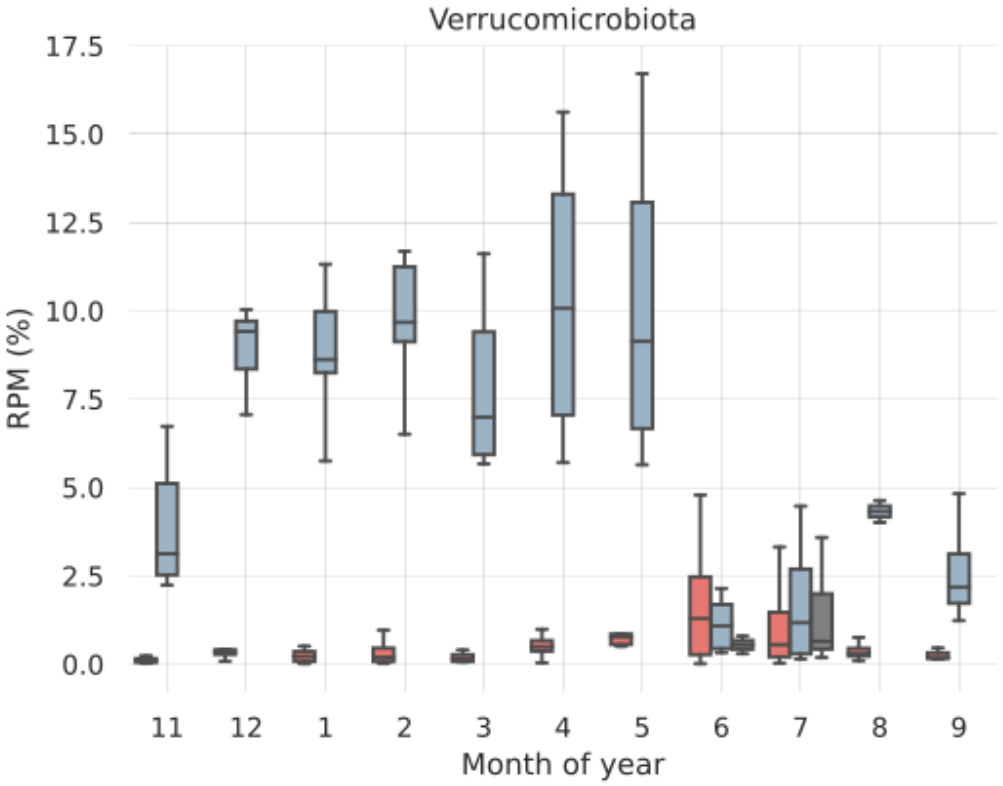

Micromonas
RPM (%)
40
35
30
25
20
15
10
5
0
11
12
1
2
3
4
5
6
7
8
9
Month of year

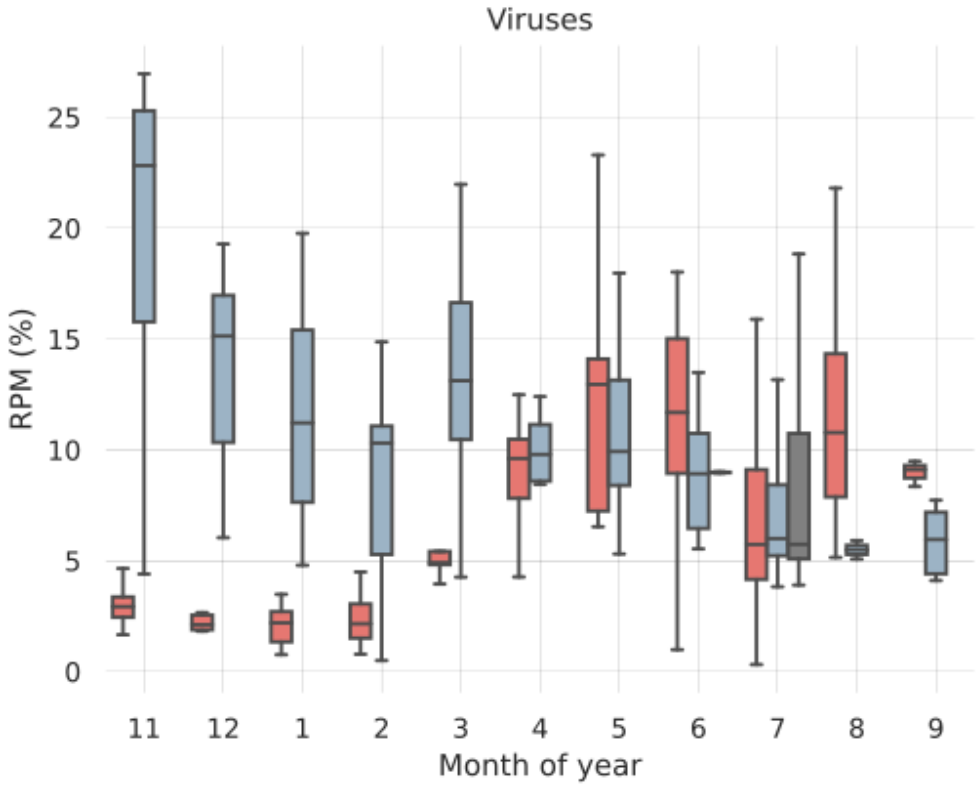

Myxococcota
RPM (%)
2.5
2.0
1.5
1.0
0.5
0.0
11
12
1
2
3
4
5
6
7
8
9
Month of year


Phaeodactylaceae
RPM (%)
14
12
10
8
6
4
2
0
11
12
1
2
3
4
5
6
7
8
9
Month of year

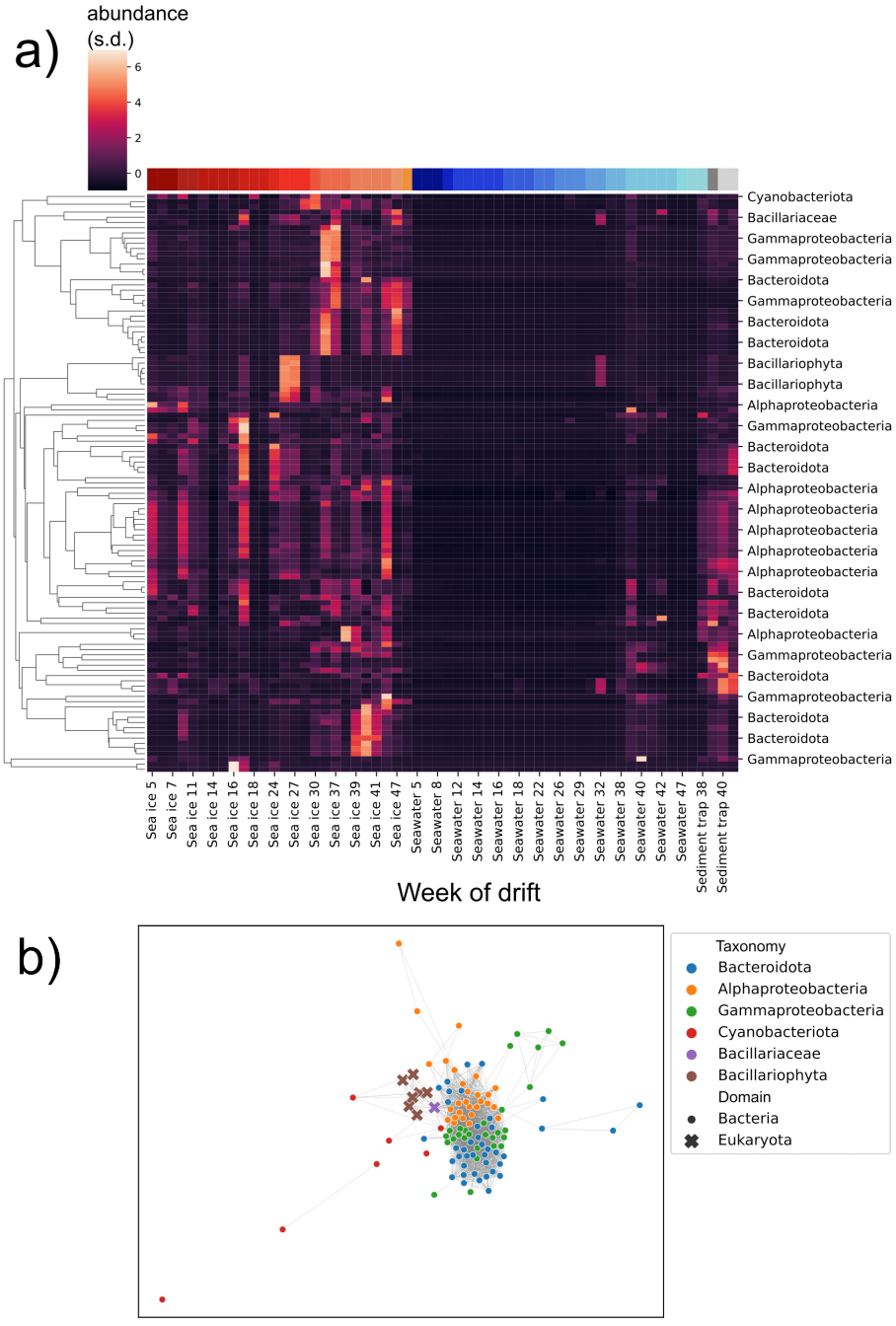

Planctomycetota
RPM (%)
0
2
4
6
8
10
11
12
1
2
3
4
5
6
7
8
9
Month of year

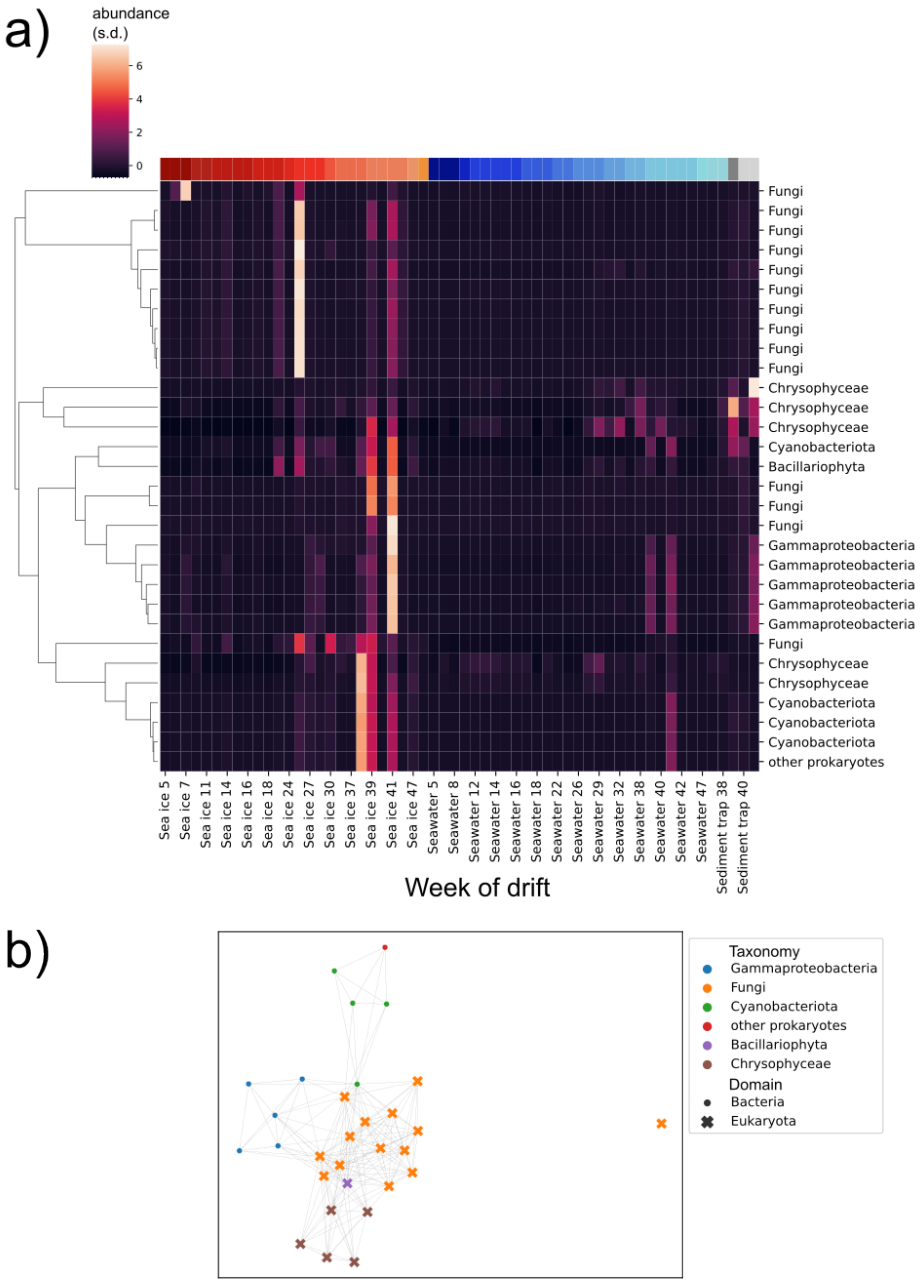

Thermoplasmatota
RPM (%)
0
1
2
3
4
5
6
11
12
1
2
3
4
5
6
7
8
9
Month of year

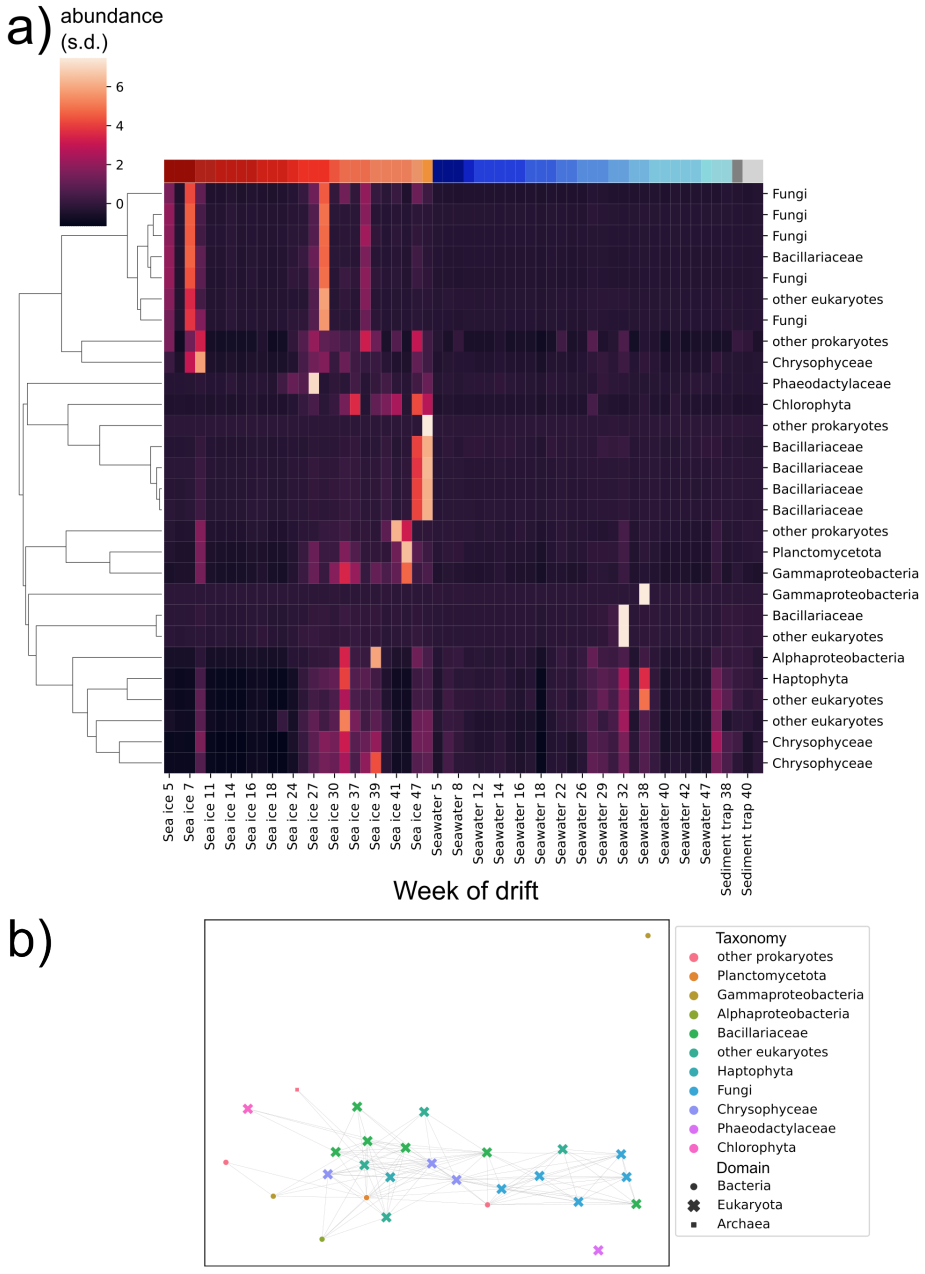

Verrucomicrobiota
RPM (%)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
11
12
1
2
3
4
5
6
7
8
9
Month of year

