Computational approaches for advancing our understanding of marine microbes

Anthony Duncan

School of Computing Sciences
University of East Anglia
Student Number 3041557

This dissertation is submitted for the degree of
Doctor of Philosophy

May 2023

I would like to dedicate this thesis to my partner Ruth, my two cats Halo & Marley, and my parents Barrie & Theresa.

Acknowledgements

I would like to first thank my supervisory team whose support, direction and advice made this project possible, Professors Vincent Moulton and Thomas Mock at UEA, Dr Richard Leggett at the Earlham Institute, and Clara Manno at the British Antarctic Survey. The coronavirus pandemic fell during the middle two years of my PhD, and though we could only meet remotely guidance from Professor Moulton helped greatly in keeping the project focussed during a stressful time. Additionally it allowed everyone to meet my cats, who joined several meetings, and were less beneficial to maintaining focus.

Funding for my research came from Next Generation Unmanned Systems Science (NEXUSS) who I appreciate backing this project despite it lacking the signature presence of marine autonomous platforms of many of their projects. My route to computational biology has been particularly indirect, starting as an undergraduate in American Literature, and my change of heading is in no small part due to the enthusiastic support of my Masters project supervisor Dr. Katharina Huber.

I am greatly indebted and thankful to those who collected and sequenced the data analysed in this thesis, the Arctic and Atlantic metagenomes having been collected Dr. Katrin Schmidt, Dr. Willem van de Poll and Dr. Klaas Timmermans and sequenced by the Joint Genome Institute (JGI). The contribution of colleagues at JGI has been invaluable: performing initial assembly and annotation, contribution of three additional genomes for our analyses by Dr. Asaf Salamov, and in making some of our results available on their algal genomics resource PhycoCosm. This valued support enabled our focus to be on recovery of eukaryote genomes, and subsequent analysis.

The public availability of metagenome data from different environments was a great benefit in this research, and I am thankful to those groups who produced and published the datasets we utilised: the Human Microbiome Project (HMP), Tara Oceans Foundation, and the data published by Tee et al [1].

Abstract

Ocean microbes are essential for marine life, forming the base of the ocean food web and contributing to biogeochemical cycling of essential nutrients. The advent of modern molecular genetics techniques has revealed a large degree of diversity among these microbial communities, and metagenomic sequencing allows insight into the total metabolic potential and phylogeny of its constituent organisms. In this thesis, we develop two new computational approaches to analysing metagenomic sequencing data in order to advance our understanding of marine microbes.

Many marine microbes cannot be grown under lab conditions, but methods to obtain metagenome assembled genomes (MAGs) from metagenomic data have been widely applied for prokaryotes. However, many of the most abundant and environmentally significant microbes are eukaryotic, for which few MAGs have been recovered. To address this gap, we designed and implemented a pipeline for automated recovery of eukaryotic MAGs. From 12 samples, we obtained 21 MAGs from lineages including diatoms and prasinophytes. Our analysis of these eukaryotes, alongside prokaryotes from the same samples, showed a demarcation between polar and non-polar communities. The highest quality MAG has been included in algal genomics resource PhycoCosm as *Micromonas* sp. AD1.

We also want to understand the functional capability of the whole microbial community, as well as individual organisms. Functions are known to be shared between organisms and pathways, so we developed an unsupervised machine learning approach using the Non-Negative Matrix Factorisation (NMF) decomposition method to identify modules of functions which reflect this expected sharing of functions. Interpreting the resulting decomposition is important for exploratory analysis, and we developed the Leave-One-Out Correlation Decrease (LOOCD) method for this task with good performance identifying shared functions. Our methods successfully recover modules in simulated sequencing data and in real world cases studies, both identifying established groups (e.g. surface and mesopelagic ocean) and having meaningful biological interpretation.

Table of contents

List of figures

List of tables

Abbreviations

AAI Average Amino Acid Identity. 70, 87, 89, 94

ANI Average Nucleotide Identity. 57, 70, 87, 89, 90, 94, 97, 100, 117, 145, 194

ASV amplicon sequence variant. 52

BLAST Basic Local Alignment Search Tool. 48–52, 54, 70

BLAT BLAST Like Alignment Tool. 48, 49

bp base pairs. 33

BUSCO Benchmarking Universal Single-Copy Orthologs. 123

BWT Burrows-Wheeler transform. 54

CAZy Carbohydrate-Active enZYmes. 62

CCA Canonical Correspondence Analsis. 125, 126, 200

cDNA complementary DNA. 35

CNN Convolutional Neural Network. 196

COG Cluster of Orthologous Groups. 61, 62, 71

CTD Conductivity, Temperature, Depth. 65

DBG deBruijn graph. 42–44, 55

DCM deep chlorophyll maximum layer. 10, 59, 60, 64, 117, 149, 178, 180, 181, 190

ddNTP dideoxyribonucleotide triphosphate. 32, 33

dNTP deoxyribonucleotide triphosphate. 20, 32

EBI European Bioinformatics Institute. 5, 148

EC Enzyme Comission. 71

ESOM emergent self-organising maps. 53

EVE EVolutionary Ecosystem. 197, 198

FACS fluorescence-activated cell sorting. 36

GO Gene Ontology. 62, 71, 72, 106, 118, 124, 183, 184

GOS Global Oceans Survey. 129, 194

GSEA Gene Set Enrichment Analysis. 141, 170, 178, 180, 181, 183, 197

GTDB Genome Taxonomy Database. 44, 57

GTDB-Tk Genome Taxonomy Database Toolkit. 57, 68, 247

Gy Gigayears. 21

HMM Hidden Markov Model. 49, 50, 60, 62, 69

HMP Human Microbiome Project. 5, 122, 148, 168–172, 190

ICA Independent Component Analysis. 129, 130

IMG Integrated Microbial Genomes & Microbiomes. 52, 61, 67, 68, 71, 73, 76, 78, 104, 199

IMM Interpolated Markov Model. 50, 51, 53

JGI Joint Genome Institute. 4, 52, 61, 64, 67–69, 71, 78, 85, 195, 198, 199, 271

KDE Kernel Density Estimate. 161, 189

KEGG Kyoto Encyclopedia of Genes and Genomes. 61, 62, 71, 123, 124, 144–146, 171, 281

KL Kullback-Leibler. 146, 149, 156, 173, 178, 182

KO Kegg Orthology. 61, 62, 71, 169, 171

LCA Lowest Common Ancestor. 50–52, 56

LDA Latent Dirichlet Allocation. 128

LOOCD Leave-One-Out Correlation Decrease. 5, 137–139, 146, 159, 160, 162–166, 171, 173, 174, 189, 190, 194

LPD Latent Process Decomposition. 128

LSU large subunit. 46

MAG metagenome assembled genome. 3–5, 44, 52, 53, 55–59, 63, 64, 69–73, 81–85, 87, 89, 90, 93–95, 97–108, 112, 114–119, 121, 145, 147, 193–195, 198–201, 247, 249, 251

MDS Multidimensional Scaling. 125, 126

MIMAG minimum information about a metagenome-assembled genome. 56

MISAG minimum information about a single amplified genome. 56

MMETSP Marine Microbial Eukaryote Transcriptome Sequencing Project. 57, 62, 70, 87, 90, 94, 95, 102

MOSAiC Multidisciplinary drifting Observatory for the Study of Arctic Climate. 116, 145–147, 156, 158, 165, 167, 168, 195, 196, 200, 201, 280, 281

mRNA messenger RNA. 19–21, 41, 46

NCBI National Center for Biotechnology Information. 85

NGS next-generation sequencing. 33, 54

NMF Non-Negative Matrix Factorisation. 3–5, 122–124, 128–132, 134–136, 141, 142, 146–149, 156, 159, 165, 167–176, 178, 186–191, 193–197, 200, 201

NTP nucleoside triphosphate. 20

OLC overlap-layout-consensus. 41–43

OM-RGC Ocean Microbial Reference Gene Catalog. 123

OTU operational taxonomic unit. 52, 127

PCA Principal Components Analysis. 104, 118, 125, 126, 129, 143, 180, 181, 196

PCoA Principal Coordinates Analysis. 74, 79, 126

PCR polymerase chain reaction. 35

Pfam Protein Family. 106, 130

pPCR real-time quantitative PCR. 35

RED Relative Evolutionary Divergence. 57

rRNA ribosomal ribonucleic acid. 20, 46, 71

RT-PCR Reverse transcription PCR. 35

SAG Single-cell Amplified Genome. 36, 56, 58, 90

SMRT Single Molecule Real Time. 35, 36

SSU small subunit. 46, 47

SVM support vector machine. 43, 58

t-SNE t-distributed Stochastic Neighbor Embedding. 126

tRNA transfer ribonucleic acid. 20, 21

WGCNA Weighted Gene Correlation Network Analysis. 122–124, 127, 128, 135, 148, 165, 167, 168, 173–176, 178, 190

WGS whole-genome shotgun sequencing. 35, 47, 148

Chapter 1

Introduction

The ocean covers approximately 70% of the earth's surface and contains 97% of its water, and provides one of the planet's largest habitats for life [5]. While on land plants are largely responsible for the production of new organic matter utilising energy from sunlight, in the oceans this role is mostly performed by photosynthetic microbes known as phytoplankton [6]. These phytoplankton form the base of the marine food web as well as being responsible for approximately 50% of the planets atmospheric oxygen [7], and communities of phytoplankton and other microbes are essential in the cycling of elements vital to life such as carbon, nitrogen and phosphorus. Phytoplankton include both bacteria and more complex but still single-celled eukaryotes. In nutrient rich water, eukaryotes often dominate these communities, with blooms of such eukaryotic phytoplankton often being visible from space. For example the calcium carbonate scales of Coccolithophores give waters a characteristic milky blue colour (Figure 1.1).

The broad aim of this thesis is to expand the tools available for computational analysis of the ocean microbes and communities which are fundamental to the functioning of the oceans, and to help understand how these organisms and communities may respond under conditions of climate change. The geographic distribution and growth rate of marine microbes is dependent on environmental conditions such as temperature and nutrient concentration. Human activity is altering these ocean conditions, with consequences for marine microbial life. Increasing CO_2 concentration in the ocean, driven by anthropogenic atmospheric CO_2 emission, has led to ocean acidification which threatens calcifying species such as the widespread Coccolithophore *Emiliania huxleyi* [8]; the Arctic has warmed 5 °C since 1900 with reducing sea-ice cover [9]; nitrogen pollution in coastal regions causes large blooms and subsequent oxygen depletion, leading to expanding hypoxic ocean dead zones [10].

While these communities of ocean microbes have been studied and monitored for centuries by programmes such as the long-running microscopy based continuous plankton

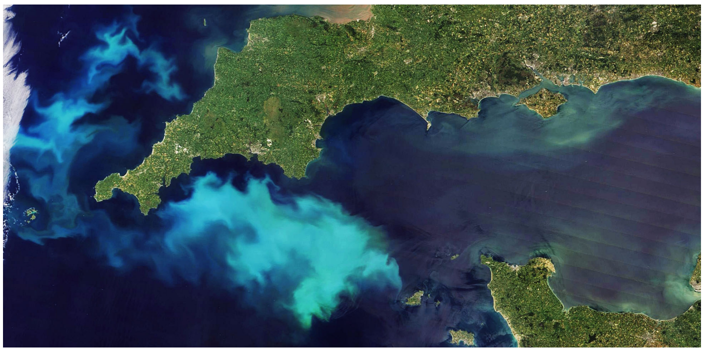


Fig. 1.1 Phytoplankton blooms off the south coast of England visible from satellite imaging from 2020. Light blue bloom is likely composed of Coccolithophores, which appear this colour due to calcium carbonate plates surrounding the organisms. (https://earthobservatory.nasa.gov/images/146897/channeling-a-bloom)

recorder [11], we study their genomic content through analysis of metagenomic sequencing data. The introduction of genomics techniques, and high throughput next-generation sequencing, allowed assembly of genomes for some marine microbes, giving new insight into their metabolism and traits. Genome assembly techniques were initially limited to those organisms which could be cultivated in isolation under lab conditions, which represent only a small proportion of the overall diversity of marine microbes [12]. Metagenomics encompasses a range of techniques to bypass this bias of culturability, and sequence and study the community as a whole [13]. At the base of metagenomic analysis is sequencing data, containing sequences of DNA from all of the organisms present in a sample taken from the environment. Large scale ocean expeditions performing metagenomic sequencing and analysis have begun to characterise the structure and function of microbes across the global ocean [4]. However data remains sparse from inaccessible but environmentally significant regions such as the Arctic [14], and challenges remain in analysing these data. We break down our aim of understanding the metagenomes of these microbial communities into two objectives.

Firstly at the level of individual genomes, we develop methods which allow us to recover and analyse genomes of uncultured eukaryotic microbes from metagenomic sequencing data. Reference genomes for eukaryotic marine microbes are currently sparse [15]. Many species appear unculturable, and some lineages have complex, difficult to assemble genomes. Eukaryotic phytoplankton have a complex evolutionary history, believed to have emerged as the result of endosymbiotic events, when a non-photosynthetic eukaryote engulfed a cyanobacterium, part of whose genetic material is retained in current eukaryotic phytoplankton in the chloroplast; many of the important lineages of such as diatoms originate from further secondary endosymbiosis, where these organisms were themselves engulfed [16]. This paucity of reference data poses challenges for meta-omic analysis; we have few organisms against which to compare new sequences to identify lineage or function. Metagenome binning methods have been developed to computationally recover draft genomes from metagenomic data, helping to grow the range of organisms for which we have genomic information [17]. Bacterial and archaeal metagenome assembled genomes (MAGs) have been generated in large volume for marine microbes [18], but few for eukaryotes; at the time of commencing this thesis we were aware of 2 such MAGs [19, 20]. and there were had been no studies specifically targetting automated recovery of multiple eukaryotic MAGs from ocean metagenome data. Our first objective was to recover eukaryotic MAGs from 12 samples spanning the Arctic and Atlantic oceans, and hence to expand the range of eukaryotic organisms for which we have genomic information.

Secondly at the level of the entire community, we aim to develop methods to produce an interpretable description of the functions and hence metabolic potential of microbial communities which captures local as well as global patterns. Taxonomic and functional annotations of meta-omic data results in high dimensional data, with potentially tens of thousands of functions or taxa identified [21]. For ocean data, while the volume of data is growing rapidly, the number of samples available for machine learning techniques remains low in comparison to "big data" fields such as computer vision or satellite remote observation. Unsupervised methods are beneficial for exploratory analysis of data, where we seek to learn underlying structures such from meta-omic data, such as groups of functions which share similar patterns of distribution across geographic space or environmental conditions. Individual functions may participate in responses to multiple conditions however, so approaches are needed which can identify local as well as global patterns. To address this we develop methods to apply the machine learning method Non-Negative Matrix Factorisation (NMF) for metagenomic analysis, to provide an reduced dimensional description of community function which reflects this underlying assumption of functions being shared among latent structures.

1.1 Thesis Structure

This thesis is focussed on analysis of ocean meta-omic data using computational tools, and so opens with two background chapters providing context first on the ocean and its resident microbes, and second on the computational tools of meta-omics.

In Chapter 2 we provide a broad outline of the global ocean and processes which shape it, including circulation, stratification, and biogeochemical cycling. A summary of the microbial life found in the oceans is given next, and salient points about their genetics and community assembly processes. The chapter closes with details of the sequencing technologies which are the source of the data meta-omic techniques seek to process.

Chapter 3 adds bioinformatics background to the preceding environmental and biological context. Computational steps which have become common in bioinformatics analyses including assembly, taxonomic classification, gene prediction and functional annotation are covered. Methods for recovering genomes from metagenome assemblies are introduced here, along with corresponding methods for assessing their quality, phylogeny and function.

Chapter 4 details the pipeline we developed for automated binning of eukaryotic MAGs. This was applied to metagenome sequencing from 12 samples spanning from the subtropical Atlantic to the Arctic Oceans, collected in 2012 by Katrin Schmidt and Klaus Valentin [22], and sequenced and assembled by Joint Genome Institute (JGI). The pipeline we developed recovered 21 eukaryotic MAGs from this data. These included organisms from environmentally significant lineages including prasinophytes and diatoms. Our results add to the binning of prokaryotes from the same data performed by colleagues at JGI. Our analysis of these MAGs shows a clear demarcation between polar and non-polar MAGs, in terms of which organisms are present and the functions they encode. Among these MAGs we also show an associated eukaryote-prokaryote pair, with functions enriched suggesting a mutualistic relationship. The results of this research have been published in an article in *Microbiome* [23], as well as contributing to a chapter in *The Molecular Life of Diatoms* [24], and under review following an invited submission at *Data in Brief.* Furthermore, one of the MAGs recovered has since been added to the PhycoCosm algal genomics resource maintained by JGI as *Micromonas* sp. AD1 [25]. Additionally I extracted sequences related to zinc-binding from the same samples studied in this chapter for analysis by colleagues, which subsequently formed part of an article published in *Nature Ecology and Evolution* [26]. My contribution was to develop the eukaryotic binning pipeline, performing all analyses of both prokaryotic and eukaryotic MAGs, and writing the resulting paper. Colleagues at JGI performed sequencing, assembly, taxonomic and functional annotation prior to binning, and reviewed paper drafts.

Chapter 5 discusses our results in applying the unsupervised machine learning technique NMFs to meta-omic data. This chapter opens with a brief background on approaches used in analysis of meta-omic data, to help illustrate where NMF differs and its potential benefits. Our methods for selecting parameters and interpreting models are detailed, along with methods of generating simulated data for evaluating our approaches. We show results on two simulated datasets, followed by three real world case studies illustrating how this method can be applied to meta-omics data, including large scale studies such as Tara Oceans. My contribution to this research was conceiving and performing rank selection experiments, generating synthetic and simulated data, conception of two new methods to interpret features in decompositions (permutation and Leave-One-Out Correlation Decrease (LOOCD)), conceiving and performing experiments to evaluate module recovery, implementing methods as a python package, implementing visualisation methods, and applying these methods to real world data as case studies. Three public sources of data are use in the case studies, from the Human Microbiome Project (HMP) [27], a study of a New Zealand river estuary [1], and the European Bioinformatics Institute (EBI) functional annotation of Tara Oceans data [3, 4].

The thesis closes in Chapter 6 with a summary of our main results. We also propose future work following on from our work, including further analysis of MAGs, additional computational tools which could aid eukaryotic binning, additional applications of NMF, and a potential approach to combining metagenome analysis and earth system models. Finally, we discuss how our results fit into the broad context and future of the field.

Chapter 2

Biological and Environmental Background

2.1 Summary

This thesis explores computational methods to understand the wealth of metagenomic data being generated from oceans microbe communities. Before exploring the computational side, this chapter will first provide an outline of the oceanic and microbiological context, in order to frame both the problems and results in later chapters. Section 2.2 provides an overview of the forces shaping the contemporary ocean and its biogeochemistry. Section 2.3 introduces the microbial population of the ocean, the basics of their biology and ecology. Section 2.4 covers the distinctive features of the planet's youngest ocean, the Arctic Ocean, to place into context the Arctic datasets which have been used in Chapter 4. Section 2.5 looks at some of the ways in which the Earth's changing climate could influence the ocean, its microbial inhabitants, and the Arctic. Finally, 2.6 describes the sequencing technologies which provide a view into the molecular activity of marine microbes, and which are the source of the data on which the computational methods of Chapter 3 operate.

2.2 Oceans

Oceans cover 70.8% of Earth's surface, and contain 97% of the planet's water. They play an important role in global processes, including cycling of elements vital to life on Earth such as carbon, nitrogen and phosphorus, regulating temperature by absorbing the high levels of equatorial solar irradiance and circulating this energy across the global ocean, and in the hydrological cycle in which moves water between atmosphere, ocean, and ground [5]. Life in the ocean is estimated to account for biomass two orders of magnitude lower than terrestrial biomass (≈6 Gt C in the ocean, ≈470 Gt C on land), but accounts for approximately 45% of the planetary primary productivity [28, 6]. This section draws on the textbook of Webb [5] to provide a broad overview of some well established important characteristics of the world's oceans.

2.2.1 Circulation

Two broad systems of currents operate in the oceans; wind driven surface currents (Figure 2.1), and density driven thermohaline circulation which includes deeper waters driven by differences in temperature and salinity (Figure 2.2).

Surface Currents

The surface currents are driven by global atmospheric circulation which is divided up into six cells, with three each side of the equator. Either side of the equator are the Hadley cells, at either poles the polar cells, and between are mid-latitude Ferrel cells. In the Hadley cells, warm air rises near the equator, cooling as it rises and eventually being forced polewards by the rising air beneath it. The cooled air eventually sinks at about 30° latitude, returning to the surface and flowing back towards the equator, heating and gaining moisture. The Coriolis effect from Earth's rotation gives these winds an easterly direction. Air in the polar cells circulates in the same easterly direction at the surface. Circulation here is also driven by convection, air rising at approximately the 60° latitude, and descending towards the poles, with the easterly direction imparted again by the Coriolis effect. Air circulation in the mid-latitude Ferrel cells moves in the opposite direction to its neighbouring polar and Hadley cells, having a westerly direction at the surface. Circulation in the Ferrel cells is driven in large part by the motion in cells to either side of it, rather than heat driven convection.

This movement of air near the surface drives the surface ocean currents. Near the equator, the easterly winds drive the north and south equatorial currents which flow toward the west. Similarly at midlatitudes, the westerly winds create currents heading toward the east. Interactions with continental landmasses and the Coriolis effect create large scale circular gyres within the oceans, shown in Figure 2.1. Water at the centre of these gyres moves very little, while water at the edges circulates around it. The Antarctic circumpolar current circulates continuously eastward around Antarctica, uninterrupted by any landmass. The Arctic Ocean is more enclosed, but the Beaufort gyre is found in the Canada basin, serving to collect a growing volume of fresh water at its centre [29].

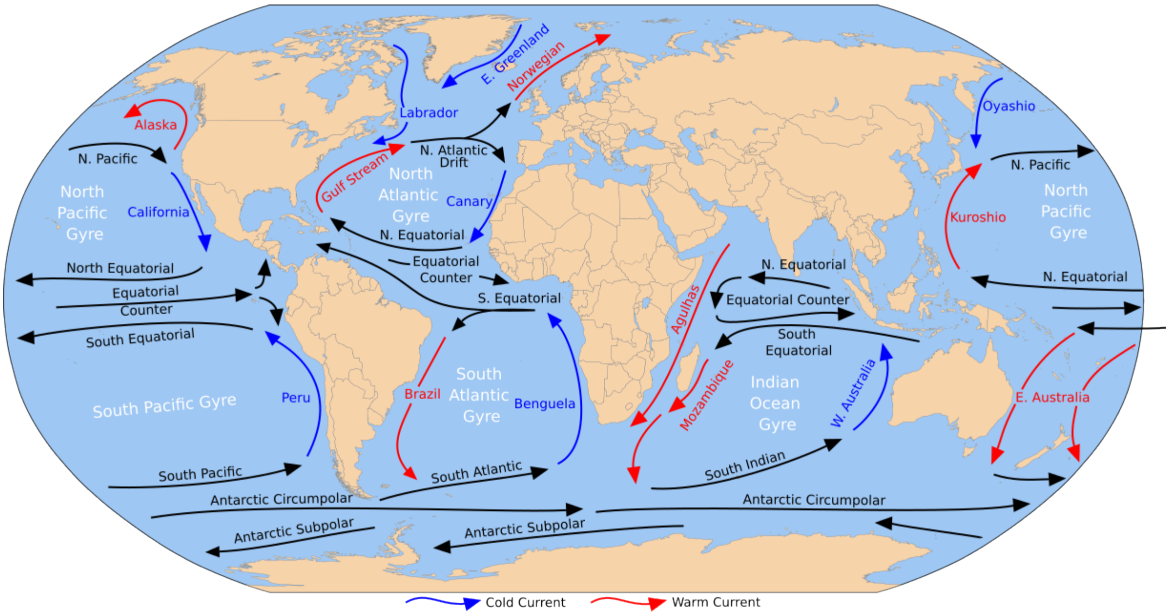


Fig. 2.1 Global surface ocean currents. Blue indicates cold currents, red indicates warm currents. Five of the large ocean gyres are labelled in white. Adapted from *Ocean surface currents*, by M. Pidwirny, 2007 (https://commons.wikimedia.org/wiki/File:Corrientes-oceanicas.png). Public domain image. [30]

Thermohaline Circulation

Thermohaline circulation is the circulation driven by gradients in density resulting from temperature and salinity conditions, and is slower than these surface currents. These currents are shown in Figure 2.2. At high latitudes during polar winters with little light, water density increases due to cooling and incorporation of fresh water into sea ice, increasing salinity of the water beneath. This dense water sinks to the ocean floor, and there moves slowly towards the equator. These cold bottom waters collect behind raised areas of the ocean floor, before spilling over often forming narrow valleys, called gateways. The cold bottom water slowly becomes less dense due to ocean mixing, and as newer more dense cold water arrives it will be forced upwards. This upwelling motion is very slow, with upwelling once thought to be broadly diffused across the oceans, but more recently that a large amount of the global upwelling occurs in the Southern Ocean [31].

2.2.2 Vertical Structuring

The ocean is vertically stratified, with layers divided by differences in density of water, shown by the coloured lines in Figure 2.3 [5]. Beneath the surface ocean, the points at which the change in temperature and salinity are at their greatest define the thermocline and halocline

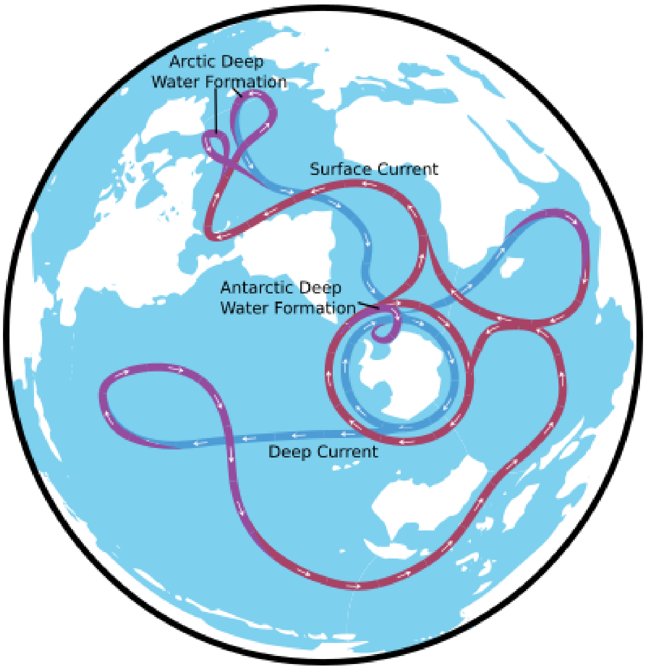


Fig. 2.2 Map illustrating thermohaline circulation. Adapted from *Map of the world's "conveyor belt"*, by Asva, 2009 (https://commons.wikimedia.org/wiki/File:Conveyor_belt.svg). Copyright under Creative Common Attribution-Share Alike 3.0 License. [32]

respectively. The pycnocline is the point of greatest change in density, which is driven by temperature and salinity, so often occurs at similar depths to the thermo- and halocline. These layers form barriers to mixing, keeping cold dense waters below the warmer less dense surface waters in most oceans. The depth of the pycnocline varies across the oceans, from between 10 to 500 metres, and can be absent in high latitude polar oceans. Strong ocean stratification prevents the mixing of nutrient rich deep waters into the sunlit ocean, and so limiting the nutrients available for microbial life. A layer of interest when considering the activity of photosynthetic microbes is the Deep Chlorophyll Maximum Layer (DCM). This is the point at which chlorophyll, taken as a measurement of primary productivity, peaks. Depth of the DCM is influenced by nutrient availability and light penetration, conditions which also impact the types of phytoplankton present.

A second factor in vertical structuring of the oceans is penetration of sunlight, shown by the horizontal coloured bands in Figure 2.3. Water rapidly absorbs light, with only 1% of light remaining at a depth of 100 metres, and no light by 1000 metres. Different wavelengths of light are absorbed differently, with green and blue light penetrating furthest, and other colours more readily absorbed. The uppermost layer, the epipelagic zone extends to the depth where 1% of surface light remains. This depth can vary depending on water conditions, in clear open oceans the photic zone will extended further than in water rich with phytoplankton.

Phytoplankton form the base of the upper ocean food web, as well as exporting organic carbon to lower layers of the ocean as deceased matter drifts downwards. The mesopelagic zone extends from the bottom of the epipelagic zone to the depth at which no light remains. With little to no light, organisms living in this layer utilise matter exported from above. Organisms move between the epipelagic and mesopelagic ocean, known as diel vertical migration, contributing to the transport of biomass between these layers. Estimates of how much life the mesopelagic ocean contains vary. A recent study based on acoustic data gave a median estimate the biomass of mesopelagic fish to be about 11 Gt [33], and order of magnitude higher than previous estimates of 1 Gt [34]. In the deep lightless ocean are the bathypelagic and abyssopelagic zones, from 1000 to 4000 metres and 4000 to 6000 metres respectively. The bathypelagic contains extremely little or no primary production, reliant on the small proportion of matter which is transferred from the shallower ocean. Across much of the ocean this layer meets the ocean floor, and can contain hydrothermal vents whose heated waters provide a supply of elements such as iron and sulphur.

2.2.3 Ocean Biogeochemisty

Microbial life in the ocean relies on the presence of a variety of critical nutrients, and understanding their cycling can help understand the biogeography and stresses on microbial communities. Here we provide and overview of the cycling of three elements important for primary production by phytoplankton, carbon, nitrogen and phosphorus, as well as some significant trace elements.

Carbon Cycle

The marine biological carbon pump starts in the surface ocean, where phytoplankton use dissolved carbon dioxide during photosynthesis, producing glucose and oxygen [5]. These are further converted into forms of organic carbon such as carbohydrates, lipids and proteins. Some plankton, such as coccolithophores, also incorporate carbon into calcium carbonate structures. Larger zooplankton prey on primary producers, transferring their organic carbon to higher trophic levels of the epipelagic foodweb. Dead phytoplankton and zooplankton fecal matter form part of the marine snow, which sinks towards the ocean floor. A large proportion of this matter is decomposed by bacteria, and returned to the oceans pool of inorganic carbon. However a small proportion, about 1%, will reach the ocean floor sediments where it can be stored for several million years [5]. This biological pump is an important process in regulating the amount of atmospheric carbon dioxide, and hence global temperature, as discussed later in Section 2.5 [35].

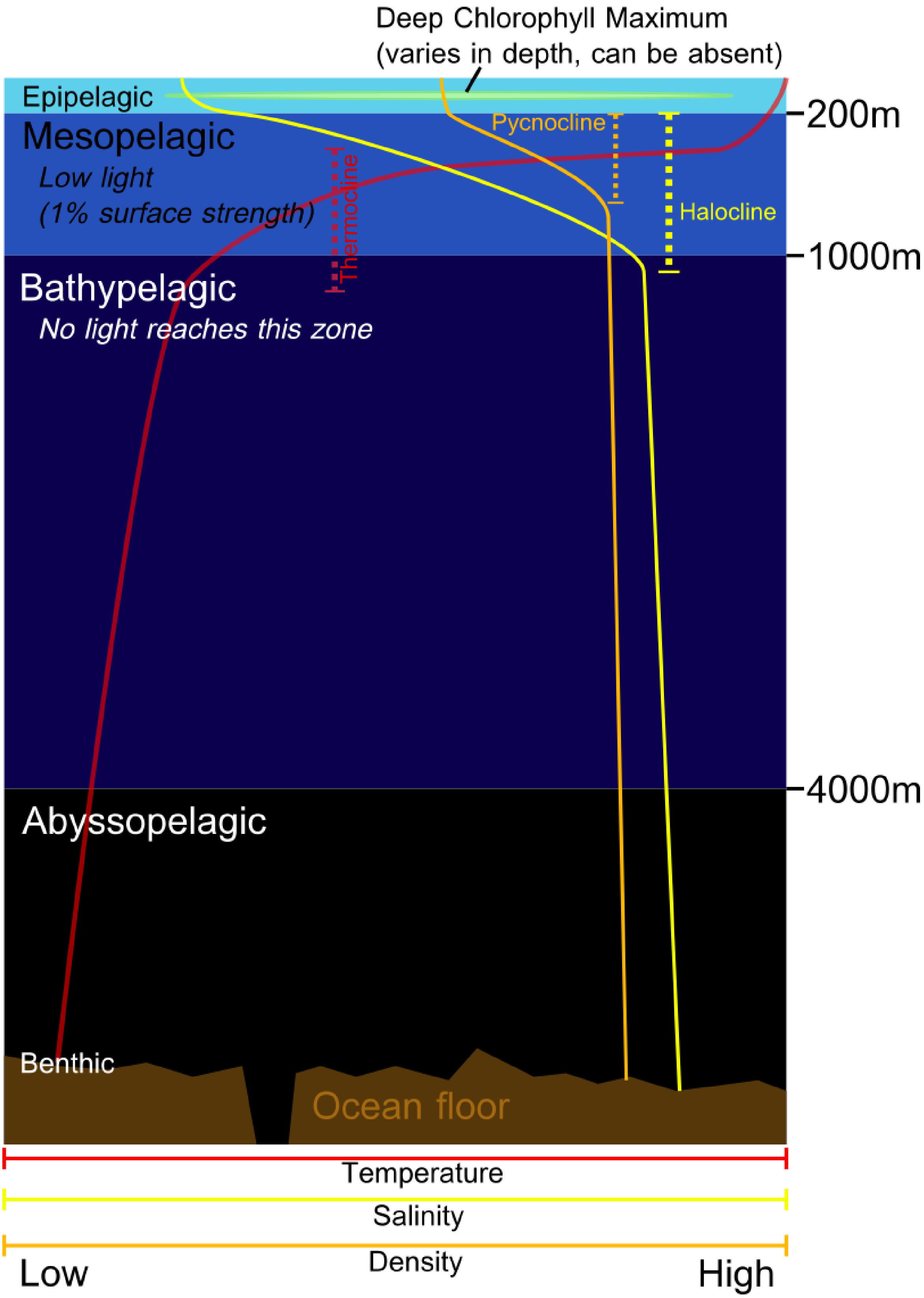


Fig. 2.3 Layers in the open ocean, and example temperature, salinity and density gradients. The deep chlorophyll maximum is indicated, but the depth can vary, and a chlorophyll maximum can be absent completely. Example temperature, salinity and density gradients are given in red, yellow, and orange lines respectively. The regions of greatest change for each are highlight with dashed lines, the thermocline, halocline, and pycnocline respectively.

Nitrogen Cycle

Nitrogen is an essential element for ocean life, being incorporated into all amino acids and hence proteins, nucleic acid, and as a major component of chlorophyll is important for primary productivity in the oceans. Pajares et al. [36] recently reviewed the state of knowledge about the marine nitrogen cycle, and summary of their findings is provided below.

Nitrogen is abundant in the atmosphere as gaseous N_2, but in this form it is not available for biological use. Nitrogen fixation is the process of converting dissolved N_2 into biologically usable ammonia (NH_3), a reaction catalysed by nitrogenase. Microbes which carry out this process are known as diazotrophs. Nitrogenase is sensitive to O_2, which is produced during photosynthesis. This leads to photosynthetic organisms carrying out nitrogen fixation at night, or in some cases creating anoxic environments in heterocysts for nitrogen fixation. Diazotrophs are known to be abundant in warm, oligotrophic surface oceans; large blooms of the cyanobacteria *Trichodesmium* form in the Tropical and North Atlantic Oceans. However, the distribution of dizaotrophs is not limited to these tropical surface waters, with biological nitrogen fixation shown to be present in a wide range of environments including the Arctic, coastal upwelling regions, and hydrothermal vents.

Nitrification is a process which converts ammonia (NH_3) to nitrate (${NO_3}^-$). This is often a two step process with each step being carried out by different organisms; ammonia is oxidised to nitrite (${NO_2}^-$), then nitrite oxidised to nitrate. Ammonia oxidation is carried out by species of both bacteria and archaea (AOB/AOA), and nitrite oxidising by bacteria (NOB). Organisms were recently discovered which carry out both steps of this process, known as commamox organisms (complete ammonia oxidation), with evidence suggesting a high relative abundance in coastal regions compared to near absence in the open ocean [37, 38]. Ammonia, nitrite and nitrate can be used by phytoplankton as sources of nitrogen, and their use by these primary producers introduces nitrogen into the ocean food web. Nitrogen fixed in the surface ocean can be recycled within the sunlit ocean, but some will be transferred to layers below by similar processes to carbon transfer; vertical migration, falling dead matter and excrement, or mixing of layers.

Biologically available nitrogen is returned to inorganic nitrogen through a corresponding set of processes. Denitrification returns nitrate to N_2, though the process is carried out in several steps where individual organisms may only conduct a subset of them. In the oceans, denitrification is largely limited to very low oxygen areas such as oxygen minimum zones (OMZ) and sediments. The anammox process returns ammonium and nitrite to N_2, and is currently only known within the order of bacteria Planctomycetales, and occurs largely in low oxygen environments such as OMZs. N-Damo (nitrate/nitrite-dependent anaerobic methane oxidation) is a more recently discovered process that couples methane (CH_4) oxidation to nitrate or nitrite reduction, resulting in production of N_2.

Phosphorus Cycle

Phosphorus is also vital for life, forming the backbone of DNA and RNA, being involved in energy transmission in ATP, and orthophosphate being used during photosynthesis. The description of the marine phosphorus cycle given below is based on the review conducted by Paytan and McLaughlin [39].

Unlike nitrogen, phosphorus cannot be fixed from the atmosphere. Instead, the main source of phosphorus is continental weathering, reaching the ocean in river inputs or through deposition of atmospheric dust into the oceans. Riverine input is significant for coastal or estuarine systems, but further from the coasts dust deposition becomes an important source. Phosphorus gets removed from the ocean system when it is buried in ocean floor sediment. In addition to these natural geological sources, human activity provides additional source of phosphorus; it is a limiting factor in plant agriculture, and the phosphorus fertiliser used in these industries can end up being washed into coastal systems. Clearing of land for agriculture can also lead to increases in mineral rich atmospheric dust, known as eolian dust, which is later deposited in the ocean.

Phosphorus is present in the oceans in dissolved or particulate forms. These can be divided further into organic and inorganic forms. Inputs arrived as particulate inorganic phosphorus (PIP) or particulate organic phosphorus (POP). Inorganic phosphorus, usually in the form of orthophospate, is the form which can be assimilated by phytoplankton, and subsequently into organic compounds. These phytoplankton are grazed on by larger zooplankton, and some phosphorus returned to the ocean dissolved organic phosphorus (DOP) pool by zooplankton excretion. DOP consists of biological products such as carbohydrates, lipids, and proteins. Microbial activity can return this organic phosphorus to inorganic form, with bacteria and some phytoplankton producing enzymes which catalyse this process. Both dissolved inorganic phosphate (DIP) and DOP can be adsorbed into sinking particulate matter, moving phosphorus between the dissolved and particulate pools.

DIP often shows a gradient with depth, with the surface ocean being depleted in DIP due to being used up by microbial productivity, and increasing in concentration with depth, accumulating in old deep waters. The opposing trend is shown by DOP, which is highest in the surface oceans where it is produced.

The hydrolysis process which converts DOP to DIP is carried out by bacteria throughout the water column, as well as by phytoplankton in the sunlit layers, resulting in very little DOP being transferred to lower depths. As DOP is abundant in the surface but not biologically available, productivity can be limited by the rate at which DOP can be remineralised to biologically available DIP. Productivity can also be influenced by variation in the phosphorus requirements of different lineages of organisms. While the average required ratio of N/P is stable at about approximately 16 (see Section 2.2.3), individual species can require a higher or lower availability of phosphorus, between 8.2 and 45. Hence the makeup of the microbial community can affect when a system become phosphorus limited.

The phosphorus cycle connects to other ocean biogeochemical cycles. For instance diazotrophic nitrogen fixation requires phosphorus, so the level of bioavailable phosphorus can limit nitrogen fixation, and hence primary productivity. As the sources and sinks of phosphorus are driven by geological processes, over long geological timescales phosphorus has been considered the limiting nutrient for long term ocean productivity. At any given point, a system may be limited by other nutrients or trace elements, but over a geologic timescale the limiting factor may be phosphorus.

Redfield Ratio

The Redfield ratio is an observed relationship which links carbon, nitrogen and phosphorus. The same ratio of carbon, nitrogen and phosphorus was observed in both marine phytoplankton organic matter, and in the water of the deep ocean, across the oceans. The seemingly static ratio was initially observed between nitrate and phosphate (a ratio between N:P of 16:1) [40], and later extended to include carbon (ratio C:N:P of 106:16:1) [41]. This ratio appears on average constant in phytoplankton and deep ocean water. Two mechanisms were initially proposed explaining this observation. Firstly that phytoplankton N:P reflects the composition of the water around it, with species having different nutrient requirements competing and the community eventually coming to reflect the nutrients conditions in the surrounding water. Second that this is maintained by biological feedback, with the activity of organisms such a diazoptrophs and denitrifying bacteria moving the nutrient composition of seawater closer to that of the phytoplankton. This connection has formed an important part of the current understanding and modelling of marine carbon and nutrient cycling [42].

Trace Elements

A number of less abundant elements also play important roles in marine microbial processes, and a lack of them can act as limiting factor on productivity. Iron is a vital nutrient for a range of cellular processes, being used in the respiratory electron transport chain, and with particular relevance for phytoplankton, in the photosynthetic electron transport chain (see Section 2.3.2). Phytoplankton growth has been estimated to be iron limited in a large area of the surface ocean, between 30% and 50% [43]. Iron limitation of phytoplankton growth has been observed in cultures under lab conditions, but also demonstrated in situ in the oceans [44]. Controversial iron fertilisation experiments introduced bioavailable iron into high nutrient low chlorophyll areas of the ocean, and showed that this could stimulate phytoplankton blooms in these systems [45].

Some trace elements are more important in specific regions and to specific lineages of organisms. Recent research suggests that zinc is an important nutrient in polar regions, and that polar phytoplankton have a raised demand for zinc [26]. Concentration of dissolved zinc is high in polar oceans, and concentrations of cellular zinc in the phytoplankton which inhabit them reflects this. Genomic evidence shows families of zinc-finger proteins are expanded in polar phytoplankton, and co-expressed with genes involved in primary metabolism such as photosynthesis or fatty acid metabolism. This was supported by metagenomic evidence as well, with density of zinc domain genes being raised in polar compared to non-polar metagenomes, and in those genes the ratio of synonymous to non-synonymous mutations suggesting selection. However the importance of zinc is region and lineage specific, zinc concentration is low in other ocean regions such as tropical oceans, and the increased zinc requirement seems specific to those species which have colonised the polar oceans, particularly the Southern Ocean.

2.3 Marine Microbes and Microbiomes

2.3.1 History

While molecular methods of studying ocean microbial communities are new, the study of these communities has a much longer history. Antonie van Leeuwenhoek was the first to observe and count "animalcules" in drops of water as early as 1675, among the many subjects he studied with his advances in microscopy [46]. From these very first observations, it has been clear that sample environments contain multiple different types of organisms, living together as a microbial community. Later Adolphe-Adrien Certes arranged for samples of deep sea sediment to be collected on the Talisman and Travailleur expeditions of 1880 to 1883, and working in the lab of Louis Pasteur obtained cultures of microbes from these field samples, demonstrating the near omnipresence of microbial life across even the most extreme of earth’s environments. An early example of experimental marine microbiology, both Certes and Paul Regnard sought to study microbial activity under different pressures, studying the rate of putrefaction of matter in chambers under various pressures [47].

Marine plankton have since been the subject of one of the worlds longest running biological monitoring programs, the Continuous Plankton Recorder (CPR) [11]. This program has been running since 1931, and using the same methods since 1958, providing multi-decadal insight into parts of the marine microbial community. Despite the roughly two and half centuries separating them, the taxonomy of species captured by the CPR is identified by microscopy, as with Leeuwenhoek's counts of animalcules. For many decades the ocean microbial population was thought to be sparse and of little importance to broader ocean processes, however Azam [48] highlights a series of advances through the 1970s and 1980s which changed how we understand microbial life in the oceans: studies of microbial respiration suggested microbes formed a large part of the oceans metabolism and were connected to consumers at higher trophic levels [49]; fluorescence based methods permitted a direct count of bacteria which was three orders of magnitude greater than previous estimates [50]; and estimates that bacteria utilised between quarter and half of primary production playing a significant role in the ocean carbon cycle [51, 52].

Molecular methods similarly represented a sea change in the study of marine microbial communities. Extracting and sequencing genomic material allowed new insight into the evolutionary history of organisms, and the diversity of microbial communities. Analysis of genes coding for subunits of ribosomal RNA (rRNA) rearranged the broad understanding of evolution into three domains, bacteria, archaea, and eukarya [53]. The development of techniques to obtain sequences from environmental samples without culturing resulted in an increased understanding of the diversity of organisms present in the ocean, which had not been evident using microscopy [54, 16].

The introduction of what has become known as next-generation sequencing provided both increased throughput and reduced cost. Using these new technologies, the range of culturable marine microbes with full genomes sequenced began to expand. Beyond those organisms which could be cultured, techniques developed for whole genome shotgun sequencing from environmental samples [13]. This approach generates reads from the genetic sequences of organisms in an environmental sample without the need for culturing, allowing insight into genetic makeup of unculturable members of natural marine communities. These sequencing and analysis tools make up the field of metagenomics, and the generation of metagenomic data from the oceans has continued to increase dramatically. High profile globe spanning expeditions like the Global Oceans Survey and Tara Oceans [55, 4] generated large volumes of data which are still being analysed and generating new results, and recent expeditions like the year round Arctic expedition MOSAiC [56] continue to expand the regions and ocean conditions for which metagenomic data is available.

As well as expeditions getting larger and reaching more challenging parts of the oceans, sequencing technology has continued to develop. Third generation sequencing platforms offer much longer reads, albeit currently with higher error rates, and portability allowing sequencing in-situ [57]. In parallel, improving techniques for single-cell sequencing are helping to further expand the range of organisms for which genomes are available, allowing sequencing of specific organisms from environmental samples without thee need to grow them in culture [58].

2.3.2 Molecular Microbiology

All known living organisms divide into three domains, bacteria, archaea and eukarya, based on their evolutionary history, shown on the left of Figure 2.4. Archaea and eukarya are more closely related to each other than they are to bacteria, though the exact relationship between these domains remains to be resolved [59]. A further grouping can be made based on the cell structure of organisms, shown in a highly simplified form on the right of Figure 2.4. Prokaryotes are the archaea and bacteria, whose genomes are organised in a circular structure in the cell's cytoplasm. Their genomes tend to be simpler, with sequences coding for genes close together without much intergenic DNA. Eukaryotes are more complex in structure, and in the organisation of their genetic material. Eukaryotes have membrane bound organelles which contain their genetic material. All eukaryotes have a membrane bound nucleus, which contains their nuclear DNA. This is organised into multiple chromosomes, which are linear with a start and end rather than circular as in prokaryotes. A much higher proportion of eukaryotes genetic material is non-coding, with genes interrupted by introns, portions of sequence which do not code for a protein. In addition to being split across multiple chromosomes, these chromosomes can be present in one or more copy, known as the ploidy of the organism. Other organelles in eukaryotic cells, such a chloroplasts and mitochondria, contain their own separate genome. These genomes are organised in a way more similar to prokaryotes, being circular and with a high proportion of coding DNA.

Another important division can be made between autotrophs and heterotrophs. Autotrophs store chemical energy in organic compounds synthesised from inorganic compounds in the environment such as CO_2 and water, a process called primary production. Primary production requires an energy source, with photosynthetic organisms utilising sunlight, where chemosynthetic organisms used energy from inorganic chemical reactions. Chemosynthesis is more common in deep sea organisms where light is scarce or absent, with the light filled surface ocean home to more photosynthetic organisms. On the land, autotrophs are dominated by multicellular plants. However in the oceans the bulk of autotrophs and primary producers are single celled microbial phytoplankton. Heterotrophs rely on other organisms for organic

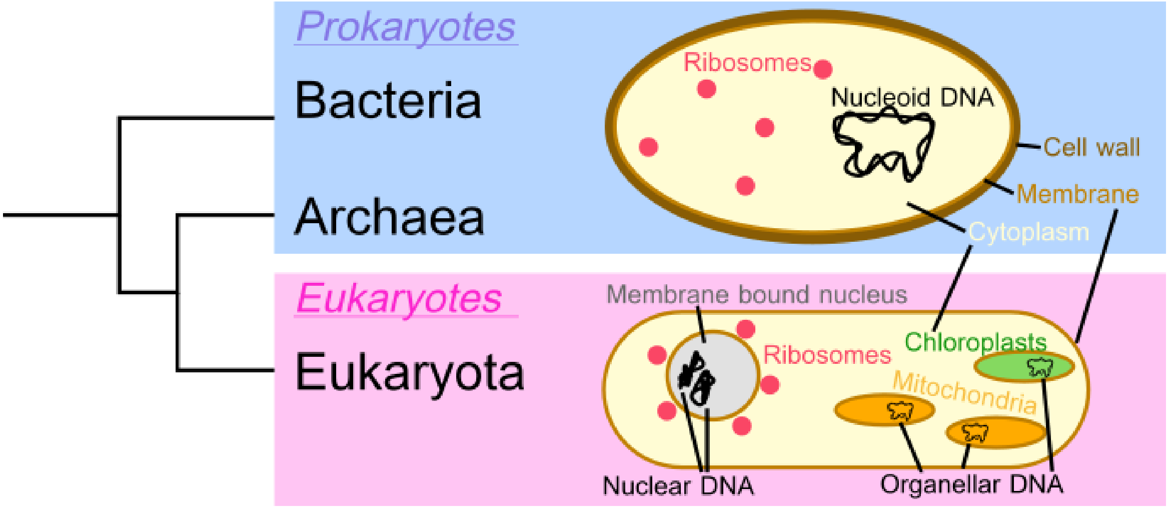


Fig. 2.4 Relationship and basic structure of the three domains of life, bacteria, archaea and eukarya. The tree on the left shows that eukarya are more closely related to archaea than to bacteria. Cell diagrams on the right show a highly simplified bacteria and phytoplankton cell, to highlight some high-level structual differences.

carbon, acquiring them by consuming either autotrophs, or other heterotrophs. Ocean heterotrophs range from microbes from all domains, up to the oceans biggest organisms like whales. Mixotrophic organisms combine both of these trophic modes and is believed to be widespread in plankton, with both a diverse range of organisms being mixotrophs and with mixotrophs being present broadly across the surface oceans [60]. In the Arctic summer, mixotrophic ciliates make up a large portion of the total chlorophyll outside the diatom blooms close to sea ice. Zooplankton preferentially feed on these ciliates, making mixotrophic organisms an important component in transferring nutrient to higher trophic levels [61].

Nucleic Acids and Proteins

There are three biopolymers of interest to us looking at microbial genetics: DNA, RNA and proteins. In a broad sense, DNA contains genes; genes encode the sequence of amino acids in a protein which has a biological function [62]. To synthesise a protein from a gene, RNA polymerase binds to DNA and creates messenger RNA (mRNA), a process called transcription. Ribosomes bind to mRNA and assemble amino acids into a protein in the order contained in the mRNA, known as translation. An overview of the structure and role of these three biopolymers is given below [63].

DNA

DNA is a double stranded molecule of two polymers, each strand composed of a sequence of deoxyribonucleotide triphosphate (dNTP) molecules. Each dNTP contains a nitrogenous base bound to a deoxyribose sugar, and three phosphate groups bound to this sugar. Four nitrogenous bases are used in dNTPs: adenine, cytosine, guanine, and thymine, commonly abbreviated A, C, G, and T. The two strands are joined by hydrogen bonds between pairs of bases: A with T, and C with G. The sequence of bases in a strand of DNA carries the genetic information. Each strand stores the same information, and by base pairing the complimentary strand of a single strand can be synthesised, allowing replication of DNA. An organism's genome is composed of one or more chromosomes containing DNA. Prokaryotic and eukaryotic genomes are organised differently. Bacteria usually have one circular chromosome containing their genome. Eukaryotes have a nuclear genome of one or more chromosomes stored in the cell nucleus. Organelles in either type of cell, such a chloroplasts or mitochondria, contain their own genome. Parts of this genome may be transferred to the nuclear genome over time.

A genome contains many genes; each gene codes for the production of a molecule which has a biological function. This coding region is flanked by untranslated regions which RNA polymerase can bind to in order to read the coding region and synthesise mRNA. The organisation of a gene is different in prokaryotes and eukaryotes, with a salient difference being the division of the coding region into introns and exons in eukaryotes. Introns are sections of the coding region which do not code for amino acids, and must be removed from the initially synthesised mRNA, leaving only the exons in the mature mRNA. This can make locating and interpreting genes in eukaryotic genomes more computationally challenging [64].

RNA

RNA is a single stranded polynucleotide, consisting a chain of nucleoside triphosphate (NTP) molecules. NTP molecules are similar to dNTPs, but using a ribose sugar instead of deoxyribose, and the base thymine is replaced with uracil, abbreviated U. Each mRNA encodes the order of amino acids in a protein. Ribosomes bind to mRNA to synthesise proteins by assembling amino acids based on the mRNA sequence. Some RNA has a direct function and is not converted to a protein in order to function. Ribosomes are largely composed of ribosomal ribonucleic acid (rRNA) which is not translated to a protein. Transfer Ribonucleic Acid (tRNA) connects amino acids and the codons in mRNA; each tRNA attaches to a specific amino acid, and the anticodon region of the tRNA pairs to the complementary codon in mRNA. The sum of the RNA in a cell is named the transcriptome.

Protein

Proteins are polymers composed of a sequence of amino acids. There are 20 standard amino acids, with some organisms using additional ones. The order of bases in a gene's coding region encodes the order of amino acids in the protein to be synthesised. Synthesis of proteins by ribosomes from mRNA is known as translation. A block of three nucleotides is a codon, each codon relates either to one amino acid or is a stop codon which terminates translation.

The relationship between codons and amino acids is called the genetic code. The genetic code is redundant, meaning more than one codon can code for the same amino acid. Genetic codes are very similar between organisms but variations exist where a codon codes for a different amino acid than usual in some organisms. The plastid and mitochondrial genomes of the recently described class of phytoplankton *Chloropicophyceae* have such a variant genetic code [65, 66].

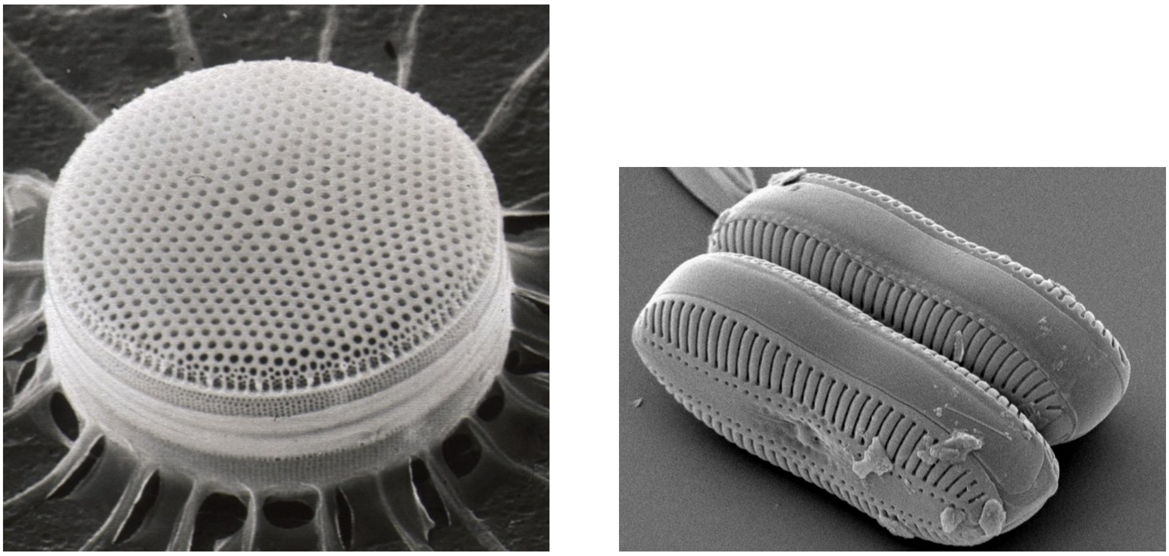
Proteins carry out most of the functions within a cell. Knowing which proteins are coded for in a genome enables us to make inferences about what functions the organism can carry out, and how it may interact with its environment.

Phytoplankton

The ocean's consumers all rely on the primary producers which form the base of the food web, so next we will give more detail on the phytoplankton which provide 45% of global primary production [6]. Phytoplankton are a group of autotrophic single-celled organisms found in aquatic environments which convert light to chemical energy via photosynthesis. Photosynthesis is believed to have been acquired in a prokaryotic ancestor of contemporary cyanobacteria by 2.7 Gigayears (Gy) ago. During the Mesozoic era eukaryotic microbes acquired photosynthesis and became dominant in phytoplankton communities [16]. Photosynthetic eukaryotes are believed to have emerged from a single endosymbiotic event when a eukaryotic cell engulfed a cyanobacterium [67]. Some photosynthetic components of the cyanobacterium have been retained in the eukaryote host as a plastid, the chloroplast, while other genes have been transferred to the nuclear genome of the host [68]. This endosymbiosis resulted in three major groups of photosynthetic algae: red algae, green algae and glaucophytes. Many contemporary species of eukaryotic phytoplankton such as diatoms and coccolithophores acquired chloroplasts through secondary endosymbiosis, where a heterotrophic eukaryote engulfed a red alga. Whether this secondary endosymbiosis of a red plastid occurred once or multiple times is unresolved [69].

The most significant and diverse groups of eukaryotic phytoplankton are diatoms, dinoflagellates and haptophytes. In terms of number of observed species, dinoflagellates and Stramenopiles (mainly represented by diatoms) make up a majority of the diversity of all phytoplankton [16].

Diatoms are a unicellular class of Stramenopiles characterised by a silica cell well called a frustule. They are a diverse group, with estimates suggesting that over 100,000 contemporary diatom species exist [70]. Morphologically, diatoms are divided into two broad groups based on frustule shape; centric, and pennate (Figure 2.5). Centric diatoms are older, with pennate diatoms evolving more recently and representing most of the species diversity. Diatoms are abundant in nutrient rich coastal areas, and estimated to account for 20% of global carbon fixation. In the Southern Ocean diatoms adapted to the fluctuating conditions in light, temperature and nutrients are the main primary producers [71]. Part of the carbon fixed by diatoms sinks to the ocean floor and becomes trapped in sediment, contributing about half of marine carbon sequestration.



(a) Centric diatom

(b) Pennate diatom

Fig. 2.5 Diatoms with pennate and centric frustules. Centric diatom image adapted from CSIRO image by Wikimedia, used under the Creative Commons Attribution 2.5 Generic license [72, 73]. Pennate diatom adapted from image in Bradbury via Wikimedia Foundation, used under Creative Commons Attribution 2.5 Generic license [74, 75].

Dinoflagellates are largely found in marine environments, and around half of the contemporary species are photosynthetic, and some photosynthetic dinoflagellates are mixotrophs, employing multiple strategies for acquiring energy and organic material [76]. They typically possess two flagella: a ribbon like transverse flagellum and a longitudinal flagellum which allows the cell some movement within the environment to find locally optimal conditions [77]. Dinoflagellates are also most abundant in coastal waters, preferring environments with rich nutrients from land or upwelling deep water. These coastal species will often include a resting stage, either as cysts or spores on the ocean floor, in their lifecycle, making them unsuited to open ocean environments. The distribution of dinoflagellates appear similar in the temperate waters either side of the equator. In polar regions heterotrophic dinoflagellates feed on summer diatom blooms, though some photosynthetic species are also present [77]. Some groups have harmful effects on other organisms through production of toxins. Large 'red tide' blooms of dinoflagellates are potentially harmful to humans who consume toxin exposed shellfish. In addition to primary production, dinoflagellates may provide a host environment for symbiont cyanobacteria to efficiently fix nitrogen, contributing to the cycling of nitrogen [78].

Haptophyta include the calcifying group of Coccolithophores, whose blooms can be visible from space (Figure 1.1), due to the characteristic chalky light blue colour given to the blooms by the calcium carbonate scales which cover the cells. Coccolithophores are estimated to provide 20% of total phytoplankton primary productivity, and the sinking of their calcium carbonate shells forms part of the ocean's carbon pump [79]. *Emiliania huxleyi* is one of the most abundant and broadly distributed of the Coccolithophores, with its range having expanded into polar waters since the beginning of the 21st century [80]. Sequencing of the *E. huxleyi* genome showed that strains exhibited a high degree of variability, and these differences in functional potential may explain why the species succeeds in a wide range of habitats [81].

Chloroplastida is a clade which ranges from the smallest known free-living eukaryote *Ostreococcus tauri* to multicellular land plants [82]. The main representatives from this clade in marine environments are the green algae, and among those only the Prasinophyte lineage is abundant in the ocean [83]. Prasinophytes show a high degree of diversity, with variation in cell shape and size, their flagellar apparatus, and their cellular functions. The largest clade of Prasinophytes is the Mamiellophyceae, which includes species which are common among the ocean's picoplankton. Micromonas are a Mamiellophyceae genus with wide distribution across coastal and open ocean environments. They have been observed to be dominant in summer Arctic waters, with one study finding Micromonas to be almost the only organism recovered in the picoplankton size fraction [84]. These polar Micromonas have recently been described as new species, *M. polaris* [85], and the known strains of Micromonas appear to be adapted to specific thermal niches with species *M. commoda* and

M. bravo having two distinct thermotypes [86]. The Ostreococcus are a smaller and simpler genus, lacking the flagellum of many other Prasinophytes. Division of strains of Ostreococcus into 'low-light' and 'high-light' ecotypes has been suggested based on lab evidence [87], though evidence from environmental samples suggests other factors may be driving its global distribution [88]. Bathycoccus is another comparatively simple genus of Mamiellophyceae with a similarly wide distribution across the surface ocean. Recently single-cell sequencing added an additional genome alongside that of *Bathycoccus prasinos*, and suggested that this newer *Bathycoccus sp.* TOSAG39-1 was adapted to deeper waters of the DCM in temperate regions than *B. prasinos* which is more prevalent in tropical waters [89].

Photosynthesis

In both bacteria and eukaryotes, photosynthesis occurs in thylakoids, which in eukaryotes are localised in the chloroplast [90]. The thylakoid membrane contains the pigment molecules which absorb light energy, and surrounds the thylakoid lumen. Four main components embedded in the thylakoid membrane take part in the light dependent parts of photosynthesis: photosystems I and II, cyctochrome b6f and ATP synthase. Water is split into protons, oxygen, and electrons in the oxygen splitting complex of photosystem II, and these electrons are transferred to chlorophyll molecules. When chlorophyll in photosystem II absorbs a photon, the resulting excited electron is moved along the electron transport chain, moving from photosystem II to cytochrome b6f onto photosystem I where additional energy is imparted again from an absorbed photon. From photosystem I, electrons are used in either cyclic or non-cyclic electron transport. Non-cyclic transport moves the electron to an enzyme which reduces $NADP^+$ to NADPH, a key component of the Calvin cycle reactions which fixes inorganic carbon to biologically usable glucose. Cyclic transport transports electrons back to cytochrome b6f, resulting in the movement of protons across the membrane. This creates a different concentration of protons either side of the membrane, which is used by ATP synthase for synthesis of ATP, a key energy source for cellular activities.

The Calvin cycle is the portion of photosynthesis which converts inorganic carbon dioxide and water to biologically available glucose, utilising ATP and NADPH produced in the thylakoid reactions. This is a three step process. First a carbon dioxide molecule is combined with RuBP, subsequently splitting into two 3-PGA molecule. This reaction between RuBP and CO_2 is catalysed by the enzyme RuBisCO, sometimes estimated to be most abundant enzyme on earth though recent research challenges this [91]. The second step uses ATP and NADPH to convert the 3-PGA to the sugar G3P, resulting in $NADP^+$ and ADP as byproducts. Some G3P molecules go to be used for synthesis of glucose, some are regenerated to RuBP to be reused in the cycle.

2.3.3 Microbial Communities

The characteristics and processes of phytoplankton and other microbial species can be individually studied and described, with some species being amenable to growing in isolation under lab conditions. This is far from their usual condition in nature however, where they will co-exist often with a broad range of microbes from other species. There are a few commonly used measures to describe the mixture of organisms in a microbial community. Richness is the number different species present in a community, regardless of their abundance. Two types of diversity are often used to describe communities, alpha and beta diversity, which seek to describe local species diversity and difference in diversity between locations respectively. The total diversity, gamma diversity, was defined as the sum of alpha and beta diversity [92]. A range of different indices have been used for alpha diversity [93], of which richness is one, as well as the Shannon Index [94] and Simpson index [95]. Beta diversity is the differentiation in species between locations, which people have sought to further partition into different components, such as replacement, richness and nestedness components [96]. A similar approach to assessing the difference between locations is using measures of similarity or dissimilarity between samples, based on either binary presence of species or abundances. The Bray-Curtis dissimilarity, Sørensen similarity and Jaccard similarity have been employed for this purpose, and more recently approaches which seek to incorporate information about the relatedness of the taxa observed such as UniFrac distance have been developed [97].

Understanding these natural communities and their diversity necessitates understanding the processes which shape the microbial communities across environmental gradients, spatially, and temporally. In the oceans, many microbes show a highly cosmopolitan distribution, appearing widely dispersed, leading to the Baas Becking hypothesis that "everything is everywhere but the environment selects" [98]. The kind of environmental selection suggested by this hypothesis plays an important role in determining the composition of communities, whereby organisms compete given their fitness for the environmental conditions. Other processes also impact community structure, with some ecologists suggesting the community assembly process to be shaped by diversification, dispersal, and drift in addition to selection [99]. Dispersal is the movement of organisms between locations, and can be either active or passive. Active dispersal, where organisms move themselves to different locations, is very limited for microbes, with flagellate eukaryotes are estimated to have an average swim speed of $186.70 \mu ms^{-1}$ [100] which is dwarfed by typical distances between ocean sampling locations. Passive dispersal is when external forces act to move organisms, such as ocean circulation or winds. The Baas Becking hypothesis suggests unlimited dispersal, with all organisms being dispersed throughout the global ocean. However marine sediments connected by currents but separated by great geographic distance shared several taxa, where more isolated bodies of water did not, suggesting a role for ocean circulation in dispersal [101]. Dormancy strategies, which let an organism be inactive under inhospitable conditions, allow for increased dispersal. Dormant organisms can transit through hostile environments over long time scales [102]. Diversification is the action of evolutionary processes in the organisms of a community. The short generation time of individual microbes allow adaptation to occur on observable timescales, with the Arctic species *Micromonas polaris* shown to adapt to increased temperature within 200 generations under lab conditions [103]. Horizontal gene transfer, the movement of genes between organisms, is believed to be frequent within the oceans [104], allowing the comparatively rapid spread of beneficial traits between organisms. This can allow species to enter new environments, such as ice dwelling diatoms which appear to have acquired genes for ice binding proteins via horizontal gene transfer [105]. Finally, drift covers stochastic changes in the community. In the oceans, some evidence suggests drift is more important in structuring prokaryotic communities than eukaroytes [106].

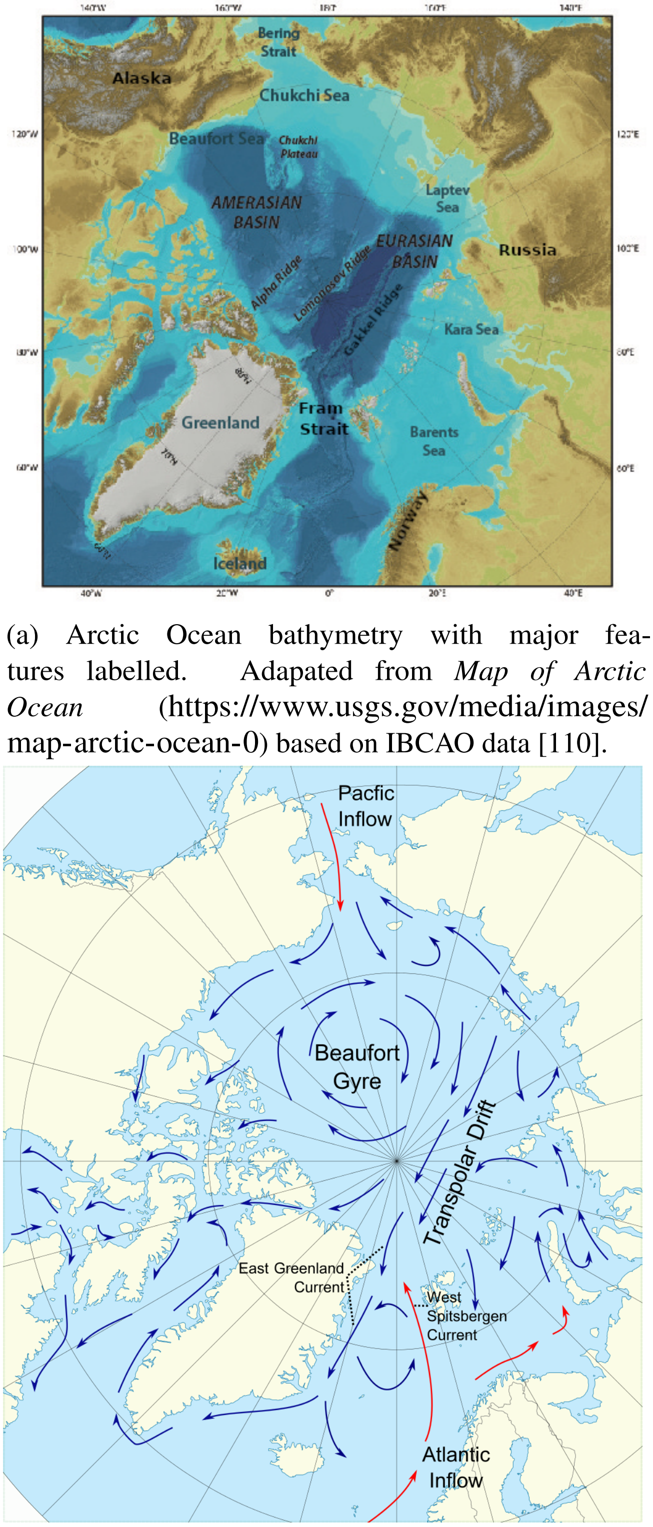
2.4 The Arctic

This section gives an overview of the Arctic Ocean and its dynamics, drawing from the review if Timmermans and Marshall [107]. The Arctic consists of landmasses including Greenland, Northern Russia, and Canada, mostly enclosing the shallow Arctic ocean and its year round sea ice. These terrestrial and marine environments share some unique conditions due to their high latitude. One way of defining the extent of the Arctic is based on sunlight, with the Arctic circle being defined as the lowest latitude at which the sun will not rise on the December solstice, and will not set during the June solstice, which is currently about 66◦33’49.2" N. During polar night the sun does not rise, and the length of this polar night increases with latitude, extending to approximately 11 weeks near the pole. High latitude regions additionally receive less solar irradiance outside of the polar night as light must travel through an increased amount of atmosphere, forming one of the main drivers of the low polar temperatures, and characterstic regions of permanent ice on land and sea.

The Arctic ocean is quite dissimilar to its neighbouring Atlantic and Pacific oceans. Rather than wide deep open areas relatively uninterrupted by land masses, the Arctic is enclosed by land, and characterised by a small number of deep basins surrounded by long shallow shelves, shown in Figure 2.6a. The Lomonosov ridge divides the two large Amerasian and Eurasian basins, each of which is further divided by the smaller Alpha and Gakkel ridges. The shallow Chucki Sea extends from Alaska and Eastern Russia, and the Barents Sea from Norway and Western Russia, and along with Greenland and Canada almost enclose the deeper Arctic waters. Stratification in the Arctic Ocean is driven primarily by salinity, in contrast to primarily temperature driven stratification in the more southern open oceans. Salinity stratified waters show more limited vertical mixing, contributing to the formation of large phytoplankton blooms in the Arctic as growing phytoplankton are not mixed out of the light rich layers of the ocean.

Advection, the movement of water masses into and around the Arctic Ocean, impacts both the physical and biological process of the Arctic, with broad currents shown in Figure 2.6b. Water is exchanged with two neighbouring oceans: with the Pacific Ocean through the Bering Strait, and with the Atlantic through the Fram Strait and Barents Sea. Within the Arctic Ocean, two major currents are the Beaufort Gyre which circulates water of Pacific origin in the Amerasian Basin, and the Transpolar Drift which transports water and ice from the East Siberian and Laptev seas towards the Fram Strait. The narrow Fram Strait between Svalbard and Greenland contains two distinct currents, the West Spitsbergen current which is warmer Atlantic water flowing into the Arctic ocean, and the East Greenland current where cold Arctic water flows into the Atlantic. Warmer inflowing Pacific or Atlantic waters can end up being at depth below colder surface waters due to the strength of the salinity driven stratification. Both nutrients and biomass are carried along with these inflowing waters, with some research estimating that a majority of zooplankton are introduced by advection [108]. Changes in advection are impacting Arctic Ocean conditions, with effects differing between the Eurasian and Amerasian basins. In the past three decades, pulses of warm Atlantic water entering the Arctic Ocean have raised the average temperature, weakening the halocline in the Eurasian Basin, such that it no longer poses as strong a barrier to heat flux from warm Atlantic Water with consequent reduced ice cover. In the Amerasian Basin, warmer inflows reduce ice cover making the warm surface water more susceptible to wind driven transport deeper into the basin, extending the warming effect. Overall as a result of changing advection the Arctic seems to be experiencing a general decrease in nutrients, except in a few areas such as the Amundsen Basin, North Chuchki Sea and Canada Basin [109].

Carmack and Wassmann [108] divide the Arctic Ocean into four ‘contiguous domains’, where areas in the same domain share conditions and processes, and as such are likely to have similar response to changes in climate, which are summarised here. The Seasonal Ice Zone is the portion that experiences seasonal freezing and thawing. This is a widening region as multiyear ice retreats, with later freezing and earlier breakup increasing light availability for microbial communities, and increasing vertical mixing through exposure to autumn storms. As the Seasonal Ice Zone retreats beyond the edge of shelves, there is the potential for the upwelling of nutrient rich waters. The Riverine Coastal Domain is characterised coastal currents that transport of freshwater around the perimeter of the Arctic Ocean. This domain receives low density freshwater input from rivers into higher density ocean water, and is deflected rightward by the Coriolis effect. Atmospheric and terrestrial effects impact this region, such as increased precipitation over land altering volume and content of runoff into the ocean waters. The Pacific-Arctic domain covers the Pacific Waters which enter the Amerasian Basin and circulate within the Beaufort Gyre and Canada basin. In comparison to Atlantic waters, these are lower salinity, higher in nutrient content, and support a distinct biological community. Warmer summer water from the Pacific can affect ice cover, as well as supplying nutrients to support phytoplankton. The pan-Arctic margin domain encompasses the shelf break which extends around the Arctic Oceans from Spitsbergen to West Greenland. Atlantic and Pacific circumpolar boundary currents are contained in this domain, and conditions can vary depending on factors such a shelf depth and width or proximity to river inputs. Despite the variability they share some climate responses, such the potential for increased nutrient upwelling as seasonal ice retreats.



(a) Arctic Ocean bathymetry with major features labelled. Adapated from *Map of Arctic Ocean* (https://www.usgs.gov/media/images/map-arctic-ocean-0) based on IBCAO data [110].

(b) Surface currents in the Arctic Ocean. Adapted from *Arctic Ocean circulation map*, by Zeimusu, 2012 (https://commons.wikimedia.org/wiki/File:Arctic_Ocean_circulation_map.svg). Copyright under Creative Common Attribution-Share Alike 3.0 License.

Fig. 2.6 Maps of Arctic bathymetry and circulation

Sea ice plays important roles beyond influencing Arctic Ocean salinity and stratification. The albedo effect of ice is an important control on both Arctic and global temperature [111]. Ice and snow reflect a high proportion of sunlight, where darker exposed water or land absorb it and lead to increasing heat. The melting of sea ice exposes more absorbent surfaces, leading to a positive Ice-Albedo feedback discussed further in Section 2.5. There is a strong seasonal pattern to sea ice, with the summer sea ice covering approximately a third of the area of the maximum winter sea ice extent.

As with many of Earth's seemingly hostile environments, sea ice provides a habitats for microbial life. Bacteria, archaea and viruses are found in small brine channels in ice, with a high ratio of viruses to bacteria, possibly providing an environment for horizontal gene transfer of traits related to ice resistance [112, 105]. Photosynthetic algae are found at the surface and bottom of sea ice as well as within it. Algal communities differ between these environments, with autotrophic flagellates typifying the surface, diatoms the bottom communities, and the ice interior communities being more mixed [113]. Some algae are incorporated into ice during ice formation, but ice communities differ between surface water and sea ice, suggesting that they are not simple snapshots of the surface water community at the time of incorporation [114]. As with sea ice and light availability, microbial communities in the Arctic show seasonal dynamics. Temperate oceans show two yearly phytoplankton blooms, in spring and later summer, where Arctic waters were typified by a single spring growing period; however decreasing ice cover has led to some regions experiencing a similar pattern of two yearly blooms [115]. Spring blooms are often dominated by diatoms, and despite their short growing season can account for more than half the annual primary productivity [116].

2.5 Climate

Natural and anthropogenic changes in the earth's climate have had, and will continue to have, significant effects on the global oceans and their inhabitants. Increased atmospheric CO_2 is contributing to rising global temperatures, but also to a decrease in ocean pH, known as ocean acidification [117]. Increased ocean acidity leads to reduced concentration of calcium carbonate minerals, an important mineral for the abundant group of calcifying phytoplankton coccoltihophores. A modelling approach showed significant changes in phytoplankton community composition, with acidification having the largest impact on the ecological function of the community [118]. Acidification has been cited as a threat to calcification in species such as the widespread coccolithophore *Emiliana huxleyi* [8]. Adaptive evolution may allow such species to respond to shifting environmental conditions; a study found all cultures exposed to increased atmospheric CO_2 showed decreased calcification, but those grown under such conditions showed improvement compared to those grown in current conditions [119].

Eukaryotic phytoplankton groups play important ecological roles in marine ecosystems, through primary production and cycling of other nutrients. The success and abundance of these species can change in response to environmental conditions, some of which are associated with anthropogenic climate change, with a decreasing abundance of dinoflagellates and rising abundance of diatoms observed over a 50 year period in the North East Atlantic and North Sea [120]. This analysis found changes in community composition were driven by the interaction of changing sea surface temperature and increasing wind during summers.

The Arctic is warming more rapidly than midlatitude regions, over twice as fast and the global average, an effect known as Arctic amplification [14]. A mixture of mechanisms and effects have been suggested to contribute to and result from Arctic amplification [121]. Reduction in sea ice extent may be both a cause an effect, due to the sea ice albedo feedback [111]. The year round extent of sea ice is decreasing at an accelerating rate, and the ice is increasingly younger and thinner compared to multiyear ice [122]. Sea ice and snow are highly reflective, preventing much of the energy from sunlight reaching the Arctic from being absorbed. When this ice melts, more absorbent water is exposed, resulting in a greater amount of energy being absorbed causing heating and further sea ice loss. Sea ice and glacial melt also represent a source of freshwater input into the Arctic, with models predicting this will strengthen the salinity based stratification of the central Arctic ocean, limiting the nutrient supply to surface oceans and hence productivity [123].

With these changes likely to continue occuring, understanding the interactions between microbial communities and environmental conditions is a important challenge in responding to anthropogenic climate change.

2.6 Sequencing

Knowing that the genome contains the information vital for cellular function and the mechanism for transfer of traits between organisms or generations, methods to determine the order of nucleotides within a genome are important for the study of molecular genetics. One of the first widely used methods was commonly called Sanger sequencing [124]. By 2008 Sanger sequencing could generate reads up to 1,000 bases in length with an accuracy per base up to 99.99% [125]. Limitations of this sequencing method include the high cost per base and low throughput. Over 13 years the Human Genome Project assembled the human genome at an estimated cost of 0.5 to 1 billion dollars [126, 127]. New technologies referred to as next-generation or second generation sequencing have been developed with higher throughput and lower cost, and have seen wide adoption.

Sequencing methods produce reads consisting of a sequence of bases observed on a piece of genetic material, either DNA or RNA. These reads are often shorter than the area of interest on the genome, a read may not contain the entirety of a gene we are interested in. Algorithms have been developed to assemble short reads into longer composite sequence (contigs), covered in more detail in Section 3.3.

2.6.1 Sanger Sequencing

Sanger sequencing is not commonly used for metagenomic sequencing. We introduce the Sanger method as a basis from which to explain the high throughput Illumina method used to sequence the samples used in Chapters 4 and 5.

The double stranded DNA to be sequenced is split apart to single stranded DNA using heat. In cells, DNA is replicated by DNA polymerases which bind to single stranded DNA and pairs each base with a corresponding dNTP. Sanger sequencing introduces a small proportion of chain-terminating dideoxyribonucleotide triphosphate (ddNTP) molecules into the medium in which replication takes place. The ddNTP molecules lack the group which allows further nucleotides to be added to the chain, causing replication to terminate. This results in sequences of different length, but which have all terminated at either the base corresponding to that in the ddNTP, or at the full length. Measuring the length of these partially replicated sequences allows us to measure at which positions in the template sequence the corresponding base occurs.

The length of fragments is measured using gel electrophoresis. DNA fragments are placed in a gel medium, and an electric field applied. DNA has a negative charge and moves towards the anode, smaller fragments moving more quickly. Positions at which fragments group can be visualised using various methods, and the positions where fragments stop indicate the positions of the corresponding base. The chain termination step must be repeated four times, once using a ddNTP with each base.

Fluorescent labels were subsequently incorporated into ddNTP, each base fluorescing a different colour [128]. When a laser is applied to fragments incorporating fluorescently labelled ddNTPs, peaks in certain wavelengths of light indicate the presence of a base. Capillary electrophoresis was developed as a miniaturised and parallelised alternative to gel electrophoresis, and combined with automated base calling using fluorescence to meet the demand for increased throughput from projects such as the Human Genome Project [129].

2.6.2 Next-generation Sequencing (NGS)

Next-generation Sequencing (NGS) technologies are a group of sequencing methods which were developed after Sanger sequencing, allowing rapid high throughput sequencing though often at the cost of shorter read lengths [125]. A recent review of next generation sequencing technologies used for sequencing environmental samples found Illumina platforms to be dominant [130]. Illumina produce a range of machines targeted at different uses: MiSeq, iSeq, and MiniSeq for targeted sequencing and small genomes; NextSeq and NovaSeq for high-throughput uses [127]. All of the samples used in this thesis were sequenced using the now discontinued high throughput Illumina HiSeq platform. Maximising the volume of sequence output is advised for metagenomic sequencing, in order to obtain sequences for rare members of the community [130]. The Illumina HiSeq platforms used generate large volumes of short high-quality reads between 100 and 150 base pairs (bp), giving paired-end reads up to 300bp. Read length has continued to improve on new platforms, with NovaSeq generating paired end reads up to 250bp each. The Illumina sequencing method is here explained in detail, and other next-generation sequencing methods covered more briefly at the end of this section.

Bently et al. [131] described sequencing using reversible terminators which is the method used in Illumina devices. Their method shares some principles with Sanger sequencing, using ddNTPs to terminate replication of single strand templates, and fluorescent markers for base calling. Reversible terminators are similar to the ddNTPs used in Sanger chain termination sequencing, except they can be returned to a non-terminating state, allowing DNA synthesis to continue. A plate with many template strands is created, with strands close to each other being clones. One end of the template is bound to primers on the plate. Reversible terminators incorporating all bases are introduced, and will bind to the free end of the template. A single ddNTP will bind, and then synthesis will be prevented as they are terminators. The plate is fluoresced and colours observed for different clusters, indicating which base was incorporated on that cluster of template strands. The plate is washed to remove unbound terminators, then the bound terminators are reversed and the dye removed, and the process repeated. This is illustrated for a single strand in Figure 2.7. This introduces a number of efficiencies compared to Sanger sequencing: terminators for all bases are introduced simultaneously, and does not require a step equivalent to electrophoresis to separate fragments by size. Read lengths tend to be shorter, during the human genome sequencing this gave read lengths of 35bp, though this has been improved over time.

Illumina is the most frequently used next generation sequencing platform, but alternative sequencing platforms are common in earlier metagenomic research. Pyrosequencing was the first next generation sequencing technology to be commercialised [133, 134] by 454

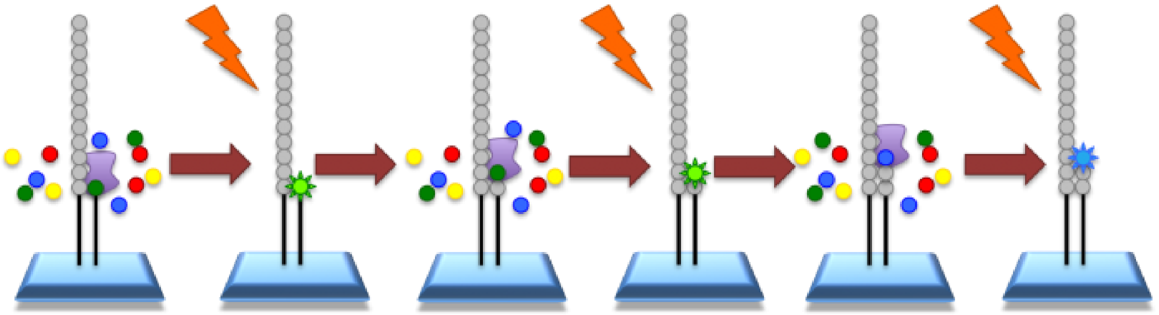


Fig. 2.7 Illumina sequencing method, image from EBI training material [132]. From left to right: A single strand template (grey) has been attached to slide, and fluorescently labelled ddNTPs introduced and the polymerase has incorporated a green ddNTP; unincorporated ddNTPs have been washed away, and the incorporated ones are flouresced; the fluorescent dye and terminator are are cleaved from the ddNTP in position one, ddNTPs are reintroduced and the polymerase incorporates a corresponding one at position two; unincorporated ddNTPs are washed away and the new dNTP in position two is fluoresced; this is repeated to the end of the template sequence. Image used under Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Life Sciences, later Roche. Amplified clonal fragments are captured on beads, and sequence bearing beads are deposited in an array of wells which fit a single bead. Nucleotides are introduced to the array in sequence, with only a single base being introduced at a time. Pyrophosphate is released as the nucleotide is incorporated, the chemistry present creates a burst of light when this occurs. Cameras monitor the wells, and changing light intensity indicates incorporation of a base. The nucleotides are not terminating, so when homopolymers (more than one consecutive identical base) exist on the template, multiple nucleotides could be incorporated. Intensity of light is used to determine the number of bases incorporated, but is prone to errors. The nucleotides are washed away, and a different nucleotide introduced, and this process repeated. Ion Torrent is a subsequent platform using a similar method, but observing the incorporation of nucleotides by measuring pH change due to release of hydrogen ions [135], and is subject to similar homopolymer errors. The 454 pyrosequencing platform has been discontinued, though IonTorrent platforms are still in production.

Unlike the other next generation sequencing methods discussed here, SOLiD is not a sequencing by synthesis approach [136]. Instead oligonucleotides of eight bases are ligated to a template sequence by DNA ligase rather than polymerase. Bases 1 to 3 and 6 to 8 of the oligonucleotide are degenerate, meaning they will pair with any base on the template sequence. The middle two bases are labelled with fluorescent dye, each colour corresponding to a pair of bases, known as two base encoding. After the incorporated oligonucleotide is observed, the final three nucleotides are cleaved, removing the dye and allowing ligation of a subsequent oligonucleotide. After repeating this, bases in position 1, 2, 6, 8... will be observed. Repeating the process with a different offset from the origin of the template strand using a different primer allows sequencing the other positions.

Two methods of preparing libraries of DNA for sequencing are common for environmental samples. whole-genome shotgun sequencing (WGS) sequencing randomly shears the target sequence into smaller fragments, each fragment is then amplified and sequenced. This method was applied to sequence isolated genomes, initially small bacterial genomes [137] and later the large human genome [138, 126]. Environmental samples can be prepared in the same way, generating fragments from across the genomes of all organisms present in the sample [139]. Fewer fragments will originate from species which are rare in the sample, requiring deep sequencing to obtain good coverage of rare species. Amplicon sequencing uses primers designed to select specific regions of the genome using polymerase chain reaction (PCR) amplification, so fragments all originate from the same region of the genome rather than being randomly distributed across the entire genome. Commonly for microbial communities phylogenetic marker genes are selected, as their use for characterising which organisms are present is well studied (see Section 3.4.2). Reverse Transcription PCR (RT-PCR) is a transcriptomics application amplifying target transcripts: the complementary DNA (cDNA) for transcripts is synthesised, and then the cDNA for target transcripts amplified. Real-time Quantitative PCR (pPCR) has become an important method in quantifying gene expression levels, and works by monitoring the RT-PCR amplification reactions, often via fluorescent labelling, allows quantitative measurements of the level at which transcripts are present [140].

2.6.3 Third Generation Sequencing

High throughput with short reads characterises next-generation sequencing platforms. Short read lengths have limitations, for instance making it difficult to resolve repetitive regions [141]. Amplification of template sequences introduces biases, with some sequences replicated more frequently than others [142]. A set of technologies sometimes called third generation sequencing have been developed, producing longer reads without the need for amplification.

Pacific Biosciences Single Molecule Real Time (SMRT) technology [143] is a sequencing by synthesis method, performing synthesis in small chambers called zero-mode waveguides. A polymerase is fixed at the bottom of the chamber, and a template strand and fluorescently labelled nucleotides are introduced. Polymerization occurs continuously, with a camera observing the polymerase. When a nucleotide is incorporated, the time it dwells near the polymerase increases, giving a change in fluorescent signal recorded by the camera. The fluorescent tag is cleaved off during incorporation, diffusing out of the observed area. This increased read length reduced the difficulty of assembling reads to a complete genome, allow assembly of finished genomes for six bacteria using a single library and SMRT sequencing [144]. Error rates are greater than short read sequencing however, with SMRT displaying an error rate of 11-14% [141].

Nanopore sequencing feeds a single template strand through a nanopore in a membrane. As the nucleotides pass through the pore they can be observed [145]. The most successfully commercialised version are the Oxford Nanopore devices [146]. An enzyme moves single stranded DNA through the nanopore, and changes in current across the nanopore as nucleotides move through is used to determine the sequences of bases. Both strands of DNA can be sequenced allowing increased accuracy over sequencing a single strand. The MinION sequencer is a portable device, and has the potential to sequence marine samples in-situ, without flash-freezing and returning to shore. MinION accuracy has increased over the year, with a current claimed accuracy of >99%, and is capable of producing reads up to 2.3 Mbp [147]. Nanopore sequencing has been used for assembly of genomes for eukaryotic microbes [148] and humans [149]. Hybrid methods to assemble combined short and long reads have been developed, using higher accuracy short reads to correct the lower accuracy long reads [150] (Section 3.3.3).

2.6.4 Single Cell Sequencing

Genetic material obtained from environmental samples comes from the whole population of organisms present, and identifying which of the resulting fragmentary reads originate from the same source organism presents a challenging computational problem discussed in Section 3.5. This problem can be bypassed using technologies which allow the sorting and isolating of individual cells from an environmental sample. Amplifying and sequencing biological material, DNA or RNA, from these single cells allows insight into the genome and transcriptional activity of individual organisms from a mixed environmental sample without the need for culturing. A commonly used cell sorting method is fluorescence-activated cell sorting (FACS), a flow-cytometry based method which can identify cells based on their optical properties [151]. Single cells are placed into droplets of fluid which are given a charge based on whether they have the target properties, and these droplets sorted into target and non-target cells by attraction to electromagnets. The range of measurements which can be made from isolated single cells now extends beyond genomic and transcriptomic, including DNA methylation and cell surface proteins among others [58]. Genomes obtained using single cell methods are often referred to as a Single-cell Amplified Genome (SAG). Using these techniques, 30 SAGs were recovered from the Tara Oceans data for eukaryotic microbes, one third of which were absent from the metagenomic sequencing efforts from the same samples, potentially capturing rare taxa other methods may omit [152].

Chapter 3

Bioinformatics Background

3.1 Summary

With the biological and oceanographic scene set in Chapter 2, this chapter turns to the computational elements of metagenomics. Section 3.2 gives a brief history and outline of some of the fundamental goals of metagenomics. Section 3.3 reviews methods of assembling short (or in some cases long) sequence reads into longer, hopefully more informative, sequences. Section 3.4 explores ways of identifying which organism the anonymous sequence fragments of metagenomics originated from, and to estimate the taxonomic composition of the sampled communities. Section 3.5 is about the rapidly developing field of genome-resolved metagenomics, which aims to recover partial genomes from metagenomic assemblies. Section 3.6 looks at identifying genes in sequence data, the first step in moving from sequences to their potential function. Finally, Section 3.7 shows that second step, ways of identifying the potential function of predicted genes.

3.2 Metagenomics

Metagenomics describes a range of techniques used to study the genomes of uncultured microbial organisms in a sample taken from the environment [13]. Next generation sequencing of the genomes of all organisms present in an environmental sample results in short sequence fragments where the species a given sequence originates from is unknown, as well as its position in the originating genome being unknown. This is in contrast to more well established single organism genomics techniques, which sequence a clonal culture consisting of organisms with a single shared ancestor. It is estimated 99% of prokaryotes are unculturable [153]. Most microbial eukaryote lineages have no cultured representative, and

51% lack a genome [154]. The difficulty in culturing these organisms makes their genomic content unavailable to techniques requiring a clonal culture to obtain genomic sequences.

Metagenomic sequencing and analysis allows insight into this unculturable majority, but presents specific computational problems. A common step for isolate genome sequencing is to assemble the short reads into longer sequences. Sequence fragments from an environmental sample cannot be assumed to assemble to single longer genome, having originated from multiple species (for more details see Section 3.3). There are difficulties in applying isolate assembly methods to metagenomic data, and even with adapted methods, assemblies will often remain fragmentary [155]. With or without assembly, identifying which species a sequence originated from is often difficult, imprecise, or impossible.

Two fundamental goals in metagenomics are obtaining taxonomic and functional profiles of the sampled community. A taxonomic profile characterises who is there: which organisms are present and how abundant each is. Functional profiles estimate what the community could do: which genes are present and their abundance. The functional profile can be separate from taxonomic profile, a gene can be detected without knowledge of which species it originated from.

Some studies aim to reconstruct nearly complete or complete genomes from metagenomic data [156]. Algorithms for inferring which sequences originate from the same species have been developed, based on characteristics of the sequences [157] or the abundance of the sequences between samples [158, 159]. Longer third-generation sequencing reads can cover large portions of a genome in a single read. Methods have been created to combine these long reads with higher quality short reads to compensate for the increased error rate [150, 148].