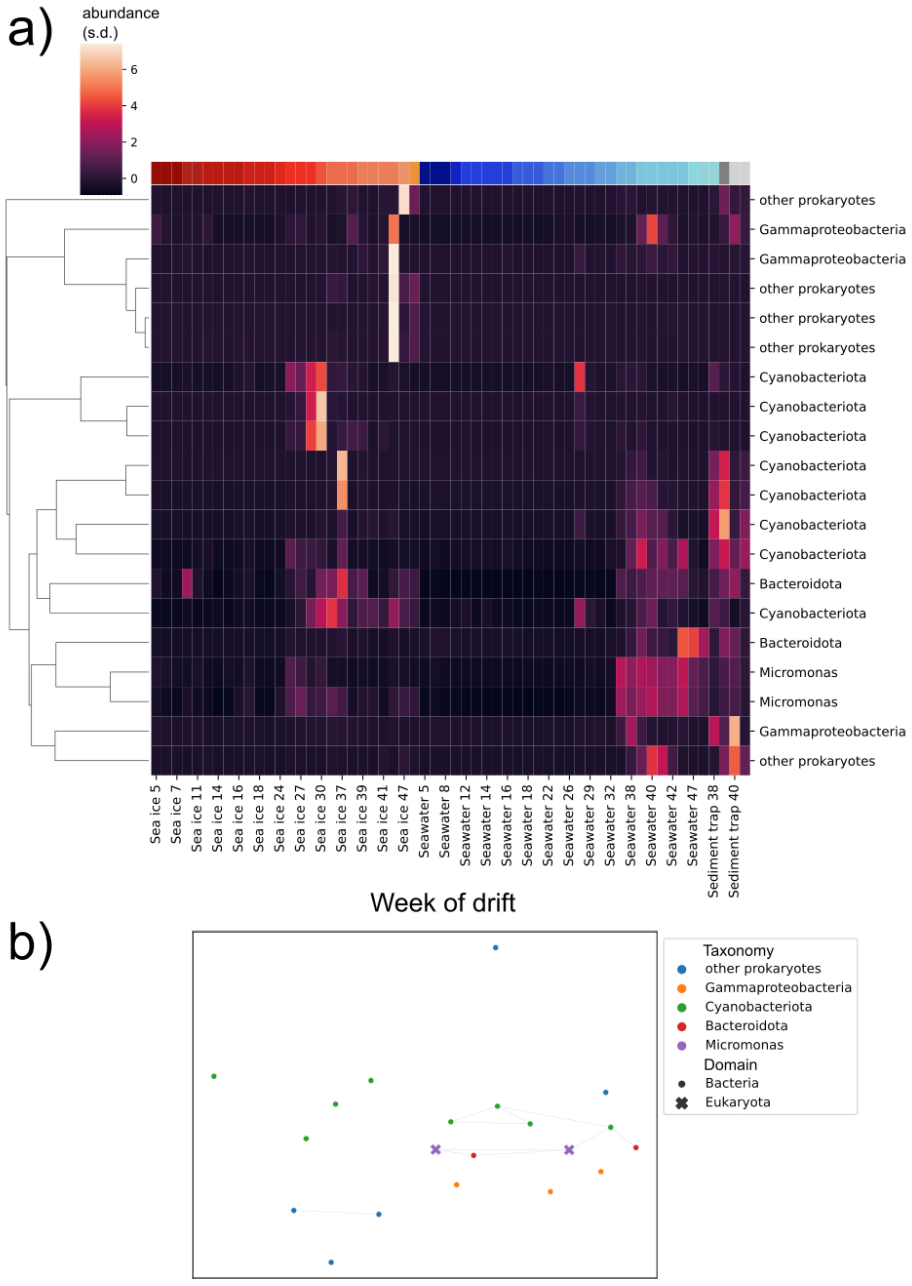

Viruses
RPM (%)
25
20
15
10
5
0
11
12
1
2
3
4
5
6
7
8
9
Month of year

C.5.1 Eukaryote-Prokaryote Network Modules




a)
abundance
(s.d.)
6
4
2
0
Fungi
Fungi
Fungi
Fungi
Fungi
Fungi
Fungi
Fungi
Fungi
Fungi
Chrysophyceae
Chrysophyceae
Chrysophyceae
Cyanobacteriota
Bacillariophyta
Fungi
Fungi
Fungi
Gammaproteobacteria
Gammaproteobacteria
Gammaproteobacteria
Gammaproteobacteria
Gammaproteobacteria
Fungi
Chrysophyceae
Chrysophyceae
Cyanobacteriota
Cyanobacteriota
Cyanobacteriota
other prokaryotes
Sea ice 5
Sea ice 7
Sea ice 11
Sea ice 14
Sea ice 16
Sea ice 18
Sea ice 24
Sea ice 27
Sea ice 30
Sea ice 37
Sea ice 39
Sea ice 41
Sea ice 47
Seawater 5
Seawater 8
Seawater 12
Seawater 14
Seawater 16
Seawater 18
Seawater 22
Seawater 26
Seawater 29
Seawater 32
Seawater 38
Seawater 40
Seawater 42
Seawater 47
Sediment trap 38
Sediment trap 40
Week of drift
b)
Taxonomy
Gammaproteobacteria
Fungi
Cyanobacteriota
other prokaryotes
Bacillariophyta
Chrysophyceae
Domain
Bacteria
Eukaryota


a)
abundance
(s.d.)
6
4
2
0
Fungi
Fungi
Fungi
Bacillariaceae
Fungi
other eukaryotes
Fungi
other prokaryotes
Chrysophyceae
Phaeodactylaceae
Chlorophyta
other prokaryotes
Bacillariaceae
Bacillariaceae
Bacillariaceae
Bacillariaceae
other prokaryotes
Planctomycetota
Gammaproteobacteria
Gammaproteobacteria
Bacillariaceae
other eukaryotes
Alphaproteobacteria
Haptophyta
other eukaryotes
other eukaryotes
Chrysophyceae
Chrysophyceae
Sea ice 5
Sea ice 7
Sea ice 11
Sea ice 14
Sea ice 16
Sea ice 18
Sea ice 24
Sea ice 27
Sea ice 30
Sea ice 37
Sea ice 39
Sea ice 41
Sea ice 47
Seawater 5
Seawater 8
Seawater 12
Seawater 14
Seawater 16
Seawater 18
Seawater 22
Seawater 26
Seawater 29
Seawater 32
Seawater 38
Seawater 40
Seawater 42
Seawater 47
Sediment trap 38
Sediment trap 40
Week of drift
b)
Taxonomy
other prokaryotes
Planctomycetota
Gammaproteobacteria
Alphaproteobacteria
Bacillariaceae
other eukaryotes
Haptophyta
Fungi
Chrysophyceae
Phaeodactylaceae
Chlorophyta
Domain
Bacteria
Eukaryota
Archaea


a)
abundance
(s.d.)
6
4
2
0
other prokaryotes
Gammaproteobacteria
Gammaproteobacteria
other prokaryotes
other prokaryotes
other prokaryotes
Cyanobacteriota
Cyanobacteriota
Cyanobacteriota
Cyanobacteriota
Cyanobacteriota
Cyanobacteriota
Cyanobacteriota
Bacteroidota
Cyanobacteriota
Bacteroidota
Micromonas
Micromonas
Gammaproteobacteria
other prokaryotes
Sea ice 5
Sea ice 7
Sea ice 11
Sea ice 14
Sea ice 16
Sea ice 18
Sea ice 24
Sea ice 27
Sea ice 30
Sea ice 37
Sea ice 39
Sea ice 41
Sea ice 47
Seawater 5
Seawater 8
Seawater 12
Seawater 14
Seawater 16
Seawater 18
Seawater 22
Seawater 26
Seawater 29
Seawater 32
Seawater 38
Seawater 40
Seawater 42
Seawater 47
Sediment trap 38
Sediment trap 40
Week of drift
b)
Taxonomy
other prokaryotes
Gammaproteobacteria
Cyanobacteriota
Bacteroidota
Micromonas
Domain
Bacteria
Eukaryota

C.6 MAG Abundance PCoA Coordinates

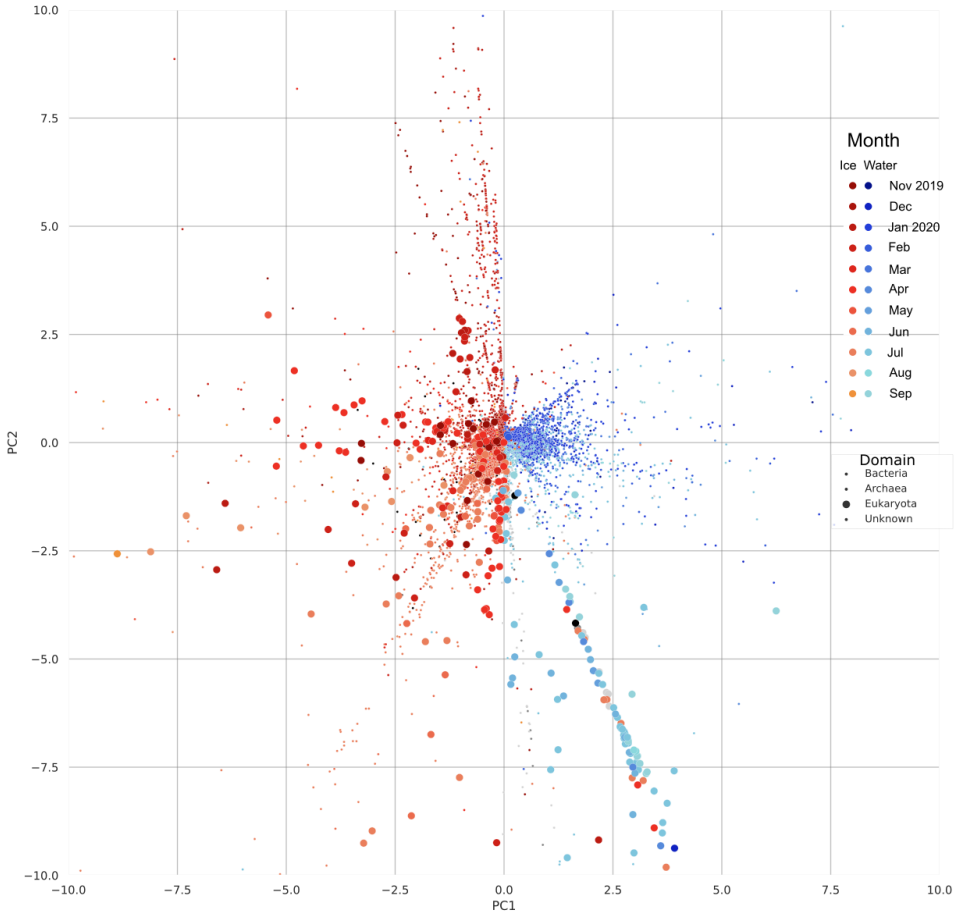


Figure C.1: Projection of MAG abundances onto the PCoA components of beta diversity; points indicate MAGs and their provenance (i.e. the sample type from which they were recovered) is indicated by the colour. Eukaryotes are indicated by larger dots.

C.7 WGCNA Abundances

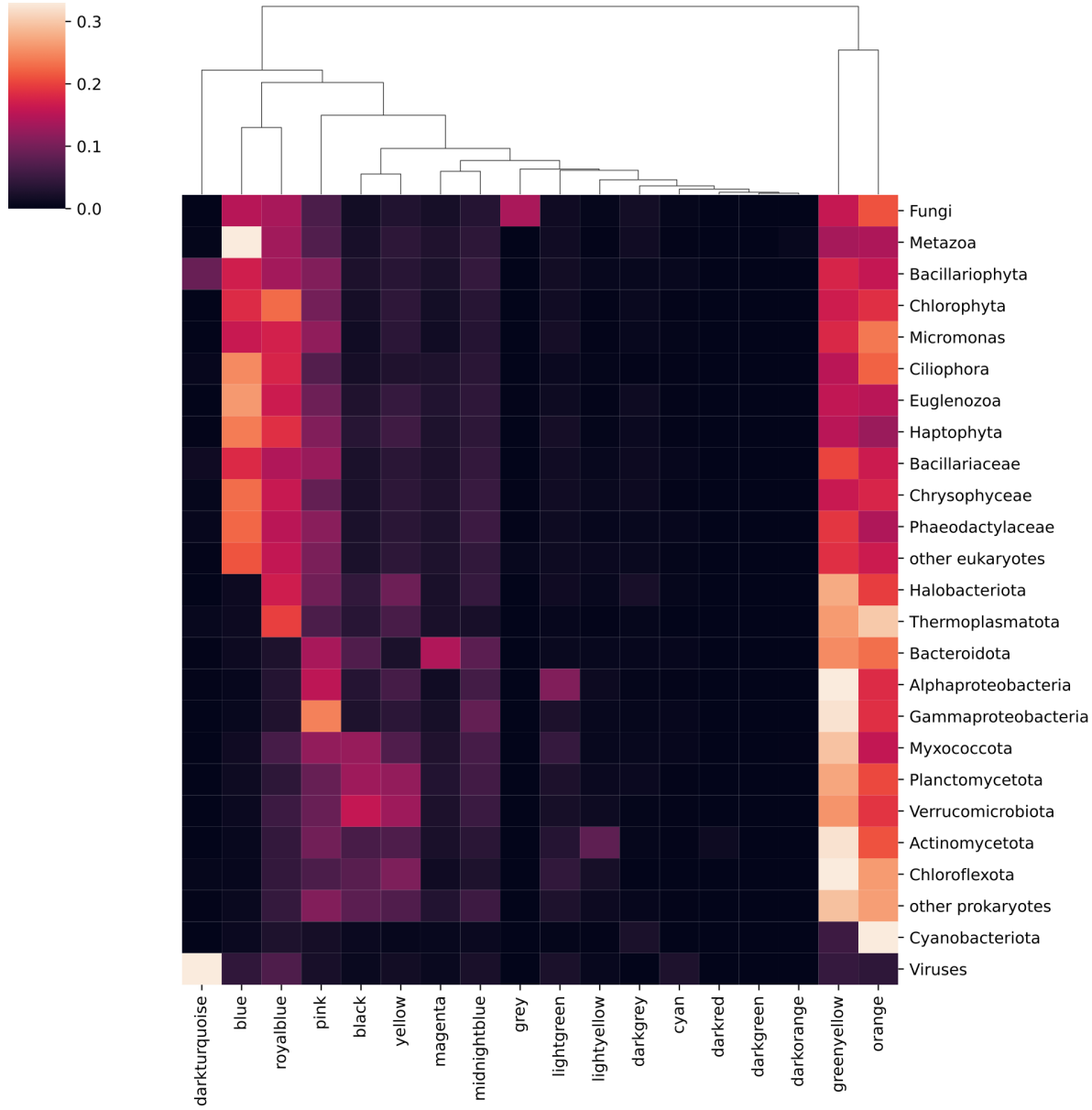


Figure C.2: Abundance of WGCNA modules in phyla. Intensity shows mean relative abundance of Pfams from that module, $\log_{10}$ scaled.

C.8 Data Availability

The following are a list of data sources which were used in Chapter 7. As this work is not yet published, not all are yet publicly availably, but, are saved in the following open access databases, and paths on the UEA cluster (relative paths are relative to /gpfs/data/-mock_lab/will/mosaic_tmp/mag_catalogue):

1. **Metagenome physical parameters**
    - Parameters such as nutrient concentrations, temperature, listed in Table 7.1.
    - ./sample_physical_parameters
    - Data was derived from the following sources: [405]–[409].
2. **Sample Metadata**
    - Parameters such as library preparation, PCR cycles, IMG/M pipeline version, assembler version, are available through the JGI web portal as per-sample accessory tables.
3. **Mag Catalogue**
    - A redundant set of MAGs, generated across all samples
    - A non-redundant set of MAGs, deduplicated at a 99% ANI level
    - Annotation files for each MAG
    - A spreadsheet containing statistics for each MAG such as sequencing depth, total number of bases, taxonomy, number of contigs, N50, etc.
    - /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/euk_dereplicated
    - /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/prok
    - /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/vir
    - /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue prok_drep_99 containing prokaryotes dereplicated with dRep at a 99% ANI level
    - /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/euk_drep_99 containing eukaryotes deduplicated with dRep at a 99% ANI level
    - /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/annotations
4. **Abundance Tables**

- Mappings from each sample to the set of MAGs, as tab separated tables. The units of these tables are either: number of bases, reads per kilobase million (RPKM), reads per million (RPM), or average coverage
- ./abundances/mag_abundances_X.tsv where X is one of the following: RPKM, RPM, number_of_bases, or abundance (which uses Strobealign's definition of abundance i.e. number of mapped bases divided by length of the contig).

5. **Network Data**

- Species networks generated from SCNIC (using SparCC correlations)
- /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/networks/scnic_v3/
- /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/networks/wgcna_genes_v1/

6. **Trees**

- Newick tree files and files used for IToL to generate tree. The prokaryotic trees used GTDB, newick files are at ./prok_gtdb/classify
- The eukaryotic tree used concatenated BUSCO genes, the newick file is at /gpfs/data/mock_lab/will/mosaic_tmp/mag_catalogue/trees

7. **Workflow Code**

- Snakemake workflow for the coassembly pipeline
- https://github.com/willboulton/mosaic-workflows
- Network analysis of species and genes
- https://github.com/willboulton/mosaic-communities-paper
- Binning visualisation method - this is part of the mosaic-workflows Github but also available through Github in VALENCE (below).
- https://github.com/willboulton/valence

genes

MDPI

Article

Genetic and Structural Diversity of Prokaryotic Ice-Binding Proteins from the Central Arctic Ocean

Johanna C. Winder [1,†], William Boulton [2,3,†], Asaf Salamov [3], Sarah Lena Eggers [4], Katja Metfies [4], Vincent Moulton [2] and Thomas Mock [1,*]

[1] School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
[2] School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
[3] DOE Joint Genome Institute, Algal and Fungal Program, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[4] Alfred Wegener Institute, Polar Biological Oceanography, Am Handelshafen 12, 27570 Bremerhaven, Germany
* Correspondence: t.mock@uea.ac.uk
† These authors contributed equally to this work.

Abstract: Ice-binding proteins (IBPs) are a group of ecologically and biotechnologically relevant enzymes produced by psychrophilic organisms. Although putative IBPs containing the domain of unknown function (DUF) 3494 have been identified in many taxa of polar microbes, our knowledge of their genetic and structural diversity in natural microbial communities is limited. Here, we used samples from sea ice and sea water collected in the central Arctic Ocean as part of the MOSAiC expedition for metagenome sequencing and the subsequent analyses of metagenome-assembled genomes (MAGs). By linking structurally diverse IBPs to particular environments and potential functions, we reveal that IBP sequences are enriched in interior ice, have diverse genomic contexts and cluster taxonomically. Their diverse protein structures may be a consequence of domain shuffling, leading to variable combinations of protein domains in IBPs and probably reflecting the functional versatility required to thrive in the extreme and variable environment of the central Arctic Ocean.

Keywords: metagenomics; MAGs; ice-binding proteins; DUF3494; domain shuffling; polar genomics; Arctic Ocean; MOSAiC expedition

Citation: Winder, J.C.; Boulton, W.; Salamov, A.; Eggers, S.L.; Metfies, K.; Moulton, V.; Mock, T. Genetic and Structural Diversity of Prokaryotic Ice-Binding Proteins from the Central Arctic Ocean. *Genes* **2023**, *14*, 363. https://doi.org/10.3390/genes14020363

Academic Editors: Joe Hoffman, Melody Clark and Svenja Heesch

Received: 14 December 2022
Revised: 24 January 2023
Accepted: 25 January 2023
Published: 30 January 2023

1. Introduction

Ice-binding proteins (IBPs) are a large group of cold-active enzymes found across all three domains of life, but little is known about their diversity in natural environments. Depending on their concentration, IBPs function in one of two dominant modes: thermal hysteresis (TH) or ice-recrystallisation inhibition (IRI) [1]. TH refers to freezing point depression, while IRI prevents the growth of larger, tissue-damaging ice crystals [2,3]. Which of these modes dominates is also thought to relate to their environmental function [1]. In prokaryotic and eukaryotic microbes, the majority of ice-binding proteins contain a ~200 amino acid domain of unknown function 3494 (DUF 3494) [4]. These DUF3494 IBPs (henceforth IBPs) are often found in psychrophilic bacteria [5], in part due to prevalent horizontal gene transfer (HGT) [4,6]. Our understanding of the function of prokaryotic IBPs is mainly derived from lab-based studies, but how widespread or representative these functions are remains unknown.

A number of bacterial IBPs have been functionally characterised, revealing varied potential environmental roles related to their structures. The Pfam library reports over 4000 IBP sequences from over 3000 taxa, the majority of which are prokaryotic [5]. Among them, 237 domain architectures are found [5]. Despite this diversity, studies of prokaryotic IBPs have largely focused on targeted, lab-based studies of single IBPs. The majority of characterised IBPs have a single domain architecture with an N-terminal signal peptide, implying secretion or membrane localisation [4,5]. Roles have been suggested for different

prokaryotic IBP structures—including the prevention of heterogeneous ice formation by the organism [7,8] and the maintenance of a liquid habitat by conserving triple junctions between ice grains [9,10]. IBPs from *Shewanella frigidimarina* and *Marinomonas primoryensis* contain bacterial immunoglobulin-like repeats which act as a tether between the cell membrane and the IBP, permitting an ice adhesion function [11,12]. Ig-like domains generally consist of two antiparallel β-sheets which twist to surround a hydrophobic core, and are often associated with bacterial adhesion to a variety of substrates [13]. In addition to these diverse functions of prokaryotic IBPs, a number of microbial eukaryotic IBPs have been functionally characterised, suggesting roles in brine channel shaping [14] and intracellular roles [15].

The diversity of IBPs may be connected to their environmental, taxonomic and specific genomic contexts. Prokaryotic IBPs are found in various frozen environments, including glacier cryoconites [16], subglacial lakes [9], polar desert soils [17] and sea ice [7]. However, the study of the natural diversity of IBPs in these environments is limited by their accessibility. This is especially true in winter, when IBPs may play an important role [18]. Metagenomics is a powerful tool to explore both the taxonomic and functional compositions of microbial communities, as well as to explore a specific group of sequences such as IBPs. Due to the limited accessibility of polar environments, only a few meta-omics studies have addressed IBPs directly or indirectly, but this has still provided insight into their ecology. These studies have revealed eukaryotic IBPs to be highly expressed in situ [19] and implicated prokaryotic IBPs in a commensal relationship with an Antarctic moss [20]. Some metagenomics studies have also referenced IBPs in passing, but without focusing further [17,21]. Metagenomics can yield the sequences of abundantly encoded IBPs, providing information about their taxonomic distribution and allowing the prediction of and comparison between their sequences and structures. This can then be used to suggest how diversity is generated within a taxon, as, for example, by domain shuffling. Gene synteny is also especially relevant in bacteria, where neighbouring genes are mostly co-transcribed in operons [22], and horizontally transferred genes can cluster in chromosomal "hotspots" [23]. Metagenome-assembled genomes (MAGs) can be used to address this topic, while avoiding the complications of culture-based approaches [24,25]. These have rarely been applied to polar contexts, despite the challenges of culturing organisms from these environments [26–29].

Here, we used metagenome-informed genomics to explore the genetic diversity and predicted structural diversity of prokaryotic DUF3494 IBPs from the central Arctic Ocean. During leg 2 of the MOSAiC expedition, 15 metagenomic samples were collected spanning the bathypelagic, mesopelagic, epipelagic, sea–ice interface and interior ice layers [30]. Unlike in many other studies, these samples were collected during polar winter. We explored the total community composition as well as the composition of IBP-encoding taxa within each environment, expecting that the metagenomes from the sea–ice interface and the interior ice would have a higher relative abundance of IBP genes compared to those from epipelagic or meso/bathypelagic environments. We characterised the diverse possible IBP domain architectures present in the samples, predicting the structures of the abundant architectures. MAGs were used to explore the genomic context of IBPs, determining which domain architectures were abundant in the genes upstream and downstream of the IBPs. We then compared the amino acid sequences of DUF3494s found in IBPs with abundant domain architectures to determine structural or taxonomic trends. Finally we compared the amino acid sequences of every DUF3494, exploring the distribution of domain architectures, signal peptide presence and transmembrane domain presence.

2. Materials and Methods

2.1. Sample Collection

Fifteen metagenome samples were collected during leg 2 of the MOSAiC expedition (collection dates between 13 January 2019 and 7 February 2020), during the Arctic winter (Figure 1). These samples were collected both from pelagic layers, with seawater collected via sampling from a CTD rosette, and from sea–ice layers. Ice samples were melted, and

50 mL of sterile filtered seatwater were added per 1 cm ice core. Samples were filtered with a Sterivex 0.22 micrometre filter, stored at −80 °C on board the Polarstern until the end of leg 2 (24 February 2020), and subsequently shipped to the Alfred Wegener Institute, at a temperature of −80 °C. The sample volumes used can be found in Supplementary Table S1. Two of the fifteen samples were created through pooling; i.e., the third in each trio of epipelagic samples was pooled from the other two (pelagic samples from the same CTD rosette). Together, these 15 metagenomic samples constituted the set of ECO-omics metagenome pilot samples. Of the 15 samples, 8 were from pelagic layers and the remaining 7 from sea-ice. Of the seawater samples, 4 were from the epipelagic, with 2 taken from a depth of 20 m and 2 from 50 m, and a further 2 samples were generated through pooling material from the other 2 replicates (see Supplementary Table S1 for details). Each pair was collected from the same CTD rosette. The remaining two seawater samples were from the meso and bathypelagic, sampled from depths of 200 and 4082 m, respectively. Of the seven sea-ice samples, five co-located samples, including the four samples labelled interior ice, were from different layers within the same ice core, from first-year ice. The remaining two samples were second-year ice from the sea–ice interface, the 0 to 5 cm bottom layer of the ice, at the interface with the ocean. Associated metadata are in Supplementary Tables S1 and S2, which also provide the IDs of the relevant GOLD databases.

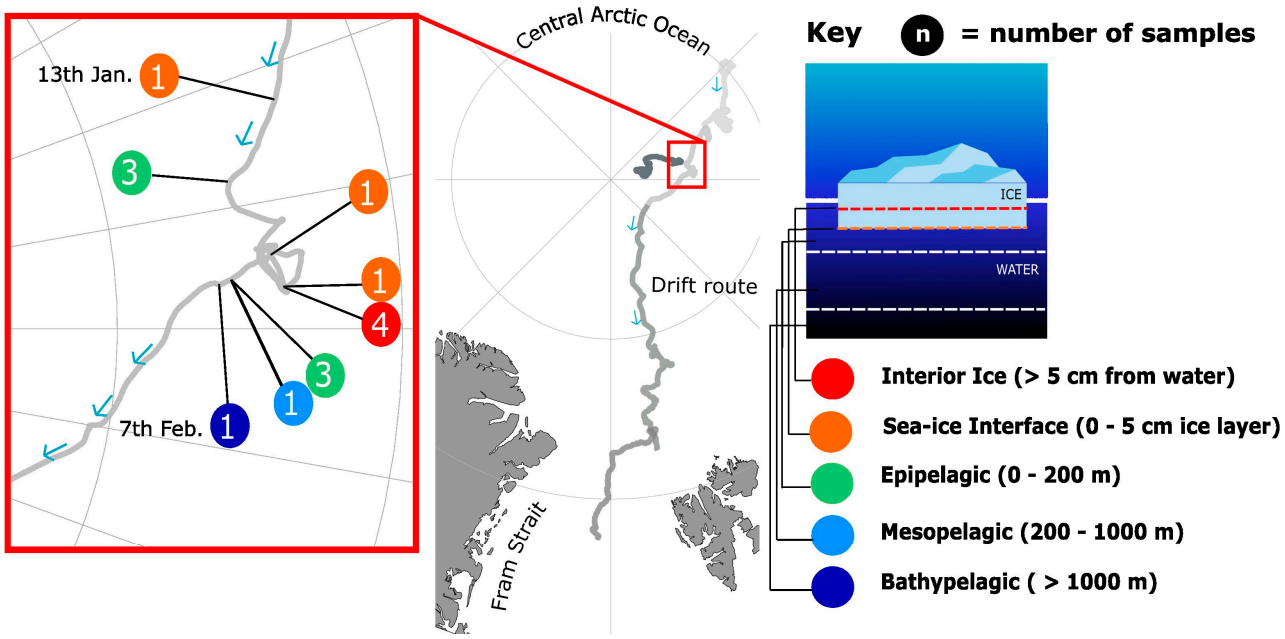


Figure 1. Drift route of the MOSAiC expedition, and description of the pilot samples. The red box shows the drift route of the RV Polarstern between the 13 January and the 7 February 2020. During this time, the 15 pilot samples were collected. Co-occurring samples (either from the same CTD rosette, or neighbouring ice cores) are shown, with the number of co-located samples from the same environment circled. The schematic diagram on the right describes the environment of each of the samples. Figure adapted under a CC BY 4.0 licence from [30].

2.2. DNA Extraction and Sequencing

The DNA was extracted at the Alfred Wegener Institute, using the Qiagen PowerWater DNA kit, following a slightly modified version of the QIAGEN DNeasy Power Water SOP v1 (QIAGEN N.V., Hilden, Germany) [31]. Samples were sent to the DoE Joint Genome Institute (JGI) for sequencing. Sequencing was performed following either the Illumina regular fragment, 300 base pair, or the Illumina low input, 300 base pair protocols (Supplementary Table S1), with the sea–ice interface and meso and bathypelagic samples following the low input protocol, and epipelagic and interior ice samples using the regular fragment protocol.