The Global Ocean Sampling expedition [139, 55] was the first large scale attempts to gather metagenomic data from ocean microbe communities. This generated 7.7 million reads making up 6.3Gbp from 41 samples taken from a transect originating in the North Atlantic, through the Panama Canal to the South Pacific. Since then, the number of samples for which metagenome sequencing data is available has continued to increase, as well as the volume of sequence data for each sample. The Tara Oceans expedition [4] took 243 samples from 68 stations, producing 7.2Tbp of data. Sanger sequencing was used by the Global Ocean Sampling expedition, so while the overall volume of sequence produced was much smaller it produced longer high quality reads compared to the short read next generation sequencing platforms used for most of the Tara Oceans samples. Along with publicly available metagenome data from projects with smaller scope, the breadth of marine metagenomic data has continued to increase. The samples analysed in Chapter 4 were sequenced between 2016 and 2019, the smallest of which generated 311 million reads containing 46.79 Gbp, approximately seven times larger than the combined GOS data. This growing body of data has necessitated the development of bioinformatics algorithms which can handle larger datasets on a feasible timescale. Several methods of relating post-processing metagenomic data to environmental parameters are widely used, and discussed in more detail later in Section 5.2.

Metatranscriptomics studies the mRNA transcripts in cells in an environmental sample. Marine bacteria typically contain about 200 mRNA transcripts, which degrade within minutes [160]. Taniguchi et al. [161] found that gene transcripts in a community correlated to the abundance of corresponding proteins, even though this did not hold within an individual cell. Sequencing transcripts from an environmental sample allows insight into which processes are active in a community under the sampling conditions [162]. With both metagenomic and transcriptomic data, the differing abundance of genes and gene transcripts can be used to assess which are being under or over expressed, an approach used to study the response of microbe communities to the Deepwater Horizon oil spill [163].

3.3 Sequence Assembly

Assembling short reads from fragmented DNA into longer sequences, known as contigs, has been well studied and many tools have been developed. These contigs can be further arranged into scaffolds, which consist of contigs and gaps of known length but unknown sequence between the contigs Assembly software developed for use on isolate genomes have been a.pplied to metagenomic data. More recently, assemblers to deal with specific problems of metagenomic assembly have emerged. A summary of assemblers for isolate and metagenome uses is given below, as metagenomic assemblers adapt methods used in isolate assemblers, and some isolate methods can be used for long read third generation sequencing data.

3.3.1 Isolate Genome Assembly

Celera [164] was one of the first commonly used assemblers, using what has become known as the overlap-layout-consensus (OLC) method. OLC compares all pairs of reads, finding overlaps where the end of one read matches the start of another. Overlapping reads are laid so that they are aligned on overlapping regions. Many assemblers approach the layout step by using a graph with a vertex for each read, and edges joining overlapping reads. This graph is searched for a Hamiltonian path (one which visits each vertex once). Locating a Hamiltonian path is known to be an NP-hard problem, implying it is hard to find such a path in practice. There may be positions where bases in overlapping reads do not match; a consensus sequence is generated by selecting most probable bases at each position. This method was developed to meet the needs of the Human Genome project [126], which used Sanger sequencing. Increasing adoption of next generation sequencing produced much shorter reads in greater volume, making the all against all comparison of reads required in the OLC approach computationally demanding [165], having at least $O(n^2)$ complexity where n is the number of reads.

Assembly using deBruijn graph (DBG) methods remove the all against all overlap comparison, and changes the task of finding a path through the graph to the less complex problem of finding an Eulerian path which visits each edge once. Pevzner et al. [166] describe the DBG they implement in the assembler EULER. Reads are divided into subsequences of length k called k-mers. Each $(k-1)$-mer is represented by a vertex in the graph, and a directed edge between two vertices (a,b) exists for each k-mer observed in reads which can be formed by appending the last character of b to a. Genomes contain low complexity or repeated regions, which cause ambiguities in the DBG. Where a repeated region is longer than the value of k, multiple possible paths through the graph exist. Real world sequencing data will contain substitution errors, where the wrong base is called for a position. Substitution errors create ‘bubbles’ or ‘spurs’ where the graph splits for the duration of k-mers containing the incorrect base. Assemblers vary in how they construct, simplify, split and traverse the DBG [167].

EULER implemented several methods of simplifying and resolving graph ambiguities. Errors in reads are filtered by removing low frequency k-mers from the graph. Reads are threaded through the graph to identify which areas of possible repeats in the graph exist in reads. Spurs are removed, and the graph split at boundaries between low and high coverage areas. Velvet [168] uses a heuristic search for bubbles and removes the lower coverage path. Unambiguous paths are collapsed to a single node. Read threading is also used to remove paths representing fewer than a threshold number of reads. Information from paired end reads are used during assembly, attempting to link long contigs joined by paired reads by finding a path between the known ends. SAOPdenovo [169, 170] implements a more memory efficient DBG representation, alongside familiar bubble resolution and spur removal methods. After contigs have been created from the graph, a scaffolding step is performed to join contigs to larger scaffolds with undetermined bases between them using information from paired end reads. The memory requirements for DBG assembly can be high, with greater than 600GB memory required to assemble a human genome [171]. Methods of reducing the memory requirement were implemented ABySS 2 and Minia [171, 172] using bloom filters, and in BCALM2 using minimiser hashing [173].

3.3.2 Metagenome Sequence Assembly

Biological communities commonly contain species of different relative abundance, from very abundant dominant species to very rare species. Consequently metagenome assembly presents a new set of specific problems [155]. Reads from species in a community will be present at different levels of coverage depending on their abundance, so coverage based strategies for identifying repeats are more complicated in a metagenomic context. Assembling reads taken from a community risks creating sequences which originate from the genomes of multiple taxa, known as chimeric sequences. Identical sections will exist in separate genomes, in well conserved areas such as regions coding for ribosomal RNA. Paths from separate genomes will join at these shared regions, giving the possibility for chimeric contigs to be generated by going from an unambiguous region into a shared repeat and exiting to region from a different genome. Coverage can distinguish between some taxa, as an abundant taxon will have greater coverage than a rare one. Rare taxa have similar low abundances and coverage level is unlikely to be distinguishing for these taxa. Capturing rare taxa requires generating enough reads that a detectable amount originate from the rare community members. Greater volumes of sequence data comes with a greater number of errors, creating large and complex DBGs.

Assembly software specialised for metagenome assembly take differing approaches to addressing some of these problems. IDBA-UD [174] and Megahit [175] partition the DBG into separate graphs where neighbouring vertices have significantly different coverage. MetaVelvet-SL takes a machine learning approach, using a support vector machine (SVM) trained to classify chimeric vertices in the graph. Megahit uses the memory efficient succinct DBG representation [176] to keep the memory requirement manageable given the increased volume of sequence data. Preprocessing techniques are incorporated in some pipelines to simplify the assembly problem, Meta-CRAM [177] identifies and removes reads originating from reference genomes, and assembles only the unclassified reads. SPAdes [178] is an assembler which was intended for assembling single-cell sequencing data, but the pipeline metaSPAdes [179] covers metagenomic assembly incorporating heuristic methods for estimating repeated regions shared between genomes.

Third generation sequencing technologies produce long reads with lower throughput, making this type of data more amenable to OLC assembly. Canu [144] is a successor to the Celera assembler adapted to handle the longer lower quality reads generated by current long read technologies. A popular alternative is Flye and it’s metagenomic counterpart metaFlye, based on constructing a graph representing repeat sections in misassembled “disjointigs” and resolving identified repeats in the graph using long reads [180, 181].

3.3.3 Hybrid Assembly

Long reads are ideal for spanning complex genomic features but can have higher error rates which has led to hybrid assembly approaches, combining the strengths of long and short reads. In approaches such as hybridSPAdes [150] short reads are converted to a DBG, and long reads are mapped to the graph and used for closing gaps and resolving repeats in the graph. Another hybrid approach is to start by assembling long reads, and subequently using the higher quality short reads to correct errors in the long read assembly through tools such as Pilon [182]. While not designed initially for metagenomic use, hybrid assembly approaches have been integrated into metagenomic analysis pipelines such as the nf-core MAG pipeline [183] and MUFFIN [182]. Combining long and short read technologies was shown to able to improve aassembly for difficult to assembly phylotypes in biogas reactor samples, improving contig length 118% [184].

3.4 Taxonomic Classification

Taxonomy is a discipline dealing with the naming and classification of biological organisms. Species are distinguished using multiple methods, including comparison of biological sequences or morphological features [185]. Classifications can be unstable and subject to revision, with researchers in different areas adopting differing ways to delineate species [186]. Several organisations maintain curated hierarchical taxonomies, such as NCBI, SILVA, and Genome Taxonomy Database (GTDB) [187], though the placement, inclusion and naming of groups varies between taxonomies [188]. Hierarchical taxonomies are often represented as trees (Figure 3.1), where some vertices are assigned a named taxonomic rank. These ranks are ordered groupings of related organisms, going from distant relationships between members of the same domain to close relationships between members of the same genus.

Taxonomic classification methods aim to estimate which organisms are present and in what quantities in the sampled community from sequencing data. These approaches can be broadly divided into two camps: those which compare query sequences to reference sequences with known origin, and those which look only at query sequences and aim to group sequences from the same species without assigning a taxonomic label.

3.4.1 Reference Based

Reference based approaches look for similarities between query sequences and those in a reference database of sequences with a known origin. If the query sequence is similar enough to a reference sequence, it could be considered as coming from that or a closely

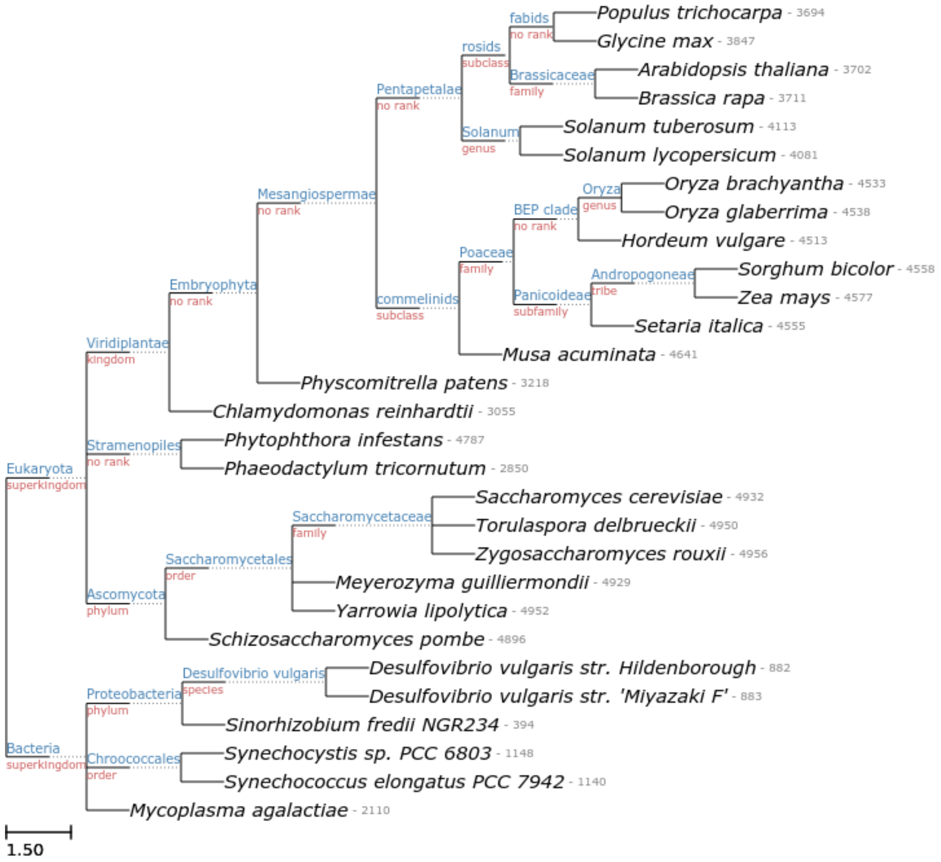


Fig. 3.1 Subset of the NCBI Taxonomy visualised as a tree by ETE Toolkit [189]. Red labels are the rank assigned to that node, blue and black labels are names assigned to the node.

related organism. These methods rely on the availability of well curated reference sequences. Many general databases exist such as NCBI RefSeq [190] as well as databases for specific environments such as MarRef and MarDB for marine prokaryotes [191]. GenBank collects all publicly available nucleotide and protein sequences, and maintain extensive databases combining these and the more curated RefSeq data, usually referred to as “nt” and “nr” respectively [192].

There are biases in what is included in these databases. Culturable organisms are overrepresented, along with organisms from more often studied environments such as the human gut microbiome [193]. Fewer reference genomes are available for marine eukaryotic than prokaryotic microbes, it is less likely that a close relative of a eukaryotic query sequence will be present in a reference database.

3.4.2 Marker Genes

Well studied genes or parts of the genome are often used as reference sequences rather than complete genomes. Variations in genes shared between species can be used to resolve which lineage an unknown sequence is most closely related to. These are often referred to as *marker genes*. Genes coding for parts of the ribosome are often used as phylogenetic marker genes [194], and databases of these marker genes are available such as SILVA [195].

The ribosome is mostly composed of rRNA divided up into two parts, the small subunit (SSU) which reads mRNA and the large subunit (LSU) which assembles amino acids. Protein synthesis is essential to cell function and so the genes coding for rRNA are ancient and shared across many organisms, and well conserved between species. Within the SSU, a smaller rRNA subunit is commonly used as a marker gene, the 16S subunit in prokayotes and 18S subunit in eukaryotes. We call the gene which codes for an rRNA subunit rDNA.

Hills et al. [196] review reasons rDNA has been widely used in phylogenetics. Genomes often contain multiple copies of rDNA. Compared to single copy genes, rDNA sequences vary little within a species, but display difference between species. This means relatively few samples from a species are required to characterise the rDNA of the species. The 16/18S SSU rDNA evolves more slowly than the LSU, and contains a combination of well conserved and hypervariable regions. The well conserved regions make it possible to design primers for PCR amplification of the desired hypervariable region. The hypervariable regions evolved more rapidly, at a different evolutionary rate for each region. Varying rates of evolution allow the evolutionary history between organisms to be discerned at different points in history using different hypervariable regions. Comparing these hypervariable regions can be used for taxonomic classification of metagenomic reads.

Reads from environmental whole-genome shotgun sequencing (WGS) sequencing can be queried against an rDNA reference database, but a majority of reads will not be from this region. In short read WGS sequencing data from three marine samples taken by the Tara Oceans expedition, between 0.01% and 0.1% of reads were 16S SSU [197]. Amplicon sequencing can target the SSU region of the genome to produce reads mostly from the SSU rDNA, or specifically from the hypervariable regions in the gene.

Limitations when using SSU amplicons for taxonomic classification of metagenomic data have been documented. Logares et al. [197] found that while 16S rDNA reads made up a small portion of the reads in environmental WGS, these fragmented reads identified 61% more OTUs than long read sequences from 16S amplicons. Another study compared 16S SSU sequences assembled from publicly available environmental WGS metagenome data to amplicon sequences, and found a minimum of 9.6% of sequences assembled from metagenome data were not present in amplicon data [198]. They found sequences which were missed by amplification tended to come from newly described baceterial lineages such as Candidate Phyla Radiation, or in archaea from outside the currently recognised phyla. These studies suggest the existence of unexplored diversity which is not captured by SSU amplicon sequencing.

Conversely, other research found that less than 50% of phyla identified in 16S SSU amplicon sequencing were recovered in metagenomic sequencing data [199]. They compare metagenomic reads to whole genome reference databases, and they note some phyla have no reference genome while 16S SSU reference sequences are available. Comparing metagenomic reads to a database of 16S rDNA reference sequences was shown to poorly recover the composition of a synthetic community [199]. They suggest this could be due to the varying level of taxonomic resolution of the hypervariable region the short reads covered, as well as observing a bias towards over representation of sequences with low GC content.

3.4.3 Sequence Similarity

Local alignment techniques seek to find subsequences which are similar in a pair of sequences. Newly obtained query sequences can be compared to a database of reference sequences, and can be used to infer taxonomic origin of a query sequence. If the query sequence has a subsequence which is similar enough to one of the reference sequences, the query sequence may come from an organism closely related to the reference sequence.

Sequences which are similar due to some shared evolutionary history are said to be homologous. Different terms are used for homologous sequences based on the origin of the homology [200]. Orthologs are homologous as they evolved from a common sequence in a shared ancestor. Paralogs are homologous due to duplication of a sequence within a genome, which could have evolved since duplication in an ancestral genome. Xeonologs are homologous due to the transfer of a sequence between species, referred to as horizontal gene transfer.

The Smith-Waterman algorithm [201] is a dynamic programming algorithm which locates the highest scoring local alignment. Gaps are allowed, representing positions in the alignment where a base has been inserted or deleted (indels). The scoring system is defined by a substitution matrix giving the score for observing a pair of bases in the same position, and a function for scoring gap length. This approach is computationally expensive and searching increasingly large sets of sequencing data against increasingly large reference databases has become impractical, prompting the development of heuristic local alignment methods.

BLAST

Basic Local Alignment Search Tool (BLAST) [202, 203] is one of the most widely used heuristic local alignment tools. For each query sequence, BLAST starts by looking for *words* within the query sequence which score above a threshold value t when compared to words in a reference sequence, using a similar scoring system to Smith-Waterman algorithm. A word is a subsequence of characters of a specified length found in a sequence. When two highly scoring words are within a specified distance, an extension step is triggered. This extension step uses a heuristically constrained version of the Smith-Waterman algorithm, restricting the search space by not allowing the score to drop below a threshold during extension. BLAST provides several metrics of alignment quality: percent identity, bit score and evalue. Percent identity is the percent of bases which match in the aligned section. Bit score is a normalised version of the score generated by the scoring system specified for the extension step, taking into account the statistical properties of the scoring system. The e-value is derived from the size of the database and the bit score, and is the number of alignments with this bitscore which would be expected by chance in a database of this size.

BLAST is a family of programs facilitating multiple types of search. BLASTN searches nucleotide query sequences against nucleotide reference sequences. Nucleotide sequences can be translated to protein sequences and searched against a protein database using BLASTX, and the reverse process translating proteins to possible nucleotide sequences via TBLASTX [202].

Other Heuristic Local Alignment Methods

Large scale sequencing projects have motivated the development of faster local alignment algorithms. BLAST Like Alignment Tool (BLAT) was developed in response to the need to align many short reads against the large human genome [204]. Where BLAST indexes the query sequence, BLAT indexes the reference sequence, to allow rapid search of many query sequences against the reference. This is suited to rapid alignment against a small number of large reference sequences, but creating and keeping many indices incurs high memory usage, making this less applicable for metagenomics. LAST [205] finds initial matches to extend based on rarity of sequence rather than a fixed metric such as score or length. This helps locate seeds within repetitive or low complexity areas which are prevented from participating in initial matches in BLAST. YASS [206] locates seeds taking into account the observed proportion of transitions and transversions of related biological sequences.

Short read aligners are intended to align a high volume of short sequencing reads against longer sequences, either reference genomes or assembled contigs. Aligning reads against contigs allows an estimate of how many reads originate from the assembled sequence, and consequently an estimate of the abundance of the taxon or genes on the contig. Bowtie [207] and BWA-MEM [208] are commonly used, with BWA-MEM showing faster computation on longer reads than Bowtie. Psuedoalignment originated in the analysis of transcriptomic data, to calculate which reference transcript a read originated and infer the abundance of those transcripts from without performing a full alignment. Kallisto [209] implements a fast, low memory psuedoalignment algorithm, and has been demonstrated to be applicable to quantification problems in metagenomics [210].

Hidden Markov Models

Detecting homologous sequences can be treated in a probabilistic manner using a Hidden Markov Model (HMM) [211]. A HMM models a process which creates a sequence of observable events, such as sequence of amino acids in a protein, by moving through a series of hidden states over time where each state has a probability of emitting a given symbol. Time in the context of biological sequences is position in the sequence. Transition between the model's hidden states is a Markov process, meaning that the probability distribution of which hidden state the model is in at time t is dependent only on the state at the previous time step $t-1$. A HMM is defined by the following elements: a set of hidden states S; for each state $s \in S$ the emission probability $e(x,s)$ that symbol x will be emitted by state s; for each $s_1, s_2 \in \binom{S}{2}$ the transition probability $t(s_1, s_2)$ of moving from state s_1 to s_2; the initial probability $i(s)$ of the model starting in state s for all $s \in S$. This underlying Markov chain allows the recursive definition of several useful probabilities in terms of these parameters.

Given a HMM H representing a set of sequences from the same group, we may want to know the probability that model H generated sequence x. This is the observation probability of x given H. The dynamic programming forward algorithm calculates the observation probability in $O(ls^2)$ time, where l is the length of the query sequence, and s is the number of hidden states in the model [211]. Query sequences with a high probability of being generated by the model are likely homologous to the group of sequences the model represents. A similar question is which sequence of hidden states has the highest probability to generate observed sequence x. This can be found using the Viterbi algorithm [212].

Calculating these probabilities requires a model with all the parameters specified. There are methods available to fit the emission, transition and initial probabilities to a set of observed sequences. The set of states used varies depending on the model's intended use, and several types of HMM are used in analysis of sequences. Profile HMMs use a linear sequence of states, with three types of state representing a match, insertion, or deletion event [213]. A state of each type is created for every position in the sequence, and probabilities for emission and transition can be estimated from a multiple sequence alignment by the frequency of symbols and gaps at positions in the alignment. Profile HMMs for protein families have been trained by Pfam [214] and can be searched using HMMER [215, 216] through a web interface or local installation.

Interpolated Markov Models (IMMs) are used in Phymm [217] for taxonomic classification. An IMM is trained on genome sequences from the taxonomic group of interest, which could be species or a higher grouping. Emitted bases are predicted by the IMM based on a varying k number of preceding bases. The varying number allows prediction based on preceding k-mers for which reliable probabilities can be generated from the training data. This approach is useful for metagenomic taxonomic classification, as a model can be trained on the single or small numbers of possibly fragmented reference genomes. For a query sequence observation probabilities are generated for each IMM in the database, and the sequence given the same taxonomy as highly scoring models. Phymm showed results comparable to BLAST, and a method combining high scoring IMMs and BLAST searches showed improved performance over either method alone [217].

Exact *k*-mer Matching

Generating local alignments or finding observation probabilities against a large reference database is computationally demanding. Wood et al. [218] proposed an algorithm to exactly match k-mers in metagenomic reads to k-mers found in reference sequences. A database is created containing each k-mer found in the reference database, and the taxonomic label of the sequence containing that k-mer. Where a k-mer exists in more than one reference sequence with different taxonomic labels, the database stores the Lowest Common Ancestor (LCA) of all the reference sequences in which the k-mer is found. Reads are split into k-mers and the LCA for each retrieved from the database. A taxonomic label is assigned based on a tree formed by restricting the general taxonomic tree to only those nodes contained in the matching *k*-mers, and giving each node a weight equal to the number of *k*-mers which had this taxonomic label. The path from leaf to root with the greatest sum of weights is assigned as the taxonomic label for the read.

Fast lookup of the LCA for a *k*-mer is achieved using minimizers to group similar *k*-mers together, as similar *k*-mers are likely to be queried one after another so may already be in CPU memory. The complete database needs to be held in memory and can be large, with the default database being 70gb. Kraken's accuracy is slightly lower than BLAST when classifying to genus level, but is between 150 and 240 times quicker than other comparable methods which attempt to classify all reads [218]. This speed makes *k*-mer matching approaches appealing for large metagenomic data.

The *k*-mer matching approach has been extended to estimate abundance at different taxonomic levels by Bracken [219], by using probabilistic approaches to distribute reads assigned above the desired rank to the nodes below. Kaiju [220] translates genomic sequences to proteins and searches against a database of proteins. Proteins tend to be more conserved than genome sequences, allowing the detection of more distant homologs. CLARK [221] implements *k*-mer matching classification aimed at improving accuracy of assignment at the genus or species level by removing *k*-mers shared between groups and classifying only using the remaining discriminative *k*-mers.

3.4.4 Assigning Taxonomy from Sequence Similarity

Taxonomy can be assigned in different ways from measures of sequence similarity. Phymm [217] assigns the query sequence the same taxonomy as the highest scoring reference IMM, Metaxa 2 [222] considers the top 5 BLAST alignments and their length and percent identity them to assign a taxonomy with a reliability score. The LCA algorithm assigns a taxonomic label based on multiple local alignments for a query sequence [223]. In this method, where local alignments above a user defined threshold are found against reference sequences with different taxonomies, the query sequence is classified as the first ancestor shared by all reference sequences. This approach is quite conservative and results in many reads being assigned at high ranks rather than more specific species or genus ranks. A weighted version of the LCA algorithm weights each reference sequence based on how many reads have a significant alignment to the sequence, and then placing each read on the taxonomic node such that the node and its descendants account for 75% of the significant alignments for the read [224]. Taxator-tk [225] uses the proportion of matching bases among local alignments for a query sequence to assign taxonomy.

3.4.5 Taxonomy Without a Reference Database

Reference free approaches seek to group sequences originating from the same operational taxonomic unit (OTU), but do not assign a position in a taxonomic ordering to the identified OTUs. These methods may identify taxa which lack a close relative in reference databases. The identified OTUs and their abundances can be used to calculate measures of diversity within and between communities without a taxonomic placement [13]. Mothur [226] clusters rDNA amplicon sequences to OTUs using several distance based clustering methods on a sparse matrix of sequence dissimilarities. PhylOTU [227] identifies OTUs from shotgun metagenomic sequencing rather than amplicon sequencing, their results identified novel taxa which had been under-represented previously due amplification bias. With the decreasing error rates of sequencing techniques, some studies have forgone the clustering steps involved in OTU analysis, and instead used exact amplicon sequence variants (ASVs) [228].

3.4.6 Pipelines

Taxonomic classification methods often requires several steps before arriving at estimates of abundance, leading to the development of pipelines which incorporate tools to perform each step targeting different situations. Metaxa [222] uses HMMER & BLAST to locate rDNA sequences in shotgun metagenomics data and predict their origin at a high level, and assign a taxonomic classification using BLAST search results against the SILVA database. QIIME [229] by default uses the RDP naive Bayes classifier [230], as well as providing other options such as BLAST based assignment. MG-RAST [231] also provides a variety of approaches, including searches against rDNA databases with BLAST, and phylogenomic reconstruction using information from the SEED database [232]. SHOGUN [233] is targeted at shallow metagenome sequencing data, and uses short read aligners and an implementation of the LCA algorithm to assign taxonomy. The JGI Integrated Microbial Genomes & Microbiomes (IMG) pipeline [234] which was used to process the data used in Chapter 4 provides taxonomic classification for each sequence by finding each gene in the sequence and labelling that with the taxonomy of the top BLAST result for the gene, and labels the sequence with the LCA of all genes on the sequence.

3.5 Metagenome Assembled Genomes (MAGs)

Assembling metagenomic reads results in contigs originating from a mix of genomes in the sampled community. Given the difficulty of culturing the majority of microbes, studies have sought to find ways to group contigs into ‘bins’ where contigs originating from the same species are placed in the same bin. Each bin is referred to as a Metagenome Assembled Genome (MAG). These draft genomes place contigs and their genes into a genomic context, providing an estimation of the metabolic potential of members of the community, as well as characteristing the community as collection of genomes in addition to a collection of genes [21]. Even after binning the genome of an organism from the community is still likely to be fragmented and incomplete, represented by multiple contigs and missing regions.

Large numbers of prokaryotic MAGs have been recovered from ocean metagenomes [156, 19]. Prior to publishing our research, there had been no similar large scale recovery of eukaryotic MAGs from ocean environments, but methods had shown success recovering eukaryotes from human gut samples [235, 236]. Two studies had recovered single eukaryotic MAGs for Prasinophytes from ocean data, among other prokaryotic MAGs; the first a species of Micromonas [237], the second a Bathycoccus [20], both from polar samples. Since then, research recovering MAGs on a large scale from the TARA Oceans data has been published, with one study reporting the recovery of 713 genomes from a combination of metagenome binning and single-cell sequencing techniques [152], another 988 MAGs from binning of data from multiple depths [238].

3.5.1 Binning Methods

Binning methods can be divided into two broad groups: those looking at the composition of contigs, and those looking at the coverage of contigs in multiple samples. Frequencies of tetranucleotides (4-mers) are known to vary in different sections of microbial genomes [239] and between genomes from different clades [240]. One approach [241] used the unsupervised learning technique of emergent self-organising maps (ESOM) to group contigs with similar tetranucleotide frequencies together. This method located 87 bacterial bins, 21 of which appeared over 90% complete. A different approach combined a *k*-means like clustering method with IMMs to bin sequences, using sequences as the data points being clustered, and an IMM trained on cluster members as the cluster representative [242]. This method requires specifying the expected number of clusters *k* before clustering, meaning binning may have to repeated for different values where estimates of *k* are not easily obtained, as is likely to be the case for environmental samples. LikelyBin [243] used a Monte-Carlo Markov Chain to estimate the master distribution of nucleotide frequencies which generated a read, and separate reads based on the estimated parameters of the distribution. One limitation encountered in their results was that more closely related organisms with similar nucleotide frequencies are poorly separated, with little separation achieved between two species of Streptococcus.

Differential coverage uses the coverage of contigs in multiple samples for binning. Organisms are assumed to be present at different abundance in different samples or samples from different locations. Coverage of contigs from the same genome can be assumed to covary between samples. This differential coverage approach was used to obtain 31 bins based on sequencing the same sample using two different DNA extraction methods, creating differing patterns of coverage which could be plotted against each other [244]. These bins were further refined using tetranucleotide frequencies to attempt to separate out species within these broad bins. Many tools to handle differential coverage binning based on more than two coverage profiles now exist. CONCOCT [245] bins all contigs using coverage and sequence composition, potentially creating many small bins requiring manual curation. CONCOCT is provided as default option in anvi'o, a meta-omics analysis platform [246]. Anvi'o further includes tools for visualisation of contig coverage in bins generated by CONCOCT or other tools, and allowing manual refinement based on this. MetaBat [157] seeks to produce fewer, higher quality bins using coverage information. Using the coverage based binning tool BinSanity, thousands of draft prokaryote genomes were recovered from coassemblies of the Tara Oceans data [156]. The consensus based method DAS Tool combines multiple binning tools, which generated prokaryotic consensus bins with results improved over any single binning algorithm [247].

Coverage of contigs is generated by aligning reads back to the assembled contigs, often referred to as short read alignment. The high volume of short reads obtained by next-generation sequencing (NGS) platforms means local alignment algorithms such as BLAST would be too slow for solving this alignment problem. Two broad approaches have been taken: hashing, and Burrows-Wheeler transform (BWT) based methods [248]. Hashing based methods locate seed alignments where a read and reference sequence have a short exact or very close match, and these seed alignments are extended. In a simple example, seeds are *k*-mers which match in the genome and read. Either the genome or reads can be indexed for all *k*-mers, and the locations where each *k*-mer occurs stored in a hash table. The other sequences can be scanned for *k*-mers in the same way, and locations of matching *k*-mers looked up in the hash table. These seed alignments can then be extended. BWT based methods include some of the most commonly used tools, such as Bowtie2 [207] and BWA [208]. The BWT of a string is a reversible permutation, which has often been used for compression. Using an FM-index [249] allows quick lookup of substrings as seed matches with comparatively low memory requirements.

Pseudoalignment tools such as Kallisto [209] have been developed recently, intended to deal with transciptomics problems. The transcriptome is the total RNA present in a cell, and when this RNA has been sequenced, a common task it to identifying how many of the resulting reads maps to each known gene sequence in that organism. Pseudoalignment methods seek to identify reference sequences a read could have originated from, without performing an exact alignment. In a metagenomic context, this allows an estimate of how many reads map to each contig, but not where on the contig the read aligns. In the case of Kallisto, this is achieved by creating a DBG representing the transcripts. A read can be represented as a path in the graph to indentify transcripts which the read is compatible with. Pseudoalignment coupled with Expectation-Maximization algorithms has been found to perform quickly and accurately for metagenomic read assignment [210]. The primary advantage of pseudoalignments for meatagenomic binning is the reduced computational costs in both time and memory requirements. Estimates of variance in coverage across a contig are not available from these tools however, which are used in some binning tools, while others such as Metabat [157] can use this simplified coverage information.

Outputs of these binning tools can vary due to both differences in algorithm, but also due to tools having slightly different approaches; some seeking to identify only good quality bins, some binning a greater porportion of the assembly, some aiming to generate potentially contaminated bins for further manual refinement. Consensus type approaches have been developed to combine the results of multiple binning algorithms, to harness the strengths and offset the weaknesses of invidual methods. DAS Tool (Dereplication, Aggregation, Scoring) scores bins from multiple binning algorithms using a scoring function based on presence and duplication of single copy genes [247]. Binning_refiner uses BLAST to seek similarity between contigs of two binning algorithm outputs, seeking to reduce contamination [250]. MetaWRAP uses Binning_refiner as a step, first splitting bins from multiple methods into low contamination variants, and selecting the best of the variants using completeness and contamination scores (see Section 3.5.2). The selected bin is then reassembled using the reads which map back to the bin assmbled using an isolate genome assembler SPAdes [178]. Both DAS_Tool and MetaWRAP use the prokaryote specific quality assessment tool CheckM as part of their pipeline, so are unsuited to eukaryotic MAGs.

Where individual assemblies are used rather than co-assemblies (see Section 3.5.3), this can result in a large number of MAGs which are highly similar, representing closely related strains present in several samples. De-replication seeks to identify MAGs which are the same, to a certain threshold, and select a single best representative for this group [251]. This steps reduces the size of data for subsequent analysis steps, but retaining highly similar MAGs can allow pangenomic and strain levels analyses.

3.5.2 Quality and Taxonomy

Quality of genome bins is usually expressed in terms of completeness and contamination. Orthologs to a list of genes which are present in a single copy in nearly all members of a taxonomic group are sought in the bin, and completeness expressed as a percentage of the expected single copy genes found. Completeness is assessed by identifying the proportion of orthologs for these single copy genes present in the bin [252]. Single copy genes for which more than one ortholog exist in the bin are possible contamination. Contamination is the percentage of the single copy genes which have two or more orthologs in the bins. Bins can be both highly complete and highly contaminated, suggesting the bin contains contigs from more than one genome. The most commonly adopted standard for reporting quality information about MAGs and Single-cell Amplified Genomes (SAGs) are minimum information about a metagenome-assembled genome (MIMAG) and minimum information about a single amplified genome (MISAG) respectively [253]. Under these standards, a medium quality MAG is one which may be composed of many short fragments, with $\geq 50\%$ completeness and $\leq 10\%$ contamination. The standards for a high quality MAG are $\geq 90\%$ completeness and $\leq 5\%$ contamination, but also requires identification of some additional elements. A high quality MAG requires presence of 23S, 16S and 5S rRNA genes, and a minimum of 18 tRNAs, though for eukaryotic MAGs the 18S rRNA gene may also be considered important. In addition to these quality statistics, the standards also suggests reporting standard assembly statistics such as N50, L50, maximum contig length etc.

For prokaryotic MAGs, CheckM [254] is commonly used to characterise quality. CheckM selects a suitable gene set specific to the lineage of the MAG rather than a broad universal set, based on the taxonomic classification it performs. The method also looks for marker sets, which consist of consistently collocated single copy genes, rather than individual marker genes. Collocated genes are likely to be retrieved together, and may give an overestimate of completeness when each genes is counted individually. EukCC [255] provides a tool with a similar aim as CheckM for eukaryotes, aiming to select a suitable set of marker genes for a MAG to provide a lineage specific estimate of quality. Lineage in EukCC is estimated based on set of 55 widely occuring single copy genes. Sequences for these genes from MAGs are placed on the tree using pplacer [256], and the LCA of all these placements selected as the appropriate lineage, and a more specific set of marker genes used for quality assessment. Initially EukCC used ab initio eukaryote specific gene prediction tool Genemark-ES [64], but has subsequently adopted metagenome specific gene prediction too MetaEuk [257].

Both EukCC and CheckM estimate lineage to select an appropriate marker genes set. However usually it will be of interest to look for more specific taxonomic identifications, as well as establish the relationships between recovered MAGs. For prokaryotic MAGs, the GTDB and associated Genome Taxonomy Database Toolkit (GTDB-Tk) has become established as a standard approach for taxonomic identification [187]. The GTDB-Tk uses sets of marker genes for bacteria and archaea, with matches identifed using HMMER [216]. Domain is decided based on which domains marker genes have the highest number of matches, and are then placed onto a domain specific reference tree using a concatenated alignment of the marker genes using pplacer [256]. Using this placement, species level taxonomies are assigned based on Average Nucleotide Identity (ANI) to references, or where placed higher in the tree Relative Evolutionary Divergence (RED) [258] is used to resolve ambiguous placements. A similar standard for eukaryotes has not yet emerged, but tools are available aiming to taxonomically identify eukaryotic MAGs. EUKulele [259] aims to provide a straightforward process for taxonomic identification of eukaryotic MAGs, using similarity search against a user defined database, providing default options including the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) database [260].

3.5.3 Assembly and Co-Assembly

Prior to binning, a choice needs to be made on whether to assemble and bin each set of reads individually, or to pool reads and use a co-assembly. The volume of metagenomic samples being sequenced is expanding, and handling the resulting sets of reads individually can be computationally expensive. In particular, aligning reads back to an assembly is time consuming and generates very large output files, and seeking to align all sets of reads back to the assemblies grows exponentially. Co-assembly is an approach which pools sets of reads from similar environments, such as ocean basins, and generates a single assembly from the pooled reads. This approach has been used for large-scale analyses which cover the global ocean [2, 238], reducing the computational costs of analysis steps following assembly. There are concerns that co-assembly collapses strain level variation, as the high proportion of shared sequence generates complex assembly graphs in which long unambiguous paths cannot be found [261, 251]. This poses problems for genome resolved approaches, as recovered genomes may be a combination of related strains. Culturable species of phytoplankton such as *Emiliania huxleyi* have been observed to have a large degree of genomic and functional variability within the species [81]. Co-assembly risks conflating or discarding this species level variation. Quality of MAGs was observed to be lower for co-assemblies when evaluated in gut samples [251]. This same study suggested methods for de-replication, a step aiming to select a single best representative from among highly similar MAGs generated from closely related organisms in different assemblies. This simplifies downstream analyses, but MAG based pangenomic analyses are becoming a practical way to examine the variation within these similar genomes, with adjustments and tools emerging to compensate for the absence of genes due to incompleteness or contamination [262–264].

3.5.4 Eukaryotic MAGs

A majority of binning studies have focussed on retrieving prokaryotic genomes. Prior to beginning this thesis, few studies aimed to retrieve draft eukaryote genomes from metagenomic data [236, 235, 20], with some using a preprocessing step to predict eukaryotic contigs before binning. One approach used a linear SVM trained to predict eukaryotic contigs [235], implemented in the package EukRep. They retrieved 4 eukaryotic genome bins with greater than 80% completeness from samples taken from a freshwater geyser. The same techniques were applied to more complex samples taken from infant gut and neonatal intensive care unit surface swabs [236]. With deep sequencing and a large number of samples, fourteen novel eukaryotic genomes were identified with median 91% completeness. Development of this pre-filtering approach has been continued, with the tool tiara utilising deep learning techniques, and aiming to both identify eukaryotic sequences, and further classify those into nuclear, plastidial and mitochondrial origin [265]. This showed similar accuracy to EukRep for nuclear eukaryotic sequences, but greatly improved identification of organellar sequences, which are vital to cellular processes of interest in the ocean such as photosynthesis. The differing gene structure between eukaryotes has also been used a signature to discern the domain of origin by whokaryote [266], using gene density and intergenic distance as features for a random forest classifier. This obtained performance slightly lower than tiara, but a model with the tiara prediction used as an additional feature appeared to perform better.

The first eukaryotic MAGs for an ocean microbe was recovered from binning of metagenomes from the Amundsen Sea in the Antarctic [237]. This study generated a single eukaryotic MAG, *Micromonas* sp. ASP10-01a, of estimated 93% completion, among a much higher number of prokaryotic MAGs. A second study targetted recovery of Bathycoccus MAGs from samples taken in the Beaufort Sea, representing the first recovered Arctic eukaryote MAG [20]. Subsequently, eukaryotic MAGs have been extracted from the TARA Oceans by two separate efforts with different methodologies, alongside our results in Chapter 4 forming the first large scale recoveries of eukaryotic MAGs from ocean environments. Delmont et al. [152] binned the data using anv'io [246], recovering 683 eukaryotic MAGs, as well as 30 SAGs, and reported the first MAG of greater than 1 Gbp in length, capturing a range of organisms from copepods to picoplankton. These MAGs remain partial compared to those from culturing efforts, with an average completeness of about 40%, assessed using the BUSCO v3 eukaryotic core gene set [252]. The MAGs and SAGs recovered represented an estimated 26% of the reads, based on recruit of reads to contigs adjusted for completeness.

This illustrates an area in which caution is required, the set of MAGs recovered is only a partial representation of the community, the majority of sequence data and potentially community members are not represented among them.

The second study by Alexander et al. [238] took an automated approach to binning, developing binning pipeline EukHeist which combines metagenomic and metatranscriptomic data. Binning was performed by MetaBat2 [157], which does not require the manual curation of bins that is core to the design of anvi'o. Assemblies were binned, and then potential eukaryotic bins identified using EukRep [236], selecting any bin with $\geq 90\%$ of its length predicted as eukaryotic. 988 MAGs were recovered in this way, though only 485 were more than 30% complete, the cutoff which was applied to consider a MAG of sufficient quality for this study. From the gene content of these MAGs, it was possible to predict whether organisms were autotrophs or heterotrophs (Section 2.3.2) from their gene content, and to identify ecological niches for recovered Stramenopiles. Neither eukaryotic binning effort utilised the pre-filtering approach, i.e. using EukRep or similar tools to identify eukaryotic contigs prior to binning. Alexander et al. did use EukRep after binning of all sequences to separate out eukaryotic MAGs from prokaryotes.

An area where the two studies differed quite widely despite using the same data is in how well marine fungi were recovered. Delmont et al. recovered a single fungal MAG from the phylum Ascomycota, while Alexander et al. recovered 16 fungal MAGs. More fungal MAGs were recovered by Alexander et al. from their mesopelagic co-assemblies, where the fungal community has been observed to be more diverse compared to the epipelagic [267]. Depth may be the cause for this difference, as Delmont et al. focused on the upper ocean, using only surface and DCM samples. The marine fungal community is comparatively poorly understood compared to its terrestrial counterpart, but believed to be broadly dispersed and a contributor to biogeochemical cycles [268]. Identifying which differences in sampling or computational methodology lead to improved recovery of fungal MAGs would help expand the understanding of this portion of the marine environment.

3.6 Gene Prediction

Reads or assembled contigs may contain partial or complete genes coding for production of proteins. Locating these genes and identifying proteins they code for allows an insight into the biological functions a community is capable of. Functional annotation can be divided into two steps: finding genes or partial genes in sequences, and identifying homologous proteins with known functions.

Ab initio gene prediction software seeks to locate protein coding genes without comparing the query sequence to reference sequences. Analysis of the Tara Oceans data found a large proportion of novel genes in sequencing of prokaryote enriched marine samples, with up to 90% of genes from the Southern Ocean DCM being novel [4]. Reference based gene identification would fail to capture many of these novel genes, making *ab initio* gene prediction tools important in less well characterised environments.

Many gene prediction tools use HMMs to model the statistical properties of coding genes, and locate parts of sequences best fitting these models. Linear relationships between the frequency of nucleotides in codon positions in genes and the nucleotide frequencies in the whole genome have been demonstrated, and utilised in the program GeneMark.hmm to set model parameters for new unannotated sequences [269]. Global nucleotide frequencies for the new sequence are calculated, and parameters for the HMM selected using the established relationships. Hidden states in the model represent start and stop codons, and coding and non-coding regions. Locating the most likely sequences of hidden states for a query sequence can be performed with the Viterbi algorithm, labelling 3-mers as the likely components of genes. Initial versions of GeneMark.hmm assumed all sequences originate from a single genome, however the heuristics used to select model parameters have been expanded for use on metagenomic data [270]. Nucleotide frequencies are predicted based only on GC content, and non-linear relationships between GC content and specific codons are used.

Errors in sequencing present problems for gene prediction. Substitution errors can create a spurious start or stop codon, or remove a genuine one. Glimmer-MG [271] accounts for some of these errors by locating low quality bases in start or stop codons, and considers a path where the previous possibly spurious stop codon did not exist. Insertion or deletion of a base during sequencing can cause all bases in the sequence to be shifted by a character, altering the following codons, known as a frameshift error. FragGeneScan [272] is another HMM-based gene prediction tool which handles frameshift errors by having hidden states for insertion and deletion in the submodels for coding regions.

Eukaryotic genomes tend to be larger and more complex, making *ab initio* gene prediction more difficult. GeneMark-ES [64] does not require a predetermined training set of genomes with similar organisation, the unlabelled query sequences are instead used for iterative unsupervised training of the model, with parameters initialised based on GC content. An intron submodel is included in the HMM to reflect the division of eukaryote genes into introns and exons. Prediction algorithms specific to genomes with these more complex organisational characteristics will be important for eukaryotes, GeneMark-ES includes features which use characteristics specific to intron splicing in some fungal genomes in the gene prediction algorithm.

In addition to tools specifically developed for metagenomic data such as MetaGene [273], single genome tools have been adapted to handle anonymous and fragmentary metagenome sequences. Glimmer-MG [271] sets model parameters for each sequence by performing a taxonomic classification using Phymm [217] rather than GC content. Meta Prodigal [274] is a pipeline containing the single genome prokaryotic gene predictor Prodigal [275] with pre-processing clustering steps to handle metagenomic data.

The JGI IMG pipeline [234] which was used to annotate metagenomes in Chapter 4 uses multiple gene predictors and outputs a majority rule consensus of the results. The gene prediction tools used are GeneMark.hmm, Prodigal, MetaGeneAnnotator, and FragGeneScan. These tools are suited to the fragemnted nature of metagenomic sequences, but predict based on simpler prokaryotic gene structure rather than the more complex eukaryotic structures.

MetaEuk [257] bridges that gap, providing a reference based method for predicting eukaryote genes in metagenomic sequences. Potential protein coding fragments of contigs (those between stop codons) are searched against a reference database of predicted proteins (e.g MMETSP [260]), and where fragments on the same contig match against the same reference sequence are considered potential exons. These exons are assessed for compatability based on their ordering on the contig, distance between them, and lack of overlapping on the reference sequence. Dynamic programming is then use to find the optimal set of exons for the reference sequence. This approach removes the limitation of tools such as GeneMark-ES which assume all sequences originate from the same genome, allowing prediction of genes with intron/exon structures in metagenome data, but as a reference based approach is limited by the extent of available databases.

3.7 Functional Annotation

Function can be assigned to predicted genes by locating similar genes or proteins using sequence similarity methods discussed in Section 3.4.3.

Individual proteins form part of larger biochemical processes, and different projects seek to associate genes or proteins with these wider processes. The Kyoto Encyclopedia of Genes and Genomes (KEGG) provides a curated database of pathways, visualised as networks showing molecular interactions within a given process such as photosynthesis. Elements in pathway can be compounds or products of a Kegg Orthology (KO), a group of genes identified as orthologous and assigned a unique KO identifier [276]. Associating a gene with a KO allows the presence of pathways in the community to be analysed. A similar pathway based database is provided by MetaCyc [277]. Cluster of Orthologous Groups (COG) [278] derived orthologous groups which contain genes from at least three lineages, and assigned each COG one of 25 high level functional categories. Sequences with similarity to those in a COG can be annotated with the specific function of those in the COG and the higher level grouping. COG is focussed on bacteria and infrequently updated, eggNOG [279] extends the method to cover a wider range of organisms, and also provide annotations from other systems such as KO terms for the orthologous groups.

Gene Ontology (GO) [280] organises functions into three directed acyclic graphs, one each for describing biological processes, molecular function, and cellular components. A given gene could receive annotations in all three graphs. The GO consortium also maintain or provide references to mappings from other annotation systems to GO terms. The Pfam protein families database consists of seed alignments for protein families, from which profile HMMs are provided [281]. Protein family entries are annotated with functional information, where this is available in literature. Searching a query protein sequence against the database of HMMs allows the unidentified protein to be associated with a protein family if high probability matches are found. InterPro seeks to intergrate a wide range of sources, including Pfam, into a single database [282]. Entries in the combined InterPro databases have profile HMMs provided, and the metadata provides links back to their source databases, and cross references to other resources such as KEGG. In addition, they maintain a tool for performing annotation using this database, InterProScan [283].

For focussed analyses, specific databases are available, or small sets of curated marker genes can be selected and identified. The Carbohydrate-Active enZYmes (CAZy) database is specific to enzymes involved in biological processing of carbohydrates [284]. For well studied processes such as nitrogen cycling, marker genes indicating the capacity for specific steps of the process are often used, such as *nirsS* or *nirK* genes indicating the presence of denitrifiers [285]. As well as process specific, environment specific databases can be utilised. In the marine context, the MMETSP and MarRef [260, 191] provide reference sets with taxonomic and functional annotation, allowing functional assignment by homology to annotated sequences in these databases.

Chapter 4

Metagenome Assembled Genomes Across the Arctic and Atlantic Oceans

4.1 Summary

This chapter first details the methods by which we obtained both eukaryotic and prokaryotic MAGs from twelve metagenomic samples taken from stations between the Arctic Ocean and the tropical Atlantic Ocean, collection as part of the Sea of Change project. This chapter is an adaptation of a paper that has been published in Microbiome [23], and at the time of being added to bioRxiv [286] represented the first study specifically targeting the recovery of multiple eukaryotic MAGs from ocean metagenomes.

The chapter begins by covering the methods used for assembly, initial annotation of the samples, and retrieval of prokaryotic and eukaryotic MAGs (Sections 4.2.2, and 4.2.3). The remaining methods Sections 4.2.4 to Section 4.2.9 cover the approaches taken to describe and analyse the recovered MAGs.

Results of these analyses are then presented in Section 4.3. After a summary of the unbinned data in Section 4.3.1, presentation of MAGs starts with an overview in Section 4.3.2. Analysis of the quality of MAGs is given in Section 4.3.3. Phylogenomic and taxonomic analysis identifies which organisms are present amongst the MAGs is in Section 4.3.4. In Section 4.3.5, distribution of the MAGs across the sampled stations is evaluated using coverage, looking at where these organisms are found. Functional analysis of the genes encoded by MAGs reveals what these organisms are capable of, and how these functions differ between polar and non-polar climates as explained in Section 4.3.6. Finally, associations between some pairs of eukaryotic and prokaryotic MAGs suggest some ways in which these

organisms interact is presented in Section 4.3.7. Section 4.4 brings the chapter to a close with a discussion of the results.

This project was a collaboration between ourselves and colleagues at JGI. Sequencing, assembly, initial gene prediction and annotation was performed by JGI. My contribution was devising and performing eukaryotic binning and read based taxonomic classification. Asaf Salamov at JGI generated three additional eukaryotic MAGs with different methodology, as discussed in Section 4.2.8. All subsequent analyses were devised in discussion with my supervisory team, and carried out by myself. The paper was drafted with my supervisory team, with much of the discussion placing results in a biological context being contributed by Thomas Mock.

4.2 Methods

The twelve samples we used to generate the MAGs were collected on two expeditions in 2012 as part of the Sea of Change project, with 6 samples from above the Arctic Circle, and 5 from the tropical and subtropical Atlantic. Samples were taken from the surface ocean or DCM layer, and filtered to select for eukaryotic size organisms. Detailed methods for collection and sequencing of these samples has been described by Schmidt [22] who collected these samples and by Martin [287], as well as summarised in our published paper [23]. The cruise and sampling strategy are summarised here.

4.2.1 Collection and DNA Extraction

Samples were collected on two RV Polarstern (Alfred-Wegener Institute for Polar and Marine Research, Bremerhaven, Germany) expeditions described by Martin et al. and Schimdt et al. [288, 22] and summarised below. During these campaigns, samples were taken from forty four stations for 18S and 16S amplicon sequencing, and for metatranscriptomic sequencing. Eleven samples from the DCM an surface layers of the ocean were selected for metagenomic sequencing. Six of these were stations within the Arctic circle, five in the tropical and sub-tropical Atlantic, shown in Figure 4.1. Arctic samples were collected on ARK-XXVII/1 (PS80) between 17th June and 9th July 2012; Atlantic samples were collected on ANT-XXIX/1 (PS81) between 1st and 24th November 2012. Samples from the Arctic were taken from between 10 and 20 metres of depth, where those in the Atlantic were deeper from between 30 and 80 metres. The sampling plan grouped samples as either surface (0-10m) or DCM (10-100m), so all the samples selected for metagenome sequencing were from the DCM group, but still display variation in depth between the two regions. Water samples were taken in 12 L Niskin bottles by a rosette sampler with attached Conductivity, Temperature, Depth (CTD) sensor, providing measurements of salinity and temperature at the time of sampling. Water samples were pre-filtered with a 100 µm mesh to remove larger zooplankton. Samples were then distributed into 1.25 l bottles, and those intended for DNA extraction were filtered with 1.2 µm polycarbonate filters, which were stored in liquid nitrogen at -80 °C until analysis. Phosphorus, nitrogen and silicate analysis was performed on duplicate samples, filtered with 0.2 µm nitrate cellulose filters and stored at -20 °C for phosphorus and nitrogen analysis, and 4 °C for silicate.

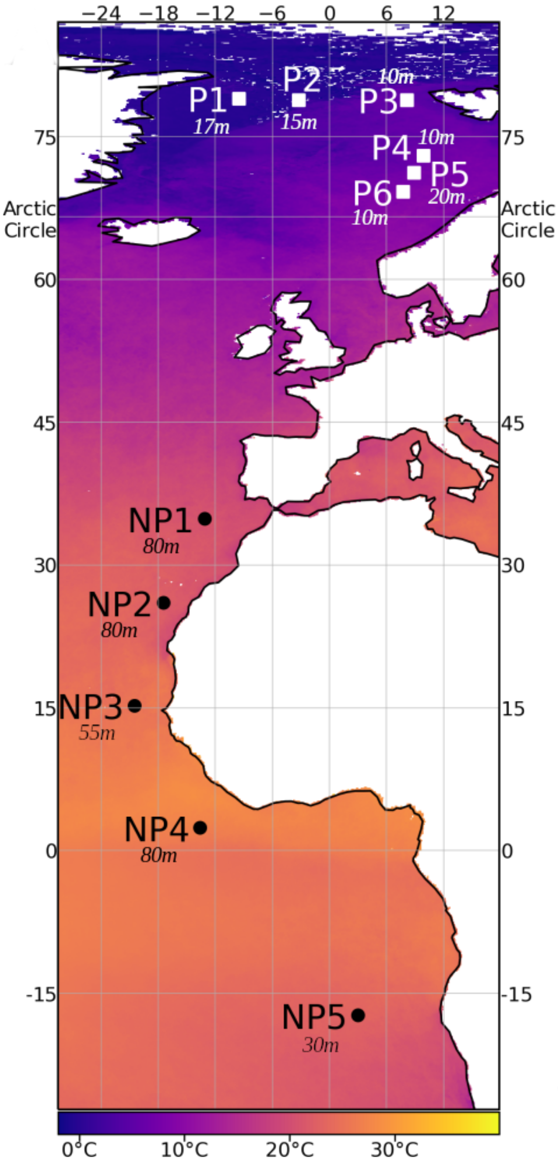


Fig. 4.1 Map showing stations from which samples were collected and sample depth [22]. Colour indicates mean annual sea surface temperature for 2012, taken from remote observation [289]. The latitude of the Arctic Circle is indicated with labelled line.

Cells were washed off the filter with pre-heated (65 °C) solution A from the kit, and the supernatant was transferred into a new tube with one small spoon of glass beads (425 µm-600 µm, acid-washed) (Sigma-Aldrich, USA). The samples were then vortexed three times in intervals of 3 s to break the cells. RNAse A was added to the samples and incubated for 30 min at 65 °C. The supernatant was transferred into a new tube, and solution B from the kit was added followed by a chloroform phase separation and an ethanol precipitation. DNA was pelleted by centrifugation and washed several times with isopropanol, air-dried, and suspended in 100 µL TE buffer. DNA concentration was measured with a Nanodrop (Thermo Fisher Scientific, Waltman, MA, USA), samples snap-frozen in liquid nitrogen and stored at $-80\,^{\circ}$C until sequencing. Description of the samples and associated metadata is available through the GOLD database [290].

As only a small number of the stations sampled on these expeditions were selected for metagenomic sequencing, in this work each station and associated sample has been relabelled to simplify referring to them. To easily differentiate between polar and non-polar locations the stations are labelled either P*n* or NP*n* respectively. Each sample is then assigned a number roughly following a path starting at the coast of Greenland and heading east and south. The first polar sample on that path is thus P1, the first non-polar sample NP1, as shown in Figure 4.1. Two duplicate samples from station P3 were selected for sequencing, these are labelled P3a and P3b. Details of sampling locations and associated metadata, including identifiers for data in repositories, is provided in Appendix A.2

4.2.2 Sequencing, Assembly and Annotation

These samples were sequenced, assembled and annotated by the JGI IMG pipeline [291], briefly summarised here. Paired end sequencing was performed on an Illumina HiSeq platform. BBDuk [292] was used to remove Illumina adapters, then BBDuk filtering and trimming applied. As a standard part of the quality control in the IMG pipeline, reads mapping to the human HG19 genome with over 93% identity were discarded. Remaining reads were assembled with MEGAHIT [175]. The quality controlled reads were mapped back to the assembly to generate coverage information using seal [293]. Some of these samples were later reassembled using SPAdes [178].

Genes were predicted by the IMG pipeline [291]. Briefly, genes were predicted from assembled contigs using prokaryotic GeneMark.hmm, MetaGeneAnnotator, Prodigal, and FragGeneScan [294, 295, 275, 272]. The number of copies of each gene is estimated from coverage of contigs generated by mapping back reads using seal [293]. Taxonomy was assigned to genes based on the top scoring USEARCH [296] result against an IMG reference database of non-redundant proteins from isolate genomes. Contigs were assigned the taxonomy of the last common ancestor of all genes on the contig, where more than 30% of genes have USEARCH hits. Where samples were later reassembled and annotated, as discussed in the previous paragraph, we use the predicted genes and estimated gene copies from the most recent assembly.

We performed taxonomic classification and abundance estimation of reads was performed using Bracken [219] and Kraken2 [297]. A custom Kraken2 database was constructed using all RefSeq genomes for bacteria, archaea, viruses, protozoa, fungi, as well as plants excluding embryophyta. Non-embryophyta plants were included to cover the green algae. Reads were taxonomically classified using Kraken, and abundance at the level of phylum was estimated with Bracken.

4.2.3 Binning

The IMG pipeline identified a number of prokaryotic bins. Samples were binned by JGI as described by Chen et al [291]. Briefly, each assembly was binned separately, using MetaBat [157] and a minimum contig size of 3,000 bp. As covered in Section 3.5.1, common forms of evidence used by binning algorithms are tetranucleotide frequencies and differential coverage. Each assembly was individually binned, differential coverage was not used, so only tetranucleotide frequency evidence was considered by the binning algorithm. Resulting bins were assessed for completion and contamination with CheckM [254], and subsequently assigned a taxonomy using GTDB-Tk [187]. While eukaryotic sequences were not excluded from this binning, all bins were labelled as archaea, bacteria or unknown by CheckM, prompting the distinct binning attempt for eukaryotes.

For the eukaryotic binning we carried out, each assembly was binned separately, the process for binning one assembly is given below, with more background on the approach and tools available in Section 3.5.4. Eukaryotic contigs were predicted with EukRep [235], which uses a linear support vector machine to classify sequences as eukaryotic or prokaryotic using k-mer frequencies. Coverage of the eukaryotic contigs was estimated by pseudoaligning the reads from each sample to the contigs using Kallisto [209]. Binning was performed using MetaBat2 [157] with the coverage information, and a minimum contig size of 1,500 bp. Completeness and contamination of resulting bins were assessed with BUSCO [252], using the eukaryota_odb9 set of genes. Bins which were less than 50% complete were discarded from further analysis. Completeness and contamination of bins was later reassessed using EukCC [255] which takes a similar approach to CheckM, seeking to identify the bin lineage and select a more specific set of single copy genes. The diagram in Figure 4.2 shows the binning and analysis pipeline.

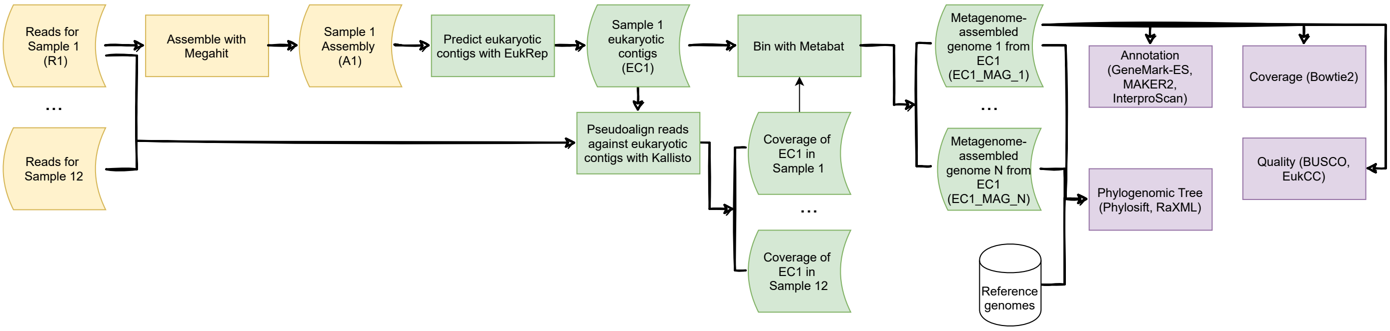

Reads for Sample 1 (R1)
...
Reads for Sample 12
Assemble with Megahit
Sample 1 Assembly (A1)
Predict eukaryotic contigs with EukRep
Sample 1 eukaryotic contigs (EC1)
Pseudoalign reads against eukaryotic contigs with Kallisto
Bin with Metabat
Coverage of EC1 in Sample 1
...
Coverage of EC1 in Sample 12
Metagenome-assembled genome 1 from EC1 (EC1_MAG_1)
...
Metagenome-assembled genome N from EC1 (EC1_MAG_N)
Reference genomes
Annotation (GeneMark-ES, MAKER2, InterproScan)
Phylogenomic Tree (Phylosift, RaXML)
Coverage (Bowtie2)
Quality (BUSCO, EukCC)

Fig. 4.2 Pipeline for eukaryotic binning. Steps in yellow were performed by the Joint Genome Institute (JGI) [291]. We performed the binning process shown in green, and analysis steps in violet.

4.2.4 Phylogenomics

To aid in the taxonomic identification of MAGs, a phylogenomic tree was constructed from both reference genomes and MAGs. PhyloSift [298] was used to identify sequences homologous to the mostly-single copy marker genes in bins and reference genomes using the HMMs provided by PhyloSift. Trees for eukaryotes and prokaryotes were constructed separately.

For eukaryotic reference genomes, all protists and green algae labelled representative from NCBI were used. This included only a single diatom genome, *Fragilariopsis cylindrus*, so two additional diatom genomes (*Thalassiosira pseudonana*, *Phaeodactylum tricornutum*) taken from JGI were included after initial analysis of MAGs suggested several were potential diatoms. For prokaryotes, all genomes in the MarRef [299] database were included. Homologous sequences were located and the best hit retained when there were multiple. The PhyloSift set of marker genes contains both genes specific to eukaryotes, and genes identified in bacteria but which have full length homologs in eukaryotes based on searches against the yeast genome. For eukaryotes, all genes were used, for prokaryotes the eukaryote specific genes were excluded. Marker genes present in less than 50% of the genomes (reference or MAG) were not used in future steps of the analysis; MAGs with fewer than 50% of marker genes which passed this threshold were then excluded. Homologous sequences were aligned against the PhyloSift models, and alignments for all genes concatenated. FastTree [300] was used to build initial phylogenomic trees for the eukaryotic and prokaryotic alignments, using the general time reversible model option. Trees including bootstrap values were subsequently constructed using RAxML [301] using the GTRCAT model approximation with 100 bootstrap replicates. The eukaryotic tree was midpoint rooted, while the prokaryotic tree was rooted between the clades containing archaea and bacteria. The resulting trees were visualized with Interactive Tree of Life Viewer [302].

Following the publication of eukaryotic MAGs generated from Tara Oceans data [152] we constructed a tree combining putative diatom MAGs from both these results, our results, and 44 reference genomes. This tree was constructed using a concatenated alignment of the genes from the eukaryota_odb_10 gene set. The same thresholds were applied, with marker genes present in less than 50% of the genomes were excluded, and genomes with less than 50% of those genes excluded. Genes were individually aligned using MUSCLE [303], the alignments concatenated and trimmed using TrimAL's -automated1 setting [304]. The tree was constructed with RAxML using the automatic model selection option PROTGAMMAAUTO with 100 bootstrap replicates.

As additional evidence for taxonomy, contigs from MAGs were searched against databases with BLAST [203], and each contig assigned a taxonomy using the MEGAN-LR algorithm [305]. Eukaryotes were searched against MMETSP [260], prokaryotes against NT. Selected clades of MAGs with close placement had pairwise ANI and Average Amino Acid Identity (AAI) calculated, using the pyani [306] BLAST based ANIb method and CompareM [254] respectively. For AAI comparison, reference protein sequences were retrieved from MarRef [191] for prokaryotes, and from PhycoCosm [25] for eukaryotes.

4.2.5 Coverage

Coverage for each eukaryotic MAG was generated by aligning reads from each sample back to the bins using Bowtie2 [207]. Detection and mean coverage were calculated from these alignments using BedTools [307]. We considered a MAG not present in a sample if the detection (proportion of bases in MAG with any read aligned) was lower than 0.5, as in Olm et al [236]. To serve as an estimate of the relative abundance of a MAG, the mean coverage was divided by the number of million reads in the sample.

Only a fraction of the species truly present will have MAGs recovered. To estimate the total proportion of the population represented by MAGs, contigs from eukaryotic and prokaryotic MAGs were concatenated to a single file, and read pairs from each sample were pseudoaligned back to this set of contigs representing all MAGs using Kallisto [209]. The proportion of the reads which mapped back to the concatenated contigs was taken as an estimate of the proportion of the reads represented by the recovered MAGs.

4.2.6 Gene Prediction

Protein coding genes were predicted as part of the IMG pipeline prior to binning (see Section 3.6). Prediction was performed using an ensemble of prokaryotic gene prediction tools: Prodigal, prokaryotic GeneMark.hmm, FragGeneScan and MetaGeneAnnotator. Each of these tools aims to identify genes with prokaryotic gene structure, and are not adapted to the more complex gene structures of eukaryotes. In addition, non protein coding features like CRISPR elements and rRNA are predicted. Eukaryotic MAGs had genes predicted using GeneMark-ES [64] in self training mode with MAKER2 [308]. GeneMark-ES starts from the assumption that all sequences provided originate from the same genome, so this step had to be performed on the generated bins, rather than contigs prior to binning.

4.2.7 Functional Annotation

The IMG pipeline annotated genes which were predicted by its ensemble of prokaryotic gene prediction tools. Protein coding genes are assigned to COG, Pfam, TIGRfam, KO, and a subset of InterPro families. Further background on function annotation databases is given in Section 3.7. The proteins are further associated with KEGG and MetaCyc pathways based on the KO terms and related Enzyme Comission (EC) numbers. GO terms for prokaryotic genes were generated using the mapping of Pfam accessions to GO terms maintained by InterPro. Protein coding genes predicted separately for eukaryotic contigs after binning lacked functional annotation from the IMG pipeline, and were annotated using InterproScan 5 [283].

4.2.8 Additional Eukaryotic MAGs Generated by JGI

Work undertaken by Asaf Salamov at JGI identified three additional eukaryotic MAGs using different methods. These three MAGs have been included in all analyses in this chapter, and a summary of their method is given here. Following assembly, contigs were searched against NR and MMETSP [260] using Mmseqs2 [309] to assign taxonomy, and prokaryotic contigs discarded. Binning was again performed using MetaBat [157] on the filtered contigs of each assembly separately. Bins were then filtered to select those with conserved taxonomic origin, retaining only those with more than 50% of contigs assigned to a single eukaryotic phylum and a total length of greater than 5 Mbp. Finally bins were filtered to remove contigs from other taxa. This resulted in recovery of three additional 3 medium quality MAGs.

4.2.9 Inter-kingdom Species Association

To investigate associations between prokaryotic and eukaryotic MAGs, we looked at the correlation between coverage of pairs of MAGs. Ordinary least squares regression was performed between each pair of eukaryote and prokaryote MAGs, and any pair with $R^2 \geq 0.7$ and p-value ≤ 0.05 retained. Examining plots for the retained pairs suggested that some of the correlations were driven by single or a small number of highly influential observations. To address this we used Cook's distance, which provides a measure of how influential an individual observation is to the results of a linear regression analysis [310]. We discarded any pairs where the regression did not did not meet the thresholds mentioned earlier after points with Cook's distance greater than 1.25 were removed.

For the pair of MAGs with the clearest association (NP2_2E and NP3_22P), we looked at the enrichment of functions in each. Enriched GO terms were identified using Fisher's exact test, comparing terms found within the MAG to terms in a set of background MAGs. The selected pair were taxonomically identified as *Bathycoccus* and *Alphaproteobacteria* respectively. For the eukaryote NP2_2E, all Prasinophyte MAGs not involved in any of identified associations were used as a background set, a total of 2 MAGs. For prokaryote NP3_22P, all Alphaproteobacteria MAGs not involved in associations were selected as background set, a total of 12 MAGs. We considered any terms overrepresented in the associated MAG with $p \leq 0.05$ in a one-sided test to be enriched in the MAG.

To check whether the identified enriched terms were specific to this associated pair, or would appear enriched regardless of association, we looked at terms enriched in two control pairs. The same background sets were retained, and MAGs not involved in any of the identified associations selected as pairs to investigate for enrichment. The first pair selected were distantly related to the background sets, the selected eukaryote was *Bacilliarophyta* P3a_4E, and prokaryote *Gammaproteobacteria* NP3_6P. To identify whether enrichments were taxonomically driven, a second closely related pair were drawn from the background sets. The selection MAGs was eukaryote *Prasinophyte* P2_1E and prokaryote *Alphaproteobacteria* P3a_15P. These two MAGs were removed from the background set in these enrichment analyses.

4.3 Results

To provide an overview of the community from which the MAGs are being drawn, this chapter opens in Section 4.3.1 with a summary of taxonomic and functional annotations of the reads and assemblies prior to binning. Sections 4.3.2 to 4.3.7 then focus in on the community members for which MAGs were recovered, presenting analyses of quality, taxonomy, distribution, function, and interaction.

4.3.1 Data Summary

Sequencing of the 12 samples resulted in 4.53 billion reads totalling 679.25 Gbp, with each sample ranging between 46.79 Gbp and 67.37 Gbp. The size of each data at each step of processing is shown in Figure 4.3. Assembling each station with MEGAHIT resulted in 42.10 million contigs totalling 23.02 Gbp. The MEGAHIT assemblies for three samples, P3b, P4 and P5 are notably smaller than the rest, being less than 1 Gbp in length. Summary statistics for reads and assemblies are provided in Appendix A.2. Reads for six samples were assembled using both MEGAHIT and later SPAdes. Assemblies from SPAdes in all cases resulted in a smaller overall length of assembly, but with the largest contigs being longer in SPAdes assemblies and a greater proportion of the assembly being in scaffolds greater than 50 kbp. While the two duplicate samples from the same station P3a and P3b generated 59.34 Gbp and 46.79 Gbp respectively, the resulting assemblies differ greatly in size, the MEGAHIT assemblies being 3.12 Gbp and 0.39 Gbp.

Taxonomy

Prior to binning, we looked at the overall taxonomic composition of the community as a whole, by taxonomically classifying both the reads and assembled contigs. Reads were taxonomically annotated by IMG for a subset of these samples. This annotation was done using different versions of the IMG pipeline, and the taxonomic composition based on reads shows grouping based on pipeline version. Comparison of these IMG read annotations would be partial and confounded by pipeline version, so we performed a read based taxonomic classification was using Kraken2 and Bracken [218, 219]. Kraken2 taxonomically classified 365.85 million (15.74%) of the read pairs. Bracken abundance estimation at the levels of superkingdom and phylum is shown in Figure 4.4.

Distribution of relative abundances at the ranks of superkingdom and phylum for polar and non-polar samples are shown in Figure 4.4. Generally, eukaryotes are more abundant in polar stations, contributing between 22% and 27% of the total abundance of reads, whereas they only contribute between 12% and 19% non-polar stations. In non-polar stations with lower abundance of eukaryotes, there is a corresponding increase in the abundance of archaea. This is most pronounced in stations NP1 and NP2, where the most southern non-polar station NP5 appears to be more similar to polar stations. Mean abundance of eukaryotes in polar and non-polar samples shows a statistically significant difference ($p = 0.000074$), assessed using

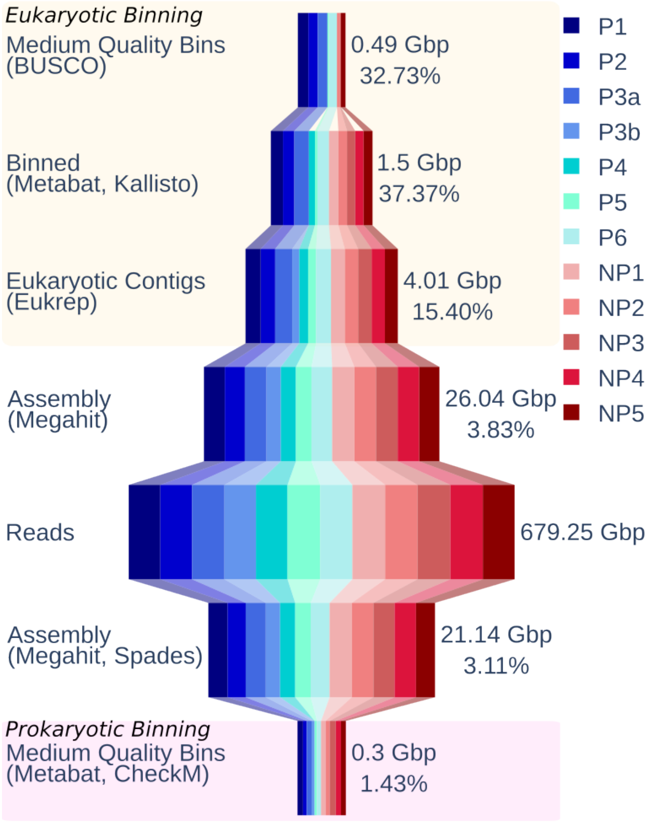


Fig. 4.3 Representation of size of data at steps of processing. Widest central bar represents quality controlled reads. Upper portion is the eukaryotic binning process, lower is the prokaryotic binning. Polar stations are red, non-polar are blue. Width represents the size of data at that step, and is log-scaled.

a T-test assuming independent samples. Similarly, abundance of archaea shows a significant difference between polar and non-polar samples ($p = 0.0091$). Differences between the other superkingdoms is not significant at $p = 0.05$.

At the rank of phylum, Proteobacteria is the most abundant, with Ascomycota the most abundant eukaryotic phylum. The most abundant species is the Cyanobacteria *Prochlorococcus marinus* with a mean relative abundance of 3.40%; the most abundant eukaryote is *Micromonas commoda* with mean relative abundance of 1.24%. The photosynthetic eukaryotic phyla Chlorophyta and Bacillariophyta generally have higher relative abundance in polar stations, with Cyanobacteria being more abundant in non-polar stations. The southernmost non-polar station *NP5* appears more similar to the polar stations, with a raised relative abundance of Bacillariophyta.

Principal Coordinates Analysis (PCoA) of the species level taxonomy of these samples was performed using the pairwise Bray-Curtis distance between samples. The results in Fig-

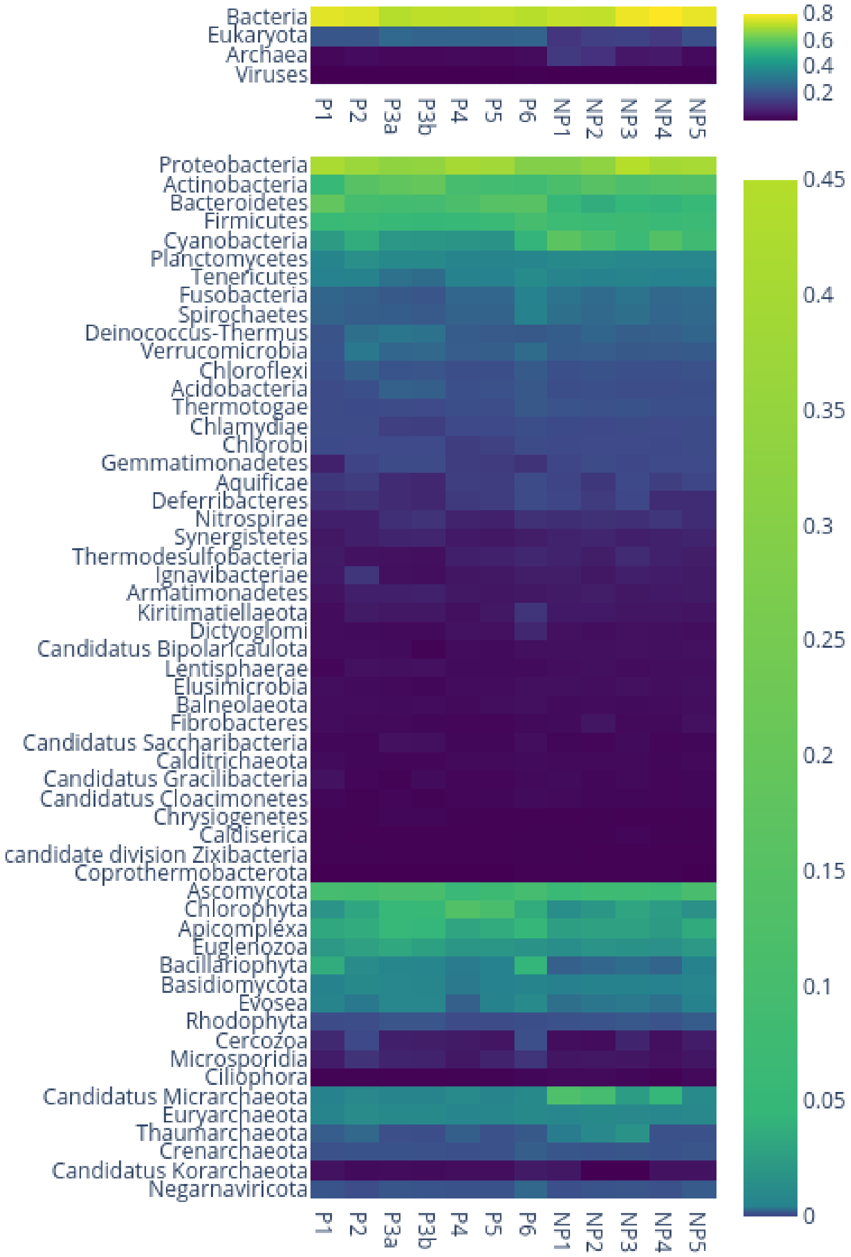


Fig. 4.4 Taxonomic composition of samples estimated by Bracken, summarised to the rank of superkingdom (top) and phylum (bottom).

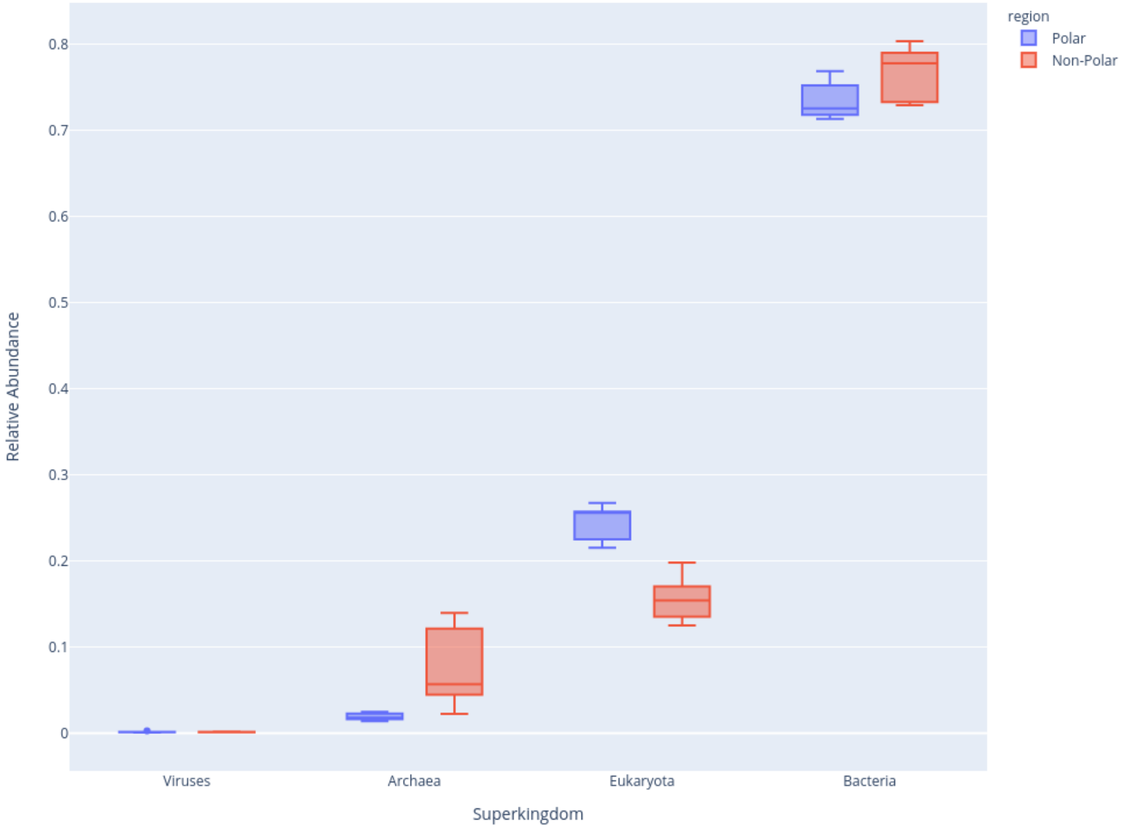


Fig. 4.5 Distribution of relative abundance estimates from Bracken for each superkingdom, divided up by polar and non-polar. Each pair was tested for differing means using a T-test assuming independent samples. p-values are Viruses $p = 0.72$, Archaea $p = 0.0091$, Eukaryota $p = 0.0000745$, Bacteria $p = 0.065$,

ure 4.6 shows show a clear separation of polar and non-polar samples along the primary axis, which explains 46.1% of variation, suggesting a clear demarcation between the taxonomic composition of polar and non-polar communities.

Function

One or more genes were predicted by the IMG pipeline on 36.76 million of the 42.10 million contigs, with 50.30 million genes predicted in total. Domains homologous to those in the Pfam database were found in 13.83 million (27.51%) of the predicted genes. Within samples, this proportion varied from 17.97% to 33%. The two samples from P3 had the lowest ratio of genes with homologous Pfam domains, both under 20%.

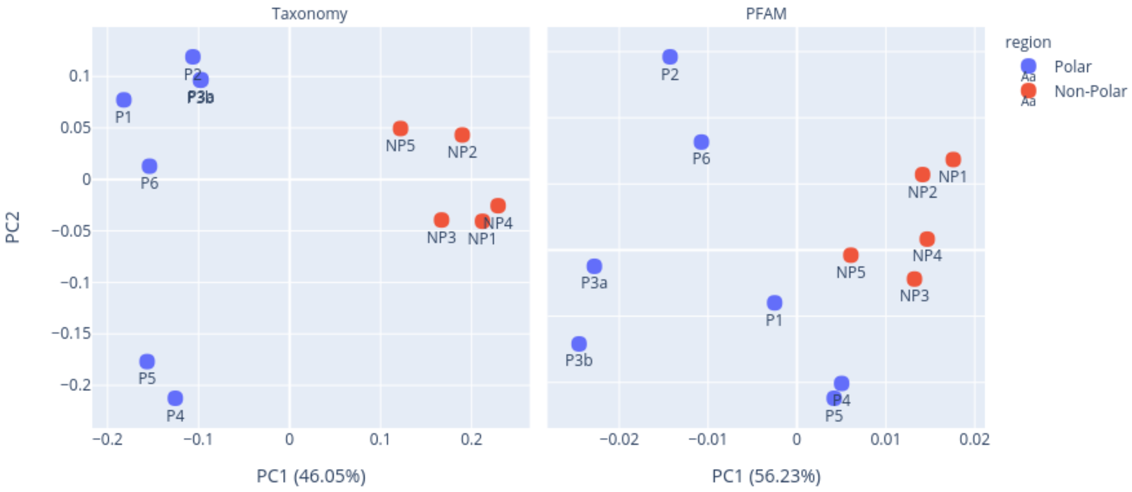


Fig. 4.6 First and second component of Principal Coordinate Analysis performed using Bray-Curtis distances between relative taxonomic abundance at species rank, and relative abundance of Pfam domains in gene annotations. Percent in axis labels is the percentage of variance explained. Taxonomy shows a clear separation between polar and non-polar samples, which is less pronounced based on function.

Taxonomic affiliations were assigned to 17.74 million of the genes, of which 28% were eukaryotic, 66% prokaryotic, 6% viral. The relative estimated gene copies by taxonomy at the rank of superkingdom and phylum is shown in Figure 4.7. These data do not describe the distribution of individuals in the sample, as the number of genes encoded in the genome of an organisms will vary with lineage. However, many of the broad trends observed in the read based classification and abundance estimates hold for these gene based data as well.

The most abundant genes were of bacterial origin followed by eukaryotes, viruses and archaea. On the phylum level, genes from Proteobacteria were most abundant with Haptista being the most abundant eukaryotic phylum followed by Chlorophyta. Generally, eukaryotic genes are more abundant in polar stations, contributing between 25 and 46% of the total genes, whereas they only contribute between 10 and 31% in non-polar stations. In non-polar stations with a lower abundance of eukaryotic genes, there is a corresponding increase in the abundance of archaea and viruses. Differences between the means of gene counts in polar and non-polar stations are statistically significant for eukaryotes, viruses and archaea assessed using a T-test at a significance level set at ≤ 0.05. Genes from photosynthetic eukaryotes such as Chlorophyta and Bacillariophyta generally have higher relative abundance in polar stations, whereas those from Cyanobacteria are more abundant in non-polar stations. The proportion of genes annotated as of fungal origin is much lower than in the read-based

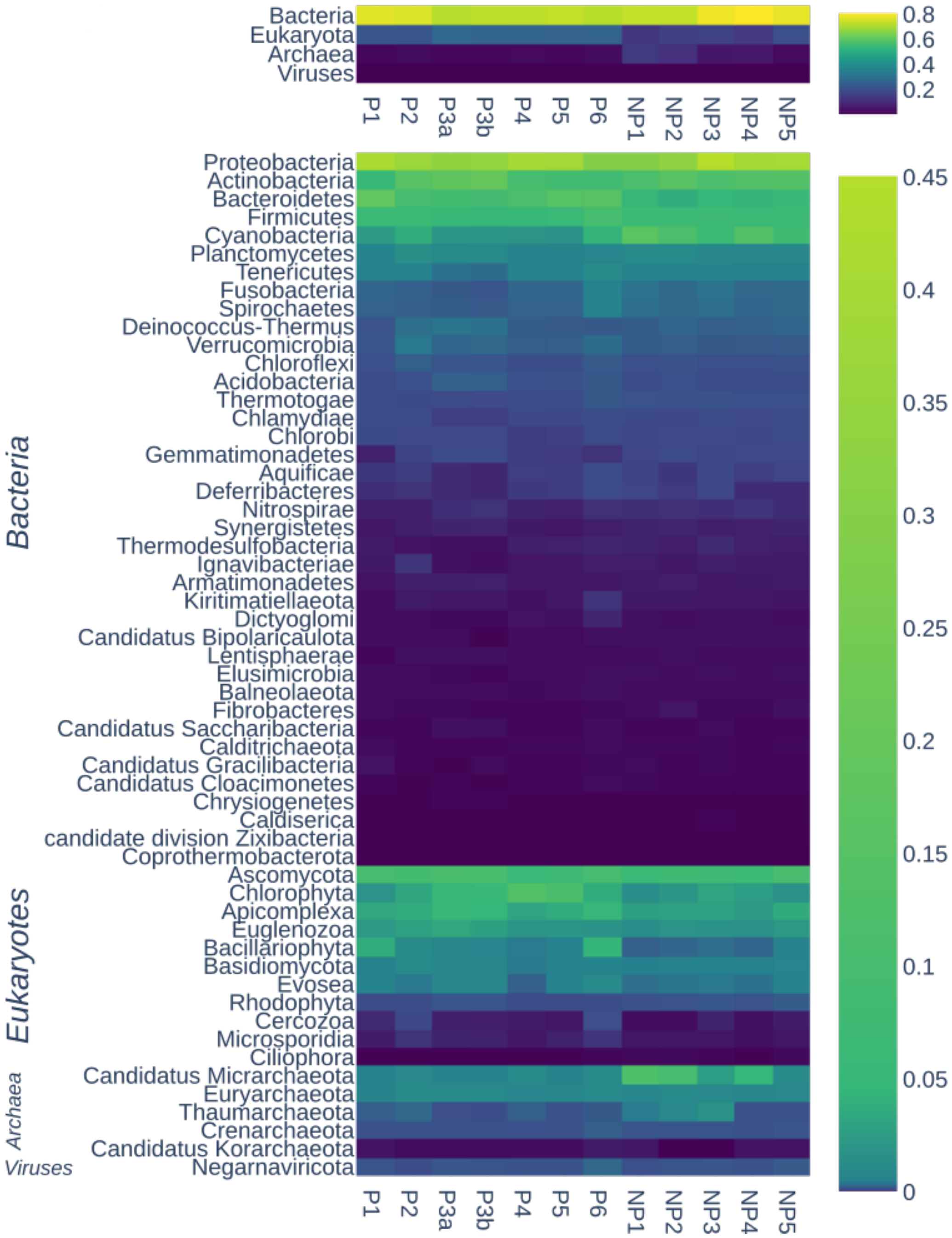


Fig. 4.7 Relative abundance of genes by taxonomic assignment. Data produced by JGI IMG pipeline [291].

estimated taxonomic abundance, where Ascomycota are estimated as the most abundant eukaryotic phylum. One possible explanation for this is that fungal genomes can vary greatly in size, but tend to encode a comparatively small number of genes [311].

A majority (87%) of the observed Pfam domains are present in both polar an non-polar samples. The proportion of domains of unknown function is higher in the domains uniquely found in either polar or non-polar stations than shared between them. Domains of unknown function constitute 16.55% of the shared domains, but 23.76% and 29.71% of domains unique to either the polar and non-polar respectively. Among domains unique to the polar samples, 63.57% were observed in only one sample, and none were in all samples. For the non-polar samples this was lower at 43% in only one sample, and 8.50% were in all samples. The domains found in both areas are more well distributed, with 57.55% being ubiquitous in every sample. The PCoA plot in Figure 4.6 is based on the relative abundance of Pfam domains in samples. The separation between polar and non-polar samples is less clear than in the ordination based on taxonomy.

A plot of the 20 GO terms in each namespace with the highest mean abundance each GO namespace is shown in Figure 4.8. Among this restricted set, terms related to ribosomal components or activity (GO:0003735, GO:0006412, GO:0005840) show greater abundance among the polar samples than non-polar. This fits with findings that phytoplankton require a higher density of ribosomes under cold conditions to meet cellular protein synthesis requirements [312]. Samples P2 and P6 show raised values for some parts of the photosynthetic machinery, those terms related to photosystem I and the thylakoid membrane (GO:0009522, GO:0009579). Among the polar stations, P2 and P6 are also the pair with high estimated abundance of Cyanobacteria.

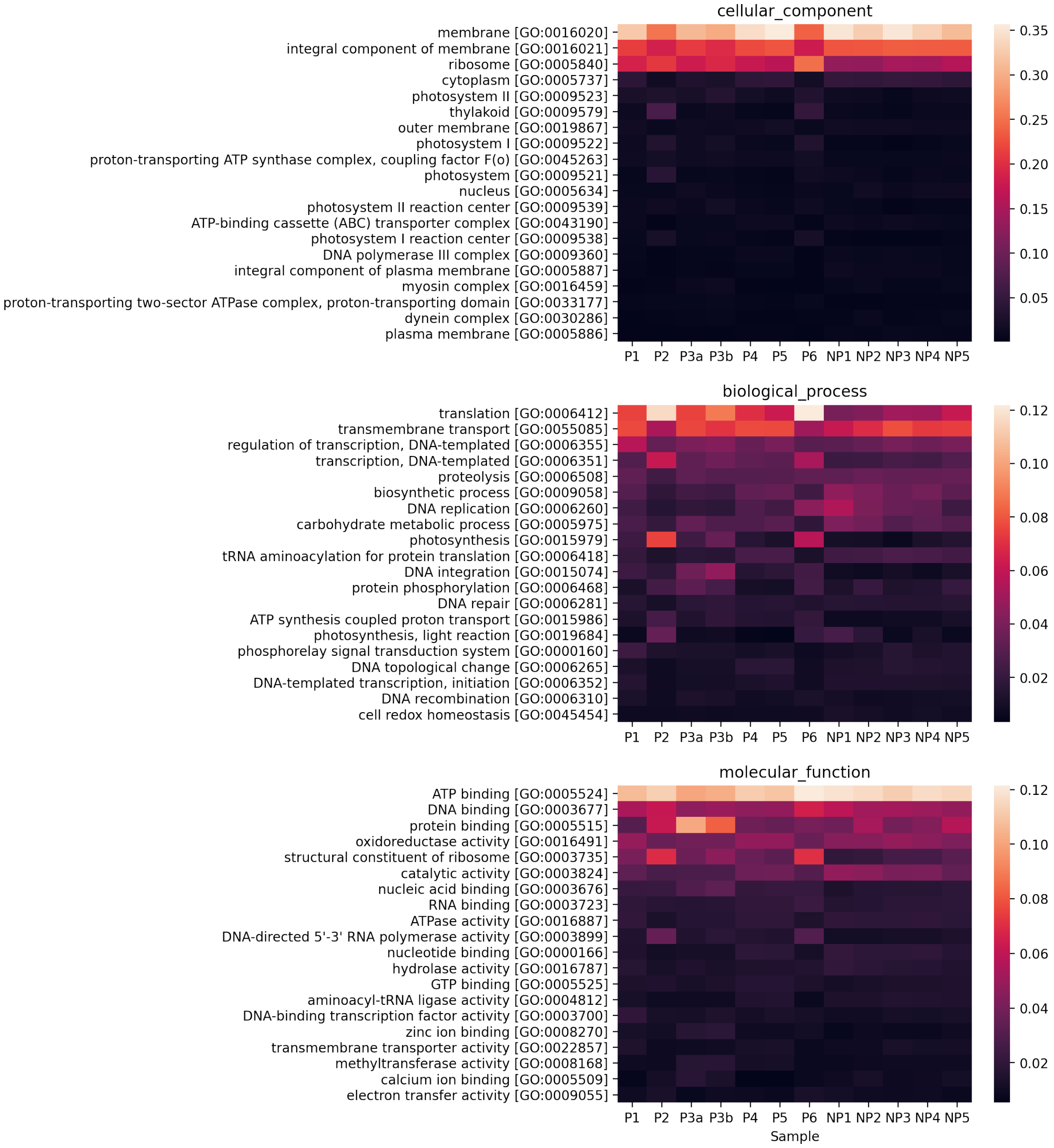


Fig. 4.8 Relative abundance of GO terms with the highest relative abundance in predicted genes. Abundance is based on estimated gene copies for genes predicted in each assembly. GO term counts were based on the mapping from Pfam to GO terms maintained by the GO Consortium. Relative abundance calculated separately for each GO namespace, as terms in each namespace can be similar or synonymous, and Pfam domains can map to multiple terms.

4.3.2 Bin Summary

Metagenome binning resulted in 143 MAGs of medium or high quality; 122 prokaryotic MAGs, and 21 eukaryotes. The MAGs total 0.79Gbp of contigs, and 8.1% of all reads mapped back to the MAGs. For the prokaryotes, these bins total 0.3Gbp, a small proportion of the assemblies as shown in Figure 4.3.

Prokaryotes were taxonomically identified using GTDB-Tk [187], and 116 were classified as bacteria, and 6 as archaea. Slightly more prokaryotic MAGs were retrieved from non-polar samples than polar, 64 and 58 respectively. All prokaryotic MAGs from polar samples were classified to at least the phylum level, and came from among Bacteroidota, Proteobacteria and Verrumicrobia. Verrumicrobia were only recovered from polar samples. The prokaryotic MAGs from non-polar samples come from a wider range of phyla, and included all of the 6 archaea. In addition to Bacteroidota and Proteobacteria, non-polar MAGs included 6 Actinobacteriota, 8 Myxococcota, 2 Patescibacteria, 5 Planctomycetota and 1 Poribacteria, all of which phyla were unique to the non-polar samples. Summary statistics for prokaryotic MAGs are provided in Appendix A.3.

Filtering the assembly for each sample to retain only eukaryotic contigs as predicted by EukRep [235] resulted in 2,151,309 contigs totalling 4.01 Gbp. A much higher proportion of polar assemblies was predicted as eukaryotic than non-polar, shown in Figure 4.9. From these we recovered 21 medium quality eukaryotic MAGs. Only four of these eukaryotic MAGs were retrieved from non-polar samples. Taxonomy was assigned to the eukaryotic MAGs based on their placement in a phylogenomic tree discussed in detail in Section 4.3.4; 8 placed with Mamiellophyceae reference genomes, 10 with Bacillariophyta, the placement of the remaining 3 was less clear. All but one of the Bacillariophyta originated from polar samples. The Mamiellophyceae break down by genus: all the polar Mamiellophyceae MAGs placed in a clade with Micromonas, the non-polar with Ostreococcus or Bathycoccus. Summary statistics for eukaryotic MAGs are shown in Table 4.1.

4.3.3 Quality

Completeness is expressed as the percentage of expected single-copy genes from a selected gene set observed in the MAG, and contamination as the percentage of single copy genes observed in two or more copies. For prokaryotes, CheckM [254] selects a suitable gene set based on the identifying the probable lineage of each MAG. We initially used BUSCO [252] and the eukaryota_odb9 gene set for eukaryotes. Later we reassessed the eukaryotes using EukCC [255], a tool taking a similar lineage specific approach to CheckM, which was published shortly after initial completion of our binning. Following established standards

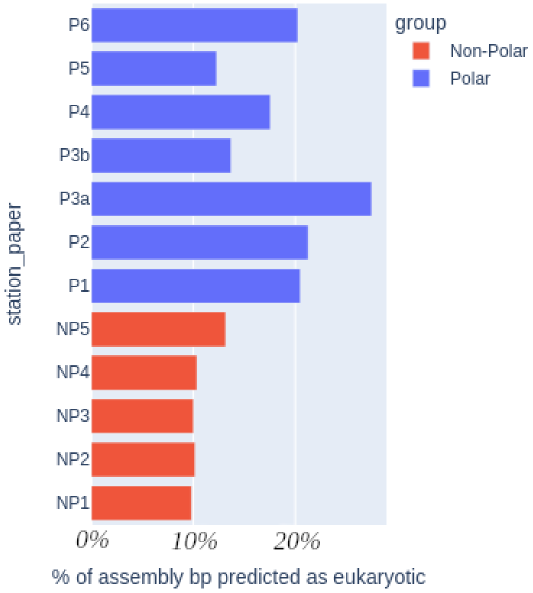


Fig. 4.9 Percentage of assembly length which was predicted as eukaryotic by EukRep [235]

[253], medium-quality for a MAG requires at least 50% completion and less than 10% contamination. The completeness, contamination, size and phyla of MAGs are shown in Figure 4.10.

Eukaryotic MAGs had a mean completeness of 67.82% and contamination of 2.82%. Details of the eukaryotic MAGs are shown in Table 4.1. The MAG with the highest completion is P2_1E at 92.97%. All but one MAG is quite fragmented, with a median L50 of 5,229 bp. The exception is P2_1E which contains many contigs longer than 50 kbp, the longest being 106 kbp. The number of predicted genes for eukaryotes ranges from 4,808 to 29,691, with a median of 12,301, with a positive correlation between length of sequence in MAGs and number of genes predicted as shown in Figure 4.11. We discard any eukaryotic MAGs which did not meet the medium quality completeness and contamination thresholds when assessed with BUSCO and eukaryota_odb9 gene set. When reassessed with EukCC, P1_3E fell slightly below the completeness threshold, from 55.8% to 48.72%, but with 0% contamination. This MAG was retained for all subsequent analyses.

Prokaryotic MAGs have a slightly greater mean completeness of 74.30% and similar contamination of 2.68%. The prokaryote with the highest completeness was Flavobacteriaceae P1_21P at 99.62% and a contamination of 2.81%. Assemblies for prokaryotic MAGs are slightly less fragmented with a median L50 of 11,402 bp and a median size of 2.23 Mbp. Part of the reason for this could be that the prokaryotic binning process used a minimum contig size of 3,000 bp, where 1,500 bp was used during eukaryotic binning. The number of predicted genes for prokaryotes ranges from 948 to 5,124, with a median of 2,254.5; as

Name	Contigs	Size Mbp	N50	L50	Longest Contig	Contigs $\geq$ 50 kbp	Completion (%)	Contamination (%)	Genes	Estimated Phylum
NP2_1E	2500	8.184211	762	3590	17504	0	64.14	3.52	5626	Chlorophyta
NP2_2E	1676	9.500694	413	7047	41273	0	65.81	0.85	4808	Chlorophyta
NP3_1E	2142	10.170215	570	5606	26327	0	70.48	0.63	5269	Chlorophyta
NP5_1E	5871	26.127275	1272	5463	56712	1	74.36	0	13678	Bacillariophyta
P1_1E	9009	39.146115	2256	5231	38710	0	74.51	3.92	19182	Bacillariophyta
P1_2E	4292	28.112524	929	9186	48653	0	74.36	0	14003	Bacillariophyta
P1_3E	8436	28.188465	2469	3646	26728	0	48.72	0	13261	Unknown
P2_1E	1539	21.112593	304	21396	105674	36	93.97	1.9	11269	Chlorophyta
P2_2E	10010	34.954696	2864	3867	43670	0	64.14	8.42	16812	Unknown
P2_3E	6287	25.87493	1677	4876	28783	0	66.67	3.92	13298	Bacillariophyta
P3a_1E	16982	58.203378	4688	3836	36404	0	68.89	2.22	29691	Haptophyta
P3a_2E	4635	20.475517	1207	5226	33641	0	62.75	5.88	11414	Bacillariophyta
P3a_3E	3223	12.128887	819	4238	42272	0	62.58	3.76	8289	Chlorophyta
P3a_4E	5446	24.3072	1389	5453	31509	0	66.67	0	14484	Bacillariophyta
P5_1E	2825	11.368259	728	4619	29122	0	56.14	1.75	7595	Chlorophyta
P6_1E	5547	29.631732	1263	7053	64386	2	78.43	1.96	12916	Bacillariophyta
P6_2E	4424	28.291357	964	8698	55053	1	70.59	7.84	12096	Bacillariophyta
P6_3E	3183	12.520615	841	4624	32566	0	61.4	1.75	8488	Chlorophyta
P1_4E	4074	18.716209	1146	5316	29967	0	58.82	0	9447	Bacillariophyta
P1_5E	4964	25.983826	1381	6015	39678	0	58.82	5.88	12446	Bacillariophyta
P2_5E	3511	17.474266	876	6052	32709	0	81.9	5.4	12301	Chlorophyta

Table 4.1 Assembly summary statistics for 21 eukaryotic MAGs

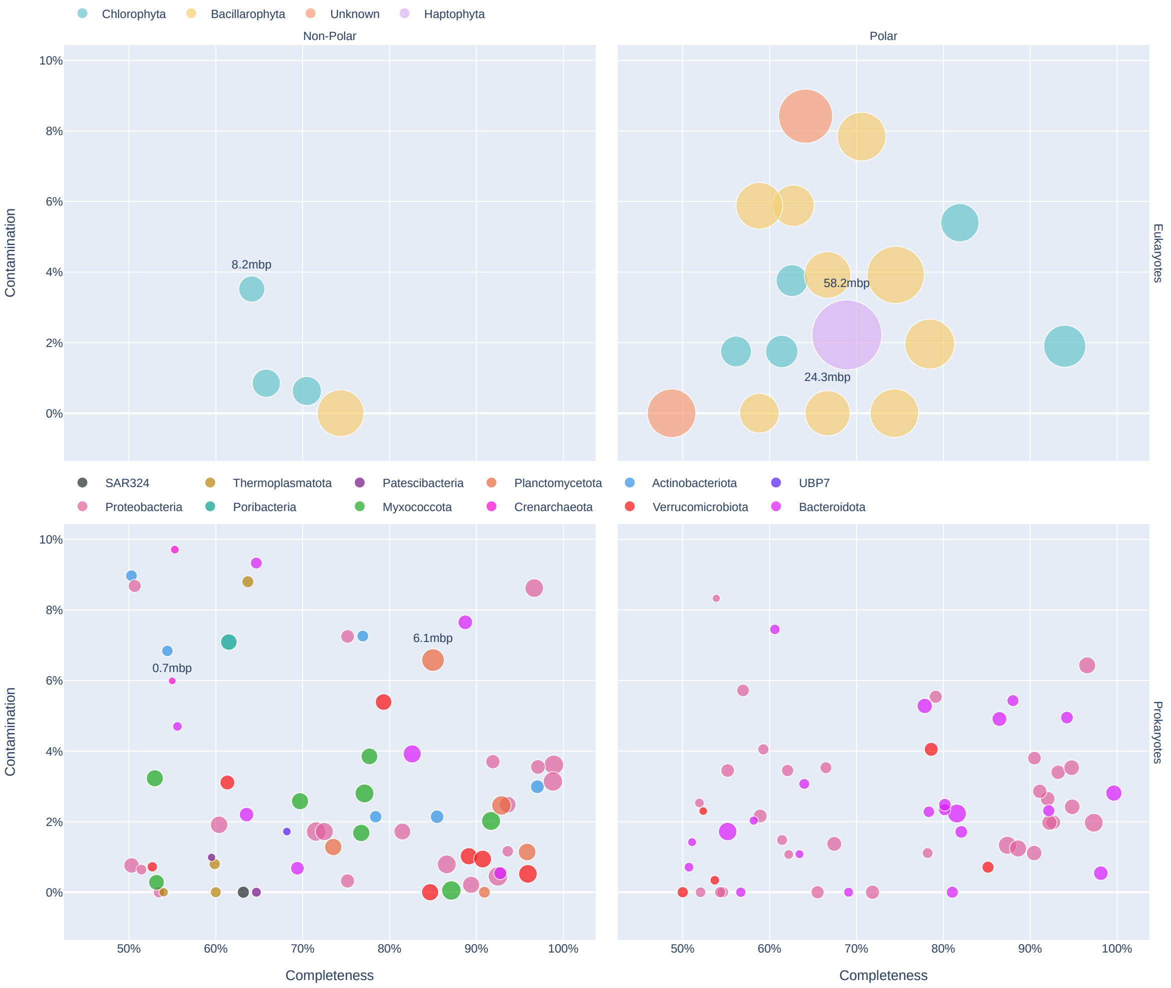


Fig. 4.10 Completeness, contamination, size and phylum of MAGs. The upper plots show eukaryote, the lower prokaryotes. The left column is non-polar MAGs, and the right polar. Area of the point represents size of the MAG. Colour shows the estimated phylum of each MAG. The vertical axis shows percent contamination, and the horizontal percent completeness.

with eukaryotes the overall length of a MAG and number of genes recovered show clear positive correlation (Figure 4.11). Summary statistics for prokaryotic MAGs are provided in Appendix A.3

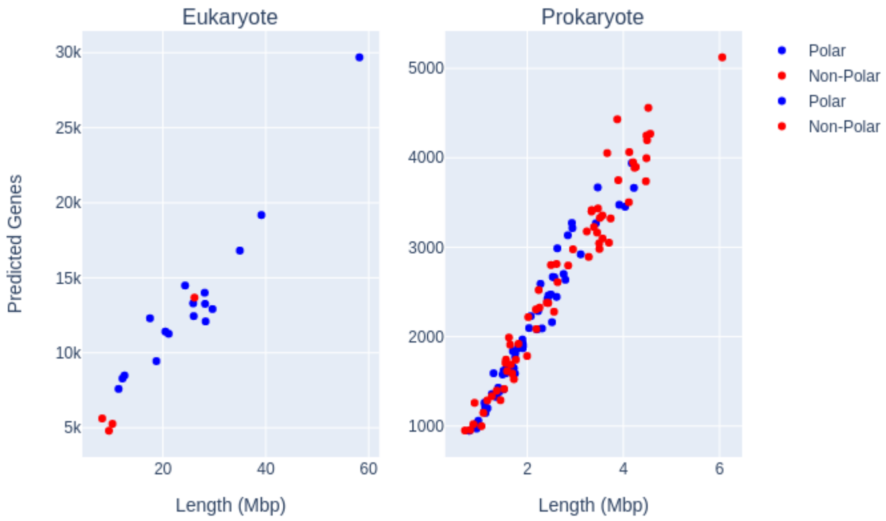


Fig. 4.11 Size of MAG in Mbp plotted against number of predicted genes. The left plot shows eukaryotes, the right prokaryotes. Colour indicates polar or non-polar origin of MAGs, blue and red respectively.

4.3.4 Phylogenomics and Taxonomy

Eukaryotes

The phylogenomic tree for eukaryotes in Figure 4.12 was constructed using concatenated alignments of 57 marker genes, a subset of those included in the PhyloSift package [298]. Representative genomes for protists and green algae were retrieved from National Center for Biotechnology Information (NCBI) in addition to two diatom genomes from JGI (*Thalassiosira pseudonana, Phaeodactylum tricornutum*), for a total of 412 reference genomes in addition to the 21 eukaryotic MAGs. A complete list of taxa included is available in Appendix A.5.

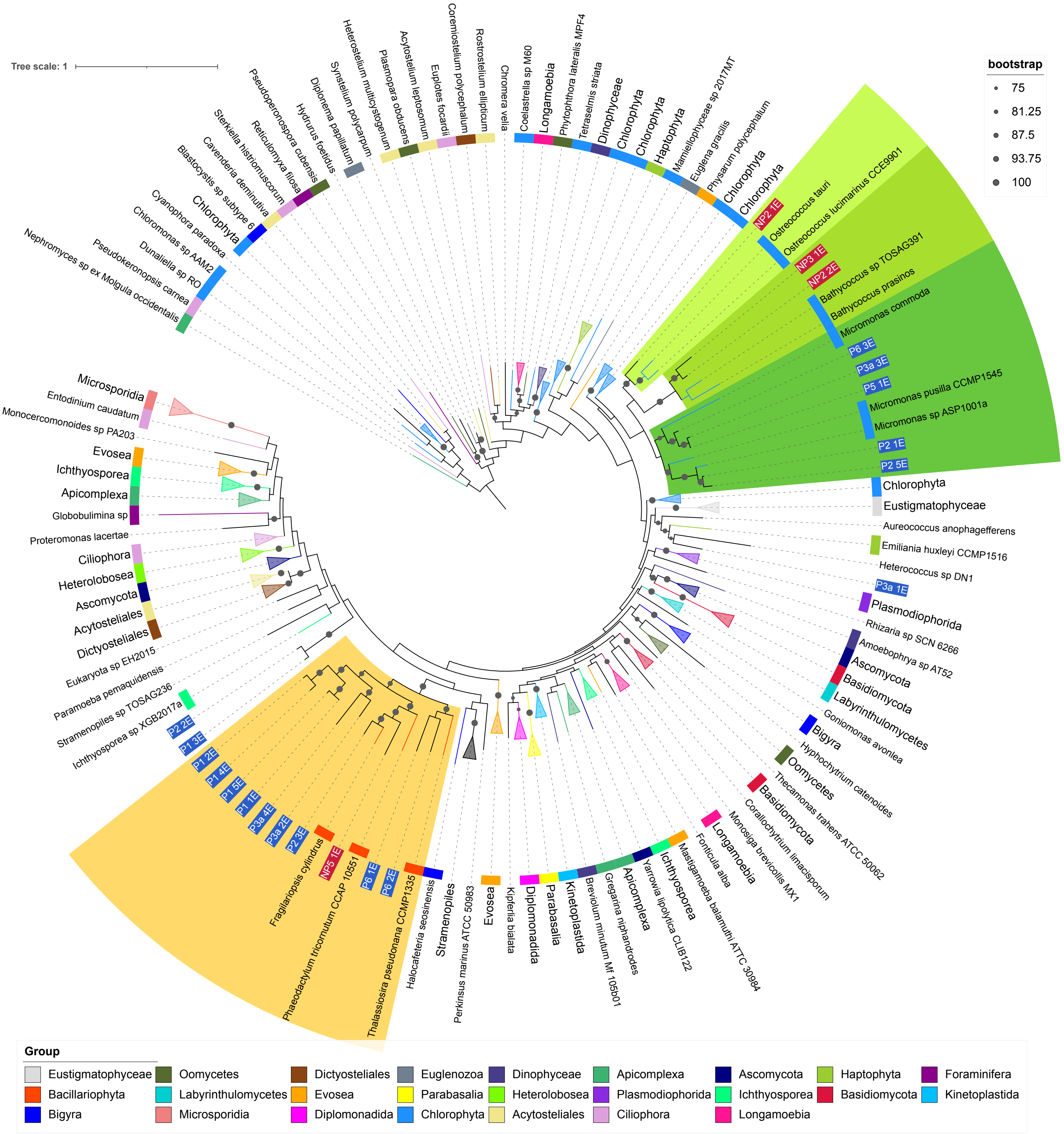


Fig. 4.12 Phylogenomic tree for eukaryotic MAGs and reference genomes. Label and inner band colour indicate taxonomy of reference genomes, using the NCBI taxonomy. MAG labels have a blue background for polar MAGS and a red background for non-polar. Clades which contained reference genomes all from the same taxonomic group have been collapsed. Coloured ranges highlight clades where MAGs place with reference genomes of a consistent taxonomy. Bootstrap values are indicated by grey dots on branches.

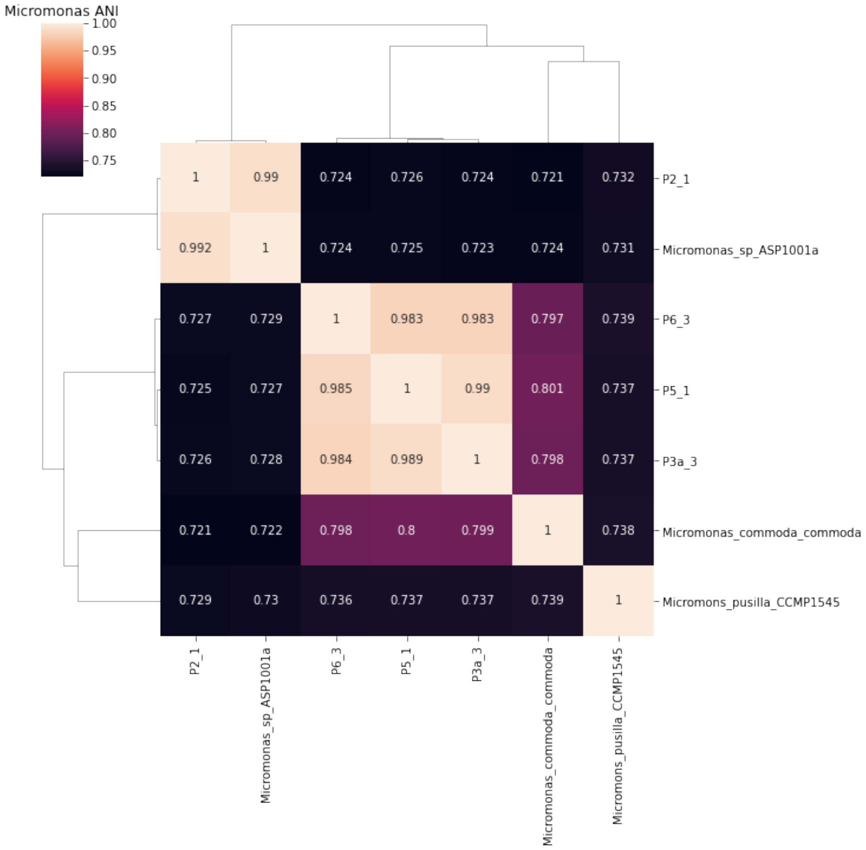
Most MAGs placed in two clades, which contain all the Bacillariophyta or Mamiellophyceae reference genomes. Branches within these clades are long, and more specific identification via this phylogenomic tree construction method seems difficult while there are still few reference genomes available for eukaryotic marine microbes. Within the Mamiellophyceae clade, three MAGs (P6_3, P5_1 and P3a_3) are closely related to one another, but relationships to the reference genomes is more distant. The Bacillariophyta clade appear to have more distant relationships, with no MAGs which appear closely related. P2_2E and P1_3E are difficult to estimate a taxonomy for, placing close to one another but distant from any of the included reference genomes.

Mamiellophyceae

Mamiellophyceae are a class of green algae, the largest clade in the Prasinophyte lineage (see Section 2.3.2). Mamiellophyceae-like MAGs appear to further divide into three clades containing reference genomes from the three genera: Micromonas, Bathycoccus, and Ostreococcus. Micromonas MAGs were only recovered from the polar samples, and Bathycoccus and Ostreococcus from non-polar samples.

Among the Micromonas MAGs some have high ANI to each other or reference genomes, shown in Figure 4.13. MAGs P2_1 and P2_4E have 99% ANI with *Micromonas* sp. 1001a, a species reconstructed from an Antarctic metagenome [237]. Three MAGs appear similar: P6_3, P5_1 and P3a_3. ANI between these MAGs is 98% or higher, and 99% between P5_1 and P3a_3. This group do not share high ANI with any of the reference genomes however. For the Mamiellophyceae AAI supports the placements in the phylogenomic trees. For instance, NP2_1E is placed close to Ostreococcus references in the three and shows the highest AAI of 73.46% to *Ostreococcus lucmarinus*.

Assignment of the contigs from Mamiellophyceae-like MAGs based on searching against MMETSP showed consistency with the taxonomy suggested by the phylogenomic tree. Summarising to the level of phylum, all but NP2_1 have over 99% of contigs assigned to Chlorophyta or a descendant. The contigs which were not assigned to Chlorophyta were either assigned to the Eukaryota node, or had no BLAST hits, and no contigs were assigned to other phyla. This suggests a consistent taxonomic origin for the sequences in these MAGs at least at the phylum level, rather than representing sequences which are not biologically related. Evidence from these BLAST searches also supports the taxonomies suggested by the phylogenomic tree at the genus level; all Mamiellophyceae MAGs had at least 87% of their contigs assigned to the genus they placed with in the phylogenomic tree. Less confirmatory evidence is available for NP2_1. Contigs with no BLAST hits made up 34.12% of the contigs. For those contigs which did have hits, 96% were assigned to Chlorophyta, with the remaining



	P2_1	Micromonas_sp_ASP1001a	P6_3	P5_1	P3a_3	Micromonas_commoda_commoda	Micromons_pusilla_CCMP1545
P2_1	1	0.99	0.724	0.726	0.724	0.721	0.732
Micromonas_sp_ASP1001a	0.992	1	0.724	0.725	0.723	0.724	0.731
P6_3	0.727	0.729	1	0.983	0.983	0.797	0.739
P5_1	0.725	0.727	0.985	1	0.99	0.801	0.737
P3a_3	0.726	0.728	0.984	0.989	1	0.798	0.737
Micromonas_commoda_commoda	0.721	0.722	0.798	0.8	0.799	1	0.738
Micromons_pusilla_CCMP1545	0.729	0.73	0.736	0.737	0.737	0.739	1

Fig. 4.13 Average nucleotide identity of all genomes placed in the Mamiellophyceae clade.

assigned across Stramenopiles, Alveolata, Haptophyta, Rhodophyta, and Cryptophyta. Trees showing assignment of contigs for a subset of the Mamiellophyceae are shown in Figure 4.14, including two with clear assignment at phylum level, and NP2_1 with less clear assignment.

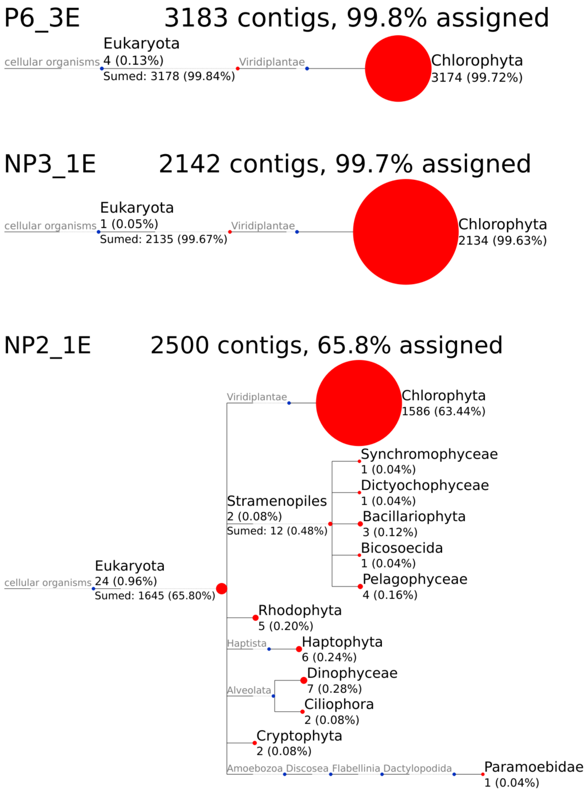


Fig. 4.14 Assignment of contigs in three Mamiellophyceae MAGs based on BLAST search against MMETSP database, summarised at phylum level. Nodes in red had at least one contig assigned. Size of node is scaled to number of contigs assigned. On internal nodes, the number above the branch is contigs assigned directly to that node, below the branch is the number assigned to that node or any of its descendents.

Bacillariophyta

The largest group of MAGs were those which placed in a clade with Bacillariophyta reference genomes, accounting for 8 of the 22 eukaryotic MAGs. Bacillariophyta are a phylum of diatoms, characterised by their silica frustules and estimated to account for 20% of ocean carbon fixation, which were introduced in Section 2.3.2. Within the Bacillariophyta clade in our tree, branches are much longer than the Mamiellophyceae clade, both between pairs of MAGs, and between MAGs and reference genomes. ANI and AAI provide extra evidence suggesting genus level identification of some of the Bacillariophyta MAGs however.

P2_3E had an ANI of 85.5% and AAI of 83.15% to *Fragilariopsis cynlindrus*, supporting their close placement (Figure 4.15). The next highest AAI among Bacillariophyta MAGs is much lower, 66.99% between NP5_1E and *Pseudo-nitzschia multiseries*. MMETSP contains sequences from Bacillariophyta taxa which currently lack a complete genome, results from searching sequences in MAGs against this database provided further evidence for taxonomy which could not be captured by the phylogenomic tree. Apart from MAG P3a_4E, all the MAGs in the Bacillariophyta clade had 85% or more of their assigned contigs classified at the level of phylum when searched against MMETSP. For P3a_4E, of the contigs which could be assigned a taxonomy (35.8%), a majority (23.58%) were classified as Bolidophyceae, a sister taxa to Bacillariophyta. When selecting reference taxa for the phylogenomic tree, no Bolidophyceae genomes were available on NCBI, however a genome has since been made available for *Triparma laevis* [313]. Additional close placement was obtained for P6_2E for which ca. 84% of contigs were classified as *Leptocylindrus danicus*, P3a_2E for which 96.14% of contigs were classified as *Minutocellus polymorphus* and P1_5E for which ca. 85% of contigs were classified as *Chaetoceros neogracilis*. P1_1E shows a high ANI to our potential Chaetoceros MAG P1_5E; however, a lower proportion of contigs in P1_1E (ca. 69%) were assigned to Chaetoceros.

The results of searching these MAGs against MMETSP showed a mean 37.37% percent of contigs in each MAG with no hits, considerably greater than 4.45% for Mamiellophyceae. Along with the long branches in the phylogenomic tree, this suggests that the Bacillariophyceae are more distant from currently available reference genomes than the green algae Mamielliophyceae.

A similar eukaryotic binning effort was published shortly after our binning work, recovering over 700 eukaryotic genomes: 683 eukaryotic MAGs along with 30 SAGs [152]. The genomes total 25.2 Gbp in length with 10,207,450 predicted genes, originating from 280 billion reads from the 798 samples from the Tara Oceans expeditions. Although our dataset is smaller at approximately 1.5% the size in terms of reads (4.5 billion reads from 12 samples), we recovered MAGs at a similar ratio of approximately 9 billion reads per Gbp recovered, compared to 11 billion reads per Gbp recovered in the Tara Ocean dataset. Thus, starting from a more restricted dataset, it is still possible to recover a comparable volume of MAGs as exemplified for Bacillariophyta MAGs. Although the number and diversity of retrieved Bacillariophyta MAGs are higher in the Tara Oceans dataset, our set of MAGs is distributed over a significant number of clades, as shown in the tree combining MAGs from both studies in Figure 4.17. Hence, smaller metagenome studies are still providing access to uncultured genomic microbial diversity and their MAGs.

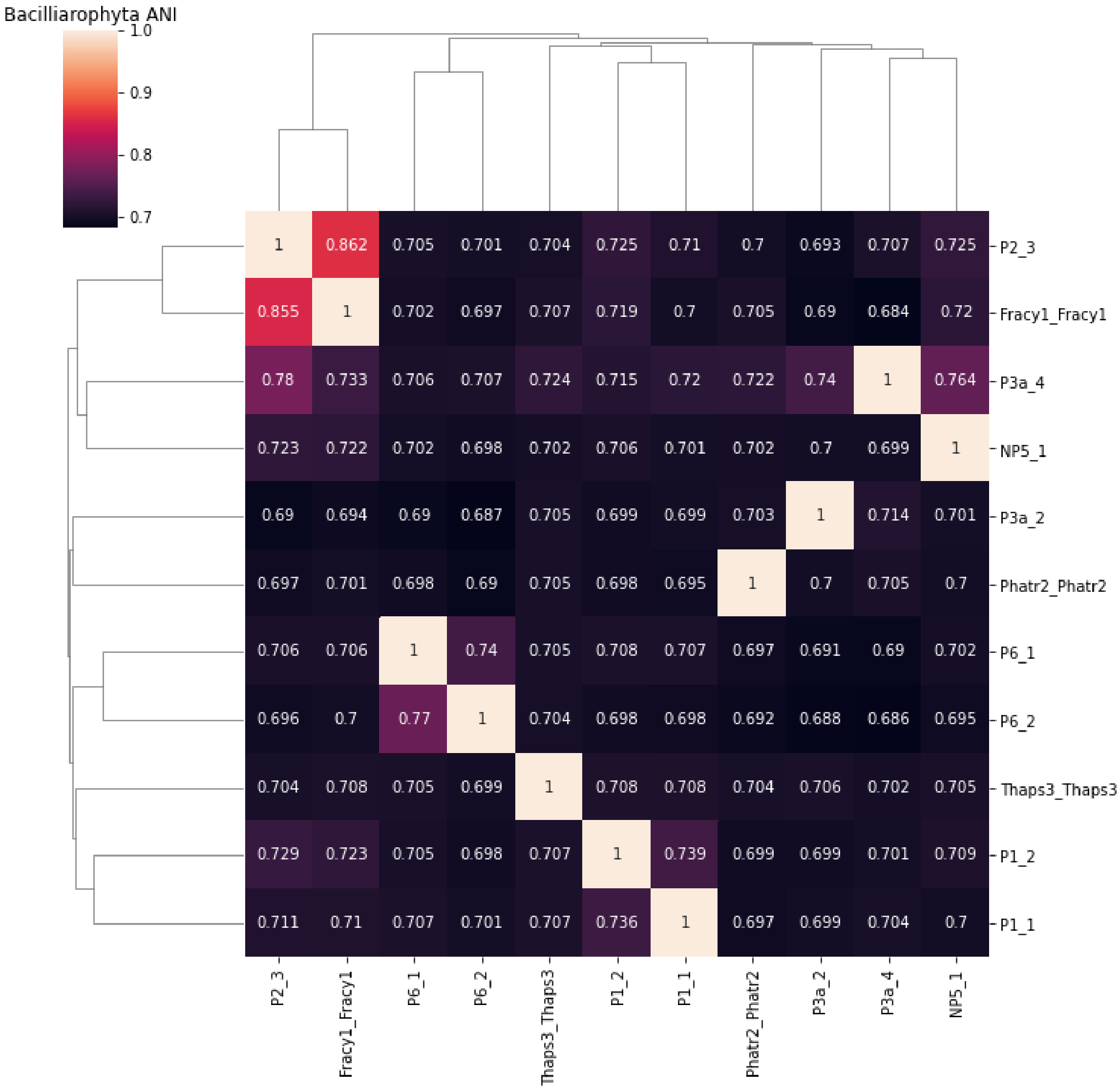


Fig. 4.15 Average nucleotide identity of all genomes placed in the Bacillariophyta clade.

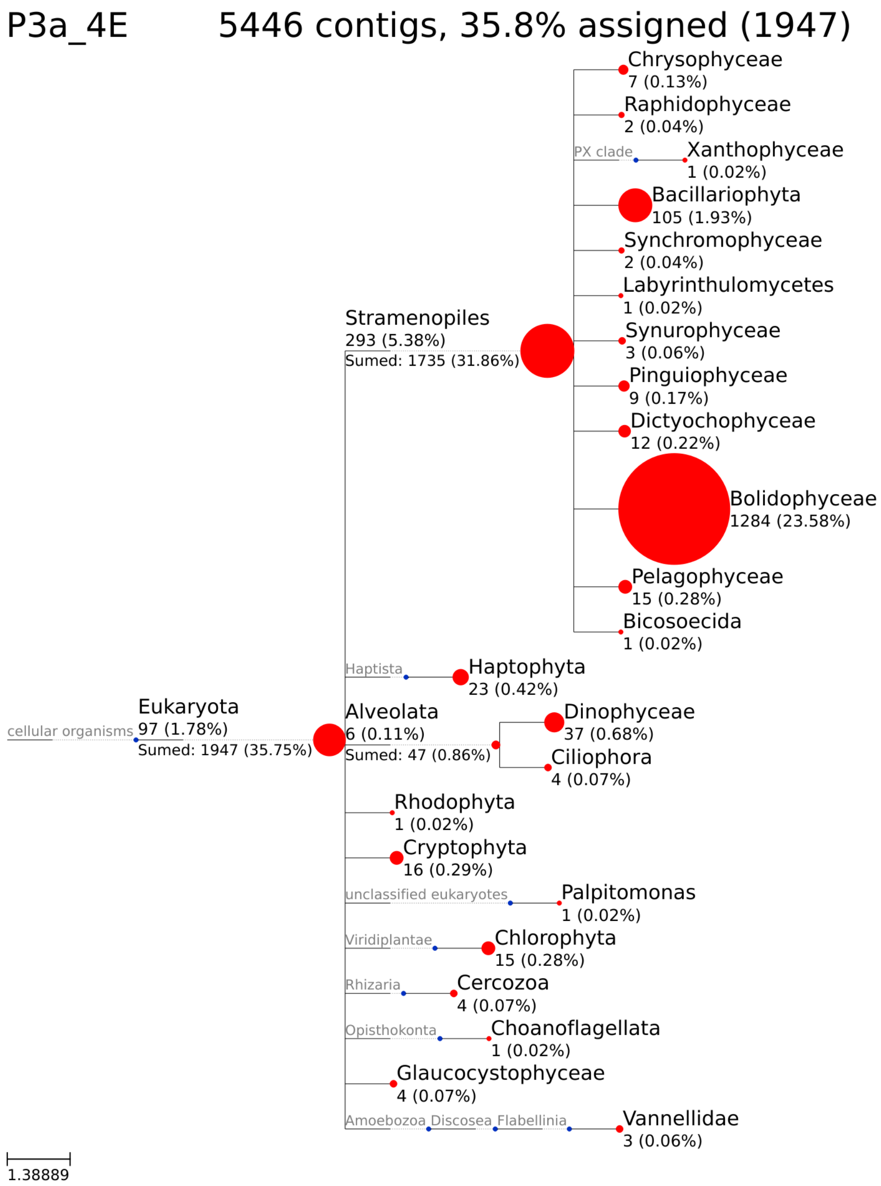


Fig. 4.16 Assignment of contigs in *P3a_4*, based on BLAST search against MMETSP database, summarised at phylum level. Nodes in read had at least one contig assigned. Size of node is scaled to number of contigs assigned. On internal nodes, number above the branch is contigs assigned directly to that node, below the branch is the number assigned to that node or any of it's descendents. The phylum with most contigs assigned is Bolidophycaeae, a phylum for which no complete genome was available at the time of carrying out the research.

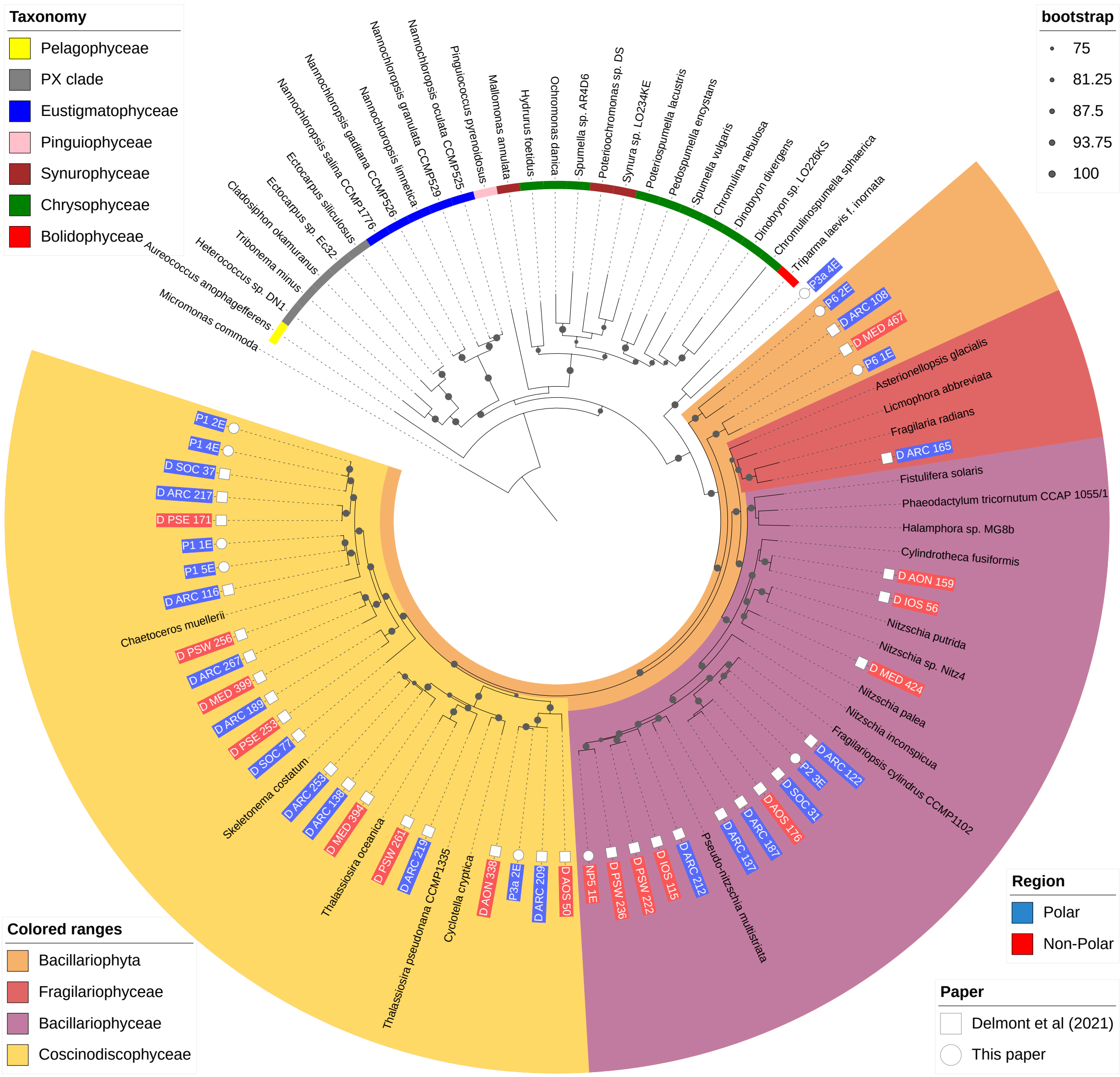


Fig. 4.17 Phylogenomic tree of combining diatom MAGs from two sources. The tree was constructed using a subset of BUSCO genes. Our Bacillariophyta MAGs, along with non-redundant Bacillariophyta MAGs from Delmont et al. [152], were included, along with 18 Bacillariophyta reference genomes, 26 other Ochrophyta references, and Micromonas commoda as an outgroup. Leaf labels indicate whether MAGs originated from polar or non-polar data, blue and red respectively. The paper the MAGs originate from is indicated by either a square or circle. Grey dots show bootstrap values.

Haptophyta

Haptophyta is a phylum of algae including the Coccolithophores who are characterised by their calcium carbonate scales, among which *Emiliania huxleyi* is one of the most abundant and broadly distributed, including expanding into polar waters (Section 2.3.2). The MAG P3a_1 placed closest to the Haptophyta *Emiliania huxleyi*. *E. huxleyi* is quite distant from the other two Haptophyta *Chrysocromulina parva* and *Chrysocromulina* sp. CCMP2291, which are from the Prymnesiales order. These two Prymnesiales placed as neighbouring leaves, and showed 97% ANI. *E. huxleyi* and *P3a_1* have much lower ANI with each other and the two Prymnesiales genomes (ca. 73%), shown in Figure 4.18. This MAG showed the highest AAI with a group of Haptophyta including Phaeocystis and Chrysochromulina species, with the highest being 62.59% AAI with *Phaeocystis antarctica*.

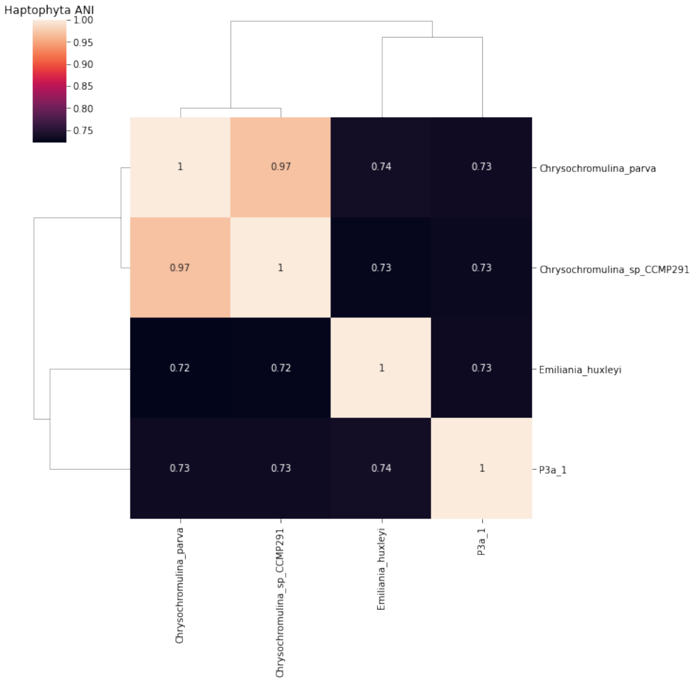


Fig. 4.18 Average Nucletotide Identity of Haptophyta reference genomes and MAGs.

Searching contigs from P3a_1 against MMETSP, a majority of contigs with hits were assigned to a range of Haptophyta taxa which included *E. huxleyi* among them, with most being assigned to *P. Antarctica*, supporting the AAI results. Contigs were also assigned to several other phyla as well, possibly due to MAG contamination.

P2_2 & P1_3

The two mags P2_2 and P1_3 are placed close to each other, but distant from any reference genomes. Searches against MMETSP assigned contigs very widely across multiple phyla, with less than 10% of contigs assigned to any taxa. Taxonomy for these two is difficult to assign based on these two forms of evidence.

Prokaryotes

The phylogenomic tree for prokaryotes in Figure 4.19 was constructed using concatenated alignments of 38 marker genes, a subset of the 40 prokaryotic marker genes included in PhyloSift gene set. 970 reference genomes for marine prokaryotes were retrieved from the MarRef database [191], a complete list of reference taxa is available in Appendix A.4. The tree includes MAGs in which 50% or more of the selected marker genes were identified, resulting in 88 of the 122 prokaryotic MAG being included. The largest group consists of 31 MAGs which placed within a clade with alpha-, beta-, and Gammaproteobacteria references. A further 24 placed with Bacteroidota, of which 17 are in clades of Flavobacteriales.

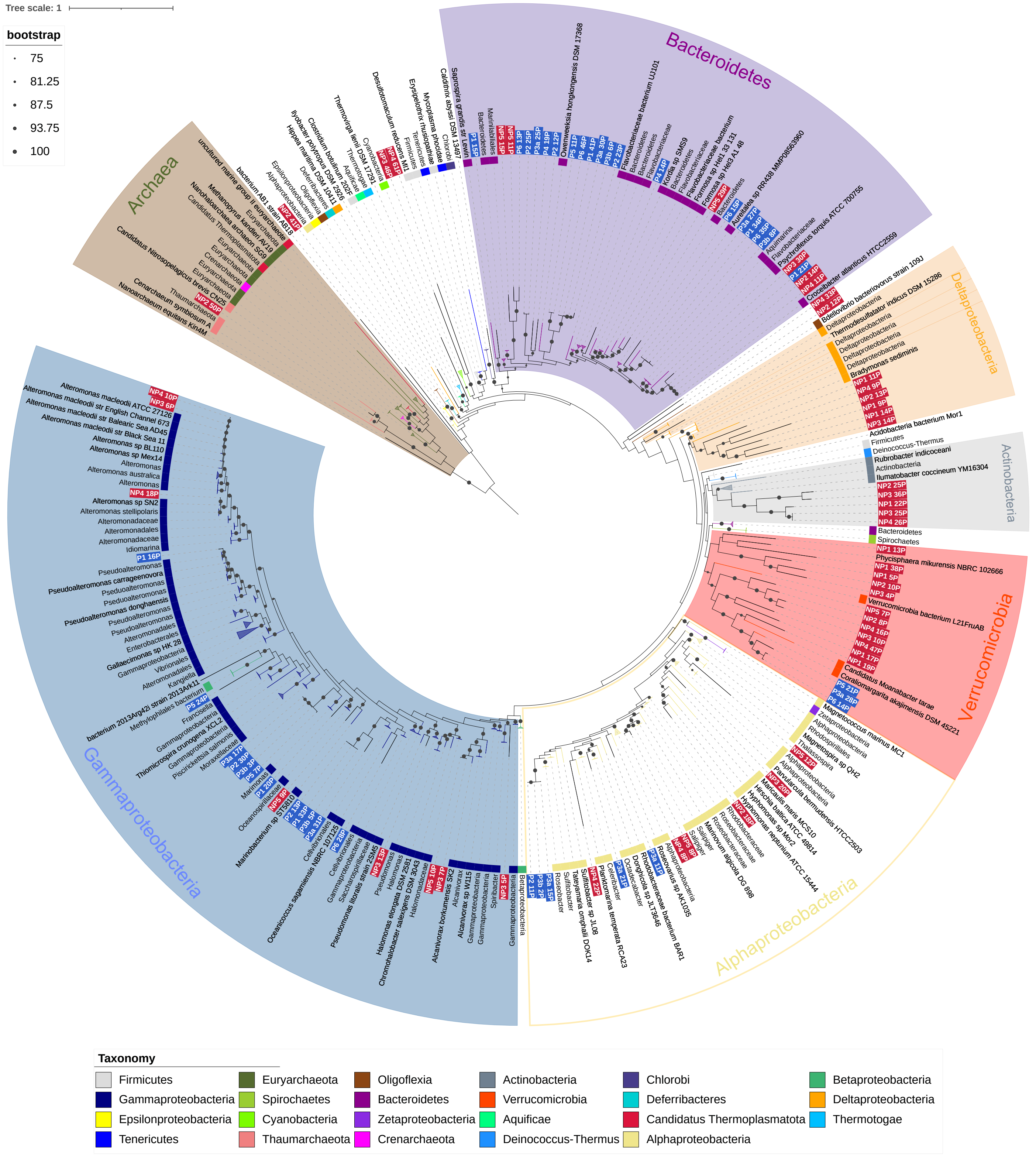


Fig. 4.19 Phylogenomic tree including prokaryotic MAGs and MarRef reference genomes. Inner band colour indicates taxonomy of reference genomes, using the NCBI taxonomy. MAG labels have a blue background for polar MAGS and a red background for non-polar. Clades which contained reference genomes all from the same taxonomic group in the legend have been collapsed; the size of triangle is scaled to the number of leaves in the collapsed clade. Collapsed clades have been given labels which encompass all the contained leaves. Bootstrap values are indicated by grey dots on branches.

The phylogenomic tree is largely in agreement with the taxonomies predicted by GTDB-Tk at the level of phylum. There are some instances where MAGs have not been placed close to any of the included references, such as NP34_33P and NP2_12P, where GTDB suggested a more specific estimate; NP2_12P was assigned to a class of Poribacteria, for which no reference genomes are included in the MarRef data.

Some MAGs recovered from different stations appear closely related to one another. NP4_10P and NP3_6P are closely related to each other as well as to multiple *Alteromonas macleodii* strains. The reference genomes for *A. macleodii* can be split into those from surface and deep ocean [314]. These MAGs have a greater than 95% ANI to three surface genomes shown in Figure 4.20, suggesting a species level relationship. The ANI between these MAGs and deep ocean *A. macleodii* is below 95%. This is supported by the assignment of contigs within the MAGs based on BLAST searches against the NT database, for both MAGs at least 89% of contigs are assigned to the *A. macleodii* node or a strain below it.

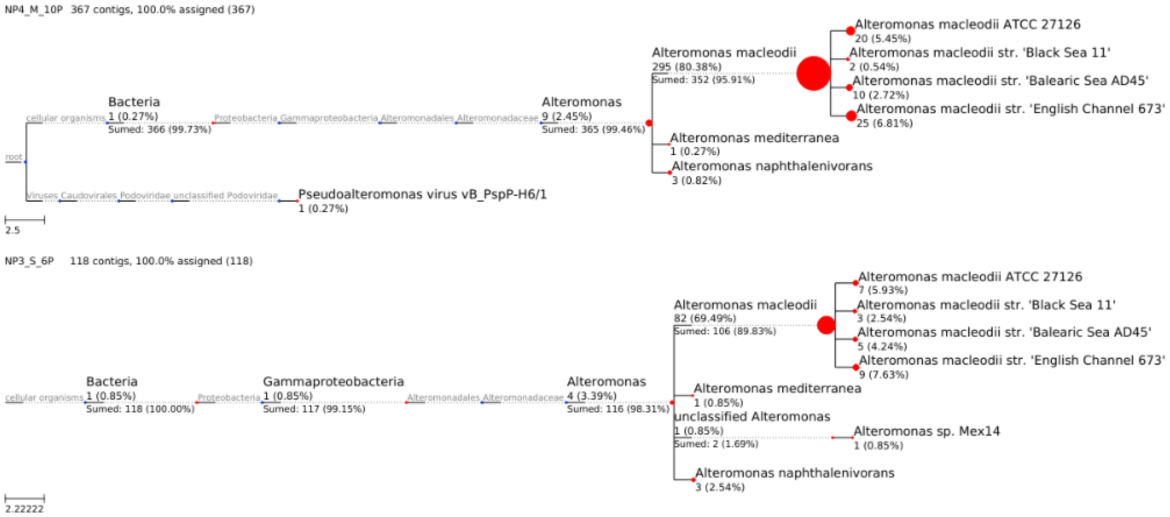
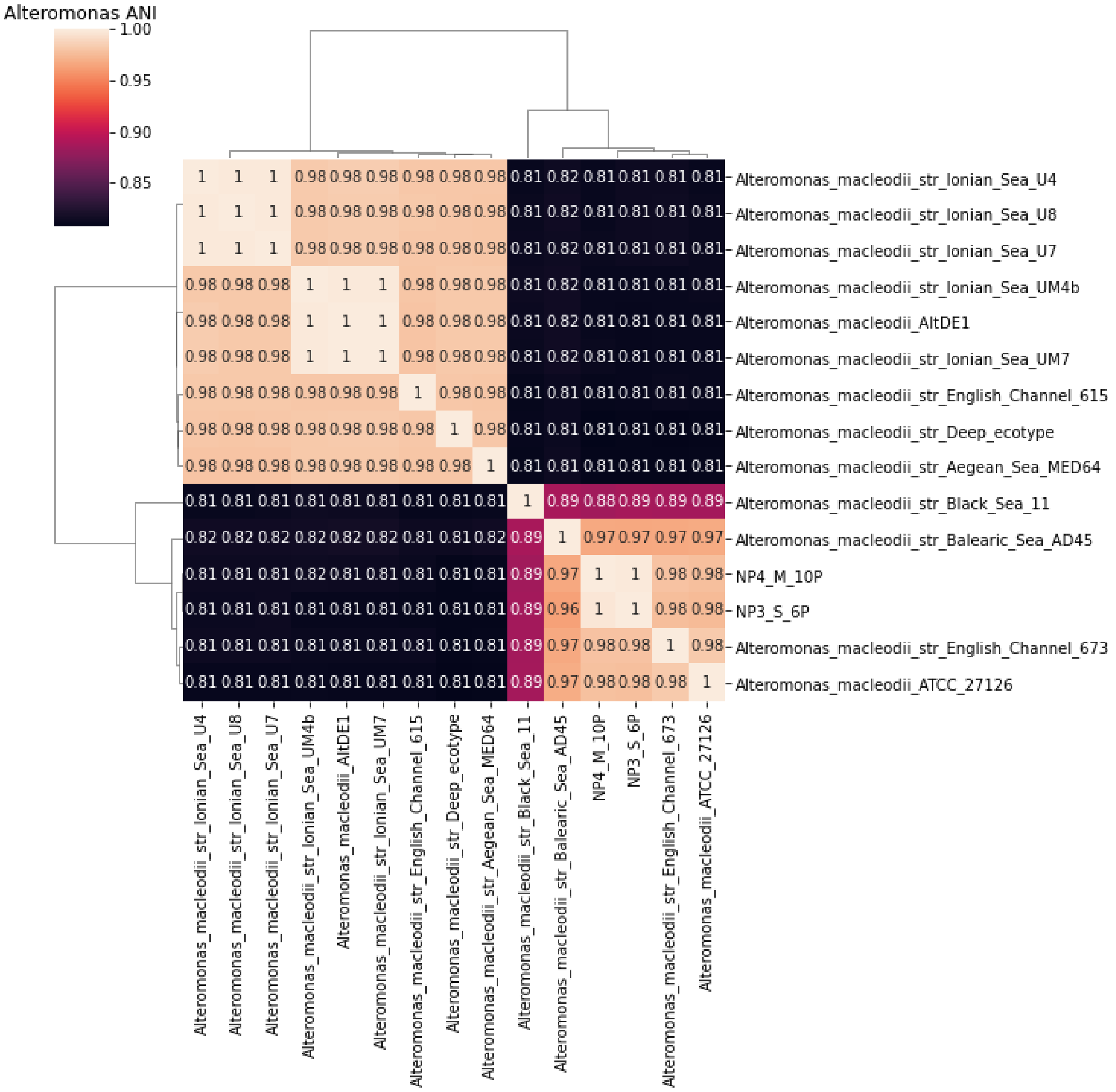


Fig. 4.21 Taxonomic placements of contigs from Alteromonas-like MAGs. Number of contigs assigned to nodes of the NCBI taxonomy by MEGAN-LR based on BLAST results against the NT database. Size of node is scaled to number of contigs assigned. Number above each branch gives the number and percentage of contigs assigned to that node, number below the branch the same but for that node or any descendents.

Other groups of MAGs display similarly close relationships to each other, but are more distant from reference genomes. Four polar MAGs which placed among Bacteroidetes, P6_35P, P3b_8P, P1_34P, and P3a_27P, share over 95% identity to each other, but less than that to their closest reference genome, an unclassified species of genus Aureitalea. The results of assigning contigs via BLAST searches is similarly mixed, most contigs being assigned to a mix of Flavobacterieaceae or uncultured bacterium. A representative example



	Alteromonas_macleodii_str_Ionian_Sea_U4	Alteromonas_macleodii_str_Ionian_Sea_U8	Alteromonas_macleodii_str_Ionian_Sea_U7	Alteromonas_macleodii_str_Ionian_Sea_UM4b	Alteromonas_macleodii_AltDE1	Alteromonas_macleodii_str_Ionian_Sea_UM7	Alteromonas_macleodii_str_English_Channel_615	Alteromonas_macleodii_str_Deep_ecotype	Alteromonas_macleodii_str_Aegean_Sea_MED64	Alteromonas_macleodii_str_Black_Sea_11	Alteromonas_macleodii_str_Balearic_Sea_AD45	NP4_M_10P	NP3_S_6P	Alteromonas_macleodii_str_English_Channel_673	Alteromonas_macleodii_ATCC_27126
Alteromonas_macleodii_str_Ionian_Sea_U4	1	1	1	0.98	0.98	0.98	0.98	0.98	0.98	0.81	0.82	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_Ionian_Sea_U8	1	1	1	0.98	0.98	0.98	0.98	0.98	0.98	0.81	0.82	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_Ionian_Sea_U7	1	1	1	0.98	0.98	0.98	0.98	0.98	0.98	0.81	0.82	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_Ionian_Sea_UM4b	0.98	0.98	0.98	1	1	1	0.98	0.98	0.98	0.81	0.82	0.81	0.81	0.81	0.81
Alteromonas_macleodii_AltDE1	0.98	0.98	0.98	1	1	1	0.98	0.98	0.98	0.81	0.82	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_Ionian_Sea_UM7	0.98	0.98	0.98	1	1	1	0.98	0.98	0.98	0.81	0.82	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_English_Channel_615	0.98	0.98	0.98	0.98	0.98	0.98	1	0.98	0.98	0.81	0.81	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_Deep_ecotype	0.98	0.98	0.98	0.98	0.98	0.98	0.98	1	0.98	0.81	0.81	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_Aegean_Sea_MED64	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	1	0.81	0.81	0.81	0.81	0.81	0.81
Alteromonas_macleodii_str_Black_Sea_11	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	1	0.89	0.88	0.89	0.89	0.89
Alteromonas_macleodii_str_Balearic_Sea_AD45	0.82	0.82	0.82	0.82	0.82	0.82	0.81	0.81	0.82	0.89	1	0.97	0.97	0.97	0.97
NP4_M_10P	0.81	0.81	0.81	0.82	0.81	0.81	0.81	0.81	0.81	0.89	0.97	1	1	0.98	0.98
NP3_S_6P	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.89	0.96	1	1	0.98	0.98
Alteromonas_macleodii_str_English_Channel_673	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.89	0.97	0.98	0.98	1	0.98
Alteromonas_macleodii_ATCC_27126	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.89	0.97	0.98	0.98	0.98	1

Fig. 4.20 Average Nucleotide Identity between Alteromonas reference genomes and MAGs which placed close to them. MAGs show higher ANI with surface than deep sea ecotypes.

of placement of contigs is shown in Figure 4.22. These four MAGs could represent members of the same novel species of Bacteroidetes.

There are few close relationships between polar and non-polar MAGs evident in the tree. The median distance from a polar MAGs to the nearest polar MAG is lower than to the nearest non-polar MAG, and the same for non-polar to non-polar shown in Figure 4.23. In both cases the difference in medians is significantly different at $p < 0.01$ using Mood's median test. One clade of Bacteroidetes is an exception, where polar MAG P1_21P appears closely related to NP2_14P, NP3_30P and NP4_11P. The closest reference is *Croecibacter*

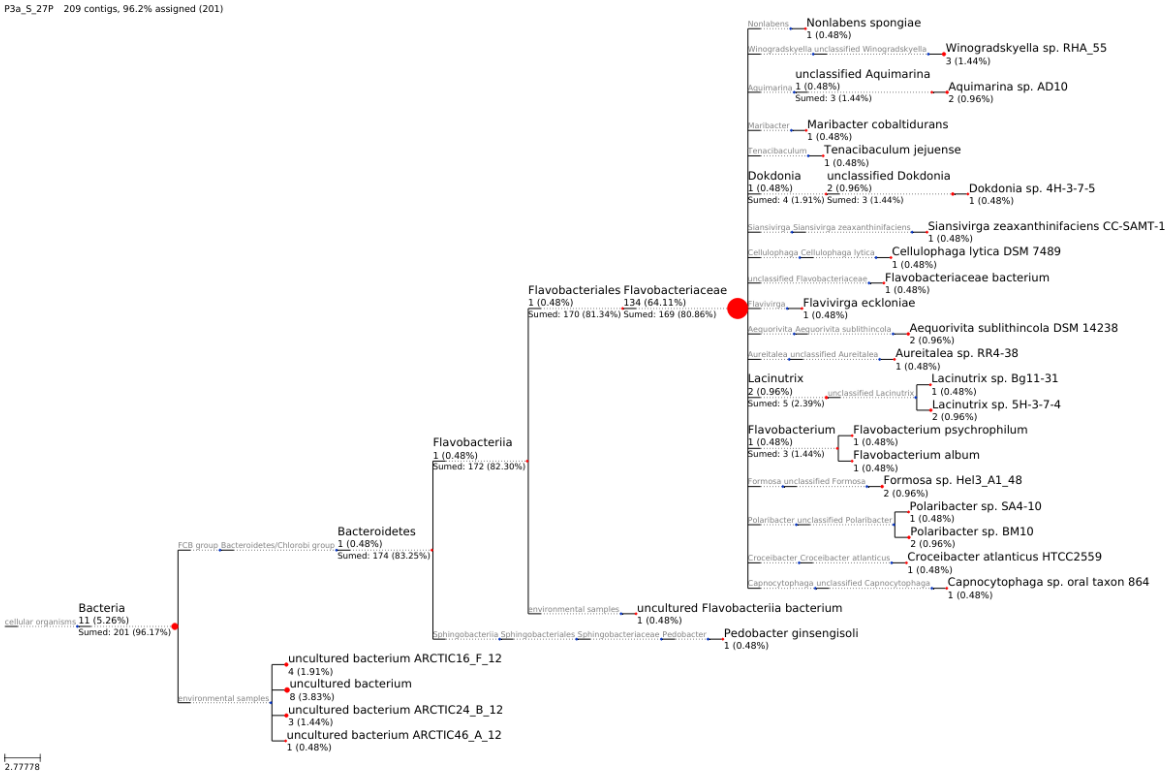


Fig. 4.22 Taxonomic placement of contigs from P3a_27P, one of the MAGs which placed closet to an Aureitalea reference genome. Number of contigs assigned to nodes of the NCBI taxonomy by MEGAN-LR based on BLAST results against the NT database. Size of node is scaled to number of contigs assigned. Number above each branch gives the number and percentage of contigs assigned to that node, number below the branch the same but for that node or any descendents.

atlanticus which is in different clade. Pairwise ANI between these mags and the *C. atlanticus* reference genome is greater than 95%, suggesting these MAGs could represent genomes of the species *C. atlanticus*.

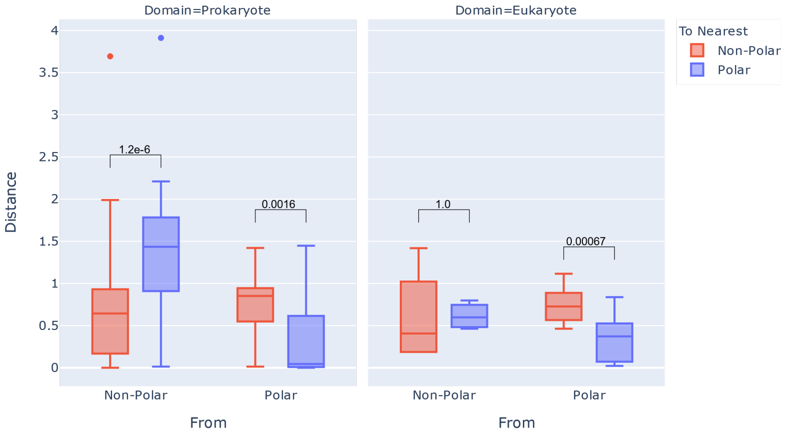


Fig. 4.23 Distribution of tree distances from MAGs to the nearest polar or non-polar MAG. Values between pairs are p-values from Mood's median test.

Some MAGs had been classified at genus level by GTDB-Tk and species level by CheckM, but for which the phylogenomic tree does not suggest a similarly specific classification. MAGs P3a_28P, P6_14P, P5_21P, P2_21P, and P6_33P were classified at genus level as Puniceicoccaceae by GTDB-Tk. In the phylogenomic tree, the first three placed closest to *Coraliomargarita akajimnesis* but with longer branches than observed between taxa from the same species elsewhere in the tree. The latter two lacked the amount of marker genes required to be included in the tree. Looking at the ANI shown in Figure 4.24 also suggests these MAGs and *C. akajimensis* are not the same species, no pair shares above 95% ANI.

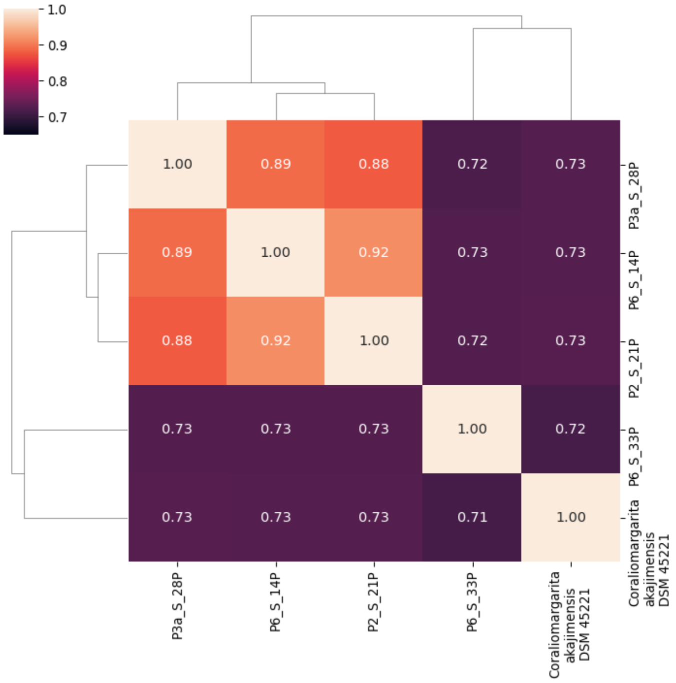


Fig. 4.24 Average Nucleotide Identity between *C. akajimensis* and MAGs which placed close to it. ANI between MAGs and reference genomes is lower than would be expected for genomes of the same species.

4.3.5 Coverage

Each MAG contains contigs which originate from a single assembly, which in turn was generated using reads sequenced from a single sample. The organism represented by this MAG may however be present in other samples, but could not be recovered from that sample's assembly due to reasons such as low abundance and hence low coverage preventing good assembly. To describe the distribution of the identified MAGs across the twelve samples, reads from each sample were mapped to each MAG and the mean coverage for contigs calculated. This mean coverage was then adjusted to account for the differing number of reads from each sample, to obtain mean coverage per million reads, as an estimate of the relative abundance of MAGs in the sampled locations. Mean coverage per million reads in each MAG is shown in Figure 4.25. We adopted the criteria from Olm et al. [236] and considered a MAG not present where less than 50% of bases in the MAG had at least one read aligned to them.

The binning process uses covarying coverage to group contigs into bins, so for highly similar MAGs a similar pattern of coverage across samples would be expected. Five closely related Micromonas MAGs P2_30, P3a_17, P5_7, P3b_3, and P4_17 show this pattern strongly, with a very similar rising and falling coverage from stations P3 to P6. Coverage of MAGs often shows a gradient across geographically close stations. There is a clear demarcation between polar and non-polar MAGs. Of the 122 prokaryotic MAGs, 116 are only present in either polar or non-polar samples. MAGs detected in both tend to be detected in samples P1 and P2 albeit with low coverage.

For eukaryotic MAGs present in a sample, the mean coverage ranged between 0.92 and 87.24, with a mean lower than that of prokaryotes at 17.68. Many of the patterns observed among the prokaryotes hold for eukaryotic MAGs also. Micromonas MAGs P6_3E, P5_1E and P3a_3 which appear similar based on the phylogenomic tree show a very similar pattern of coverage from stations P2 to P6. The demarcation between polar and non-polar stations is clear among eukaryotes, no MAG was found to be on both sides of the Arctic Circle. This coverage is limited to describing only the distribution of those members of the community for which a MAG was recovered. There are phyla which appeared abundant in gene or read based classification (Figures 4.4 and 4.7) such as Haptista, which no recovered MAG clearly belongs. These lineages have the potential to contain widespread species which would not follow the strong demarcation observed.

Approximately half of the Bacillariophyta MAGs were present at only one or two stations maximum whereas Mamiellophyceae MAGs were generally more widespread. The one non-polar Bacillariophyta MAG NP5_1 is present only in stations NP4 and NP5, the southernmost of the non-polar stations. Potential Haptophyte P3a_1E is present in three polar stations, and most abundant at P3, where the Bacillariophyta MAGs are less abundant.

Most of the Bacillariophyta MAGs have a niche occupancy, being present in one or two stations. P3a_2 and P3a_4 are more widespread. Based on BLAST searches against MMETSP, P3a_2 had 96.14% of contigs assigned to picoeukaryote *Minutocellus polymorphus* and *Triparma pacifica. M. polymorphus* has been observed by in Arctic seas [315] and sea ice [316]. Our MAG which appears similar to *M. polymorphus* is spread with coverage across the eastern stations closer to Svalbard and Norway. The other MAG which placed with Bacillariophyta with a wider spread with a peak in coverage at P4 is *P3a_4*, which had similarity to species of Bolidophyceae when searched against MMETSP, with 51.80% of its classified contigs classified as *Triparma pacifica.* Kuawata et al. [317] found high abundance of some clades of Triparma, including *T. pacifica* in similar stations between Svalbard and Greenland.

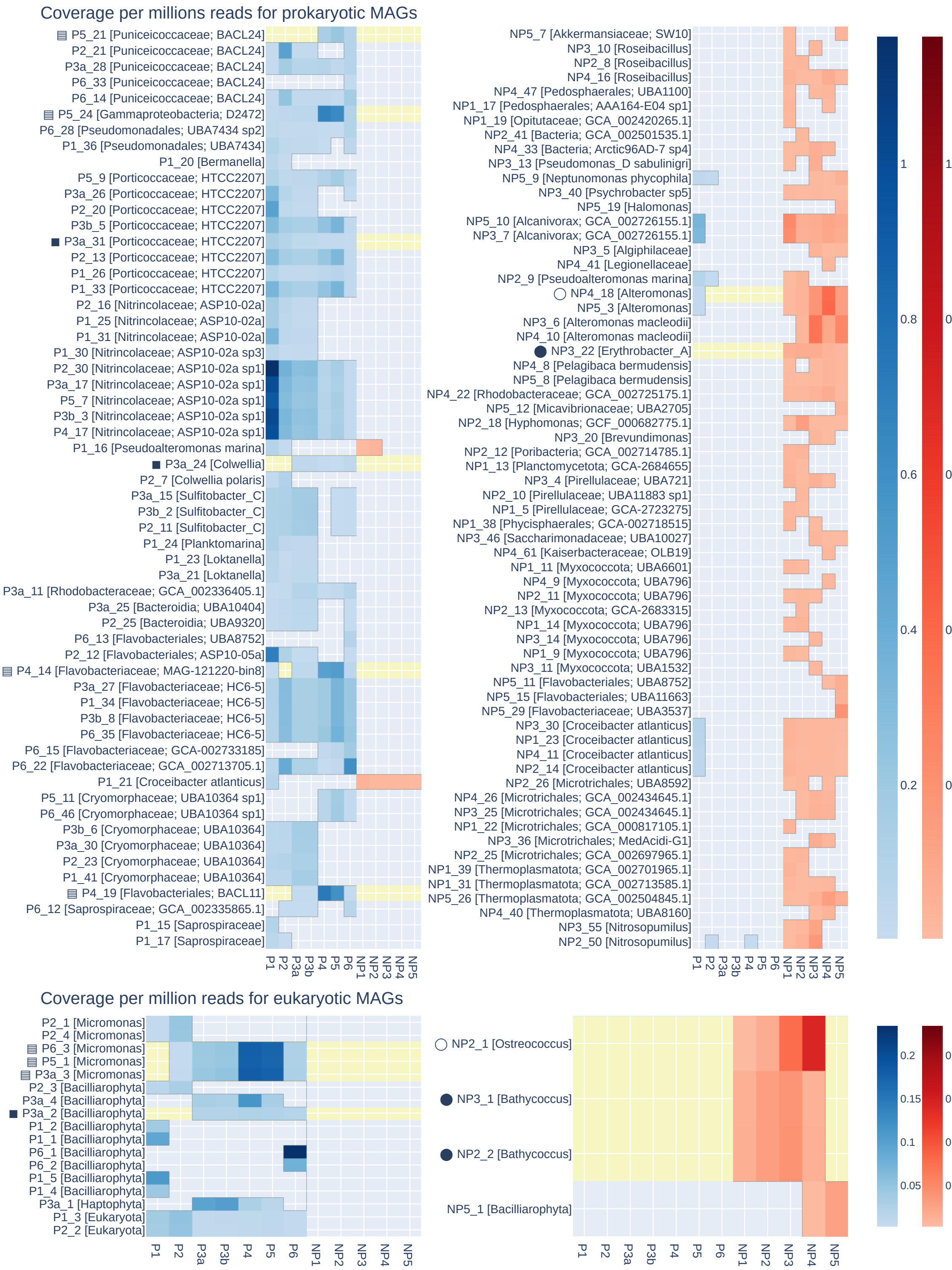


Fig. 4.25 Mean coverage of MAGs when reads from each sample mapped back to contigs in MAGs. Upper heatmaps show prokaryotes, lower show eukaryotes. Left heatmps show MAGs recovered from polar stations, right show non-polar stations. Blue cells show coverage in polar stations, red shows the coverage in non-polar stations. Where MAGs have a detection below 0.5 they are considered not to be present in a sample, and no value shown. Very few MAGs are present in both polar and non-polar stations.

4.3.6 Gene Prediction and Function

Genes were predicted for each eukaryotic MAG using GeneMark-ES [283], and predicted genes annotated using InterproScan [283]. Prokaryotic genes were predicted and annotated by the IMG pipeline prior to binning. We examined the extent to which functions are shared or unique between polar and non-polar environments, for both the population as a whole before binning, and for prokaryotic and eukaryotic MAGs separately. First we look at the Pfam domains, and then at the GO terms associated with these domains, using the mapping of Pfam to GO terms maintained by the GO Consortium.

The Principal Components Analysis (PCA) plot of the Pfam abundance in each MAG in Figure 4.26 shows mostly clear separation into taxonomic groups, supporting the broad classifications drawn from the phylogenomic tree. Clustering by taxonomy is clearer in eukaryotes than prokaryotes. The two large groups of Bacillariophyta and Mamiellophyceae are clearly separated, with the possible Bolidophyceae P3a_4 closer to P3a_1 the potential Haptophyte. Some prokaryotic groups form clear clusters, such as Bacteriodetes and Actinobacteria, while others are more spread such as the Proteobacteria. Most of the MAGs without an assigned taxonomy cluster to the right of the plot, along with a single Verrumicrobia MAG which is separate from the main cluster of Verrumicrobia.

MAGs were grouped based on the region the bin originated from, either polar or non-polar. The number of Pfams observed in these groups is shown in Figure 4.26, for the whole population before binning, and for eukaryotic and prokaryotic MAGs. The whole population showed a majority of Pfams were found in all regions, showing a widely distributed shared core of functions. Among functions unique to one area, prokaryotic MAGs had many more unique functions in the non-polar stations. This is inverted for eukaryotic MAGs which have more domains which were unique to polar stations. A majority of the eukaryotic MAGs were retrieved from polar stations, only 4 of the 21 being from non-polar stations. This imbalance could partially explain the high number of functions unique to polar eukaryotic MAGs. Prokaryotic MAGs were more balanced across the two regions, 65 from non-polar and 58 from polar stations.

Four of the five most abundant Pfam families unique to non-polar prokaryotic MAGs are PSD1, 3, 4, 5 & C. These are domains of unknown function shared by cytochrome-like proteins in the planctomycete species *Rhodopirellula baltica*. Three MAGs classified as planctomycetes were recovered, all from non-polar stations. These domains were found in 24 of the 65 non-polar prokaryotic MAGs, which were assigned to a wide taxonomic range: Acidimicrobiia, Actinobacteria, Alphaproteobacteria, Gammaproteobacteria, Planctomycetaceae, and MAGs which were classified as either Bacteria or unclassified. All five proteins are typically found together in MAGs only NP2_9 contained one (PSD3) without

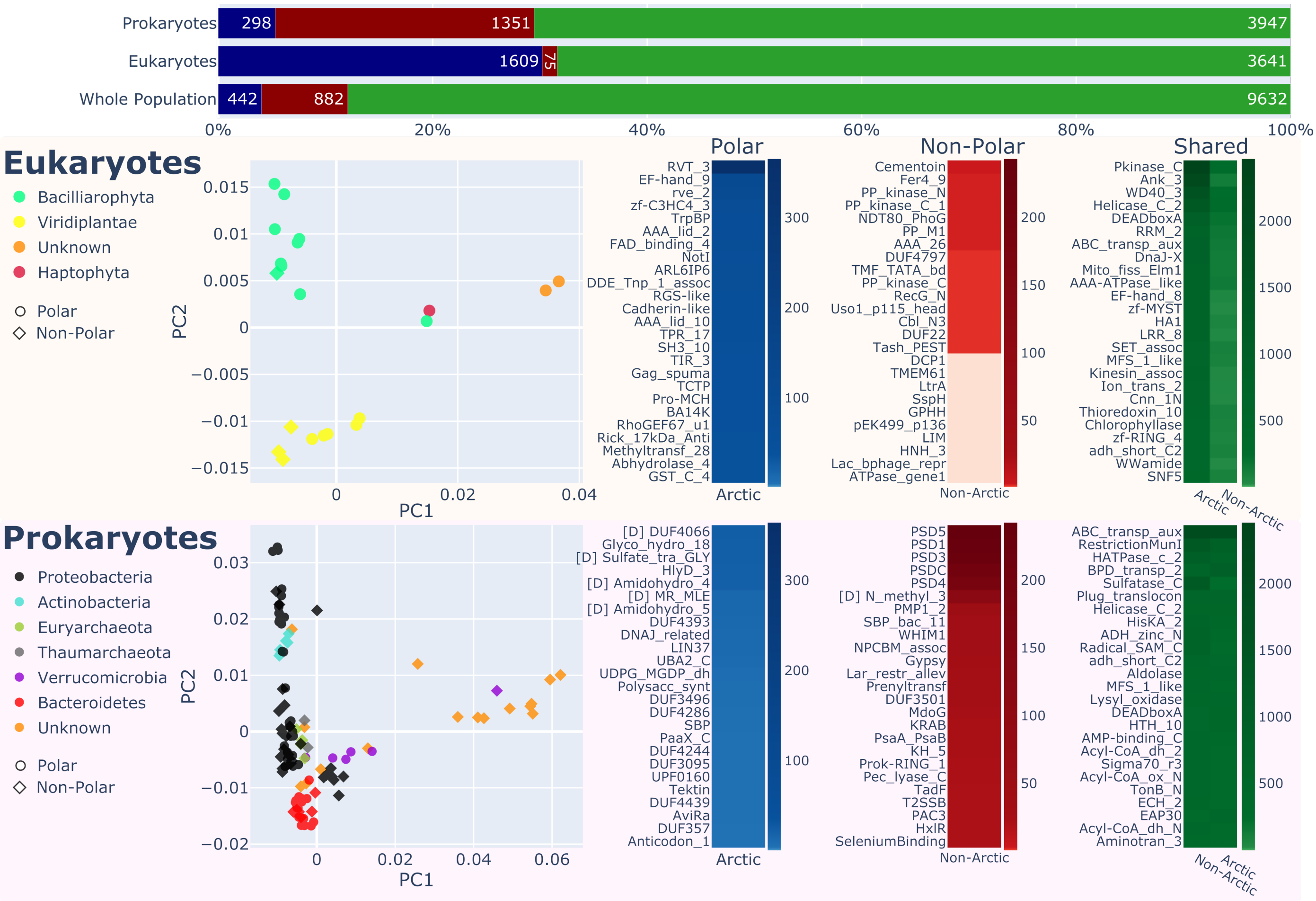


Fig. 4.26 At the top, horizontal stacked bars how the number of Pfam annotations present in both regions, labelled shared, or only found in polar or non-polar regions. Size of each bar scaled to the percentage of all domains that accounts for, number of domains in that group labelled on the bar. One set of bars is shown for the whole population prior to binning, the eukaryotic MAGs and the prokaryotic MAGs.From left to right. Below, the two shaded areas related to eukaryotic MAGs (peach, top) and prokaryotic MAGs (pink, bottom). The scatter plot shows a PCA ordination of the proportion of PFAM annotations in each MAG. Colours in these plots identify the taxonomy which had been assigned to the MAG. Heatmaps to the right indicate the 25 most abundant PFAMs which are unique to an area or among those shared.

the others being present. The three planctomycete MAGs are richer in these domains than others, accounting for 27.59% of the non-polar unique PSD domains. Along with with the 15 MAGs classified only at the level of Bacteria making up 67.93% these two groups account for a majority.

Related domains PSCyt2, PSCyt3, PSD2 are shared between polar and non-polar, being found in four polar MAGs all identified as Puniceicoccaceae. When we mapped from

Protein Family (Pfam) accessions to GO terms using the mapping maintained by InterPro, photosynthesis related terms were found only in non-polar MAGs.

Eukaryotic MAGs have more Pfams unique to polar environments. The most abundant of these is RVT_3, a domain believed to part of a retrotransposon found in plants. RVT_3 was most abundant in two of the Bacillariophyta MAGs P6_1 with 125 genes with this annotation and P6_2 with 144. This domain has been observed in complete genomes for Bacillariophyta, but in lower numbers. The IMG database shows *Phaeodactylum tricornutum* has 3 genes containing RVT_3, *Thalassiosira pseudonana* containing 1. MAGs P6_1 and P6_2 also contain a high number of of genes with rve_2 domains. This domain is less common among eukaryotes, the Pfam databases shows homologous sequences only in two species of Metazoa and Fungi. rve_2 is an integrase catalytic domain, present in transposase proteins as well as catalysing reactions involved in the integration of viral genomes into host genomes. The combination of high number of these two domains in P6_1 and P6_2 could suggest they share a high level of transposase activity.

Mapping to GO terms shows the same broad trends as for Pfams as might be expected. PCA analysis of the GO term abundance groups MAGs strongly by taxonomy for eukaryotes, with some clear grouping for prokaryotes; and the same pattern of a widely shared core functions in the whole population, and greater niche functions for polar eukaryote and non-polar prokaryote MAGs is evident.

The higher level functional summary available from GO terms showed some additional functional differences between regions. Terms related to cold exposure are among the most abundant terms observed only in polar environments. Ice binding (GO:0050825) is unique to the polar eukaryotic MAGs. Ice binding proteins have been observed in a wide range of organisms across the biological kingdoms, including diatoms and marine bacteria [318]. The proteins encompass a range of activities; among polar algae, recrystallisation inhibition has been suggested to maintain brine pockets which form during the freezing of seawater, providing a viable habitat in freezing conditions [319]. Among prokaryotes, the most abundant term unique to polar MAGs is heat shock protein binding (GO:0031072). Heat shock proteins were observed to be expressed in arctic Rhizobium species in response to heat stress [320] and in response to suboptimal temperatures in *Alicyclobacillus acidoterrestris* [321]. Terms related to photosynthetic activity in prokaryotes are unique to non-polar MAGs, with photosynthesis and photosystem II (GO:0015979, GO:0009523) among the most abundant unique terms. Some differences appear driven by the taxonomy of MAGs recovered in the two areas. Micromonas have flagellum-based motility, and Micromonas MAGs were only recovered in polar samples. Consequently, some terms related to flagella such as cilium assembly (GO:0060271), which is considered equivalent to microtubule-based

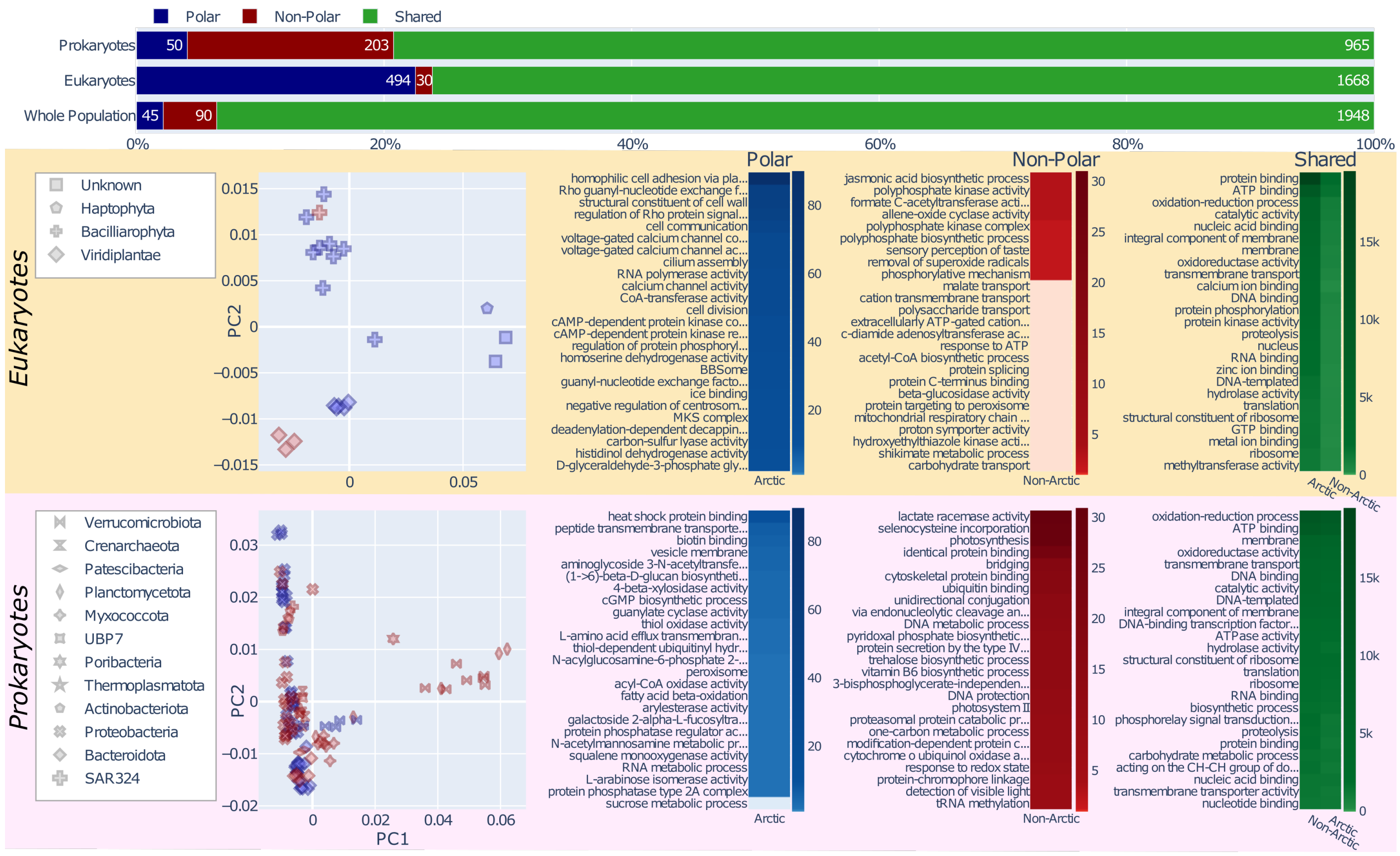


Fig. 4.27 At the top, horizontal stacked bars show the number of GO term annotations present in both regions, labelled shared, or only found in polar or non-polar regions. Size of each bar scaled to the percentage of all domains that accounts for, number of domains in that group labelled on the bar. One set of bars is shown for the whole population prior to binning, the eukaryotic MAGs and the prokaryotic MAGs. From left to right. Below, the two shaded areas related to eukaryotic MAGs (peach, top) and prokaryotic MAGs (pink, bottom). The scatter plot shows a PCA ordination of the proportion of GO terms in each MAG. Colours in these plots identify the taxonomy which had been assigned to the MAG. Heatmaps to the right indicate the 25 most abundant Pfams which are unique to an area or among those shared.

flagellum assembly, is unique to the non-polar MAGs. Some of the unique polar terms are driven by the two unidentified MAGs P2_2E and P1_3E, such as the most abundant unique polar term "homophilic cell adhesion via plasma membrane adhesion molecules" (GO:0007156), for which 95% of annotations were observed in genes from these unidentified MAGs.

4.3.7 Inter-Kingdom Associations

To identify associations between eukaryotic and prokaryotic MAGs, we looked for pairs where the coverage per million reads was correlated. We set this threshold as an $R^2 \geq 0.7$ and p-value ≤ 0.05. This initially resulted in 38 associations, however plotting these associations shows that some appear to be driven by a small number of influential high values mixed with a majority of low or zero values. Influential values were identified as those with a Cook's distance of > 1.25, and are indicated by a star in Figures 4.28, 4.29, and 4.30. Influential points were removed, and associations kept only if they met the same correlation and significance criteria.

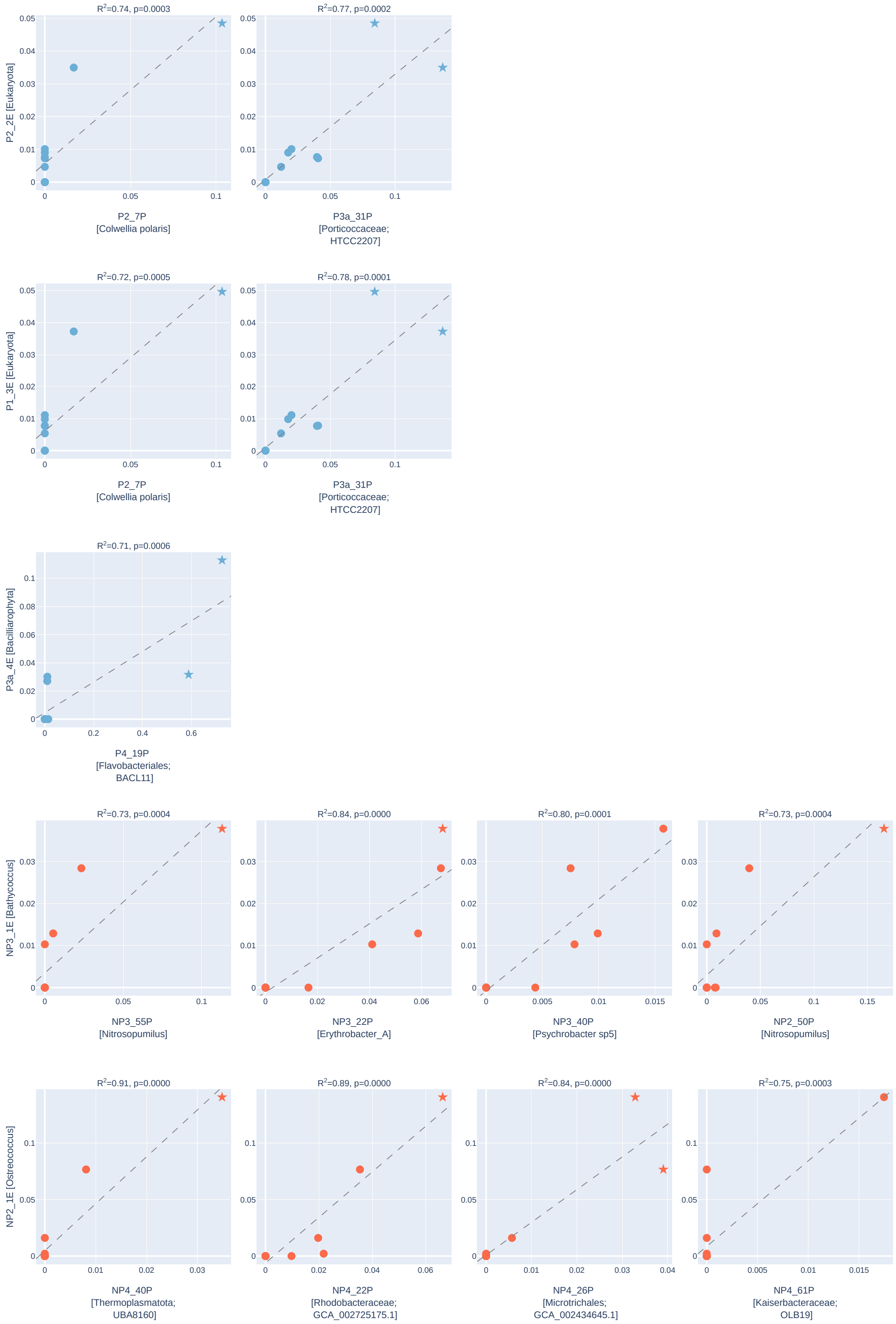


Fig. 4.28 Coverage of eukaryote/prokaryote pairs which appear associated. Some associations appear to be driven by a small number of points with high values. Influential points with Cook's distance ≥ 1.25 are indicated with stars. Split across multiple figures for legibility, see Figures 4.29 & 4.30.