For the regular protocol, the DNA was sheared to 300 bp using the Covaris LE220-Plus and size selected with SPRI using TotalPure NGS beads (Omega Bio-tek, Norcross, GA, USA). The fragments were treated with end-repair, A-tailing and the ligation of Illumina compatible adapters (IDT, Inc, Gladesville, Australia) using the KAPA-HyperPrep kit (KAPA Biosystems, Wilmington, MA, USA). The prepared libraries were quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. The sequencing of the flowcell was performed with the Illumina NovaSeq sequencer using NovaSeq XP V1.5 reagent kits, S4 flowcell, following a 2 × 151 indexed run recipe. For the low input protocol (10 ng of DNA), the procedure was the same, except that the sample was enriched using 5 cycles of PCR.

Bioinformatics Processing of Samples

The sequence quality control, assembly, annotation and binning were all performed using the IMG/M metagenome annotation pipeline (v.5.0.23) [32]. Briefly, reads were trimmed of Illumina adapters and then filtered for quality and for human or lab contamination using BBDuk (v38.79) [33], and each sample was individually assembled using metaSPAdes (v3.14.1) [34]. Genes were predicted using consensus between GeneMark (v1.05) [35], INFERNAL (v1.1.3) [36], Prodigal (v2.6.3) [37] and tRNAscan-SE (v.2.0.7) [38]. Only contigs of lengths of at least 500 base pairs were retained, representing between 61.2% and 94.7% of the total reads of the samples (Supplementary Table S3). Annotations were performed using the hmmsearch function of HMMER (3.1b2) [39], using model specific cutoffs, and with models from a range of protein databases including Pfam-A (v30) [40]. Bins were generated using metaBAT2 (v2.12.1) [41], with a minimum contig size of 1000, and assessed for completeness and contamination with CheckM (v1.0.12) [42], and bins of less than 50% completeness or above 10% contamination were discarded. The bins were taxonomically assigned using GTDB-tk (v0.2.2) [43]. The subsequent analysis of the IBPs used all genes that were annotated with the PF11999 Pfam domain. The abundance of the PF11999 was measured using reads per kilobase million (RPKM). We used the Phobius web server [44] to further annotate transmembrane domains and signal peptides, and MMSeqs2 (v01889*) [45] to scan the assemblies against both the NR and MMETSP [46] databases for taxonomic annotation. Sequences classified as eukaryotic were removed for downstream analysis.

2.3. Community Analysis

We compared the prokaryotic community compositions of the total assembly, the MAGs, and their respective IBP-producing communities across sites. We used the R packages phyloseq (v1.40.0) and ggplot2 (v3.4.0) [47,48] to plot both the total prokaryotic community composition and that of the IBP-containing community. Vegan in R (v2.6-4) [49] was used to carry out the comparisons of community composition, using permANOVA and non-metric multidimensional scaling (NMDS) to visualise them.

We then explored which bacterial orders encoded IBPs with diverse gene architectures —defined as containing >1 domain in the IBP or containing a signal peptide and/or transmembrane domain(s).

2.4. Protein Structure Prediction

The domain architectures for modelling were identified by the presence of multiple protein families (Pfams) within the same gene. We selected the five most environmentally abundant (total reads per kilobase million; RPKM) domain architectures in the total dataset for modelling. Representative IBPs for each domain architecture were further selected on the basis of their environmental abundance. The structures were modelled using AlphaFold (v2.1.1) [50], with the models reported being the highest confidence models from the AlphaFold output. Functional information about the individual domains in these IBPs was obtained from the Interpro database [5]. A conceptual figure denoting typical domain architecture was produced using Inkscape (v1.2.1).

2.5. Upstream and Downstream Gene Analysis

The domain architecture of the genes surrounding the IBPs in MAGs was determined by querying the genes found the closest, upstream or downstream, to the IBP genes, and recording their relative locations and which protein families were present. We queried which domains and domain architectures were the most abundant within these upstream and downstream genes. Genomic context and domain architecture figures were produced using Inkscape v1.2.1. As above, broader functional characterisations were obtained using InterPro [5].

2.6. Phylogenetic Analysis of IBPs

To determine how the phylogenetic relationships between IBPs varied depending on the domain architecture, environment and taxonomic assignments, we produced gene trees of the most environmentally abundant gene architectures, as well as gene trees of IBPs across all domain architectures. The alignments of the amino acid sequences of HMMER hits to the DUF3494 domain were produced using muscle (v2.0.4) [51], and low quality columns of the alignment were removed using TrimAl (v1.2) [52]. The trees were generated with FastTree (v2.1.1) [53], using the default parameters, and visualised using interactive tree of life (IToL; v6.6) [54]. We repeated this method for IBPs within MAGs. Gene trees with fewer than 60 leaves, or with multi-copy DUF3494 domain architectures, were rooted at their midpoint. For the remaining trees, we rooted the trees using an outgroup of 130 IBPs from the dinoflagellate *Polarella glacialis* [55] (accessions in Supplementary Table S4).

3. Results

3.1. Diverse Prokaryotic Communities and MAGs Encode IBP Genes

From the whole metagenome assemblies, we retrieved between 4.91×10^7 and 2.50×10^8 bacterial reads per sample and between 1.59×10^5 and 9.00×10^6 archaeal reads per sample. Of all of the assemblies, 71% could be classified to the order level. From them, we identified 207 bacterial orders and 32 archaeal orders. The most commonly identified bacterial orders were Cellvibrionales (15.2%) and Rhodobacterales (13.3%). The most common archaeal orders were Nitrosopumilales (0.88%) and Candidatus Poseidoniales (0.46%). We also retrieved 750 total medium and high quality MAGs from these samples (Figure 2c,d).

The subset of these communities in which DUF3494-containing proteins (henceforth IBPs) were found was analysed separately. In all, 85.74% of IBPs could be assigned order-level taxonomy. These IBP-encoding communities were composed of 60 bacterial orders and 5 archaeal orders. The most common bacterial orders were Flavobacteriales (1936 IBPs; 50.79%) (Bacteroidetes) and Alteromonadales (893 IBPs; 23.43%) (Gammaproteobacteria) (Figure 2). The most common archaeal orders were Methanomicrobiales (0.36%; 14 IBPs) (Euryarchaeota) and Candidatus Poseidonales (0.11%; 4 IBPs) (Candidatus Thermoplasmatota). A total of 3581 (80.54%) IBPs were found in the interior ice, 797 (17.93%) in the sea–ice interface, 60 (1.35%) in the epipelagic zone and 8 (0.18%) in the meso/bathypelagic zones.

In all, 199 IBPs were encoded by 79 MAGs. The most IBP-encoding MAGs were obtained from the interior ice habitat (67/79 MAGs), followed by the sea–ice interface (8/79 MAGs), with three and one MAGs found in the epipelagic and meso/bathypelagic environments, respectively. The order level composition of the IBP-encoding communities varied among different sites (total assembly permANOVA: F = 6.83, $p = 0.001$, R2 = 0.651; MAGs (genus level): F = 4.81, $p = 0.002$, R2 = 0.74) (Supplementary Figure S1).

3.2. Diverse IBP Structures Are Abundant in the Natural Environment

Diverse domain architectures were predicted from the genomic sequences of the IBPs. A total of 116 unique domain architectures were found in 3869 prokaryotic IBPs spanning 65 identified orders. These diverse architectures included a total of 46 protein families. Single domain IBPs were by far the most abundant in the environment (61.54% of the total environmental relative abundance, RPKM) and the most prevalent across the samples (accounting for 70.53% of the total number of IBPs), followed by double domain IBPs

(20.51% of the RPKM; 11.75% of the total number of IBPs) (Figure 3c; Table 1). Triple domain IBPs were also abundant (1.15%; 0.44%) (Figure 3d and Table 1).

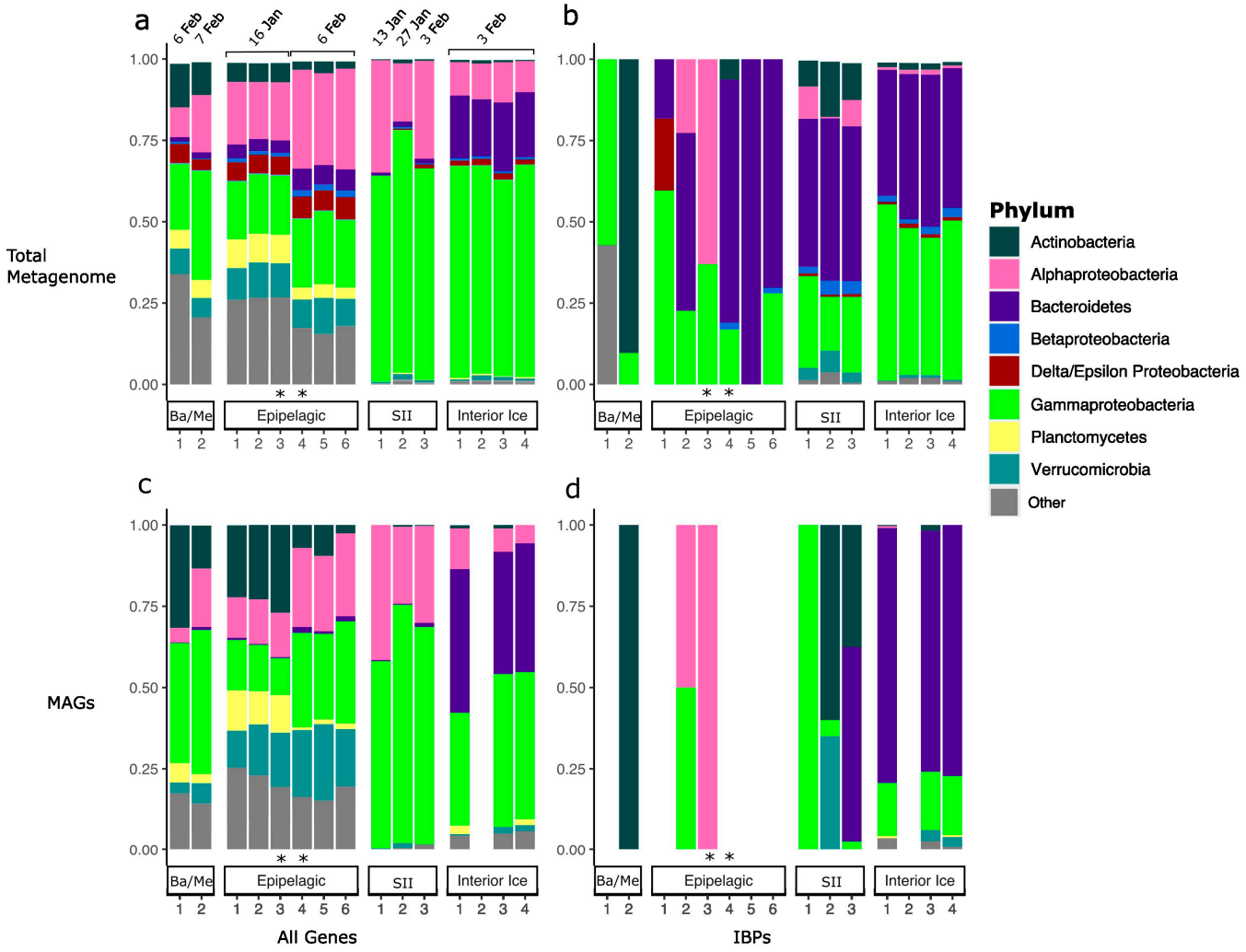


Figure 2. Total and ice-binding protein-encoding prokaryotic community composition and MAG distribution vary with environment type. Phylum-level composition (proportion of reads of prokaryotic assembly) of (**a**) prokaryotic whole communities and (**b**) prokaryotic taxa encoding at least one ice-binding protein (IBP). In the total assembly (**a**,**b**), whole communities were dominated by Gamma and Alpha- proteobacteria (pink and light green) across environments, with this becoming especially striking in the sea–ice interface and interior ice environments. Bacteroidetes (purple), Verrucomicrobia (teal) and Actinobacteria (dark green) were also variably dominant across environments. IBP-encoding communities were dominated by Actinobacteria in the meso/bathypelagic environment, and by Bacteriodetes and Gammaproteobacteria in all other environments. The taxonomic composition of all prokaryotic MAGs (**c**) and prokaryotic MAGs encoding at least one IBP (**d**) retrieved from these samples broadly mirrors the distribution of the total assembly communities. Ba/Me and SII refer to samples from the bathy/mesopelagic layers, and sea–ice interface (5 cm ice core bottom layer), respectively. The category 'other' includes all phyla with relative abundance less than 2.5%. Asterisks (*) represent samples formed by pooling. Sampling dates for all samples are indicated in panel (**a**). Note that IBP-encoding MAGs were not retrieved from each sampling location.

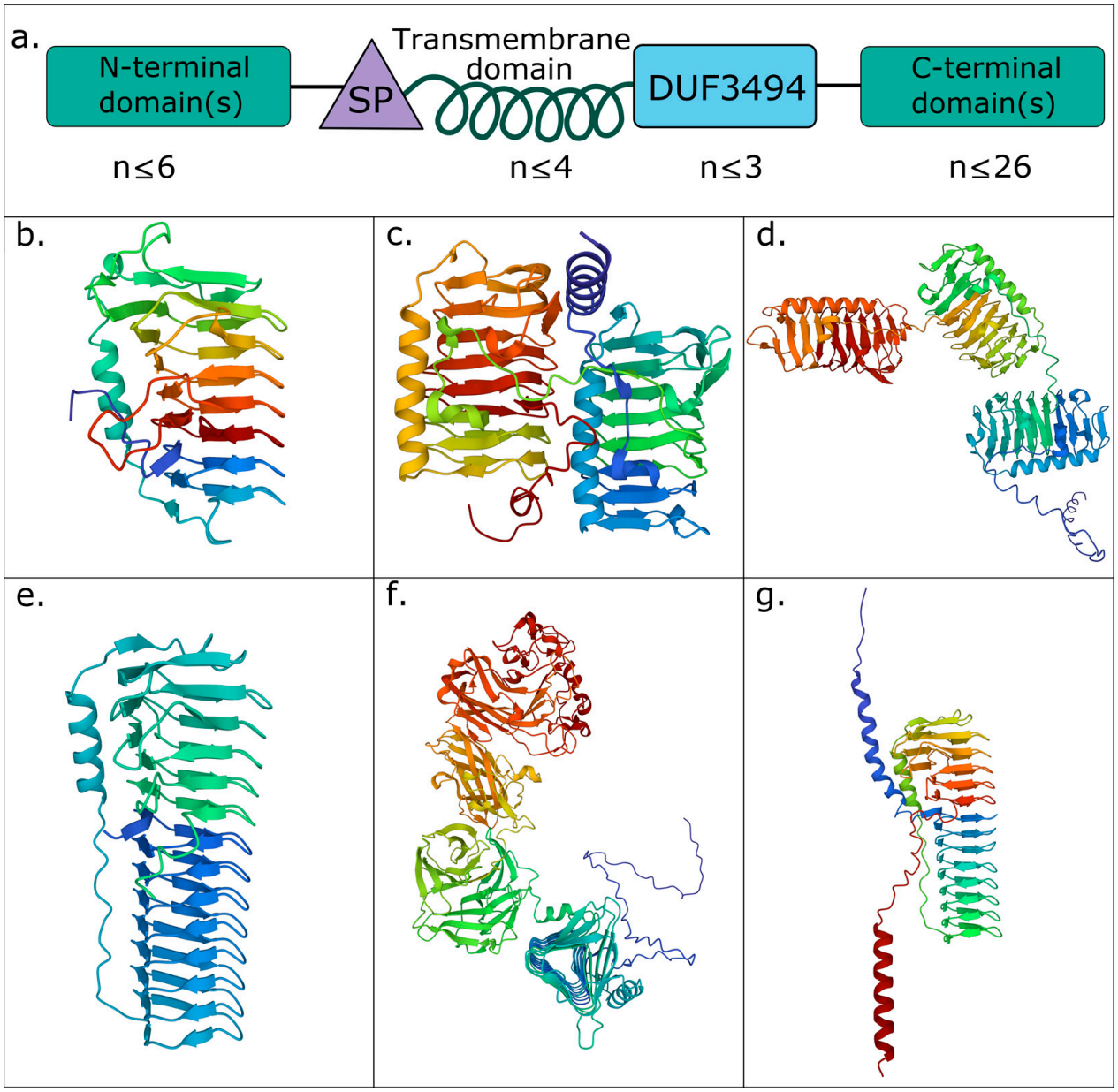


Figure 3. The structures and domain architectures of abundant prokaryotic ice-binding proteins reflect potentially diverse biological roles. (**a**) Concept diagram displaying the modular diversity of IBP domain architectures. IBPs minimally consist of a single DUF3494 domain (blue) but can consist of up to three DUF3494 domains. Where multiple DUF3494 domains are found, they are not necessarily found in immediate succession—other domains may be interspersed among them. Signal peptides (SP; yellow) and transmembrane domains (TMDs; green helix) are variably present, with up to four transmembrane domains being found in a protein. Additional N and C-terminal domains (pink; location defined by the position of the first DUF3494 domain) are also found, with up to 6 N-terminal domains or up to 26 C-terminal domains in a single protein (up to 28 domains in a single protein). The most abundant IBP domain architectures found in the total assembly were (**b**) single domain IBPs; (**c**) double domain IBPs; (**d**) triple domain IBPs; (**e**,**f**) single domain IBPs with an additional C-terminal DUF4842 (pfam16130); (**g**) single domain IBPs with an additional C-terminal PEP C-term motif (pfam07589). The most environmentally abundant representative of each domain architecture was selected for modelling. Note the length of the DUF3494 domain in (**e**–**g**). Proteins are coloured according to residue position, blue being the N-terminus and red being the C-terminus of the protein. Further information about domain architecture abundances is found in Table 1.