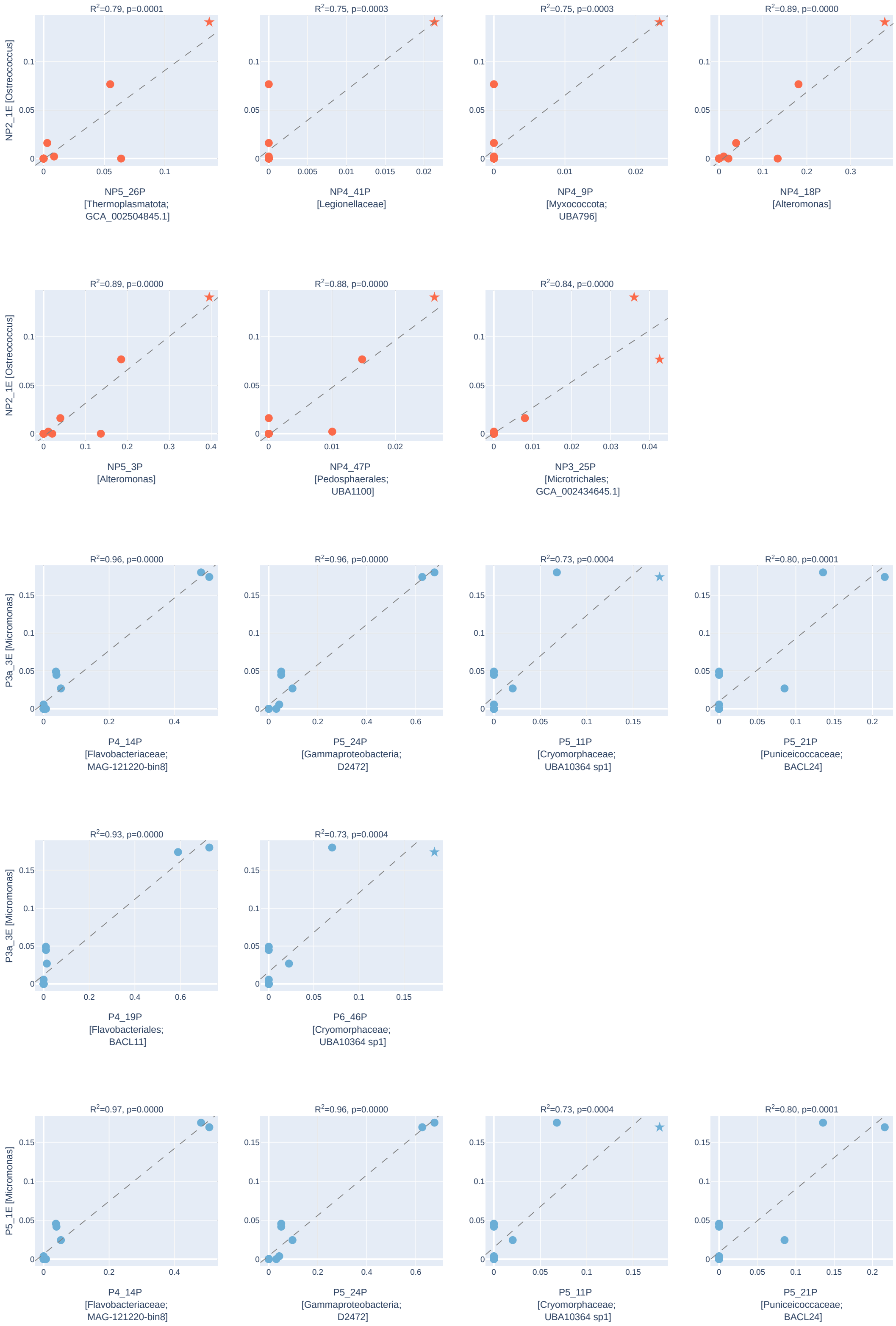


Fig. 4.29 Coverage of eukaryote/prokaryote pairs which appear associated. Some associations appear to be driven by a small number of points with high values. Influential points with Cook's distance ≥ 1.25 are indicated with stars. Split across multiple figure for legibility, see Figures 4.28 & 4.30.

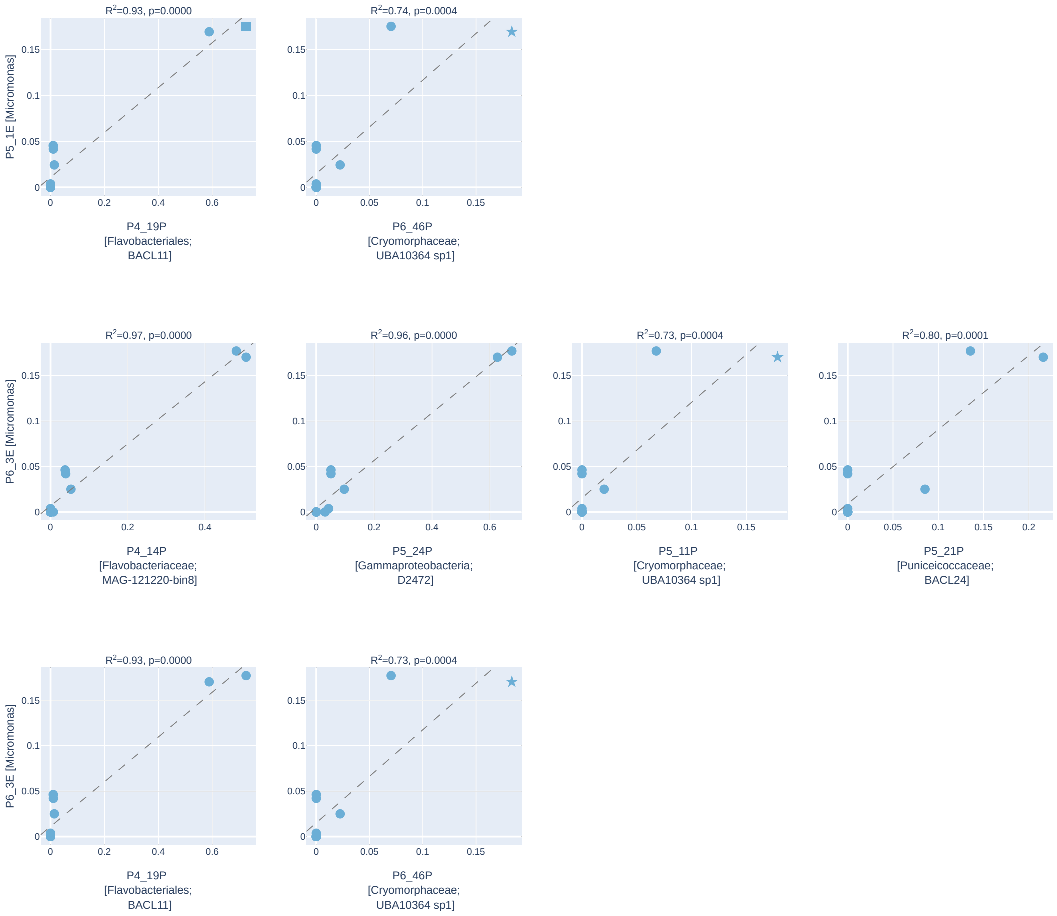


Fig. 4.30 Coverage of eukaryote/prokaryote pairs which appear associated. Some associations appear to be driven by a small number of points with high values. Influential points with Cook's distance ≥ 1.25 are indicated with stars. Split across multiple figures for legibility, see Figure 4.29 & 4.30.

This resulted in 17 inter-kingdom species associations (15-positive and 2-negative associations) between MAGs from eukaryotic phytoplankton and heterotrophic bacteria, shown in Figure 4.31. The two negative associations were between Bacillariophyta P3a_2E (which appears similar to *Minutocellus polymorphus*) and two Gammaproteobacteria MAGs P3a_1P and P3a_24P. Remaining associations were positive, with the eukaryotic MAGs members all coming from the Prasinophyte MAGs. Among the polar samples, the eukaryotes were the three highly similar Micromonas MAGs which placed closet to *M. commoda*, showing associations with a range of Flavobacteriia, Gammaproteobacteria, and Puniceicoccaceae MAGs. While these associations show high ($R^2 \leq 0.97$), they are driven by high coverage in the two samples P4 and P5 and low coverage elsewhere. In the non-polar associations, the eukaryotic MAGs were all three of the Ostreococcus and Bathycoccus. Bathycoccus NP3_1E and NP2_2E are widely distributed among the non-polar stations, and were associated with the same Erythrobacter MAG NP3_22P; Ostreococcus NP2_1E was associated with Alteromonas NP4_18P. These MAGs are widespread among the non-polar samples, the prokaryotes are observed in all non-polar samples and the eukaryotes in all but the southernmost NP5.

To further investigate the nature of the association between MAGs, we looked at which GO terms which were enriched in one pair with the strongest seeming association, NP2_2E and NP3_22P. Enriched GO terms for associated pair NP2_2E and NP3_22P are shown in Figure 4.32 as an example. The only enriched cellular components in both were the membranes: the Golgi membrane for the Bathycoccus MAG and the outer membrane for the Erythrobacter MAG. Enriched molecular functions in the Bathycoccus MAG included glycosyltransferase activity and transport of pyrimidine nucleotide sugar. The Erythrobacter MAG was characterised by a more diverse number of molecular functions with several related to transmembrane transport, hydrolase, transferase and ligase activity.

Using the same method, we looked at the enrichment of MAGs which did not participate in associations as a control set. We selected two pairs of eukaryote and prokaryote MAGs: one pair of Prasinophyceae and Alphaproteobacteria (P2_1E, P3a_15P) which are more closely related to the associated MAGs shown in Figure 4.33, one pair of Bacillariophyta and Gammaproteobacteria (P3a_4E, NP3_6P) which are more distant. In the first control set, no terms were enriched in both the control pair and the associated pair; in the second more distantly related control set, only a single term of the 82 is enriched in the associated and control eukaryote MAG, where 11 of 92 shared by the prokaryotes. This suggests that the enriched terms in the associated pair is driven by the association rather than taxonomy, as similar taxa do not have the same enriched terms.

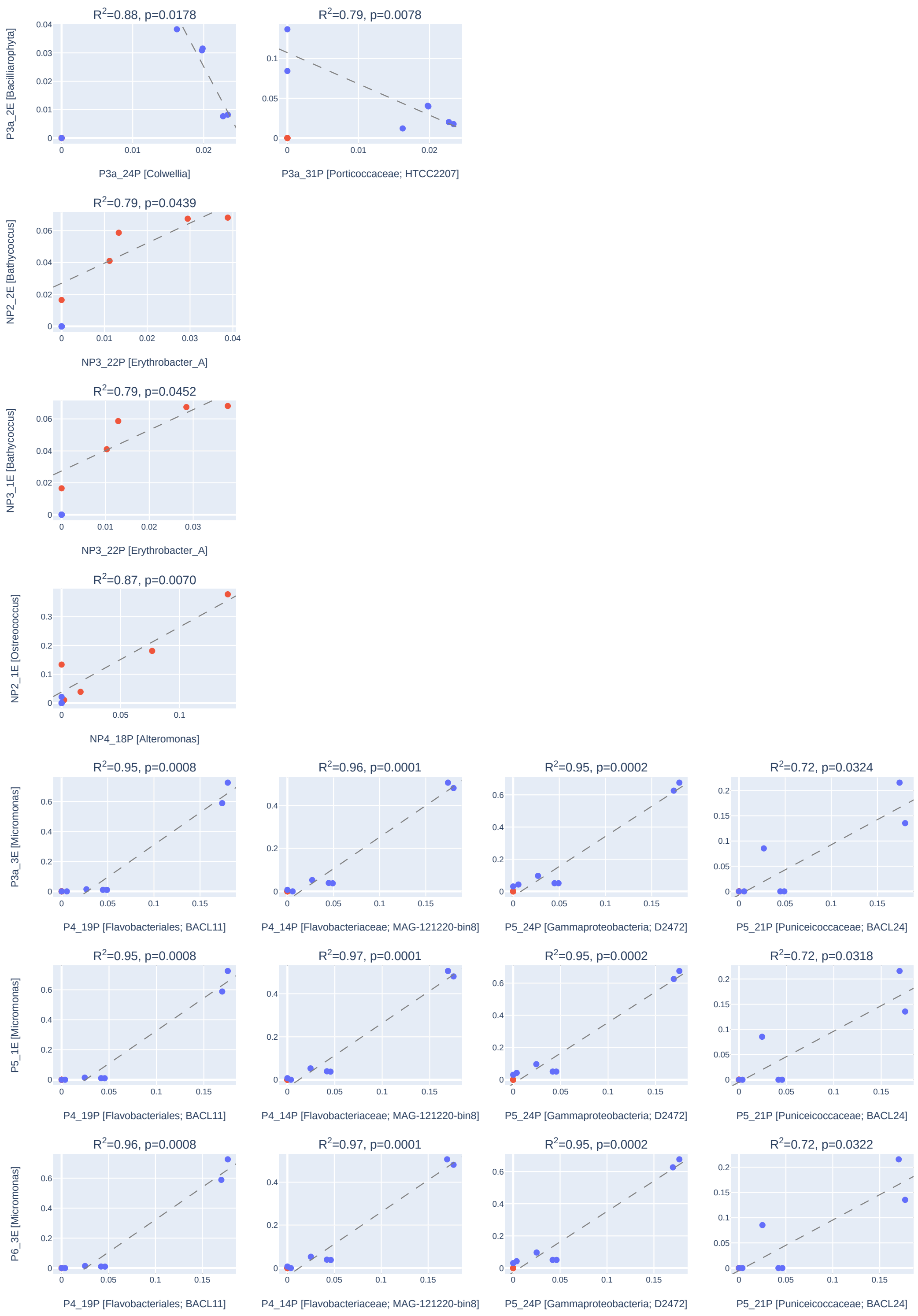


Fig. 4.31 Coverage of eukaryote/prokaryote pairs which appear associated once highly influential points are removed.

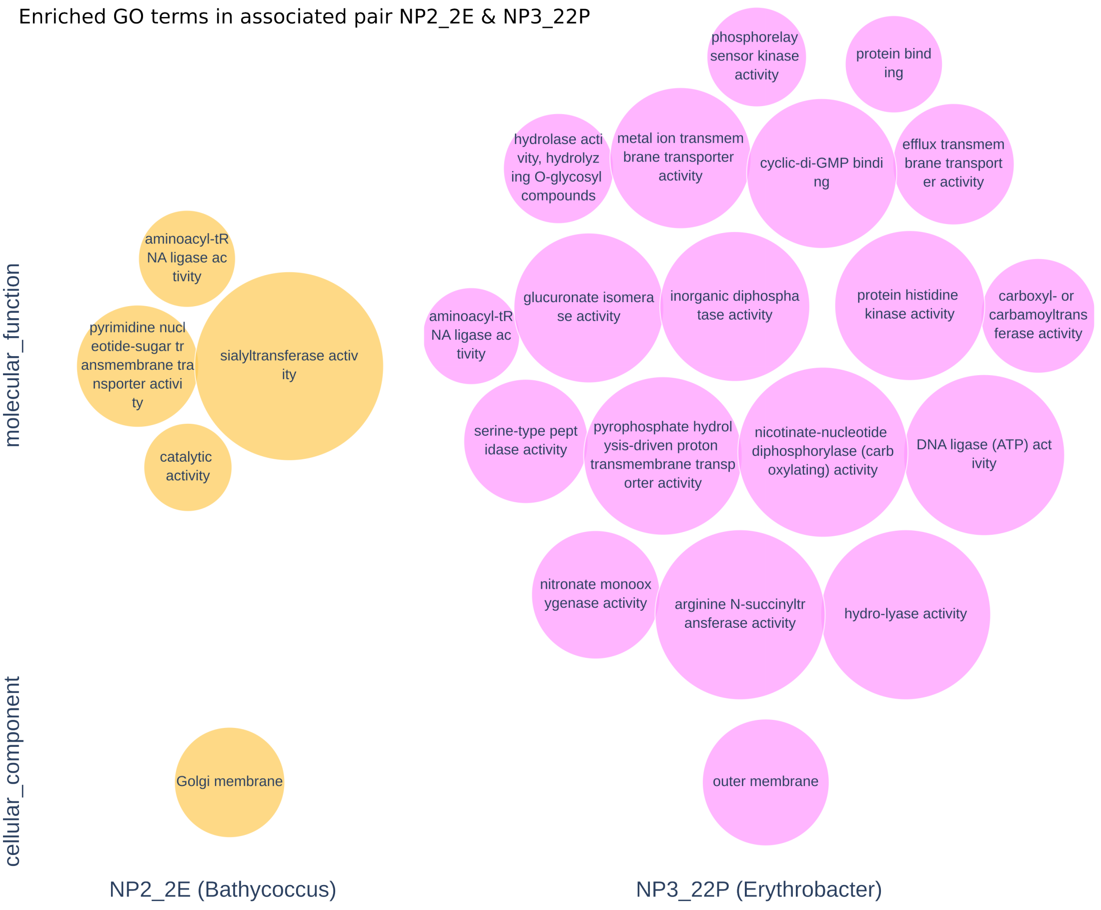


Fig. 4.32 Enriched GO terms in associated pair of MAGs. Circle is scaled to the log-odds ratio.

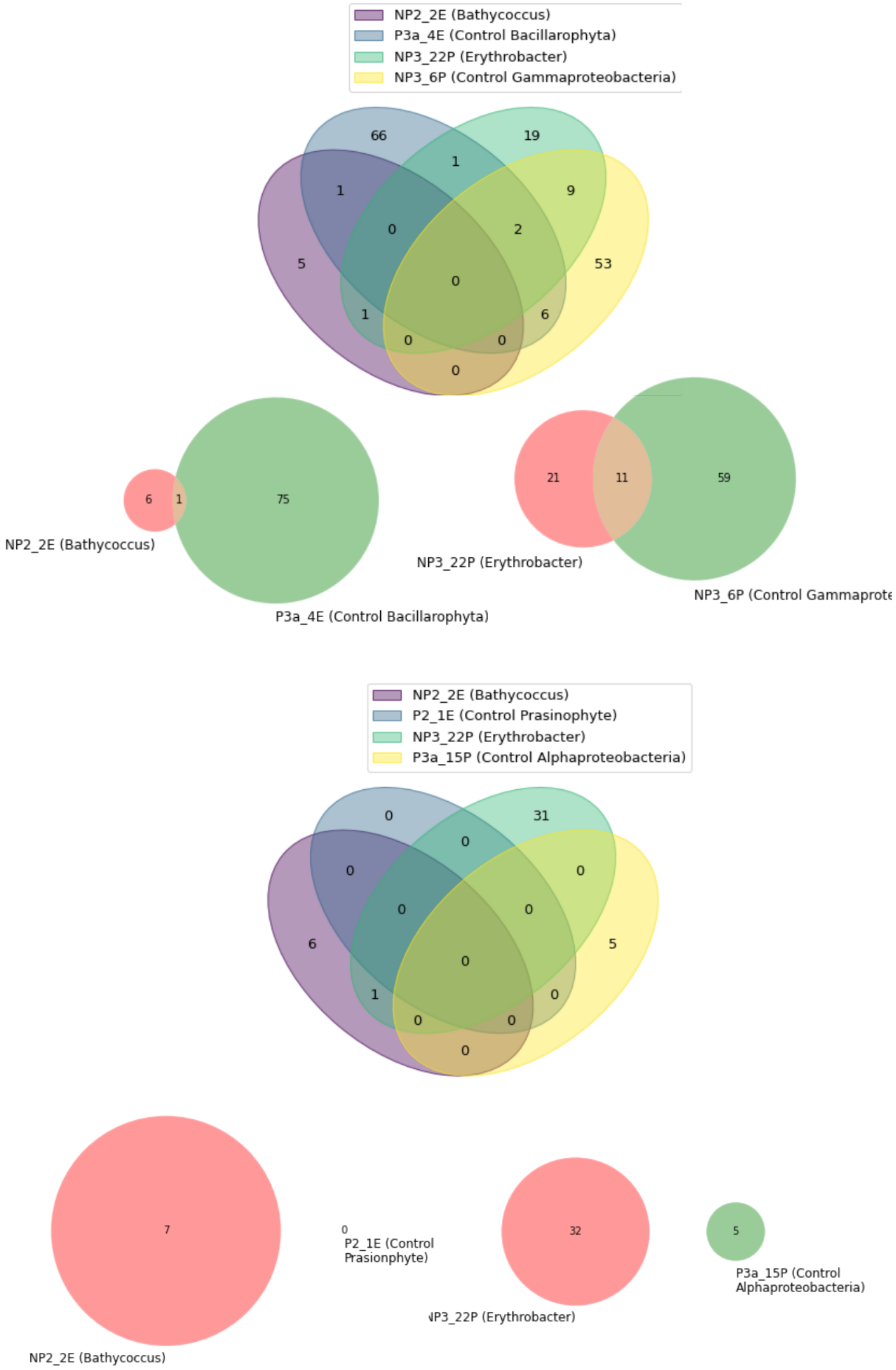


Fig. 4.33 Visualisation of terms enriched when using control MAGs, as described in Section 4.3.7. Each section of the Venn diagram relates to terms enriched in one of the MAGs. The top plots show when a distantly related control pair were selectd, the bottom plots a more closely related pair. In the distant pair, many terms were enriched in selected MAGs as expected due to being taxonomically distant from the background set. In the closer pair, no enriched terms were shared between the associated MAGs NP2_2E and NP3_22P and the controls, suggesting the enriched terms are specific to the association of the two rather than determined by taxonomy.

4.4 Discussion

Recovery of MAGs

The scale of ocean metagenomic sampling has grown rapidly, from the 44 samples and 6.25 Gbp from the Atlantic and Pacific Oceans of the Global Ocean Survey in 2007 [55], to the 243 samples and 7.2 Tbp taken across the non-polar oceans by Tara Oceans in 2015. The future promises similar growth in sampling and sequencing volume, as well as in reach through expeditions such as Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) [322] in the central Arctic Ocean (see Section 6.2.4). From a comparatively small total of 12 samples and 679.25 Gbp, we were able to recover 21 medium quality eukaryotic MAGs using an automated binning approach, coming from environmentally significant lineages such a diatoms and Prasinophytes. Comparing this to the recent recovery of 683 eukaryotic MAGs from surface Tara Oceans metagenomes [2], we recovered MAGs at a similar ratio of approximately 9 billion reads per Gbp recovered, compared to 11 billion reads per Gbp recovered in the Tara Oceans dataset. Looking at the diatoms (Figure 4.17), the diversity and volume recovered are higher in the Tara Oceans data as might be expected given the wider range and size of data used, but our MAGs placed over a significant number of clades. This demonstrates that metagenomic binning and analysis of MAGs are viable methods for smaller metagenomic studies, and can provide genomic insight into ocean eukaryotes in environmental samples.

While these methods allow genomic insight into unculturable microbes, these MAGs are an incomplete representation of the total community, recruiting 8.1% of the reads when we mapped reads back to the pooled MAGs. Recovery of MAGs was not always in correspondence with the abundance of reads from specific taxonomic groups, with some taxa which appear abundant such as Ascomycota and Apicomplexa having few MAGs recovered. This mismatch is potentially caused by a combination of factors, including sequencing depth, read length and the quality of reads most likely play a significant role in relation to genome size and complexity. The latter two factors might be the reason why we did not retrieve any MAGs from Apicomplexa such as dinoflagellates. Intraphylum diversity in combination with choice of Metabat may play a role too, as populations with low diversity and high coverage have been observed to improve the quality of MAGs recovered by Metabat [323, 324]. Green algae show low diversity, and especially members from the Prasinophytes have small genomes and are abundant in the surface ocean [325], which might explain why we retrieved several MAGs from different classes. We recovered no fungal MAGs despite their abundance in terms of reads and genes, and from the larger Tara Oceans dataset Delmont et al. [2] recovered a single fungal MAG. Notably, both our research and Delmont et al. used only samples from the surface and DCM depths. Alexander et al. [238] also used Tara Oceans data but included deeper samples, and were able to recover 16 fungal MAGs, of which 11 originated from mesopelagic co-assemblies. Abundance of fungi has been observed to increase abruptly in mesopelagic waters in comparison to surface samples [267], possibly explaining the greater recovery of fungal MAGs, even if the increased abundance is matched with raised diversity. This taxonomic mismatch extends to the prokaryotic MAGs as well. Similar to other prokaryotic binning studies [156], we failed to recover any Firmicutes MAGs, although similar studies using human gut data have done so [326].

Biogeography and Association

The distribution of MAGs between polar and non-polar environments showed a clear distinction, with very few prokaryotes crossing the Arctic circle, and no eukaryotic MAGs doing so. These two regions have major climatic differences as discussed in Section 2.4 including temperature, presence of sea ice, and water stratification. This demarcation seen in MAGs is congruent with other research, which suggests a boundary at around 15 °C mean annual temperature separating global primary production [327], and shaping microbiome taxonomic composition in metagenomic and metatranscriptomic data [288]. There is some evidence that similar pressures in the Southern Ocean result in highly similar organisms being observed there, despite the geographic distance, as evident in the high similarity (>99% ANI, Figure 4.13) between our Micromonas MAG P2_1E and *Micromonas* sp. ASP1001a, recovered from the Amundsen Sea in the Antarctic [19]. A potentially confounding factor is that our Arctic samples came from shallower samples (10-20m), while the non-Arctic samples were from deeper in the ocean (30-80m). The intention of sampling was that these depths represent the DCM at their sampling station, however the potential remains that communities at these different depths may be under different selective pressures. The depth of the DCM is known to vary with latitude, occurring at much shallower depths in high Northern latitudes, and deeper in mid latitudes [328], reflected in the sampling depths during the expedition. That photosynthetically active lineages were observed, and MAGs recovered suggest that even with depth differences, samples were both drawn from depths with photosynthetic activity at least among their functional potential. In the larger set of metatranscriptomic sequencing collected on the same expeditions, modules associated with cold polar samples (and hence lower sampling depth) and warm mid-latitude samples were identified, but GO term enrichment did not show and terms related to primary productivity enriched between the two modules [288], supporting that these may be comparable DCM samples.

This strong demarcation in taxonomic distribution is reflected in the interkingdom associations we identified, which were also exclusively between organisms originating from the same side of the Arctic Circle, and hence that mean annual temperature boundary, suggesting co-evolution under different conditions was driving the formation of the associations. The enrichment of GO terms in the associated pair we investigated showed enrichment of terms related to membrane processes, and to transport and substrate transformation, suggestive of a mutualistic relationship with exchange processes across membranes. These relationships exist between eukaryotic autotrophs and heterotrophic bacteria in phycosphere, a mucus layer surrounding phytoplankton [329].

Metabolism

The most notable distinction between metabolism in MAGs is driven by taxonomy, as shown in the visible grouping by taxonomy in the PCA plots in Figures 4.27 and 4.26, though the grouping is less clear for prokaryotes. Functional annotation of the whole population, prior to binning, suggests that the bulk of functions are shared between polar and non-polar environments, with only a small portion being unique to either region (Figure 4.26). Amongst the MAGs however, greater differences between the regional function of prokaryotes and eukaryotes, considered separately, was evident. Polar eukaryotes displayed a high number of unique Pfams, where this trend was inverted for prokaryotes with non-polar prokaryotes displaying more unique Pfams.

Among polar eukaryotes, the high abundance of transposable elements among the Pfams suggests that genomes have been forced to diversify, possible to respond to the dynamic surface ocean environment (formation of sea ice, seasonal mixing). In non-polar eukaryotes, Pfams related to phosphate acquisition and metabolism in addition to Pfams involved in iron metabolism and electron transport were among the most enriched domains in non-polar eukaryotes. The relatively low nutrient concentrations in these stratified waters might only allow eukaryotes to thrive if they have developed mechanisms for the efficient uptake of nutrients [330, 331]. Smaller-sized prokaryotes with streamlined genomes usually outcompete eukaryotes in these environments as their nutrient demand is lower [330].

The polar prokaryotes however are more challenging to describe, with the most abundant unique polar functions being typified by a high abundance of domains of unknown function. Functions unique to prokaryotes in non-polar environments have high abundance for PSD domains that are shared by chytochrome c-like proteins for electron transport as part of the respiratory chain in prokaryotes. This potentially suggests that respiratory activity is enhanced in non-polar prokaryotes compared to their polar counterparts, which would be expected according to the positive relationship between temperature and metabolic activity [332].

In contrasting the metabolism present in MAGs, a note of caution is that the taxonomy of MAGs recovered between regions shows some important differences, and some of these distinctions could be confounded by taxonomy. This is particularly the case for eukaryotes, where only a single non-polar diatom was recovered, and where the Prasinophytes are divide by genus, with Micromonas being uniquely polar.

Chapter 5

Using Non-Negative Matrix Factorisation to Identify Functional Modules

5.1 Summary

In this chapter, we move from the genome resolved approach of MAGs to look instead at ways of understanding the distribution of functions encoded in metagenomic data without resolving individual genomes. There is evidence that function is more stable in relation to environmental gradients than taxonomy [333], giving reason to think that underlying structure in functional data may be more readily computationally recovered than in taxonomic data. Taxonomic composition of communities also responds to environmental conditions, such as shifts in beta diversity seen between polar and non-polar taxa [288]. However function can be decoupled from taxonomy; functions associated with environmental conditions can be performed by differing taxa across samples driven by processes other than selection [334]. In modelling approaches genes encoded by a community appeared more stable in relation to environmental conditions than taxonomy [335, 336], and a predictive approach using all genes from ocean metagenomes, including those lacking annotation, identified 14,585 clusters of proteins strongly related to ocean environmental conditions [337], with a clear difference between polar and temperate waters. To understand microbial processes shaping the ocean, recovering patterns from the more stable functional data poses a less complex problem than the more variable taxonomy.

Differences between polar and temperate function have been identified using methods which seek a reduced dimensional description: two modules of genes, one each associated with polar and warmer waters, were identified in metatranscriptomic data collected during the same cruises as data used in Chapter 4 [288]. These cold and warm associated modules

provide insight into the molecular functions required under different temperatures, however the use of a hierarchical clustering method as part of the Weighted Gene Correlation Network Analysis (WGCNA) [338] method used limits a function to only appearing in one module. This does not reflect the underlying biology where functions or genes may be shared across organisms, metabolic pathways, or environmental niches.

Our aim in this chapter is to develop and evaluate a method to identify modules of functions in metagenomic or metatranscriptomic data which permits for the sharing of functions between the identified modules, and allows the description of the function in any sample as a mixture of the identified modules.

We do this using Non-Negative Matrix Factorisation (NMF), a matrix decomposition method with established uses in fields including computer vision and text analysis, but a more limited history of application in environmental metagenomics. Section 5.2 gives a background of approaches often used in analysis of functional data, and provides motivation for selecting decomposition and specifically NMF as an approach for identifying meaningful modules of functions and their distribution across the oceans.

The rest of the chapter is then split broadly into two parts. First, evidence is provided that NMF, and associated interpretative and visualisation methods, are capable of identifying known groups using synthetic and simulated data. Having established the efficacy of the selected methods on simulated data, we then apply these methods to illustrative real world datasets, to both show congruence with previous analyses (i.e. identifying well established groupings of samples) and to show interpretative benefits of the modules recovered.

Methods used for conducting NMF are covered in Sections 5.3.1 and 5.3.2. Interpreting the resulting matrix decomposition, to identify which features best describe a module and enriched gene sets, are then covered in Sections 5.3.3 and 5.3.4, along with visualisation methods in Section 5.3.5. In establishing that our selected methods can identify groups of functions, we generated synthetic data and also performed *in silico* simulation of metagenomic sequencing data from two communities described in Section 5.3.6.

The results in Section 5.4 provides evidence for the first part of the problem. Using the synthetic and simulated data, we establish that NMF and our associated methods can identify an appropriate rank, a key parameter for decomposition, in Section 5.4.1. From these decompositions, we assess how well recovered groups of functions resemble the true underlying groups in Section 5.4.2.

In Section 5.5 we apply the methods we developed to a range of real world data as case studies. Human associated microbiomes from different points on the body from the HMP are known to be functionally distant [27], providing a simple example dataset where these communities should be straightforward to separate. Moving to environmental microbiomes, we analyse data taken from a river estuary [1] in Section 5.5.1. Prior analysis of this data had included use of the hierarchical correlation based WGCNA method, allowing us to compare this and the non-hierarchical NMF results. Scaling up we analyse the well studied Tara Oceans data from the surface oceans in Section 5.5.3, illustrating application to larger environmental datasets. We conclude with a discussion of our results in Section 5.6.

5.2 Background

Metagenomic sequencing, followed by taxonomic and functional annotation, results in a ‘parts list’ [339] describing the organisms present and the metabolic potential encoded by their genomes; metatranscriptomic sequencing describing activity rather than potential. The total number of features, whether looking for taxa or functions, in meta-omic annotations is often very large, and growing larger with the continued revelation of previously unknown microbial diversity and function. Cataloguing genes from Tara Oceans, the Ocean Microbial Reference Gene Catalog (OM-RGC) contains over 47 million non-redundant genes [4]; annotation of the twelve samples discussed in Chapter 4 resulted in 10,957 Pfam domains being observed. These high dimensional parts lists are difficult to interpret directly, making methods to extract biological insight from these data desirable. For human microbiomes many interesting problems take the form of a supervised learning task, for instance we may wish to establish whether gut microbiome distinguishes people with inflammatory bowel disease from those without [340]. In the ocean, and particularly less well understood areas like Arctic ice communities, such binary phenotypic labels are less clear. In these environments the task is one of unsupervised learning, where we wish to identify some latent structure within the data.

While this latent structure itself is unknown, we can make assumptions about its properties based on biological knowledge. Sequences in a metagenome will originate from a set of organisms, each of whose genome encodes a set of genes. The genes (and so function) present, and their proportion, will thus be driven by which organisms (and so genomes) are present in a sample, and the abundance of these organisms. Each metagenome can be understood as a mixture of genomes, and each of these genomes can be understood as a mixture of genes. Any individual gene or function could be present in multiple genomes, with some widely shared or near universal (i.e. the Benchmarking Universal Single-Copy Orthologs (BUSCO) genes). Shifting the frame slightly and putting aside taxonomy, metabolic pathways and their genes have the same properties. Each pathway in an ontology such as Kyoto Encyclopedia of Genes and Genomes (KEGG) contain many genes (Section 3.7); the genes present in a metagenomic sample can be considered as a mixture of these pathways; each gene could be present in more than one pathway. Pulling back further, the genes in a metagenomic sample can be phrased as a mixture of underlying groups of genes, and the genes of each module describing a broad set of functions prevalent under certain ocean conditions. In common with WGCNA we refer to these computationally generated groups of functions as modules; a module is a group of functions whose abundance behaves similarly across a number of the samples studied. These modules are not intended to recover groups equivalent to pre-existing functional ontologies such as KEGG or GO, but instead the aim is that a module is a group of functions which describe activity prevalent in an environmental niche. For instance, given the differing conditions driven by vertical stratification for microbes, we could expect different modules of functions at different depths, with the mixing of these modules forming a gradient from surface to benthos.

There is reason to think this structure will be more easily identified for functional rather than taxonomic annotations, as function has been shown to be more stable across environmental gradients than taxonomy [333, 337, 335, 336]. Possibly this is due to functional redundancy; two organisms may be taxonomically distant, but both perform the same required function, leading to factors other than selection driving the local taxonomic composition. The goal for the work presented in this chapter is to implement and demonstrate that NMF is capable of identifying such modules of functions, which match this intuitive description of the underlying biology where genes or functions may be present in multiple modules. This requires demonstrating which of the proposed methods to identify the appropriate dimension for the decomposition performs well for overlapping modules (Section 5.4.1), and that from a decomposition with correct dimensions we can identify which functions belong in a module (Section 5.4.2).

First we briefly introduce several methods of analysis which have been applied to metaomics data, in order to better illustrate the motivation for our selection of NMF as the analysis method best suited given our assumptions about the underlying structure of the data.

5.2.1 Distance and Dissimilarity

Each metagenomic sample can be expressed as a vector V of length n, where n is the number of features observed, and V_i is the value for the i^{th} feature. Given m samples, it can be of interest to know how similar or dissimilar any pair of samples p and q are. When features are taxonomic units, this is analogous the idea of β-diversity in ecology (the difference between the taxonomic composition of samples). Various methods of measuring the distance between sample vectors have been employed, from the Euclidean distance, to more ecology specific measures among them the Bray-Curtis dissimilarity, UniFrac distance, and Simpson index [97]. Pairwise dissimilarities between samples are amenable to further analyses, such as hierarchical clustering or Multidimensional Scaling (MDS). They are also amenable to a variety of statistical tests, such as Mantel and partial Mantel tests, which tests the correlation between two matrices [341, 342]; correlation can be examined between matrices of functional or taxonomic distances, and potentially explanatory variables such as geographic distance. These measures of dissimilarity have been widely and productively used, however a downside is that the difference between samples is compressed to a single value, making interpreting the contribution of individual features challenging.

5.2.2 Ordination

Ordination methods seek a reduced dimensional representation of a high dimensional dataset, describing data with n features by positions along k axes, with $k \ll n$, such that similar samples are close together on the axes, and dissimilar ones separated. This is often used as a visualisation technique, allowing the data to be plotted in a readable manner on two or three dimensions, with an intuitive interpretation of close points being similar, and distant points dissimilar. Some ordination methods also serve as dimensionality reduction techniques, but the interpretability of the resulting axes varies. Details of these methods and their application in ecological contexts is reviewed in Paliy et al. [343].

Principal Components Analysis (PCA) is a statistical method which computes a number of principal components which best explain the variance in the data, with each principal component being a linear combination of the original n features, and such that the first principal component explains the largest amount of variance possible, the second the most variance not explained by that, and so on. In practice for ordinations, especially for visualisation, often only the first few principal components are used provided they explain a sufficient amount of the variance within the data. These resulting principal components can lack an intuitive interpretation in the biological context, requiring post-processing to identify features which contribute to identified groupings [344].

Canonical Correspondence Analsis (CCA) is a conceptually similar method which extends Correspondence Analysis to allow the incorporation of predictive variables via multiple regression, to analyse predictor variables for the identified axes [345]. Input for this method consists of two separate matrices for m samples, one describing the abundance of n taxa or functions in each sample, another describing values of z environmental measurements across the same m samples. Results are commonly displayed as triplots, with quantitative explanatory variables indicated as arrows over a two dimensional plot, with the perpendicular position of data points along that arrow showing it's assoication with that explanatory variable [346]. CCA is a long established tool in community ecology, predating the introduction of metagenomics methods [345]. This allows identification of similar samples, and significantly to directly identify their relationship to environmental gradients, but the contribution of features to these structure and relationships is less directly identified.

MDS is a group of methods that do not operate directly on the sample vectors, instead seeking to embed a matrix of pairwise distance measures into a k dimensional space [347]. Classical MDS is equivalent to Principal Coordinates Analysis (PCoA). This embedding seeks to minimise a loss function, often called strain or stress, and solutions found using optimisation techniques. Metric MDS assumes that the distances provided are a metric, however this is not the case for many ecological measures of distances such as Bray-Curtis dissimilarity. Non-metric MDS methods have been developed and will handle dissimilarity values which do not satisfy the criteria of a metric. Where the axes of PCA and CCA lack intuitive meaning in relation to the biological context but do relate to the original m features, MDS axes lack meaning in relation to the m features, seeking to preserve instead the distances between points.

t-distributed Stochastic Neighbor Embedding (t-SNE) is a method which can separate nonlinear data, where methods such as PCA are linear and would perform poorly in these cases [348]. The method is probability based, first constructing a probability distribution on pairs of the input data where similar objects are assigned a high probability, then defining a similar probability distribution for points in the low dimensional space, minimising the Kullback-Leibler divergence between the two distributions (Kullback-Leibler divergence is a measure of the difference of two probability distributions). t-SNE embeddings appear to be sensitive to hyperparameter selection however, and the development of guidelines for how to select appropriate values is an area of active research [349].

5.2.3 Network Analysis

Relationships between functions can be expressed as a graph, with each feature represented by a vertex, and an edge placed between related features. These networks are then amenable to a variety of topological and statistical analyses, such as identifying hub genes which have a high degree, and could be considered highly important genes among those studied. Fundamental to this approach is the method by which the network is constructed, meaning how we decide which vertices to place an edge between, along with whether and how to assign weight and direction to these edges. Jiang et al. [350] reviewed the construction of networks from omics data, assessing their applicability to microbiome data. A commonly used and computationally simple approach is to use a measure of correlation to decide between which vertices to place edges. A coefficient or significance threshold can be selected beyond which and edge with exist in the graph, and the coefficient sometimes used as a weight for this edge [351]. An alternative approach is the use of regression based models, which have the advantage of being able to account for covariates. Gaussian graphical modelling approaches have been adapted to fit the typical high dimensionality of environmental genetics data, where the number of genes is typically much higher than the number of samples available, and applied to expression and taxonomic data [352, 353]. These graphical models utilise partial correlation, the correlation of genes *a* and *b* conditioned on all remaining genes, to identify direct interactions which are not dependent on a separate variable.

5.2.4 Clustering

Grouping features which show similar distribution across samples allows description of clusters of functions or organisms which commonly exist together, and which can describe a subset of the samples. Weighted Gene Correlation Network Analysis (WGCNA) is an established technique which uses a mixture of methods to address this task, using correlation between feature profiles to define edge weights between features in a network, and using a matrix of dissimilarities based on topological overlap in this network as input for hierarchical clustering [354]. This produces a dendrogram which can be cut at a predetermined or algorithmically selected height to generate modules of genes (as defined in Section 5.1). To briefly restate, a module is a group of functions whose abundance behaves similarly across samples studied, and potentially describes microbial activity prevalent in an environmental niche. These modules are characterised by an 'eigengene', describing the weighted average expression profile of the genes of that module across the samples, allowing the description of samples in terms of how well the gene abundances for each sample correlate to each module's eigengene. The method was originally developed and applied to gene expression patterns in microarray data [355], but has been applied across environments and types of data, for instance to marine OTU data [356] and gut metatranscriptome data [357]. Recently the WGCNA method was used to identify two modules of genes in ocean metatranscriptomes spanning from pole to pole, one strongly associated with cold, another with warm conditions, showing a clear demarcation in function between polar and non-polar waters [288].

This method has a lot of properties that we are looking for, providing a description of both modules of related genes, and the relationship between those modules and samples. However given the high prevalence of gene sharing expected across modules, the use of correlation and hierarchical clustering impose some limitations. Correlation of genes abundances across all samples is used to identify related features; where features are related only in a subset of samples the correlation will be much lower.

Consider for example gene G_a involved metabolic pathway *a*, and gene G_{ab} which is involved in both metabolic pathways *a* and *b*. If among our samples we have some in which either *a* is expressed or *b* is expressed, or neither, the profiles of G_a and G_{ab} will appear poorly correlated. The assumption of global correlation for related genes does not always align with assumptions about the underlying biology being modelled, where we might expect a given gene to be present in multiple modules. Using hierarchical clustering to identify gene modules also poorly fits situations where a high degree of feature sharing can be assumed, performing a discrete assignment of each feature to a single cluster. This is addressed somewhat by using the correlation between module eigengenes and the profile of each gene in the input data as a “fuzzy membership” measure, although again shared genes are likely to show lower correlations (Section 5.3.3).

WGCNA is one among many tools in the field of gene expression which seek to identify modules, but it’s clustering approach is similar to others in being unable to identify overlapping clusters. A recent review found that the among the gene expression module identification tools tested, those which are capable of detecting overlap, such as FLAME [358], performed well compared to other clustering strategies [359]. However, they noted that decomposition approaches outperformed all clustering and biclustering methods.

5.2.5 Decomposition

Decomposition methods provide a non-hierarchical approach to identifying underlying modules permitting a more mixed description when applied to metagenomic data. These methods seek to represent a matrix X as a product of a number of smaller matrices, typically two, which we call W and H, such that $WH \approx X$. The rank k of a decomposition is the number of columns and rows allowed in W and H respectively, and for our work each describes a module. Decomposition results in a description of how much each module contributes to a sample (W) and how much each gene contributes to a module (H). While we use the NMF decomposition method, we now briefly survey other decomposition methods which have been successfully applied to biological data for context before introducing NMF in more detail in Section 5.3.1.

Latent Process Decomposition (LPD) is a Bayesian model approach, which seeks to describe each sample as a mixture of multiple underlying processes, initially applied to cDNA microarray data from cancerous cells and yeast [360]. This approach is similar to Latent Dirichlet Allocation (LDA) which has been used for topic modelling in natural language processing, which shares analogous assumptions that a given document is a mixture of topics, and topics a mixture of words, but permitting the use of continuous values rather than counts [361]. The resulting model is a probabilistic description of each identified process in terms of the contribution of genes, allowing for an understanding of the roles of both samples and genes through the identified processes. Applying it to metatranscriptomic data for prostate cancers identified a process strongly correlated to prostate-specific antigen failure, a factor which increases mortality risk among prostate cancer patients [362, 363].

Another matrix factorisation method, Independent Component Analysis (ICA), has shown utility in analysis of cancer omics data, with interpretability of the identified components being among the benefits of the method [364], and was shown to perform well in a review of module detection methods for gene expression data [359]. ICA seeks to maximise the statistical independence of the components, with multiple measurements of independence being used (such as Kullback-Leibler divergence and kurtosis), and algorithms taking different approaches to maximisation. Comparing ICA, NMF, and PCA in the context of gene expression data, Stein-O'Brien et al. note that while the methods are comparable, they identify different types of patterns (see also [365] where ICA and NMF are compared).

5.2.6 NMF

The NMF decomposition method is well established in computer vision and topic modelling, and has seen some application to metagenomic data. NMF seeks to decompose an input matrix X containing values of features for samples to two matrices W and H, such that $X \approx WH$ [366]. For an intuitive description, this often described as learning the parts that make up objects, aiming to generate both a description of which parts an object has, and what features contribute to that object. The initial application of this method was to facial decomposition, to identify basis facial features from images [366], providing a description of each facial image as a mixture of the basis facial images.

A key constraint in NMF is that no entry in matrices W and H may be negative, which provides significant benefits when interpreting the resulting model. For instance, in the facial image context, it would be difficult to interpret the meaning of an image having a negative weight for a certain type of nose; the non-negativity constraint precludes this, instead providing more interpretable situations such as a face having a mix of two types of nose.

Extending this to microbial communities, we can assume any sample is a mixture of underlying communities, and a non-negative model generates a description fitting that; a community cannot be negatively present. This approach was applied to microarray data [367], and later metagenomic data from ocean samples from the Global Oceans Survey (GOS) [368, 369]. This analysis included 45 samples with counts of 8,214 Pfams, finding that sample similarity based on the decomposition were strongly correlated with environmental distance. To our knowledge, these two previous studies are the largest application of NMF to functional profiles of ocean microbial metagenomes, and the technique has not been applied to new larger datasets such as Tara Oceans. In addition to being smaller than data generated by contemporary ocean expeditions, interpretation of the functions in identified modules was limited to manual inspection of the 100 Pfams which were highly correlated to the modules profile across sites.

A highly constrained implementation has been used to target discovery of specific fiber degradation processes in the human gut [370], using graphs of established metabolic processes to constrain the decomposition. This process relies on the availability of well founded knowledge about the environment and processes therein being studied, which is less well suited to the ocean environment where many regions and taxa remain poorly understood.

A supervised approach, aiming to separate labelled classes, was developed and applied to animal gut and human microbiome data [371], showing improved separation between classes. The animal gut microbiomes have clearly meaningful labels available (ruminant vs non-ruminant), as do the human gut (inflammatory bowel disease vs healthy). However in large scale ocean surveys classification is a less meaningful task, it being unclear what labels we would seek to separate based upon, and it would instead be preferable to use the unsupervised, unconstrained versions of NMF to discover latent structures.

Decomposition methods appear well suited to identifying gene modules in contexts with a high degree of overlap, which we assume the structure underlying metagenomic data to have. While ICA shows greater performance in gene expression data [372], metagenomic data does not express the same over-and-underexpression, and so the the negative coefficients lack as clear an intuitive meaning. These factors combined led us to investigate NMF as a module recovery tool for metagenomic data.

5.3 Methods

This section provides greater detail on how NMF works, our generation of synthetic and simulated data, and subsequent evaluation of rank selection methods. Following this, we describe measures we developed for assessing the importance of features to modules, and subsequently methods for assigning features to modules. Visualisation tools we developed are also detailed, as well as methods of looking at functional enrichment within modules. We implemented rank selection, feature importance, assignment, visualisations, and enrichment tools as a python module `metagenome-nmf` [373] with the aim of making them available to other researchers.

5.3.1 NMF

We now go into detail about how NMF works. X is an $m \times n$ matrix of metagenomic functional annotations, where for each of m samples we have measured how frequently each

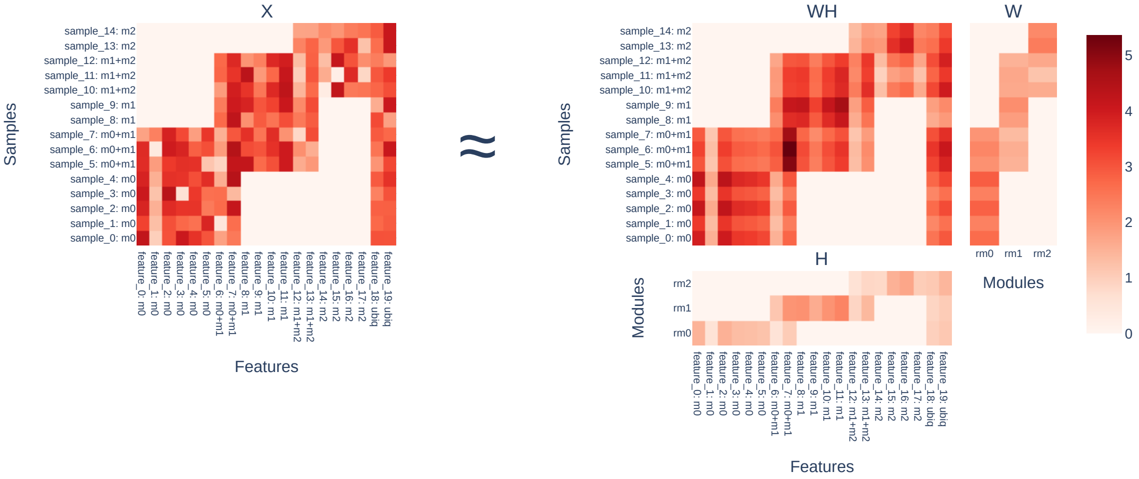


Fig. 5.1 Example of an NMF decomposition for artificial toy data with three underlying modules. The input data X is shown on the left; in this toy data a given feature can be presenting in many modules, such as feature_7 which is in modules m0 and m1. The matrix decomposition with rank 3 is shown to the right. W has a row for each sample, and a column for each module; H a column for each feature, and a row for each module. The product WH is approximates input data X. In decomposed matrix H, it can be seen that feature_7 has non-zero weights in both recovered modules rm0 and rm1. Similarly, a given sample can be a mix of modules, with sample_6 being a mixture of m0 and m1. In decomposed matrix W sample_6 has non-zero weights for recovered modules rm0 and rm1.

of n functions was observed. We want to represent each sample as a linear combination of k underlying functional modules. Additionally, we want to describe each of the k modules in terms of the functions which contribute to it. To achieve this, we can seek a matrix decomposition $X_{m\times n} \approx W_{m\times k}H_{k\times n}$, with the constraint that no entry in W or H be negative. This non-negativity constraint fits our intuitive assumptions; it has no meaning for a module of functions cannot be negatively present in a sample. A desirable property in the context of meta-omics is that the resulting description is more amenable to intuitive interpretation than the high-dimensional source data, so a model with $k \ll n$ is sought.

Relating this back to the biological case, each column in matrix W represents one of the k modules, where W_{ij} is the weight of module j in sample i. Correspondingly, each row in H describes each module as a combination of the n features, where H_{ij} is the weight of feature j in module i. Samples in W may have an above 0 weight for more than one module, rather than a discrete clustering. An example decomposition with overlapping features and samples is show in Figure 5.1. Much of the NMF literature has these matrices transposed in their descriptions, with X having features on rows, and hence W describing feature weights in modules; we adopt the transposed approach in common with the scikit-learn [374] implementation of NMF we use.

This decomposition $X \approx WH$ is typically found by iteratively updating W and H minimising an objective function. Part of the objective function typically includes a measure of the distance between X and the decomposition WH. The Frobenius norm is an extension of the Euclidean distance to the case of matrices. Kullback-Leibler divergence is a statistical measure of difference between two probability distributions, and has shown to be equivalent to the initially proposed divergence in Lee and Seung [366]. Itakura-Saito divergence has shown good performance in audio processing tasks. The two distance measures we considered are the Frobenius norm and Kullback-Leibler divergence, as the majority of literature surrounding Itakura-Saito divergence is utilising it on audio spectra rather than data similar to metagenomics. The objective function for the Frobenius norm takes the form

$$O_F(X,WH) = \tfrac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - (WH)_{ij}\right)^2 \tag{5.1}$$

The objective function for the Kullback-Leibler divergence is similar, taking the form

$$O_{KL}(X,WH) = \sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}\log\left(\frac{X_{ij}}{(WH)_{ij}}\right) - X_{ij} + (WH)_{ij}\right) \tag{5.2}$$

The W and H matrices are initialised with random values, and iterative updates continue until a local minimum has been reached (i.e. no or very small change in the objective function) or after a fixed number of iterations. Two commonly used approaches for making these updates are the multiplicative update and coordinate descent methods. Implementations of these update methods can vary, and the summary below gives the implementation of NMF we use which is provided by the python package scikit-learn [374–376]. The multiplicative update method was initially proposed by Lee and Seung [366], and is generalised by Févotte et al. [375] to the three different objective functions discussed in the previous paragraph where parameter β takes a different value: 2, 1, 0 for Euclidean, Kullback-Leibler and Itakura Saito divergence respectively. Updates are iteratively made to W and H as

$$H \leftarrow H.\frac{W^T\left[(WH)^{\cdot(\beta-2)}.X\right]}{W^T\,[WH]^{\cdot(\beta-1)}}] \tag{5.3}$$

$$W \leftarrow W . \frac{\left[(WH)^{.(\beta-2)} .X \right] H^T}{[WH]^{.(\beta-1)} H^T} \tag{5.4}$$

where . indicates an element-wise operation, and the division is also element-wise. This is continued for a fixed number of iterations, or until convergence, where convergence is defined as $\frac{e_{n-1}-e_n}{e_0} < \theta$, where e_n is the objective function at iteration n, and θ some threshold defaulting to $1e-4$. While scikit-learn does implement a coordinate descent update method, it does not include a version which is generalised to $\beta \in \{0,1,2\}$, being limited to the FAST HALS method based on a Euclidean objective function only. Given this limitation we use the multiplactive update method, as the Frobenius norm is poorly suited to sparse datasets [377], which is likely to be the case for both microbial taxa and function which have a long tail of rare taxa or functions.

5.3.2 Rank Selection

One of the key parameters for decomposition is selecting k, the rank of the decomposition. The process of generating a decomposition requires this to be specified, an appropriate value cannot be identified during execution. Some contexts may suggest appropriate values for the number of modules in the decomposition, k. For instance a study looking at samples from three kinds of leukemia suggests 3 as an appropriate value for k [367]. This is not the case for environmental metagenome datasets, where we are often starting with few assumptions about which samples are likely to be functionally similar and how many modules are likely to be latent in the data. Ocean microbe communities in particular display low functional distance across large geographic distances, in contrast to much more pronounced distances between nasal and gut samples from the same individual. Given this, criteria are needed to identify the most appropriate values of k for a given input X. We implemented a range of rank selection criteria, to evaluate which appeared most suited to simulated and real world meta-omic data.

A simple heuristic can be based on the values of the objective function across values of k. The objective functions will tend to decrease as k increases, so simply looking to optimise the objective function is inappropriate. An alternative set of approaches look at the stability of classification of samples resulting from multiple random initialisations. Some papers have sought an elbow point in this objective function as k increases [378]. Two approaches [367, 379] are based on the stability of sample classification, where each sample s is assigned to one of k groups based on the highest weight in $W_{s,\cdot}$. For each random initialisation, a consensus matrix C is constructed, where $C_{i,j} = 1$ if samples i and j are assigned to the same group, 0 if not. Matrix $\bar{C}$ is the average of these connectivity matrices across all initialisations, each entry taking a value between 0 and 1 indicating the frequency with which a pair of samples were assigned to the same group. From $\bar{C}$ two measures have been derived to evaluate factorisation stability. Brunet et al. [367] used the cophenetic correlation between distances induced from $\bar{C}$, and distances resulting from average linkage hierarchical clustering of $\bar{C}$. Kim et al. [379] uses a dispersion measure defined as

$$p = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} 4\left(\bar{C}_{ij} - 1/2\right)^2$$

Jiang et al. [369] define the concordance index, an approach which is not based on discrete classification of samples. This approach intuitively fits the environmental metagenomics context, where we are seeking to identify subcommunities of genes or taxa which mix in some proportion to create our observed community; we expect each sample to be a mixture of these underlying subcommunities rather than a clear representative of one of them. The index is based on similarity matrix S is given by $S = \bar{H}^T \bar{H}$ where $\bar{H}$ is H with each column divided by its Euclidean norm. The concordance index is given by $1 - D$, where D is mean squared difference between off-diagonal entries of S from different random initialisations. To the best of our knowledge, this method has not been evaluated in comparison to other rank selection criteria.

Muzzarrelli et al. [372] provides a review of rank selection methods. One approach evaluated is split-half validation, where X is split randomly in half to X_a, X_b and a factorisations $W_a H_a, W_b H_b$ learnt from each half, and the identified subcommunities matched up. We implement this matching using the Hungarian algorithm [380] based on Euclidean distances between modules in H_a and H_b. Each feature is assigned to the module for which it has greatest weight, and the similarity of these assignments assessed using the adjusted Rand Index [381, 382]. The mean adjusted Rand Index across multiple random initialisations is used to select rank. A conceptually similar approach starts with randomly permuting each feature individually, and learns a factorisation from the original and permuted data across values of k [383]. The slopes of the objective function are compared, and k selected as the lowest value for which the slope of the original matrix is lower than that of the permuted matrix.

Muzarelli et al. introduce the idea of imputation based rank selection [372]. This is based on variants of NMF which can assign weights to entries in X. By setting weights of some entries to 0, these are effectively held out from the learning process, and the quality of a factorisation can be evaluated by comparing the imputed values in WH to their values in X. The two metrics they detail are based on mean square error (MSE) between values of the held-out entries in X and WH over multiple random initialisations. The median and median absolute deviation of the MSE for each value of k can be used as rank selection criteria. While this criteria performed well in their review, available implementations for weighted NMF suitable for incorporation in our python package had execution times which made them impractical for application to data on the scale expected of metagenomic or metatranscriptomic sequencing, so we omitted these methods from our evaluation [384, 385].

5.3.3 Feature Interpretation

For a given decomposition $X \approx WH$, matrix H has n columns for each feature, and k rows, with H_{ki} giving the weight of feature i in underlying module k. For metagenomic data, in particular functional annotations, the number of features can be very high making manual inspection of H to identify important features for each module impractical. Simple approaches looking at the features with highest weight only capture which features are most abundant – a feature which is highly but equally abundant in all modules may be less informative than a rare feature which has low abundance in only a few modules. We have implemented and evaluated several techniques for identifying important features.

Feature Importance

Specificity [369] captures the extent to which a feature i is evenly represented across all modules, or is represented by only one module, taking a value between 0 and 1 respectively. Specificity is defined as

$$\sigma\left(H_{,i}\right) = \frac{\sqrt{k} - \sum |H_{ji}| / \sqrt{\sum H_{ji}^2}}{\sqrt{k} - 1}$$

This poorly addresses one of the fundamental properties we seek to capture, where features can be shared by multiple modules, so instead we look for alternatives which can capture this sharing.

Correlation looks at correlation between the column vector $W_{,j}$ and $X_{,i}$ [368]. This is similar to the idea of module membership in WGCNA, where correlation between the module eigengene and the profile of each gene in the input data is used to give a fuzzy idea of how much a gene belongs to a given module. We used Pearson correlation in our implementation. This method looks at correlation across all samples, which presents limitations in situations where features can be present in multiple modules. For example, if feature i is present in underlying modules a and b, the column vectors $W_{,a}$ and $W_{,b}$ will correlate poorly with a sample which contains a mix of both modules, illustrated in Figure 5.2.

We developed an alternative method of assessing feature significance in modules which compares feature weights in the selected model to those learnt from randomly permuted

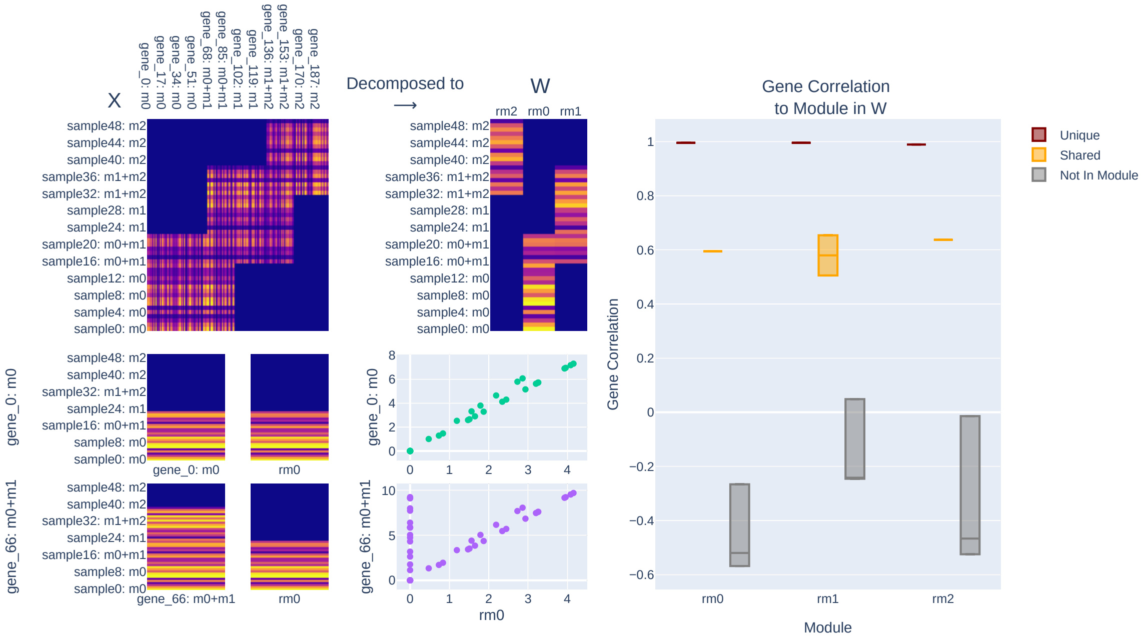


Fig. 5.2 Illustration that features which appear in multiple modules can have lower correlation between distribution in X and module profile in W. Input matrix X (top left) is toy data we generated with 50 samples and 200 features, and 3 modules. We allowed each feature to be present in either one or two modules, and each sample could contain one or two modules. Features and samples are labelled to indicate the modules they are in i.e. gene_66: m0+m1 is in modules m1 and m2. Each feature in a module is perfectly correlated (Pearson's $r = 1$) to each other feature in the module, for the subset of samples where that module is present. The sample matrix W resulting from an NMF decomposition of X is shown. Below, two example genes are selected, one which is present only in module m0 (gene_0), and one which is present in modules m0 and m1 (gene_66). The scatter plots show the weight for each sample in recovered module rm0 in W plotted against the weight for the two genes in X across samples. Gene gene_66 has a number of samples for which there is an above 0 weight in X (as it is part of module m1 as well), but for which the weight in recovered module rm0 is 0 (as it represents only underlying module m0). On the far right is a box plot showing for each module the correlation of genes which are unique, shared, or not included in a module, showing that shared genes have a much lower correlation than thos which are unique.

data in which the underlying structure has been disrupted. In outline, we learn models from permuted data, and fit a normal distribution to the module weights in H in these permuted models, and use the probability of this distribution generating the weight in the selected model. More precisely, we start with data X with dimensions $m \times n$ (m samples and n features) and define X^p as X with values for each feature randomly permuted. $H(X^p)$ is the feature weight matrix learnt from X^p, with entry $H(X^p)_{j,i}$ the weight for feature i in module j learnt from X^p. We learn $H(X^p)$ for r different random permutations, and concatenate the resulting matrices to H^r, a with dimensions $kr \times n$ where k is the rank of the model, with column $H^r_{,i}$ containing the weights for all modules for feature i. A normal distribution is fitted to each column of this matrix, $\mathscr{N}(H^n_{,i})$, and the probability of observing $H(X)_{m,i}$ taken as the measure of importance of feature i in module j, $perm(i,j)$.

We also introduce Leave-One-Out Correlation Decrease (LOOCD) as a method of identifying important features which may be shared among many modules. The basis of the approach is comparing correlation of feature values across samples in X and the complete model WH, and correlation between features in X and WH^{-j} where the column and row corresponding to module j is removed from W and H respectively. We define LOOCD for feature i in module j as

$$loocd(i,j) = r(X_i, (WH)_i^{-j}) - r(X_i, (WH)_i)$$

where $r(a,b)$ is the Pearson correlation coefficient between vectors a and b. Figure 5.3 illustrates this in situations where a feature is and is not important to a module.

We applied these three methods to synthetic data which we generated with known properties, with results shown in Figure 5.14.

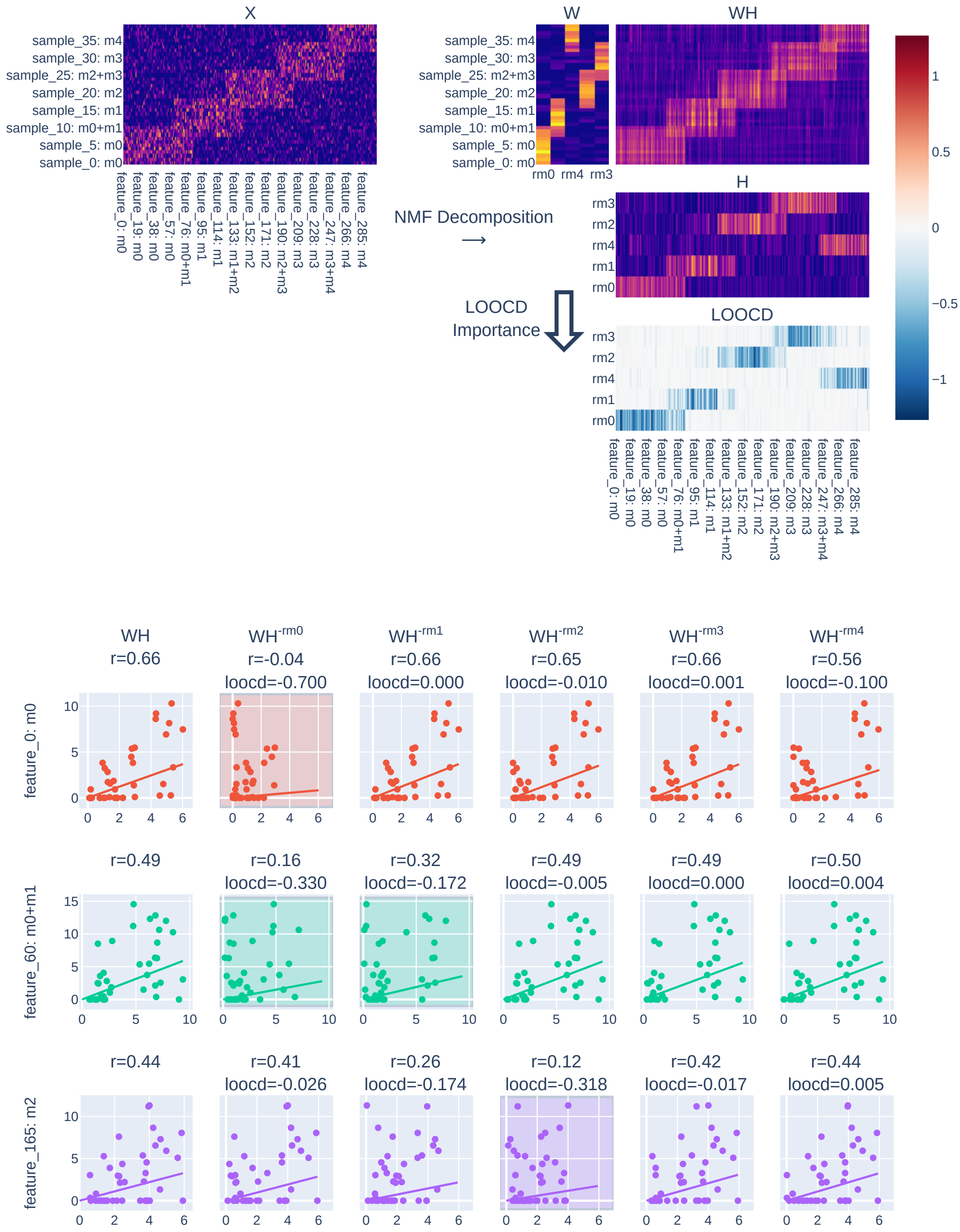


Fig. 5.3 Example of Leave-One-Out Correlation Decrease (LOOCD) feature importance method (see Section 5.3.3) for features which are unique and shared between latent modules. The top shows toy synthetic data we constructed with 5 modules where features and samples overlap (Section 5.3.6). Features and samples are labelled to indicate the modules they are in i.e. gene_66: m0+m1 is in modules m1 and m2. The resulting decomposition and LOOCD values are shown to the right. Scatter plots at the bottom show the relationship between three example features in X (one on each row) and WH with each module removed. The first column of scatter plots is relationship between the feature and full WH, other columns are each with one module removed. Lines show an ordinary least squares line of best fit. Scatter plots where the feature is part of the left out module are highlighted with background colour. For these, the correlation decrease is greater than for modules which the feature is not part of.

Feature Assignment

The underyling module structure in our synthetic data and simulated data is binary, with each feature either being part of or not part of a given module. Correlation, permutation, and LOOCD give a continuous measure of the importance of features to each module; from these we want to determine which features belong to each module, allowing for a feature to belong to multiple modules. Three methods were evaluated: a simple threshold, a greedy assignment approach, and a kernel density estimate based method.

From each of the correlation, permutation, and LOOCD, for every feature i we can generate an ordered list L^i, where each entry is a pair identifying the module j and importance value $imp(i,j)$, ordered by descending importance. From this, we implemented three methods of determining whether a feature should be assigned to a module.

The threshold approach is to set some value above or below which the feature will be assigned as part of the module. This relies on being able to identify a threshold value which will be stable or predictable across data with varying rank.

We developed a simple greedy assignment algorithm which is carried out for each feature, incrementally adding modules which improve the correlation between the model and X. This method is conceptually similar to LOOCD, being based around correlation between X and WH with some dimensions of WH witheld. The algorithm is given in Algortihm 1, and described below. At each iteration, the feature with the greatest importance value is added to the set of module features, and the correlation between the restricted model and X evaluated. If the correlation has not improved beyond a certain threshold d then the previous set is returned. The rationale is that when including a module in the model is not improving the relationship between the model and the source data, that and following modules do not contribute to the description of feature i in the model.

We developed a kernel density estimate method which is based on the observation that importance measures of correlation and permutation for a module tended to form a two peaked distribution. Features known to be in the underlying module were more frequently on one side of the central minima, and those not belonging to the module more frequently on the other. Where I^m is the importance measures of all features for module m, a kernel density estimate is produced using the gaussian_kde method of the scipy pacakge [386], which estimates bandwidth using Scott's Rule [387]. The minima is located using the argrelextrma function of the scipy package [386], and features on the side of this minima indicating greater importance assigned to module m. Figure 5.15 shows this method applied to example synthetic data we generated as explained in Section 5.3.6.

Scoring the recovered modules requires a method which handles the overlapping nature of the underlying and recovered modules. As such commonly used clustering scores such at

Algorithm 1: Greedy Module Assignment

```
Data: $X \simeq WH$, feature index $i$, threshold $d$
Result: $M$, a set of modules feature $i$ is assigned to
begin
    $L^i \longleftarrow$ ordered list of module identifier and importance for feature $i$
    $M \longleftarrow \emptyset$
    $c \longleftarrow 0$
    for $m, s \in L^i$ do
        $M' \longleftarrow M \cup \{m\}$
        /* Calculate Pearson correlation between $WH$ restricted to
           feature $i$ and modules $M'$ with values in $X$ for feature $i$
           */
        $c' \leftarrow corr(W_{,M'}H_{M',i}, X_{,i})$
        if $c' - c < d$ then
            return $M$
        $M \longleftarrow M'$
        $c \longleftarrow c'$
    return $M$
```

the Rand index, or classification scores such as precision, recall and the derived F1 score, are not suitable. Instead we use a scoring method applied to biclustering, relevance and recovery, to compare recovered and underlying modules [388]. For set of known modules M and observed modules M', where for $m \in M$, m is a set of the features belong to a single module, relevance is defined as

$$relevance = \frac{1}{|M'|} \sum_{m' \in M'} \max_{m \in M} \left(\frac{|m' \cap m|}{|m' \cup m|} \right)$$

and recovery using the same method with M and M' reversed. Descriptively, for every recovered module m', the true module m is found with which it is most similar by looking the maximum Jaccard index ($\frac{|m' \cap m|}{|m' \cup m|}$). The Jaccard index takes a value between 0 and 1, where 0 is no elements in the intersection, and 1 when $m' = m$. The sum of these similarities is divided by the number of recovered modules, giving a mean score for all recovered modules. Relevance scores how the recovered modules match up to the true modules; if there are 5 true modules, but only 2 modules are recovered with each complete (i.e. containing all expected elements), the relevance score would be 1. Recovery scores to what extent the true modules were recovered; in the example above, recovery would be below 1, depending on the intersection of the true modules.

5.3.4 Functional Enrichment

For functional metagenomic data, the ability to summarise to a higher level which processes are enriched or depleted within the modules identified is a useful step in interpreting the models provided by NMF. We did not develop new approaches for such enrichment analyses, but as part of the python module implemented the prerank version of the Gene Set Enrichment Analysis (GSEA) method [389]. We use the Pearson correlation between rows of H and X as input to the prerank method, using the GSEA implementation provided by GSEApy [390, 391]. We implemented methods for identifying GO term enrichment in data where features are Pfam domains and Interpro accessions, and for identifying KEGG pathway enrichment where the features are KEGG orthologs; the implementation can also accept any custom sets. The result is a table with each row detailing a feature which either enriched or purified in a modules, with the normalised enrichment score. GSEA corrects for multiple testing generating a false discover rate q-value, and we use default significance threshold to be 0.05. Additionally, we implemented visualisation tools to plot this table of enriched terms as a heatmap, to output scatter plots of the underlying correlations, and produce GSEA diagrams for gene sets.

5.3.5 Visualisation

Visualising the model generated by NMF is a helpful interpretative step. The first visualisation we use is a heatmap triplot simultaneously displaying W, H and either X or WH, with columns and rows reordered, with the aim of visually revealing overlapping block structures. Ordering is performed on W and H separately, then applied to the larger matrix X or WH. We hierarchically cluster W and H based on the affinity matrix using average linkage, although provide a parameter for specifying alternate linkage methods, and to use Euclidean distances instead. The leaf list of the resulting dendrogram is used to reorder the relevant matrix. The optimal leaf reordering method [392] provided by scikit-learn can provide an improved ordering but is computationally expensive on data with dimensional typical of metagenomic lists of functions, so we offer it in the python module as a parameter which is disabled by default. An affinity matrix A is generated as described in Maetschke et al. [393], however their suggested reordering based on the Fiedler vector derived from the Laplacian of A failed to result in a visually apparent recovery of the overlapping block structure in our testing. However, hierarchical clustering of A using average linkage did display recovery of the structure beyond using Euclidean distances, so we used this method. An example of the heatmap triplot visualisation is shown in Figure 5.4

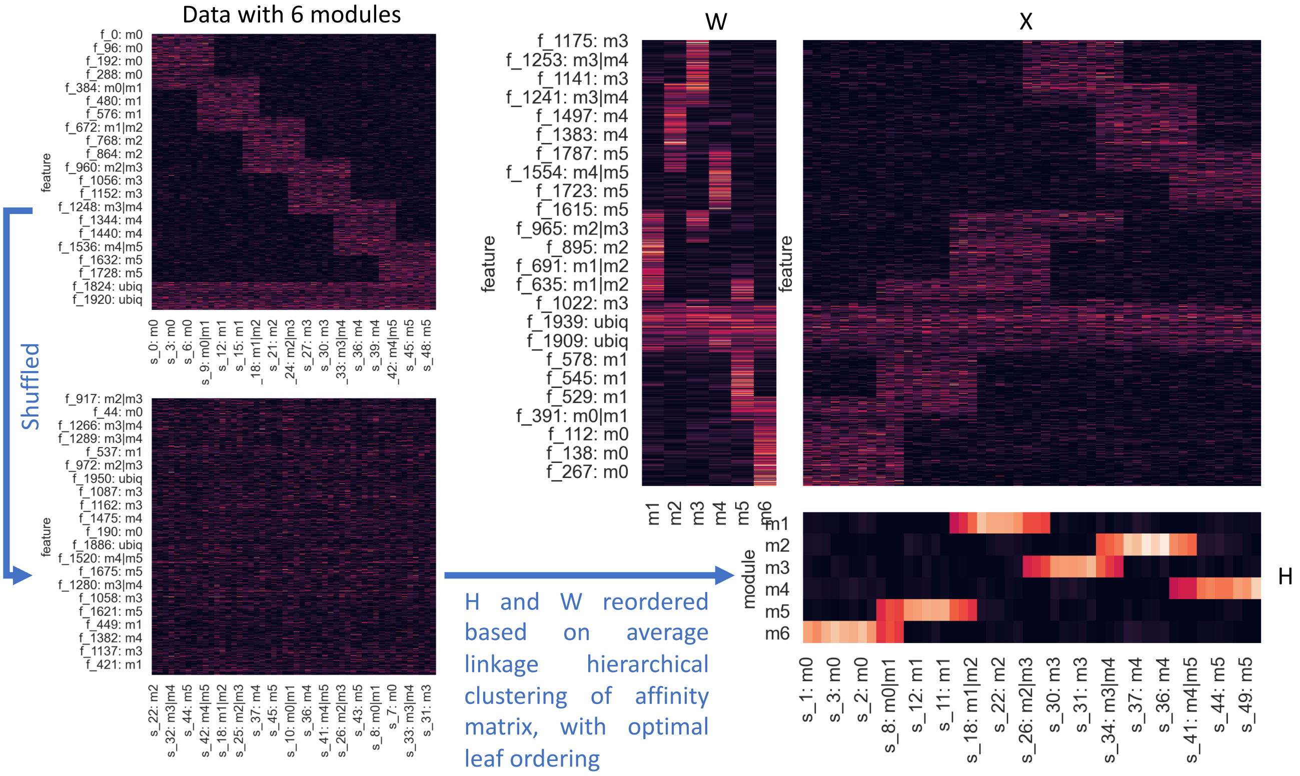


Fig. 5.4 Example of a heatmap triplot for an NMF model (Section 5.3.5). Top left shows synthetic data we constructed to have 6 overlapping modules (Section 5.3.6). When shuffled (bottom left), no structure is visible from this data. An NMF model is then learnt from the shuffled data, and the H and W matrices reordered by hierarchically clustering the affinity matrix using average linkage, and leaves of the resulting dendrogram ordered using optimal leaf ordering. Right shows the resulting reordered W, H, and X, visually recovering some of the underlying overlapping block structure.

For environmental samples, relating the model back to the sampling location can help interpret results. We use two methods of presenting the weight of modules in each sample on a map projection. First is representing each point as a pie chart on a map, with the radius of each section proportional to the weight of each module at that station. Second is an adaptation of a method mapping the module weights for each sample to a point on the RGB colourscale [394]. This provides a high-level sense of functional similarity of sites based on visual colour similarity as an at-a-glance visualisation. A model with $k = 3$ could be mapped simply to an RGB colourscale, selecting one module to correspond to each of red, blue, and green. The process used in Richter et al. [395] used the first three axes of PCA to provide values for each colour channel. In brief: data is transformed using the Box-Cox transformation to have Gaussian-like distributions to mitigate the effect of outliers and scaled to have zero mean and unit variance; PCA is performed and the first three components taken; each component is scaled to have 0 mean and unit variance; the scaled components are decorrelated using the Mahalanobis transform; each component is then mapped to 0-255 on one channel of the colourscale. Our adaptation is to scale the amount of space on each channel of the colourscale to match the amount of variance explained by each component of the PCA. If the third component explained only a small amount of variance, it can perceptually have a large influence on the resulting colour when using the full range of the channel. If $ve(pc_1)$ is the variance explained by the first component, we allow the first component to always use the full space 0-255, then permit the second and third components to use a proportion of the space $\frac{ve(pc_2)}{ve(pc_1)}$ and $\frac{ve(pc_3)}{ve(pc_1)}$ respectively, centred on the midpoint of the scale.

5.3.6 Datasets and Data Simulation

Synthetic Data

To evaluate rank selection and feature identification methods, we create synthetic data with a known underlying number of modules. Our assumption about data from environmental metagenomics is that a some features (genes, taxa) will be present in multiple modules, and that samples will contain some mixture of modules. Further, we assume that there will be some features which are present in all samples, in functional terms representing core cellular functions present in all environments (e.g. translation). Hence in our synthetic data we allow modules to overlap in both samples (modules can be present in multiple samples) and in features (a feature can be present in multiple modules), and for a proportion of features to be ubiquitously present. To reflect this, we create data with an overlapping block structure, with some proportion of features represented in all samples (Figure 5.5). Each entry which is part of a module block is filled with a uniform random value between 0 and 1. Functions and taxa

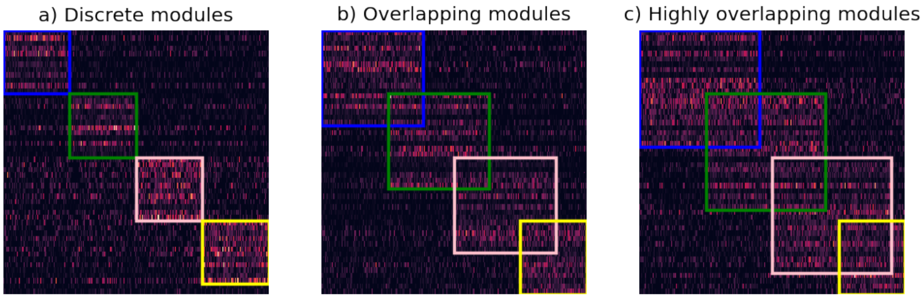


Fig. 5.5 Heatmaps of synthetic data, all of rank 4 (Section 5.3.6). The coloured outlines highlight the modules. a) shows discrete modules, with no overlap on samples or features. b) shows 40% of samples and features overlapping; for c) this value is 60%.

are not equally abundant, some may only be rare or in few copies, so we scale each feature by multiplying it by a random value between 1 and 10. Normally distributed noise is then added to each entry, and any resulting negative entries are set to 0. We have used synthetic data with different parameters, varying: standard deviation of noise applied, proportion of features in 2 modules, proportion of samples in 2 modules, proportion of ubiquitous features, number of underlying modules k.

Simulated Data

Synthetic data provides a simple test case, but it is desirable to have a test case closer to the intended use in ocean meta-omic data. Suitable test data requires that we know the ground truth of which genes should be placed together in a module, and ideally in what ratio each module was present in each sample. Real world data which has been sequenced and annotated lacks this ground truth, we do not know exactly what modules or their abundance generated Tara Oceans data for example. Instead we seek to simulate similar sequencing data for which we define the modules and ratio at which they mix in samples. The recovered modules can then be compared to ground truth used in generating the simulations. With this objective, we simulated metagenomic sequencing for two communities, each based on genomes of ocean bacteria. Firstly, a simple community composed of 5 bacteria from among the KEGG organisms, their name and KEGG abbreviation given below:

- *Alteromonas macleodii* English Channel 673, amg
- *Hydrogenovibrio crunogenus*, tcx
- *Prochlorococcus marinus* subsp. marinus CCMP1375, pma

- *Colwellia psychrerythraea*, cps
- *Trichodesmium erythraeum*, ter

Each of these is a marine bacteria, and the combination, was chosen to provide some functions which are shared by subset of the organisms, and some which are unique. *P. marinus* and *T. erythraeum* are both autotrophs sharing the capability to carry out photosynthesis, however *P. marinus* is a diazotroph responsible for a large amount of nitrogen fixation across the ocean, while *T. erythraeum* has a comparatively small genome. *H. crunogenus* is a sulfur oxidising bacterium isolated from a hydrothermal vent, and as the only sulfur oxidising bacterium contain a number of unique functions. *A. macleodii* is a species divided into surface and deep ecotypes, with the one we use being among the surface strains; it is a heterotroph, and so the functions it encodes will differ from the surface autotrophs *P. marinus* and *T. erythraeum*. Finally, *C. psychrerythraea* is a psychrophile capable of growing in low temperatures, with adaptations to enable this which would not be expected to be shared by the other temperate organisms selected. Some organisms share niches or metabolic functions, while some originate from unique conditions or perform unique functions, providing a mix of shared and unique functions in the underlying modules for this simulation. For each of these organisms, the KEGG database provides a list of the KEGG orthologs in the genome, providing a ground truth for which features should be in each identified module. Results of rank selection and module recovery methods applied to this community are shown in Figure 5.12 and Figure 5.19 respectively.

The second community simulation is intended to move closer to the focus of this thesis on Arctic microbial communities, and the data presented and analysed in 4, and to which we would hope to apply similar methods in future work. This community was based on pilot data from Arctic samples taken during the recent MOSAiC expedition [56] at four different depths: ice, sea-ice interface, surface and deep ocean (for more details on MOSAiC data see Section 6.2.4). We define our underlying modules as a set of KEGG organisms, and assign each organism an abundance in that community. Each underlying module is based on one of the Arctic samples; we identified KEGG organisms closest to the taxonomic classification of MAGs in each sample using the NCBI taxonomy and ANI where there were multiple candidates. Each KEGG organism in the module was assigned an abundance equal to the average coverage of the MAG it was similar to. The composition of each community in terms of genomes is given in Appendix B.1. Results of rank selection and module recovery methods applied to this community are shown in Figure 5.13 and Figure 5.19 respectively.

In this more complex case, the ground truth we seek to evaluate against is the set of KEGG orthologs contained in any genome in the module.

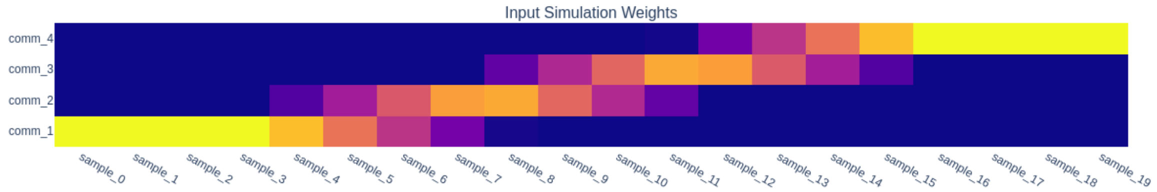


Fig. 5.6 Weights of communities in MOSAiC derived simulation. comm_1 is derived from surface sea-ice, comm_2 from ice water interface, comm_3 from surface water and comm_4 from the mesopelagic.

Simulation Methodology

For each of these communities we simulated a number of samples, where sequencing for each sample was simulated using CAMISIM [396], a metagenomic community read simulation pipeline using ART [397] for read simulation. The default profile for read simulation was used, generating Illumina 150 bp paired-end reads with a HiSeq 2500 error profile. Reads were quality controlled and merged using fastp [398]. Genes were predicted from reads using FragGeneScan [272], and annotated using KofamScan [399]. Counts of KEGG Orthologs in each sample were produced from this annotation. For this five genome community, CAMISIM randomly selected the abundance of each genome per sample, and we generated 2 Gbp of reads per sample, for 25 samples. For the MOSAiC derived community, the abundance of genomes was not determined by CAMISIM in this case, we provided abundance based on the linearly interpolated abundance of modules multiplied by the abundance of organisms per module. The relative abundance of communities is shown in Figure 5.6, and the derived relative abundance of genomes is shown in Appendix B.1. The intention of this was to provide a simple representation of communities mixing along a depth gradient. Again 2 Gbp of reads were generated per sample. A schematic of the simulation for this simulated dataset is shown in Figure 5.7.

For both cases, in evaluation NMF models were built using Kullback-Leibler (KL) divergence, and k equal to the true number of modules, 5 and 4 respectively. KEGG orthologs were assigned to a module where there was a LOOCD of ≤ -0.05, and recovered and true modules matched up the Hungarian algorithm with Jaccard distance as the cost. For each pair of recovered and true modules, we calculate precision and recall.

5.3.7 Real World Case Study Data

We used synthetic and simulated data to provide some validation of the NMF methods we have developed. We then applied NMF to a range of real world case studies, to show performance on true meta-omics data generated from environmental samples. The selected

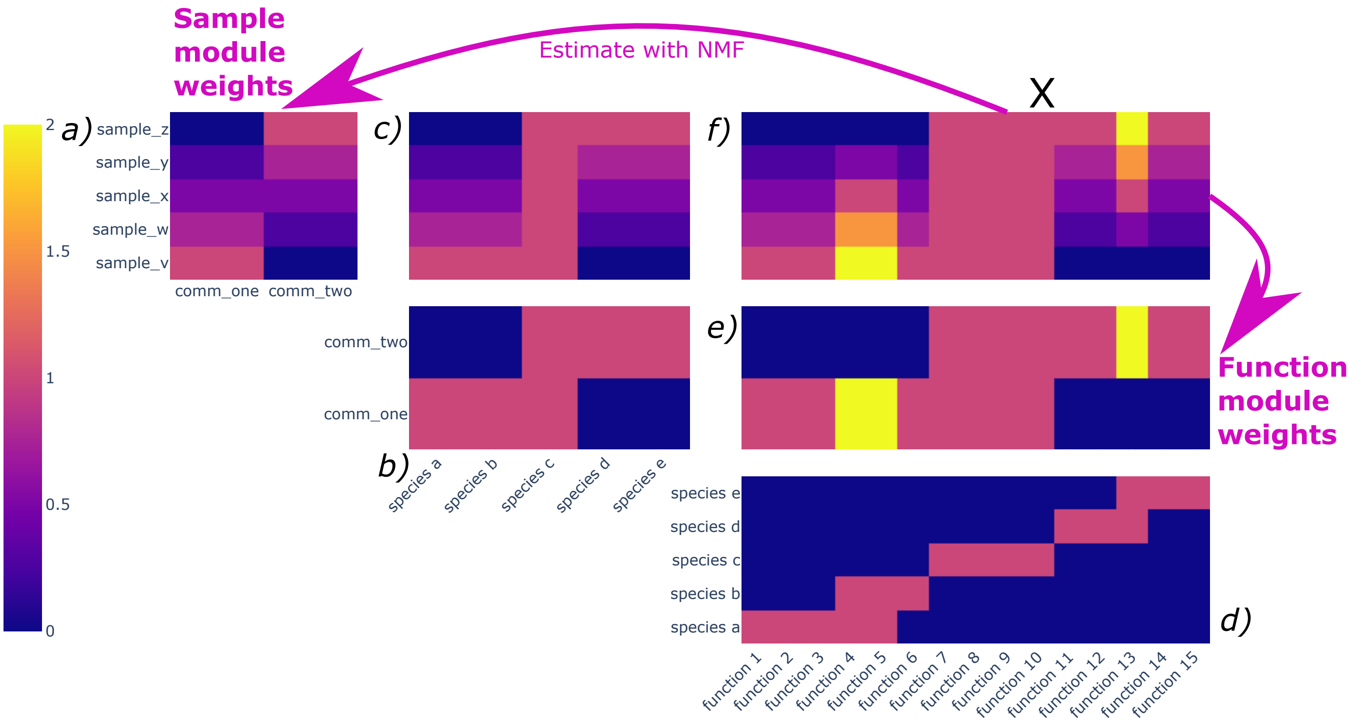


Fig. 5.7 A schematic diagram showing our construction of the simulated data based on pilot MOSAiC data [56] (Section 5.3.6). We define a), the relative abundance of each module in each sample as a linear gradient, roughly emulating a transition of communities with water depth. This is one of the matrices we hope to recover using NMF. For each underlying community, we define which genomes are present, and in what abundance, based on the MAGs generated from those MOSAiC samples. Table c) is the abundance of each genome in each sample derived from a) and b), provided to CAMISIM to determine volume of reads to be simulated from each genome. We do not seek to recover c). In d), the functions of each species are shown, with some overlapping. We do not seek a recovery at species level resolution, instead seeking to recover table e), the functions present in each module, which the pooled functions in its constituent genomes. Table f) is the counts of functions in the data simulated by CAMISIM, and is used as our input matrix X.

datasets span a range of types of meta-omic data, size of dataset, and environment of origin. Properties of each of the datasets as well as discussion of the reasons for their selection are as follows.

Human Microbiome Project (HMP)

The Human Microbiome Project has provided a wealth of data on the microbial life associated with different parts of the human body [400]. With the end goal of assessing the functions of the global ocean in mind, we chose to use the functional annotation of genes recovered in samples taken from multiple different points of the body in the HMP1 stage of the project [27, 400]. It has been well established that human associated microbial communities from different points on the body are highly functionally distant [27]. This provides us with a real world dataset, for which each sample can be labelled with a sampling site, where given the functional distance, we would expect the NMF approach to be able to identify modules related to these labels and identify a clear separation.

Waiwera River Estuary Water and Sediment

Part of the reason for exploring a matrix decomposition approach is to obtain modules which permit sharing of features compared to heirarchical clustering approaches. We selected data taken along the Waiwera river estuary on the South Island of New Zealand [1] primarily as this data had been previously analysed using WGCNA thus allowing a comparison of modules generated with NMF. Samples were taken along a salinity gradient from fresh, to brackish, and eventually marine waters, where we might expect a mixing of underlying functional modules in response to this gradient. Both sediment and water were sampled, providing in addition to an environmental gradient two very distinct environments. In comparison to the Human Microbiome Project Whole-genome Shotgun Sequencing (WGS) sequencing approach, this data is much smaller through looking at a selected of key marker genes for biologically important processes in the river environment, such as nitrogen cycling.

TARA Ocean Surface Ocean Metagenomic Data

Metagenomic sequencing of the samples collected during the Tara Oceans expeditions has expanded our understanding of the genes and functions across the global ocean [4, 2]. Taxonomic and functional annotations of much of the sequencing data has been made available publically by EBI, annotated using their MGnify pipeline [3]. As a case study, we selected the functional annotation of metagenomes filtered for prokaryotic size fraction organisms (MGnify Study MGYS00000410), as functional annotations of the eukaryotic size fraction were not publically available. This data represents a step up in scale from the river estuary data, with 248 samples and abundance of 16,349 InterPro entries for each sample.