Table 1. Abundant IBP domain architectures from the total assembly. We grouped IBPs from the total assembly by their domain architecture and summed the abundance (reads per kilobase million; RPKM) for each IBP with that architecture. We then collected their protein family names (Pfam) from the Interpro database [5]. Using information provided by Interpro, we organised each Pfam into a broader functional grouping. Note that the abundances were summed across all environments, comprising two samples from the bathy/mesopelagic zone, six samples from the epipelagic, three from the sea–ice interface and four from the interior ice.

Domain Architecture	Protein Family (Pfam)	Broader Function	Abundance (RPKM)	Abundance (%)
pfam11999	DUF3494	IBP	4983.46	61.54
pfam11999_pfam11999	DUF3494	IBP	1660.68	20.51
pfam11999_pfam16130	DUF4842 *	β-barrel Ig fold	413.49	5.11
pfam11999_pfam07589	PEP C-term motif	Sorting/ Exopolysaccharides	209.68	2.59
pfam11999_pfam11999_pfam11999	DUF3494	IBP	93.11	1.15
pfam11999_pfam11999_pfam01345	DUF11	Cell wall-related	63.74	0.79
pfam11999_pfam02010	REJ domain	Membrane associated	58.44	0.72
pfam04519_pfam11999	Polymer-forming cytoskeletal	Cytoskeleton	47.68	0.59
pfam11999_pfam11999_pfam13517_pfam13517_pfam07593	FG-GAP-like repeat	Cell adhesion	44.35	0.55
	ASPIC and UnbV	Cell adhesion		
pfam11999_pfam11999_pfam13517_pfam13517_pfam13517_pfam07593	FG-GAP-like repeat	Cell adhesion	25.93	0.32
	ASPIC and UnbV	Cell adhesion		0.31
pfam11999_pfam03797	Autotransporter β-domain	Secretion	24.83	0.30
pfam13205_pfam13205_pfam11999	BIg-like domain *	Tethering	24.45	0.26
pfam11999_pfam11999_pfam02412_pfam02412_ pfam02412_pfam02412_pfam02412	Thrombospondin type 3 repeat	Cell adhesion	21.36	0.25
pfam11999_pfam01391	Collagen triple helix repeat	Cell adhesion	20.53	0.24
pfam11999_pfam07603	DUF1566	Unknown	19.42	0.23
pfam11999_pfam02494_pfam02494_pfam02494	HYR domain	Cell adhesion	18.99	0.23
pfam14341_pfam11999	PilX N-terminal	Cell adhesion	18.87	0.21
pfam04862_pfam11999	DUF642	Unknown (thought to be exclusive to plants)	16.73	0.20
pfam11999_pfam04862	DUF642	Unknown (thought to be exclusive to plants)	13.87	0.17
pfam11999_pfam01345	DUF11	Cell wall-related	13.17	0.16

* These domains are considered "Ig-like".

Some differences in the structure of the DUF3494 domain were observed. All modelled proteins contained the discontinuous right-handed β-solenoid with three flat faces and a braced α helix. However, the length of the β-solenoid varied, with longer solenoids containing 14 coils found in some of the most environmentally abundant IBPs (Figure 3e–g).

In the natural environment, IBPs containing a protein family classified as immunoglobulin-like make up a large proportion of the environmental relative abundance (467.69 RPKM; 6.62%), but they were not as prevalent across samples, accounting for only 3.57% of all of the IBPs found. They were, therefore, likely found in highly abundant individual IBPs rather than in a large number of distinct IBPs with the same architectures. Of these proteins, 15.22% contained a transmembrane domain, 45.65% contained a signal peptide and 7.25% contained both. In all, 84.06% of these IBPs came from interior ice, 14.49% from the sea–ice interface and 1.45% from the epipelagic zone.

The most prevalent domain architectures across the samples consisted of protein families whose role involves cell adhesion and exopolysaccharides. IBPs with these domain architectures had an abundance of 411.20 RPKM (5.82% of the environmental relative abundance), constituting 4.88% of all IBPs found. This is reflected in certain architectures with large numbers of repeated domains. The most striking examples of them among our samples are double domain IBPs with up to 26 C-terminal thrombospondin type-3 repeats, and single and double domain IBPs containing up to 7 C- or N-terminal bacterial immunoglobulin-like (BIg) domains. A total of 33.93% of IBPs with an adhesion function contained a TMD, while 48.68% contained a signal peptide and 32.8% contained both. As for their origins, 84.66% of these IBPs came from interior ice, 12.70% from the sea–ice interface and 2.65% from the epipelagic zone.

Other broadly abundant functions or protein families found in these domain architectures include protein families with no known or suggested functions (461.2 RPKM/6.53% abundance; 3.15% prevalence across samples), calcium binding proteins (54.88 RPKM/0.78%; 1.16%) and trafficking/secretion-related proteins (81.25 RPKM/1.15%; 0.70%).

In the MAGs, the most abundant domain architectures were single domain (55.32%; 56.28%), double domain (23.57%; 20.60%), single domain with a DUF4842 (7.86%; 4.52%), triple domain (1.64%; 1.51%) and single domain with a PEP C-term motif (1.58%; 3.02%).

3.3. The Genomic Context of IBPs Suggests Mechanisms for Generating Diversity

MAGs were used to explore the protein families present in the genes flanking IBP genes. Forty-six of the seventy-nine MAGs contained >1 IBP. The highest number of IBPs in a single MAG was nine (e.g., Figure 4b). In MAGs with multiple IBPs, these IBPs were frequently found in the same contig, immediately upstream or downstream of one another. Furthermore, these IBPs often had identical domain architectures, e.g., double domains (Figure 4a,d).

One hundred and three unique domain architectures were found downstream of IBPs (Figure 4). The most frequent domain architectures found downstream of IBPs in MAGs were single domain IBPs (6.12%) and double domain IBPs (4.76%). Following this, the five most abundant downstream domain architectures contained small solute membrane transport proteins (MFS; pfam07690; 3.40%), bacterial 2-component systems containing a DNA binding domain and a response regulator receiver domain (pfam04397_pfam00072; 2.72%), an antioxidant enzyme (AhpC/TSA family; pfam00578; 2.04%), DNA topoisomerase (pfam01131_pfam01751; 2.04%) and a phosphodiesterase (pfam01663; 2.04%).

The most frequent individual domains found downstream of IBPs, rather than whole architectures, were generally characterised by IBPs and repeats. They are as follows: IBPs (pfam11999; 10.82%), FG-GAP repeats (pfam14312; 3.03%), response regulator receiver domains (pfam00072; 2.60%), DNA binding domains (pfam04397; 2.60%) and bacterial transferase hexapeptides (pfam00132; 2.16%).

One hundred and seven unique domain architectures were found upstream of IBPs (Figure 4). The most frequent domain architectures found upstream of IBPs in MAGs were single domain IBPs (5.19%) and double domain IBPs (5.19%). Following this, the five most abundant upstream domain architectures were transposases (IS66 family; pfam03050; 2.60%), aminotransferases (pfam00155; 1.95%), autoregulatory aminotransferases (pfam00155_pfam00392; 1.95%), post-translational sulfatase modification domains (SUMF1; pfam03781; 1.95%), and DUF3050 (pfam11251; 1.95%).

The most frequent individual domains found upstream of IBPs were IBPs (pfam11999; 10.83%), a metal-binding motif (pfam11617; 8.33%), a β-propellor repeat (pfam07676; 4.17%), a WG repeat motif (pfam14903; 3.33%) and an aminotransferase (pfam00155; 2.50%).

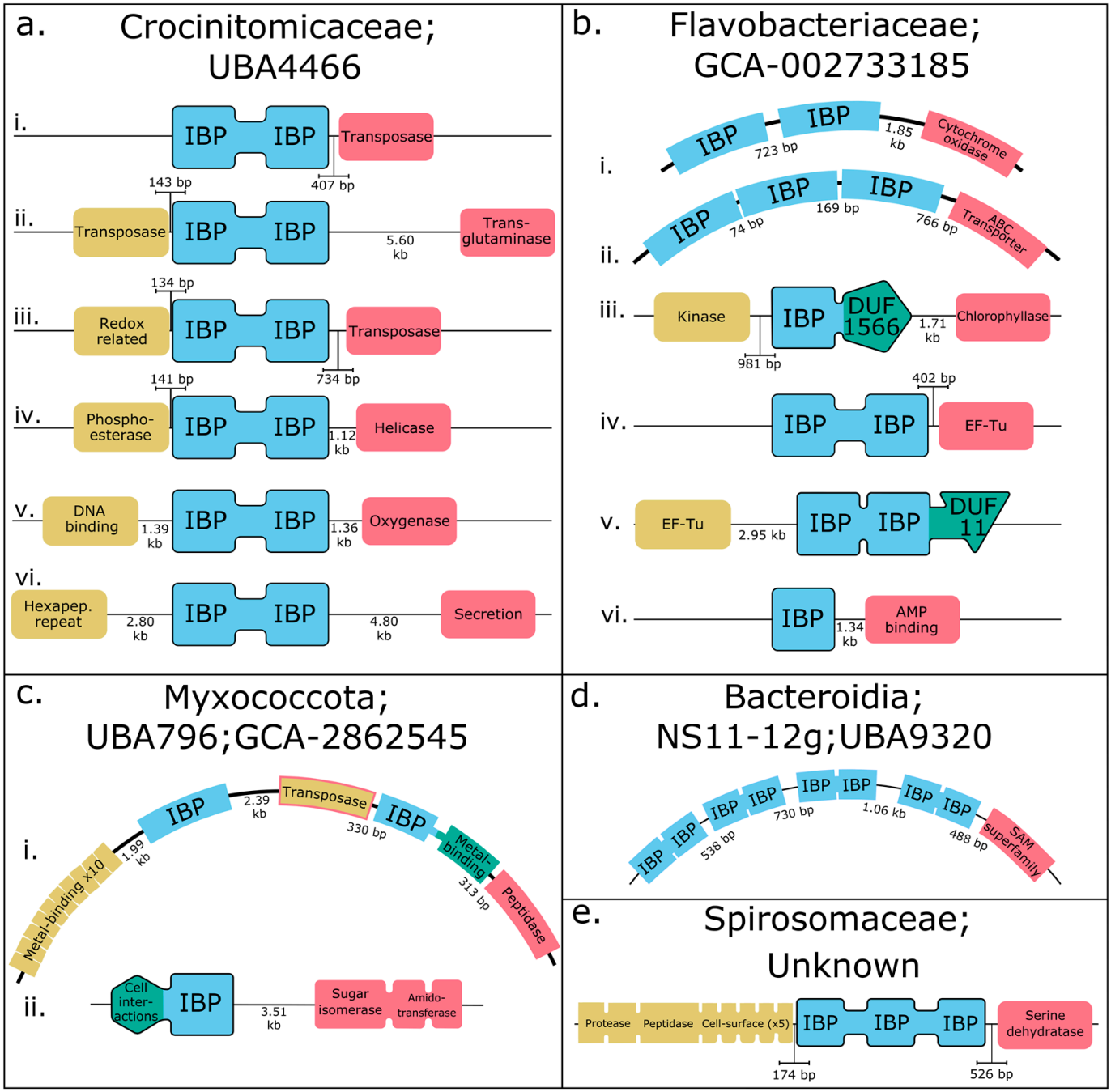


Figure 4. Genomic context of ice-binding proteins (IBPs) in metagenome-assembled genomes from the central Arctic Ocean. Genes upstream of IBPs are in yellow, IBPs are in blue, and genes downstream of IBPs are in pink. Double and triple domain IBPs are linked by blue connectors. Non-IBP domains within the IBP genes are in green. Each line is a single contig. Numbers between domains refer to the genomic distance between them. If an upstream or downstream domain is not shown, the IBP gene is the first or last gene in its contig. Contigs not listed in a specific order. (**a**) All six IBPs from Crocinitomicaceae (family) UBA4466 are double domain IBPs (ddIBPs). (i) ddIBP with a transposase immediately (407 bp) downstream. (ii) ddIBP with a transposase immediately (143 bp) upstream and a transglutaminase 5.60 kb downstream. (iii) ddIBP with a redox-related domain immediately (134 bp) upstream and a transposase nearby (734 bp) downstream. (iv) ddIBP with a phosphoesterase immediately (141 bp) upstream and a helicase 1.12 kb downstream. (v) ddIBP with a DNA binding domain 1.39 kb upstream and an oxygenase domain 1.36 kb downstream. (vi) ddIBP with hexapeptide repeats 2.80 kb upstream and a secretion-related domain 4.80 kb downstream. (**b**) The nine IBPs from the Flavobacteriaceae (family) GCA-002733185 are found in a variety of genomic contexts. (i) Two single

domain (sd) IBPs are found adjacent to each other at the start of a contig, separated by 723 bp. A cytochrome oxidase domain is found 1.85 kb downstream of the second sdIBP. (ii) Three sdIBPs are found adjacent to each other at the start of a contig, separated by 74 and 169 bp, respectively. An ABC transporter domain is found 766 downstream. (iii) A sdIBP containing a DUF1566 domain has a kinase 981 bp upstream and a chlorophyllase 1.71 kb downstream. (iv) A ddIBP is found at the start of a contig, with an elongation factor Tu domain 402 bp downstream. (v) A ddIBP containing a DUF11 domain is found at the end of a contig, with an elongation factor Tu domain 2.95 kb upstream. (vi) A sdIBP is found at the start of a contig with an AMP-binding domain 1.34 kb downstream. (**c**) The three IBPs from Myxococcota (family); UBA796; GCA-2862545. (i) Two sdIBPs are found in the same contig, separated by a transposase. The second sdIBP also contains a metal-binding domain. A protein containing 10 metal-binding domains is found 1.99 kb upstream of the first sdIBP. A transposase is 2.39 kb downstream of the first sdIBP and 330 bp upstream of the second sdIBP. The second sdIBP is followed by a peptidase 313 bp downstream. (ii) A sdIBP with a cell-interaction related domain is found at the start of a contig, with a protein containing sugar isomerase and amidotransferase domains found 3.51 kb downstream. (**d**) All four IBPs from Bacteroidia; NS11-12g; UBA9320 are ddIBPs and are found adjacent to each other at the start of the same contig, separated by 538 bp, 730 bp and 1.06 kb, respectively, and a SAM superfamily domain is found 488 bp downstream of the fourth ddIBP. (**e**) The only IBP from Spirosomaceae (order unknown) is a triple domain IBP. A protein containing a protease domain, a peptidase domain and five cell-surface related domains is found 174 bp upstream of the IBP and a serine dehydratase domain is found 526 bp downstream. Data used to produce this figure are in Supplementary Table S5.

3.4. Phylogenetic Distribution of Abundant Domain Architectures Implicates Domain Shuffling

The sequences of the DUF3494 domain(s) of IBPs with the most abundant domain architectures were compared (Figure 5 and Table 1). A number of these most abundant domain architectures did not appear to be present in a wide variety of individual IBPs—rather, individual IBPs with these architectures were highly abundant.

In all, 2886 single domain IBPs were found. Of these, just 39.92% contained a signal peptide, and 90.6% contained no transmembrane domain, while 8.34% contained one TMD, 0.90% contained two TMDs and 0.07% contained three TMDs. A total of 2.94% contained both an SP and at least one TMD. Of the 2886 total IBPs, 2501 could be classified to the order level. Among them, the five most abundant orders encoding this domain architecture were Flavobacteriales (1130 IBPs), Alteromonadales (714 IBPs), Burkholderiales (112 IBPs), Oceanospirillales (71 IBPs) and Acidimicrobiales (47 IBPs). As for their origin, 78.55% of these IBPs came from interior ice, 19.82% from the sea–ice interface, 1.42% from the epipelagic zone and 0.021% from the meso/bathypelagic zones.

A total of 455 double domain IBPs were found (Figure 5a), 57.80% of which contained a signal peptide, implying secretion. Meanwhile, 96.04% of them contained no transmembrane domain (TMD), while 3.74% contained one TMD and 0.22% contained two TMDs. Only 0.44% contained both a TMD and an SP, while 38.68% contained neither. In all, 335 IBPs with this domain architecture could be classified to the order level. The five most abundant orders were Flavobacteriales (278 IBPs), Cytophagales (10 IBPs), Alteromonadales (8 IBPs), Acidimicrobiales (6 IBPs), Burkholderiales, Cellvibrionales, Solirubrobacterales, Streptomycetales and Thiotrichales (3 each). As for their origins, 85.06% of these IBPs came from interior ice, 14.07% from the sea–ice interface and 0.88% from the epipelagic zone.

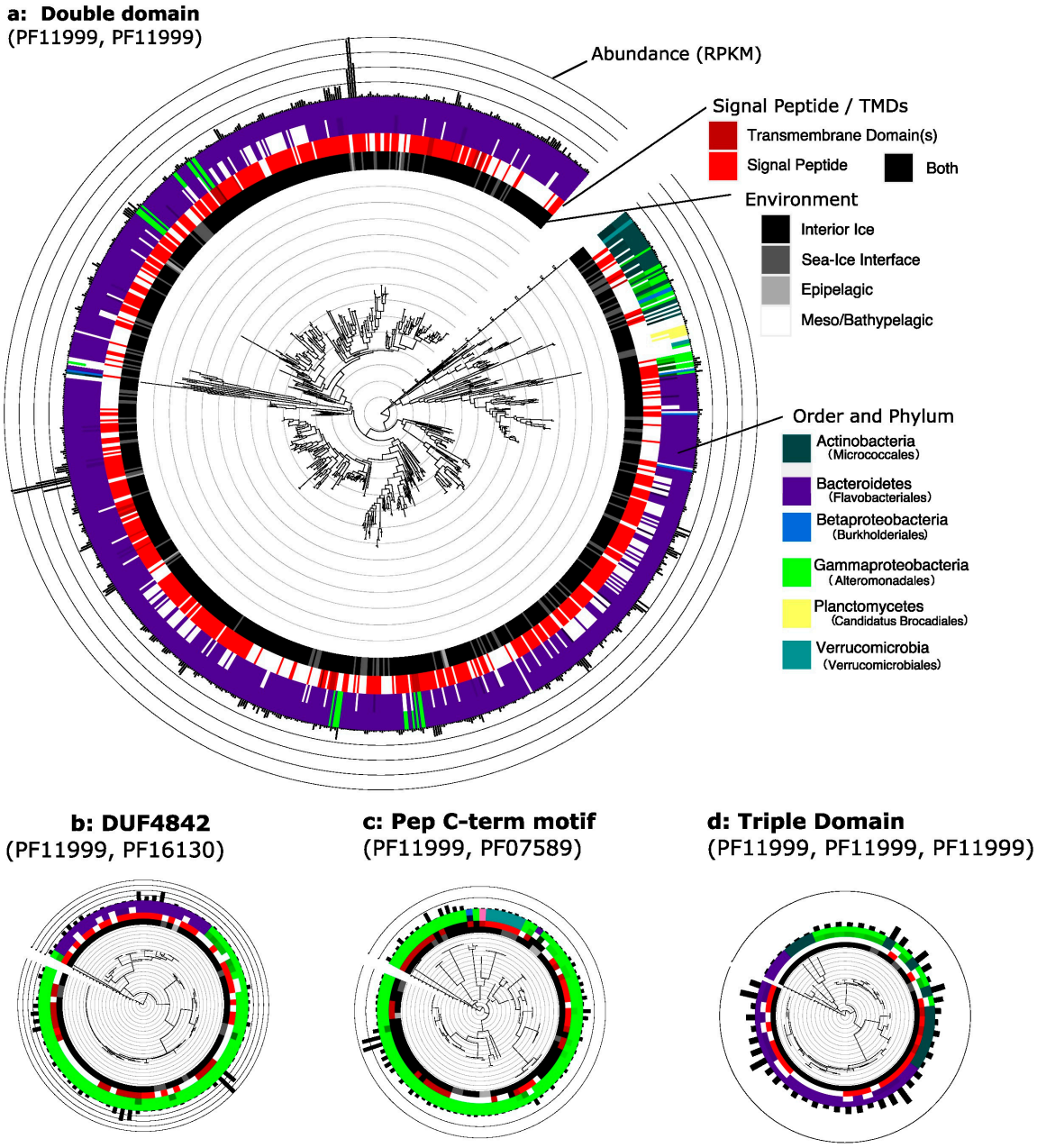


Figure 5. Abundant ice-binding protein domain architectures have distinct phylogenetic distributions. Gene trees of the DUF3494 domains present in the four environmentally most abundant domain architectures excluding single domain ice-binding proteins (IBPs). Trees are annotated with (from centre out) environment (black: interior ice, dark grey: sea–ice interface, light grey: epipelagic, white: meso/bathypelagic), signal peptide (SP: bright red) or transmembrane domain (TMD: dark red) presence (both: black), order-level and phylum-level classification (Bacteroidetes: purple, Gammaproteobacteria: bright green, Verrucomicrobia: teal, Actinobacteria: forest green, Betaproteobacteria: dark blue) and abundance (reads per kilobase million, demarcated in multiples of 10). Orders are coloured in shades of their parent phylum colour to show diversity within a phylum; the dominant order is shown specified in brackets in the legend and uses the same shade as the parent phylum. White gaps signify where the order was unknown. (**a**) Double domain IBPs (ddIBPs) mainly come from a single order of Bacteroidetes (Flavobacteriales), with the presence of SP and TMDs not appearing to be associated with taxonomy. (**b**) IBPs containing a C-terminal DUF4842 (pfam16130) come from Gammaproteobacteria and a single order of Bacteroidetes. TMDs are abundant only in one of the two clades of Bacteroidetes IBPs, and IBP abundance is distributed across both phyla. (**c**) IBPs containing

a PEP C-term motif mainly come from Gammaproteobacteria and Verrucomicrobia, with one from each of Alphaproteobacteria, Betaproteobacteria and Bacteroidetes. The majority of these IBPs contain an SP and/or a TMD. (**d**) Triple domain IBPs (tdIBPs) come from Bacteroidetes, Gammaproteobacteria and Actinobacteria. Most tdIBPs from Bacteroidetes cluster within a single clade, which is most similar to a monophyletic clade of Actinobacteria. TdIBPs from Gammaproteobacteria are found in a clade which also contains three Actinobacteria tdIBPs. Although the majority of tdIBPs are from Bacteroidetes and are found in a monophyletic clade, the tdIBPs which are the most different from this group also contain tdIBPs from Bacteroidetes.

In all, 86 single domain IBPs contained a DUF4842 (pfam16130) the function of which is unknown, but which contains a β-barrel immunoglobulin fold (Figure 5b). Of that total, 37.21% contained a signal peptide, while the majority lacked one, suggesting that many of these proteins may be intracellular. Similarly, only 19.77% contained a transmembrane domain, and none contained both a signal peptide and transmembrane domain, while 43.02% contained neither. Of the 75 IBPs with this domain architecture which could be classified to the order level, 51 were found within the Alteromonadales, 17 within the Flavobacteriales, 6 within the Cellvibrionales and 1 within the Vibrionales. In total, 88.37% of these IBPs came from interior ice, 10.47% from the sea–ice interface and 1.16% from the epipelagic zone.

A total of 92 single domain IBPs contained a PEP C-term motif (pfam07589) that is exopolysaccharide-related (Figure 5c). Among them, 53.26% contained a signal peptide, implying secretion, and 61.96% contained one transmembrane domain (TMD), while 3.26% contained two TMDs. In all, 34.78% contained both a transmembrane domain and a signal peptide, while 16.30% contained neither. Of the 89 IBPs with this domain architecture which could be classified to the order level, 70 were found within Alteromonadales, 8 within Oceanospirillales, 6 within Verrucomicrobiales and 2 in Methylococcales, and the remaining 3 were found in Ferrovales, Rhodobacterales and Thiotrichales, respectively. As for their origins, 83.70% of these IBPs came from interior ice, 10.87% from the sea–ice interface and 5.43% from the epipelagic zone.

In all, 23 double domain IBPs contained a DUF11 (pfam01345), whose function is unknown but is thought to be cell-wall related. Of them, 73.91% contained a signal peptide, implying that the majority were secreted. Only 8.70% contained a transmembrane domain, and none contained both an SP and a TMD; 17.39% contained neither. Of the 22 IBPs with this domain architecture which could be classified to the order level, 20 were found within the Flavobacteriales and 2 were found within the Saprospirales. A total of 82.61% of these IBPs came from interior ice, and 17.39% from the sea–ice interface.

Seventeen triple domain IBPs were found (Figure 5d). Of these, 41.18% contained a signal peptide. Only 5.88% contained a transmembrane domain, and none contained both an SP and a TMD; 52.94% contained neither. Of the 12 IBPs with this domain architecture which could be classified to the order level, 5 were found within Flavobacteriales, 4 within Micrococcales, 2 within Cytophagales and Cellvibrionales and 1 within Thiotricales. A total of 94.12% of these IBPs came from interior ice, and 5.88% from the sea–ice interface.

3.5. IBPs from the Total Assembly Cluster Taxonomically

A large amount of structural diversity was distributed across the tree of 3869 IBP sequences (Figure 6). In the total assembly, 43.47% of IBPs contained a signal peptide, while 9.10% contained one TMD, 0.85% contained two TMDs, and 0.05% contained three TMDs. Only 3.23% of the IBPs contained both an SP and at least one TMD, and 49.75% contained neither.

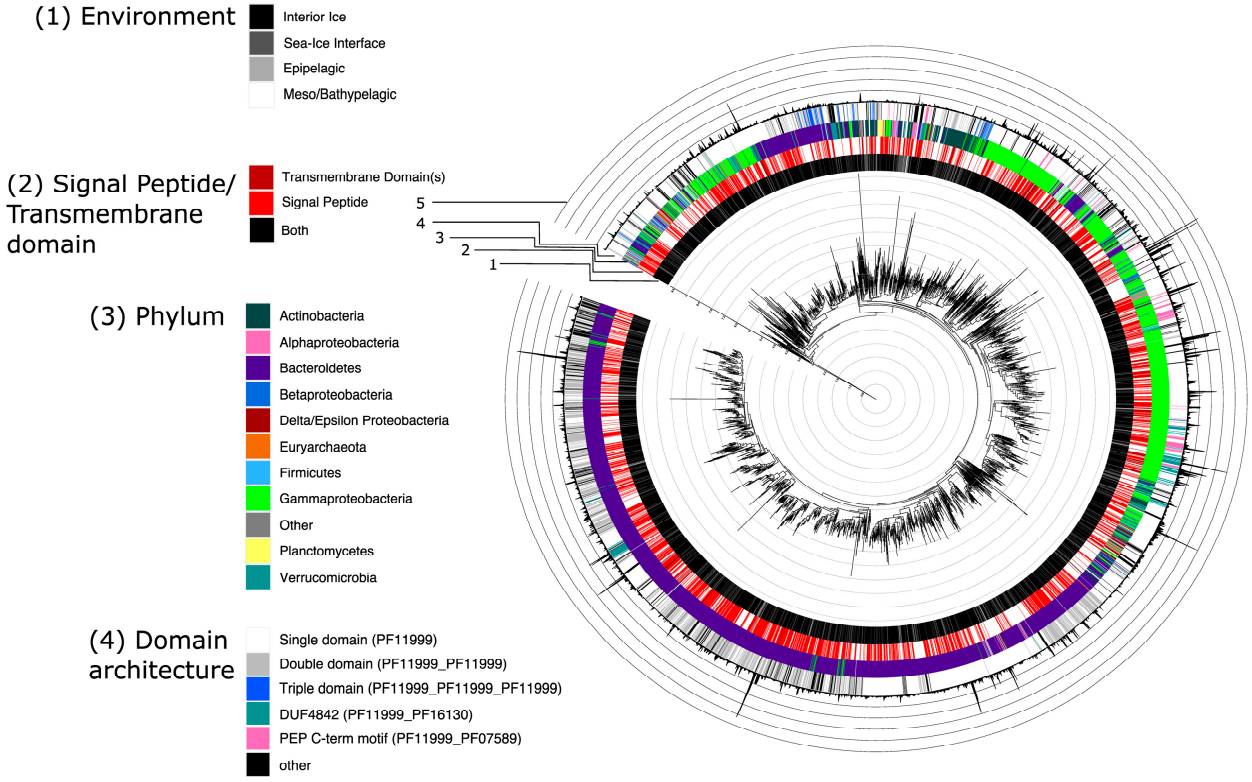


Figure 6. Structurally diverse prokaryotic IBPs are phylogenetically widely distributed. Trees of prokaryotic DUF3494 domains, either just found in MAGs (top), or in the total assembly. Trees are annotated with (from centre out) 1. Environment (black: interior ice, dark grey: sea–ice interface, light grey: epipelagic, white: meso/bathypelagic); 2. Signal peptide (SP: bright red) or transmembrane domain (TMD: dark red) presence (both: black); 3. Phylum-level classification (Bacteroidetes: purple, Gammaproteobacteria: bright green, Verrucomicrobia: teal, Actinobacteria: forest green, Betaproteobacteria: dark blue); 4. Domain architecture (single domain: white, double domain: grey, triple domain: blue, DUF4842: teal, PEP C-term motif: pink, other protein family: black); 5. Abundance (reads per kilobase million, demarcated in multiples of 10). Most IBPs are found in sea–ice environments. Signal peptide presence is evenly distributed across the tree. The majority of IBPs are found within Bacteroidetes or Gammaproteobacteria. Many IBPs from Bacteroidetes are more similar to each other than to IBPs from Gammaproteobacteria and other phyla. IBPs from Gammaproteobacteria cluster less closely, and are interspersed with IBPs from other phyla. Broadly, the more monophyletic, recently branched clade of the Bacteroidetes IBPs appears to be enriched with double domain IBPs and less individually abundant domain architectures, with one clustered group of DUF4842-containing IBPs also present. Conversely, the majority of IBPs with PEP C-term motifs are found within Gammaproteobacteria. Triple domain architectures are mainly found in IBPs from Gammaproteobacteria and the second-largest clade of Betaproteobacteria. Note that some of the most dissimilar IBPs come from the same phyla. IBP abundance appears to be relatively evenly distributed across phyla.