Samples in this analysis were taken at multiple depths, providing an additional chance to validate that NMF can be used to separate environments we expect to be functionally different. The DCM, characterised by primary production, we expect to be functionally distant from those of mesopelagic samples. Following this validation, we can use NMF to analyse the distribution of function across a single ocean layer, here selecting to look at the surface ocean. To our knowledge, NMF has not been applied to this data previously, providing an opportunity to demonstrate how NMF can produce interpretible models of large scale environmental meta-omics data.

5.4 Results

5.4.1 Rank Selection

One of the key tasks in decomposition is to identify an appropriate rank k. Generally the objective function will continue to decrease as k increases; with the goal of finding a low dimensional, interpretable model, we want to find the lowest value for k which captures the latent structure in the data. The methods proposed for evaluating which k best achieves this were covered in Section 5.3.2. We evaluated how well these methods identified appropriate values of k for datasets where the true number of underlying modules is known; first on synthetic data, and then simulated metagenomic sequencing data.

Synthetic Data

Synthetic data was generated which varied along the following parameters:

- k - Number of clusters.
- o_s, o_f - The overlap between clusters in the rows (o_s) and columns (o_f). This is expressed as a proportion of the rows or columns which are in two clusters.
- u - Standard deviation of normally distributed noise applied with mean 0. Entries which are in a module are given a uniformly distributed value between 0 and 10 before this noise is applied.

Synthetic datasets were generated with the properties shown in Table 5.1, for $k = 2..12$, with 100 samples and 500 features, with 50 features being ubiquituous. For each set of properties, 100 matrices were generated, and model selection run once for each matrix, searching ranges of k between 2 and 15, using KL divergence and the multiplicative update solver. For each run, we took as the rank selected for k where the value was highest. The

label	o_s	o_f	u
discrete_lownoise	0	0	1
lo_f_lownoise	0	0.1	1
lo_sf_lownoise	0.1	0.1	1
lo_s_lownoise	0.1	0	1
mod_f_lownoise	0	0.4	1
mod_sf_lownoise	0.4	0.4	1
mod_s_lownoise	0.4	0	1
discrete_highnoise	0	0	4
lo_f_highnoise	0	0.1	4
lo_sf_highnoise	0.1	0.1	4
lo_s_highnoise	0.1	0	4
mod_f_highnoise	0	0.4	4
mod_sf_highnoise	0.4	0.4	4
mod_s_highnoise	0.4	0	4

Table 5.1 Parameters used for generating synthetic data for rank selection evaluation (Section 5.3.6). 100 matrices were generated for each set of parameters for $k = 2..12$. Labels are assigned to each set of 100 matrices to indicate noise level, and which dimensions overlap in that dataset. Results of model selection experiments on this synthetic data is are shown in Figure 5.8 and Figure 5.10.

number of times each method selected either the correct value for k in each case is shown in Figure 5.8, and where it selected a rank ± 1 in Figure 5.9.

The permutation based model selection method performed well in both low and high noise datasets for lower ranks, but with performance declining for higher ranks in the noisier datasets. Modules which are shared between samples and between features have the lowest performance, and in the most complex dataset mod_sf_highnoise the correct rank was not identified in any datasets with rank 9 or above. The permutation method tended to underestimate the rank of data, shown by the early peaks in Figure 5.10 for mod_n_highnoise. Among the two methods based on consensus matrices, dispersion and cophenetic correlation, dispersion shows generally higher performance, though both methods have uneven performance across ranks in mod_s_highnoise and mod_sf_lownoise datasets.

Examining the plot of dispersion and cophenetic correlation for one of these datasets shows (Figure 5.10) for $k < 5$ a downward trend initially with no peak, while for $k \geq 5$ there is a peak visible. This trend is most evident in the synthetic datasets where modules overlap in both samples and features, suggesting this method of rank selection would be best suited when domain specific knowledge suggests that samples would form discrete groups.

Split-half validation performs poorly in cases where modules overlap in features and samples, with the correct rank not identified for any $k > 4$ in the mod_sf_highnoise dataset.

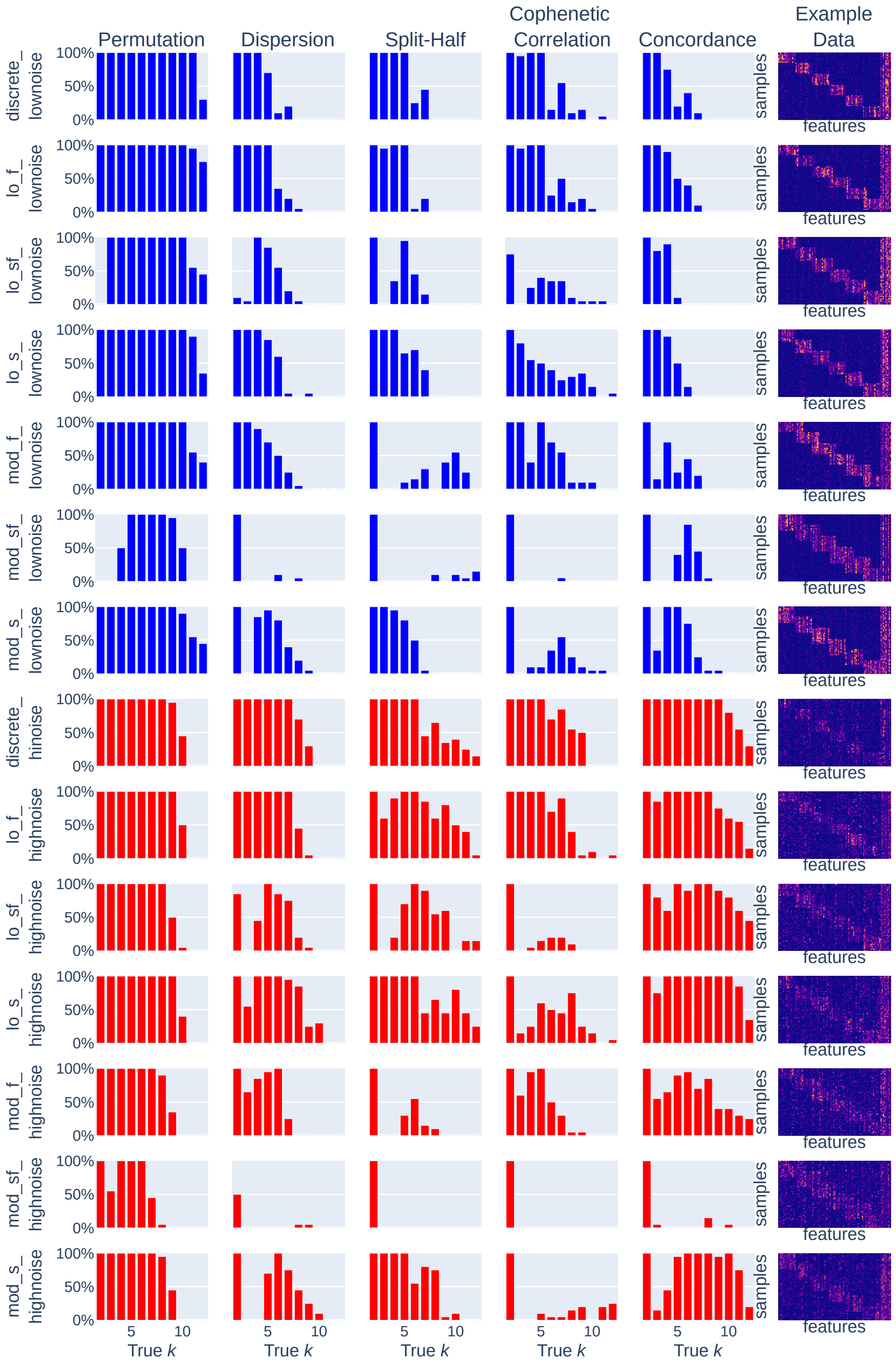


Fig. 5.8 Times a given model selection criteria peaked at the true value of k in the synthetic data we generated (Table 5.1). Vertical axes are the percentage of times correct k was selected. Horizontal axis is true value of k. Dataset labels are shown at the left of the plots, method labels at the top of the plots. The rightmost column shows an illustrative example dataset. Colour indicates low noise (blue) and high noise (red).

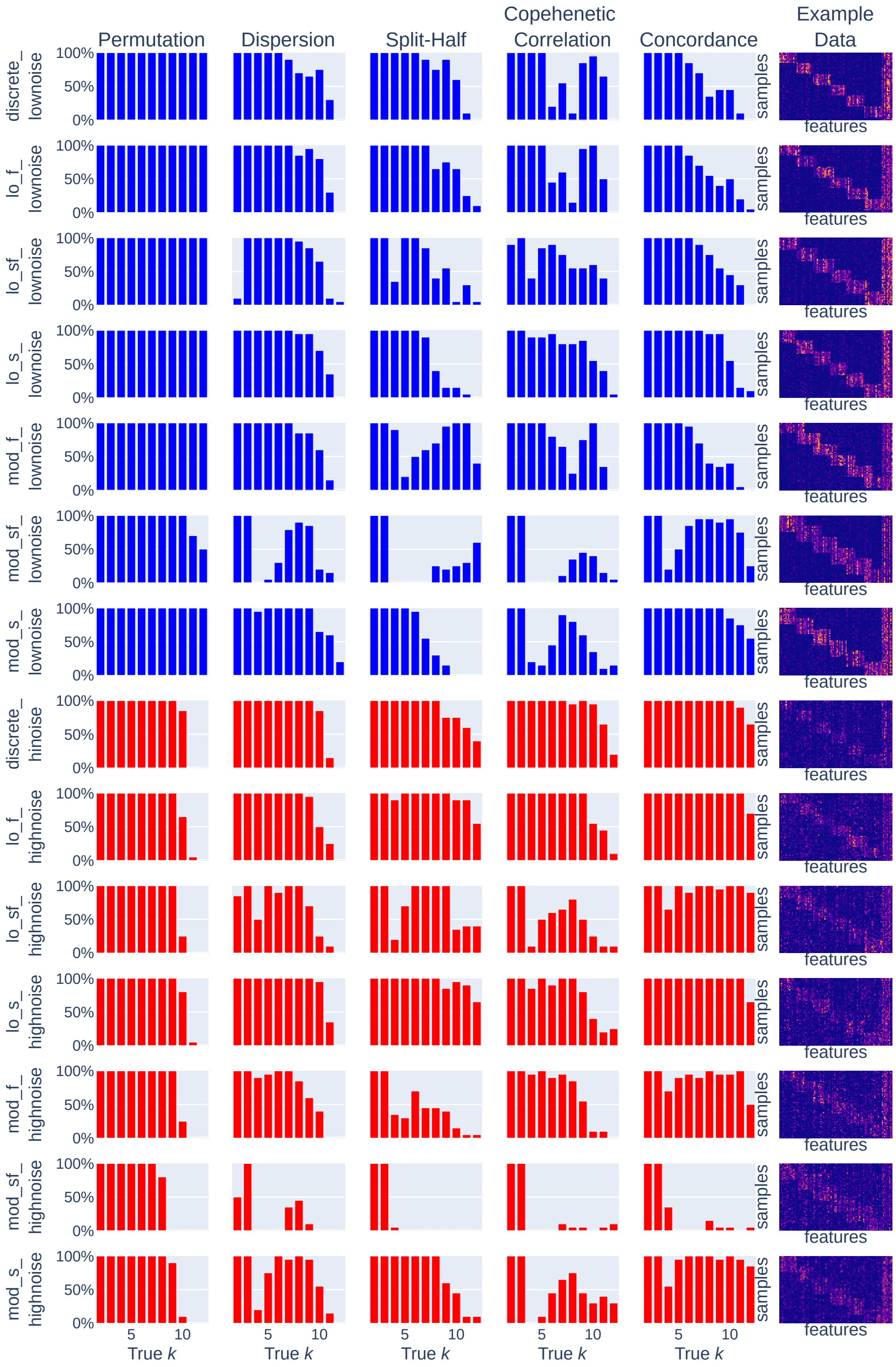


Fig. 5.9 Times a given model selection criteria peaked at ±1 of the true value of k in the synthetic data we generated (Table 5.1). Vertical axes are the percentage of times correct k was selected. Horizontal axis is true value of k. Dataset labels are shown at the left of the plots, method labels at the top of the plots. The rightmost column shows an illustrative example dataset. Colour indicates low noise (blue) and high noise (red).

Looking to the plot of values over ranks of k, in the simpler dataset discrete_lownoise the split half method displayed peaks near the true value of k, however this pattern becomes much less clear in more overlapping noisier data.

Concordance displays an improved performance in high noise datasets compared to the low noise ones, seemingly performing poorly even on the simplest discrete_lownoise case. Examining the plots of the concordance index over values of k however shows that in the discrete_lownoise case there is a clear elbow point which is easily identified by eye at the true ranks of k, but for higher rank of k the peak value occurs slightly after this elbow point. In noisier, more complex data (the example of mod_s_highnoise is shown in Figure 5.10) there is a peak at the true value of k, with values declining again after. For the concordance index, selecting an elbow point or peak value appears a better approach rather than solely looking for the maximum value.

No one rank selection method clearly outperformed the others. However for low values of k, the permutation method performed well on a wide range of the test datasets, though performing poorly as k increased. The concordance index did not always assume its maximum value at the correct rank, but consistently displayed identifiable peaks or elbow points at the correct rank where other methods show no such signal (such as in mod_s_highnoise in Figure 5.10). For data with an unknown latent rank, consensus between methods would provide strong evidence for a suitable rank; where consensus is not achieved, peak or elbow points in the concordance index appear to be the most consistent signal indicating suitable rank.

Visualisation can assist in assessing a suitable rank where there is not clear consensus among methods. Taking an example of a single synthetic dataset with 6 modules from mod_sf_hignoise, this is illustrated in Figure 5.11. No clear consensus is available between the different model selection criteria. However, both consensus based methods peak at $k = 3$, permutation peaks at $k = 6$, and the concordance index has peaks at both. Using the heatmap triplots and ordering techniques covered in Section 5.3.5, it is visually apparent that additional structure is recovered in $k = 6$ compared to $k = 3$, suggesting $k = 6$ as a more appropriate rank. Exploring the values of k near those suggested by the rank selection methods can help confirm a suitable value of k; for our example data, looking at $k = 7$ the additional module m7 has few features with high weight, and the samples with high weight for the module are scattered. Looking to one module fewer, $k = 5$, if we retain the ordering of sample from $k = 6$, it is evident that two of the modules have been combined; m1 and m4 in the $k = 6$ model have been combined in m4 in the $k = 5$ module, the block highlighted yellow in Figure 5.11. In summary, selecting an appropriate rank for the decomposition can be aided by a combination of rank selection criteria and visualisation, but requires researcher investigation and judgement.

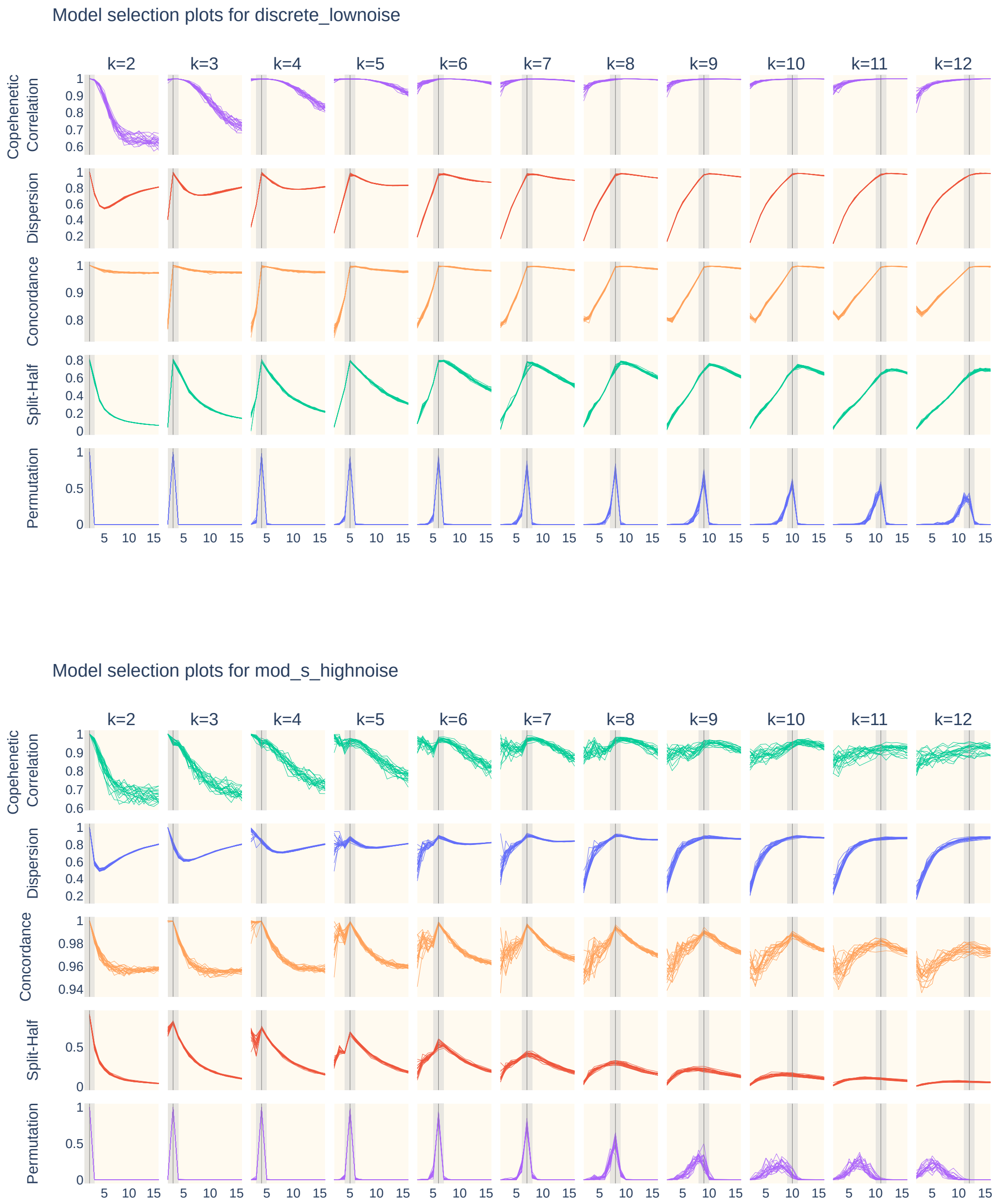


Fig. 5.10 Rank selection criteria over values of k for synthetic data we generated (Table 5.1, Section 5.3.6). Two synthetic datasets are shown, discrete_lownoise (top) and mod_s_highnoise (bottom). Each column is a different true latent rank, each row a different rank selection method. The vertical grey line indicates the true latent rank, with the grey band showing ∓ 1. Peaks or elbow points are evident near the correct rank for many methods in the simpler discrete_lownoise case, but are less clear in mod_s_highnoise for higher ranks, with only the concordance index showing indication of true rank. Similar plots for all synthetic datasets are available in Appendix B.1

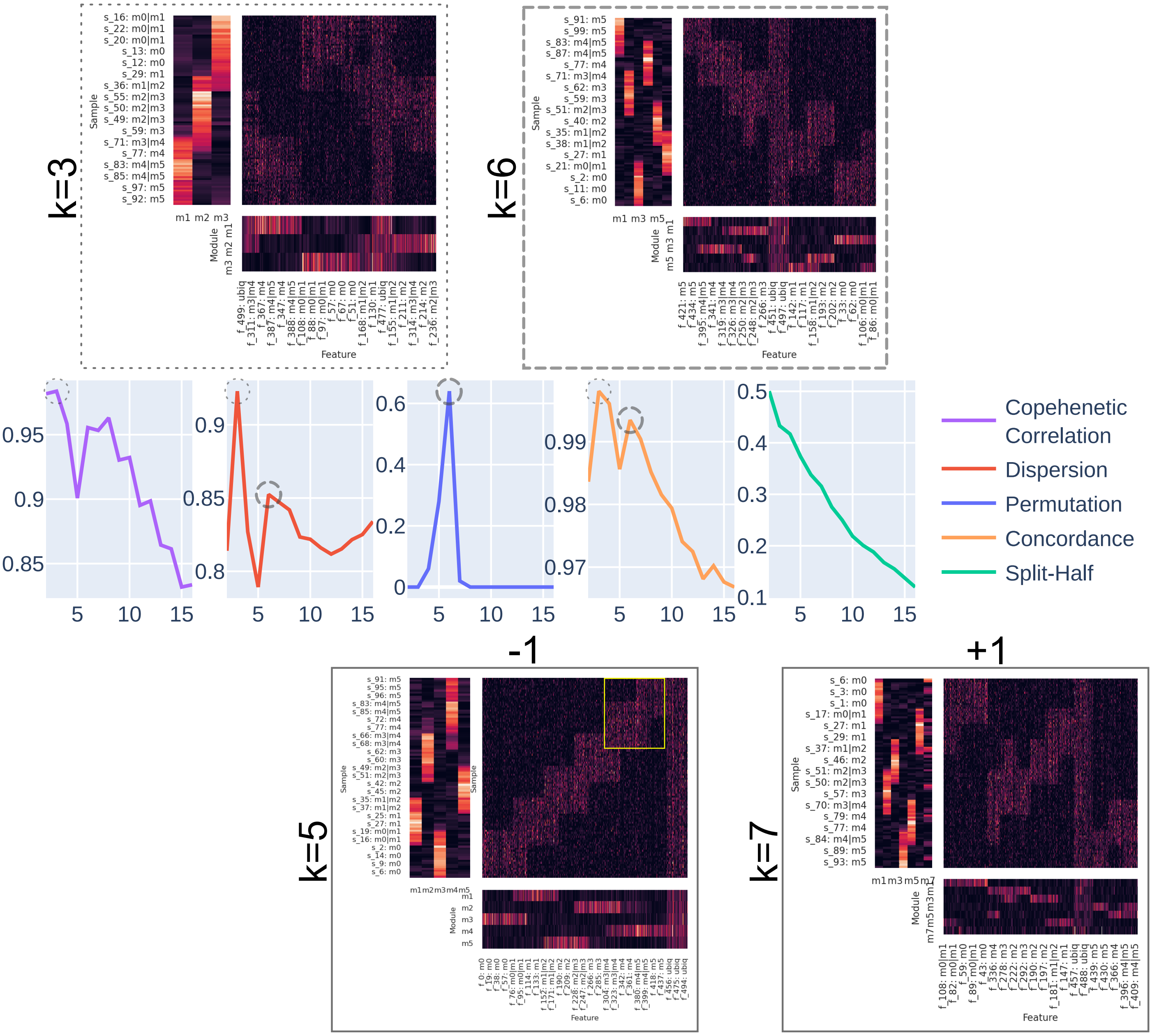


Fig. 5.11 Rank selection performed for a single matrix taken from one of the synthetic dataset we constructed to evaluate rank selection methods, mod_sf_highnoise (Table 5.1, Section 5.3.6). The central line plots show rank selection criteria across values of k. The points circled with short or long dashes indicate points at which multiple methods show peaks, at $k = 3$ and $k = 6$ (short and long dashes respectively). Above heatmaps visualise the decompositions for these ranks, illustrating the recovery of additional structure in the $k = 6$ decomposition. Below, the same type of visualisation is shown for one rank either side of $k = 6$. In the $k = 7$ decomposition, the additional module m7 has few features with high weight and scattered weight in samples. In the $k = 5$ decomposition, module m4 can be seen to represent two of the underlying modules (block outlined in yellow).

Simulated Data

The two simulated datasets detailed in Section 5.3.6 provide a test case closer to a real world omics dataset, however unlike a real world dataset the true value of k is known.

Model selection was run for the five genome simulated community, using KL beta divergence, multiplicative update, 100 iterations for each value of k, and searching $k = 2..10$; results are shown in Figure 5.12. For the smaller five genome community there is some consensus displayed: peaks appear at $k = 5$ for concordance index, permutation and split half method. The consensus matrix based methods did not agree, with a peak in cophenetic correlation at $k = 7$. Visualising $k = 7$ shows two modules which are highly similar in terms of feature and sample weights (m3, m7), and one module which is given low weight across a large number of samples (m1), suggesting it may be too large a rank.

The simulated dataset derived from the MOSAiC pilot samples is more complex, each sample being a mixture of four communities of genomes; rank selection and accompanying visualisations are shown in Figure 5.13. Again there was some consensus between model selection criteria, with peaks at $k = 4$ in concordance, cophenetic correlation and dispersion. In this case however, the permutation and split-half methods had no peak after $k = 2$. Visualising $k = 2$, it is appears that among the features on the far right of the plot, there is more variation than is captured by module m1 alone, the features presents vary from sample_8 to sample_19. Comparing this to $k = 4$, in the sample matrix W the previous two modules have each been split into two, and the features show less immediately obvious undescribed variation though with the higher number of features this is more difficult to assess. Going a rank higher to $k = 5$, again it appears m1 has been split into two modules, m1 and m2. The features with higher weight in new module m2 appear quite widespread in X, though whether this additional module is redundant is less clear than in the smaller five genome community.

In both simulated metagenomic datasets, the concordance index showed a peak at the true latent rank. Alongside its good performance in synthetic data, this suggests the concordance index as the more consistent indicator of suitable rank for data expected to have a high degree of sharing, both features being shared among modules, and modules shared among samples. While there was agreement between the concordance index and other criteria in both simulated data, which criteria showed agreement was different between the two. As the concordance index performed well in both synthetic and simulated data, we adopt the approach of looking for agreement between the concordance index and another method as a strong sign of the suitable rank for NMF decomposition.

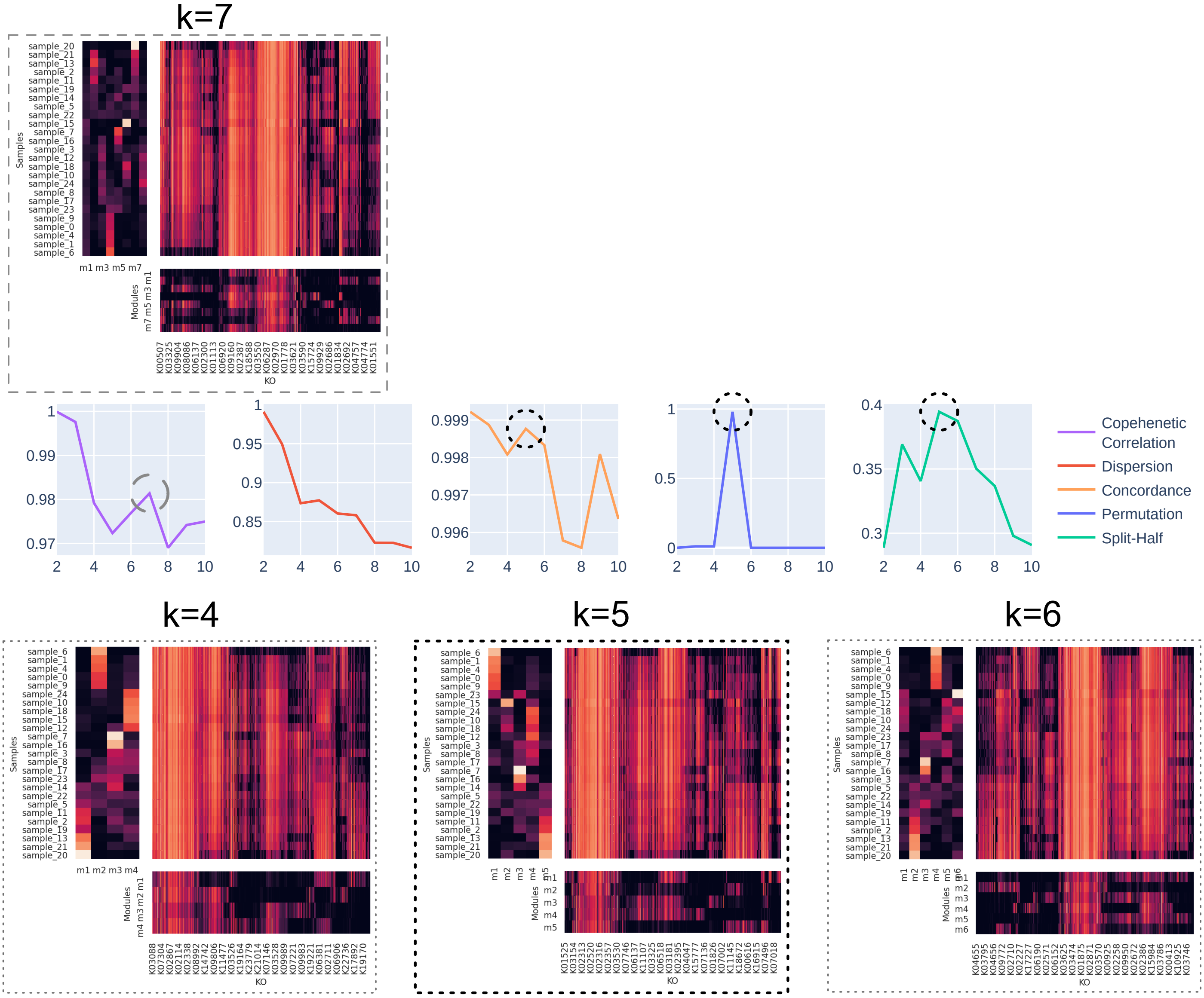


Fig. 5.12 Rank selection performed for the 5 genome simulated community (Section 5.3.6). Line plots in the middle show the value of model selection criteria over values of k searched. Concordance index, permutation and split half validation showed peaks at $k = 5$, indicated by the short dashed circle. The cophenetic correlation has a peak at $k = 7$, indicating by long dashed circle. The top heatmap triplot visualises the decomposition for $k = 7$. Module m1 has low weight across all samples, and few functions with high weight outside those which appear in all samples; m2 and m7 have have high weight in a similar set of samples and have a similar set of functions with high weight. The bottom row of heatmap triplots shows plots for the suggested rank of $k = 5$ in the centre, and ranks one above and below.

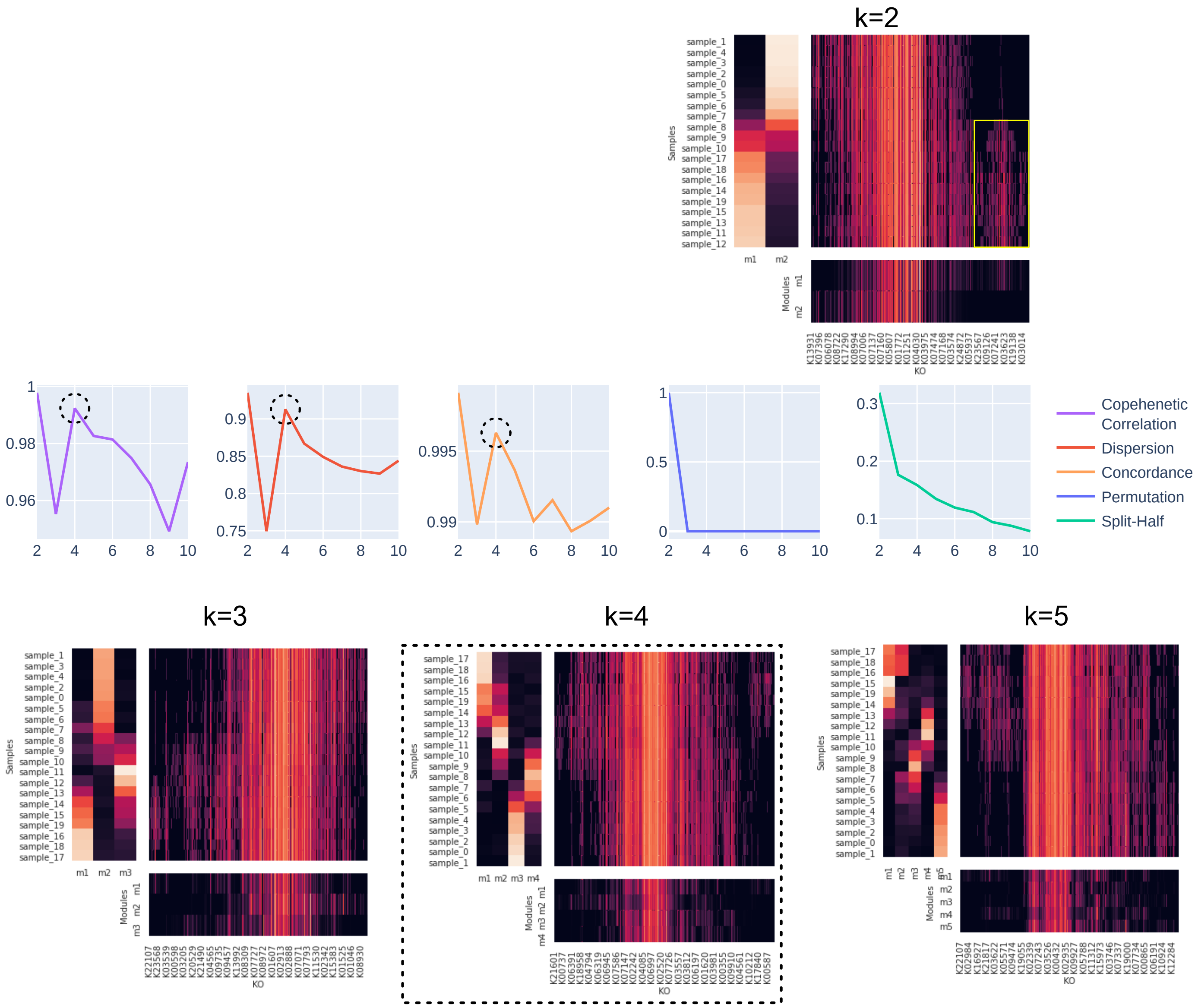


Fig. 5.13 Rank selection performed for the MOSAiC derived simulated community. This community has rank 4, with latent communities based on samples from different depths (Section 5.3.6). Line plots in the middle show the value of model selection criteria over values of k searched. Concordance index, cophenetic correlation, and dispersion showed peaks at $k = 5$, indicated by the short dashed circle. Permutation and split-half had no peak after $k = 2$. The top heatmap triplot visualises the decomposition for $k = 2$, where there appears to be a block of functions whose variation is undescribed by module m1, highlighted by the yellow block. Bottom heatmap triplots show the suggested rank $k = 4$, and one rank above and below. At $k = 5$, m1 seems to have been split into two modules m1 and m2, but functions with high weight in m2 appearing to be mostly those present in most samples.

5.4.2 Feature Assignment

Identifying which features describe a recovered module is an important step in interpreting matrix decompositions resulting from NMF. With the assumption that in the underlying structure of meta-omics data many features will be shared (genes belonging to multiple modules), we sought measures of feature importance which identified both unique and shared features. As explained in Section 5.3.3, we evaluated three methods (correlation, LOOCD, and permutation) in synthetic data to explore which would be most suited in the context of feature sharing. Figure 5.14 illustrates the performance of these three methods in synthetic data with overlapping features and samples. In an ideal measure, it would be possible to identify some point below which features do not belong to the module, and above which they do, whether unique, shared or ubiquitous. The box plots in Figure 5.14 illustrate that no measure showed such a clear cutting point, with the tails of the distribution of values for features not belonging to the underlying module and those unique to the module overlapping to some extent. However our exploratory analyses showed that typically LOOCD had the clearest distinction, with a number of outlier features not belonging to the module with high values, but the majority taking very low values below those typical of the shared and unique features, and below the upper three quartiles of the ubiquitous features.

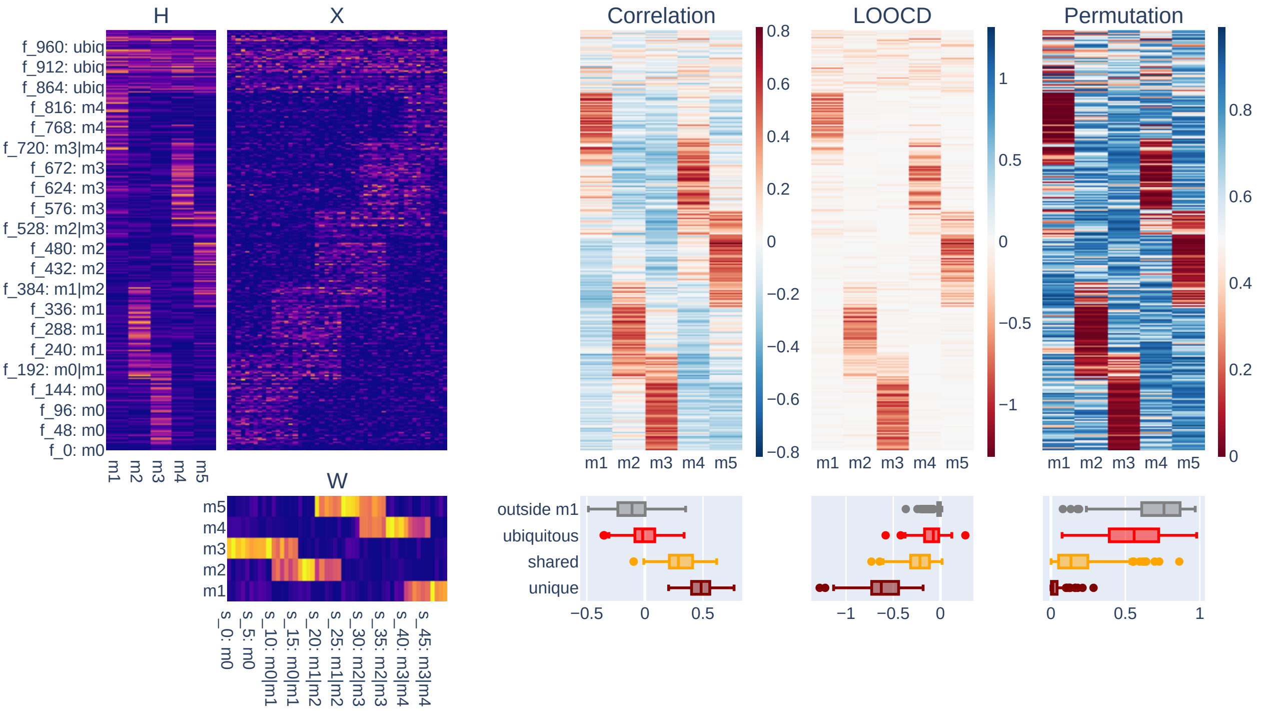


Fig. 5.14 Feature importance measures (Section 5.3.3) for a model learnt on synthetic data we generated with 5 underlying modules (Section 5.3.6). On the left is the data (X), and model (W, H), which is transposed to aid visualisation. Features are on the rows, and samples on the columns. Features and samples are labelled to indicate the modules they are in i.e. feature f_192: m0|m1 is in modules m1 and m2. To the right are heatmaps showing the values of the importance measures correlation, Leave-One-Out Correlation Decrease (LOOCD), and permutation for features matrix H. Colour scales have been chosen so that red indicates a feature which is more important to a module, blue or white less important. Beneath is a box plot showing the distribution of values for features in four categories. Firstly, features which are unique to underlying module m1 (maroon), those shared between m1 and another module (orange), features in all modules (red), and those features which do not belong in m1 (grey).

Values generated by the permutation and correlation methods tended to be distributed with two peaks, with features one side of the distribution tending to be those which belonged to the underlying module. We explored using this property to identify a suitable cutting point, by estimating a probability density function using kernel density estimation (Section 5.3). This method showed ability to identify shared and unique genes, but comparatively poor identification of ubiquitous features, as shown in Figure 5.15.

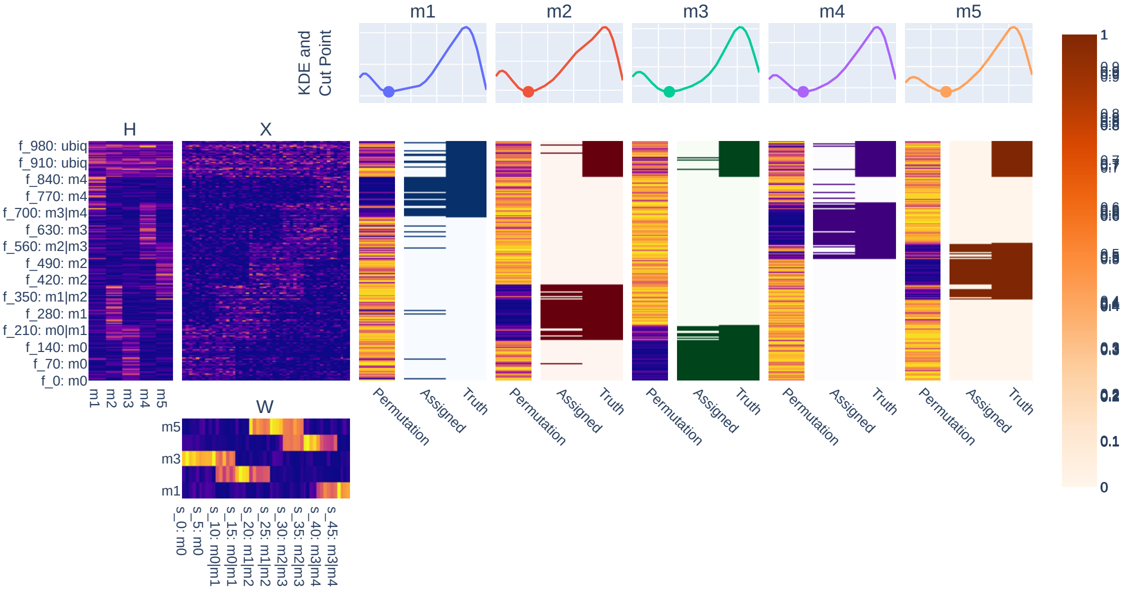


Fig. 5.15 Assignment of features based on the permutation importance measure with a threshold value identified through fitting a Kernel Density Estimate (KDE), as described in Section 5.3.3. On the left is the data (X), and model (W, H), which is transposed to aid visualisation. Matrix X is synthetic data with 5 modules we generated as described in 5.3.6. Features are on the rows, and samples on the columns. To the right are heatmaps showing the permutation importance values, and second heatmap showing features assigned using the KDE method, and the true underlying module. Above each of these a plot of the KDE, with the identified threshold value indicated with a point.

Two other methods of moving from measures of importance to binary assignment of genes to modules were explored, a greedy assignment algorithm and establishing a simple threshold value, evaluating their performance using relevance and recovery. LOOCD tends to show the highest peaks for relevance and recovery in synthetic data, for instance in example synthetic data in Figure 5.16 peaking at 0.69, in comparison to 0.59 and 0.62 for correlation and permutation respectively. Both the greedy algorithm and KDE methods achieved outcomes lower than would be possible by selecting a suitable static threshold. The KDE approach performed very poorly in combination LOOCD, as these values did not form a distribution with two clear peaks, preventing identification of an appropriate cutting point.

For a static threshold value to be useful, it would have to be stable across multiple datasets and across data with a different underlying number of modules. As LOOCD performed best in initial analyses, we evaluated performance of LOOCD with a threshold of -0.05, with results shown in Figure 5.17. For ranks 2 to 10, we generated 50 datasets with 100 samples and 1500 features, with 20% of samples and features being in multiple modules, with features randomly scaled and normally distributed noise with a standard deviation of 4 applied (see methods in Section 5.3.6). Overall relevance and recovery declined as rank increased, but the peak remained stable around -0.05, suggesting this to be a suitable default value for assigning features to modules.

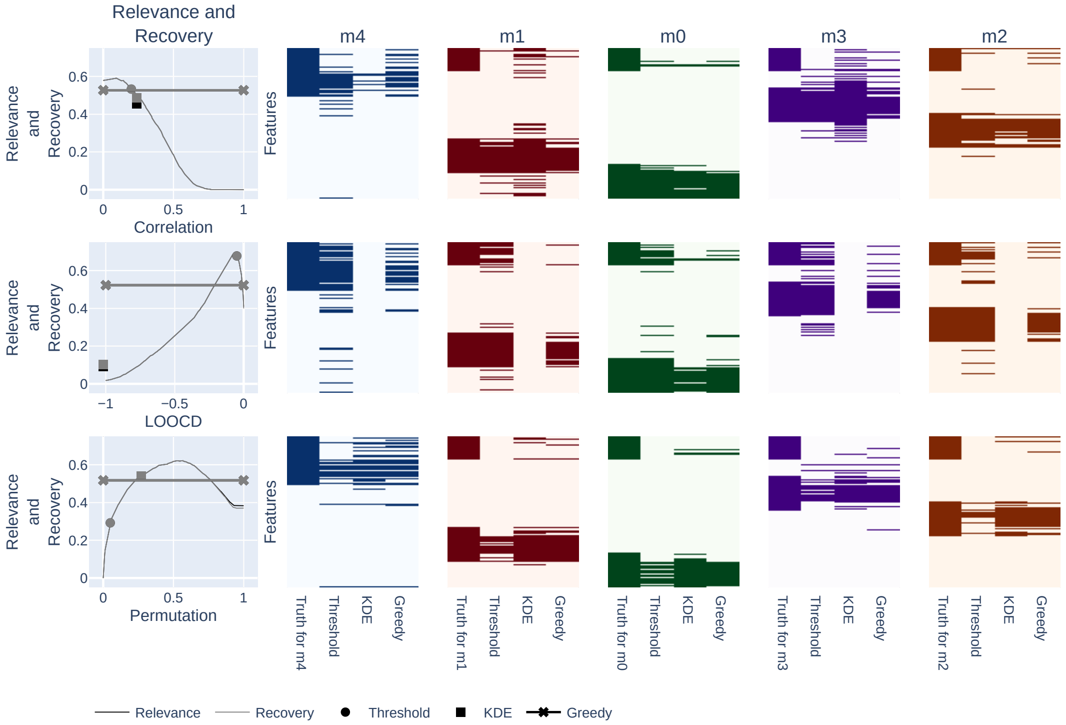


Fig. 5.16 Module recovery achieved by combinations of importance measures and feature assignment methods (Section 5.3.3). Each row shows results for a different importance measure: correlation, Leave-One-Out Correlation Decrease (LOOCD), and permutation respectively. The leftmost plot shows relevance and recovery on the vertical axis, and importance value on the horizontal. The curve plots relevance (grey) and recovery (black) when different threshold values are selected for assigning features to modules. In this case, relevance and recovery behaved very similarly, so at most points only one line is visible. A circle indicates the relevance and recovery achieved by using a default threshold value; a square the performance of the KDE method and the mean values of the thresholds identified for each module; and a horizontal line indicates the performance of the greedy assignment method (which does not select a threshold value). The heatmaps to the right show for each module the ground truth of which features are in each module, then the features assigned by each of the three methods.

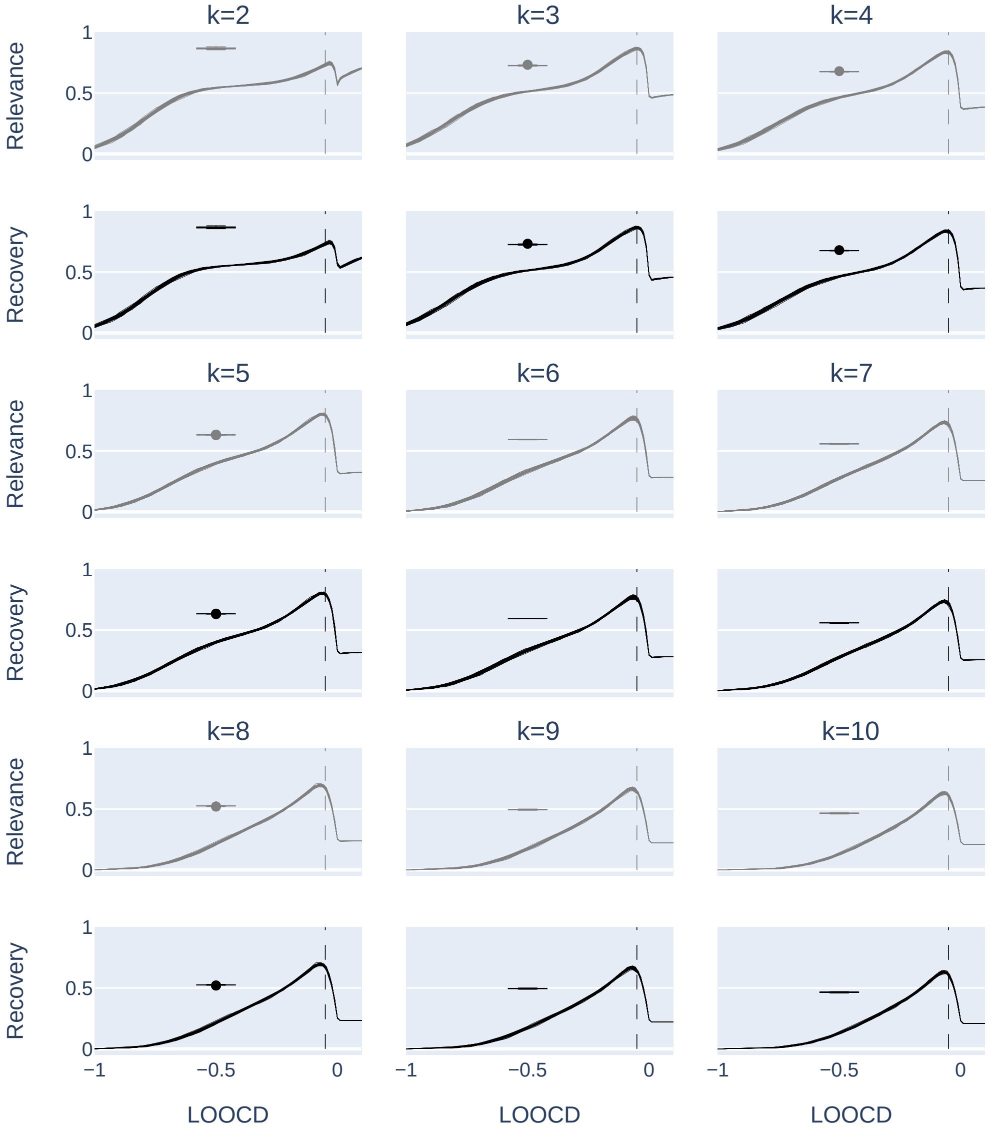


Fig. 5.17 Relevance and recovery of modules obtained using Leave-One-Out Correlation Decrease (LOOCD) and a threshold of -0.05. For each rank, 50 synthetic matrices (Section 5.3.6) were used with scores for each plotted as an individual line. The dashed line at -0.05 LOOCD is the proposed default threshold. A box plot shows the score achieved by the greedy assignment method, though there is not much variation in the scores so the box is quite compressed.

This combination of LOOCD with a threshold of -0.05 was applied to the two communities which were simulated in silico, the five genome and MOSAiC based communities (Section 5.3.6). In the five genome community, relevance and recovery again peaked at a low threshold value of approximately -0.005 with relevance and recovery 0.79. Using the proposed default threshold of -0.05 obtained a relevance and recovery of 0.73, shown in Figure 5.18.

For visual inspection of the feature assignment, we matched each recovered module to an underlying genome using the Hungarian algorithm with the Jaccard distance as costs. Each of these matched pairs had their precision and recall calculated, and the results shown in Figure 5.19. For comparison, WGCNA was also run for each dataset. WGCNA has a number of options and parameters; for our analysis of the simulated data we used the option to create a signed network, a soft power threshold of 10 for the five genome community, and 16 for the MOSAiC based community, a minimum module size of 30, and dynamic tree cutting.

Figure 5.19 shows strong correspondence between our assignments and the underlying genomes. Looking at the simpler five genome simulated community, for each genome, there are some false negatives (functions which should be present but were not identified), but fewer false positives. Across the five module and genome pairs, features are classified with a mean precision of 0.94 and mean recall 0.77. Comparing this to the modules identified by the hierarchical WGCNA method, it performs similarly well in areas with unique functions (turquoise, pink and blue modules), but shared genes are often split across multiple WGCNA modules. A group of functions shared by three of the genomes (amg, cps, tcx) is split across 9 different modules in by WGCNA. However, in the NMF modules the functions are assigned 55.9% to all three, 21.8% to two, and 20.7% to only one of the corresponding modules. For this group in the WGCNA modules, a large proportion (55.3%) are assigned to the red module. These red module functions are similar to those assigned to all 3 NMF modules, so the two approaches capture some of the same information. However, the WGCNA red module contains functions originating from the other two genomes (pma, ter), making intuitive interpretation of this module more challenging. In the area of ubiquitous functions, those which are present in all the underlying genomes, our function assignment method performs similarly well, with 50.4% being assigned to all 5 modules, and 16.5% to 4 modules. The modules of functions identified using NMF and LOOCD are partial, but offer an interpretive benefit, particularly in regions of high sharing, over a hierarchical approach.

We also simulated a more complex community based on pilot MOSAiC sampling, with four underlying modules. Each of these modules is a mix of genomes, and each simulated sample is a mixture of these modules. As with the simpler community, we evaluate the correspondence between the functions assigned to the recovered modules, and those present

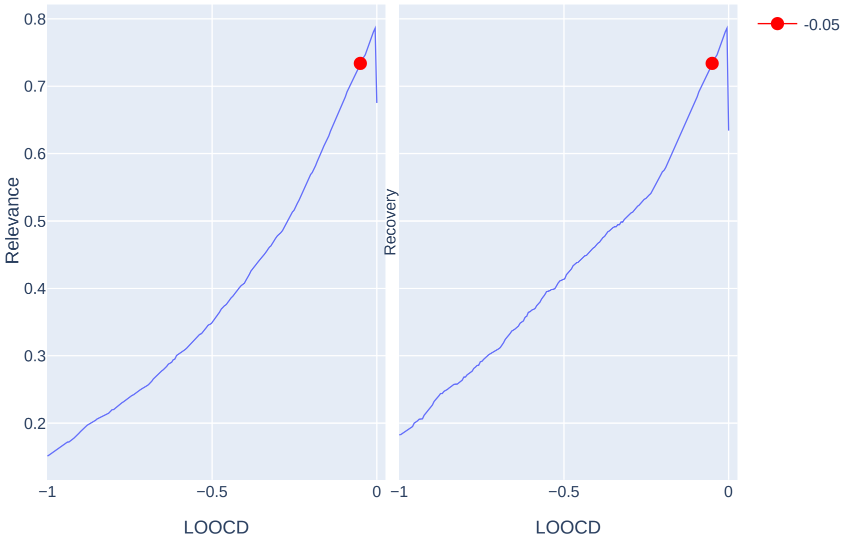


Fig. 5.18 Recovery and relevance scores for the 5 genome simulated community (Section 5.3.6) across different Leave-One-Out Correlation Decrease (LOOCD) thresholds. The default threshold of -0.05 is indicated by a red point, slightly beyond the points of peak score.

in the genomes of the latent modules. This simulated community is characterised by a much smaller proportion of unique functions; 68.9% of functions are found in all of the latent modules, and only 7.8% are unique to one. For the ubiquitous functions, fewer were assigned to all recovered modules than in the simpler community; 23.7% to all four recovered modules, 37.9% to three, 24.3% to two, and 13.9% to only one. The reduced proportion of shared functions assigned to all modules is reflected in the recall, with a mean recall of 0.68. However, the precision remains similarly high as it was in our simpler community, with mean precision of 0.98. The functions unique to one latent module are more well identified than those shared. For latent module comm_4, 94.4% of the unique functions are correctly identified. This compares favourably to the WGCNA classification, where only 55% of the unique functions are grouped together in the same module M3. Overall fewer functions were assigned a module by WGCNA, with many unassigned, represented by the grey module M0. In this more complex simulated community, the recovered modules remain partial. The high precision suggests that confidence can be merited that those functions assigned to a module belong together.

For our simulated communities, we know for each sample the ratio at which the latent modules are mixed to generate the sample. Another way to evaluate our results is by comparing the correlation between these known mixing ratios, and the relative weights for

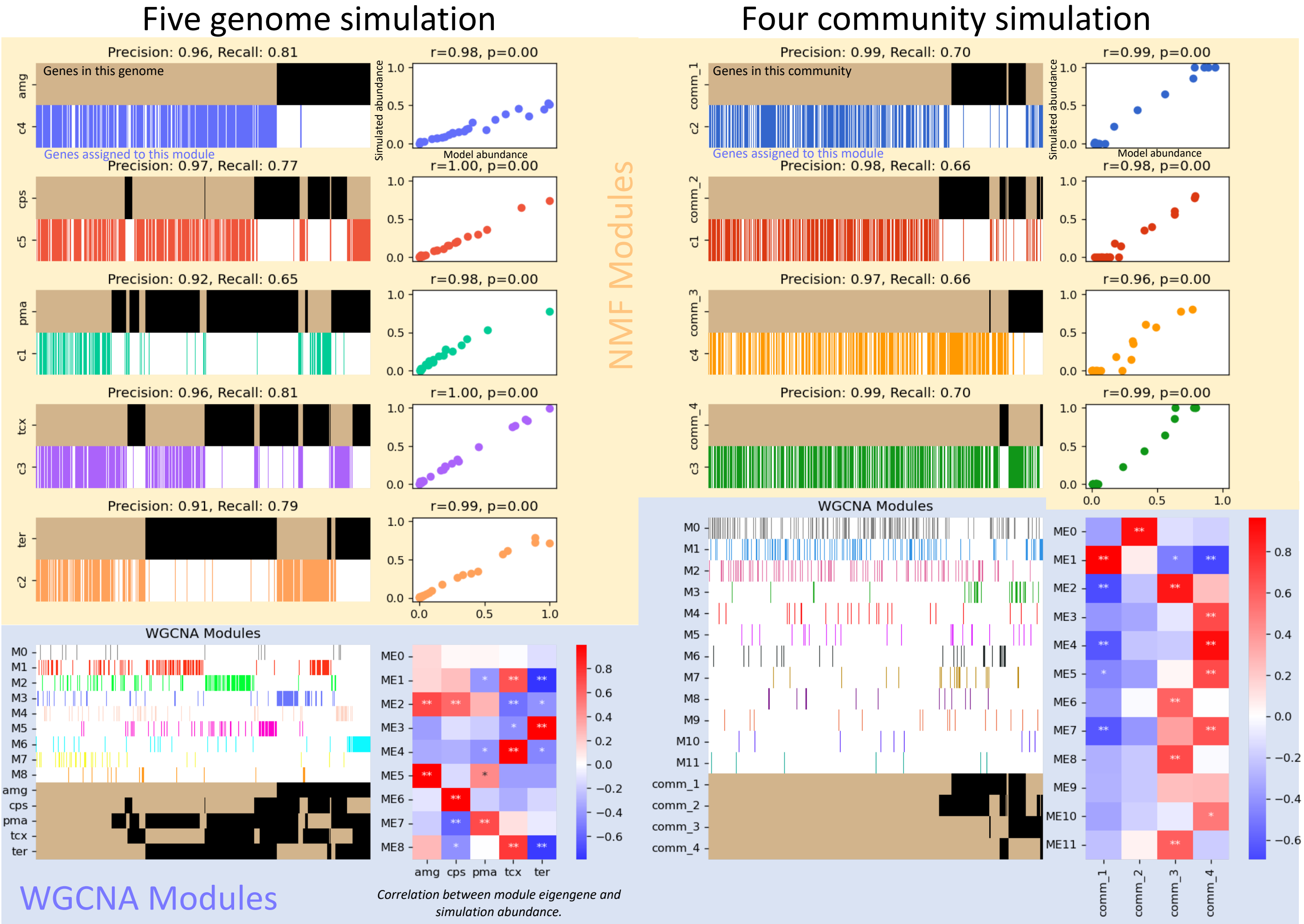


Fig. 5.19 Comparison of modules recovered by NMF (peach background) and WGCNA (blue background). The 5 genome simulated community is on the left, the MOSAiC derived community on the right. For the NMF recovered modules, each recovered module is paired up with one of the underlying modules, and heatmap showing the ground truth of which features should be in the module in tan versus those which are assigned in the recovered module in colour. A scatter plot also shows the correlation between the relative weight of the module in W and the relative abundance used in generating the underlying simulations. The WGCNA modules are not paired up as there isn't a 1:1 relationship between underlying and recovered modules. Again tan represents the true underlying modules, and coloured strips the recovered modules. The grey M0 module is features which have not been assigned to a module. The heatmap to the right shows the correlation between each recovered module eigengene and the abundance of each underlying module used in generating the simulated data.

each module in each sample. For the five genome community, $r \geq 0.97$ for all modules. Recovered modules correlate poorly with latent modules other than the one they were matched with, with the mean off-pairing correlation being -0.23, with a maximum of 0.47. The MOSAiC derived community also showed strong correlation with $r \geq 0.96$ for all modules, and off-pairing correlations were again weaker, with a mean -0.31 and maximum of -0.0038.

The WGCNA modules can be evaluated in a similar way, by relating the module eigengene to latent module mixing ratios. Some recovered modules show similar correlations with an intuitive relationship to the underlying genomes; for example, in the five genome community the turquoise module M6 largely consists of functions unique to the cps genome, and it's strongly positively correlated to the mixing ratio of cps. Outside of unique functions, the correlations are poorer; the red module M1 contains many functions shared between the genomes pma and ter, however it has a negative correlation to the mixing ratio of ter, as it also contains more functions shared by the other 3 genomes, amg, cps & tcx. In the MOSAiC derived community, latent module comm_2 has only a strong positive correlation to the unassigned functions in the grey module M0, and no significant correlation to the recovered modules. The opposite is found in comm_4, being significantly positively correlated to 5 of the recovered modules.

The modules identified by NMF give a comparatively straightforward description of the distribution of the underlying modules across samples for our simulated data.

5.5 Real World Case Studies

5.5.1 HMP

Microbiomes from different parts of the body are known to be taxonomically distinct, though their function is more stable with a consistent core of housekeeping functions [27], with site specific functionality being consistent between individuals even where taxonomic composition varied [401]. Given these established, stable diferences in function between sampling locations on the body, we should be able to recover modules representing these functions and indicating a clear separation between sampling locations. We used data collected during the HMP1 stage of the HMP, where samples were collected from different locations on the body from healthy individuals. Each of the 686 samples is labelled with a precise location (e.g tongue dorsum, anterior nares), which can be further grouped into four broad groups of oral, skin, gut, and vaginal samples.

HMP functional data has been investigated previously using NMF, however this is often with the goal of distinguishing between binary classes. Cai et al. [371] applied a supervised modification of NMF to distinguish between two individuals who were sampled over a time series in a different part of the HMP, as well as between gut samples of healthy and diseased individuals in a different set of data from the MetaHIT project [402]. A constrained approach incorporating experimentally established expert curated knowledge about metabolic pathways was applied to study fiber degradation in the gut [370], focusing on a single environment rather than broad functional differences between environments. Extending NMF to allow joint analysis of taxonomic and functional data provided clear separation between the broad sampling locations [403]. Given that function appears more stable across sampling location that does taxonomy, we would hope to be able to achieve the same using only functional data.

As input data we used the relative abundance of KO terms in each sample, a total of 13,325 functions. Figure 5.20 shows results for rank selection, relationship of the decomposition to the sampling locations, and functional enrichment analysis performed on this data. Model selection showed a clear peak of the concordance index at $k = 4$, remaining high for $k = 5$ before declining. Inspecting the decompositions for both of these ranks, rank 4 shows a clear separation into the four broad groupings. At rank 5, the oral samples split into two modules, one highly weighted in mostly buccal mucosa samples, and the other representing mostly the remaining oral samples. This clear separation between known classes suggests the techniques we have selected are able to identify highly distinct underlying communities, as was shown by joint decomposition of taxonomic and functional data [403]. Additionally, enrichment analysis reveals biologically meaningful information based on the indentified modules; for example pathways related to biofilm formation are enriched in oral samples, fitting with the surface associated communities which form in the oral cavity.

We investigated whether NMF could further split a single of these broad environments into the specific sampling locations, when the input data is restricted to only samples from that group. Oral samples were selected, as other groups are more imbalanced with a large majority of samples originating from one of the specific locations (anterior nares for skin; posterior fornix for vagina). Samples from the oral cavity originate from 9 different specific locations, with the 123 from the tongue dorsum, 115 from supragingival plaque (teeth above the gum), 109 from the buccal mucosa (cheek), and less than 10 from all other locations. Results are shown in Figure 5.21. Rank selection showed peaks in cophenetic correlation and dispersion at $k = 4$, and similarly the concordance index peaks at $k = 2$ and $k = 4$ before declining, with the split-half method also showing a peak at $k = 4$. Inspecting the decomposition for $k = 4$, W has module m1 mostly associated with the buccal mucosa

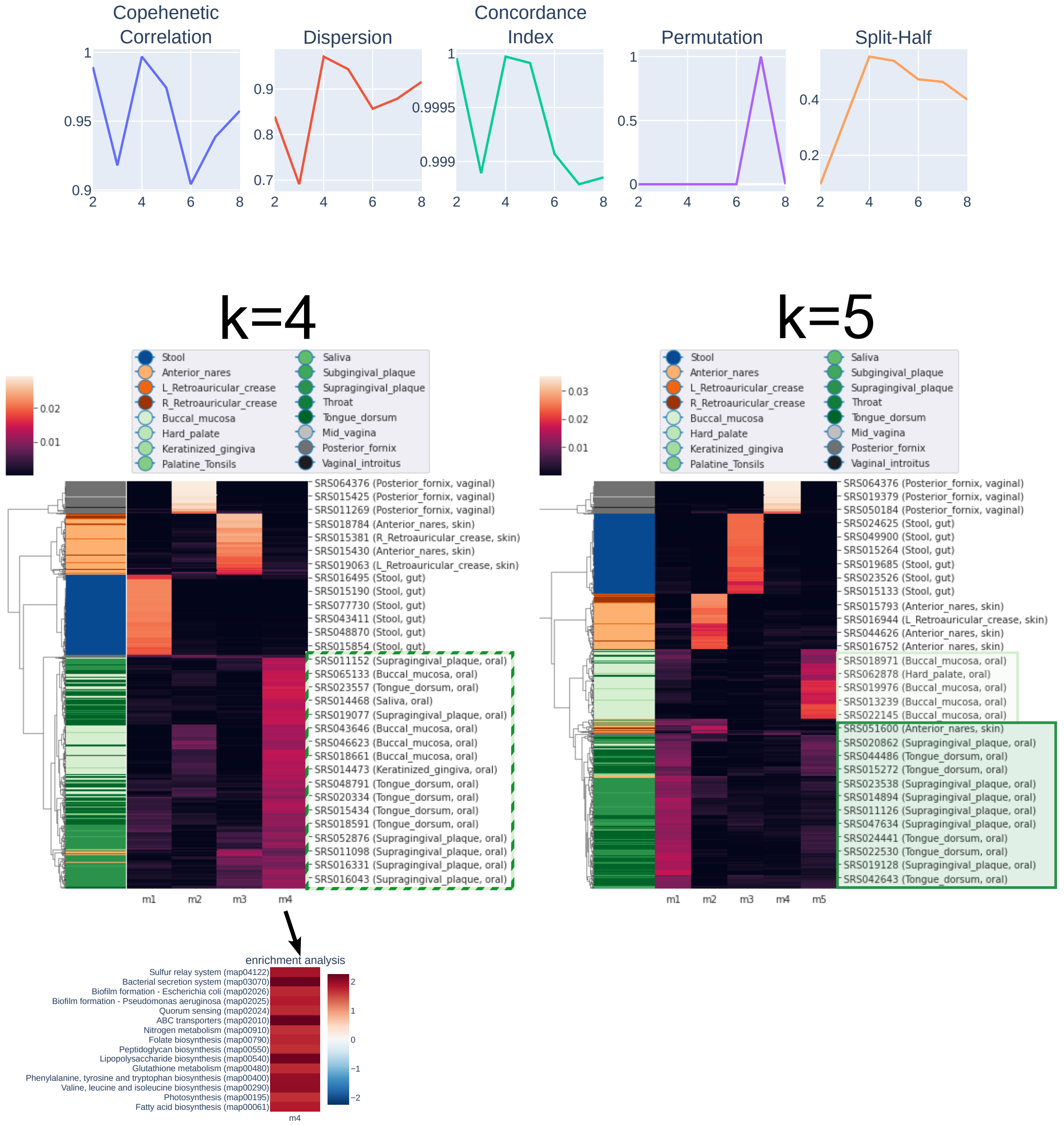


Fig. 5.20 NMF methods applied to the Human Microbiome Project (HMP) functional data [27, 400], described in Section 5.3.7. Top shows model selection, where all methods other than permutation showed a peak at $k = 4$, and remaining high for $k = 5$ in most methods. Below are shows sample matrices W for decompositions with $k = 4$ and $k = 5$. The coloured ribbon indicates the location the sample was taken from. The four broad categories of gut, vaginal, skin, and oral are indicated by blues, greens, greys, and reds respectively. The more detailed locations are shown by shades within those colours. It is evident that each recovered module relates strongly to one of the broad categories; for $k = 5$, the additional module mostly represents the buccal mucosa samples. The mixed group of oral samples in the $k = 4$ decomposition, represented m4, is highlighted by a striped green box; in the $k = 5$ decomposition, the largely buccal mucosa and other oral sample groups are highlighted by light and dark green boxes respectively. At the bottom, a GSEA enrichment for $k = 4$ for module m4, associated with oral samples, shows enrichments for biologically meaningful pathways such as biofilm formation.

samples, and m4 mostly associated with supragingival plaque. The final large group, tongue dorsum samples, appears to be represented by two modules, m2 and m3, where only a subset of the tongue dorsum samples has high weight for module m3. KEGG pathways for biofilm formation are enriched in module m3, suggesting this module may be a module of functions from more established surface associated coatings. The extent of tongue coating can vary between individuals, and is associated with different microbial communities, and can be influenced temporarily by activities such as brushing of teeth or use of a tongue scraper [404]. Examining the KO terms which are assigned to m4 after using the LOOCD method with the default -0.05 threshold, module m3 has 2,104, smaller than the other tongue dorsum module m2 which has 4,875. Approximately half of the functions in m3 (1,040) are not assigned to m2, and 540 are unique to m3, suggesting m3 represents accessory functions active in a subset of the oral samples. This is supported by looking at the decompositions for $k = 3$, which produce a module with high weight for a mix of supragingival plaque and tongue dorsum samples, and a separate module with high weight in a small number of tongue dorsum samples, suggesting a subset of tongue dorsum samples with distinct function.

Using the HMP as a case study, we have shown the ability of NMF to recover well established groupings from functional data. Additionally, applying this to the less distinct oral samples illustrated a strength of the non-hierarchical approach of NMF in describing tongue dorsum samples as a mixture of two modules of functions, allowing a more clear separation of tongue and supragingival plaque samples.

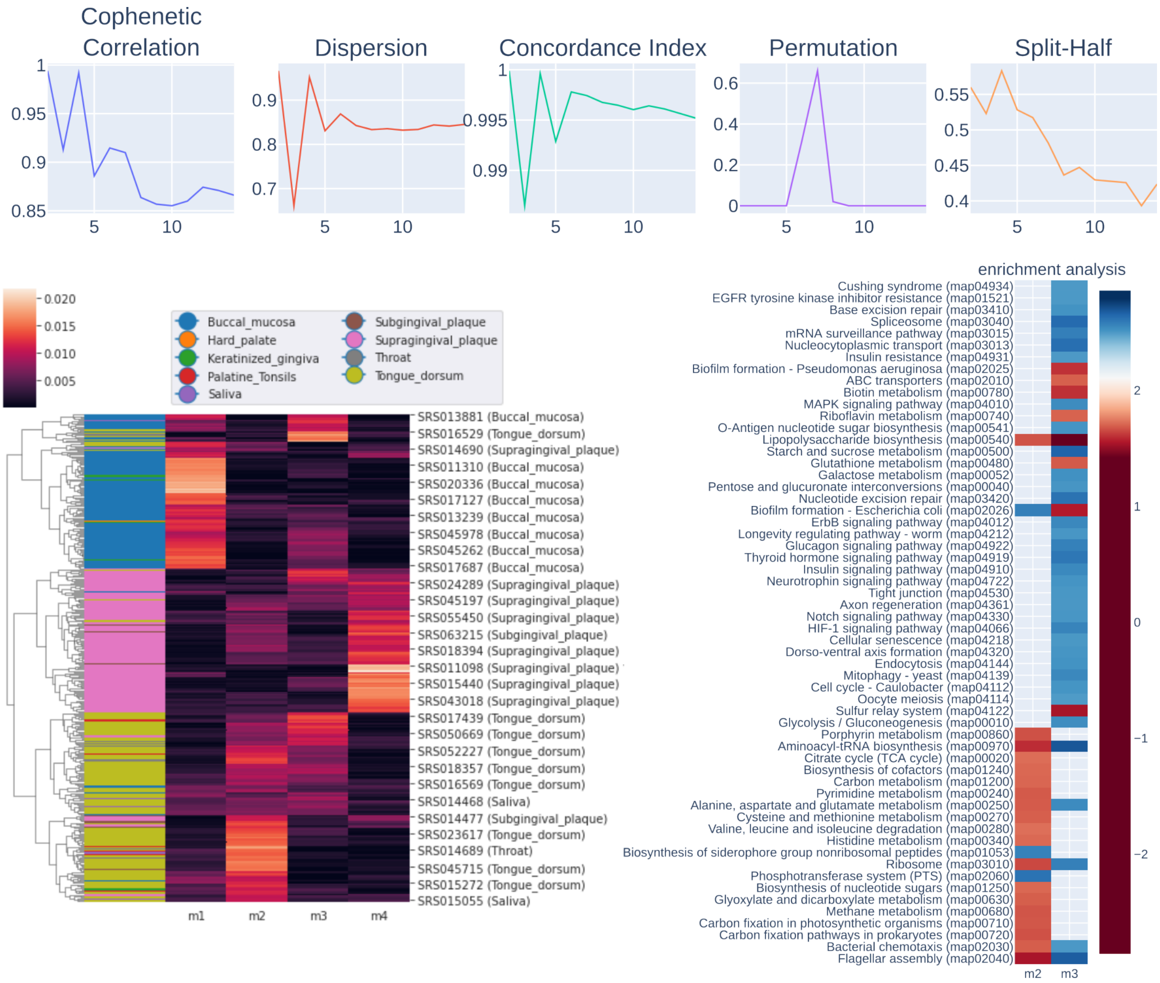


Fig. 5.21 NMF methods applied to the Human Microbiome Project (HMP) functional data for oral samples only [27, 400], described in Section 5.3.7. Top shows model selection, where all methods other than permutation showed a peak at $k = 4$. Below shows sample matrices W for decomposition with $k = 4$. The coloured ribbon indicates the location the sample was taken from. Each module associates with one of the larger classes, though the grouping is less clear than observed in earlier results on the broad categories of gut, skin etc. At the bottom right is and enrichment for the two buccal mucosa associated modules m2 and m3, showing for m3 many terms depleted and a few terms such as biofilm formation enriched.

5.5.2 Waiwera River Estuary Water and Sediment

We analysed the data presented in Tee at al. [1] using these NMF techniques. This data comes from a 5km portion along the Waiwera estuary, with samples taken from both sediment and water. These microbiomes are thus separated into two distinct groups (sediment, water), and along a salinity gradient from fresh to marine waters, which would be expected to drive the functional composition of the communities. There are 30 samples in total, 9 water and 21 sediment, and functional annotation is against a set of 83 genes which participate in important metabolic processes. Compared to the human microbiome data presented previously, the number of samples is much smaller, which is typically to be expected for environmental data. Tee et al. [1] analysed this functional data using several methods including applying WGCNA to the metatranscriptomes, finding 6 modules, and based on these module eigengenes a separation observable between sediment and water, and within each a separation by salinity.

We used log-scaled transcripts per million mapped reads for each marker gene as input data, and repeated the WGCNA analysis of Tee at al. using parameters stated in the paper [1]. NMF decompositions were performed using the multiplicative update method with KL divergence. Genes were assigned to each module using the LOOCD method with the default threshold of -0.05. During rank selection, shown in Figure 5.22, cophenetic correlation, dispersion and the concordance index were highest at $k = 2$; inspecting these decompositions showed modules corresponding to the material sampled (sediment and water). To identify modules within each of these materials sampled, we looked at higher ranks indicated by the model selection methods. The cophenetic correlation and dispersion have lower peaks at $k = 4$ followed by $k = 7$, and the concordance index at $k = 7$.

Inspecting the decomposition for rank $k = 7$, where there is some agreement between three of the model selection methods, shows a distinction between sediment and water samples, with three and four modules representing each respectively. The NMF decomposition and WGCNA modules are shown in Figure 5.23. The three water modules split up between fresh water samples (m1) and those from brackish & marine waters (m2 and m3). The sediment samples have one module with high weight in fresh water samples (m4), and the remaining brackish samples are described by a mix of two modules (m5 and m6). The final module, m7, appears largely spread with low weight across the sediment samples. Some samples have a mixed description in terms of module weights. The sediment samples from station 3 are a mixture of modules m5 and m6; water samples from stations 3 to 6 are a mixture of m2 and m3.

Both WGCNA and NMF identify a separation between sediment and water, and fresh and marine or brackish samples within those. However, the overlapping assignment of

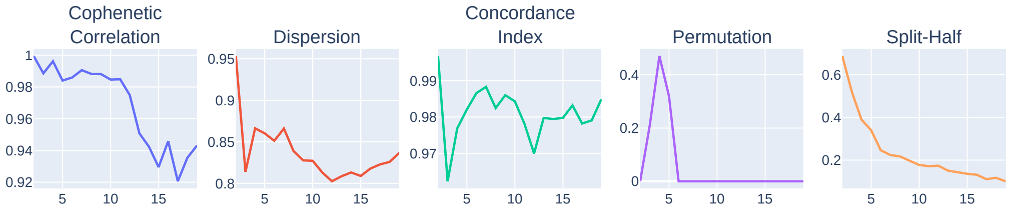


Fig. 5.22 Rank selection for $k = 2..20$ for the Waiwera river estuary data. Cophenetic correlation, dispersion, concordance, and split-half methods all took their highest values at $k = 2$. Subsequent peaks occurred again at $k = 4$ for cophenetic correlation, dispersion and permutation methods, and $k = 7$ for concordance.

genes allows us to identify some functional differences which are not apparent from the hierarchically recovered WGCNA modules. Below we look in detail at some of the additional information visible through the NMF modules and how it can suggest biologically meaningful interpretations.

In the input data, some genes appear ubiquitous, such as TrkA. The hierarchical approach places this gene in the yellow module, where LOOCD assignment on the NMF modules assigns it to 6 of the 7 recovered modules. This illustrates the potential for NMF to provide an intuitive description in cases of high gene sharing. The assignment of genes involved in nitrification to the recovered modules highlight some metabolic differences among the sediment microbiomes. Nitrification is commonly a two-step process, oxidation of ammonia to nitrite carried out by ammonia oxidising archaea or bacteria (AOA or AOB), and subsequently oxidation of nitrite into nitrate by nitrite-oxidising bacteria. Some Commamox organisms incorporate both steps. Nitrogen cycling was presented in more detail in Section 2.2.3. The nitratation genes nxrAB are assigned to modules m4 and m5, but absent in module m6. Samples from station 7, the sediment closest to the marine waters, have very high weights for module m6, suggesting reduced nitrification activity in these samples. In the WGCNA modules, these genes are assigned to the brown module, whose eigengene has even positive values across the brackish samples, which does not highlight the reduced transcription of nitrification genes in samples from station 7.

The gain or loss of genes in components can describe differing strategies. Na+/H+ antiporter genes mnhBEFG related to osmoregulation are assigned to our modules representing brackish samples, modules m3, m5, and m6. These functions are absent in the freshwater modules m1 and m4. These genes are all assigned to the yellow WGCNA module, whose eigengene has negative values for the brackish sediment samples from station 3. In the

NMF decomposition, these stations 3 sediment samples are described mostly by mixtures of modules m5, m6, and m7, where m5 and m6 both have both the mnhBEFG genes assigned assigned, suggesting this as a function shared across all brackish sediments sampled. Local patterns can be identified among the NMF modules, such as the presence of nitrate reduction genes narGH genes in fresh water samples, albeit it at low abundance, in addition to their abundance in sediment. These genes are not assigned to the brackish and marine water associated modules m2 and m3, but are found in the sediment and fresh water associated modules; however the same genes are in the brown WGCNA module, whose eigengene is negative for the fresh water samples, not capturing the increased transcription of these genes in the fresh water environments sampled.

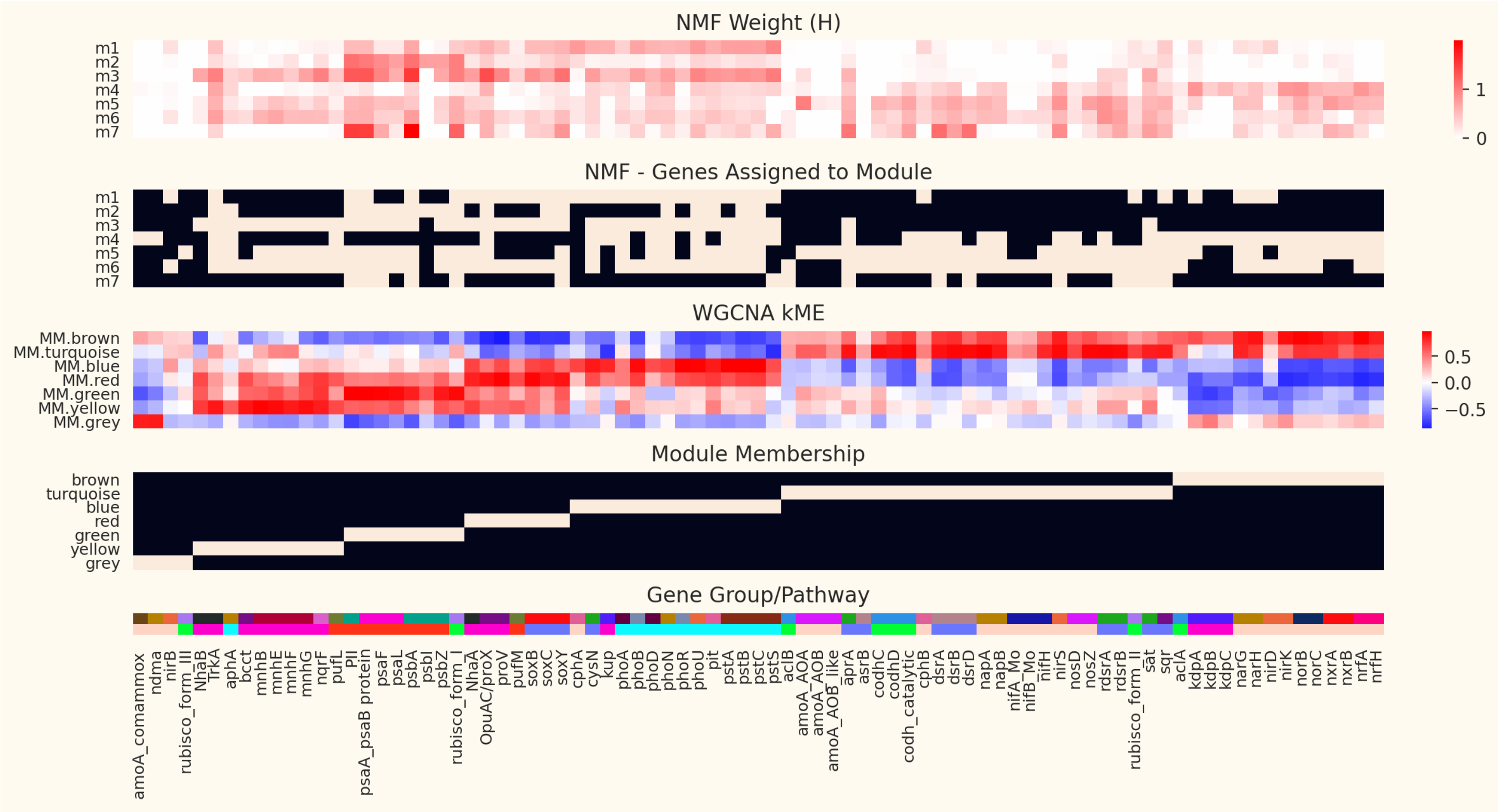
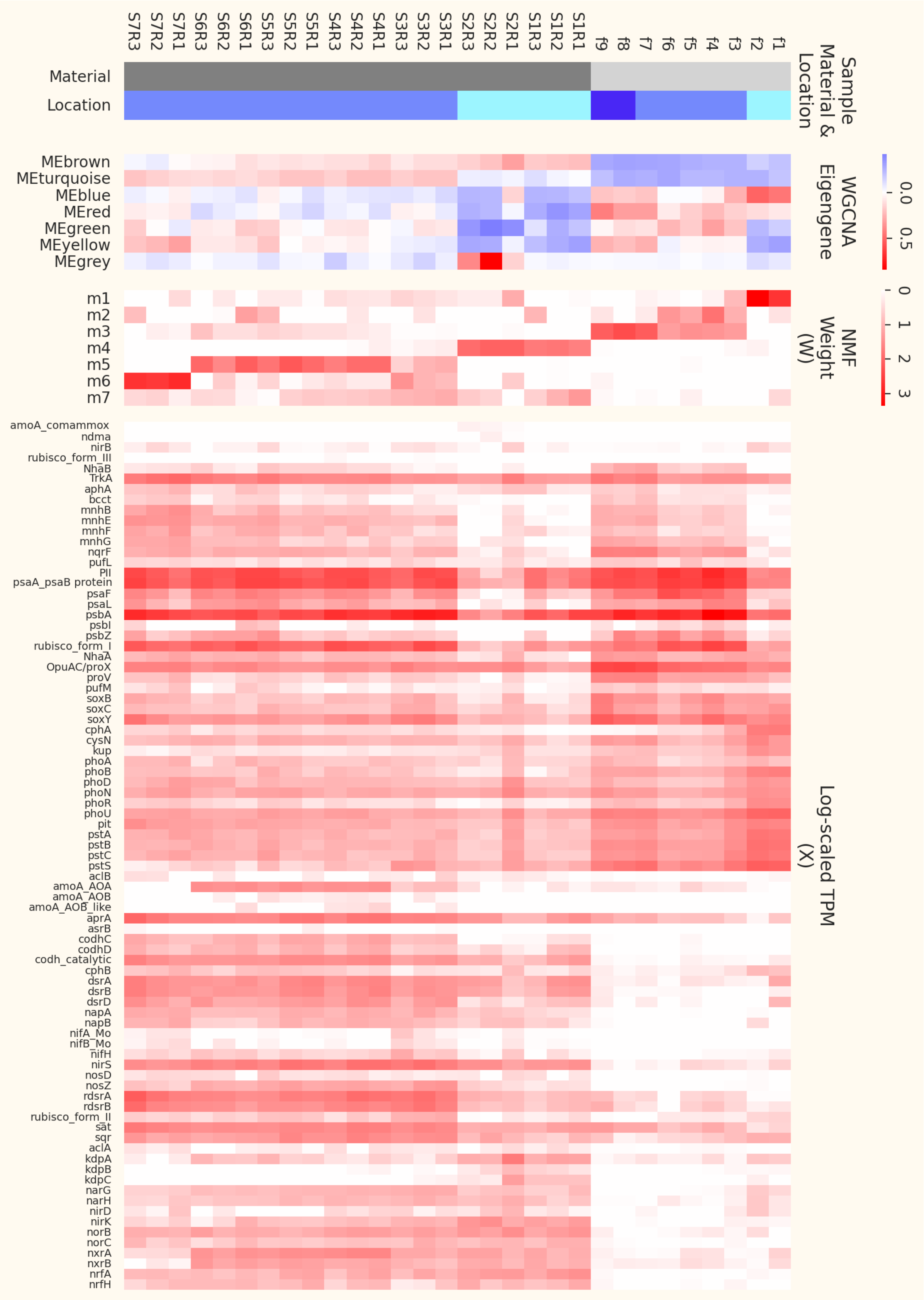


Fig. 5.23 Comparison of modules from WGCNA and NMF analysis of data taken from the Waiwera river estuary [1]. This Figure spread across two pages for legibility. Samples were taken at nine station along the river estuary from both the sediment and water column, along a gradient from fresh to marine waters, with station 1 being fresh water and 9 marine. One water sample (f1-9) was sequenced from each station; three sediment samples were taken from each station up to 7 (S1-7), sediment being unavailable for the marine stations 8 and 9. A heatmap indicates the sample material and location: light grey for water, dark grey for sediment; turquoise for fresh, light blue for brackish, and dark blue for marine stations. NMF modules weights for matrices W and H are shown as heatmaps. The similar matrices for WGCNA are shown alongside for comparison: the module eigengene is roughly equivalent to the W matrix, and kME is the correlation of the eigenegene to each gene, a fuzzy measure of how related each gene is to a module. Genes are group by pathways described in Appendix B.1 so related genes are close together, with coloured ribbons indicating which genes are in the same pathways at two levels.


S7R3
S7R2
S7R1
S6R3
S6R2
S6R1
S5R3
S5R2
S5R1
S4R3
S4R2
S4R1
S3R3
S3R2
S3R1
S2R3
S2R2
S2R1
S1R3
S1R2
S1R1
f9
f8
f7
f6
f5
f4
f3
f2
f1
Sample
Material &
Location
Material
Location
MEbrown
MEturquoise
MEblue
MEred
MEgreen
MEyellow
MEgrey
WGCNA
Eigengene
0.0
0.5
m1
m2
m3
m4
m5
m6
m7
NMF
Weight
(W)
0
1
2
3
amoA_comammox
ndma
nirB
rubisco_form_III
NhaB
TrkA
aphA
bcct
mnhB
mnhE
mnhF
mnhG
nqrF
pufL
PII
psaA_psaB protein
psaF
psaL
psbA
psbI
psbZ
rubisco_form_I
NhaA
OpuAC/proX
proV
pufM
soxB
soxC
soxY
cphA
cysN
kup
phoA
phoB
phoD
phoN
phoR
phoU
pit
pstA
pstB
pstC
pstS
aclB
amoA_AOA
amoA_AOB
amoA_AOB_like
aprA
asrB
codhC
codhD
codh_catalytic
cphB
dsrA
dsrB
dsrD
napA
napB
nifA_Mo
nifB_Mo
nifH
nirS
nosD
nosZ
rdsrA
rdsrB
rubisco_form_II
sat
sqr
aclA
kdpA
kdpB
kdpC
narG
narH
nirD
nirK
norB
norC
nxrA
nxrB
nrfA
nrfH
Log-scaled TPM
(X)

This case study illustrates two broad points. Firstly that NMF separates samples in a way that is congruous with both biological expectation and results of other methods such as WGCNA, identifying modules associated with sediment and different salinity conditions. Secondly, that the non-discrete assignment of genes can produce additional functional insight into the environments being studied otherwise not evident, such as reduced nitrification activity in the sediment of station 7, or the ubiquity of osmoregulation related activity in brackish samples.

5.5.3 TARA Oceans Surface Ocean Metagenomic Data

Our first approach to the TARA Oceans data was to use data from different depths as an additional validation that the approaches we have selected could distinguish environments known to be functionally distinct. We selected all samples from the deep chlorophyll maximum layer (DCM) and mesopelagic, as the former is characterised by raised primary productivity and photosynthesis, while the low light conditions in the mesopelagic mean these processes are not expected to be present at this depth, making the communities at the two depths quite functionally different. Where there were multiple samples from the same station and depth, one was randomly retained, due to our preliminary work showing models with replicates could end up producing modules representing only the replicates; a module only describing one sampling point is not particularly informative. This resulted in 69 samples, with 39 DCM and 30 mesopelagic samples, from stations shown in Figure 5.24. The expectation is that the models built from data relating to these two distinct ocean regions will identify modules whose weights are clearly divided between the two layers, and for which weights of features support the expected divided in metabolism between the two regions.

The data was log scaled so entry $y_i = ln(y_i + 1)$, to maintain the approximately 30% of values which are 0. Features in this data are the relative abundance of InterPro entries. A model was trained using the KL beta loss function, model selection performed using 100 iterations for each method, for $k = 2..20$ with 100 iterations for each value of k, results are shown in Figure 5.24. Most of the methods peak at $k = 2$, with further peaks at $k = 6$. Inspecting the W matrix for $k = 2$ (Figure 5.25) shows a majority of mesopelagic samples assigned a high weight for module m1 with very low weight for m2, and the inverse for DCM samples. Using this model as a simple classifier, assigning each sample to the component it has the highest weight for and assigning m1 to be mesopelagic and m2 to be DCM, results in 97% accuracy. We also identify functional differences in the two modules which fit with this distinction between the light filled DCM and the dark mesopelagic. GSEA analysis showed m2 to be enriched for a number of photosystem components, reflecting the absence

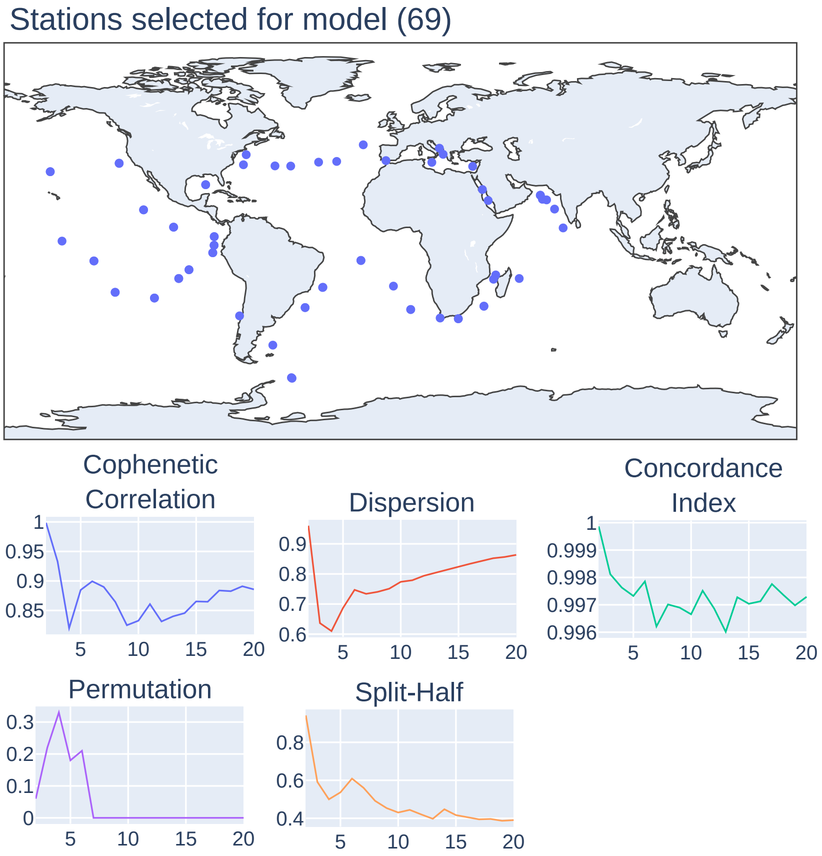


Fig. 5.24 Top: Location of samples selected from the EBI annotation of Tara Oceans expedition for inclusion [3, 4]. Bottom: Model selection values for $k = 2..20$, showing peaks at $k = 2$ and $k = 6$ for all methods but permutation, which peaks at $k = 4$. Dispersion has a peak at $k = 6$ but continues to rise after this.