The most abundant orders in which signal-peptide-containing IBPs were found were Flavobacteriales (49.59%), Alteromonadales (29.42%), Oceanospirillales (3.14%), Burkholderiales (2.69%) and Cytophagales (1.72%). For IBPs without signal peptides, the most abundant orders were Flavobacteriales (47.94%), Alteromonadales (25.98%), Burkholderiales (4.45%), Oceanospirillales (2.57%), and Micrococales (2.14%).

IBPs with TMDs were mostly found in the order Alteromonadales (58.58%), followed by Flavobacteriales (16.18%), Oceanospirillales (3.88%), Thiotrichales (2.27%) and Acidimicrobiales (1.94%). Those without TMDs were mostly found in Flavobacteriales (51.81%),

followed by Alteromonadales (24.09%), Burkholderiales (3.94%), Oceanospirillales (2.69%) and Micrococcales (1.59%).

A total of 25.41% of IBPs had a diverse domain architecture, defined as a protein with more than one domain (IBP or other protein family) (Figure 6). Of these proteins with a diverse domain architecture, 80.26% could be classified to the order level. The five most abundant orders among them were Flavobacteriales (55.26%), Alteromonadales (21.67%), Oceanospirillales (2.41%), Cellvibrionales (2.41%) and Cytophagales (2.16%). Of the single domain IBPs (i.e., those without a diverse domain architecture), 83.75% could be classified to order level. The five most abundant orders among them were Flavobacteriales (46.13%), Alteromonadales (29.29%), Burkholderiales (4.39%), Oceanospirillales (2.94%) and Micrococcales (1.66%).

4. Discussion

Our findings suggest that the structural diversity of prokaryotic IBPs is associated with their taxonomy. By surveying the complement of ice-binding proteins encoded by prokaryotic sea ice and marine communities during an Arctic winter, we compared ecological and individual-scale observations. We queried environmentally abundant IBP domain architectures, linking these to broader functions as well as to genomic context and taxonomy. The IBPs were encoded by a diverse subset of communities and MAGs. IBPs containing immunoglobulin-like domains and domains involved in cell adhesion were abundant. The genomic context of the IBPs was dominated by other IBPs. The taxonomic clustering of the IBPs was sometimes also reflected in the variable presence of signal peptides and transmembrane domains. Together, these results provide new insight into the previously underexplored natural diversity of prokaryotic IBPs in the central Arctic Ocean. Furthermore, these results highlight the value of MAGs as a complement to whole metagenomes [56], especially in study regions lacking abundant reference genomes [29].

Ig-like domains and cell adhesion-related domains were the most abundant non-ice binding domains found in the IBPs. The presence of bacterial immunoglobulin (BIg) domains in IBPs has previously been attributed to an adhesin function. These domains act as a flexible tether between the IBP and the cell [11,12]. The ice-tethering function of IBPs is thought to hold bacterial cells in close proximity to the ice, where oxygen and nutrient conditions are favourable [11]. However, IBPs that play an ice-tethering role typically contain both a membrane anchor and a signal peptide [12]. The ice-binding domain extends out via a tether which is anchored in the cell membrane. Although this has been suggested as a dominant function of IBPs previously [4,12], our results provide minimal evidence for it, as only small proportions of IBPs containing Ig or Ig-like domains had both a signal peptide and a transmembrane domain. Conversely, specific adhesion-related domains (e.g., collagen triple helix repeat) were prevalent. These domains can contribute to the ability of biofilms to bind the extracellular matrix [57–60]. Sea ice harbours microbial biofilms embedded in extracellular polymeric substances [61]. These have been suggested to create microenvironments where nutrients accumulate [62]. Our results indicate a potential role for IBPs in anchoring biofilms to ice.

Single (sd) and double (dd) domain IBPs were by far the most abundant protein domain architectures; however, less than half of the sdIBPs contained a signal peptide. Although there are other pathways for secretion that are not SP-mediated [63], this suggests that a significant proportion of sdIBPs are intracellular. An intracellular IBP has been found in the plastid membrane of a sea-ice diatom [15]; however, its biological role is unknown. It is possible that these putatively intracellular IBPs function to limit the formation of intracellular ice, or play a role in the poorly-understood freezing perception [64,65]; however, this merits further exploration. Conversely, two-thirds of ddIBPs contained signal peptides, suggesting that these play an extracellular role. DUF3494 ddIBPs have been physico-chemically characterised and shown to not be inherently more active than sdIBPs [10]. Ice crystals in sea ice have variable plane orientations [66]. It is therefore possible that multidomain IBPs

are preferentially secreted in order to increase the likelihood of successful adsorption to diverse ice planes.

In the MAGs, IBPs were most frequently flanked by other IBPs. Gene synteny can be used as a method of inferring the biological function of proteins [67–69]. Genes which cluster in prokaryotic genomes may be part of the same operon. The genes within these hypothetical operons may be connected in various ways; most notably, they may be part of the same metabolic pathway, be part of a shared non-metabolic (e.g., regulatory) pathway, or physically interact [67]. By clustering closely in the genome, IBPs, which are thought to be regulated in response to external conditions, i.e., freezing [70], could be co-regulated. Given that many IBPs are secreted and therefore function in a comparatively vast environment, they may require large volumes of protein in order to adapt rapidly to the environment [3,71]. Encoding tandem IBPs may be a mechanism to allow the bulk production of these proteins [72]. Furthermore, non-DUF3494 IBPs sometimes form multimers which function more effectively than monomeric IBPs [73,74]. Given that this is a known feature of proteins which cluster in bacterial genes, it is possible that tandem IBPs result in multimeric protein formation.

IBPs clustered taxonomically when comparing abundant domain architectures (Figure 5). There were differences between the taxonomic distributions of double domain, DUF4843-containing, PEP C-term motif-containing and triple domain IBPs. In many cases, the IBP sequences clustered according to taxonomy. Bacteria obtain IBPs via horizontal gene transfer [75]. However, our results imply that, after this acquisition, the host organisms may utilise domain shuffling to adapt the IBPs for their specific habitat and lifestyle. This is further supported by the taxonomic patterns of TMD presence. For example, transmembrane domains are abundant in PEP C-term motif-containing IBPs from Gammaproteobacteria, but clearly absent from Verrucomicrobia IBPs. PEP C-term motifs are found in biofilms and are used for protein sorting through association with exopolysaccharides in Gram negative bacteria [76]. It has been proposed that the PEP C-term motifs are necessary for bacterial aggregate formation [77]—the presence of a TMD may, therefore, alter the way that these function.

IBPs also clustered taxonomically when all sequences, not just those with abundant domain architectures, were compared (Figure 6), with IBPs from Bacteroidetes forming two distinct groups. The less basal of these groups is enriched in double domain IBPs compared to other groups, and the majority of signal peptide-containing IBPs were found in Flavobacteria (Bacteroidetes). If signal peptide addition is linked to specific taxa, this would have consequences for the ecology of IBPs, as intracellular IBPs and secreted IBPs presumably have very different roles. The observation that IBPs form taxonomic groupings has implications for their evolution, and we suggest that sdIBPs may act as building blocks for their host organisms to duplicate and shuffle into diverse architectures and, subsequently, functions.

Supplementary Materials: The following supporting information can be downloaded at:https://www.mdpi.com/article/10.3390/genes14020363/s1, Figure S1: Non-metric multidimensional scaling of order-level community composition across different environments; Figure S2: PF11999 (DUF3494) domains are one of the most differentially abundant Pfams when comparing ice and water.; Figure S3: Trees of IBPs from MAGs and total assembly.; Table S1: Sample location, processing and sequencing data.; Table S2: Sample IDs (Label used in this paper, MOSAiC, GOLD, JGI, and IMG/M, IDs); Table S3: Assembly statistics; Table S4: List of *Polarella glacialis* IBP accessions from the NCBI Short Read Archive (SRA), BioProject accession PRJEB33539; Table S5: Genomic context of IBPs from selected MAGs.; Table S6: Abundance of samples and genes of interest and their taxonomic assignments.

Author Contributions: Conceptualization, J.C.W., W.B. and T.M.; methodology, J.C.W. and W.B.; software, W.B.; validation, W.B. and J.C.W.; formal analysis, J.C.W. and W.B.; investigation, J.C.W. and W.B.; resources, W.B., A.S. and S.L.E.; data curation, W.B.; writing—original draft preparation, J.C.W. and W.B.; writing—review and editing, J.C.W., W.B., T.M., V.M. and K.M.; visualisation, J.C.W. and W.B.; supervision, T.M. and V.M.; project administration, T.M. and K.M.; funding acquisition, T.M., V.M. and K.M. All authors have read and agreed to the published version of the manuscript.

Funding: W.B. was supported by the Natural Environment Research Council and ARIES DTP [grant number NE/S007334/1]. J.C.W. was supported by the UKRI Biotechnology and Biological Sciences Research Council Norwich Research Park Biosciences Doctoral Training Partnership [grant number BB/T008717/1]. The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Data used in this manuscript were produced as part of the international Multidisciplinary drifting Observatory for the Study of the Arctic Climate (MOSAiC) with the tag MOSAiC20192020 (expedition AWI_PS_122_00). This work was also supported through the Research Council of Norway through project HAVOC (grant no 280292), and the National Science Foundation through grant OPP-1735862.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Sequence data and metadata are available through the DOE JGI web portal under proposal ID 505419. Sample metagenome IDs are listed in Supplementary Table S2. All data used for this study are in Supplementary Table S6.

Acknowledgments: We thank the Alfred-Wegener-Institut Helmholtz-Zentrum für Polar-und Meeresforschung (AWI) for the operation of RV Polarstern (https://doi.org/10.17815/jlsrf-3-163, accessed on 14 November 2022). Furthermore, we acknowledge the MOSAiC team (https://doi.org/10.5281/zenodo.5179738, accessed on 20 November 2022). Among the MOSAiC team, we especially acknowledge those responsible for collecting and processing samples from Leg 2 of the MOSAiC expedition, including Allison Fong and Clara Hoppe (AWI). Further thanks go to the members of the HAVOC project for their contributions, especially to the ice coring, including Marc Oggier (University of Alaska), Mats Granskog (Norwegian Polar Institute), Lasse Olsen (Norwegian Polar Institute), Sinhue Torres-Valdes (AWI) and Dmitry Divine (University of Bergen).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gruneberg, A.K.; Graham, L.A.; Eves, R.; Agrawal, P.; Oleschuk, R.D.; Davies, P.L. Ice Recrystallization Inhibition Activity Varies with Ice-Binding Protein Type and Does Not Correlate with Thermal Hysteresis. *Cryobiology* **2021**, *99*, 28–39. [CrossRef]
2. Yeh, Y.; Feeney, R.E. Antifreeze Proteins: Structures and Mechanisms of Function. *Chem. Rev.* **1996**, *96*, 601–618. [CrossRef]
3. Yu, S.O.; Brown, A.; Middleton, A.J.; Tomczak, M.M.; Walker, V.K.; Davies, P.L. Ice Restructuring Inhibition Activities in Antifreeze Proteins with Distinct Differences in Thermal Hysteresis. *Cryobiology* **2010**, *61*, 327–334. [CrossRef]
4. Vance, T.D.R.; Bayer-Giraldi, M.; Davies, P.L.; Mangiagalli, M. Ice-Binding Proteins and the 'Domain of Unknown Function' 3494 Family. *FEBS J.* **2019**, *286*, 855–873. [CrossRef]
5. Blum, M.; Chang, H.-Y.; Chuguransky, S.; Grego, T.; Kandasaamy, S.; Mitchell, A.; Nuka, G.; Paysan-Lafosse, T.; Qureshi, M.; Raj, S.; et al. The InterPro Protein Families and Domains Database: 20 Years On. *Nucleic Acids Res.* **2021**, *49*, D344–D354. [CrossRef]
6. Raymond, J.A.; Janech, M.G.; Mangiagalli, M. Ice-Binding Proteins Associated with an Antarctic Cyanobacterium, Nostoc Sp. HG1. *Appl. Environ. Microbiol.* **2021**, *87*, e02499-20. [CrossRef]
7. Hanada, Y.; Nishimiya, Y.; Miura, A.; Tsuda, S.; Kondo, H. Hyperactive Antifreeze Protein from an Antarctic Sea Ice Bacterium Colwellia Sp. Has a Compound Ice-Binding Site without Repetitive Sequences. *FEBS J.* **2014**, *281*, 3576–3590. [CrossRef]
8. Raymond, J.A.; Fritsen, C.; Shen, K. An Ice-Binding Protein from an Antarctic Sea Ice Bacterium. *FEMS Microbiol. Ecol.* **2007**, *61*, 214–221. [CrossRef]
9. Raymond, J.A.; Christner, B.C.; Schuster, S.C. A Bacterial Ice-Binding Protein from the Vostok Ice Core. *Extremophiles* **2008**, *12*, 713–717. [CrossRef]
10. Wang, C.; Pakhomova, S.; Newcomer, M.E.; Christner, B.C.; Luo, B.-H. Structural Basis of Antifreeze Activity of a Bacterial Multi-Domain Antifreeze Protein. *PLoS ONE* **2017**, *12*, e0187169. [CrossRef]
11. Guo, S.; Stevens, C.A.; Vance, T.D.R.; Olijve, L.L.C.; Graham, L.A.; Campbell, R.L.; Yazdi, S.R.; Escobedo, C.; Bar-Dolev, M.; Yashunsky, V.; et al. Structure of a 1.5-MDa Adhesin That Binds Its Antarctic Bacterium to Diatoms and Ice. *Sci. Adv.* **2017**, *3*, e1701440. [CrossRef]

12. Vance, T.D.R.; Graham, L.A.; Davies, P.L. An Ice-Binding and Tandem Beta-Sandwich Domain-Containing Protein in Shewanella Frigidimarina Is a Potential New Type of Ice Adhesin. *FEBS J.* **2018**, *285*, 1511–1527. [CrossRef]
13. Chatterjee, S.; Basak, A.J.; Nair, A.V.; Duraivelan, K.; Samanta, D. Immunoglobulin-Fold Containing Bacterial Adhesins: Molecular and Structural Perspectives in Host Tissue Colonization and Infection. *FEMS Microbiol. Lett.* **2021**, *368*, fnaa220. [CrossRef]
14. Bayer-Giraldi, M.; Weikusat, I.; Besir, H.; Dieckmann, G. Characterization of an Antifreeze Protein from the Polar Diatom Fragilariopsis Cylindrus and Its Relevance in Sea Ice. *Cryobiology* **2011**, *63*, 210–219. [CrossRef]
15. Gwak, Y.; Jung, W.; Lee, Y.; Kim, J.S.; Kim, C.G.; Ju, J.-H.; Song, C.; Hyun, J.-K.; Jin, E. An Intracellular Antifreeze Protein from an Antarctic Microalga That Responds to Various Environmental Stresses. *FASEB J.* **2014**, *28*, 4924–4935. [CrossRef]
16. Singh, P.; Hanada, Y.; Singh, S.M.; Tsuda, S. Antifreeze Protein Activity in Arctic Cryoconite Bacteria. *FEMS Microbiol. Lett.* **2014**, *351*, 14–22. [CrossRef]
17. Goordial, J.; Davila, A.; Greer, C.W.; Cannam, R.; DiRuggiero, J.; McKay, C.P.; Whyte, L.G. Comparative Activity and Functional Ecology of Permafrost Soils and Lithic Niches in a Hyper-Arid Polar Desert. *Environ. Microbiol.* **2017**, *19*, 443–458. [CrossRef]
18. Krembs, C.; Eicken, H.; Junge, K.; Deming, J.W. High Concentrations of Exopolymeric Substances in Arctic Winter Sea Ice: Implications for the Polar Ocean Carbon Cycle and Cryoprotection of Diatoms. *Deep Sea Res. Part Oceanogr. Res. Pap.* **2002**, *49*, 2163–2181. [CrossRef]
19. Uhlig, C.; Kilpert, F.; Frickenhaus, S.; Kegel, J.U.; Krell, A.; Mock, T.; Valentin, K.; Beszteri, B. In Situ Expression of Eukaryotic Ice-Binding Proteins in Microbial Communities of Arctic and Antarctic Sea Ice. *ISME J.* **2015**, *9*, 2537–2540. [CrossRef]
20. Raymond, J.A. Dependence on Epiphytic Bacteria for Freezing Protection in an Antarctic Moss, Bryum Argenteum. *Environ. Microbiol. Rep.* **2016**, *8*, 14–19. [CrossRef]
21. Koo, H.; Hakim, J.A.; Bej, A.K. Metagenomic Analysis of Microbial Cold Stress Proteins in Polar Lacustrine Ecosystems. In *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2016; pp. 837–844. ISBN 978-1-119-00481-3.
22. Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G.D.; Maltsev, N. The Use of Gene Clusters to Infer Functional Coupling. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 2896–2901. [CrossRef] [PubMed]
23. Oliveira, P.H.; Touchon, M.; Cury, J.; Rocha, E.P.C. The Chromosomal Organization of Horizontal Gene Transfer in Bacteria. *Nat. Commun.* **2017**, *8*, 841. [CrossRef]
24. Parks, D.H.; Rinke, C.; Chuvochina, M.; Chaumeil, P.-A.; Woodcroft, B.J.; Evans, P.N.; Hugenholtz, P.; Tyson, G.W. Recovery of Nearly 8,000 Metagenome-Assembled Genomes Substantially Expands the Tree of Life. *Nat. Microbiol.* **2017**, *2*, 1533–1542. [CrossRef] [PubMed]
25. Tully, B.J.; Sachdeva, R.; Graham, E.D.; Heidelberg, J.F. 290 Metagenome-Assembled Genomes from the Mediterranean Sea: A Resource for Marine Microbiology. *PeerJ* **2017**, *5*, e3558. [CrossRef]
26. Tytgat, B.; Verleyen, E.; Obbels, D.; Peeters, K.; Wever, A.D.; D'hondt, S.; Meyer, T.D.; Criekinge, W.V.; Vyverman, W.; Willems, A. Bacterial Diversity Assessment in Antarctic Terrestrial and Aquatic Microbial Mats: A Comparison between Bidirectional Pyrosequencing and Cultivation. *PLoS ONE* **2014**, *9*, e97564. [CrossRef] [PubMed]
27. Duncan, A.; Barry, K.; Daum, C.; Eloe-Fadrosh, E.; Roux, S.; Schmidt, K.; Tringe, S.G.; Valentin, K.U.; Varghese, N.; Salamov, A.; et al. Metagenome-Assembled Genomes of Phytoplankton Microbiomes from the Arctic and Atlantic Oceans. *Microbiome* **2022**, *10*, 67. [CrossRef] [PubMed]
28. Cao, S.; Zhang, W.; Ding, W.; Wang, M.; Fan, S.; Yang, B.; Mcminn, A.; Wang, M.; Xie, B.; Qin, Q.-L.; et al. Structure and Function of the Arctic and Antarctic Marine Microbiota as Revealed by Metagenomics. *Microbiome* **2020**, *8*, 47. [CrossRef]
29. Royo-Llonch, M.; Sánchez, P.; Ruiz-González, C.; Salazar, G.; Pedrós-Alió, C.; Sebastián, M.; Labadie, K.; Paoli, L.; Ibarbalz, F.M.; Zinger, L.; et al. Compendium of 530 Metagenome-Assembled Bacterial and Archaeal Genomes from the Polar Arctic Ocean. *Nat. Microbiol.* **2021**, *6*, 1561–1574. [CrossRef] [PubMed]
30. Mock, T.; Boulton, W.; Balmonte, J.-P.; Barry, K.; Bertilsson, S.; Bowman, J.; Buck, M.; Bratbak, G.; Chamberlain, E.J.; Cunliffe, M.; et al. Multiomics in the Central Arctic Ocean for Benchmarking Biodiversity Change. *PLoS Biol.* **2022**, *20*, e3001835. [CrossRef]
31. Ottesen, A.; Kocurek, B. QIAGEN DNeasy Power Water SOP. Available online: https://www.protocols.io/view/qiagen-dneasy-power-water-sop-bztap6ie (accessed on 6 December 2022).
32. Clum, A.; Huntemann, M.; Bushnell, B.; Foster, B.; Foster, B.; Roux, S.; Hajek, P.P.; Varghese, N.; Mukherjee, S.; Reddy, T.B.K.; et al. DOE JGI Metagenome Workflow. *mSystems* **2021**, *6*, e00804-20. [CrossRef] [PubMed]
33. Bushnell, B. BBMap: A Fast, Accurate, Splice-Aware Aligner. 2014. Available online: https://www.osti.gov/biblio/1241166 (accessed on 7 December 2022).
34. Nurk, S.; Meleshko, D.; Korobeynikov, A.; Pevzner, P.A. MetaSPAdes: A New Versatile Metagenomic Assembler. *Genome Res.* **2017**, *27*, 824–834. [CrossRef]
35. Lukashin, A.V.; Borodovsky, M. GeneMark.Hmm: New Solutions for Gene Finding. *Nucleic Acids Res.* **1998**, *26*, 1107–1115. [CrossRef]
36. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-Fold Faster RNA Homology Searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef]
37. Hyatt, D.; Chen, G.-L.; LoCascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef]
38. Chan, P.P.; Lowe, T.M. TRNAscan-SE: Searching for TRNA Genes in Genomic Sequences. In *Gene Prediction: Methods and Protocols*; Kollmar, M., Ed.; Methods in Molecular Biology; Springer: New York, NY, USA, 2019; pp. 1–14. ISBN 978-1-4939-9173-0.

39. Mistry, J.; Finn, R.D.; Eddy, S.R.; Bateman, A.; Punta, M. Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions. *Nucleic Acids Res.* **2013**, *41*, e121. [CrossRef]
40. Finn, R.D.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Mistry, J.; Mitchell, A.L.; Potter, S.C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; et al. The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Res.* **2016**, *44*, D279–D285. [CrossRef]
41. Kang, D.D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies. *PeerJ* **2019**, *7*, e7359. [CrossRef]
42. Parks, D.H.; Imelfort, M.; Skennerton, C.T.; Hugenholtz, P.; Tyson, G.W. CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Res.* **2015**, *25*, 1043–1055. [CrossRef]
43. Chaumeil, P.-A.; Mussig, A.J.; Hugenholtz, P.; Parks, D.H. GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database. *Bioinformatics* **2020**, *36*, 1925–1927. [CrossRef]
44. Käll, L.; Krogh, A.; Sonnhammer, E.L.L. Advantages of Combined Transmembrane Topology and Signal Peptide Prediction—The Phobius Web Server. *Nucleic Acids Res.* **2007**, *35*, W429–W432. [CrossRef]
45. Steinegger, M.; Söding, J. Mmseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef]
46. Keeling, P.J.; Burki, F.; Wilcox, H.M.; Allam, B.; Allen, E.E.; Amaral-Zettler, L.A.; Armbrust, E.V.; Archibald, J.M.; Bharti, A.K.; Bell, C.J.; et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* **2014**, *12*, e1001889. [CrossRef]
47. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016. ISBN 978-3-319-24277-4.
48. McMurdie, P.J.; Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **2013**, *8*, e61217. [CrossRef]
49. 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. Vegan: Community Ecology Package. R Package Version 2.5–7 2020. 2022. Available online: https://cran.r-project.org/package=vegan (accessed on 9 December 2022).
50. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
51. Edgar, R.C. MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC Bioinform.* **2004**, *5*, 113. [CrossRef] [PubMed]
52. Capella-Gutiérrez, S.; Silla-Martínez, J.M.; Gabaldón, T. TrimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses. *Bioinformatics* **2009**, *25*, 1972–1973. [CrossRef] [PubMed]
53. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **2010**, *5*, e9490. [CrossRef]
54. Letunic, I.; Bork, P. Interactive Tree Of Life (ITOL) v5: An Online Tool for Phylogenetic Tree Display and Annotation. *Nucleic Acids Res.* **2021**, *49*, W293–W296. [CrossRef] [PubMed]
55. Stephens, T.G.; González-Pech, R.A.; Cheng, Y.; Mohamed, A.R.; Burt, D.W.; Bhattacharya, D.; Ragan, M.A.; Chan, C.X. Genomes of the Dinoflagellate Polarella Glacialis Encode Tandemly Repeated Single-Exon Genes with Adaptive Functions. *BMC Biol.* **2020**, *18*, 56. [CrossRef]
56. Grossart, H.-P.; Massana, R.; McMahon, K.D.; Walsh, D.A. Linking Metagenomics to Aquatic Microbial Ecology and Biogeochemical Cycles. *Limnol. Oceanogr.* **2020**, *65*, S2–S20. [CrossRef]
57. Foster, T.J.; Geoghegan, J.A.; Ganesh, V.K.; Höök, M. Adhesion, Invasion and Evasion: The Many Functions of the Surface Proteins of Staphylococcus Aureus. *Nat. Rev. Microbiol.* **2014**, *12*, 49–62. [CrossRef] [PubMed]
58. Kawashima, T.; Kawashima, S.; Tanaka, C.; Murai, M.; Yoneda, M.; Putnam, N.H.; Rokhsar, D.S.; Kanehisa, M.; Satoh, N.; Wada, H. Domain Shuffling and the Evolution of Vertebrates. *Genome Res.* **2009**, *19*, 1393–1403. [CrossRef]
59. Loftus, J.C.; Smith, J.W.; Ginsberg, M.H. Integrin-Mediated Cell Adhesion: The Extracellular Face. *J. Biol. Chem.* **1994**, *269*, 25235–25238. [CrossRef]
60. Lukomski, S.; Bachert, B.A.; Squeglia, F.; Berisio, R. Collagen-like Proteins of Pathogenic Streptococci. *Mol. Microbiol.* **2017**, *103*, 919–930. [CrossRef] [PubMed]
61. Underwood, G.J.C.; Fietz, S.; Papadimitriou, S.; Thomas, D.N.; Dieckmann, G.S. Distribution and Composition of Dissolved Extracellular Polymeric Substances (EPS) in Antarctic Sea Ice. *Mar. Ecol. Prog. Ser.* **2010**, *404*, 1–19. [CrossRef]
62. Roukaerts, A.; Deman, F.; Van der Linden, F.; Carnat, G.; Bratkic, A.; Moreau, S.; Lannuzel, D.; Dehairs, F.; Delille, B.; Tison, J.-L.; et al. The Biogeochemical Role of a Microbial Biofilm in Sea Ice: Antarctic Landfast Sea Ice as a Case Study. *Elem. Sci. Anthr.* **2021**, *9*, 00134. [CrossRef]
63. Kuchler, K.; Thorner, J. Membrane Translocation of Proteins without Hydrophobic Signal Peptides. *Curr. Opin. Cell Biol.* **1990**, *2*, 617–624. [CrossRef]
64. Krembs, C.; Deming, J.W. The Role of Exopolymers in Microbial Adaptation to Sea Ice. In *Psychrophiles: From Biodiversity to Biotechnology*; Margesin, R., Schinner, F., Marx, J.-C., Gerday, C., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; pp. 247–264. ISBN 978-3-540-74335-4.
65. Lamers, J.; van der Meer, T.; Testerink, C. How Plants Sense and Respond to Stressful Environments1 [OPEN]. *Plant Physiol.* **2020**, *182*, 1624–1635. [CrossRef]

66. Weeks, W.F.; Gow, A.J. Preferred Crystal Orientations in the Fast Ice along the Margins of the Arctic Ocean. *J. Geophys. Res. Oceans* **1978**, *83*, 5105–5121. [CrossRef]
67. Huynen, M.; Snel, B.; Lathe, W.; Bork, P. Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Res.* **2000**, *10*, 1204–1210. [CrossRef]
68. Suyama, M.; Bork, P. Evolution of Prokaryotic Gene Order: Genome Rearrangements in Closely Related Species. *Trends Genet.* **2001**, *17*, 10–13. [CrossRef] [PubMed]
69. Yelton, A.P.; Thomas, B.C.; Simmons, S.L.; Wilmes, P.; Zemla, A.; Thelen, M.P.; Justice, N.; Banfield, J.F. A Semi-Quantitative, Synteny-Based Method to Improve Functional Predictions for Hypothetical and Poorly Annotated Bacterial and Archaeal Genes. *PLoS Comput. Biol.* **2011**, *7*, e1002230. [CrossRef] [PubMed]
70. Jung, W.; Campbell, R.L.; Gwak, Y.; Kim, J.I.; Davies, P.L.; Jin, E. New Cysteine-Rich Ice-Binding Protein Secreted from Antarctic Microalga, *Chloromonas* Sp. *PLoS ONE* **2016**, *11*, e0154056. [CrossRef] [PubMed]
71. Jin, Y.; DeVries, A.L. Antifreeze Glycoprotein Levels in Antarctic Notothenioid Fishes Inhabiting Different Thermal Environments and the Effect of Warm Acclimation. *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **2006**, *144*, 290–300. [CrossRef] [PubMed]
72. Scott, G.K.; Hayes, P.H.; Fletcher, G.L.; Davies, P.L. Wolffish Antifreeze Protein Genes Are Primarily Organized as Tandem Repeats That Each Contain Two Genes in Inverted Orientation. *Mol. Cell. Biol.* **1988**, *8*, 3670–3675. [CrossRef]
73. Mahatabuddin, S.; Hanada, Y.; Nishimiya, Y.; Miura, A.; Kondo, H.; Davies, P.L.; Tsuda, S. Concentration-Dependent Oligomerization of an Alpha-Helical Antifreeze Polypeptide Makes It Hyperactive. *Sci. Rep.* **2017**, *7*, 42501. [CrossRef]
74. Sun, T.; Lin, F.-H.; Campbell, R.L.; Allingham, J.S.; Davies, P.L. An Antifreeze Protein Folds with an Interior Network of More Than 400 Semi-Clathrate Waters. *Science* **2014**, *343*, 795–798. [CrossRef] [PubMed]
75. Sorhannus, U. Evolution of Antifreeze Protein Genes in the Diatom Genus Fragilariopsis: Evidence for Horizontal Gene Transfer, Gene Duplication and Episodic Diversifying Selection. *Evol. Bioinform.* **2011**, *7*, EBO.S8321. [CrossRef]
76. Haft, D.H.; Paulsen, I.T.; Ward, N.; Selengut, J.D. Exopolysaccharide-Associated Protein Sorting in Environmental Organisms: The PEP-CTERM/EpsH System. Application of a Novel Phylogenetic Profiling Heuristic. *BMC Biol.* **2006**, *4*, 29. [CrossRef] [PubMed]
77. Gao, N.; Xia, M.; Dai, J.; Yu, D.; An, W.; Li, S.; Liu, S.; He, P.; Zhang, L.; Wu, Z.; et al. Both Widespread PEP-CTERM Proteins and Exopolysaccharides Are Required for Floc Formation of Zoogloea Resiniphila and Other Activated Sludge Bacteria. *Environ. Microbiol.* **2018**, *20*, 1677–1692. [CrossRef] [PubMed]