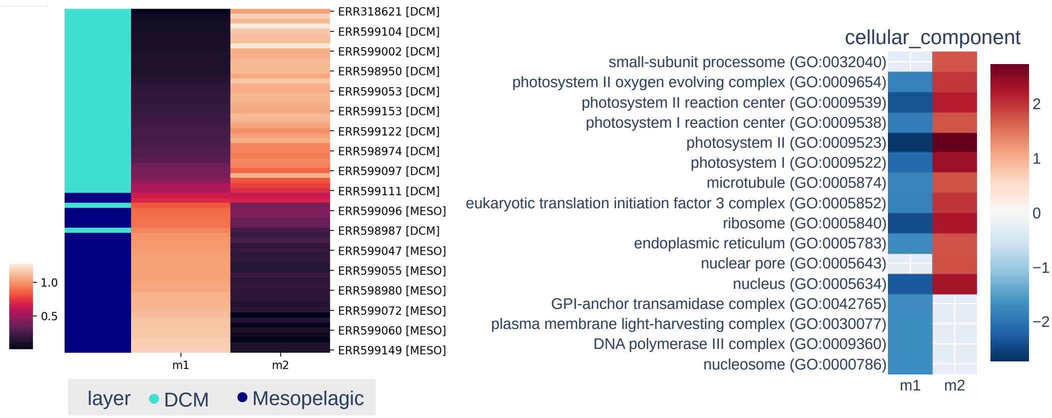


Fig. 5.25 Results of decomposition of Tara Oceans data [3, 4] from DCM and Mesopelagic depths. Left: W matrix and coloured ribbon indicating the depth the sample originates from, showing module m2 associated with DCM samples and m1 with the mesopelagic. Right: Gene Set Enrichment Analysis (GSEA) results showing DCM associated m2 enriched for photosynthetic components, reflecting an expected functional difference between the two depths.

of photosynthesis in the mesopelagic. Our model appears to distinguish between the two sets of samples we expected to be quite distinct.

There were also peaks at $k = 6$ in most model selection criteria, so we inspected the decomposition at this rank to see if it contained additional useful information (Figure 5.26). Again, most modules are associated with a given depth. The mesopelagic samples are associated with m1 and m4. Most DCM samples are a mixture of m2, m3, and m5. Plotting the module weights in two dimensions using PCA shows three groups mainly separating by depth, with three samples which are out of place. Enrichment analysis of this decomposition again shows some photosystem components enriched in all these DCM modules. Module m3 appears to represent eukaryote specific functions not found in the other DCM modules, such as endoplasmic reticulum, nucleus, and MCM complex. This module has high weight in the two southernmost samples (green on maps in Figure 5.25). Samples were size fractionated to select for prokaryotes, being filtered for either $0.22 - 1.16\,\mu\text{m}$ or $0.22 - 3\,\mu\text{m}$, however some eukaryotic picoplankton such as species of Micromonas or Bathycoccus still fall within this larger cell size range. Module m3 appears to reflect samples where a greater portion of picoplankton community is eukaryotic. The two mesopelagic modules m1 and m4 are again characterised largely by depletion of terms, with only one term appearing enriched (chromosome). The final module m6 has high weight in a mixture of depths, and is enriched mainly for membrane related terms. This six module decomposition offers some additional insight, separating out eukaryotic function to a separate module, while still separating the depth mostly as expected.

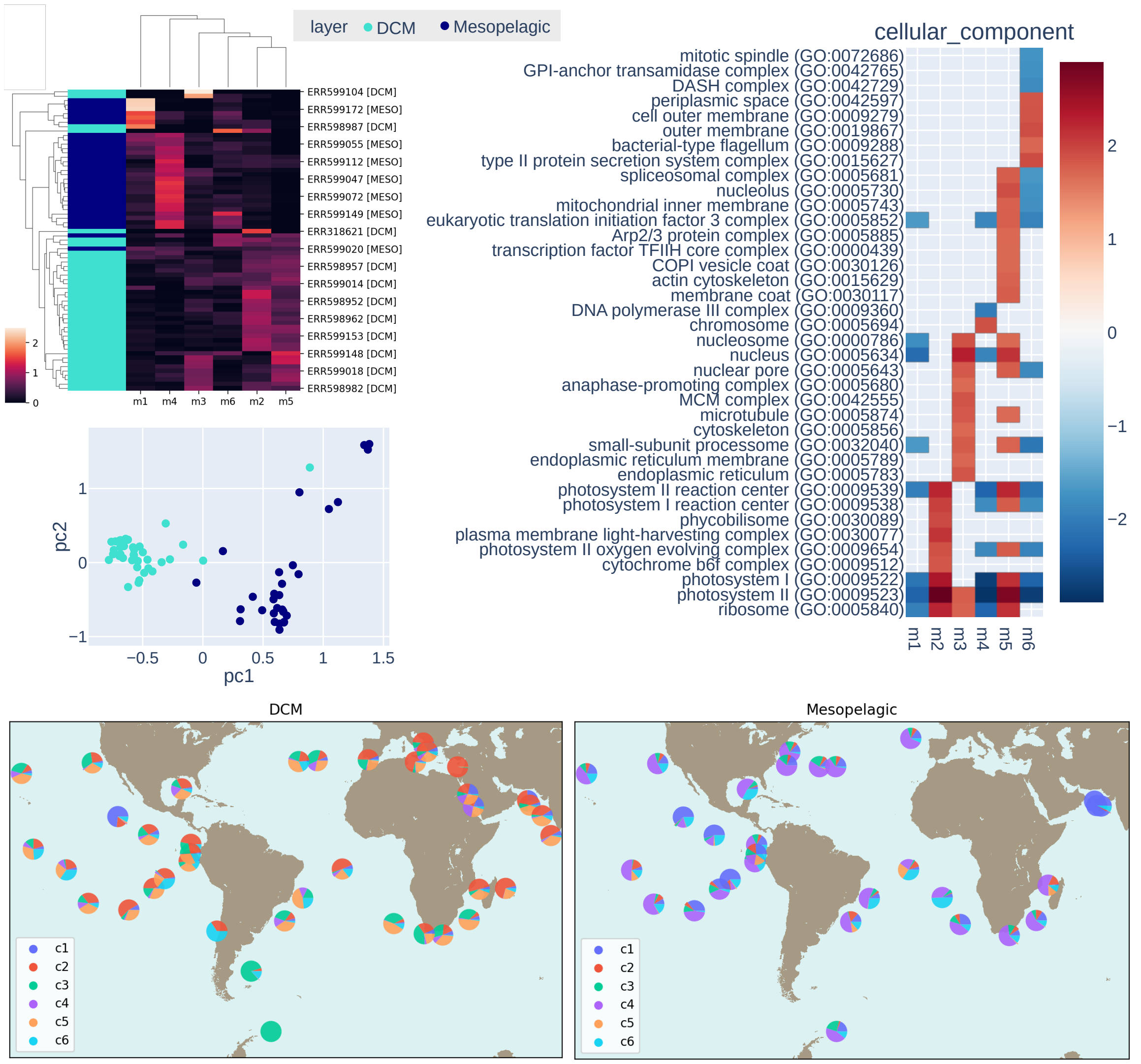


Fig. 5.26 Results of decomposition of Tara Oceans data from DCM and Mesopelagic depths [3, 4]. Top left: W matrix and coloured ribbon indicating the depth the sample originates from, Top right: GSEA enrichment analysis of modules, showing m3 enriched for eukaryotic function, m2 and m5 for photosynthesis, and m6 for membrane related terms Middle left: PCA ordination of W, with colour indicating depth of sample. Bottom: Maps showing proportion of each module in each station, separated by depth, with left being DCM and right mesopelagic.

Having demonstrated recovery of a meaningful separation between samples expected to show clear functional difference, we applied the same methods to analyse only surface ocean samples from the same dataset. As previously, where duplicate samples existed we discarded one sample at random. Three samples which had been sequenced using pyrosequencing rather than Illumina sequencing were also removed. Only samples labelled as surface water were included, all of which were taken at a depth of 5 m. The two southernmost stations were excluded as outliers, as based on the two depth decomposition and preliminary analyses these two stations were highly functionally distinct, causing problems identifying an appropriate rank for the decomposition. This resulted in a total of 58 samples, originating from 22 Longhurst provinces, with the largest number from the South Pacific subtropical gyre (11 samples). Each sample has been labelled with the Longhurst province it originates from in addition to the sample identifier. Relative abundances of InterPro entries were used as input data.

Rank selection was performed for $k = 2..20$, using the multiplicative update solver and KL divergence, with 100 iterations for each value of k, with results shown in Figure 5.27. Both consensus based methods show peaks at $k = 3$ and $k = 9$, though dispersion continues to climb after this point. The concordance index also shows a peak at $k = 3$ and after this as $k = 6$ and $k = 9$. Permutation similarly shows a peak nearby at $k = 8$ and remaining high for $k = 9$. The split-half selection method appears less clear, with similarly high values between $k = 8$ and $k = 18$.

Looking at the modules for $k = 9$, some geographic patterns are apparent in Figure 5.28. Module m3 has high weight throughout the North Atlantic; m8 taking a high weight in the South Atlantic; and m4 similarly high through the South Pacific and Indian Ocean. Not all modules have such geographic grouping: module m2 has high weight in two Southern samples which are closer to land, off the western coasts of Chile and South Africa; and module m9 has high weight in both the Mediterranean and more southern samples from the Pacific.

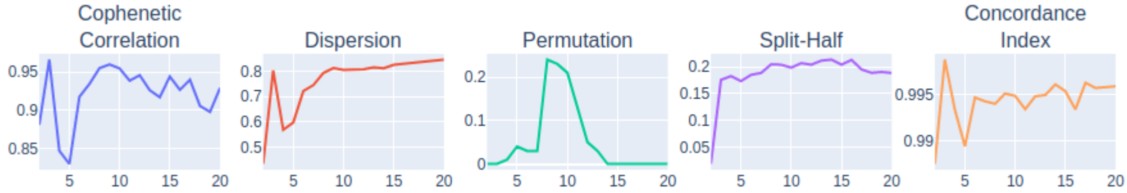


Fig. 5.27 Rank selection for Tara Oceans surface data [3, 4]

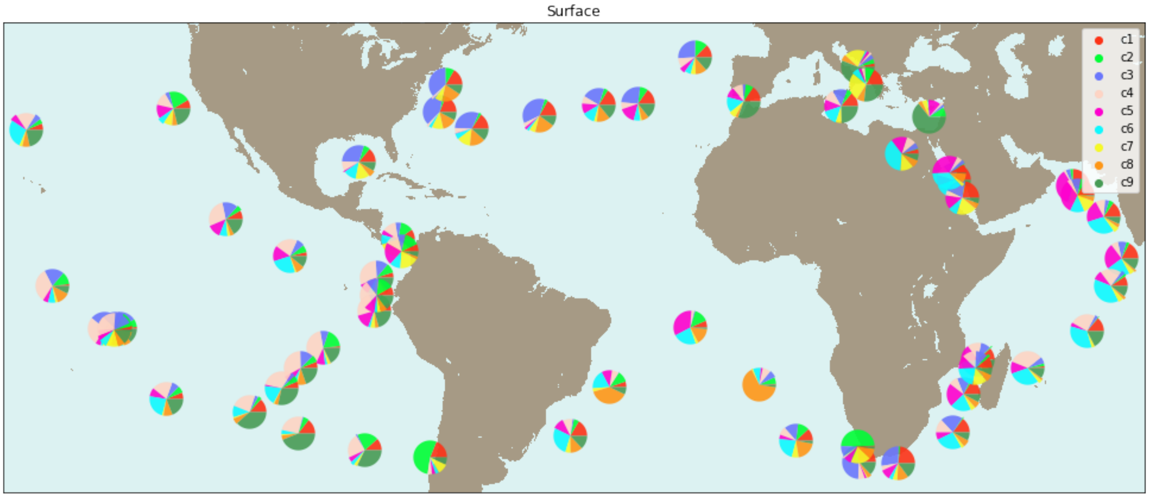


Fig. 5.28 Map of module weights for Tara Oceans surface modules for $k = 9$ [3, 4]

GSEA enrichment of these nine modules is shown in Figure 5.29 for the biological process namespace of the Gene Ontology (GO). Three of the modules (m3, m4, and m5) are enriched for photosynthesis, and this the most widely shared enriched term, other terms being enriched in at most 2 modules. Module m2 which has high weight in two samples closer to coasts has only a single term enriched for transpositional recombination which occurs via a DNA intermediate.

Weights of model components can be related to environmental measurements such as temperature, or measurements of activity such as chlorophyll-a concentration. Correlation of the weight of each component to in-situ measurements are shown in Figure 5.30. Modules m4, m5, and m6 each have a positive correlation with temperature individually; however the sum of these three modules is strongly correlated to temperature, suggesting that heat response could be a mixture of these modules functions (Figure 5.31). Of these modules positively correlated with temperature, only m4 is also positively correlated with chlorophyll-a concentration. Module m3 is also positively correlated with chlorophyll concentration, but has a negative correlation with temperature. Again, the sum of modules m3 and m4 is more strongly correlated than either individually, allowing a separation of functions associated with primary production into cold (m3) and warm (m4) modules. Correlation of these module combinations are shown in Figure 5.31

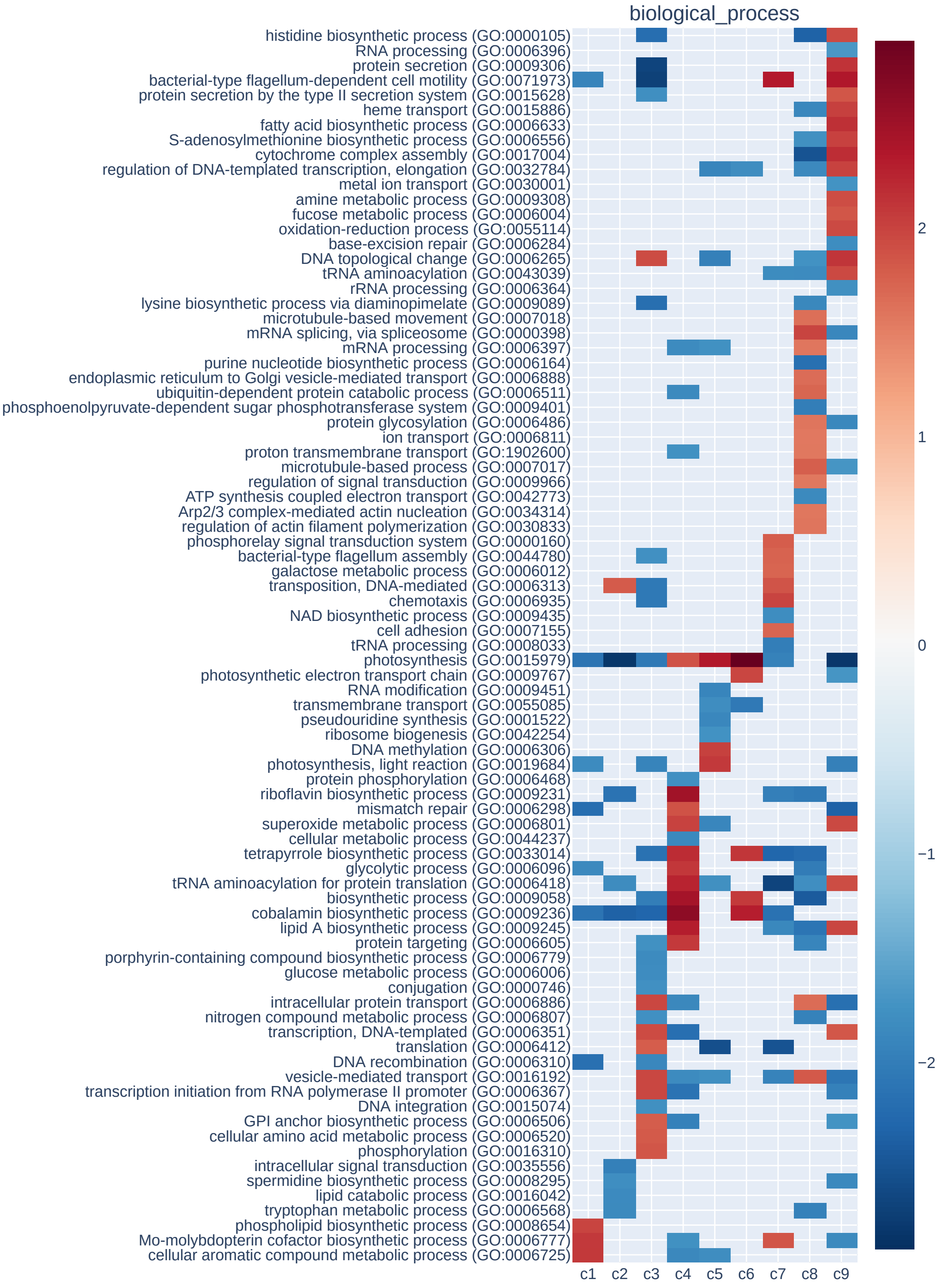


Fig. 5.29 Enrichment of GO terms in modules for Tara Ocean surface decomposition with $k = 9$ [3, 4]. Limited to the GO biological process namespace.

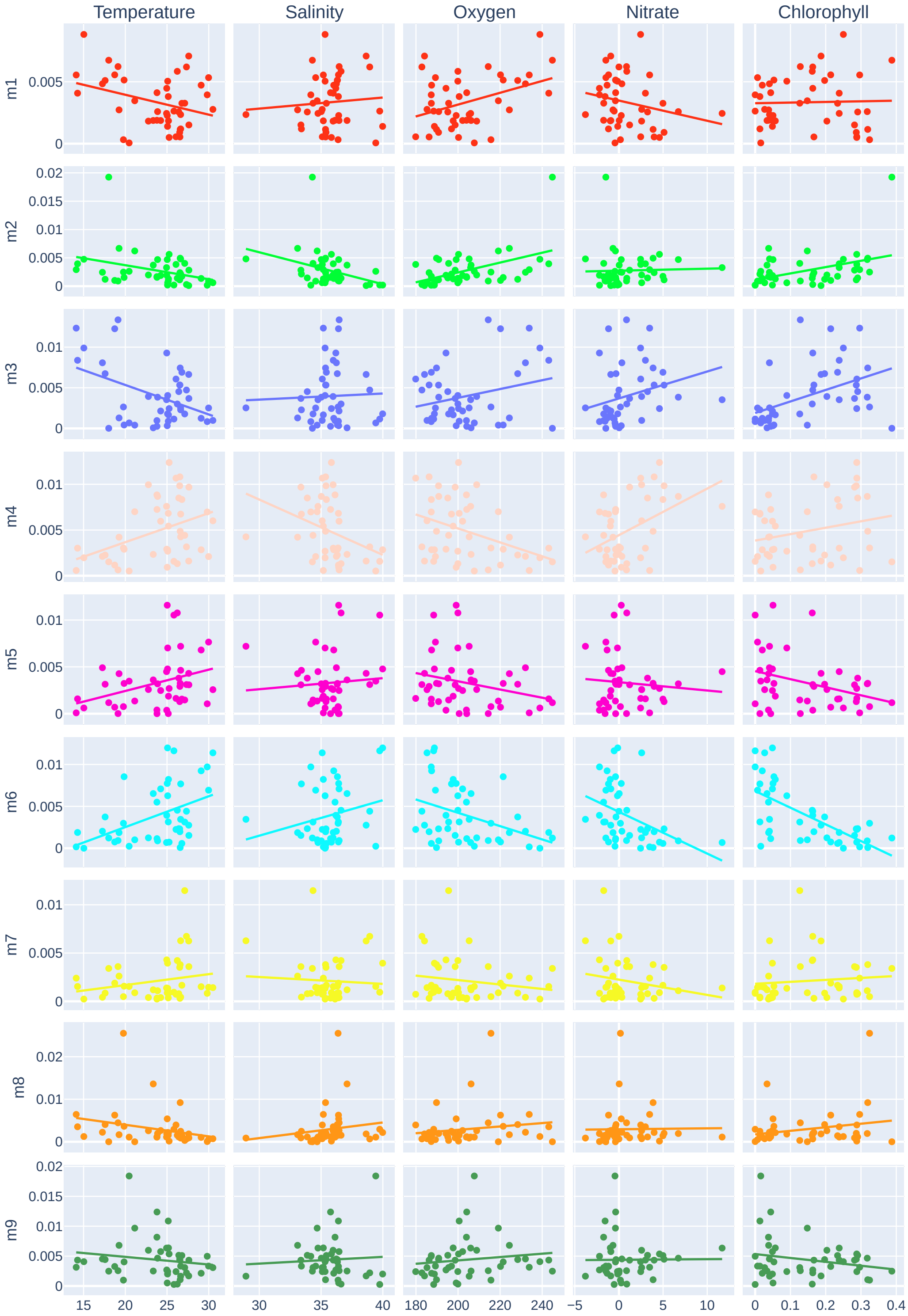


Fig. 5.30 Correlation of Tara Ocean surface decomposition and measurements of environmental conditions taken in-situ. Line of best fit based on ordinary least square regression.

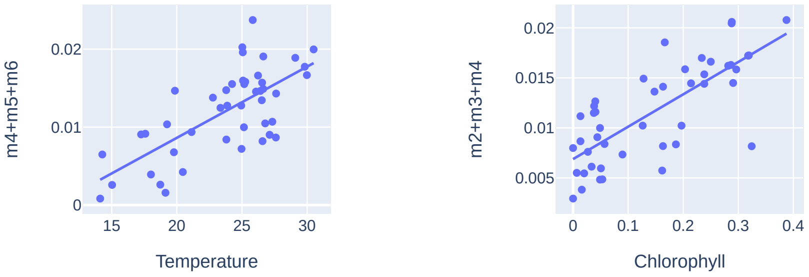


Fig. 5.31 Correlation of combinations of Tara Ocean surface modules and environmental conditions taken in situ. Line of best fit based on ordinary least square regression.

The weights of modules can also provide useful input for predictive models. We performed a multiple linear regression using the ordinary least squares method with chlorophyll *a* concentration as the dependent variable and module weights as explanatory variables, with a resulting $R^2 = 0.744$. Chlorophyll *a* concentration is often used an estimate of the activity of a microbial community,specifically the primary production being carried out, as chlorophyll *a* is a distinctively pigmented component of the photosynthetic machinery. As a simple measure of how well this model would fit to new data, we used the duplicate samples which had not been included as part of the decomposition learning. Module weights were determined from the InterPro abundances of the duplicate samples, and then chlorophyll *a* concentration predicted using these module weights and the linear regression, with results shown by red points in Figure 5.32. Chlorophyll *a* concentration was predicted well for all duplicate stations except one, which had an in-situ measurement far outside the range of those included in the decomposition learning, measuring 1.55 where the maximum among stations included was 0.39. These duplicate samples collected simultaneously with those from which the decomposition was learnt are not a robust test set, but it shows that the reduced dimensions of NMF could be further explored as inputs for predictive methods connecting function more directly to microbial activity.

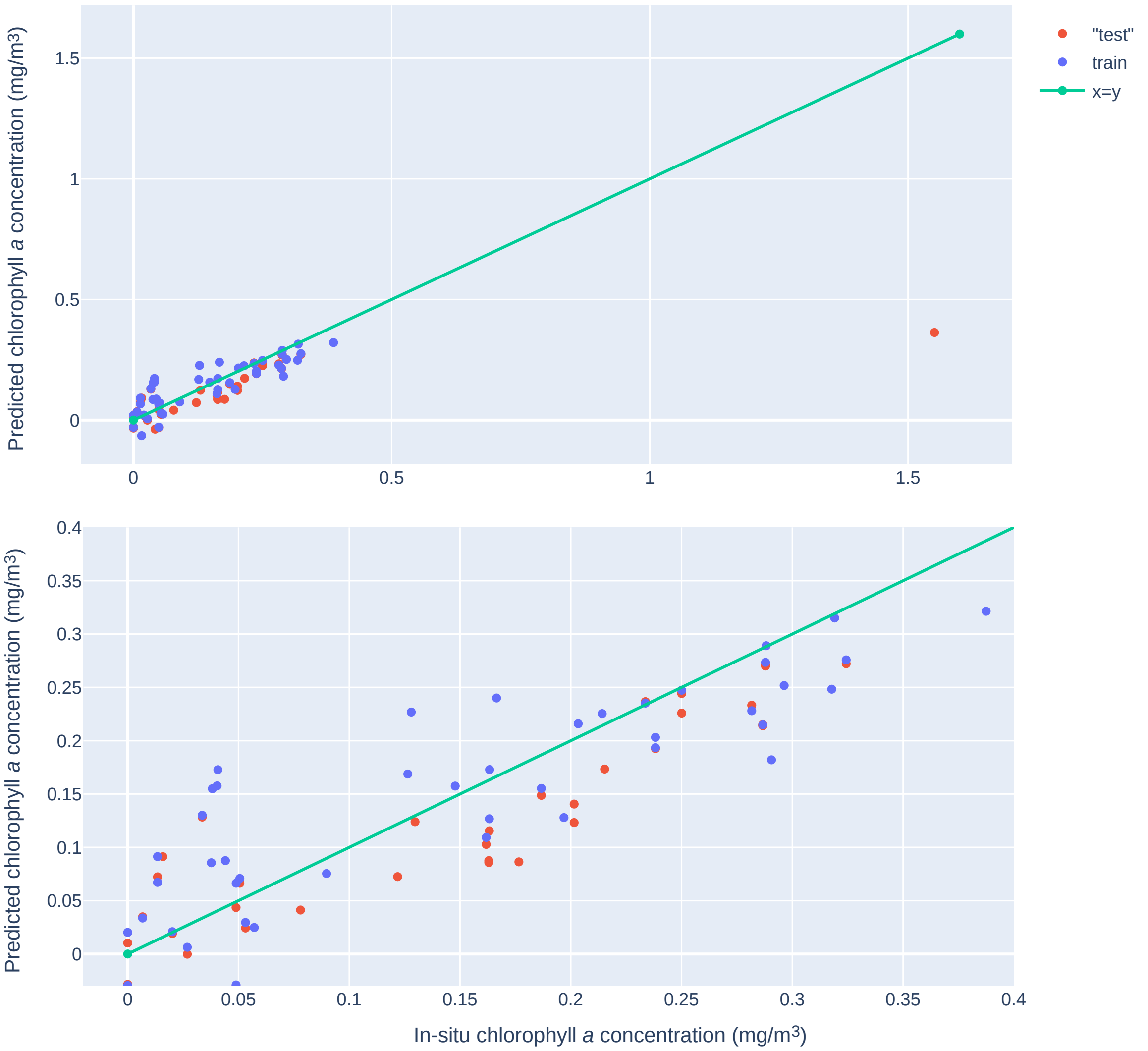


Fig. 5.32 Chlorophyll *a* concentration (mg/m^3) predicted by linear regression is shown on the vertical axis, with in-situ observation of chlorophyll *a* concentration on the horizontal axis. An ordinary least square regression was carried out using module weights from NMF decomposition of the Tara Oceans surface data [3, 4] as predictor variables. Blue points are samples included in learning the decomposition; red are duplicate samples which were not included in decomposition learning. The top plot includes one duplicate which had an in-situ observation far outside the range included in the training samples; bottom shows the same plot with this point excluded.

5.6 Discussion

5.6.1 Rank Selection

Selecting an appropriate rank is an important task in NMF decomposition for which a variety of methods have been proposed. Several methods were compared in a previous study [372] however this analysis did not include a method proposed to be well suited to data with overlapping latent modules, the concordance index [369]. Additionally, this review evaluated performance on synthetic datasets with either discrete latent modules, or datasets generated from randomly filled W and H matrices. We started from the assumption that the underlying structure in meta-omic data would have a structure of overlapping blocks, which is not covered by these two synthetic datasets. Additionally to our best knowledge, an evaluation of the concordance index has not been previously performed for any data, and the other methods have not been evaluated on simulated sequencing data. Our evaluation found that on synthetic data many of the methods tested had peaks or elbow points one above or below of the true rank, however in datasets with more overlap and noise the concordance index retained this signal more clearly than the other methods tested. The permutation method performed well for data with low noise, but more poorly in data with more overlap and noise, terminating the search at too small a rank. Rank selection criteria thus approximately locate the true rank of the data, with the concordance performing best across data with different properties. In application to real world data, decompositions either side of the suggested rank identified by the rank selection criteria should be inspected to identify the most suitable value. Further visualisation of matrices can identify whether added modules are informative, such as refining or splitting a module in comparison to a lower rank, or whether uninformative, such as low weight across all samples or weight for only a single sample, and should form a vital step in identifying a suitable decomposition rank.

To support these results in synthetic data, we simulated metagenomic sequencing of two different communities and evaluated the model selection performance on this data which is closer to a natural community. Between these two simulations we observed that while the maximum value of the concordance index occurs before the true rank, in both cases it showed a peak at the true rank. No other method had a peak at the correct rank in both simulations. Based on these results, for application to real world data in Section 5.5 we suggest a strategy of seeking agreement between the concordance index and any of the other methods when selecting an appropriate rank, or if only using one criteria to use the concordance index. We demonstrated in these case studies that this rank selection approach identified ranks corresponding to known meaningful groupings in three different real world scenarios. Additionally, exploring other ranks with high values provided informative refinement the decompositions in the Tara Oceans case study (Section 5.5.3) separating out the contribution of eukaryote specific function in a subset of surface samples.

5.6.2 Feature Importance

Given an NMF decomposition, it is desirable to be able to know which features are important to each module. The weight of a feature may be a poor indication of this however, as an abundant but uninformative feature may have higher weight than a rare interesting feature. Similarly we illustrated issues with correlation between module weights over samples and feature abundances in input data as a way of assessing feature importance. Where features are shared between latent modules they were shown to have lower correlation even in data where within a latent module the features have perfect correlation (Figure 5.2). We introduced two new ways to assess the importance of features, aiming to improve identification of shared features: permutation based and LOOCD. All methods assigned lower values to shared and ubiquitous features, however LOOCD showed the lowest crossover between values for features not in a module and those which are (Figure 5.14).

In real world applications, it is necessary to know what value for each of these importance measurements distinguishes relevant features. Testing the three methods we developed (a greedy algorithm, a KDE method, and a simple threshold) in combination with the three feature importance measures we found that LOOCD performed best in identifying which features belong to a module if a correct threshold can be identified. However the KDE method failed to identify suitable thresholds, and the greedy assignment (which does not work by selecting a threshold) performed worse than the KDE method. We demonstrated that a threshold of value of -0.05 for LOOCD gave close to the maximum recovery and relevance scores across multiple synthetic datasets with a range of latent ranks, providing a stable default threshold. Together this provides a new method of identifying which features to consider important when interpreting the features of an NMF decomposition which is more robust in data with high feature sharing. This was supported by analysis of the simulated data, in which recovery of the underlying functional modules, including shared and ubiquitous function, was demonstrated using LOOCD and the default threshold.

We began from the assumption that in the ocean and other natural microbial environments, many functions would be shared across latent modules; the LOOCD method we developed showed the best performance in recovering modules in synthetic and simulated data with such overlapping. In combination with a threshold of -0.05 beyond which to consider a feature as important to a module, we suggest this method is well suited to interpret features in decompositions of real world data metagenomic data.

5.6.3 Summary

In this chapter we have developed and illustrated techniques for applying NMF to metagenome data from microbial communities, showing it's applicability both in computationally simulated data and real world data from multiple environments. While the NMF decomposition method itself is not new, its application in metagenomics has been infrequently explored. We have developed approaches suited to data where features, in our case functions of microbes, will be widely shared between the latent modules we seek to recover. To our best knowledge, the concordance index [369] had not been evaluated, either in isolation or in comparative studies. Further, comparative studies had not used data with overlapping module structures. We showed that the concordance index to be the best performing rank selection method in our experiments, and suggest this an appropriate tool for identifying suitable ranks of NMF decomposition. The feature weight matrix H of NMF decompositions remain high dimensional, and biological interpretation requires methods which can identify features important to each module. We developed the new LOOCD method, which better identifies which features which are important to a module including those which are shared or ubiquitous, in contrast to other methods such as correlation which less well capture shared features.

In three real world case studies we showed that these methods represents a promising technique for exploratory analysis of the growing volume of environmental metagenomic data. We were able to show in Section 5.5.1 the identification of functional modules corresponding to communities with established differences in HMP data, where samples from different locations on the body are shown to be functionally distinct. We further illustrated that within oral samples, where functional distance between sample location is less distinct, the methods identified modules relating to the locations within the mouth and functional terms enriched within those modules (Figure 5.21). Similar applicability was shown in data from a study of the Waiwera river estuary [1], and we showed interpretive benefits in comparison to the discrete WGCNA methods of identifying modules. Analysis of EBI annotation of the Tara Oceans data in Section 5.5.1 showed that our methods can handle large scale, ocean metagenome data analysis, describing the functional modules characterising the DCM and mesopelagic ocean. We present an initial analysis of surface ocean samples from the same data, and show it has promise as a method for relating function and environmental conditions through correlations of module weights and in-situ measurements, and through regression analysis of primary productivity measured via chlorophyll concentration. More work is left to do in interpreting this surface ocean analysis in collaboration with biologists, but we have demonstrated the potential of our methods for understanding ocean metagenome data.

A recent study used NMF for a similar purpose, analysing Arctic metagenomic and metatranscriptomic data [405]. Identifying $k = 4$ as a suitable rank for both metagenomic and -transcriptomic data, the four sub-metagenomes (or modules in the terminology used in this thesis) were shown to correspond to a vertical region from surface to deep waters, and enzymes related to aromatic compound degradation found to be highly specific to humic-rich fluorescence DOM maximum samples. Rank selection was performed based on cophenetic correlation and dispersion, two methods as discussed in Section 5.3.2 are based on assigning each sample to a single class, and evaluating stability of clustering between random initialisations. Where strong separation is expected this may be a suitable method, and the Arctic Ocean may fit this assumption, displaying strong vertical stratification. While this is clear in the metagenomic data, with visualisations of the consensus matrix showing clear stable groups, this is less apparent in metatranscriptomic data. For the metatranscriptomic profiles, cophenetic correlation and dispersion indicate different suitable ranks, with dispersion taking it's lowest value at $k = 4$, and visualisation of the consensus matrix showing less consistent clustering than in the metagenomic data, suggesting an alternative rank selection such as the concordance index method may be applicable to this data. Biological interpretation of the decomposition used an approach explicitly inspired by specificity [369], and showed interesting insight into which functions were unique to a particular module. Visualisations of the function matrix suggest patterns of features shared between two or three modules, so there may be additional biological insight possible through methods such as those explored here which better capture these shared features. This application of NMF to Arctic data demonstrates both that there is a desire for methods which produce interpretable models of ocean microbial function which can be served by NMF, and that techniques which are suited to overlapping underlying structures may be beneficial when strong separation is not observed.

Chapter 6

Discussion and Future Work

6.1 Summary

In this thesis we have developed two methods for analysing meta-omic sequencing of environmental samples, with a focus on marine microbes. Firstly we developed a pipeline to generate MAGs for eukaryotic microbes from metagenomic sequencing of marine microbial communities, providing draft genomes for uncultured members of these natural communities. Secondly, we developed methods based on NMF decomposition for describing the distribution of functions across the ocean which permits functions to be shared among modules, and demonstrated the applicability of these methods in simulated and real world data.

More specifically in Chapter 4 we described the methods used to recover eukaryotic MAGs from 12 sets of metagenomic reads from the Atlantic and Arctic oceans, generating in total 21 eukaryotic MAGs. The methods we used were similar to those employed for prokaryotic binning, but incorporating a step separating out eukaryotic contigs using EukRep [235]. Additionally, we showed that use of psuedo-alignment tools such as Kallisto [209] can provide a sufficient coverage estimate for metagenomic binning tools, with reduced computational cost compared to short read alignment tools. We also presented analysis of both these eukaryotes and the 122 prokaryotes, showing their quality, taxonomy, distribution, function, and association between MAGs from the two kingdoms.

These analyses showed a clear distinction between the MAGs recovered from polar and non-polar samples, with no eukaryotes crossing the Arctic circle, supporting breakpoints identified in 16S/18S and metatranscriptome beta diversity at approximately 9.5 °C and 13 °C respectively in samples from the same expeditions [288]. Functional annotation of these MAGs showed a greater number of unique functions in polar eukaryotes, suggesting that a dynamic surface ocean with seasonal mixing and sea-ice formation requires these genomes to diversify. An associated eukaryote and prokaryote pair we identified were enriched for membrane processes related to transport, suggesting exchange between the two organisms, fitting with a mutualistic relationship.

In Chapter 5 we described methods of applying NMF to environmental meta-omics data. While this method itself is not new, having a history of applications in other domains such as computer vision and document analysis, its application to environmental meta-omics data has been limited; the last application to ocean data we are aware of was to Global Oceans Survey (GOS) data. A key problem in decomposition methods is selection of an appropriate rank; we evaluated the performance of several rank selection methods on both synthetic and simulated sequencing data with features which appear in multiple modules. Our evaluation identified the Concordance Index, which had not been included in previous comparisons, as the most consistent of the measures tested. Having obtained a decomposition of appropriate rank, we developed new methods of evaluating how relevant each feature is to a module in a context of shared features, and from these to classify whether a feature is in a module or not. Our LOOCD measure of feature importance showed improved performance over the other two tested (correlation and permutation based), particularly for shared features. Alongside this we identified a suitable value to use as a cut-off for classification of features using LOOCD, which appeared stable across ranks in synthetic data and performing well in simulated sequencing data. This chapter concluded with three cases studies, showing application of these techniques to meta-omics data of increasing complexity.

6.2 Future Work

6.2.1 Pangenomic Analysis of Micromonas MAGs

Eukaryotic MAGs for ocean microbes are now being generated on a large scale [2, 238, 23], and recovering closely related genomes. Our results in Chapter 4 recovered 5 Micromonas MAGs, with 20 and 26 in two binning analyses of Tara Oceans data [152, 238]. We found a high similarity (>98% ANI) between our Arctic Micromonas MAG P2_1E and a MAG previously recovered from the Antarctic *Micromonas* sp. ASP10-01a [19].

Species of picoeukaryotes in the genus Micromonas span a very wide latitudinal and thermal range, including the polar adapted *Micromonas polaris*, for which no complete genome is yet available [85]. The genus appears to divide into thermotypes, with temperature determining their distribution [86]. Changing ocean conditions in the Arctic such as warming and ocean acidification have been predicted to increase the role of Micromonas in this region, and experiments showed that Micromonas are capable of adapting to shifting thermal conditions [103].

The growing volumes of MAGs provide the potential to explore the existing functional diversity among Micromonas beyond those cultured strains. MAGs are incomplete, posing difficulties for pangenomic analyses: an absent gene could be a false negative, where it is absent due to the MAG's incompleteness, or truly absent. Tools and methods taking this into consideration have been developed and applied for prokaryotic MAGs [263], where large numbers of closely related MAGs have been available for a longer time. Extending these approaches to Micromonas, and hence eukaryotes, will help reveal environment specific traits among a genus with global importance.

6.2.2 Superkingdom Prediction of Metagenomic Reads

Current PhD student William Boulton and colleagues at JGI have begun generating eukaryotic MAGs from the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) metagenomics data (Section 6.2.4). This is a much larger set of data than our 12 samples, making techniques for reducing the size of data at steps valuable. Part of the process they have adopted is to identify potentially eukaryotic reads, which can then be assembled separately from the prokaryotic reads. This can be achieved using reference based methods, however for classification of reads this can be time or memory intensive. A tool like Kraken [218] intended for taxonomic classification of reads is rapid but with high memory requirement, and aims for a level of taxonomic resolution not required for separation at the superkingdom level. For assembled contigs, tools such as EukRep and Tiara [235, 265] provide rapid classification into a small number of classes (eukaryotic, prokaryote, plastid) through machine learning techniques (support vector machines and neural networks), but require a minimum length of contig for classification. Taking a similar machine learning approach to train a model for superkingdom level classification of metagenomic reads has the potential to provide a computationally cheaper way to filter eukaryotic metagenomic reads prior to assembly, reducing the volume of data to be handled at an early point in eukaryotic binning efforts.

6.2.3 NMF Modules as a Feature Extraction Method

The case study of the surface Tara Oceans data in Section 5.5.3 showed that a linear regression model fitted to the module weights explained approximately 74% of variation among in-situ chlorophyll-a concentration. We propose that the explanatory power of this reduced dimension representation could be used as a feature extraction method when applying other machine learning techniques to meta-omic data. Feature extraction methods seek to construct a reduced number of features derived from original high dimensional data, seeking to provide a lower dimension input with reduced redundancy for a subsequent machine learning technique. Ordination approaches such as PCA have been commonly employed as feature extraction methods [406].

Deep learning methods such as Convolutional Neural Networks (CNNs) have led to significant advances in applications where the number of samples is much greater than the number of features, including in Earth systems science [407]. However while ocean metagenome sampling and sequencing is expanding rapidly, the number of features (taxa, genes, functions) seems set to remain greater than the number of samples for the near future. Extracting a lower dimensional representation of high dimensional structures such as metagenome taxonomy has shown improved performance in metagenome classification tasks [408], and NMF has been used for this purpose in classification of images of medical diagnosis [409]. NMF modules could provide value as an interpretable feature extraction method for deep learning regression models aiming to relate metagenomic profiles to measures of microbial activity (e.g. chlorophyll-a, CO_2 flux).

6.2.4 Meta-omics Informed Earth Systems Modelling in the Central Arctic

Metagenomic and metatranscriptomic data from large scale ocean expeditions such as the Global Ocean Survey [55], Tara Oceans [4] and Sea of Change [312] have provided insight into the traits of ocean communities which underlie transformation of matter and energy in their environments. This trait information has been used to develop *in silico* models linking microbial activity and ocean biogeochemical cycles, allowing predictions to be made under different conditions of warming [250, 312]. However the smaller number of studies of polar microbiomes, including ours in Chapter 4, have shown distinct differences between polar microbes and function and their non-polar counterparts [23, 288]. Existing models are therefore difficult to apply to polar environments, as they do not reflect the observed evolutionary novelty and associated traits of polar organisms.

Recently, the Multidisciplinary drifting Observatory for the Study of Arctic Climate (MOSAiC) expedition has radically expanded the amount of data available for the central Arctic Ocean [56]. This year-round expedition used ‘RV Polarstern’ as a drifting research platform frozen into the ice and was completed in October 2020, linking observations across the climate of the highly inaccessible central Arctic Ocean with an estimated 10 Tbp of sequence data from marine microbes. Microbial communities from both ice and water were sampled at multiple depths. From this expedition, we have access to $\geq$400 genomic and transcriptomic samples each from across the Arctic Ocean with linked physical (e.g.

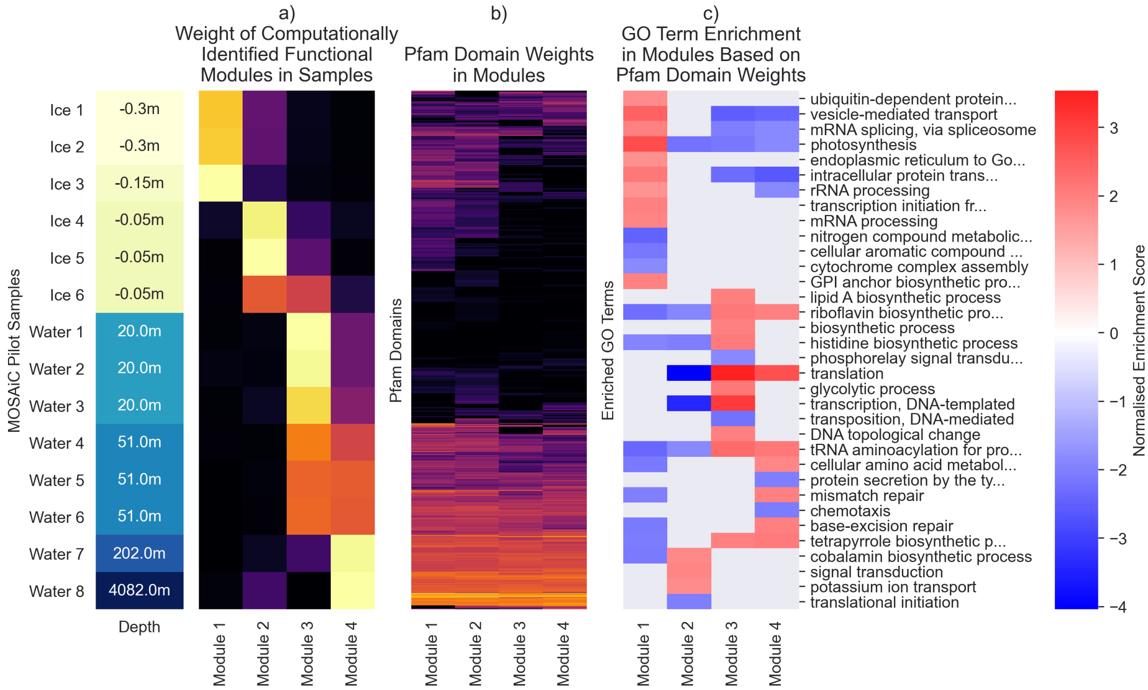


Fig. 6.1 Application of NMF methods to MOSAiC pilot sequencing [56]. On the left is an indication of depths from which samples were collected. a) shows the W matrix giving the weight of modules for each sample. A clear separation is visible between the two depths of ice. Water samples are similarly represented by two modules, one shallower one deeper, with the samples at 51 metres being a mixture of the two. b) shows the H feature matrix, illustrating a mix of widely shared and unique features in modules. c) is GSEA enrichment of GO terms based on the H matrix, showing a high level view of the processes enriched or depleted in each module, such a photosynthesis being enriched in the upper ice samples.

temperature, salinity, currents) and biogeochemical measurements (e.g., nutrients, carbon export). This linked meta-omic and physical data over a whole year provides a unique opportunity for developing a cell model tailored to polar microbes of the Arctic Ocean, allowing simulation of how warming may affect the Arctic ecosystem including its food web and associated biogeochemical cycles. Our preliminary analysis of metagenomic sequencing of 14 pilot samples shows that the NMF approach discussed in Chapter 5 can identify modules associated with different sample depths from ice to the bathypelagic ocean, and the traits within these modules (Figure 6.1)

In collaboration with Professor Tim Lenton at the University of Exeter, we have written a Leverhulme Trust proposal to develop a novel polar cell model, building on the EVolutionary Ecosystem (EVE) cell model that links omics-informed traits with biogeochemical cycles. This model was informed by a transcriptomic-derived relationship between temperature and the biosynthesis rate of proteins across major latitudinal zones of the oceans [410, 312].

Based on the elemental composition of ribosomes, the EVE model reproduced phytoplankton growth strategies, cell size and N:P stoichiometry based on a representation of fundamental cellular and biophysical constraints. However, the existing EVE model does not represent polar-specific traits and their plasticity including their ability to evolve under conditions of selection. Furthermore, it only represents photoautotrophic microbes. Consequently, the underpinning cell model will need to be modified to reflect the strategies and mechanisms (e.g. storage strategies of eukaryotic algae, mixotrophy, heterotrophy) of a wide range of organisms (bacteria, archaea, protists) in divergent conditions (in ice, protracted darkness) as they were encountered during the MOSAiC expedition.

6.3 Outlook

Advances in metagenomic and metatranscriptomic sequencing have allowed access to the genetic material of natural microbial communities, including their unculturable majority. A wide range of computational methods have been developed to generate insight into the taxonomy, function and activity represented in this genetic data. Some steps have started to become standard in metagenome analysis, such as assembly, and taxonomic and functional annotation. The output of these remains complex however, analysis techniques are required to help interpret the information within and between samples.

Obtaining genomes for the unculturable majority of organisms is a problem being tackled using multiple approaches. Some of these address issues which are presented by recovering MAGs from short read sequencing. Single-cell sequencing reduces the risk of contamination and chimeric sequences within the genomes recovered [58]. Third generation long read sequencing technologies can produce long sequences of a genome without need for assembly and binning (Section 2.6.3). Projects such as 100 Diatom Genomes [411] are focussed on using new and existing techniques to expand the diversity of organisms for which we have reference genomes. Do sometimes fragmented and highly incomplete MAGs have a role in light of these developing techniques? There exists a huge amount of second generation sequencing data from environmental samples, and automated metagenome binning is a comparatively inexpensive method through which genomic information can be obtained from this already existing data. There is reasonable caution about the inclusion of MAGs into reference databases [412, 413], and in light of this manual curation and validation of MAGs is a time consuming but valuable step. Of our eukaryotic MAGs, only one has been added to JGI's PhycoCosm genomes as *Micromonas* sp. AD1 and labelled as a metagenome extracted assembly; this was the most complete genome we recovered, with clear taxonomic identification of a majority of contigs; the genes predicted were also assessed independently by JGI and were highly similar to our initially gene models. Many of the lower completeness MAGs with less confirmatory evidence for lineage would be unsuitable for inclusion in such a resource, but there is a role for high quality MAGs in contributing to these public database.

Recovery of eukaryotic MAGs is a significant advance in metagenomics, as eukaryotic plankton play vital roles in ocean processes. However the proportion of the community recovered in the current set of studies remains low, with 8.1% of reads in our data mapping back to the eukaryotic MAGs and 11.8% of the total reads in Delmont et al. [2], limiting the extent to which metagenomic binning can describe the overall community. Coverage has been cited as a key limitation in eukaryotic MAGs recovery [235], and increasing sequencing depth to obtain sufficient coverage of rarer taxa could prove cost-prohibitive. Computational tools to simplify the difficult of individual steps of the binning process, such as we suggested for assembly in Section 6.2.2 may play role in advancing the proportion of the community we can describe with these methods. Recovery of prokaryotic MAGs has become a common part of metagenome analysis and integrated as a step in some pipelines such as IMG [291]. With the importance of eukaryotic plankton in the oceans, it seems likely that eukaryotic MAGs will also become a common step in analysis of ocean metagenome data. This will however require settling on some tools and databases as standard, to make results more easily comparable between studies; EukCC [255] seems to have been adopted for quality estimation, but methods of taxonomic identification and estimating MAG abundance vary between studies currently.

Methods for combining analysis of species, trait and environmental data such as joint species distribution models have been successfully developed and applied in ecology [414]; MAGs represent a method of providing a species-trait link which can allow such analyses in metagenomic data. Genetic data tends to described the presence or absence of thousands of genes or functions rather than a smaller number of traits, so ways to infer traits from genomes will be valuable in this application. Some studies have already done this, such as predicting trophic mode from genomic content of a MAG [238] or growth rate from MAG codon usage bias [415], and it may be possible to infer broad environmentally linked traits like freezing resistance from genes involved such as those coding for ice-binding proteins. While characterising the pangenome is difficult using MAGs, the increasing volume of eukaryotic MAGs has started to permit to a limited extent the analysis of population genomics for those MAGs sharing a lineage, such as as identifying population structure and which genes are under selection in Arctic Chaetoceros [416], and the recovery of further MAGs will allow this approach to be applied more broadly.

Analysis of functional metagenomic data presents its own unique challenges, but looking to methods applied in other dissimilar fields can suggests ways to overcome these. We applied NMF, a method more commonly used for parts-based analysis in fields such as document analysis and computer vision to obtain a similar description of metagenomic data; as mentioned above joint species distribution models could provide a powerful way to analyse MAG data. Looking to a related field, a recent model utilising eDNA amplicon sequencing estimated changes in organisms abundances, accounting for covariates and error in sampling and sequencing [417]. This could be adapted to metagenomics to offer a way of understanding the seasonal dynamics in full-year metagenome studies such as MOSAiCs's. In fields other than metagenomics this cross-domain influence has been impactful: in artificial intelligence research, game-playing agents have been incorporated in searching for improved matrix multiplication algorithms [418]. Established methods such as CCA or correlation based network analysis and clustering remain useful and have proven to be powerful tools, as well as having the benefit of familiarity and a wide range of well supported implementations. However with the growing volume of functional data and MAGs, novel or adapted tools will need to be introduced to best harness the potential of these data. We propose that NMF is one such tool among others such as statistical network models or joint species distribution models.

Application of NMF in this thesis focussed largely on metagenomic data, describing the functional potential of the whole community. Some of this overall functional repertoire may not represent metabolism which is active under the conditions during sampling. Organisms may be present but dormant, with some studies suggesting widespread seed banks of dormant organisms which may thrive given shifts in conditions [419], or more local processes such as a small number of cells surviving through unfavourable seasonal dynamics [420]. Some evidence supports functional potential being stable across environmental conditions in comparison to taxonomic composition [333], and looking at the abundance of functions rather than presence/absence may avoid the potentially 'noisy' functions contributed by low abundance dormant organisms. Metatranscriptomic sequencing captures the activity of the community at the time of sampling, revealing which parts of the functional potential were being transcribed. Many functions encoded in dormant or otherwise less inactive cells will not contribute to this data, removing a source of potentially uninformative features. Saelens et al. [359] note that gene expression data tends to be characterised by local as well as global patterns, and by functions which play roles in multiple pathways. In a recent study of Arctic metagenome and metatranscriptome data using NMF [405] discussed in Section 5.6.3, the consensus matrix constructed for rank selection showed consistent clustering for metagenomic data, but was less stable in metatranscriptomic data from the same samples. Metatranscriptomic data may eliminate some of the background noise of low inactive functions, but contain a greater degree of overlapping and local patterns. As such decomposition approaches such as NMF may be well suited to analysis of metatranscriptomic data, but benefit from associated rank selection and interpretive techniques which capture the local and overlapping patterns.

Our work in Chapter 4 included only seven samples from the Arctic ocean, and from this small set of samples generated 16 medium quality MAGs. Within these MAGs we replicated the observation of Martin et al. [288] from analysis of the metatranscriptomic sequencing of samples taken during the same expeditions that there is a strong demarcation between Arctic and temperate and subtropical microbial communities, both in terms of taxa and functions. Other metagenomic analyses show that the Arctic is home to considerable genetic novelty, with a recent study of the Tara Ocean Arctic samples identifying 441 prokaryotic MAGs from novel species [421]. The Arctic ocean is among the planet's most inaccessible environments, and consequently our understanding of the microbial communities and their activity are incomplete. Human driven climate change affects the Arctic at an accelerated rate, with the possibility for tipping points triggering abrupt non-linear change in the Arctic and beyond [121, 422]. Data from the recently completed MOSAiC expedition covers the Central Arctic Ocean across a full season including the Arctic winter, complementing the Tara Ocean Arctic collected during a circumnavigation of the Arctic from May to October. Analysis of this influx of Arctic data is undoubtedly a complex undertaking given the established novelty, but will advance our understanding of the unique functioning of polar and Arctic microbes, and their interactions with broader ocean processes. In the longer term, feeding this knowledge forward into predictive climate and earth systems models as discussed in Section 6.2.4 can help in evaluating and planning around the impact of continued climate change.

References

[1] H. S. Tee, D. Waite, G. Lear, and K. M. Handley, “Microbial river-to-sea continuum: Gradients in benthic and planktonic diversity, osmoregulation and nutrient cycling,” *Microbiome*, vol. 9, p. 190, Sept. 2021.

[2] 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, p. 100123, May 2022.

[3] A. L. Mitchell, A. Almeida, M. Beracochea, M. Boland, J. Burgin, G. Cochrane, M. R. Crusoe, V. Kale, S. C. Potter, L. J. Richardson, E. Sakharova, M. Scheremetjew, A. Korobeynikov, A. Shlemov, O. Kunyavskaya, A. Lapidus, and R. D. Finn, “MGnify: The microbiome analysis resource in 2020,” *Nucleic Acids Research*, vol. 48, pp. D570–D578, Jan. 2020.

[4] 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, T. O. Coordinators, 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, and P. Bork, “Structure and function of the global ocean microbiome,” *Science*, vol. 348, p. 1261359, May 2015.

[5] P. Webb, *Introduction to Oceanography*. Roger Williams University, 2021.

[6] C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski, “Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components,” *Science*, vol. 281, pp. 237–240, July 1998.

[7] S. Petrovskii, Y. Sekerci, and E. Venturino, “Regime shifts and ecological catastrophes in a model of plankton-oxygen dynamics under the climate change,” *Journal of Theoretical Biology*, vol. 424, pp. 91–109, July 2017.

[8] L. Beaufort, I. Probert, T. de Garidel-Thoron, E. M. Bendif, D. Ruiz-Pino, N. Metzl, C. Goyet, N. Buchet, P. Coupel, M. Grelaud, B. Rost, R. E. M. Rickaby, and C. de Vargas, "Sensitivity of coccolithophores to carbonate chemistry and ocean acidification," *Nature*, vol. 476, pp. 80–83, Aug. 2011.

[9] M. Ardyna and K. R. Arrigo, "Phytoplankton dynamics in a changing Arctic Ocean," *Nature Climate Change*, vol. 10, pp. 892–903, Oct. 2020.

[10] M. Le Moal, C. Gascuel-Odoux, A. Ménesguen, Y. Souchon, C. Étrillard, A. Levain, F. Moatar, A. Pannard, P. Souchu, A. Lefebvre, and G. Pinay, "Eutrophication: A new wine in an old bottle?," *Science of The Total Environment*, vol. 651, pp. 1–11, Feb. 2019.

[11] P. C. Reid, J. M. Colebrook, J. B. L. Matthews, and J. Aiken, "The Continuous Plankton Recorder: Concepts and history, from Plankton Indicator to undulating recorders," *Progress in Oceanography*, vol. 58, pp. 117–173, Aug. 2003.

[12] M. S. Rappé and S. J. Giovannoni, "The Uncultured Microbial Majority," *Annual Review of Microbiology*, vol. 57, no. 1, pp. 369–394, 2003.

[13] J. C. Wooley, A. Godzik, and I. Friedberg, "A Primer on Metagenomics," *PLOS Computational Biology*, vol. 6, p. e1000667, Feb. 2010.

[14] J. Cohen, X. Zhang, J. Francis, T. Jung, R. Kwok, J. Overland, T. J. Ballinger, U. S. Bhatt, H. W. Chen, D. Coumou, S. Feldstein, H. Gu, D. Handorf, G. Henderson, M. Ionita, M. Kretschmer, F. Laliberte, S. Lee, H. W. Linderholm, W. Maslowski, Y. Peings, K. Pfeiffer, I. Rigor, T. Semmler, J. Stroeve, P. C. Taylor, S. Vavrus, T. Vihma, S. Wang, M. Wendisch, Y. Wu, and J. Yoon, "Divergent consensuses on Arctic amplification influence on midlatitude severe winter weather," *Nature Climate Change*, vol. 10, pp. 20–29, Jan. 2020.

[15] S. J. Sibbald and J. M. Archibald, "More protist genomes needed," *Nature Ecology & Evolution*, vol. 1, p. 0145, Apr. 2017.

[16] N. Simon, A.-L. Cras, E. Foulon, and R. Lemée, "Diversity and evolution of marine phytoplankton," *Comptes Rendus Biologies*, vol. 332, pp. 159–170, Feb. 2009.

[17] C. Yang, D. Chowdhury, Z. Zhang, W. K. Cheung, A. Lu, Z. Bian, and L. Zhang, "A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 6301–6314, Jan. 2021.

[18] B. J. Tully, R. Sachdeva, E. D. Graham, and J. F. Heidelberg, "290 metagenome-assembled genomes from the Mediterranean Sea: A resource for marine microbiology," *PeerJ*, vol. 5, p. e3558, July 2017.

[19] T. O. Delmont, C. Quince, A. Shaiber, Ö. C. Esen, S. T. Lee, M. S. Rappé, S. L. McLellan, S. Lücker, and A. M. Eren, "Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes," *Nature Microbiology*, vol. 3, p. 804, July 2018.

[20] N. Joli, A. Monier, R. Logares, and C. Lovejoy, “Seasonal patterns in Arctic prasinophytes and inferred ecology of Bathycoccus unveiled in an Arctic winter metagenome,” *The ISME Journal*, vol. 11, pp. 1372–1385, June 2017.

[21] C. Frioux, D. Singh, T. Korcsmaros, and F. Hildebrand, “From bag-of-genes to bag-of-genomes: Metabolic modelling of communities in the era of metagenome-assembled genomes,” *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1722–1734, Jan. 2020.

[22] K. Schmidt, *Thermal Adaptation of* Thalassiosira Pseudonana *Using Experimental Evolution Approaches*. PhD thesis, University of East Anglia, 2017.

[23] 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,” *Microbiome*, vol. 10, p. 67, Apr. 2022.

[24] T. Mock, K. Hodgkinson, T. Wu, V. Moulton, A. Duncan, C. van Oosterhout, and M. Pichler, “Structure and Evolution of Diatom Nuclear Genes and Genomes,” in *The Molecular Life of Diatoms* (A. Falciatore and T. Mock, eds.), pp. 111–145, Cham: Springer International Publishing, 2022.

[25] I. V. Grigoriev, R. D. Hayes, S. Calhoun, B. Kamel, A. Wang, S. Ahrendt, S. Dusheyko, R. Nikitin, S. J. Mondo, A. Salamov, I. Shabalov, and A. Kuo, “PhycoCosm, a comparative algal genomics resource,” *Nucleic Acids Research*, vol. 49, pp. D1004–D1011, Jan. 2021.

[26] N. Ye, W. Han, A. Toseland, Y. Wang, X. Fan, D. Xu, C. van Oosterhout, I. V. Grigoriev, A. Tagliabue, J. Zhang, Y. Zhang, J. Ma, H. Qiu, Y. Li, X. Zhang, and T. Mock, “The role of zinc in the adaptive evolution of polar phytoplankton,” *Nature Ecology & Evolution*, pp. 1–14, June 2022.

[27] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, M. G. Giglio, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, J. Versalovic, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-Vercoe, E. J. Alm, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, V. Bonazzi, J. Paul Brooks, G. A. Buck, C. J. Buhay, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. G. Chain, I.-M. A. Chen, L. Chen, S. Chhibba, K. Chu, D. M. Ciulla, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, M. A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. DeSantis, C. Deal, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, D. J. Dooling, S. P. Dugan, W. Michael Dunne, A. Scott Durkin, R. C. Edgar, R. L. Erlich, C. N. Farmer, R. M. Farrell, K. Faust, M. Feldgarden, V. M. Felix, S. Fisher, A. A. Fodor, L. J. Forney, L. Foster, V. Di Francesco, J. Friedman, D. C. Friedrich, C. C. Fronick, L. L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M. Y. Giovanni, J. M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, S. Kinder Haake, B. J. Haas, H. A. Hamilton, E. L. Harris, T. A. Hepburn, B. Herter, D. E. Hoffmann, M. E. Holder, C. Howarth, K. H.

Huang, S. M. Huse, J. Izard, J. K. Jansson, H. Jiang, C. Jordan, V. Joshi, J. A. Katancik, W. A. Keitel, S. T. Kelley, C. Kells, N. B. King, D. Knights, H. H. Kong, O. Koren, S. Koren, K. C. Kota, C. L. Kovar, N. C. Kyrpides, P. S. La Rosa, S. L. Lee, K. P. Lemon, N. Lennon, C. M. Lewis, L. Lewis, R. E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C.-C. Lo, C. A. Lozupone, R. Dwayne Lunsford, T. Madden, A. A. Mahurkar, P. J. Mannon, E. R. Mardis, V. M. Markowitz, K. Mavromatis, J. M. McCorrison, D. McDonald, J. McEwen, A. L. McGuire, P. McInnes, T. Mehta, K. A. Mihindukulasuriya, J. R. Miller, P. J. Minx, I. Newsham, C. Nusbaum, M. O'Laughlin, J. Orvis, I. Pagani, K. Palaniappan, S. M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K. S. Pollard, M. Pop, M. E. Priest, L. M. Proctor, X. Qin, J. Raes, J. Ravel, J. G. Reid, M. Rho, R. Rhodes, K. P. Riehle, M. C. Rivera, B. Rodriguez-Mueller, Y.-H. Rogers, M. C. Ross, C. Russ, R. K. Sanka, P. Sankar, J. Fah Sathirapongsasuti, J. A. Schloss, P. D. Schloss, T. M. Schmidt, M. Scholz, L. Schriml, A. M. Schubert, N. Segata, J. A. Segre, W. D. Shannon, R. R. Sharp, T. J. Sharpton, N. Shenoy, N. U. Sheth, G. A. Simone, I. Singh, C. S. Smillie, J. D. Sobel, D. D. Sommer, P. Spicer, G. G. Sutton, S. M. Sykes, D. G. Tabbaa, M. Thiagarajan, C. M. Tomlinson, M. Torralba, T. J. Treangen, R. M. Truty, T. A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V. Ward, W. Warren, M. A. Watson, C. Wellington, K. A. Wetterstrand, J. R. White, K. Wilczek-Boney, Y. Wu, K. M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B. P. Youmans, L. Zhang, Y. Zhou, Y. Zhu, L. Zoloth, J. D. Zucker, B. W. Birren, R. A. Gibbs, S. K. Highlander, B. A. Methé, K. E. Nelson, J. F. Petrosino, G. M. Weinstock, R. K. Wilson, O. White, and The Human Microbiome Project Consortium, "Structure, function and diversity of the healthy human microbiome," *Nature*, vol. 486, pp. 207–214, June 2012.

[28] Y. M. Bar-On, R. Phillips, and R. Milo, "The biomass distribution on Earth," *Proceedings of the National Academy of Sciences*, vol. 115, pp. 6506–6511, June 2018.

[29] A. Proshutinsky, R. Krishfield, J. M. Toole, M.-L. Timmermans, W. Williams, S. Zimmermann, M. Yamamoto-Kawai, T. W. K. Armitage, D. Dukhovskoy, E. Golubeva, G. E. Manucharyan, G. Platov, E. Watanabe, T. Kikuchi, S. Nishino, M. Itoh, S.-H. Kang, K.-H. Cho, K. Tateyama, and J. Zhao, "Analysis of the Beaufort Gyre Freshwater Content in 2003–2018," *Journal of Geophysical Research: Oceans*, vol. 124, no. 12, pp. 9658–9689, 2019.

[30] D. M. Pidwirny, "Image of the ocean currents," Aug. 2007.

[31] JR. Toggweiler and R. M. Key, "Ocean circulation: Thermohaline circulation," *Encyclopedia of Atmospheric Sciences*, vol. 4, pp. 1549–1555, 2001.

[32] Avsa, "Map of the world's "conveyor belt".," Nov. 2009.

[33] X. Irigoien, T. A. Klevjer, A. Røstad, U. Martinez, G. Boyra, J. L. Acuña, A. Bode, F. Echevarria, J. I. Gonzalez-Gordillo, S. Hernandez-Leon, S. Agusti, D. L. Aksnes, C. M. Duarte, and S. Kaartvedt, "Large mesopelagic fishes biomass and trophic efficiency in the open ocean," *Nature Communications*, vol. 5, p. 3271, Feb. 2014.

[34] J. Gjøsæter and K. Kawaguchi, "A review of the world resources of mesopelagic fish," tech. rep., Food and Agriculture Organization of the United Nations, 1980.

[35] J. A. Raven and P. G. Falkowski, “Oceanic sinks for atmospheric CO2,” *Plant, Cell & Environment*, vol. 22, no. 6, pp. 741–755, 1999.

[36] S. Pajares and R. Ramos, “Processes and Microorganisms Involved in the Marine Nitrogen Cycle: Knowledge and Gaps,” *Frontiers in Marine Science*, vol. 6, 2019.

[37] F. Xia, J.-G. Wang, T. Zhu, B. Zou, S.-K. Rhee, and Z.-X. Quan, “Ubiquity and Diversity of Complete Ammonia Oxidizers (Comammox),” *Applied and Environmental Microbiology*, vol. 84, pp. e01390–18, Nov. 2018.

[38] H. Daims, E. V. Lebedeva, P. Pjevac, P. Han, C. Herbold, M. Albertsen, N. Jehmlich, M. Palatinszky, J. Vierheilig, A. Bulaev, R. H. Kirkegaard, M. von Bergen, T. Rattei, B. Bendinger, P. H. Nielsen, and M. Wagner, “Complete nitrification by Nitrospira bacteria,” *Nature*, vol. 528, pp. 504–509, Dec. 2015.

[39] A. Paytan and K. McLaughlin, “The Oceanic Phosphorus Cycle,” *Chemical Reviews*, vol. 107, pp. 563–576, Feb. 2007.

[40] A. C. Redfield, *On the Proportions of Organic Derivatives in Sea Water and Their Relation to the Composition of Plankton*, vol. 1. University Press of Liverpool, 1934.

[41] A. Redfield, “The Biological Control of Chemical Facotrs in the Environment,” *American Scientist*, vol. 46, no. 3, pp. 230A–221, 1958.

[42] T. M. Lenton and A. J. Watson, “Redfield revisited: 1. Regulation of nitrate, phosphate, and oxygen in the ocean,” *Global Biogeochemical Cycles*, vol. 14, no. 1, pp. 225–248, 2000.

[43] R. Sutak, J.-M. Camadro, and E. Lesuisse, “Iron Uptake Mechanisms in Marine Phytoplankton,” *Frontiers in Microbiology*, vol. 11, 2020.

[44] X. Gao, C. Bowler, and E. Kazamia, “Iron metabolism strategies in diatoms,” *Journal of Experimental Botany*, vol. 72, pp. 2165–2180, Mar. 2021.

[45] P. W. Boyd, T. Jickells, C. S. Law, S. Blain, E. A. Boyle, K. O. Buesseler, K. H. Coale, J. J. Cullen, H. J. W. de Baar, M. Follows, M. Harvey, C. Lancelot, M. Levasseur, N. P. J. Owens, R. Pollard, R. B. Rivkin, J. Sarmiento, V. Schoemann, V. Smetacek, S. Takeda, A. Tsuda, S. Turner, and A. J. Watson, “Mesoscale Iron Enrichment Experiments 1993-2005: Synthesis and Future Directions,” *Science*, vol. 315, pp. 612–617, Feb. 2007.

[46] F. N. Egerton, “Leeuwenhoek as a founder of animal demography,” *Journal of the History of Biology*, vol. 1, pp. 1–22, Mar. 1968.

[47] A. Adler and E. Dücker, “When Pasteurian Science Went to Sea: The Birth of Marine Microbiology,” *Journal of the History of Biology*, vol. 51, no. 1, pp. 107–133, 2018.

[48] F. Azam, “Introduction, history, and overview: The ‘methods’ to our madness,” in *Methods in Microbiology*, vol. 30 of *Marine Microbiology*, pp. 1–12, Academic Press, Jan. 2001.

[49] L. R. Pomeroy, “The Ocean’s Food Web, A Changing Paradigm,” *BioScience*, vol. 24, pp. 499–504, Sept. 1974.

[50] J. E. Hobbie, R. J. Daley, and S. Jasper, “Use of nuclepore filters for counting bacteria by fluorescence microscopy,” *Applied and Environmental Microbiology*, vol. 33, pp. 1225–1228, May 1977.

[51] Hagström, U. Larsson, P. Hörstedt, and S. Normark, “Frequency of Dividing Cells, a New Approach to the Determination of Bacterial Growth Rates in Aquatic Environments,” *Applied and Environmental Microbiology*, vol. 37, pp. 805–812, May 1979.

[52] P. J. leB Williams, “Incorpooration of microheterotrophic processes into the classical paradigm of the planktonic food web,” *Kieler Meeresforschungen - Sonderheft*, vol. 5, pp. 1–28, 1981.

[53] C. R. Woese, O. Kandler, and M. L. Wheelis, “Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, pp. 4576–4579, June 1990.

[54] N. R. Pace, “A Molecular View of Microbial Diversity and the Biosphere,” *Science*, vol. 276, pp. 734–740, May 1997.

[55] 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,” *PLoS Biology*, vol. 5, p. e77, Mar. 2007.

[56] 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,” *PLOS Biology*, vol. 20, p. e3001835, Oct. 2022.

[57] E. L. van Dijk, Y. Jaszczyszyn, D. Naquin, and C. Thermes, “The Third Revolution in Sequencing Technology,” *Trends in Genetics*, vol. 34, pp. 666–681, Sept. 2018.

[58] T. Stuart and R. Satija, “Integrative single-cell analysis,” *Nature Reviews Genetics*, vol. 20, pp. 257–272, May 2019.

[59] T. Gabaldón, "Origin and Early Evolution of the Eukaryotic Cell," *Annual Review of Microbiology*, vol. 75, no. 1, pp. 631–647, 2021.

[60] D. K. Stoecker, P. J. Hansen, D. A. Caron, and A. Mitra, "Mixotrophy in the Marine Plankton," *Annual Review of Marine Science*, vol. 9, no. 1, pp. 311–335, 2017.

[61] D. K. Stoecker and P. J. Lavrentyev, "Mixotrophic Plankton in the Polar Seas: A Pan-Arctic Review," *Frontiers in Marine Science*, vol. 5, 2018.

[62] T. A. Brown, *Introduction to Genetics: A Molecular Approach*. Garland Science, 2012.

[63] P. G. Higgs and T. K. Attwood, *Bioinformatics and Molecular Evolution*. John Wiley & Sons, Incorporated, 2005.

[64] V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, and M. Borodovsky, "Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training," *Genome Research*, vol. 18, pp. 1979–1990, Jan. 2008.

[65] A. L. dos Santos, T. Pollina, P. Gourvil, E. Corre, D. Marie, J. L. Garrido, F. Rodríguez, M.-H. Noël, D. Vaulot, and W. Eikrem, "Chloropicophyceae, a new class of picophytoplanktonic prasinophytes," *Scientific Reports*, vol. 7, p. 14019, Oct. 2017.

[66] M. Turmel, A. Lopes dos Santos, C. Otis, R. Sergerie, and C. Lemieux, "Tracing the Evolution of the Plastome and Mitogenome in the Chloropicophyceae Uncovered Convergent tRNA Gene Losses and a Variant Plastid Genetic Code," *Genome Biology and Evolution*, vol. 11, pp. 1275–1292, Apr. 2019.

[67] P. G. Falkowski, M. E. Katz, A. H. Knoll, A. Quigg, J. A. Raven, O. Schofield, and F. J. R. Taylor, "The Evolution of Modern Eukaryotic Phytoplankton," *Science*, vol. 305, pp. 354–360, July 2004.

[68] W. Martin, B. Stoebe, V. Goremykin, S. Hansmann, M. Hasegawa, and K. V. Kowallik, "Gene transfer to the nucleus and the evolution of chloroplasts," *Nature*, vol. 393, p. 162, May 1998.

[69] P. J. Keeling, "Chromalveolates and the Evolution of Plastids by Secondary Endosymbiosis1," *Journal of Eukaryotic Microbiology*, vol. 56, no. 1, pp. 1–8, 2009.

[70] D. G. Mann and S. J. M. Droop, "Biodiversity, biogeography and conservation of diatoms," in *Biogeography of Freshwater Algae: Proceedings of the Workshop on Biogeography of Freshwater Algae, Held during the Fifth International Phycological Congress, Qingdao, China, June 1994* (J. Kristiansen, ed.), Developments in Hydrobiology, pp. 19–32, Dordrecht: Springer Netherlands, 1996.

[71] T. Mock, R. P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland, B. J. Ward, A. E. Allen, C. L. Dupont, S. Frickenhaus, F. Maumus, A. Veluchamy, T. Wu, K. W. Barry, A. Falciatore, M. I. Ferrante, A. E. Fortunato, G. Glöckner, A. Gruber, R. Hipkin, M. G. Janech, P. G. Kroth, F. Leese, E. A. Lindquist, B. R. Lyon, J. Martin, C. Mayer, M. Parker, H. Quesneville, J. A. Raymond, C. Uhlig, R. E. Valas, K. U. Valentin, A. Z. Worden, E. V. Armbrust, M. D.

Clark, C. Bowler, B. R. Green, V. Moulton, C. van Oosterhout, and I. V. Grigoriev, "Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*," *Nature*, vol. 541, pp. 536–540, Jan. 2017.

[72] CSIRO, "SEM diatom - CSIRO Science Image - CSIRO Science Image." https://www.scienceimage.csiro.au/image/7632.

[73] "CSIRO Scienc eImage 7632 SEM diatom (cropped)." https://commons.wikimedia.org/wiki/File:Detail,_CSIRO_ScienceImage_7632_SEM_diatom_(cropped).

[74] J. Bradbury, "Nature's Nanotechnologists: Unveiling the Secrets of Diatoms," *PLOS Biology*, vol. 2, p. e306, Oct. 2004.

[75] M. A. Tiffany, "Scanning Electron Micrographs of Diatoms.," Dec. 2004.

[76] M. Hoppenrath, "Dinoflagellate taxonomy — a review and proposal of a revised classification," *Marine Biodiversity*, vol. 47, pp. 381–403, June 2017.

[77] F. J. R. Taylor, M. Hoppenrath, and J. F. Saldarriaga, "Dinoflagellate diversity and distribution," *Biodiversity and Conservation*, vol. 17, pp. 407–418, Feb. 2008.

[78] R. A. Foster, E. J. Carpenter, and B. Bergman, "Unicellular Cyanobionts in Open Ocean Dinoflagellates, Radiolarians, and Tintinnids: Ultrastructural Characterization and Immuno-Localization of Phycoerythrin and Nitrogenase1," *Journal of Phycology*, vol. 42, no. 2, pp. 453–463, 2006.

[79] J. Henderiks, D. Sturm, L. Šupraha, and G. Langer, "Evolutionary Rates in the Haptophyta: Exploring Molecular and Phenotypic Diversity," *Journal of Marine Science and Engineering*, vol. 10, p. 798, June 2022.

[80] A. Winter, J. Henderiks, L. Beaufort, R. E. M. Rickaby, and C. W. Brown, "Poleward expansion of the coccolithophore Emiliania huxleyi," *Journal of Plankton Research*, vol. 36, pp. 316–325, Mar. 2014.

[81] 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," *Nature*, vol. 499, pp. 209–213, July 2013.

[82] F. Burki, A. J. Roger, M. W. Brown, and A. G. B. Simpson, "The New Tree of Eukaryotes," *Trends in Ecology & Evolution*, vol. 35, pp. 43–55, Jan. 2020.

[83] F. Leliaert, D. R. Smith, H. Moreau, M. D. Herron, H. Verbruggen, C. F. Delwiche, and O. De Clerck, "Phylogeny and Molecular Evolution of the Green Algae," *Critical Reviews in Plant Sciences*, vol. 31, pp. 1–46, Jan. 2012.

[84] S. Balzano, D. Marie, P. Gourvil, and D. Vaulot, "Composition of the summer photosynthetic pico and nanoplankton communities in the Beaufort Sea assessed by T-RFLP and sequences of the 18S rRNA gene from flow cytometry sorted samples," *The ISME Journal*, vol. 6, pp. 1480–1498, Aug. 2012.

[85] N. Simon, E. Foulon, D. Grulois, C. Six, Y. Desdevises, M. Latimier, F. Le Gall, M. Tragin, A. Houdan, E. Derelle, F. Jouenne, D. Marie, S. Le Panse, D. Vaulot, and B. Marin, "Revision of the Genus Micromonas Manton et Parke (Chlorophyta, Mamiellophyceae), of the Type Species M. pusilla (Butcher) Manton & Parke and of the Species M. commoda van Baren, Bachy and Worden and Description of Two New Species Based on the Genetic and Phenotypic Characterization of Cultured Isolates," *Protist*, vol. 168, pp. 612–635, Nov. 2017.

[86] D. Demory, A.-C. Baudoux, A. Monier, N. Simon, C. Six, P. Ge, F. Rigaut-Jalabert, D. Marie, A. Sciandra, O. Bernard, and S. Rabouille, "Picoeukaryotes of the Micromonas genus: Sentinels of a warming ocean," *The ISME Journal*, vol. 13, pp. 132–146, Jan. 2019.

[87] F. Rodríguez, E. Derelle, L. Guillou, F. Le Gall, D. Vaulot, and H. Moreau, "Ecotype diversity in the marine picoeukaryote Ostreococcus (Chlorophyta, Prasinophyceae)," *Environmental Microbiology*, vol. 7, no. 6, pp. 853–859, 2005.

[88] E. Demir-Hilton, S. Sudek, M. L. Cuvelier, C. L. Gentemann, J. P. Zehr, and A. Z. Worden, "Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus," *The ISME Journal*, vol. 5, pp. 1095–1107, July 2011.

[89] T. Vannier, J. Leconte, Y. Seeleuthner, S. Mondy, E. Pelletier, J.-M. Aury, C. de Vargas, M. Sieracki, D. Iudicone, D. Vaulot, P. Wincker, and O. Jaillon, "Survey of the green picoalga Bathycoccus genomes in the global ocean," *Scientific Reports*, vol. 6, p. 37900, Nov. 2016.

[90] K. John T. O., *Light and Photosynthesis in Aquatic Ecosystems.*, vol. 3rd ed. Cambridge University Press, 2011.

[91] Y. M. Bar-On and R. Milo, "The global mass and average rate of rubisco," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 4738–4743, Mar. 2019.

[92] R. J. Whittaker, K. J. Willis, and R. Field, "Scale and species richness: Towards a general, hierarchical theory of species diversity," *Journal of Biogeography*, vol. 28, no. 4, pp. 453–470, 2001.

[93] A. D. Willis, "Rarefaction, Alpha Diversity, and Statistics," *Frontiers in Microbiology*, vol. 10, 2019.

[94] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[95] E. H. Simpson, "Measurement of Diversity," *Nature*, vol. 163, pp. 688–688, Apr. 1949.

[96] P. Legendre, "Interpreting the replacement and richness difference components of beta diversity," *Global Ecology and Biogeography*, vol. 23, no. 11, pp. 1324–1334, 2014.

[97] C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight, “Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities,” *Applied and Environmental Microbiology*, vol. 73, pp. 1576–1585, Mar. 2007.

[98] R. De Wit and T. Bouvier, “‘Everything is everywhere, but, the environment selects’; what did Baas Becking and Beijerinck really say?,” *Environmental Microbiology*, vol. 8, no. 4, pp. 755–758, 2006.

[99] D. R. Nemergut, S. K. Schmidt, T. Fukami, S. P. O’Neill, T. M. Bilinski, L. F. Stanish, J. E. Knelman, J. L. Darcy, R. C. Lynch, P. Wickey, and S. Ferrenberg, “Patterns and Processes of Microbial Community Assembly,” *Microbiology and Molecular Biology Reviews*, vol. 77, pp. 342–356, Sept. 2013.

[100] M. F. V. Rodrigues, M. Lisicki, and E. Lauga, “The bank of swimming organisms at the micron scale (BOSO-Micro),” *PLOS ONE*, vol. 16, p. e0252291, June 2021.

[101] A. L. Müller, J. R. de Rezende, C. R. J. Hubert, K. U. Kjeldsen, I. Lagkouvardos, D. Berry, B. B. Jørgensen, and A. Loy, “Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents,” *The ISME Journal*, vol. 8, pp. 1153–1165, June 2014.

[102] M. Mestre and J. Höfer, “The Microbial Conveyor Belt: Connecting the Globe through Dispersion and Dormancy,” *Trends in Microbiology*, vol. 29, pp. 482–492, June 2021.

[103] I. Benner, A. J. Irwin, and Z. V. Finkel, “Capacity of the common Arctic picoeukaryote Micromonas to adapt to a warming ocean,” *Limnology and Oceanography Letters*, vol. 5, no. 2, pp. 221–227, 2020.

[104] L. D. McDaniel, E. Young, J. Delaney, F. Ruhnau, K. B. Ritchie, and J. H. Paul, “High Frequency of Horizontal Gene Transfer in the Oceans,” *Science*, vol. 330, pp. 50–50, Oct. 2010.

[105] J. A. Raymond and H. J. Kim, “Possible Role of Horizontal Gene Transfer in the Colonization of Sea Ice by Algae,” *PLOS ONE*, vol. 7, p. e35968, May 2012.

[106] R. Logares, I. M. Deutschmann, P. C. Junger, C. R. Giner, A. K. Krabberød, T. S. B. Schmidt, L. Rubinat-Ripoll, M. Mestre, G. Salazar, C. Ruiz-González, M. Sebastián, C. de Vargas, S. G. Acinas, C. M. Duarte, J. M. Gasol, and R. Massana, “Disentangling the mechanisms shaping the surface ocean microbiota,” *Microbiome*, vol. 8, p. 55, Apr. 2020.

[107] M.-L. Timmermans and J. Marshall, “Understanding Arctic Ocean Circulation: A Review of Ocean Dynamics in a Changing Climate,” *Journal of Geophysical Research: Oceans*, vol. 125, no. 4, p. e2018JC014378, 2020.

[108] E. Carmack and P. Wassmann, “Food webs and physical–biological coupling on pan-Arctic shelves: Unifying concepts and comprehensive perspectives,” *Progress in Oceanography*, vol. 71, pp. 446–477, Oct. 2006.

[109] I. V. Polyakov, M. B. Alkire, B. A. Bluhm, K. A. Brown, E. C. Carmack, M. Chierici, S. L. Danielson, I. Ellingsen, E. A. Ershova, K. Gårdfeldt, R. B. Ingvaldsen, A. V. Pnyushkov, D. Slagstad, and P. Wassmann, "Borealization of the Arctic Ocean in Response to Anomalous Advection From Sub-Arctic Seas," *Frontiers in Marine Science*, vol. 7, 2020.

[110] M. Jakobsson, L. A. Mayer, C. Bringensparr, C. F. Castro, R. Mohammad, P. Johnson, T. Ketter, D. Accettella, D. Amblas, L. An, J. E. Arndt, M. Canals, J. L. Casamor, N. Chauché, B. Coakley, S. Danielson, M. Demarte, M.-L. Dickson, B. Dorschel, J. A. Dowdeswell, S. Dreutter, A. C. Fremand, D. Gallant, J. K. Hall, L. Hehemann, H. Hodnesdal, J. Hong, R. Ivaldi, E. Kane, I. Klaucke, D. W. Krawczyk, Y. Kristoffersen, B. R. Kuipers, R. Millan, G. Masetti, M. Morlighem, R. Noormets, M. M. Prescott, M. Rebesco, E. Rignot, I. Semiletov, A. J. Tate, P. Travaglini, I. Velicogna, P. Weatherall, W. Weinrebe, J. K. Willis, M. Wood, Y. Zarayskaya, T. Zhang, M. Zimmermann, and K. B. Zinglersen, "The International Bathymetric Chart of the Arctic Ocean Version 4.0," *Scientific Data*, vol. 7, p. 176, Dec. 2020.

[111] C. W. Thackeray and A. Hall, "An emergent constraint on future Arctic sea-ice albedo feedback," *Nature Climate Change*, vol. 9, pp. 972–978, Dec. 2019.

[112] J. W. Deming and R. Eric Collins, "Sea ice as a habitat for Bacteria, Archaea and viruses," in *Sea Ice*, ch. 13, pp. 326–351, John Wiley & Sons, Ltd, 2017.

[113] M. A. van Leeuwe, L. Tedesco, K. R. Arrigo, P. Assmy, K. Campbell, K. M. Meiners, J.-M. Rintala, V. Selz, D. N. Thomas, and J. Stefels, "Microalgal community structure and primary production in Arctic and Antarctic sea ice: A synthesis," *Elementa: Science of the Anthropocene*, vol. 6, p. 4, Jan. 2018.

[114] I. Werner, J. Ikävalko, and H. Schünemann, "Sea-ice algae in Arctic pack ice during late winter," *Polar Biology*, vol. 30, pp. 1493–1504, Oct. 2007.

[115] M. Ardyna, M. Babin, M. Gosselin, E. Devred, L. Rainville, and J.-É. Tremblay, "Recent Arctic Ocean sea ice loss triggers novel fall phytoplankton blooms," *Geophysical Research Letters*, vol. 41, no. 17, pp. 6207–6212, 2014.

[116] A. Tammilehto, P. C. Watts, and N. Lundholm, "Isolation by Time During an Arctic Phytoplankton Spring Bloom," *The Journal of Eukaryotic Microbiology*, vol. 64, pp. 248–256, Mar. 2017.

[117] S. C. Doney, V. J. Fabry, R. A. Feely, and J. A. Kleypas, "Ocean Acidification: The Other CO2 Problem," *Annual Review of Marine Science*, vol. 1, no. 1, pp. 169–192, 2009.

[118] S. Dutkiewicz, J. J. Morris, M. J. Follows, J. Scott, O. Levitan, S. T. Dyhrman, and I. Berman-Frank, "Impact of ocean acidification on the structure of future phytoplankton communities," *Nature Climate Change*, vol. 5, pp. 1002–1006, Nov. 2015.

[119] K. T. Lohbeck, U. Riebesell, and T. B. H. Reusch, "Adaptive evolution of a key phytoplankton species to ocean acidification," *Nature Geoscience*, vol. 5, pp. 346–351, May 2012.

[120] S. L. Hinder, G. C. Hays, M. Edwards, E. C. Roberts, A. W. Walne, and M. B. Gravenor, “Changes in marine dinoflagellate and diatom abundance under climate change,” *Nature Climate Change*, vol. 2, pp. 271–275, Apr. 2012.

[121] M. Previdi, K. L. Smith, and L. M. Polvani, “Arctic amplification of climate change: A review of underlying mechanisms,” *Environmental Research Letters*, vol. 16, p. 093003, Sept. 2021.

[122] J. Stroeve and D. Notz, “Changing state of Arctic sea ice across all seasons,” *Environmental Research Letters*, vol. 13, p. 103001, Sept. 2018.

[123] J. R. Farmer, D. M. Sigman, J. Granger, O. M. Underwood, F. Fripiat, T. M. Cronin, A. Martínez-García, and G. H. Haug, “Arctic Ocean stratification set by sea level and freshwater inputs since the last ice age,” *Nature Geoscience*, vol. 14, pp. 684–689, Sept. 2021.

[124] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proceedings of the National Academy of Sciences*, vol. 74, pp. 5463–5467, Dec. 1977.

[125] J. Shendure and H. Ji, “Next-generation DNA sequencing,” *Nature Biotechnology*, vol. 26, pp. 1135–1145, Oct. 2008.

[126] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. G. Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri,

J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, "The Sequence of the Human Genome," *Science*, vol. 291, pp. 1304–1351, Feb. 2001.

[127] J. A. Reuter, D. V. Spacek, and M. P. Snyder, "High-Throughput Sequencing Technologies," *Molecular Cell*, vol. 58, pp. 586–597, May 2015.

[128] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. H. Kent, and L. E. Hood, "Fluorescence detection in automated DNA sequence analysis," *Nature*, vol. 321, p. 674, June 1986.

[129] A. T. Woolley and R. A. Mathies, "Ultra-High-Speed DNA Sequencing Using Capillary Electrophoresis Chips," *Analytical Chemistry*, vol. 67, pp. 3676–3680, Oct. 1995.

[130] C. Quince, A. W. Walker, J. T. Simpson, N. J. Loman, and N. Segata, "Shotgun metagenomics, from sampling to analysis," *Nature Biotechnology*, vol. 35, pp. 833–844, Sept. 2017.

[131] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon,

J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith, "Accurate whole human genome sequencing using reversible terminator chemistry," *Nature*, vol. 456, pp. 53–59, Nov. 2008.

[132] EMBL-EBI Train online, "Illumina sequencing." https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/illumina-, Sept. 2012.

[133] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, p. 376, Sept. 2005.

[134] J. M. Rothberg and J. H. Leamon, "The development and impact of 454 sequencing," *Nature Biotechnology*, vol. 26, pp. 1117–1124, Oct. 2008.

[135] B. Merriman, I. T. R. Team, and J. M. Rothberg, "Progress in Ion Torrent semiconductor chip based sequencing," *Electrophoresis*, vol. 33, no. 23, pp. 3397–3417, 2012.

[136] S. Ambardar, R. Gupta, D. Trakroo, R. Lal, and J. Vakhlu, "High Throughput Sequencing: An Overview of Sequencing Chemistry," *Indian Journal of Microbiology*, vol. 56, pp. 394–404, Dec. 2016.

[137] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and E. Al, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, pp. 496–512, July 1995.

[138] J. L. Weber and E. W. Myers, "Human Whole-Genome Shotgun Sequencing," *Genome Research*, vol. 7, pp. 401–409, Jan. 1997.

[139] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W.

Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith, “Environmental Genome Shotgun Sequencing of the Sargasso Sea,” *Science*, vol. 304, pp. 66–74, Apr. 2004.

[140] S. Taylor, M. Wakem, G. Dijkman, M. Alsarraj, and M. Nguyen, “A practical approach to RT-qPCR—Publishing data that conform to the MIQE guidelines,” *Methods*, vol. 50, pp. S1–S5, Apr. 2010.

[141] M. O. Pollard, D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, “Long reads: Their purpose and place,” *Human Molecular Genetics*, vol. 27, pp. R234–R241, Aug. 2018.

[142] D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke, “Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries,” *Genome Biology*, vol. 12, p. R18, Feb. 2011.

[143] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, “Real-Time DNA Sequencing from Single Polymerase Molecules,” *Science*, vol. 323, pp. 133–138, Jan. 2009.

[144] S. Koren, G. P. Harhay, T. P. Smith, J. L. Bono, D. M. Harhay, S. D. Mcvey, D. Radune, N. H. Bergman, and A. M. Phillippy, “Reducing assembly complexity of microbial genomes with single-molecule sequencing,” *Genome Biology*, vol. 14, p. R101, Sept. 2013.

[145] D. Branton, D. W. Deamer, A. Marziali, H. Bayley, S. A. Benner, T. Butler, M. Di Ventra, S. Garaj, A. Hibbs, X. Huang, S. B. Jovanovich, P. S. Krstic, S. Lindsay, X. S. Ling, C. H. Mastrangelo, A. Meller, J. S. Oliver, Y. V. Pershin, J. M. Ramsey, R. Riehn, G. V. Soni, V. T. Cossa, M. Wanunu, M. Wiggin, and J. A. Schloss, “The potential and challenges of nanopore sequencing,” in *Nanoscience and Technology*, pp. 261–268, Co-Published with Macmillan Publishers Ltd, UK, Aug. 2009.

[146] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, “The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community,” *Genome Biology*, vol. 17, p. 239, Nov. 2016.

[147] Y. Wang, Y. Zhao, A. Bollas, Y. Wang, and K. F. Au, “Nanopore sequencing technology, bioinformatics and applications,” *Nature Biotechnology*, vol. 39, pp. 1348–1365, Nov. 2021.

[148] S. Goodwin, J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. C. Schatz, and W. R. McCombie, “Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome,” *Genome Research*, vol. 25, pp. 1750–1756, Jan. 2015.

[149] M. Jain, S. Koren, K. H. Miga, J. Quick, A. C. Rand, T. A. Sasani, J. R. Tyson, A. D. Beggs, A. T. Dilthey, I. T. Fiddes, S. Malla, H. Marriott, T. Nieto, J. O'Grady, H. E. Olsen, B. S. Pedersen, A. Rhie, H. Richardson, A. R. Quinlan, T. P. Snutch, L. Tee, B. Paten, A. M. Phillippy, J. T. Simpson, N. J. Loman, and M. Loose, "Nanopore sequencing and assembly of a human genome with ultra-long reads," *Nature Biotechnology*, vol. 36, pp. 338–345, Apr. 2018.

[150] D. Antipov, A. Korobeynikov, J. S. McLean, and P. A. Pevzner, "hybridSPAdes: An algorithm for hybrid assembly of short and long reads," *Bioinformatics*, vol. 32, pp. 1009–1015, Apr. 2016.

[151] A. Cossarizza, H.-D. Chang, A. Radbruch, M. Akdis, I. Andrä, F. Annunziato, P. Bacher, V. Barnaba, L. Battistini, W. M. Bauer, S. Baumgart, B. Becher, W. Beisker, C. Berek, A. Blanco, G. Borsellino, P. E. Boulais, R. R. Brinkman, M. Büscher, D. H. Busch, T. P. Bushnell, X. Cao, A. Cavani, P. K. Chattopadhyay, Q. Cheng, S. Chow, M. Clerici, A. Cooke, A. Cosma, L. Cosmi, A. Cumano, V. D. Dang, D. Davies, S. De Biasi, G. Del Zotto, S. D. Bella, P. Dellabona, G. Deniz, M. Dessing, A. Diefenbach, J. Di Santo, F. Dieli, A. Dolf, V. S. Donnenberg, T. Dörner, G. R. Ehrhardt, E. Endl, P. Engel, B. Engelhardt, C. Esser, B. Everts, A. Dreher, C. S. Falk, T. A. Fehniger, A. Filby, S. Fillatreau, M. Follo, I. Förster, J. Foster, G. A. Foulds, P. S. Frenette, D. Galbraith, N. Garbi, M. D. García-Godoy, J. Geginat, K. Ghoreschi, L. Gibellini, C. Goettlinger, C. S. Goodyear, A. Gori, J. Grogan, M. Gross, A. Grützkau, D. Grummitt, J. Hahn, Q. Hammer, A. E. Hauser, D. L. Haviland, D. Hedley, G. Herrera, M. Herrmann, F. Hiepe, T. Holland, P. Hombrink, J. P. Houston, B. F. Hoyer, B. Huang, C. A. Hunter, A. Iannone, H.-M. Jäck, B. Jávega, S. Jonjic, K. Juelke, S. Jung, T. Kaiser, T. Kalina, B. Keller, S. Khan, D. Kienhöfer, T. Kroneis, D. Kunkel, C. Kurts, P. Kvistborg, J. Lannigan, O. Lantz, A. Larbi, S. L. Gut-Landmann, M. D. Leipold, M. K. Levings, V. Litwin, Y. Liu, M. Lohoff, G. Lombardi, L. Lopez, A. Lovett-Racke, E. Lubberts, B. Ludewig, E. Lugli, H. T. Maecker, G. Martrus, G. Matarese, C. Maueröder, M. McGrath, I. McInnes, H. E. Mei, F. Melchers, S. Melzer, D. Mielenz, K. Mills, D. Mirrer, J. Mjösberg, J. Moore, B. Moran, A. Moretta, L. Moretta, T. R. Mosmann, S. Müller, W. Müller, C. Münz, G. Multhoff, L. E. Munoz, K. M. Murphy, T. Nakayama, M. Nasi, C. Neudörfl, J. Nolan, S. Nourshargh, J.-E. O'Connor, W. Ouyang, A. Oxenius, R. Palankar, I. Panse, P. Peterson, C. Peth, J. Petriz, D. Philips, W. Pickl, S. Piconese, M. Pinti, A. G. Pockley, M. J. Podolska, C. Pucillo, S. A. Quataert, T. R. D. J. Radstake, B. Rajwa, J. A. Rebhahn, D. Recktenwald, E. B. Remmerswaal, K. Rezvani, L. G. Rico, J. P. Robinson, C. Romagnani, A. Rubartelli, B. Ruckert, J. Ruland, S. Sakaguchi, F. Sala-de-Oyanguren, Y. Samstag, S. Sanderson, B. Sawitzki, A. Scheffold, M. Schiemann, F. Schildberg, E. Schimisky, S. A. Schmid, S. Schmitt, K. Schober, T. Schüler, A. R. Schulz, T. Schumacher, C. Scotta, T. V. Shankey, A. Shemer, A.-K. Simon, J. Spidlen, A. M. Stall, R. Stark, C. Stehle, M. Stein, T. Steinmetz, H. Stockinger, Y. Takahama, A. Tarnok, Z. G. Tian, G. Toldi, J. Tornack, E. Traggiai, J. Trotter, H. Ulrich, M. van der Braber, R. A. van Lier, M. Veldhoen, S. Vento-Asturias, P. Vieira, D. Voehringer, H.-D. Volk, K. von Volkmann, A. Waisman, R. Walker, M. D. Ward, K. Warnatz, S. Warth, J. V. Watson, C. Watzl, L. Wegener, A. Wiedemann, J. Wienands, G. Willimsky, J. Wing, P. Wurst, L. Yu, A. Yue, Q. Zhang, Y. Zhao, S. Ziegler, and J. Zimmermann, "Guidelines for the use of flow cytometry and cell sorting in immunological studies," *European Journal of Immunology*, vol. 47, pp. 1584–1797, Oct. 2017.

[152] T. O. Delmont, M. Gaia, D. D. Hinsinger, P. Fremont, A. F. Guerra, A. M. Eren, C. Vanni, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. D. Silva, M. Wessner, B. Noel, J.-M. Aury, T. O. Coordinators, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, and O. Jaillon, "Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics," *bioRxiv*, p. 2020.10.15.341214, Oct. 2020.

[153] P. D. Schloss and J. Handelsman, "Metagenomics for studying unculturable microorganisms: Cutting the Gordian knot," *Genome Biology*, vol. 6, p. 229, Aug. 2005.

[154] J. del Campo, M. E. Sieracki, R. Molestina, P. Keeling, R. Massana, and I. Ruiz-Trillo, "The others: Our biased perspective of eukaryotic genomes," *Trends in Ecology & Evolution*, vol. 29, pp. 252–259, May 2014.

[155] M. Ayling, M. D. Clark, and R. M. Leggett, "New approaches for metagenome assembly with short reads," *Briefings in Bioinformatics*, Feb. 2019.

[156] 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, p. 170203, Jan. 2018.

[157] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities," *PeerJ*, vol. 3, p. e1165, Aug. 2015.

[158] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, N. J. Loman, A. F. Andersson, and C. Quince, "CONCOCT: Clustering cONtigs on COverage and ComposiTion," *arXiv:1312.4038 [q-bio]*, Dec. 2013.

[159] E. D. Graham, J. F. Heidelberg, and B. J. Tully, "BinSanity: Unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation," *PeerJ*, vol. 5, p. e3035, Mar. 2017.

[160] M. A. Moran, B. Satinsky, S. M. Gifford, H. Luo, A. Rivers, L.-K. Chan, J. Meng, B. P. Durham, C. Shen, V. A. Varaljay, C. B. Smith, P. L. Yager, and B. M. Hopkinson, "Sizing up metatranscriptomics," *The ISME Journal*, vol. 7, pp. 237–243, Feb. 2013.

[161] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, "Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells," *Science*, vol. 329, pp. 533–538, July 2010.

[162] J. Frias-Lopez, Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. DeLong, "Microbial community gene expression in ocean surface waters," *Proceedings of the National Academy of Sciences*, vol. 105, pp. 3805–3810, Mar. 2008.

[163] O. U. Mason, T. C. Hazen, S. Borglin, P. S. G. Chain, E. A. Dubinsky, J. L. Fortney, J. Han, H.-Y. N. Holman, J. Hultman, R. Lamendella, R. Mackelprang, S. Malfatti, L. M. Tom, S. G. Tringe, T. Woyke, J. Zhou, E. M. Rubin, and J. K. Jansson, "Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill," *The ISME Journal*, vol. 6, pp. 1715–1727, Sept. 2012.

[164] 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, pp. 2196–2204, Mar. 2000.

[165] K. Paszkiewicz and D. J. Studholme, "De novo assembly of short sequence reads," *Briefings in Bioinformatics*, vol. 11, pp. 457–472, Sept. 2010.

[166] P. A. Pevzner, H. Tang, and M. S. Waterman, "An Eulerian path approach to DNA fragment assembly," *Proceedings of the National Academy of Sciences*, vol. 98, pp. 9748–9753, Aug. 2001.

[167] J. R. Miller, S. Koren, and G. Sutton, "Assembly algorithms for next-generation sequencing data," *Genomics*, vol. 95, pp. 315–327, June 2010.

[168] D. R. Zerbino and E. Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research*, vol. 18, pp. 821–829, Jan. 2008.

[169] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang, "De novo assembly of human genomes with massively parallel short read sequencing," *Genome Research*, vol. 20, pp. 265–272, Jan. 2010.

[170] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang, "SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler," *GigaScience*, vol. 1, p. 18, Dec. 2012.

[171] S. D. Jackman, B. P. Vandervalk, H. Mohamadi, J. Chu, S. Yeo, S. A. Hammond, G. Jahesh, H. Khan, L. Coombe, R. L. Warren, and I. Birol, "ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter," *Genome Research*, vol. 27, pp. 768–777, Jan. 2017.

[172] R. Chikhi and G. Rizk, "Space-efficient and exact de Bruijn graph representation based on a Bloom filter," *Algorithms for Molecular Biology*, vol. 8, p. 22, Sept. 2013.

[173] R. Chikhi, A. Limasset, and P. Medvedev, "Compacting de Bruijn graphs from sequencing data quickly and in low memory," *Bioinformatics*, vol. 32, pp. i201–i208, June 2016.

[174] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth," *Bioinformatics*, vol. 28, pp. 1420–1428, June 2012.

[175] 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," *Bioinformatics*, vol. 31, pp. 1674–1676, May 2015.

[176] A. Bowe, T. Onodera, K. Sadakane, and T. Shibuya, “Succinct de Bruijn Graphs,” in *Algorithms in Bioinformatics* (B. Raphael and J. Tang, eds.), Lecture Notes in Computer Science, pp. 225–235, Springer Berlin Heidelberg, 2012.

[177] M. Kim, X. Zhang, J. G. Ligo, F. Farnoud, V. V. Veeravalli, and O. Milenkovic, “MetaCRAM: An integrated pipeline for metagenomic taxonomy identification and compression,” *BMC Bioinformatics*, vol. 17, p. 94, Feb. 2016.

[178] 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,” *Journal of Computational Biology*, vol. 19, pp. 455–477, Apr. 2012.

[179] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, “metaSPAdes: A new versatile metagenomic assembler,” *Genome Research*, p. gr.213959.116, Mar. 2017.

[180] M. Kolmogorov, J. Yuan, Y. Lin, and P. A. Pevzner, “Assembly of long, error-prone reads using repeat graphs,” *Nature Biotechnology*, vol. 37, pp. 540–546, May 2019.

[181] M. Kolmogorov, D. M. Bickhart, B. Behsaz, A. Gurevich, M. Rayko, S. B. Shin, K. Kuhn, J. Yuan, E. Polevikov, T. P. L. Smith, and P. A. Pevzner, “metaFlye: Scalable long-read metagenome assembly using repeat graphs,” *Nature Methods*, vol. 17, pp. 1103–1110, Nov. 2020.

[182] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl, “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement,” *PLOS ONE*, vol. 9, p. e112963, Nov. 2014.

[183] S. Krakau, D. Straub, H. Gourlé, G. Gabernet, and S. Nahnsen, “Nf-core/mag: A best-practice pipeline for metagenome hybrid assembly and binning,” *NAR Genomics and Bioinformatics*, vol. 4, p. lqac007, Mar. 2022.

[184] J. A. Frank, Y. Pan, A. Tooming-Klunderud, V. G. H. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope, “Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data,” *Scientific Reports*, vol. 6, p. 25373, May 2016.

[185] N. Hubert and R. Hanner, “DNA barcoding, species delineation and taxonomy: A historical perspective,” *DNA Barcodes*, vol. 3, no. 1, pp. 44–58, 2015.

[186] S. T. Garnett and L. Christidis, “Taxonomy anarchy hampers conservation,” *Nature News*, vol. 546, p. 25, June 2017.

[187] 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*, vol. 36, pp. 1925–1927, Mar. 2020.

[188] M. Balvočiūtė and D. H. Huson, “SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare?,” *BMC Genomics*, vol. 18, p. 114, Mar. 2017.

[189] J. Huerta-Cepas, F. Serra, and P. Bork, "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data," *Molecular Biology and Evolution*, vol. 33, pp. 1635–1638, June 2016.

[190] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 35, pp. D61–D65, Jan. 2007.

[191] 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," *Nucleic Acids Research*, vol. 46, pp. D692–D699, Jan. 2018.

[192] E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B. W. Trawick, K. D. Pruitt, and S. T. Sherry, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 50, pp. D20–D26, Jan. 2022.

[193] C. Simon and R. Daniel, "Metagenomic Analyses: Past and Future Trends," *Appl. Environ. Microbiol.*, vol. 77, pp. 1153–1161, Feb. 2011.

[194] A. E. Pérez-Cobas, L. Gomez-Valero, and C. Buchrieser, "Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses," *Microbial Genomics*, vol. 6, p. mgen000409, July 2020.

[195] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner, "SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Research*, vol. 35, pp. 7188–7196, Dec. 2007.

[196] D. M. Hillis and M. T. Dixon, "Ribosomal DNA: Molecular Evolution and Phylogenetic Inference," *The Quarterly Review of Biology*, vol. 66, pp. 411–453, Dec. 1991.

[197] R. Logares, S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera, H. Sarmento, P. Hingamp, H. Ogata, C. de Vargas, G. Lima-Mendez, J. Raes, J. Poulain, O. Jaillon, P. Wincker, S. Kandels-Lewis, E. Karsenti, P. Bork, and S. G. Acinas, "Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities," *Environmental Microbiology*, vol. 16, no. 9, pp. 2659–2671, 2014.

[198] E. A. Eloe-Fadrosh, N. N. Ivanova, T. Woyke, and N. C. Kyrpides, "Metagenomics uncovers gaps in amplicon-based detection of microbial diversity," *Nature Microbiology*, vol. 1, p. 15032, Apr. 2016.

[199] M. Tessler, J. S. Neumann, E. Afshinnekoo, M. Pineda, R. Hersch, L. F. M. Velho, B. T. Segovia, F. A. Lansac-Toha, M. Lemke, R. DeSalle, C. E. Mason, and M. R. Brugler, "Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing," *Scientific Reports*, vol. 7, p. 6589, July 2017.

[200] W. M. Fitch, “Homology: A personal view on some of the problems,” *Trends in Genetics*, vol. 16, pp. 227–231, May 2000.

[201] T. F. Smith and M. S. Waterman, “Comparison of biosequences,” *Advances in Applied Mathematics*, vol. 2, pp. 482–489, Dec. 1981.

[202] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, pp. 403–410, Oct. 1990.

[203] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: A new generation of protein database search programs,” *Nucleic Acids Research*, vol. 25, pp. 3389–3402, Sept. 1997.

[204] W. J. Kent, “BLAT—The BLAST-Like Alignment Tool,” *Genome Research*, vol. 12, pp. 656–664, Jan. 2002.

[205] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, “Adaptive seeds tame genomic sequence comparison,” *Genome Research*, vol. 21, pp. 487–493, Jan. 2011.

[206] L. Noé and G. Kucherov, “YASS: Enhancing the sensitivity of DNA similarity search,” *Nucleic Acids Research*, vol. 33, pp. W540–W543, July 2005.

[207] 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, p. R25, Mar. 2009.

[208] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM,” *arXiv:1303.3997 [q-bio]*, Mar. 2013.

[209] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nature Biotechnology*, vol. 34, pp. 525–527, May 2016.

[210] L. Schaeffer, H. Pimentel, N. Bray, P. Melsted, and L. Pachter, “Pseudoalignment for metagenomic read assignment,” *Bioinformatics*, vol. 33, pp. 2082–2088, July 2017.

[211] B.-J. Yoon, “Hidden Markov Models and their Applications in Biological Sequence Analysis,” *Current Genomics*, vol. 10, pp. 402–415, Sept. 2009.

[212] G. D. Forney, “The viterbi algorithm,” *Proceedings of the IEEE*, vol. 61, pp. 268–278, Mar. 1973.

[213] S. R. Eddy, “Profile hidden Markov models.,” *Bioinformatics*, vol. 14, pp. 755–763, Jan. 1998.

[214] A. Bateman, E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. L. Sonnhammer, “The Pfam Protein Families Database,” *Nucleic Acids Research*, vol. 28, pp. 263–266, Jan. 2000.

[215] S. R. Eddy, “A new generation of homology search tools based on probabilistic inference,” in *Genome Informatics 2009*, pp. 205–211, Imperial College Press, Oct. 2009.

[216] R. D. Finn, J. Clements, and S. R. Eddy, “HMMER web server: Interactive sequence similarity searching,” *Nucleic Acids Research*, vol. 39, pp. W29–W37, July 2011.

[217] A. Brady and S. L. Salzberg, “Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models,” *Nature Methods*, vol. 6, pp. 673–676, Sept. 2009.

[218] D. E. Wood and S. L. Salzberg, “Kraken: Ultrafast metagenomic sequence classification using exact alignments,” *Genome Biology*, vol. 15, p. R46, Mar. 2014.

[219] J. Lu, F. P. Breitwieser, P. Thielen, and S. L. Salzberg, “Bracken: Estimating species abundance in metagenomics data,” *PeerJ Computer Science*, vol. 3, p. e104, Jan. 2017.

[220] P. Menzel, K. L. Ng, and A. Krogh, “Fast and sensitive taxonomic classification for metagenomics with Kaiju,” *Nature Communications*, vol. 7, p. 11257, Apr. 2016.

[221] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi, “CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers,” *BMC Genomics*, vol. 16, p. 236, Mar. 2015.

[222] J. Bengtsson-Palme, M. Hartmann, K. M. Eriksson, C. Pal, K. Thorell, D. G. J. Larsson, and R. H. Nilsson, “Metaxa2: Improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data,” *Molecular Ecology Resources*, vol. 15, no. 6, pp. 1403–1414, 2015.

[223] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, “MEGAN analysis of metagenomic data,” *Genome Research*, vol. 17, pp. 000–000, Jan. 2007.

[224] B. Buchfink, D. H. Huson, and C. Xie, “MetaScope - Fast and accurate identification of microbes in metagenomic sequencing data,” *arXiv:1511.08753 [q-bio]*, Nov. 2015.

[225] J. Dröge, I. Gregor, and A. C. McHardy, “Taxator-tk: Precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods,” *Bioinformatics*, vol. 31, pp. 817–824, Mar. 2015.

[226] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. V. Horn, and C. F. Weber, “Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities,” *Applied and Environmental Microbiology*, vol. 75, pp. 7537–7541, Dec. 2009.

[227] T. J. Sharpton, S. J. Riesenfeld, S. W. Kembel, J. Ladau, J. P. O’Dwyer, J. L. Green, J. A. Eisen, and K. S. Pollard, “PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data,” *PLOS Computational Biology*, vol. 7, p. e1001061, Jan. 2011.

[228] B. J. Callahan, P. J. McMurdie, and S. P. Holmes, “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis,” *The ISME Journal*, vol. 11, pp. 2639–2643, Dec. 2017.

[229] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight, "QIIME allows analysis of high-throughput community sequencing data," *Nature Methods*, vol. 7, pp. 335–336, May 2010.

[230] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy," *Appl. Environ. Microbiol.*, vol. 73, pp. 5261–5267, Aug. 2007.

[231] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, EM. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and RA. Edwards, "The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, p. 386, Sept. 2008.

[232] R. Overbeek, R. Olson, G. D. Pusch, G. J. Olsen, J. J. Davis, T. Disz, R. A. Edwards, S. Gerdes, B. Parrello, M. Shukla, V. Vonstein, A. R. Wattam, F. Xia, and R. Stevens, "The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST)," *Nucleic Acids Research*, vol. 42, pp. D206–D214, Jan. 2014.

[233] B. Hillmann, G. A. Al-Ghalith, R. R. Shields-Cutler, Q. Zhu, D. M. Gohl, K. B. Beckman, R. Knight, and D. Knights, "Evaluating the Information Content of Shallow Shotgun Metagenomics," *mSystems*, vol. 3, pp. e00069–18, Oct. 2018.

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

[235] 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," *Genome Research*, Mar. 2018.

[236] 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, p. 26, Feb. 2019.

[237] T. O. Delmont, A. M. Eren, J. H. Vineis, and A. F. Post, "Genome reconstructions indicate the partitioning of ecological functions inside a phytoplankton bloom in the Amundsen Sea, Antarctica," *Frontiers in Microbiology*, vol. 6, 2015.

[238] 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 dataset illuminate trophic modes and biogeography of ocean plankton," *bioRxiv*, p. 2021.07.25.453713, July 2021.

[239] P. A. Noble, R. W. Citek, and O. A. Ogunseitan, “Tetranucleotide frequencies in microbial genomes,” *Electrophoresis*, vol. 19, pp. 528–535, Apr. 1998.

[240] B. Siranosian, S. Perera, E. Williams, C. Ye, C. de Graffenried, and P. Shank, “Tetranucleotide usage highlights genomic heterogeneity among mycobacteriophages,” *F1000Research*, vol. 4, Oct. 2015.

[241] K. C. Wrighton, B. C. Thomas, I. Sharon, C. S. Miller, C. J. Castelle, N. C. VerBerkmoes, M. J. Wilkins, R. L. Hettich, M. S. Lipton, K. H. Williams, P. E. Long, and J. F. Banfield, “Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla,” *Science*, vol. 337, pp. 1661–1665, Sept. 2012.

[242] D. R. Kelley and S. L. Salzberg, “Clustering metagenomic sequences with interpolated Markov models,” *BMC Bioinformatics*, vol. 11, p. 544, Nov. 2010.

[243] A. Kislyuk, S. Bhatnagar, J. Dushoff, and J. S. Weitz, “Unsupervised statistical clustering of environmental shotgun sequences,” *BMC Bioinformatics*, vol. 10, p. 316, Oct. 2009.

[244] M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen, “Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes,” *Nature Biotechnology*, vol. 31, pp. 533–538, June 2013.

[245] 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,” *Nature Methods*, vol. 11, pp. 1144–1146, Nov. 2014.

[246] A. M. Eren, Ö. 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, p. e1319, Oct. 2015.

[247] 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,” *Nature Microbiology*, vol. 3, p. 836, July 2018.

[248] S. Canzar and S. L. Salzberg, “Short Read Mapping: An Algorithmic Tour,” *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 105, pp. 436–458, Mar. 2017.

[249] P. Ferragina and G. Manzini, “Opportunistic data structures with applications,” in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pp. 390–398, Nov. 2000.

[250] Y. song, Q. Yao, X. Yao, J. Wright, G. Wang, T. Hazen, B. Turner, M. Tfaily, L. Paša-Tolic, E. Johnston, M. Kim, K. Konstantinos, C. Pan, and M. Mayes, “Metagenomics-informed soil biogeochemical models projected less carbon loss in tropical soils in response to climate warming,” *Researchsquare*, Oct. 2021.

[251] M. R. Olm, C. T. Brown, B. Brooks, and J. F. Banfield, “dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication,” *The ISME Journal*, vol. 11, pp. 2864–2868, Dec. 2017.

[252] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, “BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs,” *Bioinformatics*, vol. 31, pp. 3210–3212, Oct. 2015.

[253] 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, The Genome Standards Consortium, N. C. Kyrpides, L. Schriml, G. M. Garrity, P. Hugenholtz, G. Sutton, P. Yilmaz, F. Meyer, F. O. Glöckner, J. A. Gilbert, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, 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,” *Nature Biotechnology*, vol. 35, pp. 725–731, Aug. 2017.

[254] 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,” *Genome Research*, vol. 25, pp. 1043–1055, Jan. 2015.

[255] P. Saary, A. L. Mitchell, and R. D. Finn, “Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC,” *Genome Biology*, vol. 21, p. 244, Sept. 2020.

[256] F. A. Matsen, R. B. Kodner, and E. Virginia Armbrust, “Pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree,” *BMC Bioinformatics*, vol. 11, pp. 538–553, Jan. 2010.

[257] E. Levy Karin, M. Mirdita, and J. Söding, “MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics,” *Microbiome*, vol. 8, p. 48, Apr. 2020.

[258] 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,” *Nature Biotechnology*, vol. 36, pp. 996–1004, Nov. 2018.

[259] A. I. Krinos, S. K. Hu, N. R. Cohen, and H. Alexander, “EUKulele: Taxonomic annotation of the unsung eukaryotic microbes,” *arXiv:2011.00089 [q-bio]*, Oct. 2020.

[260] 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," *PLOS Biology*, vol. 12, p. e1001889, June 2014.

[261] C. Quince, S. Nurk, S. Raguideau, R. James, O. S. Soyer, J. K. Summers, A. Limasset, A. M. Eren, R. Chikhi, and A. E. Darling, "Metagenomics Strain Resolution on Assembly Graphs," *bioRxiv*, p. 2020.09.06.284828, Sept. 2020.

[262] M. Buck, M. Mehrshad, and S. Bertilsson, "mOTUpan: A robust Bayesian approach to leverage metagenome-assembled genomes for core-genome estimation," *NAR Genomics and Bioinformatics*, vol. 4, p. lqac060, Sept. 2022.

[263] L. M. Ward, P. M. Shih, and W. W. Fischer, "MetaPOAP: Presence or absence of metabolic pathways in metagenome-assembled genomes," *Bioinformatics*, vol. 34, pp. 4284–4286, Dec. 2018.

[264] T. Li and Y. Yin, "Critical assessment of pan-genomics of metagenome-assembled genomes," *bioRxiv*, p. 2022.01.13.476228, Jan. 2022.

[265] M. Karlicki, S. Antonowicz, and A. Karnkowska, "Tiara: Deep learning-based classification system for eukaryotic sequences," *Bioinformatics*, vol. 38, pp. 344–350, Jan. 2022.

[266] L. J. Pronk and M. H. . Medema, "Whokaryote: Distinguishing eukaryotic and prokaryotic contigs in metagenomes based on gene structure," *Microbial Genomics*, vol. 8, p. 000823, May 2021.

[267] S. E. Morales, A. Biswas, G. J. Herndl, and F. Baltar, "Global Structuring of Phylogenetic and Functional Diversity of Pelagic Fungi by Depth and Temperature," *Frontiers in Marine Science*, vol. 6, 2019.

[268] A. Amend, G. Burgaud, M. Cunliffe, V. P. Edgcomb, C. L. Ettinger, M. H. Gutiérrez, J. Heitman, E. F. Y. Hom, G. Ianiri, A. C. Jones, M. Kagami, K. T. Picard, C. A. Quandt, S. Raghukumar, M. Riquelme, J. Stajich, J. Vargas-Muñiz, A. K. Walker, O. Yarden, and A. S. Gladfelter, "Fungi in the Marine Environment: Open Questions and Unsolved Problems," *mBio*, vol. 10, Apr. 2019.

[269] J. Besemer and M. Borodovsky, "Heuristic approach to deriving models for gene finding," *Nucleic Acids Research*, vol. 27, pp. 3911–3920, Oct. 1999.

[270] W. Zhu, A. Lomsadze, and M. Borodovsky, “Ab initio gene identification in metagenomic sequences,” *Nucleic Acids Research*, vol. 38, pp. e132–e132, July 2010.

[271] D. R. Kelley, B. Liu, A. L. Delcher, M. Pop, and S. L. Salzberg, “Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering,” *Nucleic Acids Research*, vol. 40, pp. e9–e9, Jan. 2012.

[272] M. Rho, H. Tang, and Y. Ye, “FragGeneScan: Predicting genes in short and error-prone reads,” *Nucleic Acids Research*, vol. 38, pp. e191–e191, Nov. 2010.

[273] H. Noguchi, J. Park, and T. Takagi, “MetaGene: Prokaryotic gene finding from environmental genome shotgun sequences,” *Nucleic Acids Research*, vol. 34, pp. 5623–5630, Nov. 2006.

[274] D. Hyatt, P. F. LoCascio, L. J. Hauser, and E. C. Uberbacher, “Gene and translation initiation site prediction in metagenomic sequences,” *Bioinformatics*, vol. 28, pp. 2223–2230, Sept. 2012.

[275] 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, p. 119, Mar. 2010.

[276] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Research*, vol. 28, pp. 27–30, Jan. 2000.

[277] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham, P. Kaipa, A. S. Karthikeyan, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, S. Paley, L. Popescu, A. Pujar, A. G. Shearer, P. Zhang, and P. D. Karp, “The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases,” *Nucleic Acids Research*, vol. 38, pp. D473–D479, Jan. 2010.

[278] R. L. Tatusov, E. V. Koonin, and D. J. Lipman, “A Genomic Perspective on Protein Families,” *Science*, vol. 278, pp. 631–637, Oct. 1997.

[279] J. Huerta-Cepas, D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork, “eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences,” *Nucleic Acids Research*, vol. 44, pp. D286–D293, Jan. 2016.

[280] 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, pp. 25–29, May 2000.

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

[282] R. D. Finn, T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H.-Y. Chang, Z. Dosztányi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Huang, X. Huang, I. Letunic, R. Lopez, S. Lu, A. Marchler-Bauer, H. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, I. Xenarios, L.-S. Yeh, S.-Y. Young, and A. L. Mitchell, "InterPro in 2017—beyond protein family and domain annotations," *Nucleic Acids Research*, vol. 45, pp. D190–D199, Jan. 2017.

[283] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, and S. Hunter, "InterProScan 5: Genome-scale protein function classification," *Bioinformatics*, vol. 30, pp. 1236–1240, May 2014.

[284] E. Drula, M.-L. Garron, S. Dogan, V. Lombard, B. Henrissat, and N. Terrapon, "The carbohydrate-active enzyme database: Functions and literature," *Nucleic Acids Research*, vol. 50, pp. D571–D577, Jan. 2022.

[285] M. M. M. Kuypers, H. K. Marchant, and B. Kartal, "The microbial nitrogen-cycling network," *Nature Reviews Microbiology*, vol. 16, pp. 263–276, May 2018.

[286] A. Duncan, "Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle," May 2020.

[287] K. Martin, *Bioinformatics Approaches for Assessing Microbial Communities in the Surface Ocean*. PhD thesis, University of East Anglia, 2018.

[288] K. Martin, K. Schmidt, A. Toseland, C. A. Boulton, K. Barry, B. Beszteri, C. P. D. Brussaard, A. Clum, C. G. Daum, E. Eloe-Fadrosh, A. Fong, B. Foster, B. Foster, M. Ginzburg, M. Huntemann, N. N. Ivanova, N. C. Kyrpides, E. Lindquist, S. Mukherjee, K. Palaniappan, T. B. K. Reddy, M. R. Rizkallah, S. Roux, K. Timmermans, S. G. Tringe, W. H. van de Poll, N. Varghese, K. U. Valentin, T. M. Lenton, I. V. Grigoriev, R. M. Leggett, V. Moulton, and T. Mock, "The biogeographic differentiation of algal microbiomes in the upper ocean from pole to pole," *Nature Communications*, vol. 12, p. 5483, Sept. 2021.

[289] "Moderate-resolution Imaging Spectroradiometer (MODIS) Aqua 11µm Day/Night Sea Surface Temperature Data; 2014 Reprocessing," tech. rep., NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group.

[290] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, H. Y. Katta, A. Mojica, I.-M. A. Chen, N. C. Kyrpides, and T. Reddy, "Genomes OnLine database (GOLD) v.7: Updates and new features," *Nucleic Acids Research*, vol. 47, pp. D649–D659, Aug. 2019.

[291] I.-M. A. Chen, K. Chu, K. Palaniappan, M. Pillay, A. Ratner, J. Huang, M. Huntemann, N. Varghese, J. R. White, R. Seshadri, T. Smirnova, E. Kirton, S. P. Jungbluth, T. Woyke, E. A. Eloe-Fadrosh, N. N. Ivanova, and N. C. Kyrpides, "IMG/M v.5.0: An

integrated data management and comparative analysis system for microbial genomes and microbiomes," *Nucleic Acids Research*, vol. 47, pp. D666–D677, Jan. 2019.

[292] B. Bushnell, "BBTools software package," *URL http://sourceforge. net/projects/bbmap*, 2014.

[293] L. Pireddu, S. Leo, and G. Zanetti, "SEAL: A distributed short read mapping and duplicate removal tool," *Bioinformatics*, vol. 27, pp. 2159–2160, Aug. 2011.

[294] A. V. Lukashin and M. Borodovsky, "GeneMark.hmm: New solutions for gene finding," *Nucleic Acids Research*, vol. 26, pp. 1107–1115, Feb. 1998.

[295] H. Noguchi, T. Taniguchi, and T. Itoh, "MetaGeneAnnotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes," *DNA Research*, vol. 15, pp. 387–396, Dec. 2008.

[296] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26, pp. 2460–2461, Oct. 2010.

[297] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," *Genome Biology*, vol. 20, p. 257, Nov. 2019.

[298] A. E. Darling, G. Jospin, E. Lowe, F. A. Matsen, H. M. Bik, and J. A. Eisen, "PhyloSift: Phylogenetic analysis of genomes and metagenomes," *PeerJ*, vol. 2, Jan. 2014.

[299] 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," *Nucleic Acids Research*, vol. 46, pp. D692–D699, Jan. 2018.

[300] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments," *PLOS ONE*, vol. 5, p. e9490, Mar. 2010.

[301] A. Stamatakis, "RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies," *Bioinformatics*, vol. 30, pp. 1312–1313, May 2014.

[302] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v4: Recent updates and new developments," *Nucleic Acids Research*, vol. 47, pp. W256–W259, July 2019.

[303] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, pp. 1792–1797, Mar. 2004.

[304] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses," *Bioinformatics*, vol. 25, pp. 1972–1973, Aug. 2009.

[305] D. H. Huson, B. Albrecht, C. Bağcı, I. Bessarab, A. Górska, D. Jolic, and R. B. H. Williams, "MEGAN-LR: New algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs," *Biology Direct*, vol. 13, p. 6, Apr. 2018.

[306] L. Pritchard, P. Cock, and Ö. Esen, "Pyani v0. 2.8: Average nucleotide identity (ANI) and related measures for whole genome comparisons," 2019.

[307] A. R. Quinlan, “BEDTools: The Swiss-Army Tool for Genome Feature Analysis,” *Current Protocols in Bioinformatics*, vol. 47, no. 1, pp. 11.12.1–11.12.34, 2014.

[308] C. Holt and M. Yandell, “MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects,” *BMC Bioinformatics*, vol. 12, p. 491, Dec. 2011.

[309] M. Steinegger and J. Söding, “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets,” *Nature Biotechnology*, vol. 35, pp. 1026–1028, Nov. 2017.

[310] R. D. Cook, “Cook’s Distance,” in *International Encyclopedia of Statistical Science* (M. Lovric, ed.), pp. 301–302, Berlin, Heidelberg: Springer, 2011.

[311] T. K. Mohanta and H. Bae, “The diversity of fungal genome,” *Biological Procedures Online*, vol. 17, p. 8, Apr. 2015.

[312] A. Toseland, S. J. Daines, J. R. Clark, A. Kirkham, J. Strauss, C. Uhlig, T. M. Lenton, K. Valentin, G. A. Pearson, V. Moulton, and T. Mock, “The impact of temperature on marine phytoplankton resource allocation and metabolism,” *Nature Climate Change*, vol. 3, pp. 979–984, Nov. 2013.

[313] A. Kuwata, K. Saitoh, Y. Nakamura, M. Ichinomiya, and N. Sato, “Draft Whole-Genome Sequence of Triparma laevis f. inornata (Parmales, Bolidophyceae), Isolated from the Oyashio Region, Western North Pacific Ocean,” *Microbiology Resource Announcements*, vol. 9, pp. e00367–20, Aug. 2020.

[314] E. Ivars-Martínez, G. D’auria, F. Rodríguez-Valera, C. Sánchez-Porro, A. Ventosa, I. Joint, and M. Mühling, “Biogeography of the ubiquitous marine bacterium *Alteromonas macleodii* determined by multilocus sequence analysis,” *Molecular Ecology*, vol. 17, no. 18, pp. 4092–4106, 2008.

[315] T. Belevich, L. Ilyash, I. Milyutina, A. Troitsky, and V. Sergeeva, “Picoplanktonic diatoms of Russian Arctic Seas revealed by metagenome analyze,” *Issues of modern algology*, pp. 105–110, Jan. 2019.

[316] T. A. Belevich, L. V. Ilyash, I. A. Milyutina, M. D. Logacheva, D. V. Goryunov, and A. V. Troitsky, “Photosynthetic Picoeukaryotes in the Land-Fast Ice of the White Sea, Russia,” *Microbial Ecology*, vol. 75, pp. 582–597, Apr. 2018.

[317] A. Kuwata, K. Yamada, M. Ichinomiya, S. Yoshikawa, M. Tragin, D. Vaulot, and A. Lopes dos Santos, “Bolidophyceae, a Sister Picoplanktonic Group of Diatoms – A Review,” *Frontiers in Marine Science*, vol. 5, 2018.

[318] M. Bar Dolev, I. Braslavsky, and P. L. Davies, “Ice-Binding Proteins and Their Function,” *Annual Review of Biochemistry*, vol. 85, no. 1, pp. 515–542, 2016.

[319] J. A. Raymond and D. Remias, “Ice-Binding Proteins in a Chrysophycean Snow Alga: Acquisition of an Essential Gene by Horizontal Gene Transfer,” *Frontiers in Microbiology*, vol. 10, 2019.

[320] J. Cloutier, D. Prévost, P. Nadeau, and H. Antoun, “Heat and cold shock protein synthesis in arctic and temperate strains of rhizobia.,” *Applied and Environmental Microbiology*, vol. 58, pp. 2846–2853, Sept. 1992.

[321] L. Jiao, J. Ran, X. Xu, and J. Wang, “Heat, acid and cold stresses enhance the expression of DnaK gene in Alicyclobacillus acidoterrestris,” *Food Research International*, vol. 67, pp. 183–192, Jan. 2015.

[322] 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, p. 00062, Feb. 2022.

[323] I. Luque, M. L. Riera-Alberola, A. Andújar, and J. A. G. Ochoa de Alda, “Intraphylum Diversity and Complex Evolution of Cyanobacterial Aminoacyl-tRNA Synthetases,” *Molecular Biology and Evolution*, vol. 25, pp. 2369–2389, Nov. 2008.

[324] B. Papudeshi, J. M. Haggerty, M. Doane, M. M. Morris, K. Walsh, D. T. Beattie, D. Pande, P. Zaeri, G. G. Z. Silva, F. Thompson, R. A. Edwards, and E. A. Dinsdale, “Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes,” *BMC Genomics*, vol. 18, p. 915, Nov. 2017.

[325] A. Z. Worden, J.-H. Lee, T. Mock, P. Rouzé, M. P. Simmons, A. L. Aerts, A. E. Allen, M. L. Cuvelier, E. Derelle, M. V. Everett, E. Foulon, J. Grimwood, H. Gundlach, B. Henrissat, C. Napoli, S. M. McDonald, M. S. Parker, S. Rombauts, A. Salamov, P. V. Dassow, J. H. Badger, P. M. Coutinho, E. Demir, I. Dubchak, C. Gentemann, W. Eikrem, J. E. Gready, U. John, W. Lanier, E. A. Lindquist, S. Lucas, K. F. X. Mayer, H. Moreau, F. Not, R. Otillar, O. Panaud, J. Pangilinan, I. Paulsen, B. Piegu, A. Poliakov, S. Robbens, J. Schmutz, E. Toulza, T. Wyss, A. Zelensky, K. Zhou, E. V. Armbrust, D. Bhattacharya, U. W. Goodenough, Y. V. de Peer, and I. V. Grigoriev, “Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas,” *Science*, vol. 324, pp. 268–272, Apr. 2009.

[326] D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson, “Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life,” *Nature Microbiology*, vol. 2, pp. 1533–1542, Nov. 2017.

[327] M. J. Behrenfeld, R. T. O'Malley, D. A. Siegel, C. R. McClain, J. L. Sarmiento, G. C. Feldman, A. J. Milligan, P. G. Falkowski, R. M. Letelier, and E. S. Boss, "Climate-driven trends in contemporary ocean productivity," *Nature*, vol. 444, pp. 752–755, Dec. 2006.

[328] M. Cornec, H. Claustre, A. Mignot, L. Guidi, L. Lacour, A. Poteau, F. D'Ortenzio, B. Gentili, and C. Schmechtig, "Deep Chlorophyll Maxima in the Global Ocean: Occurrences, Drivers and Characteristics," *Global Biogeochemical Cycles*, vol. 35, no. 4, p. e2020GB006759, 2021.

[329] S. A. Amin, M. S. Parker, and E. V. Armbrust, "Interactions between Diatoms and Bacteria," *Microbiology and Molecular Biology Reviews*, vol. 76, pp. 667–684, Sept. 2012.

[330] M. W. Lomas, J. A. Bonachela, S. A. Levin, and A. C. Martiny, "Impact of ocean phytoplankton diversity on phosphate uptake," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 17540–17545, Dec. 2014.

[331] T. J. Browning, E. P. Achterberg, J. C. Yong, I. Rapp, C. Utermann, A. Engel, and C. M. Moore, "Iron limitation of microbial phosphorus acquisition in the tropical North Atlantic," *Nature Communications*, vol. 8, p. 15465, May 2017.

[332] A. P. F. Pires, R. D. Guariento, T. Laque, F. A. Esteves, and V. F. Farjalla, "The negative effects of temperature increase on bacterial respiration are independent of changes in community composition," *Environmental Microbiology Reports*, vol. 6, no. 2, pp. 131–135, 2014.

[333] S. Louca, L. W. Parfrey, and M. Doebeli, "Decoupling function and taxonomy in the global ocean microbiome," *Science*, vol. 353, pp. 1272–1277, Sept. 2016.

[334] S. Louca, S. M. S. Jacques, A. P. F. Pires, J. S. Leal, D. S. Srivastava, L. W. Parfrey, V. F. Farjalla, and M. Doebeli, "High taxonomic variability despite stable functional structure across microbial communities," *Nature Ecology & Evolution*, vol. 1, pp. 1–12, Dec. 2016.

[335] V. J. Coles, M. R. Stukel, M. T. Brooks, A. Burd, B. C. Crump, M. A. Moran, J. H. Paul, B. M. Satinsky, P. L. Yager, B. L. Zielinski, and R. R. Hood, "Ocean biogeochemistry modeled with emergent trait-based genomics," *Science*, vol. 358, pp. 1149–1154, Dec. 2017.

[336] P. E. Larsen, F. R. Collart, D. Field, F. Meyer, K. P. Keegan, C. S. Henry, J. McGrath, J. Quinn, and J. A. Gilbert, "Predicted Relative Metabolomic Turnover (PRMT): Determining metabolic turnover from a coastal marine metagenomic dataset," *Microbial Informatics and Experimentation*, vol. 1, p. 4, June 2011.

[337] E. Faure, S.-D. Ayata, and L. Bittner, "Towards omics-based predictions of planktonic functional composition from environmental data," *Nature Communications*, vol. 12, p. 4361, July 2021.

[338] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, p. 559, Dec. 2008.

[339] J. Raes, I. Letunic, T. Yamada, L. J. Jensen, and P. Bork, "Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data," *Molecular Systems Biology*, vol. 7, p. 473, Jan. 2011.

[340] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V. Ward, J. A. Reyes, S. A. Shah, N. LeLeiko, S. B. Snapper, A. Bousvaros, J. Korzenik, B. E. Sands, R. J. Xavier, and C. Huttenhower, "Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment," *Genome Biology*, vol. 13, p. R79, Apr. 2012.

[341] N. Mantel, "The Detection of Disease Clustering and a Generalized Regression Approach," *Cancer Research*, vol. 27, pp. 209–220, Feb. 1967.

[342] P. E. Smouse, J. C. Long, and R. R. Sokal, "Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence," *Systematic Zoology*, vol. 35, no. 4, pp. 627–632, 1986.

[343] O. Paliy and V. Shankar, "Application of multivariate statistical techniques in microbial ecology," *Molecular Ecology*, vol. 25, no. 5, pp. 1032–1057, 2016.

[344] S. Ciucci, Y. Ge, C. Durán, A. Palladini, V. Jiménez-Jiménez, L. M. Martínez-Sánchez, Y. Wang, S. Sales, A. Shevchenko, S. W. Poser, M. Herbig, O. Otto, A. Androutsellis-Theotokis, J. Guck, M. J. Gerl, and C. V. Cannistraci, "Enlightening discriminative network functional modules behind Principal Component Analysis separation in differential-omic science studies," *Scientific Reports*, vol. 7, p. 43946, Mar. 2017.

[345] C. J. F. ter Braak, "Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis," *Ecology*, vol. 67, no. 5, pp. 1167–1179, 1986.

[346] P. Legendre and L. Legendre, "Chapter 11 - Canonical analysis," in *Developments in Environmental Modelling* (P. Legendre and L. Legendre, eds.), vol. 24 of *Numerical Ecology*, pp. 625–710, Elsevier, Jan. 2012.

[347] N. C. Kenkel and L. Orloci, "Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies: Some New Results," *Ecology*, vol. 67, no. 4, pp. 919–928, 1986.

[348] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[349] R. Gove, L. Cadalzo, N. Leiby, J. M. Singer, and A. Zaitzeff, "New guidance for using t-SNE: Alternative defaults, hyperparameter selection automation, and comparative evaluation," *Visual Informatics*, vol. 6, pp. 87–97, June 2022.

[350] D. Jiang, C. R. Armour, C. Hu, M. Mei, C. Tian, T. J. Sharpton, and Y. Jiang, "Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities," *Frontiers in Genetics*, vol. 10, 2019.

[351] J. Friedman and E. J. Alm, "Inferring Correlation Networks from Genomic Survey Data," *PLOS Computational Biology*, vol. 8, p. e1002687, Sept. 2012.

[352] J. Schäfer and K. Strimmer, "An empirical Bayes approach to inferring large-scale gene association networks," *Bioinformatics*, vol. 21, pp. 754–764, Mar. 2005.

[353] A. Cougoul, X. Bailly, and E. C. Wit, "MAGMA: Inference of sparse microbial association networks," Feb. 2019.

[354] B. Zhang and S. Horvath, "A General Framework for Weighted Gene Co-Expression Network Analysis," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, Aug. 2005.

[355] A. Ghazalpour, S. Doss, B. Zhang, S. Wang, C. Plaisier, R. Castellanos, A. Brozell, E. E. Schadt, T. A. Drake, A. J. Lusis, and S. Horvath, "Integrating genetic and network analysis to characterize genes related to mouse weight," *PLoS genetics*, vol. 2, p. e130, Aug. 2006.

[356] J. M. Wilson, S. Y. Litvin, and J. M. Beman, "Microbial community networks associated with variations in community respiration rates during upwelling in nearshore Monterey Bay, California," *Environmental Microbiology Reports*, vol. 10, no. 3, pp. 272–282, 2018.

[357] J. Wang, J. Zhang, W. Liu, H. Zhang, and Z. Sun, "Metagenomic and metatranscriptomic profiling of Lactobacillus casei Zhang in the human gut," *npj Biofilms and Microbiomes*, vol. 7, pp. 1–10, July 2021.

[358] L. Fu and E. Medico, "FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data," *BMC Bioinformatics*, vol. 8, p. 3, Jan. 2007.

[359] W. Saelens, R. Cannoodt, and Y. Saeys, "A comprehensive evaluation of module detection methods for gene expression data," *Nature Communications*, vol. 9, p. 1090, Mar. 2018.

[360] S. Rogers, M. Girolami, C. Campbell, and R. Breitling, "The latent process decomposition of cDNA microarray data sets," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, pp. 143–156, Apr. 2005.

[361] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, p. 1608, Sept. 2016.

[362] B.-A. Luca, V. Moulton, C. Ellis, D. R. Edwards, C. Campbell, R. A. Cooper, J. Clark, D. S. Brewer, and C. S. Cooper, "A novel stratification framework for predicting outcome in patients with prostate cancer," *British Journal of Cancer*, vol. 122, pp. 1467–1476, May 2020.

[363] H. Lilja, D. Ulmert, and A. J. Vickers, "Prostate-specific antigen and prostate cancer: Prediction, detection and monitoring," *Nature Reviews Cancer*, vol. 8, pp. 268–278, Apr. 2008.

[364] N. Sompairac, P. V. Nazarov, U. Czerwinska, L. Cantini, A. Biton, A. Molkenov, Z. Zhumadilov, E. Barillot, F. Radvanyi, A. Gorban, U. Kairov, and A. Zinovyev, "Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets," *International Journal of Molecular Sciences*, vol. 20, p. 4414, Jan. 2019.

[365] G. P. Way, M. Zietz, V. Rubinetti, D. S. Himmelstein, and C. S. Greene, “Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations,” *Genome Biology*, vol. 21, p. 109, May 2020.

[366] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[367] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 4164–4169, Mar. 2004.

[368] X. Jiang, M. G. I. Langille, R. Y. Neches, M. Elliot, S. A. Levin, J. A. Eisen, J. S. Weitz, and J. Dushoff, “Functional Biogeography of Ocean Microbes Revealed through Non-Negative Matrix Factorization,” *PLOS ONE*, vol. 7, p. e43866, Sept. 2012.

[369] X. Jiang, J. S. Weitz, and J. Dushoff, “A non-negative matrix factorization framework for identifying modular patterns in metagenomic profile data,” *Journal of Mathematical Biology*, vol. 64, pp. 697–711, Mar. 2012.

[370] S. Raguideau, S. Plancade, N. Pons, M. Leclerc, and B. Laroche, “Inferring Aggregated Functional Traits from Metagenomic Data Using Constrained Non-negative Matrix Factorization: Application to Fiber Degradation in the Human Gut Microbiota,” *PLOS Computational Biology*, vol. 12, p. e1005252, Dec. 2016.

[371] Y. Cai, H. Gu, and T. Kenney, “Learning Microbial Community Structures with Supervised and Unsupervised Non-negative Matrix Factorization,” *Microbiome*, vol. 5, p. 110, Aug. 2017.

[372] L. Muzzarelli, S. Weis, S. B. Eickhoff, and K. R. Patil, “Rank Selection in Non-negative Matrix Factorization: Systematic comparison and a new MAD metric,” in *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, July 2019.

[373] “Apduncan/nmf_package: Set of tools for performing Non-Negative Matrix Factorisation on functional annotations of (mainly ocean) metagenome data.” https://github.com/apduncan/nmf_package.

[374] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[375] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the beta-divergence,” Mar. 2011.

[376] A. Cichocki and A.-H. Phan, “Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations,” *IEICE Transactions*, vol. 92-A, pp. 708–721, Mar. 2009.

[377] L. T. K. Hien and N. Gillis, “Algorithms for Nonnegative Matrix Factorization with the Kullback–Leibler Divergence,” *Journal of Scientific Computing*, vol. 87, p. 93, May 2021.

[378] P. M. Kim and B. Tidor, “Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data,” *Genome Research*, vol. 13, pp. 1706–1718, Jan. 2003.

[379] H. Kim and H. Park, “Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis,” *Bioinformatics*, vol. 23, pp. 1495–1502, June 2007.

[380] H. W. Kuhn, “The Hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[381] W. M. Rand, “Objective Criteria for the Evaluation of Clustering Methods,” *Journal of the American Statistical Association*, vol. 66, pp. 846–850, Dec. 1971.

[382] L. Hubert and P. Arabie, “Comparing partitions,” *Journal of Classification*, vol. 2, pp. 193–218, Dec. 1985.

[383] A. Frigyesi and M. Höglund, “Non-Negative Matrix Factorization for the Analysis of Complex Gene Expression Data: Identification of Clinically Relevant Tumor Subtypes,” *Cancer Informatics*, vol. 6, pp. 275–292, May 2008.

[384] V. D. Blondel, N.-d. Ho, and P. van Dooren, “Weighted nonnegative matrix factorization and face feature extraction,” 2007.

[385] S. Nanda, “wNMF: Weighted Non-Negative Matrix Factorization,” July 2022.

[386] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, and P. van Mulbregt, “SciPy 1.0: Fundamental algorithms for scientific computing in Python,” *Nature Methods*, vol. 17, pp. 261–272, Mar. 2020.

[387] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, Mar. 2015.

[388] K. Eren, M. Deveci, O. Küçüktunç, and Ü. V. Çatalyürek, “A comparative analysis of biclustering algorithms for gene expression data,” *Briefings in Bioinformatics*, vol. 14, pp. 279–292, May 2013.

[389] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 15545–15550, Oct. 2005.

[390] Z. Fang, “GSEApy,” Oct. 2022.

[391] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstråle, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genetics*, vol. 34, pp. 267–273, July 2003.

[392] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola, "Fast optimal leaf ordering for hierarchical clustering," *Bioinformatics*, vol. 17, pp. S22–S29, June 2001.

[393] S. R. Maetschke, K. S. Kassahn, J. A. Dunn, S.-P. Han, E. Z. Curley, K. J. Stacey, and M. A. Ragan, "A visual framework for sequence analysis using n-grams and spectral rearrangement," *Bioinformatics*, vol. 26, pp. 737–744, Mar. 2010.

[394] D. Richter, R. Watteaux, T. Vannier, J. Leconte, P. Frémont, G. Reygondeau, N. Maillet, N. Henry, G. Benoit, A. Fernandez-Guerra, S. Suweis, R. Narci, C. Berney, D. Eveillard, F. Gavory, L. Guidi, K. Labadie, E. Mahieu, J. Poulain, S. Romac, S. Roux, C. Dimier, S. Kandels, M. Picheral, S. Searson, S. Pesant, J.-M. Aury, J. Brum, C. Lemaitre, E. Pelletier, P. Bork, S. Sunagawa, L. Karp-Boss, C. Bowler, M. Sullivan, E. Karsenti, M. Mariadassou, I. Probert, P. Peterlongo, P. Wincker, C. de Vargas, M. R. D&apos, alcalà, D. Iudicone, and O. Jaillon, "Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems," Feb. 2020.

[395] D. J. Richter, R. Watteaux, T. Vannier, J. Leconte, P. Frémont, G. Reygondeau, N. Maillet, N. Henry, G. Benoit, O. Da Silva, T. O. Delmont, A. Fernàndez-Guerra, S. Suweis, R. Narci, C. Berney, D. Eveillard, F. Gavory, L. Guidi, K. Labadie, E. Mahieu, J. Poulain, S. Romac, S. Roux, C. Dimier, S. Kandels, M. Picheral, S. Searson, Tara Oceans Coordinators, S. Pesant, J.-M. Aury, J. R. Brum, C. Lemaitre, E. Pelletier, P. Bork, S. Sunagawa, F. Lombard, L. Karp-Boss, C. Bowler, M. B. Sullivan, E. Karsenti, M. Mariadassou, I. Probert, P. Peterlongo, P. Wincker, C. de Vargas, M. Ribera d'Alcalà, D. Iudicone, and O. Jaillon, "Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems," *eLife*, vol. 11, p. e78129, Aug. 2022.

[396] A. Fritz, P. Hofmann, S. Majda, E. Dahms, J. Dröge, J. Fiedler, T. R. Lesker, P. Belmann, M. Z. DeMaere, A. E. Darling, A. Sczyrba, A. Bremges, and A. C. McHardy, "CAMISIM: Simulating metagenomes and microbial communities," *Microbiome*, vol. 7, p. 17, Feb. 2019.

[397] W. Huang, L. Li, J. R. Myers, and G. T. Marth, "ART: A next-generation sequencing read simulator," *Bioinformatics*, vol. 28, pp. 593–594, Feb. 2012.

[398] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "Fastp: An ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, pp. i884–i890, Sept. 2018.

[399] T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, and H. Ogata, "KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold," *Bioinformatics*, vol. 36, pp. 2251–2252, Apr. 2020.

[400] L. M. Proctor, H. H. Creasy, J. M. Fettweis, J. Lloyd-Price, A. Mahurkar, W. Zhou, G. A. Buck, M. P. Snyder, J. F. Strauss, G. M. Weinstock, O. White, C. Huttenhower,

and The Integrative HMP (iHMP) Research Network Consortium, “The Integrative Human Microbiome Project,” *Nature*, vol. 569, pp. 641–648, May 2019.

[401] X. C. Morgan, N. Segata, and C. Huttenhower, “Biodiversity and functional genomics in the human microbiome,” *Trends in Genetics*, vol. 29, pp. 51–58, Jan. 2013.

[402] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang, “A human gut microbial gene catalogue established by metagenomic sequencing,” *Nature*, vol. 464, pp. 59–65, Mar. 2010.

[403] X. Jiang, X. Hu, and W. Xu, “Joint Analysis of Functional and Phylogenetic Composition for Human Microbiome Data,” in *Bioinformatics Research and Applications* (M. Basu, Y. Pan, and J. Wang, eds.), Lecture Notes in Computer Science, (Cham), pp. 346–356, Springer International Publishing, 2014.

[404] S. Roldán, D. Herrera, and M. Sanz, “Biofilms and the tongue: Therapeutical approaches for the control of halitosis,” *Clinical Oral Investigations*, vol. 7, pp. 189–197, Dec. 2003.

[405] T. Grevesse, C. Guéguen, V. E. Onana, and D. A. Walsh, “Degradation pathways for organic matter of terrestrial origin are widespread and expressed in Arctic Ocean microbiomes,” *Microbiome*, vol. 10, p. 237, Dec. 2022.

[406] M. Pechenizkiy, A. Tsymbal, and S. Puuronen, “PCA-based feature transformation for classification: Issues in medical diagnostics,” in *Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems*, pp. 535–540, June 2004.

[407] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, “Deep learning and process understanding for data-driven Earth system science,” *Nature*, vol. 566, pp. 195–204, Feb. 2019.

[408] T. H. Nguyen, E. Prifti, Y. Chevaleyre, N. Sokolovska, and J.-D. Zucker, “Disease Classification in Metagenomics with 2D Embeddings and Deep Learning,” June 2018.

[409] P. Padilla, M. Lopez, J. M. Gorriz, J. Ramirez, D. Salas-Gonzalez, and I. Alvarez, “NMF-SVM Based CAD Tool Applied to Functional Brain Images for the Diagnosis of Alzheimer’s Disease,” *IEEE Transactions on Medical Imaging*, vol. 31, pp. 207–216, Feb. 2012.

[410] T. Mock, S. J. Daines, R. Geider, S. Collins, M. Metodiev, A. J. Millar, V. Moulton, and T. M. Lenton, “Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes,” *Global Change Biology*, vol. 22, no. 1, pp. 61–75, 2016.

[411] T. Mock, “100 Diatom Genomes.” https://jgi.doe.gov/csp-2021-100-diatom-genomes/, Oct. 2020.

[412] J. Vollmers, S. Wiegand, F. Lenk, and A.-K. Kaster, "How clear is our current view on microbial dark matter? (Re-)assessing public MAG & SAG datasets with MDM-cleaner," *Nucleic Acids Research*, vol. 50, p. e76, July 2022.

[413] A. Shaiber and A. M. Eren, "Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories," *mBio*, vol. 10, June 2019.

[414] L. J. Pollock, R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy, "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)," *Methods in Ecology and Evolution*, vol. 5, no. 5, pp. 397–406, 2014.

[415] J. L. Weissman, E.-R. O. Dimbo, A. I. Krinos, C. Neely, Y. Yagües, D. Nolin, S. Hou, S. Laperriere, D. A. Caron, B. Tully, H. Alexander, and J. A. Fuhrman, "Estimating global variation in the maximum growth rates of eukaryotic microbes from cultures and metagenomes via codon usage patterns," Oct. 2022.

[416] C. Nef, M.-A. Madoui, É. Pelletier, and C. Bowler, "Whole-genome scanning reveals environmental selection mechanisms that shape diversity in populations of the epipelagic diatom Chaetoceros," *PLOS Biology*, vol. 20, p. e3001893, Nov. 2022.

[417] A. Diana, E. Matechou, J. Griffin, D. Yu, M. Luo, M. Tosa, A. Bush, and R. Griffiths, "eDNAPlus: A unifying modelling framework for DNA-based biodiversity monitoring," Nov. 2022.

[418] A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R. Ruiz, J. Schrittwieser, G. Swirszcz, D. Silver, D. Hassabis, and P. Kohli, "Discovering faster matrix multiplication algorithms with reinforcement learning," *Nature*, vol. 610, pp. 47–53, Oct. 2022.

[419] S. M. Gibbons, J. G. Caporaso, M. Pirrung, D. Field, R. Knight, and J. A. Gilbert, "Evidence for a persistent microbial seed bank throughout the global ocean," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 4651–4655, Mar. 2013.

[420] M. Wietz, C. Bienhold, K. Metfies, S. Torres-Valdés, W.-J. von Appen, I. Salter, and A. Boetius, "The polar night shift: Seasonal dynamics and drivers of Arctic Ocean microbiomes revealed by autonomous sampling," *ISME Communications*, vol. 1, pp. 1–12, Dec. 2021.

[421] 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," *Nature Microbiology*, vol. 6, pp. 1561–1574, Dec. 2021.

[422] T. M. Lenton, "Arctic Climate Tipping Points," *AMBIO*, vol. 41, pp. 10–22, Feb. 2012.

Appendix A

Appendices for Chapter 4

A.1 Sample Identifiers

Sample	Latitude	Longitude	Sample Date	JGI Project ID	IMG Taxon ID Megahit	IMG Taxon ID metaSpades	NCBI Project ID	SRA ID	Depth metres	Temperature °C	Salinity PSU	$\frac{Nitrate}{Nitrite}$	Phosphate	Silicate
P1	79.0225	-9.52472	06/07/2012	1102218	3300009432	3300027849	365111	SRP111696	17	-1.0337	31.0274	0	0.47	2.48
P2	78.86694	-3.22861	09/07/2012	1102222	3300009441	3300027810	365113	SRP111701	15	-0.8805	32.3454	3.3	0.79	5.47
P3a	78.86972	8.11222	20/06/2012	1102220	3300009544	3300027883	365112	SRP111694	10	5.269	35.0693	0.51	0.49	2.75
P3b	78.8697	8.1122	20/06/2012	1021520	3300002154		330320	SRP080437	10	5.269	35.0693	0.51	0.49	2.75
P4	73.01889	9.85667	18/06/2012	1021526	3300002153		366135	SRP099322	20	6.0186	35.1528	6.55	0.88	3.9
P5	71.20083	8.86667	18/06/2012	1021523	3300002186		366134	SRP099317	10	7.1834	35.1344	4.83	0.77	3.62
P6	69.23028	7.73028	18/06/2012	1102224	3300009436	3300027833	365114	SRP111700	10	9.0976	34.7321	2.46	0.46	1.52
NP1	34.876	-13.1352	04/11/2012	1125692	3300012953		406185	SRP129762	80	21.78	36.65	0.269427308	0	0.292683967
NP2	26.049	-17.4585	06/11/2012	1125694	3300012952		406186	SRP129813	80	24.62	36.9	0.516096977	0.066033049	0.405116155
NP3	15.249	-20.515	09/11/2012	1102230	3300009593	3300027906	365117	SRP111708	55	28.6	35.64	16.42369279	0.682339671	3.367016894
NP4	2.405	-13.602	13/11/2012	1102232	3300009790		365118	SRP098326	80	27.1	35.61	6.458482733	0.3717252	2.064853304
NP5	-17.283	2.9768	20/11/2012	1102234	3300009550	3300027859	365119	SRP111718	30	18.82	35.97	0	0.19	2.63

Table A.1 Sample identifiers and metadata for metagenome samples

A.2 Assembly Summary

Sample	Assembler	Contigs	Total	N50	L50	Max Contig	Scaffolds	% in scaf.
			Gbp			Kbp	>50Kbp	>50Kbp
P2	MEGAHIT	3059358	1.943396	724724	669	127.604	95	0.32
P2	SPADES	1933411	1.145116305	364784	658	524.554	118	0.9
P3a	MEGAHIT	4536098	3.120726	1076012	716	147.864	59	0.14
P3a	SPADES	3353304	1.983904784	703210	645	311.607	92	0.43
P1	MEGAHIT	3524824	2.504409	805269	757	991.905	149	0.53
P1	SPADES	2557771	1.590622916	494293	686	1057.628	243	1.55
NP2	MEGAHIT	5501464	3.015199	1486442	566	1107	66	0.25
NP1	MEGAHIT	6177919	3.515052	1609672	588	899.176	139	0.43
P6	MEGAHIT	3448017	2.345505	785850	697	259.275	101	0.32
P6	SPADES	2221927	1.423805502	377529	748	310.762	264	1.44
P5	MEGAHIT	1034413	0.562281	206977	567	115.416	36	0.43
P4	MEGAHIT	879176	0.50267	149007	624	95.796	22	0.28
NP5	MEGAHIT	3387596	2.075333	921231	615	517.097	129	0.6
NP5	SPADES	2895169	1.473653634	727696	497	517.111	145	0.89
P3b	MEGAHIT	791494	0.391017	186459	497	101.069	5	0.27
NP3	MEGAHIT	5072980	3.145297	1364176	620	497.503	239	0.73
NP3	SPADES	5181937	2.618668657	1364985	490	576.479	199	0.75
NP4	MEGAHIT	4681656	2.917293	1268050	623	360.328	98	0.31

Table A.2 Summary statistics for metagenomic assemblies

A.3 Prokaryotic MAG Summary

MAG ID	IMG Bin ID	Quality	GTDB-Tk Lineage	CheckM Lineage	Completeness	Contamination	Bases	Genes	Contigs
NP4_26P	3300009790_26	HQ	Bacteria; Actinobacteriota; Acidimicrobiia; Microtrichales; TK06; MedAcidi-G3; GCA_002434645.1	Bacteria ; Actinobacteria	97.01	2.99	2255374	2325	66
NP4_41P	3300009790_41	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Legionellales; Legionellaceae; None; None		93.6	1.16	1580637	1615	83
NP4_8P	3300009790_8	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Pelagibaca; Pelagibaca bermudensis	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae ; Pelagibaca ; Pelagibaca bermudensis	92.47	0.45	4518587	4559	58
NP4_9P	3300009790_9	HQ	Bacteria; Myxococcota; UBA796; UBA796; UBA796; UBA796; None	Bacteria	91.68	2.02	4258701	3898	237
NP4_16P	3300009790_16	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Verrucomicrobiales; Akkermansiaceae; Roseibacillus; None	Bacteria	89.12	1.02	3501992	2981	397
NP4_10P	3300009790_10	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Alteromonas; Alteromonas macleodii	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Alteromonadaceae ; Alteromonas ; Alteromonas macleodii	86.58	0.79	4200884	3950	367
NP4_11P	3300009790_11	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; Croceibacter; Croceibacter atlanticus		82.6	3.92	3873269	4431	569
NP4_33P	3300009790_33	MQ	Bacteria; SAR324; SAR324; SAR324; NAC60-12; Arctic96AD-7; Arctic96AD-7 sp4	Bacteria ; Proteobacteria	63.17	0	1763599	1742	282
NP4_40P	3300009790_40	MQ	Archaea; Thermoplasmatota; MGII; MGII; MGIIA; UBA562; UBA8160		59.87	0.8	1519318	1415	201
NP4_61P	3300009790_61	MQ	Bacteria; Patescibacteria; Paceibacteria; UBA9983; Kaiserbacteraceae; OLB19; None	Bacteria	59.51	0.99	816027	956	91
NP4_47P	3300009790_47	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Pedosphaerales; UBA1100; UBA1100; None	Bacteria	52.69	0.72	1263197	1334	257
NP4_22P	3300009790_22	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; GCA-002705045; GCA_002725175.1	Bacteria ; Proteobacteria ; Alphaproteobacteria	50.67	8.68	2021848	2219	370
NP4_18P	3300009790_18	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Alteromonas; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Alteromonadaceae ; Alteromonas ; Alteromonas sp. SN2	50.3	0.76	2850860	2796	399
P3b_2P	3300002154_2	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Sulfitobacter_C; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae	94.83	2.42	2928419	3272	207
P3b_6P	3300002154_6	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Cryomorphaceae; UBA10364; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales	92.13	2.31	1754849	1821	221
P3b_3P	3300002154_3	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; ASP10-02a sp1	Bacteria ; Proteobacteria ; Gammaproteobacteria	92	2.65	2565387	2664	210
P3b_5P	3300002154_5	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; unclassified ; unclassified ; unclassified ; gamma proteobacterium HTCC2207	90.48	3.8	2226673	2286	210
P3b_8P	3300002154_8	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; HC6-5; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales	78.33	2.28	1579624	1674	235
NP5_10P	3300027859_10	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Alcanivoracaceae; Alcanivorax; GCA_002726155.1	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; Alcanivoracaceae ; Alcanivorax ; Alcanivorax sp. DG881	93.59	2.48	3336769	3400	304

NP5_15P	3300027859_15	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; UA16; UBA11663; None	Bacteria ; Bacteroidetes	92.74	0.54	1996931	1784	134
NP5_12P	3300027859_12	HQ	Bacteria; Proteobacteria; Alphaproteobacteria; Micavibrionales; Micavibrionaceae; UBA2705; None	Bacteria ; Proteobacteria ; Alphaproteobacteria	91.89	3.7	2436683	2376	38
NP5_7P	3300027859_7	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Verrucomicrobiales; Akkermansiaceae; SW10; None	Bacteria	90.71	0.94	3734917	3322	409
NP5_9P	3300027859_9	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; Neptunomonas; Neptunomonas phycophila	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; Oceanospirillaceae	89.39	0.21	3510332	3328	46
NP5_11P	3300027859_11	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; UA16; UBA8752; None	Bacteria ; Bacteroidetes	88.71	7.65	2557483	2279	281
NP5_3P	3300027859_3	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Alteromonas; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Alteromonadaceae ; Alteromonas ; Alteromonas naphthalenivorans	71.55	1.72	4479979	4249	351
NP5_8P	3300027859_8	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Pelagibaca; Pelagibaca bermudensis	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae ; Pelagibaca ; Pelagibaca bermudensis	60.38	1.91	3663361	4053	643
NP5_29P	3300027859_29	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; UBA3537; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae ; Formosa ; Formosa sp. Hel3_A1_48	55.58	4.7	1088070	1148	156
NP5_26P	3300027859_26	MQ	Archaea; Thermoplasmatota; MGII; MGII; MGIIB; UBA11751; GCA_002504845.1	Archaea ; Euryarchaeota ; unclassified ; unclassified ; unclassified ; unclassified ; uncultured marine group II euryarchaeote	54	0	1047207	1000	85
NP5_19P	3300027859_19	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Halomonadaceae; Halomonas; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; Halomonadaceae ; Halomonas	53.45	0	1548523	1705	292
NP1_13P	3300012953_13	MQ	Bacteria; Planctomycetota; UBA1135; UBA2386; UBA2386; GCA-2684655; None	Bacteria	95.83	1.14	3699837	3051	48
NP1_38P	3300012953_38	MQ	Bacteria; Planctomycetota; Phycisphaerae; Phycisphaerales; SM1A02; GCA-002718515; None	Bacteria	90.91	0	1686369	1589	70
NP1_9P	3300012953_9	MQ	Bacteria; Myxococcota; UBA796; UBA796; None; None; None	Bacteria	87.11	0.05	4491777	4197	460
NP1_22P	3300012953_22	MQ	Bacteria; Actinobacteriota; Acidimicrobiia; Microtrichales; TK06; MedAcidi-G3; GCA_000817105.1	Bacteria ; Actinobacteria	85.47	2.14	2181758	2305	159
NP1_5P	3300012953_5	MQ	Bacteria; Planctomycetota; Planctomycetes; Pirellulales; Pirellulaceae; GCA-2723275; None	Bacteria ; Planctomycetes ; Planctomycetia ; Planctomycetales ; Planctomycetaceae	85	6.58	6056143	5124	760
NP1_17P	3300012953_17	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Pedosphaerales; AAA164-E04; AAA164-E04; AAA164-E04 sp1	Bacteria	79.31	5.39	3278878	2892	313
NP1_11P	3300012953_11	MQ	Bacteria; Myxococcota; UBA796; UBA9615; UBA9615; UBA6601; None	Bacteria	77.11	2.8	4231802	3889	628
NP1_14P	3300012953_14	MQ	Bacteria; Myxococcota; UBA796; UBA796; None; None; None	Bacteria	76.74	1.68	3563224	3355	340
NP1_23P	3300012953_23	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; Croceibacter; Croceibacter atlanticus		69.39	0.68	2188858	2085	26
NP1_31P	3300012953_31	MQ	Archaea; Thermoplasmatota; MGII; MGII; MGIIB; UBA496; GCA_002713585.1	Archaea ; Euryarchaeota ; unclassified ; unclassified ; unclassified ; unclassified	63.69	8.8	1722331	1525	163
NP1_19P	3300012953_19	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Opitutales; Opitutaceae; UBA5691; GCA_002420265.1	Bacteria	61.33	3.11	2630549	2610	488
NP1_39P	3300012953_39	MQ	Archaea; Thermoplasmatota; MGII; MGII; MGIIB; UBA501; GCA_002701965.1	Archaea ; Euryarchaeota ; unclassified ; unclassified ; unclassified ; unclassified	60	0	1443393	1290	143

NP2_8P	3300012952_8	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Verrucomicrobiales; Akkermansiaceae; Roseibacillus; None	Bacteria	95.92	0.52	4113173	3503	384
NP2_25P	3300012952_25	MQ	Bacteria; Actinobacteriota; Acidimicrobiia; Microtrichales; MedAcidi-G1; MedAcidi-G1; GCA_002697965.1	Bacteria ; Actinobacteria ; Acidimicrobiia ; unclassified ; unclassified ; unclassified	76.92	7.26	1645047	1910	241
NP2_18P	3300012952_18	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Hyphomonadaceae; Hyphomonas; GCF_000682775.1	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Hyphomonadaceae ; Hyphomonas ; Hyphomonas sp. L-53-1-40	75.16	0.32	2405472	2379	9
NP2_10P	3300012952_10	MQ	Bacteria; Planctomycetota; Planctomycetes; Pirellulales; Pirellulaceae; UBA11883; UBA11883 sp1	Bacteria ; Planctomycetes ; Planctomycetia ; Planctomycetales ; Planctomycetaceae	73.52	1.28	3567095	3098	568
NP2_9P	3300012952_9	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Pseudoalteromonas; Pseudoalteromonas marina	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Pseudoalteromonadaceae ; Pseudoalteromonas	72.49	1.72	3894124	3750	239
NP2_11P	3300012952_11	MQ	Bacteria; Myxococcota; UBA796; UBA796; UBA796; None; None	Bacteria	69.69	2.58	3471424	3434	524
NP2_41P	3300012952_41	MQ	Bacteria; UBP7; UBA6624; UBA6624; UBA6624; UBA6624; GCA_002501535.1	Bacteria	68.18	1.72	879918	1021	135
NP2_14P	3300012952_14	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; Croceibacter; Croceibacter atlanticus	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae ; Croceibacter ; Croceibacter atlanticus	63.54	2.2	2493551	2801	349
NP2_12P	3300012952_12	MQ	Bacteria; Poribacteria; WGA-4E; WGA-4E; UBA9662; TMED15; GCA_002714785.1	Bacteria	61.5	7.09	3238826	3177	615
NP2_50P	3300012952_50	MQ	Archaea; Crenarchaeota; Nitrososphaeria; Nitrososphaerales; Nitrosopumilaceae; Nitrosopumilus; None	Archaea ; Thaumarchaeota ; unclassified ; Nitrosopumilales ; Nitrosopumilaceae ; Candidatus Nitrosopumilus	54.98	5.99	705983	951	115
NP2_13P	3300012952_13	MQ	Bacteria; Myxococcota; UBA796; UBA796; UBA796; GCA-2683315; None		53.17	0.28	2951437	2977	596
NP2_26P	3300012952_26	MQ	Bacteria; Actinobacteriota; Acidimicrobiia; Microtrichales; UBA11606; UBA8592; None	Bacteria ; Actinobacteria	50.29	8.97	1619219	1990	302
NP3_5P	3300027906_5	HQ	Bacteria; Proteobacteria; Gammaproteobacteria; Nevskiales; Algiphilaceae; None; None	Bacteria ; Proteobacteria ; Gammaproteobacteria	98.91	3.61	4558824	4269	66
NP3_6P	3300027906_6	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Alteromonas; Alteromonas macleodii	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Alteromonadaceae ; Alteromonas ; Alteromonas macleodii	98.8	3.14	4479563	3996	118
NP3_20P	3300027906_20	HQ	Bacteria; Proteobacteria; Alphaproteobacteria; Caulobacterales; Caulobacteraceae; Brevundimonas; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Caulobacterales ; Caulobacteraceae ; Brevundimonas	97.08	3.55	2609799	2813	152
NP3_7P	3300027906_7	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Alcanivoracaceae; Alcanivorax; GCA_002726155.1	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; Alcanivoracaceae ; Alcanivorax ; Alcanivorax sp. DG881	96.65	8.62	4122650	4064	208
NP3_4P	3300027906_4	HQ	Bacteria; Planctomycetota; Planctomycetes; Pirellulales; Pirellulaceae; UBA721; None	Bacteria ; Planctomycetes ; Planctomycetia ; Planctomycetales ; Planctomycetaceae	92.87	2.46	4467928	3738	344
NP3_10P	3300027906_10	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Verrucomicrobiales; Akkermansiaceae; Roseibacillus; None	Bacteria	84.67	0	3495999	3044	432
NP3_13P	3300027906_13	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas_D; Pseudomonas_D sabulinigri	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Pseudomonadales ; Pseudomonadaceae ; Pseudomonas ; Pseudomonas sabulinigri	81.47	1.72	3387632	3226	25
NP3_25P	3300027906_25	MQ	Bacteria; Actinobacteriota; Acidimicrobiia; Microtrichales; TK06; MedAcidi-G3; GCA_002434645.1	Bacteria ; Actinobacteria	78.4	2.14	1817413	1920	144
NP3_14P	3300027906_14	MQ	Bacteria; Myxococcota; UBA796; UBA796; None; None; None	Bacteria	77.68	3.85	3341013	3416	471

NP3_22P	3300027906_22	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Sphingomonadales; Sphingomonadaceae; Erythrobacter_A; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Sphingomonadales ; Erythrobacteraceae ; Erythrobacter	75.17	7.25	2238345	2522	401
NP3_46P	3300027906_46	MQ	Bacteria; Patescibacteria; Saccharimonadia; Saccharimonadales; Saccharimonadaceae; UBA10027; None	Bacteria	64.67	0	1168482	1287	23
NP3_30P	3300027906_30	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; Croceibacter; Croceibacter atlanticus	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae ; Croceibacter ; Croceibacter atlanticus	64.66	9.33	1656548	1688	155
NP3_55P	3300027906_55	MQ	Archaea; Crenarchaeota; Nitrososphaeria; Nitrososphaerales; Nitrosopumilaceae; Nitrosopumilus; None	Archaea ; Thaumarchaeota ; unclassified ; Nitrosopumilales ; Nitrosopumilaceae ; Nitrosopumilus	55.28	9.71	904614	1260	162
NP3_36P	3300027906_36	MQ	Bacteria; Actinobacteriota; Acidimicrobiia; Microtrichales; MedAcidi-G1; MedAcidi-G1; None	Bacteria ; Actinobacteria	54.43	6.84	1557760	1743	227
NP3_11P	3300027906_11	MQ	Bacteria; Myxococcota; UBA4248; UBA7976; UBA1532; None; None		52.98	3.23	3453122	3164	690
NP3_40P	3300027906_40	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Psychrobacter; Psychrobacter sp5	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Pseudomonadales ; Moraxellaceae ; Psychrobacter ; Psychrobacter sp. TB15	51.44	0.64	1365662	1396	272
P1_21P	3300027849_21	HQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; Croceibacter; Croceibacter atlanticus	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae	99.62	2.81	3112211	2921	18
P1_16P	3300027849_16	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Pseudoalteromonas; Pseudoalteromonas marina	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Pseudoalteromonadaceae ; Pseudoalteromonas	97.31	1.97	4173936	3941	64
P1_20P	3300027849_20	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Saccharospirillaceae; Bermanella; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Oceanospirillales ; Oceanospirillaceae ; Bermanella ; Bermanella marisrubri	96.55	6.43	3429227	3269	125
P1_24P	3300027849_24	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Planktomarina; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae ; Roseobacter ; Roseobacter sp. LE17	92.65	1.98	2530500	2668	151
P1_15P	3300027849_15	MQ	Bacteria; Bacteroidota; Bacteroidia; Chitinophagales; Saprospiraceae; None; None	Bacteria ; Bacteroidetes	81.56	2.23	4219346	3664	553
P1_25P	3300027849_25	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; None	Bacteria ; Proteobacteria ; Gammaproteobacteria	71.84	0	2420077	2421	153
P1_41P	3300027849_41	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Cryomorphaceae; UBA10364; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales	69.09	0	1172628	1200	143
P1_23P	3300027849_23	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Loktanella; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae ; Loktanella	67.44	1.37	2625479	2988	335
P1_30P	3300027849_30	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; ASP10-02a sp3	Bacteria ; Proteobacteria ; Gammaproteobacteria	66.48	3.53	1701074	1838	266
P1_34P	3300027849_34	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; HC6-5; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae	63.99	3.07	1396029	1429	206
P1_33P	3300027849_33	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Cellvibrionales ; Porticoccaceae ; unclassified	61.44	1.48	1411191	1382	62
P1_17P	3300027849_17	MQ	Bacteria; Bacteroidota; Bacteroidia; Chitinophagales; Saprospiraceae; None; None	Bacteria ; Bacteroidetes	55.17	1.72	4035185	3453	317
P1_26P	3300027849_26	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Cellvibrionales	55.17	3.45	2199280	2081	271

P1_31P	3300027849_31	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; None	Bacteria ; Proteobacteria ; Gammaproteobacteria	54.62	0	1550253	1590	152
P1_36P	3300027849_36	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; UBA7434; UBA7434; None	Bacteria ; Proteobacteria ; Gammaproteobacteria	52.05	0	1354030	1326	172
P2_13P	3300027810_13	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Cellvibrionales ; Porticoccaceae ; unclassified	93.21	3.4	2493520	2471	199
P2_11P	3300027810_11	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Sulfitobacter_C; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae	90.45	1.11	2844391	3134	258
P2_7P	3300027810_7	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Colwellia; Colwellia polaris	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Colwelliaceae ; Colwellia	87.37	1.33	3914611	3475	65
P2_12P	3300027810_12	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; UA16; ASP10-05a; None	Bacteria ; Bacteroidetes	86.45	4.91	2610970	2445	319
P2_16P	3300027810_16	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; None	Bacteria ; Proteobacteria ; Gammaproteobacteria	65.52	0	2074814	2230	313
P2_30P	3300027810_30	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; ASP10-02a sp1	Bacteria ; Proteobacteria ; Gammaproteobacteria	62.2	1.07	1133751	1148	65
P2_23P	3300027810_23	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Cryomorphaceae; UBA10364; None	Bacteria ; Bacteroidetes	60.61	7.45	1300054	1591	265
P2_25P	3300027810_25	MQ	Bacteria; Bacteroidota; Bacteroidia; NS11-12g; UBA9320; UBA9320; None	Bacteria ; Bacteroidetes	56.68	0	1263188	1358	241
P2_20P	3300027810_20	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Cellvibrionales ; Porticoccaceae ; unclassified	54.31	0	1507785	1621	261
P2_21P	3300027810_21	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Opitutales; Puniceicoccaceae; BACL24; None	Bacteria ; Verrucomicrobia ; Opitutae ; Puniceicoccales ; Puniceicoccaceae ; Coraliomargarita ; Coraliomargarita akajimensis	50	0	1503876	1409	70
P3a_15P	3300027883_15	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Sulfitobacter_C; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae	94.76	3.53	2939519	3214	175
P3a_17P	3300027883_17	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; ASP10-02a sp1	Bacteria ; Proteobacteria ; Gammaproteobacteria	92.2	1.97	2755864	2701	118
P3a_11P	3300027883_11	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; HIMB11; GCA_002336405.1	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae	88.6	1.24	3466324	3669	438
P3a_30P	3300027883_30	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Cryomorphaceae; UBA10364; None	Bacteria ; Bacteroidetes	88.01	5.43	1692652	1661	216
P3a_28P	3300027883_28	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Opitutales; Puniceicoccaceae; BACL24; None	Bacteria ; Verrucomicrobia ; Opitutae ; Puniceicoccales ; Puniceicoccaceae ; Coraliomargarita ; Coraliomargarita akajimensis	85.14	0.71	1744461	1588	16
P3a_25P	3300027883_25	MQ	Bacteria; Bacteroidota; Bacteroidia; NS11-12g; UBA9320; UBA9320; UBA10404	Bacteria ; Bacteroidetes	82.06	1.71	1907560	1871	164
P3a_27P	3300027883_27	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; HC6-5; None	Bacteria ; Bacteroidetes	80.14	2.34	1747367	1771	209
P3a_26P	3300027883_26	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Cellvibrionales ; Porticoccaceae ; unclassified	62.07	3.45	1801705	1871	254

P3a_31P	3300027883_31	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Cellvibrionales ; Porticoccaceae ; unclassified	59.28	4.05	1488021	1577	284
P3a_21P	3300027883_21	MQ	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Loktanella; None	Bacteria ; Proteobacteria ; Alphaproteobacteria ; Rhodobacterales ; Rhodobacteraceae	58.91	2.16	2276430	2591	364
P3a_24P	3300027883_24	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales; Alteromonadaceae; Colwellia; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; Alteromonadales ; Colwelliaceae ; Colwellia	56.94	5.72	1900294	1968	342
P6_13P	3300027833_13	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; UA16; UBA8752; None	Bacteria ; Bacteroidetes	98.12	0.54	2515496	2162	113
P6_22P	3300027833_22	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; GCA-002733185; GCA_002713705.1	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae	81.03	0	1722552	1648	108
P6_15P	3300027833_15	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; GCA-002733185; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae	80.17	2.48	1908748	1920	116
P6_14P	3300027833_14	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Opitutales; Puniceicoccaceae; BACL24; None	Bacteria ; Verrucomicrobia ; Opitutae ; Puniceicoccales ; Puniceicoccaceae ; Coraliomargarita ; Coraliomargarita akajimensis	78.6	4.05	2306538	2092	86
P6_28P	3300027833_28	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; UBA7434; UBA7434; UBA7434 sp2	Bacteria ; Proteobacteria ; Gammaproteobacteria	78.19	1.11	1401065	1419	220
P6_12P	3300027833_12	MQ	Bacteria; Bacteroidota; Bacteroidia; Chitinophagales; Saprospiraceae; UBA1994; GCA_002335865.1	Bacteria ; Bacteroidetes	77.85	5.28	2792486	2636	382
P6_33P	3300027833_33	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Opitutales; Puniceicoccaceae; BACL24; None	Bacteria ; Verrucomicrobia ; Opitutae ; Puniceicoccales ; Puniceicoccaceae ; Coraliomargarita ; Coraliomargarita akajimensis	53.69	0.34	1102244	1162	195
P6_46P	3300027833_46	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Cryomorphaceae; UBA10364; UBA10364 sp1	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales	51.08	1.42	942690	1027	181
P6_35P	3300027833_35	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; HC6-5; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae	50.72	0.71	1121342	1224	178
P5_11P	3300002186_11	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Cryomorphaceae; UBA10364; UBA10364 sp1	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales	94.22	4.95	1912866	1895	157
P5_7P	3300002186_7	HQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; ASP10-02a sp1	Bacteria ; Proteobacteria ; Gammaproteobacteria	91.09	2.86	2440954	2462	128
P5_9P	3300002186_9	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Porticoccaceae; HTCC2207; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; unclassified ; unclassified ; unclassified ; gamma proteobacterium HTCC2207	79.12	5.54	2037789	2096	234
P5_24P	3300002186_24	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; SAR86; D2472; D2472; None	Bacteria ; Proteobacteria ; Gammaproteobacteria ; unclassified ; unclassified ; unclassified ; SAR86 cluster bacterium SAR86E	53.86	8.33	796012	948	139
P5_21P	3300002186_21	MQ	Bacteria; Verrucomicrobiota; Verrucomicrobiae; Opitutales; Puniceicoccaceae; BACL24; None	Bacteria ; Verrucomicrobia ; Opitutae ; Puniceicoccales ; Puniceicoccaceae ; Coraliomargarita ; Coraliomargarita akajimensis	52.36	2.3	886404	983	169
P4_19P	3300002153_19	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; BACL11; None; None	Bacteria ; Bacteroidetes	63.44	1.08	945481	972	64
P4_14P	3300002153_14	MQ	Bacteria; Bacteroidota; Bacteroidia; Flavobacteriales; Flavobacteriaceae; MAG-121220-bin8; None	Bacteria ; Bacteroidetes ; Flavobacteriia ; Flavobacteriales ; Flavobacteriaceae	58.17	2.03	982604	1060	92
P4_17P	3300002153_17	MQ	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Nitrincolaceae; ASP10-02a; ASP10-02a sp1	Bacteria ; Proteobacteria ; Gammaproteobacteria	51.92	2.53	1109978	1261	224

Table A.3 Sample identifiers and metadata

A.4 Reference Taxa in Prokaryotic Tree

Name	Phylum	Class	NCBI Taxid
Vibrio fischeri ES114	Proteobacteria	Gammaproteobacteria	312309
Altererythrobacter sp. ZODW24	Proteobacteria	Alphaproteobacteria	1872480
Pseudoalteromonas atlantica ECSMB14104	Proteobacteria	Gammaproteobacteria	342610
Corynebacterium sphenisci DSM 44792	Actinobacteriota	Actinobacteria	1437874
Pseudomonas aeruginosa Ocean-1155	Proteobacteria	Gammaproteobacteria	287
Rhodococcus ruber P14	Actinobacteriota	Actinobacteria	1830
Nocardia seriolae UTF1	Actinobacteriota	Actinobacteria	37332
Halopenitus persicus CBA1233	Halobacterota	Halobacteria	1048396
Synechococcus sp. WH 8102	Cyanobacteria	Cyanobacteriia	1131
Alteromonas sp. SN2	Proteobacteria	Gammaproteobacteria	232
Vibrio harveyi strain QT520	Proteobacteria	Gammaproteobacteria	669
Oscillibacter valericigenes Sjm18-20	Firmicutes_A	Clostridia	351091
Alteromonas macleodii str. ’English Channel 673’	Proteobacteria	Gammaproteobacteria	28108
Streptomyces sp. HNM0039	Actinobacteriota	Actinobacteria	1931
Desulfococcus multivorans strain DSM 2059	Desulfobacterota	Desulfobacteria	897
Pseudomonas fluorescens PF08	Proteobacteria	Gammaproteobacteria	294
Thermotoga maritima MSB8	Thermotogota	Thermotogae	243274
Pseudoalteromonas sp. SM9913	Proteobacteria	Gammaproteobacteria	53249
Edwardsiella tarda strain KC-Pc-HB1	Proteobacteria	Gammaproteobacteria	1027364
Olleya sp. Bg11-27	Bacteroidota	Bacteroidia	1906788
Thermococcus gammatolerans EJ3	Euryarchaeota	Thermococci	593117
Marinobacter similis A3d10	Proteobacteria	Gammaproteobacteria	1420916
Vibrio natriegens strain CCUG 16373	Proteobacteria	Gammaproteobacteria	691
Gramella sp. SH35	Bacteroidota	Bacteroidia	1931228
Pelolinea submarina strain MO-CFX1	Chloroflexota	Anaerolineae	913107
Vibrio gazogenes ATCC 43942	Proteobacteria	Gammaproteobacteria	687
Euzebya sp. DY32-46	Actinobacteriota	Actinobacteria	1971409
Bacillus sp. Alg07	Firmicutes	Bacilli	1409
Phaeobacter inhibens P78	Proteobacteria	Alphaproteobacteria	999548
Methanococcus voltae A3	Euryarchaeota	Methanococci	456320
Hydrogenophaga sp. LPB0072	Proteobacteria	Gammaproteobacteria	1904254
Rhodococcus erythropolis PR4	Actinobacteriota	Actinobacteria	234621
Alteromonas macleodii str. ’Ionian Sea U8’	Proteobacteria	Gammaproteobacteria	28108
Tenacibaculum maritimum strain NCIMB 2154	Bacteroidota	Bacteroidia	107401
Phaeobacter piscinae P36	Proteobacteria	Alphaproteobacteria	1580596
Cycloclasticus sp. P1	Proteobacteria	Gammaproteobacteria	2024830
Vibrio anguillarum 178/90	Proteobacteria	Gammaproteobacteria	882102
Psychrobacter sp. YP14	Proteobacteria	Gammaproteobacteria	56811
Shewanella baltica OS185	Proteobacteria	Gammaproteobacteria	402882
Campylobacter peloridis LMG 23910	Campylobacterota	Campylobacteria	1388753
Pseudoalteromonas haloplanktis TAC125	Proteobacteria	Gammaproteobacteria	326442
Streptomyces sp. GBA 94-10	Actinobacteriota	Actinobacteria	378518
Plantactinospora sp. BB1	Actinobacteriota	Actinobacteria	2071627
Kangiella profundi strain FT102	Proteobacteria	Gammaproteobacteria	1561924
Candidatus Nitrosopumilus sp. AR2	Crenarchaeota	Nitrososphaeria	1027373
Vibrio parahaemolyticus 20130629002S01	Proteobacteria	Gammaproteobacteria	670
Pseudoalteromonas sp. R3	Proteobacteria	Gammaproteobacteria	1709477
Sulfitobacter sp. D7	Proteobacteria	Alphaproteobacteria	1903071
Muricauda ruestringensis DSM 13258	Bacteroidota	Bacteroidia	886377
Lacinutrix venerupis DOK2-8	Bacteroidota	Bacteroidia	420889
Vibrio anguillarum S2 2/9	Proteobacteria	Gammaproteobacteria	989499
Paraglaciecola psychrophila 170	Proteobacteria	Gammaproteobacteria	1129794
Maribacter sp. B1	Bacteroidota	Bacteroidia	1897614
Piscirickettsia salmonis strain EM-90	Proteobacteria	Gammaproteobacteria	1435375
Marinovum algicola DG 898	Proteobacteria	Alphaproteobacteria	988812
Shewanella piezotolerans WP3	Proteobacteria	Gammaproteobacteria	225849
Desulfovibrio desulfuricans ND132	Desulfobacterota_A	Desulfovibrionia	876
Thermosipho sp. 1063	Thermotogota	Thermotogae	1968895
Pseudoalteromonas sp. OCN003	Proteobacteria	Gammaproteobacteria	53249
Phaeobacter inhibens P70	Proteobacteria	Alphaproteobacteria	383629
Phaeobacter piscinae P42	Proteobacteria	Alphaproteobacteria	1580596
Pseudomonas xanthomarina strain LMG 23572	Proteobacteria	Gammaproteobacteria	271420

Vibrio sp. dhg	Proteobacteria	Gammaproteobacteria	678
Oceanobacillus iheyensis HTE831	Firmicutes	Bacilli	221109
Chlorobium phaeovibrioides DSM 265	Bacteroidota	Chlorobia	290318
Synechococcus sp. KORDI-49	Cyanobacteria	Cyanobacteriia	1131
Phaeobacter inhibens P10	Proteobacteria	Alphaproteobacteria	383629
Aquimarina sp. BL5	Bacteroidota	Bacteroidia	1872586
Tenacibaculum jejuense KCTC 22618	Bacteroidota	Bacteroidia	584609
Streptococcus thermophilus APC151	Firmicutes	Bacilli	1308
Pyrodictium delaneyi strain Su06	Crenarchaeota	Thermoprotei	1273541
Methanocaldococcus sp. FS406-22	Euryarchaeota	Methanococci	2152917
Deferribacter desulfuricans SSM1	Deferribacterota	Deferribacteres	639282
Moritella viscosa 06/09/139	Proteobacteria	Gammaproteobacteria	80854
Bdellovibrio bacteriovorus strain 109J	Bdellovibrionota	Bdellovibrionia	959
Hyphomonas neptunium ATCC 15444	Proteobacteria	Alphaproteobacteria	228405
Arthrospira sp. TJSD092	Cyanobacteria	Cyanobacteriia	2153484
Croceicoccus naphthovorans strain PQ-2	Proteobacteria	Alphaproteobacteria	1348774
Maricaulis maris MCS10	Proteobacteria	Alphaproteobacteria	2774595
Nitrosococcus oceani ATCC 19707	Proteobacteria	Gammaproteobacteria	323261
Vibrio alginolyticus J207	Proteobacteria	Gammaproteobacteria	314288
Erythrobacter sp. YH-07	Proteobacteria	Alphaproteobacteria	1042
Shewanella algae KC-Na-R1	Proteobacteria	Gammaproteobacteria	38313
Aeromonas salmonicida S68	Proteobacteria	Gammaproteobacteria	645
Methanopyrus kandleri AV19	Euryarchaeota	Methanopyri	190192
Gillisia sp. Hel1_33_143	Bacteroidota	Bacteroidia	2018084
Bacillus anthracis MCCC 1A01412	Firmicutes	Bacilli	1396
Flavobacterium arcticum SM1502	Bacteroidota	Bacteroidia	1784713
Vibrio alginolyticus K09K1	Proteobacteria	Gammaproteobacteria	663
Thermococcus eurythermalis A501	Euryarchaeota	Thermococci	1505907
Colwellia sp. Arc7-D	Proteobacteria	Gammaproteobacteria	2161872
Pusillimonas sp. T7-7	Proteobacteria	Gammaproteobacteria	1979962
Synechococcus sp. RCC307	Cyanobacteria	Cyanobacteriia	1131
Belliella baltica DSM 15883	Bacteroidota	Bacteroidia	866536
Phaeobacter inhibens P59	Proteobacteria	Alphaproteobacteria	999548
Nitrosococcus watsonii C-113	Proteobacteria	Gammaproteobacteria	473531
Phaeobacter inhibens P51	Proteobacteria	Alphaproteobacteria	383629
Alteromonas macleodii str. 'Ionian Sea U7'	Proteobacteria	Gammaproteobacteria	28108
Halorhabdus tiamatea SARL4B	Halobacterota	Halobacteria	1033806
Bacillus anthracis MCCC 1A02161	Firmicutes	Bacilli	1396
Lactococcus garvieae strain 122061	Firmicutes	Bacilli	999552
Halorubrum trapanicum CBA1232	Halobacterota	Halobacteria	29284
Flavobacterium psychrophilum strain VQ50	Bacteroidota	Bacteroidia	96345
bacterium 2013Arg42i strain 2013Ark11	Proteobacteria	Gammaproteobacteria	1561003
Paracoccus sp. BM15	Proteobacteria	Alphaproteobacteria	267
Sulfurimonas denitrificans DSM 1251	Campylobacterota	Campylobacteria	326298
Shewanella sp. ANA-3	Proteobacteria	Gammaproteobacteria	50422
Saccharospirillum mangrovi HK-33	Proteobacteria	Gammaproteobacteria	2161747
Mycobacterium rhodesiae NBB3	Actinobacteriota	Actinobacteria	710685
Salinibacter ruber RM158	Bacteroidota	Rhodothermia	761659
Gramella sp. MAR_2010_102	Bacteroidota	Bacteroidia	1931228
Thermosediminibacter oceani DSM 16646	Firmicutes_A	Thermovenabulia	555079
Phaeobacter inhibens P88	Proteobacteria	Alphaproteobacteria	999548
Aciduliprofundum sp. MAR08-339	Thermoplasmatota	Thermoplasmata	2060325
Stappia sp. ES.058	Proteobacteria	Alphaproteobacteria	1870903
Alteromonas macleodii str. 'Balearic Sea AD45'	Proteobacteria	Gammaproteobacteria	28108
Vibrio cholerae strain Env-390	Proteobacteria	Gammaproteobacteria	1093790
Hippea maritima DSM 10411	Campylobacterota	Desulfurellia	760142
Draconibacterium orientale strain FH5	Bacteroidota	Bacteroidia	1168034
Sulfitobacter sp. AM1-D1	Proteobacteria	Alphaproteobacteria	1903071
Prauserella marina DSM 45268	Actinobacteriota	Actinobacteria	530584
Vibrio anguillarum 601/90	Proteobacteria	Gammaproteobacteria	105261
Alteromonas mediterranea strain RG65	Proteobacteria	Gammaproteobacteria	314275
Marinomonas sp. MWYL1	Proteobacteria	Gammaproteobacteria	1904862
Pseudoalteromonas phenolica strain KCTC 12086	Proteobacteria	Gammaproteobacteria	161398
Prochlorococcus marinus str. MIT 9312	Cyanobacteria	Cyanobacteriia	45397
Vibrio anguillarum PF4	Proteobacteria	Gammaproteobacteria	990314
Enterococcus faecalis TY1	Firmicutes	Bacilli	1351
Francisella halioticida DSM 23729	Proteobacteria	Gammaproteobacteria	549298
Candidatus Nitrosopumilus sp. NF5	Crenarchaeota	Nitrososphaeria	1027373
Simiduia agarivorans SA1	Proteobacteria	Gammaproteobacteria	1117647
Pseudorhodoplanes sinuspersici RIPI110	Proteobacteria	Alphaproteobacteria	1235591
Actinoalloteichus hymeniacidonis strain HPA177(T) (=DSM 45092(T))	Actinobacteriota	Actinobacteria	340345
Synechococcus sp. CC9902	Cyanobacteria	Cyanobacteriia	1131
Streptomyces sp. S063	Actinobacteriota	Actinobacteria	1931
Shewanella halifaxensis HAW-EB4	Proteobacteria	Gammaproteobacteria	271098
Candidatus Thioglobus singularis strain GG2	Proteobacteria	Gammaproteobacteria	1427364

Pseudomonas libanensis strain DMSP-1	Proteobacteria	Gammaproteobacteria	75588
Piscirickettsia salmonis strain PM37984A	Proteobacteria	Gammaproteobacteria	1238
Candidatus Nitrosopumilus koreensis AR1	Crenarchaeota	Nitrososphaeria	1229908
Halolamina sediminis strain halo7	Halobacterota	Halobacteria	1480675
Siansivirga zeaxanthinifaciens CC-SAMT-1	Bacteroidota	Bacteroidia	762954
Vibrio anguillarum ATCC-68554	Proteobacteria	Gammaproteobacteria	55601
Vibrio parahaemolyticus R14	Proteobacteria	Gammaproteobacteria	1394561
Flavobacterium psychrophilum FPG3	Bacteroidota	Bacteroidia	1452724
Aquimarina sp. AD1	Bacteroidota	Bacteroidia	1872586
Sulfitobacter sp. SK025	Proteobacteria	Alphaproteobacteria	1903071
Pseudoalteromonas aliena EH1	Proteobacteria	Gammaproteobacteria	247523
Morganella morganii KC-Tt-01	Proteobacteria	Gammaproteobacteria	1239989
Thiobacimonas profunda JLT2016	Proteobacteria	Alphaproteobacteria	1229727
Sphingorhabdus sp. Alg231-15	Proteobacteria	Alphaproteobacteria	1922222
Thermovirga lienii DSM 17291	Synergistota	Synergistia	580340
Desulfobacula toluolica Tol2	Desulfobacterota	Desulfobacteria	651182
Flavobacteriaceae bacterium MAR_2010_188	Bacteroidota	Bacteroidia	572194
Pelobacter carbinolicus DSM 2380	Desulfuromonadota	Desulfuromonadia	338963
Halomonas sp.R57-5	unclassified	unclassified	1610576
Phaeobacter piscinae P18	Proteobacteria	Alphaproteobacteria	1580596
Winogradskyella sp. PG-2	Bacteroidota	Bacteroidia	1883156
Synechococcus sp. KORDI 52	Cyanobacteria	Cyanobacteriia	1131
Octadecabacter arcticus 238	Proteobacteria	Alphaproteobacteria	391616
Oceanithermus profundus DSM 14977	Deinococcota	Deinococci	670487
Piscirickettsia salmonis AY3864B	Proteobacteria	Gammaproteobacteria	1238
Oceanicoccus sagamiensis NBRC 107125	Proteobacteria	Gammaproteobacteria	716816
Marinomonas mediterranea MMB-1	Proteobacteria	Gammaproteobacteria	119864
Alteromonas macleodii str. 'Ionian Sea UM7'	Proteobacteria	Gammaproteobacteria	28108
Desulfovibrio africanus str. Walvis Bay	Desulfobacterota_A	Desulfovibrionia	1262666
Mehyloceanibacter caenitepidi Gela4	Proteobacteria	Alphaproteobacteria	1384459
Bacillus sp. Pc3	Firmicutes	Bacilli	1409
Corynebacterium maris DSM 45190	Actinobacteriota	Actinobacteria	1224163
Piscirickettsia salmonis strain PM49811B	Proteobacteria	Gammaproteobacteria	1238
Alcanivorax borkumensis SK2	Proteobacteria	Gammaproteobacteria	393595
Erythrobacter litoralis strain DSM 8509	Proteobacteria	Alphaproteobacteria	39960
Staphylococcus aureus SJTUF_J27	Firmicutes	Bacilli	1280
Vibrio natriegens strain CCUG 16371	Proteobacteria	Gammaproteobacteria	1219067
Aliivibrio salmonicida LFI1238	Proteobacteria	Gammaproteobacteria	316275
Synechococcus sp. CC9605	Cyanobacteria	Cyanobacteriia	1131
Synechococcus sp. WH 8103	Cyanobacteria	Cyanobacteriia	1131
Rhodovulum sulfidophilum DSM 1374	Proteobacteria	Alphaproteobacteria	1188256
Halomonas venusta strain MA-ZP17-13	Proteobacteria	Gammaproteobacteria	44935
Tessaracoccus flavescens SST-39	Actinobacteriota	Actinobacteria	399497
Candidatus Atelocyanobacterium thalassa isolate ALOHA	Cyanobacteria	Cyanobacteriia	1453429
Edwardsiella tarda 080813	Proteobacteria	Gammaproteobacteria	636
Filomicrobium sp. W1	Proteobacteria	Alphaproteobacteria	2024831
Corynebacterium stationis strain 622=DSM 20302	Actinobacteriota	Actinobacteria	1705
Marinobacter sp. BSs20148	Proteobacteria	Gammaproteobacteria	50741
Thermococcus onnurineus NA1	Euryarchaeota	Thermococci	523850
Formosa sp. Hel3_A1_48	Bacteroidota	Bacteroidia	2018467
Flavobacterium psychrophilum OSU THCO2-90	Bacteroidota	Bacteroidia	96345
Bathymodiolus thermophilus thioautotrophic gill symbiont EPR9N	Proteobacteria	Gammaproteobacteria	2360
Pyrococcus furiosus DSM 3638	Euryarchaeota	Thermococci	186497
Nitrosococcus halophilus Nc 4	Proteobacteria	Gammaproteobacteria	472759
Vibrio coralliilyticus OCN014	Proteobacteria	Gammaproteobacteria	190893
Photobacterium gaetbulicola Gung47	Proteobacteria	Gammaproteobacteria	658445
Phaeobacter inhibens P80	Proteobacteria	Alphaproteobacteria	383629
Arcobacter sp. PSE-93	Campylobacterota	Campylobacteria	1872629
Mycobacterium chubuense NBB4	Actinobacteriota	Actinobacteria	710421
Roseobacter denitrificans OCh 114	Proteobacteria	Alphaproteobacteria	375451
Maribacter sp. MAR_2009_60	Bacteroidota	Bacteroidia	394221
Ferrimonas balearica DSM 9799	Proteobacteria	Gammaproteobacteria	550540
Prochlorococcus sp. MIT 0604	Cyanobacteria	Cyanobacteriia	1220
Chlorobium phaeobacteroides BS1	Bacteroidota	Chlorobia	331678
Vibrio anguillarum A023	Proteobacteria	Gammaproteobacteria	55601
Woeseia oceani strain XK5	Proteobacteria	Gammaproteobacteria	1548547
Shewanella japonica KCTC 22435	Proteobacteria	Gammaproteobacteria	93973
Sediminicola sp. YIK13	Bacteroidota	Bacteroidia	2511163
Vibrio anguillarum MVAV6203	Proteobacteria	Gammaproteobacteria	55601
Synechococcus sp. KORDI-100	Cyanobacteria	Cyanobacteriia	1131
Haloferax mediterranei ATCC 33500	Halobacterota	Halobacteria	523841
Erythrobacter gangjinensis strain JCM 15420	Proteobacteria	Alphaproteobacteria	502682
Arcanobacterium phocae strain DSM 10002	Actinobacteriota	Actinobacteria	131112
Hyphomonas sp. Mor2	Proteobacteria	Alphaproteobacteria	87
Owenweeksia hongkongensis DSM 17368	Bacteroidota	Bacteroidia	926562

Yersinia ruckeri YRB	Proteobacteria	Gammaproteobacteria	29486
Parvularcula bermudensis HTCC2503	Proteobacteria	Alphaproteobacteria	314260
Alteromonas macleodii str. 'Deep ecotype'	Proteobacteria	Gammaproteobacteria	28108
Shewanella baltica OS155	Proteobacteria	Gammaproteobacteria	325240
Alteromonas sp. MB-3u-76	Proteobacteria	Gammaproteobacteria	232
Vibrio parahaemolyticus strain FORC_018	Proteobacteria	Gammaproteobacteria	670
Alteromonadaceae bacterium 2141T.STBD.0c.01a	Proteobacteria	Gammaproteobacteria	650235
Alteromonas mediterranea strain AR43	Proteobacteria	Gammaproteobacteria	314275
Thalassolituus oleivorans strain K188	Proteobacteria	Gammaproteobacteria	187493
Silicimonas algicola strain KC90	Proteobacteria	Alphaproteobacteria	1826607
Haloplanus sp. CBA1112	Halobacterota	Halobacteria	1961696
Amycolatopsis albispora WP1	Actinobacteriota	Actinobacteria	1804986
Thermodesulfatator indicus DSM 15286	Desulfobacterota	Thermodesulfobacteria	667014
Granulosicoccus antarcticus IMCC 3135	Proteobacteria	Gammaproteobacteria	1192854
Methanosarcina siciliae C2J	Halobacterota	Methanosarcinia	1434118
Piscirickettsia salmonis strain CGR02	Proteobacteria	Gammaproteobacteria	1238
Brucella ceti TE10759-12	Proteobacteria	Alphaproteobacteria	120577
Phaeobacter gallaeciensis 2.10(Roseobacter gallaeciensis)	Proteobacteria	Alphaproteobacteria	383629
Hahella sp. KA22	Proteobacteria	Gammaproteobacteria	1628392
Ignicoccus islandicus DSM 13165	Crenarchaeota	Thermoprotei	940295
Aquimarina sp. AD10	Bacteroidota	Bacteroidia	1872586
Synechococcus sp. WH 8020	Cyanobacteria	Cyanobacteriia	1131
Rhodobacter sp. LPB0142	Proteobacteria	Alphaproteobacteria	633146
Brucella pinnipedialis B2/94	Proteobacteria	Alphaproteobacteria	120576
Vibrio scophthalmi strain VS-05	Proteobacteria	Gammaproteobacteria	190895
Micromonospora sp. WMMA2032	Actinobacteriota	Actinobacteria	1876
Pseudoalteromonas piscicida DE1-A	Proteobacteria	Gammaproteobacteria	43662
Bacillus subtilis subsp. spizizenii SW83	Firmicutes	Bacilli	909461
Vibrio chagasii strain ECSMB14107	Proteobacteria	Gammaproteobacteria	170679
Methanococcus maripaludis S2	Euryarchaeota	Methanococci	267377
Yersinia aldovae 670-83	Proteobacteria	Gammaproteobacteria	29483
Acidobacteria bacterium Mor1	Acidobacteriota	Mor1	1660251
Gramella flava JLT2011	Bacteroidota	Bacteroidia	1229726
Streptococcus iniae SF1	Firmicutes	Bacilli	1318633
Vibrio alfacsensis CAIM 1831	Proteobacteria	Gammaproteobacteria	1074311
Haliangium ochraceum DSM 14365	Myxococcota	Polyangia	502025
Alcanivorax sp. W11-5	Proteobacteria	Gammaproteobacteria	1872427
Kordia sp. SMS9	Bacteroidota	Bacteroidia	1965332
Celeribacter baekdonensis LH4	Proteobacteria	Alphaproteobacteria	1208323
Haloplanus sp. CBA1113	Halobacterota	Halobacteria	1961696
Agrococcus jejuensis strain DSM 22002	Actinobacteriota	Actinobacteria	399736
Shewanella baltica OS223	Proteobacteria	Gammaproteobacteria	407976
Methanococcus maripaludis X1	Euryarchaeota	Methanococci	1053692
Aliivibrio salmonicida strain VS224	Proteobacteria	Gammaproteobacteria	40269
Streptomyces violaceoruber S21	Actinobacteriota	Actinobacteria	1935
Erythrobacter litoralis HTCC2594	Proteobacteria	Alphaproteobacteria	314225
Tenacibaculum sp. LPB0136	Bacteroidota	Bacteroidia	1906242
Candidatus Pelagibacter ubique HTCC1062	Proteobacteria	Alphaproteobacteria	335992
Alcanivorax dieselolei B5	Proteobacteria	Gammaproteobacteria	930169
Kangiella sediminilitoris strain KCTC 23892	Proteobacteria	Gammaproteobacteria	1144748
Vibrio anguillarum NB10	Proteobacteria	Gammaproteobacteria	55601
Celeribacter indicus P73	Proteobacteria	Alphaproteobacteria	1208324
Thalassococcus sp. SH-1	Proteobacteria	Alphaproteobacteria	2017482
Shewanella baltica OS678	Proteobacteria	Gammaproteobacteria	693973
Francisella sp. FSC1006	Proteobacteria	Gammaproteobacteria	2047875
Fibrella aestuarina strain BUZ 2	Bacteroidota	Bacteroidia	1166018
Methylophilales bacterium MBRSF5	Proteobacteria	Gammaproteobacteria	1623448
Corynebacterium phocae strain M408/89/1	Actinobacteriota	Actinobacteria	161895
Pseudomonas sp. MT-1	Proteobacteria	Gammaproteobacteria	306
Devosia sp. I507	Proteobacteria	Alphaproteobacteria	2083786
Thermosipho melanesiensis BI429	Thermotogota	Thermotogae	391009
Thermosipho melanesiensis strain 431	Thermotogota	Thermotogae	46541
Lacinutrix sp. 5H-3-7-4	Bacteroidota	Bacteroidia	1937692
Pyrococcus abyssi GE5	Euryarchaeota	Thermococci	272844
Sulfurovum sp. NBC37-1	Campylobacterota	Campylobacteria	1969726
Planococcus kocurii strain ATCC 43650	Firmicutes	Bacilli	1374
Vibrio fischeri MJ11	Proteobacteria	Gammaproteobacteria	388396
Sulfitobacter sp. JL08	Proteobacteria	Alphaproteobacteria	1903071
Bacillus subtilis subsp. subtilis BS155	Firmicutes	Bacilli	909461
Thiomicrospira sp. S5	Proteobacteria	Gammaproteobacteria	1803865
Vibrio parahaemolyticus 160807	Proteobacteria	Gammaproteobacteria	670
Rubrobacter indicoceani SCSIO 08198	Actinobacteriota	Rubrobacteria	2051957
Vibrio mediterranei strain 117-T6	Proteobacteria	Gammaproteobacteria	689
Mariprofundus aestuarium strain CP-5	Proteobacteria	Zetaproteobacteria	1921086
Photobacterium profundum SS9	Proteobacteria	Gammaproteobacteria	298386

Indioceanicola profundi SCSIO 08040	Proteobacteria	Alphaproteobacteria	2220096
Pyrococcus horikoshii OT3	Euryarchaeota	Thermococci	70601
Calothrix parasitica NIES-267	Cyanobacteria	Cyanobacteriia	1973486
Aeromicrobium sp. A1-2	Actinobacteriota	Actinobacteria	1743116
Palaeococcus pacificus DY20341	Euryarchaeota	Thermococci	1343739
Haloferax gibbonsii strain ARA6	Halobacterota	Halobacteria	35746
Zunongwangia profunda SM-A87	Bacteroidota	Bacteroidia	398743
Prochlorococcus marinus bv. HNLC1	Cyanobacteria	Cyanobacteriia	1219
Rhodobacter sphaeroides strain AB29	Proteobacteria	Alphaproteobacteria	1063
Geobacillus sp. 12AMOR1	Firmicutes	Bacilli	1891658
Methyloceanibacter sp. wino2	Proteobacteria	Alphaproteobacteria	2170729
Shewanella marisflavi EP1	Proteobacteria	Gammaproteobacteria	260364
Rhodothermaceae bacterium MEBiC09517	Bacteroidota	Rhodothermia	2026787
Phaeobacter gallaeciensis P129	Proteobacteria	Alphaproteobacteria	60890
Paraoerskovia marina strain DSM 22126	Actinobacteriota	Actinobacteria	545619
Pseudomonas stutzeri strain 19SMN4	Proteobacteria	Gammaproteobacteria	316
Sulfurimonas autotrophica DSM 16294	Campylobacterota	Campylobacteria	563040
Corynebacterium marinum DSM 44953	Actinobacteriota	Actinobacteria	1224162
Piscirickettsia salmonis strain PM51819A	Proteobacteria	Gammaproteobacteria	1238
Marinobacter sp. es.042	Proteobacteria	Gammaproteobacteria	225847
Maritalea myrionectae HL2708#5	Proteobacteria	Alphaproteobacteria	454601
Nitratiruptor sp. SB155-2	Campylobacterota	Campylobacteria	2081525
Fervidobacterium pennivorans strain DYC	Thermotogota	Thermotogae	93466
Spirochaeta sp. L21-RPul-D2	Spirochaetota	Spirochaetia	28185
Thalassolituus oleivorans R6-15	Proteobacteria	Gammaproteobacteria	187493
Ilyobacter polytropus DSM 2926	Fusobacteriota	Fusobacteriia	572544
Sphingopyxis sp. LPB0140	Proteobacteria	Alphaproteobacteria	1908224
Alteromonas sp. Mex14	Proteobacteria	Gammaproteobacteria	232
Streptococcus agalactiae	Firmicutes	Bacilli	1311
Alteromonas stellipolaris strain PQQ-44	Proteobacteria	Gammaproteobacteria	233316
Desulfotomaculum reducens MI-1	Firmicutes_B	Desulfotomaculia	59610
Kosmotoga olearia TBF 19.5.1	Thermotogota	Thermotogae	651457
Polaribacter sp. ALD11	Bacteroidota	Bacteroidia	1920175
Ilumatobacter coccineum YM16-304	Actinobacteriota	Acidimicrobiia	467094
Streptomyces spongiicola HNM0071	Actinobacteriota	Actinobacteria	1690221
Phaeobacter gallaeciensis DSM 26640	Proteobacteria	Alphaproteobacteria	1423144
Ruegeria sp. TM1040	Proteobacteria	Alphaproteobacteria	1879320
Roseibacterium elongatum DSM 19469	Proteobacteria	Alphaproteobacteria	1294273
Desulfovibrio hydrothermalis AM13	Desulfobacterota_A	Desulfovibrionia	1121451
Vibrio anguillarum 4299	Proteobacteria	Gammaproteobacteria	55601
Marinobacter hydrocarbonoclasticus ATCC 49840	Proteobacteria	Gammaproteobacteria	1902815
Flavobacterium psychrophilum v4-33	Bacteroidota	Bacteroidia	96345
Methanococcus maripaludis C7	Euryarchaeota	Methanococci	426368
Streptomyces sp. CC0208	Actinobacteriota	Actinobacteria	1931
Prochlorococcus marinus str. MIT 9313	Cyanobacteria	Cyanobacteriia	45397
Pyrobaculum aerophilum str. IM2	Crenarchaeota	Thermoprotei	178306
Vibrio anguillarum 6018/1	Proteobacteria	Gammaproteobacteria	882102
Nonlabens sp. MB-3u-79	Bacteroidota	Bacteroidia	1888209
Salinimonas sp. N102	Proteobacteria	Gammaproteobacteria	1929415
Vibrio alginolyticus K04M5	Proteobacteria	Gammaproteobacteria	663
Methanoplanus petrolearius DSM 11571	Halobacterota	Methanomicrobia	679926
Reinekea forsetii Hel1_31_D35	Proteobacteria	Gammaproteobacteria	1336806
Flexibacter litoralis DSM 6794	Bacteroidota	Bacteroidia	880071
Erythrobacter flavus VG1	Proteobacteria	Alphaproteobacteria	95172
Hoeflea sp. IMCC20628	Proteobacteria	Alphaproteobacteria	1940281
Vibrio anguillarum PF4	Proteobacteria	Gammaproteobacteria	990314
Microbacterium sp. LKL04	Actinobacteriota	Actinobacteria	51671
Desulfovibrio piezophilus strain C1TLV30	Desulfobacterota_A	Desulfovibrionia	879567
Marinobacter aquaeolei VT8	Proteobacteria	Gammaproteobacteria	1163748
Thermotoga sp. Cell2	Thermotogota	Thermotogae	28240
Trichodesmium erythraeum IMS101	Cyanobacteria	Cyanobacteriia	203124
Sedimenticola sp. SIP-G1	Proteobacteria	Gammaproteobacteria	1940285
Shewanella psychrophila WP2	Proteobacteria	Gammaproteobacteria	225848
Maribacter sp. HTCC2170	Bacteroidota	Bacteroidia	1897614
Salinibacter ruber DSM 13855	Bacteroidota	Rhodothermia	309807
Francisella sp. FDC440	Proteobacteria	Gammaproteobacteria	2047875
Gramella forsetii KT0803	Bacteroidota	Bacteroidia	411154
Phaeobacter sp. LSS9	Proteobacteria	Alphaproteobacteria	1902409
Psychrobacter sp. P11G3	Proteobacteria	Gammaproteobacteria	56811
Candidatus Endolissoclinum patella L2	Proteobacteria	Alphaproteobacteria	1263978
Planococcus donghaensis strain DSM 22276	Firmicutes	Bacilli	414778
Prochlorococcus marinus subsp. marinus str. CCMP1375	Cyanobacteria	Cyanobacteriia	142554
Rhodococcus sp. 008	Actinobacteriota	Actinobacteria	1831
Vibrio shilonii QT6D1	Proteobacteria	Gammaproteobacteria	45658
Enterobacter cloacae E3442	Proteobacteria	Gammaproteobacteria	550

Tamlana sp. UJ94	Bacteroidota	Bacteroidia	1969468
Mycobacterium chelonae CCUG 47445	Actinobacteriota	Actinobacteria	1460372
Celeribacter ethanolicus strain TSPH2	Proteobacteria	Alphaproteobacteria	1758178
Pseudoalteromonas translucida KMM 520	Proteobacteria	Gammaproteobacteria	1315283
Prochlorococcus marinus str. MIT 9215	Cyanobacteria	Cyanobacteriia	45397
Aliivibrio wodanis 06/09/139	Proteobacteria	Gammaproteobacteria	80852
Polaribacter sp. Hel1_33_78	Bacteroidota	Bacteroidia	1920175
Cellulophaga baltica NN016038	Bacteroidota	Bacteroidia	1348585
Shewanella violacea DSS12	Proteobacteria	Gammaproteobacteria	637905
Plantactinospora sp. BC1	Actinobacteriota	Actinobacteria	1928616
Pseudomonas stutzeri strain 273	Proteobacteria	Gammaproteobacteria	316
Streptomyces sp. CNQ-509	Actinobacteriota	Actinobacteria	1931
Phycisphaera mikurensis NBRC 102666	Planctomycetota	Phycisphaerae	1142394
Thiolapillus brandeum Hiromi 1	Proteobacteria	Gammaproteobacteria	1076588
Pseudoalteromonas rubra strain SCSIO 6842	Proteobacteria	Gammaproteobacteria	43658
Dinoroseobacter shibae DFL 12	Proteobacteria	Alphaproteobacteria	398580
Oceanisphaera profunda SM1222	Proteobacteria	Gammaproteobacteria	1416627
Kushneria marisflavi SW32	Proteobacteria	Gammaproteobacteria	157779
Bacillus amyloliquefaciens SH-B74	Firmicutes	Bacilli	1390
Phaeobacter gallaeciensis strain JL2886	Proteobacteria	Alphaproteobacteria	60890
Shewanella sp. W3-18-1	Proteobacteria	Gammaproteobacteria	50422
Pseudoalteromonas espejiana ATCC 29659	Proteobacteria	Gammaproteobacteria	1314869
Nanohaloarchaea archaeon SG9	Nanohaloarchaeota	Nanosalinia	1737403
Kyrpidia sp. EA-1	Firmicutes_K	Alicyclobacillia	478801
Bacillus safensis strain KCTC 12796BP	Firmicutes	Bacilli	561879
Rhodovulum sp. P5	Proteobacteria	Alphaproteobacteria	34009
Thermotoga maritima strain Tma200	Thermotogota	Thermotogae	2336
Yersinia ruckeri NHV_3758	Proteobacteria	Gammaproteobacteria	29486
Methanococcus maripaludis DSM 2067	Euryarchaeota	Methanococci	267377
Prochlorococcus marinus str. NATL1A	Cyanobacteria	Cyanobacteriia	1219
Lysobacter maris HZ9B	Proteobacteria	Gammaproteobacteria	1605891
Vibrio coralliilyticus SNUTY-1	Proteobacteria	Gammaproteobacteria	190893
Polaribacter reichenbachii 6Alg 8	Bacteroidota	Bacteroidia	996801
Synechococcus sp. CC9311	Cyanobacteria	Cyanobacteriia	1131
Celeribacter manganoxidans strain DY25	Proteobacteria	Alphaproteobacteria	1411902
Piscirickettsia salmonis strain AY6297B	Proteobacteria	Gammaproteobacteria	1238
Vibrio parahaemolyticus BB22OP	Proteobacteria	Gammaproteobacteria	696485
Moorea producens JHB	Cyanobacteria	Cyanobacteriia	1454205
Agarivorans gilvus strain WH0801	Proteobacteria	Gammaproteobacteria	680279
Vibrio vulnificus strain 93U204	Proteobacteria	Gammaproteobacteria	672
Streptococcus parauberis SPOF3K	Firmicutes	Bacilli	1348
Alteromonas macleodii str. 'Ionian Sea UM4b'	Proteobacteria	Gammaproteobacteria	28108
Flavobacteriaceae bacterium AU392	Bacteroidota	Bacteroidia	1871037
Pyrococcus yayanosii CH1	Euryarchaeota	Thermococci	529709
Nitrosopumilus maritimus SCM1	Crenarchaeota	Nitrososphaeria	436308
Pseudoalteromonas donghaensis HJ51	Proteobacteria	Gammaproteobacteria	621376
Methanotorris igneus Kol 5	Euryarchaeota	Methanococci	880724
Kangiella koreensis DSM 16069	Proteobacteria	Gammaproteobacteria	523791
Streptomyces sp. PVA 94-07	Actinobacteriota	Actinobacteria	594563
Methanobacterium sp. MO-MB1	Euryarchaeota	Methanobacteria	2164
Vibrio anguillarum PF4	Proteobacteria	Gammaproteobacteria	990314
Altererythrobacter atlanticus strain 26DY36	Proteobacteria	Alphaproteobacteria	1267766
Cenarchaeum symbiosum A	Crenarchaeota	Nitrososphaeria	414004
Vibrio anguillarum T265	Proteobacteria	Gammaproteobacteria	55601
Piscirickettsia salmonis strain PM21567A	Proteobacteria	Gammaproteobacteria	1238
Cyclobacterium marinum DSM 745	Bacteroidota	Bacteroidia	880070
Hahella chejuensis KCTC 2396	Proteobacteria	Gammaproteobacteria	349521
Salegentibacter sp. T436	Bacteroidota	Bacteroidia	1903072
Phaeobacter inhibens P66	Proteobacteria	Alphaproteobacteria	999548
Thalassotalea sp. LPB0090	Proteobacteria	Gammaproteobacteria	1897616
Pseudoalteromonas tunicata D2	Proteobacteria	Gammaproteobacteria	314281
Vibrio anguillarum M3	Proteobacteria	Gammaproteobacteria	882944
Nonlabens sp. MIC269	Bacteroidota	Bacteroidia	1888209
Gramella sp. MAR_2010_147	Bacteroidota	Bacteroidia	1931228
Mycoplasma phocidae 105	Firmicutes	Bacilli	142651
Krokinobacter sp. 4H-3-7-5	Bacteroidota	Bacteroidia	2024995
Thermococcus sp. EXT12c	Euryarchaeota	Thermococci	35749
Denitrovibrio acetiphilus DSM 12809	Deferribacterota	Deferribacteres	522772
Ruegeria sp. AD91A	Proteobacteria	Alphaproteobacteria	1879320
Thermococcus sp. ES1	Euryarchaeota	Thermococci	35749
Flavobacterium psychrophilum strain MH1	Bacteroidota	Bacteroidia	96345
Persephonella marina EX-H1	Aquificota	Aquificae	309805
Vibrio anguillarum 91-8-178	Proteobacteria	Gammaproteobacteria	55601
Rhodobacteraceae bacterium BAR1	Proteobacteria	Alphaproteobacteria	1904441
Piscirickettsia salmonis strain AY6532B	Proteobacteria	Gammaproteobacteria	1238

Haloarcula sp. CBA1115	Halobacterota	Halobacteria	44098
Marinobacter adhaerens HP15	Proteobacteria	Gammaproteobacteria	351348
Dokdonia sp. PRO95	Bacteroidota	Bacteroidia	2024995
Aquiflexum balticum DSM 16537	Bacteroidota	Bacteroidia	758820
Alcanivorax xenomutans P40	Proteobacteria	Gammaproteobacteria	1094342
Acaryochloris marina MBIC11017	Cyanobacteria	Cyanobacteriia	329726
Vibrio alginolyticus strain ZJ-T	Proteobacteria	Gammaproteobacteria	663
Vibrio anguillarum PF430-3	Proteobacteria	Gammaproteobacteria	55601
Mesorhizobium sp. B7	Proteobacteria	Alphaproteobacteria	1871066
Thermococcus litoralis DSM 5473	Euryarchaeota	Thermococci	523849
Altererythrobacter namhicola strain JCM 16345	Proteobacteria	Alphaproteobacteria	645517
Thermotoga neapolitana DSM 4359	Thermotogota	Thermotogae	309803
Helicobacter cetorum MIT 99-5656	Campylobacterota	Campylobacteria	138563
Renibacterium salmoninarum ATCC 33209	Actinobacteriota	Actinobacteria	288705
Paraphotobacterium marinum NSCS20N07D	Proteobacteria	Gammaproteobacteria	1009845
Phaeobacter inhibens 2.10	Proteobacteria	Alphaproteobacteria	221822
Mycobacterium pseudoshottsii JCM 15466	Actinobacteriota	Actinobacteria	1136880
Vibrio coralliilyticus RE98	Proteobacteria	Gammaproteobacteria	190893
Lacinutrix sp. Bg11-31	Bacteroidota	Bacteroidia	1486034
Kytococcus sedentarius DSM 20547	Actinobacteriota	Actinobacteria	1526571
Staphylothermus hellenicus DSM 12710	Crenarchaeota	Thermoprotei	591019
Vibrio campbellii 1114GL	Proteobacteria	Gammaproteobacteria	680
Prochlorococcus marinus str. MIT 9301	Cyanobacteria	Cyanobacteriia	45397
Salinibacter ruber M8	Bacteroidota	Rhodothermia	761659
Methylophilales bacterium MBRSG12	Proteobacteria	Gammaproteobacteria	1623449
Flavobacterium psychrophilum strain CSF259-93	Bacteroidota	Bacteroidia	96345
Piscirickettsia salmonis strain PM22180B	Proteobacteria	Gammaproteobacteria	1238
Vibrio anguillarum PF7	Proteobacteria	Gammaproteobacteria	55601
Thioflavicoccus mobilis 8321	Proteobacteria	Gammaproteobacteria	765912
Alteromonas macleodii AltDE1	Proteobacteria	Gammaproteobacteria	1004786
Vibrio anguillarum 775	Proteobacteria	Gammaproteobacteria	882102
Vibrio parahaemolyticus R13	Proteobacteria	Gammaproteobacteria	1288784
Idiomarina sp. OT37-5b	Proteobacteria	Gammaproteobacteria	2100422
Tessaracoccus sp. NSG39	Actinobacteriota	Actinobacteria	1971211
Formosa haliotis strain LMG 28520	Bacteroidota	Bacteroidia	1555194
Phaeobacter inhibens P72	Proteobacteria	Alphaproteobacteria	999548
Rhodobacter sphaeroides strain AB27	Proteobacteria	Alphaproteobacteria	1063
Methanococcoides methylutens MM1	Halobacterota	Methanosarcinia	1434104
Piscirickettsia salmonis strain PM58386B	Proteobacteria	Gammaproteobacteria	1238
Photobacterium damselae Phdp Wu-1	Proteobacteria	Gammaproteobacteria	38293
Colwellia sp. PAMC 20917	Proteobacteria	Gammaproteobacteria	56799
Piscirickettsia salmonis strain PM23019A	Proteobacteria	Gammaproteobacteria	1238
Weissella ceti strain WS74	Firmicutes	Bacilli	759620
Methanosarcina sp. MTP4	Halobacterota	Methanosarcinia	2213
Dokdonia sp. Dokd-P16	Bacteroidota	Bacteroidia	2173169
Mariprofundus ferrinatatus strain CP-8	Proteobacteria	Zetaproteobacteria	1921087
Geobacillus kaustophilus HTA426	Firmicutes	Bacilli	235909
Rhodococcus sp. B7740	Actinobacteriota	Actinobacteria	1831
Streptococcus iniae strain YSFST01-82	Firmicutes	Bacilli	1346
Pseudoalteromonas luteoviolacea strain S4054	Proteobacteria	Gammaproteobacteria	1129367
Synechococcus sp. NIES-970	Cyanobacteria	Cyanobacteriia	1131
Pyrolobus fumarii 1A	Crenarchaeota	Thermoprotei	694429
Formosa agariphila KMM 3901	Bacteroidota	Bacteroidia	1347342
Sphingopyxis alaskensis RB2256	Proteobacteria	Alphaproteobacteria	317655
Lactococcus garvieae ATCC 49156	Firmicutes	Bacilli	420890
Prochlorococcus sp. MIT 0801	Cyanobacteria	Cyanobacteriia	1220
Marinitoga piezophila KA3	Thermotogota	Thermotogae	1545835
Arcticibacterium luteifluviistationis SM1504	Bacteroidota	Bacteroidia	1784714
Salinibacter ruber P18	Bacteroidota	Rhodothermia	761659
Pseudoalteromonas issachenkonii strain KCTC 12958	Proteobacteria	Gammaproteobacteria	152297
Phaeobacter inhibens P57	Proteobacteria	Alphaproteobacteria	999548
Acinetobacter venetianus VE-C3	Proteobacteria	Gammaproteobacteria	52133
Streptomyces sp. ADI95-16	Actinobacteriota	Actinobacteria	1244134
Salinigranum rubrum GX10	Halobacterota	Halobacteria	755307
Phaeobacter inhibens P74	Proteobacteria	Alphaproteobacteria	999548
Marinitoga sp. 1137	Thermotogota	Thermotogae	225937
Epibacterium mobile strain EPIB1	Proteobacteria	Alphaproteobacteria	379347
Methanocaldococcus fervens AG86	Euryarchaeota	Methanococci	573064
Vibrio parahaemolyticus UCM-V493	Proteobacteria	Gammaproteobacteria	670
Sulfitobacter sp. SK012	Proteobacteria	Alphaproteobacteria	1903071
Haloplanus aerogenes strain JCM 16430	Halobacterota	Halobacteria	660522
Streptomyces sp. 452	Actinobacteriota	Actinobacteria	271448
Geoglobus acetivorans SBH6	Halobacterota	Archaeoglobi	565033
Candidatus Ruthia magnifica str. Cm (Calyptogena magnifica)	Proteobacteria	Gammaproteobacteria	386487
Nonlabens marinus S1-08	Bacteroidota	Bacteroidia	930802

Synechococcus sp. WH 7803	Cyanobacteria	Cyanobacteriia	1131
Alteromonas sp. RW2A1	Proteobacteria	Gammaproteobacteria	232
Tenericutes bacterium MZ-XQ	Firmicutes	Bacilli	2231116
Alcaligenes faecalis J481	Proteobacteria	Gammaproteobacteria	511
Pseudoalteromonas sp. 1_2015MBL_MicDiv strain 15DKN1	Proteobacteria	Gammaproteobacteria	1720343
Thermococcus sp. CDGS	Euryarchaeota	Thermococci	35749
Vibrio campbellii 20130629003S01	Proteobacteria	Gammaproteobacteria	680
Carnobacterium sp. 17-4	Firmicutes	Bacilli	48221
Erythrobacter sp. Alg231-14	Proteobacteria	Alphaproteobacteria	1922225
Glaciecola nitratireducens FR1064	Proteobacteria	Gammaproteobacteria	1085623
Vibrio furnissii NCTC 11218	Proteobacteria	Gammaproteobacteria	903510
Vibrio anguillarum 87-9-117	Proteobacteria	Gammaproteobacteria	55601
Marinobacter sp. Arc7-DN-1	Proteobacteria	Gammaproteobacteria	50741
Chromohalobacter salexigens DSM 3043	Proteobacteria	Gammaproteobacteria	290398
Celeribacter marinus strain IMCC12053	Proteobacteria	Alphaproteobacteria	1397108
Candidatus Nitrosopelagicus brevis CN25	Crenarchaeota	Nitrososphaeria	1410606
Prochlorococcus marinus str. MIT 9211	Cyanobacteria	Cyanobacteriia	45397
Alteromonas mediterranea strain CP49	Proteobacteria	Gammaproteobacteria	314275
Vibrio anguillarum VIB12	Proteobacteria	Gammaproteobacteria	55601
Muricauda lutaonensis strain CC-HSB-11	Bacteroidota	Bacteroidia	516051
Erysipelothrix rhusiopathiae KC-Sb-R1	Firmicutes	Bacilli	1648
Nanoarchaeum equitans Kin4-M	Nanoarchaeota	Nanoarchaeia	160232
Pelobacter acetylenicus DSM 3247	Desulfuromonadota	Desulfuromonadia	29542
Chromobacterium sp. IIBBL 112-1	Proteobacteria	Gammaproteobacteria	306190
Phaeobacter gallaeciensis P75	Proteobacteria	Alphaproteobacteria	60890
Methylomonas methanica MC09	Proteobacteria	Gammaproteobacteria	857087
Echinicola vietnamensis DSM 17526	Bacteroidota	Bacteroidia	926556
Pseudoalteromonas agarivorans Hao 2018	Proteobacteria	Gammaproteobacteria	176102
Phaeobacter gallaeciensis P63	Proteobacteria	Alphaproteobacteria	60890
Methanosarcina siciliae HI350	Halobacterota	Methanosarcinia	1434119
Pseudoalteromonas tunicata D2	Proteobacteria	Gammaproteobacteria	87626
Serratia marcescens KS10	Proteobacteria	Gammaproteobacteria	615
Magnetococcus marinus MC-1	Proteobacteria	Magnetococcia	1288970
Methanohalobium evestigatum Z-7303	Halobacterota	Methanosarcinia	2322
Roseobacter denitrificans FDAARGOS_309	Proteobacteria	Alphaproteobacteria	2434
Methanocaldococcus infernus ME	Euryarchaeota	Methanococci	573063
Flavobacteriaceae bacterium UJ101	Bacteroidota	Bacteroidia	1150389
Desulfurobacterium thermolithotrophum DSM 11699	Aquificota	Desulfurobacteriia	868864
Helicobacter cetorum MIT 00-7128	Campylobacterota	Campylobacteria	138563
Alteromonas sp. BL110	Proteobacteria	Gammaproteobacteria	232
Vibrio owensii V180403	Proteobacteria	Gammaproteobacteria	696485
Vibrio rotiferianus B64D1	Proteobacteria	Gammaproteobacteria	670
Pelagibacterium halotolerans B2	Proteobacteria	Alphaproteobacteria	1082931
Halobacteriovorax sp. BALOs_7	Bdellovibrionota	Bacteriovoracia	2109558
Synechococcus sp. PCC 7002	Cyanobacteria	Cyanobacteriia	2269060
Lactococcus garvieae Lg2	Firmicutes	Bacilli	1363
Alteromonas australica DE170	Proteobacteria	Gammaproteobacteria	589873
Bacillus sp. Y-01	Firmicutes	Bacilli	385524
Rhodobacter sphaeroides strain AB24	Proteobacteria	Alphaproteobacteria	1063
Candidatus Nitrosopumilus sp. D3C	Crenarchaeota	Nitrososphaeria	1027373
Pseudodesulfovibrio profundus 500-1	Desulfobacterota_A	Desulfovibrionia	57320
Bacillus cereus CC-1	Firmicutes	Bacilli	1396
Arthrobacter sp. PAMC25486	Actinobacteriota	Actinobacteria	1667
Photobacterium angustum LC1-200	Proteobacteria	Gammaproteobacteria	661
Psychrobacter sp. G	Proteobacteria	Gammaproteobacteria	56811
Dokdonia sp. MED134	Bacteroidota	Bacteroidia	2024995
Clostridium botulinum 202F	Firmicutes_A	Clostridia	1415774
Serratia marcescens EL1	Proteobacteria	Gammaproteobacteria	615
Methylophaga nitratireducenticrescens GP59	Proteobacteria	Gammaproteobacteria	754476
Pseudomonas aeruginosa Ocean-1175	Proteobacteria	Gammaproteobacteria	287
Colwellia sp. MT41	Proteobacteria	Gammaproteobacteria	56799
Jannaschia sp. CCS1	Proteobacteria	Alphaproteobacteria	1966345
Spiribacter salinus M19-40	Proteobacteria	Gammaproteobacteria	1335746
Rhodococcus sp. WMMA185	Actinobacteriota	Actinobacteria	1831
alpha proteobacterium HIMB59	Proteobacteria	Alphaproteobacteria	744985
Ignicoccus hospitalis KIN4/I	Crenarchaeota	Thermoprotei	453591
Pseudoalteromonas carrageenovora IAM 12662 strain ATCC 43555	Proteobacteria	Gammaproteobacteria	1314868
Psychrobacter sp. PRwf-1	Proteobacteria	Gammaproteobacteria	56811
Shewanella sp. MR-4	Proteobacteria	Gammaproteobacteria	50422
Vibrio anguillarum 51/82/2	Proteobacteria	Gammaproteobacteria	882944
Spirochaeta thermophila DSM 6578	Spirochaetota	Spirochaetia	869211
Alteromonas australica H 17	Proteobacteria	Gammaproteobacteria	589873
Coraliomargarita akajimensis DSM 45221	Verrucomicrobiota	Verrucomicrobiae	583355
Octadecabacter temperatus strain SB1	Proteobacteria	Alphaproteobacteria	1458307
Sulfitobacter sp. SK011	Proteobacteria	Alphaproteobacteria	1903071

Pseudomonas stutzeri CCUG 29243	Proteobacteria	Gammaproteobacteria	1196835
Pyrococcus sp. ST04	Euryarchaeota	Thermococci	33866
Ruegeria pomeroyi DSS-3	Proteobacteria	Alphaproteobacteria	89184
Novosphingobium sp. PP1Y	Proteobacteria	Alphaproteobacteria	1874826
Thermotoga maritima strain Tma100	Thermotogota	Thermotogae	2336
Piscirickettsia salmonis strain AY3800B	Proteobacteria	Gammaproteobacteria	1238
Psychrobacter sp. AntiMn-1	Proteobacteria	Gammaproteobacteria	56811
Shewanella baltica OS117	Proteobacteria	Gammaproteobacteria	693970
Pontimonas salivibrio CL-TW6	Actinobacteriota	Actinobacteria	1159327
Shewanella sp. MR-7	Proteobacteria	Gammaproteobacteria	50422
Flavobacterium sp. LPB0076	Bacteroidota	Bacteroidia	239
Microbulbifer agarilyticus GP101	Proteobacteria	Gammaproteobacteria	260552
Vibrio alginolyticus ATCC 33868	Proteobacteria	Gammaproteobacteria	663
Salinibacter ruber SP73	Bacteroidota	Rhodothermia	761659
Altererythrobacter dongtanensis strain KCTC 22672	Proteobacteria	Alphaproteobacteria	692370
Shewanella amazonensis SB2B	Proteobacteria	Gammaproteobacteria	326297
Vibrio nigripulchritudo str. SFn1	Proteobacteria	Gammaproteobacteria	691
Robiginitalea biformata HTCC2501	Bacteroidota	Bacteroidia	313596
Nodularia spumigena CCY9414	Cyanobacteria	Cyanobacteriia	313624
Phaeobacter gallaeciensis P128	Proteobacteria	Alphaproteobacteria	60890
Vibrio campbellii BoB-90	Proteobacteria	Gammaproteobacteria	680
Vibrio anguillarum HI610	Proteobacteria	Gammaproteobacteria	55601
Streptomyces sp. SCSIO 03032	Actinobacteriota	Actinobacteria	1931
Euzebyella marina RN62	Bacteroidota	Bacteroidia	1761453
Gallaecimonas sp. HK-28	Proteobacteria	Gammaproteobacteria	1972664
Vibrio campbellii LA16-V1	Proteobacteria	Gammaproteobacteria	680
Yersinia ruckeri strain Big Creek 74	Proteobacteria	Gammaproteobacteria	29486
Piscirickettsia salmonis strain AY6492A	Proteobacteria	Gammaproteobacteria	1238
Erythrobacter sp. s21-N3	Proteobacteria	Alphaproteobacteria	1042
Magnetospira sp. QH-2	Proteobacteria	Alphaproteobacteria	1897614
Polaribacter reichenbachii KCTC 23969	Bacteroidota	Bacteroidia	996801
Colwellia psychrerythraea 34H	Proteobacteria	Gammaproteobacteria	167879
Staphylococcus delphini strain NCTC12225	Firmicutes	Bacilli	53344
Pseudoalteromonas issachenkonii KMM 3549	Proteobacteria	Gammaproteobacteria	1315274
Caldithrix abyssi DSM 13497	Calditrichota	Calditrichia	880073
Hyphomicrobium nitrativorans NL23	Proteobacteria	Alphaproteobacteria	1029756
Shewanella baltica BA175	Proteobacteria	Gammaproteobacteria	693974
Phaeobacter inhibens P83	Proteobacteria	Alphaproteobacteria	999548
Thermotoga sp. RQ2	Thermotogota	Thermotogae	28240
Prochlorococcus marinus str. MIT 9515	Cyanobacteria	Cyanobacteriia	45397
Alteromonas stellipolaris LMG 21856	Proteobacteria	Gammaproteobacteria	1160720
Shewanella sediminis HAW-EB3	Proteobacteria	Gammaproteobacteria	271097
Geoglobus ahangari strain 234	Halobacterota	Archaeoglobi	113653
Teredinibacter turnerae T7901	Proteobacteria	Gammaproteobacteria	377629
Shewanella denitrificans OS217	Proteobacteria	Gammaproteobacteria	318161
Prochlorococcus marinus bv. HNLC2	Cyanobacteria	Cyanobacteriia	1219
Pseudovibrio sp. FO-BEG1	Proteobacteria	Alphaproteobacteria	1909297
Nocardia seriolae strain EM150506	Actinobacteriota	Actinobacteria	37332
Vibrio anguillarum DSM 21597	Proteobacteria	Gammaproteobacteria	882102
Hermovibrio ammonificans HB-1	Aquificota	Desulfurobacteriia	228745
Prochlorococcus marinus subsp. pastoris str. CCMP1986	Cyanobacteria	Cyanobacteriia	142479
Marinobacter psychrophilus strain 20041	Proteobacteria	Gammaproteobacteria	330734
Mollicutes bacterium (Candidatus Izimaplasma) HR1	Firmicutes	Bacilli	37628
Phaeobacter inhibens P48	Proteobacteria	Alphaproteobacteria	999548
Vibrio campbellii (harveyi) ATCC BAA-1116	Proteobacteria	Gammaproteobacteria	314289
Vibrio anguillarum 90-11-287	Proteobacteria	Gammaproteobacteria	55601
Halogeometricum borinquense DSM 11551	Halobacterota	Halobacteria	469382
Halanaeroarchaeum sulfurireducens strain M27-SA2	Halobacterota	Halobacteria	1604004
Shewanella woodyi ATCC 51908	Proteobacteria	Gammaproteobacteria	392500
Methanosarcina siciliae T4/M	Halobacterota	Methanosarcinia	1434120
Bacillus infantis NRRL B-14911	Firmicutes	Bacilli	324767
Rhodobiaceae bacterium SMS8	Proteobacteria	Alphaproteobacteria	2026785
Campylobacter insulaenigrae strain NCTC12927	Campylobacterota	Campylobacteria	1031564
Gammaproteobacteria bacterium DM2	Proteobacteria	Gammaproteobacteria	1738444
Yangia sp. CCB-MM3	Proteobacteria	Alphaproteobacteria	2078585
Microbulbifer sp. A4B17	Proteobacteria	Gammaproteobacteria	359370
Ferroglobus placidus DSM 10642	Halobacterota	Archaeoglobi	589924
Marinobacter salarius strain HL2708#2	Proteobacteria	Gammaproteobacteria	1420917
Pseudoalteromonas piscicida DE2-A	Proteobacteria	Gammaproteobacteria	43662
Pseudoalteromonas atlantica T6c	Proteobacteria	Gammaproteobacteria	342610
Aureitalea sp. RR4-38	Bacteroidota	Bacteroidia	1872661
bacterium symbiont of Cryptopsaras couesii	Proteobacteria	Gammaproteobacteria	1927128
Flammeovirga sp. MY04	Bacteroidota	Bacteroidia	1978526
Vibrio owensii XSBZ03	Proteobacteria	Gammaproteobacteria	28173
Piscirickettsia salmonis strain PM25344B	Proteobacteria	Gammaproteobacteria	1238

Thalassospira sp. CSC3H3	Proteobacteria	Alphaproteobacteria	1912094
Yangia pacifica YSBP01	Proteobacteria	Alphaproteobacteria	311180
Altererythrobacter ishigakiensis strain NBRC 107699	Proteobacteria	Alphaproteobacteria	476157
Salinibacter ruber M1	Bacteroidota	Rhodothermia	761659
Cellulophaga lytica DSM 7489	Bacteroidota	Bacteroidia	867900
Vibrio sp. Ex25	Proteobacteria	Gammaproteobacteria	678
Methanocaldococcus jannaschii DSM 2661	Euryarchaeota	Methanococci	243232
Kosmotoga pacifica strain SLHLJ1	Thermotogota	Thermotogae	1330330
Candidatus Thioglobus sp. EF1	Proteobacteria	Gammaproteobacteria	2026721
Confluentimicrobium sp. EMB200-NS6	Proteobacteria	Alphaproteobacteria	1872125
Prochlorococcus marinus str. MIT 9303	Cyanobacteria	Cyanobacteriia	45397
Streptomyces sp. CMB-StM0423	Actinobacteriota	Actinobacteria	1931
Nocardiopsis dassonvillei strain NOCA502F	Actinobacteriota	Actinobacteria	2015
Halomonas sp. SF2003	Proteobacteria	Gammaproteobacteria	2136172
Myroides profundi D25	Bacteroidota	Bacteroidia	480520
Shewanella baltica OS195	Proteobacteria	Gammaproteobacteria	399599
Vibrio jasicida 090810c	Proteobacteria	Gammaproteobacteria	1280002
Aeromonas salmonicida S121	Proteobacteria	Gammaproteobacteria	645
Hirschia baltica ATCC 49814	Proteobacteria	Alphaproteobacteria	582402
Thermococcus barophilus CH5	Euryarchaeota	Thermococci	55802
Salinicoccus sp. BAB 3246	Firmicutes	Bacilli	1871624
Cyclobacterium amurskyense strain KCTC 12363	Bacteroidota	Bacteroidia	320787
Luteimonas sp. JM171	Proteobacteria	Gammaproteobacteria	1124597
Cellulophaga algicola DSM 14237	Bacteroidota	Bacteroidia	688270
Vibrio vulnificus FORC_053	Proteobacteria	Gammaproteobacteria	672
Nitratifractor salsuginis DSM 16511	Campylobacterota	Campylobacteria	749222
Thiomicrospira crunogena XCL-2	Proteobacteria	Gammaproteobacteria	39765
Bradymonas sediminis FA350	Myxococcota	Bradimonadia	1548548
Anoxybacter fermentans strain DY22613	Firmicutes_F	Halanaerobiia	1323375
Novosphingobium pentaromativorans US6-1	Proteobacteria	Alphaproteobacteria	205844
Salinibacter ruber P13	Bacteroidota	Rhodothermia	761659
Nonlabens sediminis NBRC 100970	Bacteroidota	Bacteroidia	323273
Archaeoglobus veneficus SNP6	Halobacterota	Archaeoglobi	693661
Paenibacillus durus DSM 1735	Firmicutes_I	Bacilli_A	44251
Alteromonas mediterranea strain CP48	Proteobacteria	Gammaproteobacteria	314275
Haloarcula hispanica ATCC 33960	Halobacterota	Halobacteria	634497
Actinoalloteichus sp. GBA129-24	Actinobacteriota	Actinobacteria	1872128
Bordetella sp. HZ20	Proteobacteria	Gammaproteobacteria	28081
Hyperthermus butylicus DSM 5456	Crenarchaeota	Thermoprotei	415426
Massilia sp. YMA4	Proteobacteria	Gammaproteobacteria	1882437
Bacillus velezensis strain 9912D	Firmicutes	Bacilli	492670
Pseudomonas pohangensis strain DSM 17875	Proteobacteria	Gammaproteobacteria	364197
Alteromonas macleodii ATCC 27126	Proteobacteria	Gammaproteobacteria	529120
Pseudanabaena sp. PCC 7367	Cyanobacteria	Cyanobacteriia	1153
Lacimicrobium alkaliphilum strain KCTC 32984	Proteobacteria	Gammaproteobacteria	1937692
Desulfobacterium autotrophicum HRM2	Desulfobacterota	Desulfobacteria	177437
Methanobacterium sp. MZ-A1	Euryarchaeota	Methanobacteria	2164
Maribacter sp. 1_2014MBL_MicDiv	Bacteroidota	Bacteroidia	1897614
Vibrio anguillarum VIB43	Proteobacteria	Gammaproteobacteria	55601
Paenibacillus sp. LPB0068	Firmicutes_I	Bacilli_A	58172
Saprospira grandis str. Lewin	Bacteroidota	Bacteroidia	1008
Methanohalophilus halophilus strain Z-7982	Halobacterota	Methanosarcinia	2177
Methanococcus vannielii SB	Euryarchaeota	Methanococci	406327
Vibrio alginolyticus strain ATCC 33787	Proteobacteria	Gammaproteobacteria	674977
Micromonospora krabiensis strain DSM 45344	Actinobacteriota	Actinobacteria	307121
Methanoculleus marisnigri JR1	Halobacterota	Methanomicrobia	368407
Catenovulum sp. CCB-QB4	Proteobacteria	Gammaproteobacteria	2172099
Cycloclasticus zancles 78-ME	Proteobacteria	Gammaproteobacteria	1329899
Methanococcus aeolicus Nankai-3	Euryarchaeota	Methanococci	42879
Candidatus Thioglobus singularis NP1	Proteobacteria	Gammaproteobacteria	1427364
Methanocaldococcus vulcanius M7	Euryarchaeota	Methanococci	579137
Methanothermococcus okinawensis IH1	Euryarchaeota	Methanococci	647113
Shewanella pealeana ATCC 700345	Proteobacteria	Gammaproteobacteria	398579
Zobellella denitrificans F13-1	Proteobacteria	Gammaproteobacteria	347534
Candidatus Vesicomyosocius okutanii HA	Proteobacteria	Gammaproteobacteria	412965
Vibrio campbellii BoB-53	Proteobacteria	Gammaproteobacteria	680
Candidatus Endolissoclinum faulkneri L5	Proteobacteria	Alphaproteobacteria	1401328
Candidatus Nitrosoarchaeum limnia SFB1	Crenarchaeota	Nitrososphaeria	886738
Erythrobacter sp. KY5	Proteobacteria	Alphaproteobacteria	1042
Vibrio coralliilyticus	Proteobacteria	Gammaproteobacteria	190893
Methanosarcina sp. WH1	Halobacterota	Methanosarcinia	2213
Thermotoga sp. RQ7	Thermotogota	Thermotogae	28240
Altererythrobacter epoxidivorans CGMCC 1.7731	Proteobacteria	Alphaproteobacteria	361183
Marinobacter sp. CP1	Proteobacteria	Gammaproteobacteria	50741
Vibrio vulnificus FORC_037	Proteobacteria	Gammaproteobacteria	672

Wenzhouxiangella marina strain KCTC 42284	Proteobacteria	Gammaproteobacteria	1579979
Halioglobus japonicus NBRC 107739	Proteobacteria	Gammaproteobacteria	930805
Paenibacillus sp. CAA11	Firmicutes_I	Bacilli_A	1532905
Aeromonas salmonicida S44	Proteobacteria	Gammaproteobacteria	645
Vibrio cholerae Sa5Y	Proteobacteria	Gammaproteobacteria	666
Archaeoglobus fulgidus DSM 4304	Halobacterota	Archaeoglobi	224325
Methanosarcina acetivorans C2A	Halobacterota	Methanosarcinia	188937
Vibrio campbellii	Proteobacteria	Gammaproteobacteria	680
Shewanella loihica PV-4	Proteobacteria	Gammaproteobacteria	359303
Marinobacterium sp. ST58-10	Proteobacteria	Gammaproteobacteria	1902815
Alteromonas macleodii str. 'Aegean Sea MED64'	Proteobacteria	Gammaproteobacteria	28108
Psychromonas sp. CNPT3	Proteobacteria	Gammaproteobacteria	1884585
Phaeobacter inhibens P92	Proteobacteria	Alphaproteobacteria	999548
Psychroflexus torquis ATCC 700755	Bacteroidota	Bacteroidia	313595
Nonlabens sp. Hel1_33_55	Bacteroidota	Bacteroidia	1888209
Thermococcus nautili strain 30-1	Euryarchaeota	Thermococci	195522
Flavivirga eckloniae ECD14	Bacteroidota	Bacteroidia	1803846
Weissella ceti strain WS08	Firmicutes	Bacilli	759620
Desulfovibrio salexigens DSM 2638	Desulfobacterota_A	Desulfovibrionia	526222
Roseobacter litoralis Och 149	Proteobacteria	Alphaproteobacteria	391595
Alteromonas macleodii str. 'Ionian Sea U4'	Proteobacteria	Gammaproteobacteria	28108
Marinobacter salarius R9SW1	Proteobacteria	Gammaproteobacteria	1420917
Pseudoalteromonas piscicida DE2-B	Proteobacteria	Gammaproteobacteria	43662
Halolamina aestuarii strain Hb3	Proteobacteria	Gammaproteobacteria	1480675
Thermotoga maritima MSB8	Thermotogota	Thermotogae	243274
Vibrio anguillarum JLL237	Proteobacteria	Gammaproteobacteria	55601
Prosthecochloris sp. GSB1	Bacteroidota	Chlorobia	290513
Altererythrobacter sp. B11	Proteobacteria	Alphaproteobacteria	1872480
Brucella ceti TE28753-12	Proteobacteria	Alphaproteobacteria	120577
Vibrio cholerae FORC_055	Proteobacteria	Gammaproteobacteria	666
Vibrio alginolyticus K06K5	Proteobacteria	Gammaproteobacteria	663
Vibrio breoganii strain FF50	Proteobacteria	Gammaproteobacteria	553239
Citromicrobium sp. JL477	Proteobacteria	Alphaproteobacteria	2024827
Polaribacter sp. LPB0003	Bacteroidota	Bacteroidia	1920175
Bacteriovorax marinus SJ	Bdellovibrionota	Bacteriovoracia	862908
Rhodococcus sp. H-CA8f	Actinobacteriota	Actinobacteria	1831
Spiribacter sp. UAH-SP71	Proteobacteria	Gammaproteobacteria	1930901
Actinoalloteichus sp. ADI127-17	Actinobacteriota	Actinobacteria	1872128
Rhodothermus marinus SG0.5JP17-172	Bacteroidota	Rhodothermia	29549
Edwardsiella tarda EIB202	Proteobacteria	Gammaproteobacteria	498217
Salinimonas sp. HMF8227	Proteobacteria	Gammaproteobacteria	1929415
Glaciecola sp. 4H-3-7+YE-5	Proteobacteria	Gammaproteobacteria	983545
Thermaerobacter marianensis DSM 12885	Firmicutes_E	Thermaerobacteria	644966
Campylobacter lari strain Slaughter Beach	Campylobacterota	Campylobacteria	201
Nonlabens spongiae JCM 13191	Bacteroidota	Bacteroidia	331648
Alteromonas macleodii str. 'Black Sea 11'	Proteobacteria	Gammaproteobacteria	28108
Cellulophaga baltica 18	Bacteroidota	Bacteroidia	1348584
Olleya aquimaris DAU311	Bacteroidota	Bacteroidia	639310
Staphylothermus marinus F1	Crenarchaeota	Thermoprotei	399550
Cycloclasticus sp. PY97N	Proteobacteria	Gammaproteobacteria	2024830
uncultured marine group II euryarchaeote	Thermoplasmatota	Poseidoniia	274854
Thiocystis violascens DSM 198	Proteobacteria	Gammaproteobacteria	765911
Kangiella geojedonensis strain YCS-5	Proteobacteria	Gammaproteobacteria	914150
Tistrella mobilis KA081020-065	Proteobacteria	Alphaproteobacteria	171437
Vibrio sp. EJY3	Proteobacteria	Gammaproteobacteria	689
Piscirickettsia salmonis PSCGR01	Proteobacteria	Gammaproteobacteria	1238
Vibrio anguillarum 425	Proteobacteria	Gammaproteobacteria	882102
Verrucosispora maris AB-18-032	Actinobacteriota	Actinobacteria	1003110
Halomonas sp. A3H3	Proteobacteria	Gammaproteobacteria	1486246
Vibrio alginolyticus K04M3	Proteobacteria	Gammaproteobacteria	663
Paenibacillus donghaensis KCTC 13049	Firmicutes_I	Bacilli_A	414771
Archaeoglobus profundus DSM 5631	Halobacterota	Archaeoglobi	572546
Rivularia sp. PCC 7116	Cyanobacteria	Cyanobacteriia	2047365
Thermococcus barophilus MP	Euryarchaeota	Thermococci	391623
Halobacteriovorax marinus BE01	Bdellovibrionota	Bacteriovoracia	97084
Thermococcus sp. AM4	Euryarchaeota	Thermococci	35749
Pyrococcus sp. NA2	Euryarchaeota	Thermococci	33866
Planktomarina temperata RCA23	Proteobacteria	Alphaproteobacteria	666509
Croceibacter atlanticus HTCC2559	Bacteroidota	Bacteroidia	216432
Vibrio anguillarum VA1	Proteobacteria	Gammaproteobacteria	55601
Phaeobacter gallaeciensis P73	Proteobacteria	Alphaproteobacteria	60890
Nonlabens dokdonensis DSW-6	Bacteroidota	Bacteroidia	328515
Thermotoga maritima MSB8	Thermotogota	Thermotogae	243274
Vibrio anguillarum Ba35	Proteobacteria	Gammaproteobacteria	55601
Echinicola strongylocentroti MEBiC08714	Bacteroidota	Bacteroidia	1795355

Helicobacter sp. MIT 01-6242	Campylobacterota	Campylobacteria	218
Vibrio anguillarum 91-7154	Proteobacteria	Gammaproteobacteria	55601
Flavobacterium psychrophilum strain PG2	Bacteroidota	Bacteroidia	96345
Edwardsiella tarda	Proteobacteria	Gammaproteobacteria	636
Aciduliprofundum boonei T469	Thermoplasmatota	Thermoplasmata	439481
Streptomyces niveus SCSIO 3406	Actinobacteriota	Actinobacteria	193462
Alcanivorax sp. E4	Proteobacteria	Gammaproteobacteria	1799644
Flexistipes sinusarabici DSM 4947	Deferribacterota	Deferribacteres	717231
Moritella yayanosii DB21MT 5	Proteobacteria	Gammaproteobacteria	69539
Nautilia profundicola AmH	Campylobacterota	Campylobacteria	598659
Vibrio alginolyticus K01M1	Proteobacteria	Gammaproteobacteria	663
Roseovarius sp. AK1035	Proteobacteria	Alphaproteobacteria	1486281
Vibrio scophthalmi strain VS-12	Proteobacteria	Gammaproteobacteria	45658
Myxococcus fulvus HW-1	Myxococcota	Myxococcia	33
Vibrio owensii 1700302	Proteobacteria	Gammaproteobacteria	696485
Dokdonia donghaensis DSW-1	Bacteroidota	Bacteroidia	326320
Flavobacterium sp. MEBiC07310	Bacteroidota	Bacteroidia	239
Microcella alkaliphila JAM AC0309	Actinobacteriota	Actinobacteria	279828
Methanocaldococcus sp. JH146	Euryarchaeota	Methanococci	2152917
Aeropyrum camini SY1 = JCM 12091	Crenarchaeota	Thermoprotei	1198449
bacterium AB1 strain AB1-8	UBP7	UBA6624	1898108
Brucella sp. 141012304	Proteobacteria	Alphaproteobacteria	52132
Synechococcus sp. WH 8109	Cyanobacteria	Cyanobacteriia	1131
Vibrio anguillarum 261/91	Proteobacteria	Gammaproteobacteria	990314
alpha proteobacterium HIMB5	Proteobacteria	Alphaproteobacteria	859653
Prochlorococcus marinus str. NATL2A	Cyanobacteria	Cyanobacteriia	1219
Pelagibaca abyssi JLT2014	Proteobacteria	Alphaproteobacteria	1250539
Salinibacter ruber SP38	Bacteroidota	Rhodothermia	761659
Sphingorhabdus flavimaris YGSMI21	Proteobacteria	Alphaproteobacteria	266812
Desulfocapsa sulfexigens DSM 10523	Desulfobacterota	Desulfobulbia	1167006
Candidatus Puniceispirillum marinum IMCC1322	Proteobacteria	Alphaproteobacteria	488538
Octadecabacter antarcticus 307	Proteobacteria	Alphaproteobacteria	391626
Flavobacterium psychrophilum V3-5	Bacteroidota	Bacteroidia	96345
Moorea producens PAL-8-15-08-1	Cyanobacteria	Cyanobacteriia	1155739
Endosymbiont of unidentified scaly snail isolate Monju	Proteobacteria	Gammaproteobacteria	1248727
Vibrio vulnificus CECT 4999	Proteobacteria	Gammaproteobacteria	1051646
Gramella sp. LPB0144	Bacteroidota	Bacteroidia	1931228
Planococcus maritimus strain DSM 17275	Firmicutes	Bacilli	192421
Flavobacterium psychrophilum FPG101	Bacteroidota	Bacteroidia	1452725
Vibrio vulnificus FORC_036	Proteobacteria	Gammaproteobacteria	216895
Halomonas elongata DSM 2581	Proteobacteria	Gammaproteobacteria	768066
Idiomarina loihiensis L2TR	Proteobacteria	Gammaproteobacteria	283942
Thermococcus sp. 4557	Euryarchaeota	Thermococci	35749
Thermococcus sp. CL1	Euryarchaeota	Thermococci	35749
Micromonospora tulbaghiae CNY-010	Actinobacteriota	Actinobacteria	479978
Klebsiella pneumoniae subsp. pneumoniae KC-Pl-HB1	Proteobacteria	Gammaproteobacteria	573
Shewanella frigidimarina NCIMB 400	Proteobacteria	Gammaproteobacteria	318167
Archaeoglobus fulgidus DSM 8774	Halobacterota	Archaeoglobi	1344584
Aeropyrum pernix K1	Crenarchaeota	Thermoprotei	272557
Leisingera methylohalidivorans DSM 14336	Proteobacteria	Alphaproteobacteria	1246
Phaeobacter inhibens P54	Proteobacteria	Alphaproteobacteria	999548
Verrucomicrobia bacterium L21-Fru-AB	Verrucomicrobiota	Kiritimatiellae	2026799
Phaeobacter piscinae P13	Proteobacteria	Alphaproteobacteria	1580596
Vibrio anguillarum S3 4/9	Proteobacteria	Gammaproteobacteria	882944
Desulfotalea psychrophila LSv54	Desulfobacterota	Desulfobulbia	177439
Cyanothece sp. ATCC 51142	Cyanobacteria	Cyanobacteriia	2649277
Oleiphilus messinensis ME102	Proteobacteria	Gammaproteobacteria	141451
Shewanella benthica DB21MT-2	Proteobacteria	Gammaproteobacteria	43661
Altererythrobacter marensis strain KCTC 22370	Proteobacteria	Alphaproteobacteria	543877
Pseudomonas litoralis strain 2SM5	Proteobacteria	Gammaproteobacteria	797277
Saccharophagus degradans 2-40	Proteobacteria	Gammaproteobacteria	86304
Vibrio anguillarum CNEVA NB11008	Proteobacteria	Gammaproteobacteria	55601
Thermotoga maritima MSB8	Thermotogota	Thermotogae	243274
Flavobacterium psychrophilum V4-24	Bacteroidota	Bacteroidia	96345
Nautilia profundicola strain PV-1	Campylobacterota	Campylobacteria	244787
Clostridiales bacterium 70B-A	Firmicutes_A	Clostridia	
Tateyamaria omphalii DOK1-4	Proteobacteria	Alphaproteobacteria	299262
Salinispora tropica CNB-440	Actinobacteriota	Actinobacteria	168695
Marinifilaceae bacterium SPP2	Bacteroidota	Bacteroidia	869210
Alteromonas stellipolaris strain PQQ-42	Proteobacteria	Gammaproteobacteria	233316
Algibacter sp. HZ22	Bacteroidota	Bacteroidia	1872428
Methanococcus maripaludis C6	Euryarchaeota	Methanococci	444158
Francisella sp. TX077310	Proteobacteria	Gammaproteobacteria	2047875
Chromobacterium sp. IIBBL 274-1	Proteobacteria	Gammaproteobacteria	306190
Vibrio natriegens strain CCUG 16374	Proteobacteria	Gammaproteobacteria	691

Weissella ceti strain WS105	Firmicutes	Bacilli	759620
Streptococcus iniae QMA0248	Firmicutes	Bacilli	1346
Phaeobacter inhibens DOK1-1	Proteobacteria	Alphaproteobacteria	221822
Vibrio anguillarum strain 90-11-286	Proteobacteria	Gammaproteobacteria	55601
Cellulophaga lytica strain HI1	Bacteroidota	Bacteroidia	979
Psychromonas ingrahamii 37	Proteobacteria	Gammaproteobacteria	357804
Oleispira antarctica RB-8	Proteobacteria	Gammaproteobacteria	188908
Methylocaldum marinum S8	Proteobacteria	Gammaproteobacteria	1432792
Polaribacter sp. MED152	Bacteroidota	Bacteroidia	1920175
Neorickettsia helminthoeca str. Oregon	Proteobacteria	Alphaproteobacteria	33994
Tenacibaculum dicentrarchi strain AY7486TD	Bacteroidota	Bacteroidia	669041
Spongiibacter sp. IMCC21906	Proteobacteria	Gammaproteobacteria	2024860
Vibrio anguillarum 87-9-116	Proteobacteria	Gammaproteobacteria	55601
Donghicola sp. JLT3646	Proteobacteria	Alphaproteobacteria	1929294
Serinicoccus sp. JLT9	Actinobacteriota	Actinobacteria	1871625
Chloroherpeton thalassium ATCC 35110	Bacteroidota	Chlorobia	517418
Candidatus Thioglobus singularis PS1	Proteobacteria	Gammaproteobacteria	1125411
Wenyingzhuangia fucanilytica strain CZ1127	Bacteroidota	Bacteroidia	1790137
Rhodoferax ferrireducens T118	Proteobacteria	Gammaproteobacteria	338969
Rhodovulum sp. MB263	Proteobacteria	Alphaproteobacteria	34009
Mycobacterium stephanolepidis NJB0901	Actinobacteriota	Actinobacteria	1520670
Vibrio tubiashii ATCC 19109	Proteobacteria	Gammaproteobacteria	678
Vibrio natriegens NBRC 15636 = ATCC 14048 = DSM 759	Proteobacteria	Gammaproteobacteria	1219067
Bacterioplanes sanyensis NV9	Proteobacteria	Gammaproteobacteria	1249553
Alteromonas sp. RKMC-009	Proteobacteria	Gammaproteobacteria	232
Pontibacter actiniarum DSM 19842	Bacteroidota	Bacteroidia	323450
Phaeobacter piscinae P71	Proteobacteria	Alphaproteobacteria	1580596
Jeotgalibacillus sp. D5	Firmicutes	Bacilli	1898383
Pseudomonas plecoglossicida XSDHY-P	Proteobacteria	Gammaproteobacteria	70775
Thalassolituus oleivorans MIL-1	Proteobacteria	Gammaproteobacteria	187493
Actinoalloteichus sp. AHMU CJ021	Actinobacteriota	Actinobacteria	2072503
Shewanella livingstonensis strain LMG 19866	Proteobacteria	Gammaproteobacteria	150120
Pseudoalteromonas agarivorans DSM 14585	Proteobacteria	Gammaproteobacteria	1312369
Vibrio anguillarum PF4	Proteobacteria	Gammaproteobacteria	55601
Prochlorococcus marinus AS9601	Cyanobacteria	Cyanobacteriia	1219
Vibrio coralliilyticus strain 58	Proteobacteria	Gammaproteobacteria	909421
Phaeobacter gallaeciensis P11	Proteobacteria	Alphaproteobacteria	60890
Piscirickettsia salmonis strain PM31429B	Proteobacteria	Gammaproteobacteria	1238
Formosa sp. Hel1_33_131	Bacteroidota	Bacteroidia	2018467
Methylophilales bacterium MBRSH7	Proteobacteria	Gammaproteobacteria	2546201
Vibrio parahaemolyticus strain CHN25	Proteobacteria	Gammaproteobacteria	1211705
Vibrio alginolyticus K08M3	Proteobacteria	Gammaproteobacteria	663
Marinobacter sp. Hb8	Proteobacteria	Gammaproteobacteria	50741
Planococcus plakortidis strain DSM 23997	Firmicutes	Bacilli	1038856
Halobacillus mangrovi KTB 131	Firmicutes	Bacilli	402384
Vibrio campbellii DS40M4	Proteobacteria	Gammaproteobacteria	680
Erythrobacter seohaensis strain SW-135	Proteobacteria	Alphaproteobacteria	266951
Haloquadratum walsbyi DSM 16790	Halobacterota	Halobacteria	362976
Marivirga tractuosa DSM 4126	Bacteroidota	Bacteroidia	643867
Thermoplasma volcanium GSS1	Thermoplasmatota	Thermoplasmata	273116
Phaeobacter inhibens BS107	Proteobacteria	Alphaproteobacteria	221822
Psychrobacter sp. P11F6	Proteobacteria	Gammaproteobacteria	56811
Vibrio anguillarum 9014/8	Proteobacteria	Gammaproteobacteria	990314
Candidatus Pelagibacter sp. IMCC9063	Proteobacteria	Alphaproteobacteria	2024849
Marinithermus hydrothermalis DSM 14884	Deinococcota	Deinococci	443254
Streptococcus parauberis KCTC 11537	Firmicutes	Bacilli	936154
Aeromonas salmonicida subsp. salmonicida A449	Proteobacteria	Gammaproteobacteria	29491
Cellulophaga lytica strain DAU203	Bacteroidota	Bacteroidia	979
Thermotoga sp. 2812B	Thermotogota	Thermotogae	28240
Tenacibaculum mesophilum strain DSM 13764	Bacteroidota	Bacteroidia	104268
Rhodothermus marinus DSM 4252	Bacteroidota	Rhodothermia	518766
Salinispora arenicola CNS-205	Actinobacteriota	Actinobacteria	168697
Vibrio alginolyticus K08M4	Proteobacteria	Gammaproteobacteria	663
Halorubrum sp. PV6	Halobacterota	Halobacteria	634157
Nodularia spumigena UHCC 0039	Cyanobacteria	Cyanobacteriia	1914872
Rhodobacter sphaeroides strain AB25	Proteobacteria	Alphaproteobacteria	1063
Cellvibrionaceae bacterium 017	Proteobacteria	Gammaproteobacteria	2026723
Vibrio alginolyticus K10K4	Proteobacteria	Gammaproteobacteria	663
Alteromonas macleodii str. ’English Channel 615’	Proteobacteria	Gammaproteobacteria	28108
Vibrio natriegens NBRC 15636	Proteobacteria	Gammaproteobacteria	1889773
Prosthecochloris sp. CIB 2401	Bacteroidota	Chlorobia	290513
Haloquadratum walsbyi C23	Halobacterota	Halobacteria	768065
Alcaligenes aquatilis QD168	Proteobacteria	Gammaproteobacteria	323284
Candidatus Moanabacter tarae TARA_B100001123	Verrucomicrobiota	Verrucomicrobiae	2200854
Thalassospira indica PB8B	Proteobacteria	Alphaproteobacteria	1891279

Flagellimonas sp. HME9304	Bacteroidota	Bacteroidia	2058762
Cohaesibacter sp. ES.047	Proteobacteria	Alphaproteobacteria	2026570
Streptomyces luteoverticillatus strain CGMCC 15060	Actinobacteriota	Actinobacteria	66425
Thermosipho sp. 1070	Thermotogota	Thermotogae	1968895
Marinomonas posidonica IVIA-Po-181	Proteobacteria	Gammaproteobacteria	936476
Micromonospora aurantiaca 110B(2018)	Actinobacteriota	Actinobacteria	47850
Sulfobacillus acidophilus TPY	Firmicutes_E	Sulfobacillia	1051632
Anderseniella sp. Alg231-50	Proteobacteria	Alphaproteobacteria	1922226

Table A.4 Reference taxa included in prokaryotic phylogenomic tree construction (Figure 4.12).

A.5 Reference Taxa in Eukaryotic Tree

Name	Taxonomy	Assembly	Source
Emiliania huxleyi CCMP1516	Eukaryota;Protists;Other Protists	GCA_000372725.1	NCBI
Leishmania major strain Friedlin	Eukaryota;Protists;Kinetoplasts	GCA_000002725.2	NCBI
Trypanosoma brucei gambiense DAL972	Eukaryota;Protists;Kinetoplasts	GCA_000210295.1	NCBI
Trypanosoma cruzi	Eukaryota;Protists;Kinetoplasts	GCA_000209065.1	NCBI
Giardia lamblia ATCC 50803	Eukaryota;Protists;Other Protists	GCA_000002435.1	NCBI
Entamoeba histolytica HM-1:IMSS	Eukaryota;Protists;Other Protists	GCA_000208925.2	NCBI
Eimeria tenella	Eukaryota;Protists;Apicomplexans	GCA_000499545.1	NCBI
Cryptosporidium parvum Iowa II	Eukaryota;Protists;Apicomplexans	GCA_000165345.1	NCBI
Plasmodium chabaudi chabaudi	Eukaryota;Protists;Apicomplexans	GCA_900002335.1	NCBI
Toxoplasma gondii ME49	Eukaryota;Protists;Apicomplexans	GCA_000006565.2	NCBI
Plasmodium berghei ANKA	Eukaryota;Protists;Apicomplexans	GCA_900002375.1	NCBI
Plasmodium knowlesi strain H	Eukaryota;Protists;Apicomplexans	GCA_000006355.1	NCBI
Plasmodium vivax	Eukaryota;Protists;Apicomplexans	GCA_000002415.2	NCBI
Babesia bovis	Eukaryota;Protists;Apicomplexans	GCA_000165395.1	NCBI
Theileria annulata	Eukaryota;Protists;Apicomplexans	GCA_000003225.1	NCBI
Theileria parva	Eukaryota;Protists;Apicomplexans	GCA_000165365.1	NCBI
Plasmodium falciparum 3D7	Eukaryota;Protists;Apicomplexans	GCA_000002765.2	NCBI
Dictyostelium discoideum AX4	Eukaryota;Protists;Other Protists	GCA_000004695.1	NCBI
Plasmodium yoelii	Eukaryota;Protists;Apicomplexans	GCA_900002385.1	NCBI
Phytophthora sojae	Eukaryota;Protists;Other Protists	GCA_000149755.2	NCBI
Tetrahymena thermophila SB210	Eukaryota;Protists;Other Protists	GCA_000189635.1	NCBI
Phytophthora ramorum	Eukaryota;Protists;Other Protists	GCA_002968915.1	NCBI
Plasmodium reichenowi	Eukaryota;Protists;Apicomplexans	GCA_001601855.1	NCBI
Neospora caninum Liverpool	Eukaryota;Protists;Apicomplexans	GCA_000208865.2	NCBI
Leishmania infantum JPCM5	Eukaryota;Protists;Kinetoplasts	GCA_000002875.2	NCBI
Trichomonas vaginalis G3	Eukaryota;Protists;Other Protists	GCA_000002825.1	NCBI
Naegleria gruberi	Eukaryota;Protists;Other Protists	GCA_000004985.1	NCBI
Physarum polycephalum	Eukaryota;Protists;Other Protists	GCA_000413255.3	NCBI
Paramecium tetraurelia	Eukaryota;Protists;Other Protists	GCA_000165425.1	NCBI
Acanthamoeba castellanii str. Neff	Eukaryota;Protists;Other Protists	GCA_000313135.1	NCBI
Perkinsus marinus ATCC 50983	Eukaryota;Protists;Other Protists	GCA_000006405.1	NCBI
Phytophthora infestans T30-4	Eukaryota;Protists;Other Protists	GCA_000142945.1	NCBI
Blastocystis hominis	Eukaryota;Protists;Other Protists	GCA_000151665.1	NCBI
Cyanophora paradoxa	Eukaryota;Protists;Other Protists	GCA_004431415.1	NCBI
Euglena gracilis	Eukaryota;Protists;Other Protists	GCA_900893395.1	NCBI
Ichthyophthirius multifiliis	Eukaryota;Protists;Other Protists	GCA_000220395.1	NCBI
Sterkiella histriomuscorum	Eukaryota;Protists;Other Protists	GCA_001273305.2	NCBI
Entamoeba invadens IP1	Eukaryota;Protists;Other Protists	GCA_000330505.1	NCBI
Entamoeba dispar SAW760	Eukaryota;Protists;Other Protists	GCA_000209125.2	NCBI
Aureococcus anophagefferens	Eukaryota;Protists;Other Protists	GCA_000186865.1	NCBI
Monosiga brevicollis MX1	Eukaryota;Protists;Other Protists	GCA_000002865.1	NCBI
Leishmania braziliensis MHOM/BR/75/M2904	Eukaryota;Protists;Kinetoplasts	GCA_000002845.2	NCBI
Capsaspora owczarzaki ATCC 30864	Eukaryota;Protists;Other Protists	GCA_000151315.2	NCBI
Ascogregarina taiwanensis	Eukaryota;Protists;Apicomplexans	GCA_000172235.1	NCBI
Cryptosporidium muris RN66	Eukaryota;Protists;Apicomplexans	GCA_000006515.1	NCBI
Cavenderia fasciculata	Eukaryota;Protists;Other Protists	GCA_000203815.1	NCBI
Hyaloperonospora arabidopsidis Emoy2	Eukaryota;Protists;Other Protists	GCA_000173235.2	NCBI
Saprolegnia parasitica CBS 223.65	Eukaryota;Protists;Other Protists	GCA_000151545.2	NCBI
Dictyostelium firmibasis	Eukaryota;Protists;Other Protists	GCA_000277485.1	NCBI
Dictyostelium citrinum	Eukaryota;Protists;Other Protists	GCA_000286055.1	NCBI
Dictyostelium intermedium	Eukaryota;Protists;Other Protists	GCA_000277465.1	NCBI
Polysphondylium violaceum	Eukaryota;Protists;Other Protists	GCA_000277445.1	NCBI
Astrammina rara	Eukaryota;Protists;Other Protists	GCA_000211355.2	NCBI

Thecamonas trahens ATCC 50062	Eukaryota;Protists;Other Protists	GCA_000142905.1	NCBI
Dictyostelium purpureum	Eukaryota;Protists;Other Protists	GCA_000190715.1	NCBI
Gregarina niphandrodes	Eukaryota;Protists;Apicomplexans	GCA_000223845.4	NCBI
Leishmania donovani	Eukaryota;Protists;Kinetoplasts	GCA_000227135.2	NCBI
Phytophthora cinnamomi	Eukaryota;Protists;Other Protists	GCA_001314365.1	NCBI
Plasmodium fragile	Eukaryota;Protists;Apicomplexans	GCA_000956335.1	NCBI
Paramecium caudatum	Eukaryota;Protists;Other Protists	GCA_000715435.1	NCBI
Plasmodium gallinaceum	Eukaryota;Protists;Apicomplexans	GCA_900005855.1	NCBI
Tetrahymena malaccensis 436	Eukaryota;Protists;Other Protists	GCA_000231845.2	NCBI
Proteromonas lacertae	Eukaryota;Protists;Other Protists	GCA_002245135.1	NCBI
Tetrahymena borealis	Eukaryota;Protists;Other Protists	GCA_000260095.1	NCBI
Tetrahymena elliotti 4EA	Eukaryota;Protists;Other Protists	GCA_000231825.2	NCBI
Trypanosoma congolense	Eukaryota;Protists;Kinetoplasts	GCA_002287245.1	NCBI
Trypanosoma vivax Y486	Eukaryota;Protists;Kinetoplasts	GCA_000227375.1	NCBI
Sarcocystis neurona	Eukaryota;Protists;Apicomplexans	GCA_000727475.1	NCBI
Mastigamoeba balamuthi ATTC 30984	Eukaryota;Protists;Other Protists	GCA_000765095.1	NCBI
Entamoeba moshkovskii	Eukaryota;Protists;Other Protists	GCA_002914575.1	NCBI
Trypanosoma rangeli	Eukaryota;Protists;Kinetoplasts	GCA_003719475.1	NCBI
Sphaeroforma arctica JP610	Eukaryota;Protists;Other Protists	GCA_001186125.1	NCBI
Plasmopara viticola	Eukaryota;Protists;Other Protists	GCA_001695595.3	NCBI
Theileria equi strain WA	Eukaryota;Protists;Apicomplexans	GCA_000342415.1	NCBI
Reticulomyxa filosa	Eukaryota;Protists;Other Protists	GCA_000512085.1	NCBI
Crithidia mellificae	Eukaryota;Protists;Kinetoplasts	GCA_002216565.1	NCBI
Nannochloropsis limnetica	Eukaryota;Protists;Other Protists	GCA_001614225.1	NCBI
Nannochloropsis oculata CCMP525	Eukaryota;Protists;Other Protists	GCA_004335455.1	NCBI
Nannochloropsis gaditana CCMP526	Eukaryota;Protists;Other Protists	GCA_000240725.1	NCBI
Nannochloropsis granulata CCMP529	Eukaryota;Protists;Other Protists	GCA_004335405.1	NCBI
Entamoeba nuttalli P19	Eukaryota;Protists;Other Protists	GCA_000257125.1	NCBI
Babesia microti strain RI	Eukaryota;Protists;Apicomplexans	GCA_000691945.2	NCBI
Phytophthora parasitica INRA-310	Eukaryota;Protists;Other Protists	GCA_000247585.2	NCBI
Chromera velia	Eukaryota;Protists;Other Protists	GCA_000585135.1	NCBI
Pseudoperonospora cubensis	Eukaryota;Protists;Other Protists	GCA_000252605.1	NCBI
Hammondia hammondi	Eukaryota;Protists;Apicomplexans	GCA_000447165.1	NCBI
Phytomonas serpens 9T	Eukaryota;Protists;Kinetoplasts	GCA_000331125.1	NCBI
Eimeria maxima	Eukaryota;Protists;Apicomplexans	GCA_000499605.1	NCBI
Bremia lactucae	Eukaryota;Protists;Other Protists	GCA_004359215.1	NCBI
Eimeria acervulina	Eukaryota;Protists;Apicomplexans	GCA_000499425.1	NCBI
Fonticula alba	Eukaryota;Protists;Other Protists	GCA_000388065.2	NCBI
Nannochloropsis oceanica	Eukaryota;Protists;Other Protists	GCA_004519485.1	NCBI
Leishmania amazonensis	Eukaryota;Protists;Kinetoplasts	GCA_005317125.1	NCBI
Vitrella brassicaformis CCMP3155	Eukaryota;Protists;Other Protists	GCA_001179505.1	NCBI
Oxytricha trifallax	Eukaryota;Protists;Other Protists	GCA_000711775.1	NCBI
Plasmodium vinckei vinckei	Eukaryota;Protists;Apicomplexans	GCA_000709005.1	NCBI
Blastocystis sp. subtype 4	Eukaryota;Protists;Other Protists	GCA_000743755.1	NCBI
Symbiodinium sp. clade A Y106	Eukaryota;Protists;Other Protists	GCA_003297005.1	NCBI
Plasmodium sp. gorilla clade G2	Eukaryota;Protists;Apicomplexans	GCA_900097015.1	NCBI
Endotrypanum monterogeii	Eukaryota;Protists;Kinetoplasts	GCA_000333855.2	NCBI
Leishmania panamensis	Eukaryota;Protists;Kinetoplasts	GCA_000755165.1	NCBI
Plasmodium cynomolgi strain B	Eukaryota;Protists;Apicomplexans	GCA_000321355.1	NCBI
Saprolegnia diclina VS20	Eukaryota;Protists;Other Protists	GCA_000281045.1	NCBI
Angomonas deanei	Eukaryota;Protists;Kinetoplasts	GCA_001659865.1	NCBI
Pythium iwayamai DAOM BR242034	Eukaryota;Protists;Other Protists	GCA_000387465.2	NCBI
Pythium aphanidermatum DAOM BR444	Eukaryota;Protists;Other Protists	GCA_000387445.2	NCBI
Pythium arrhenomanes ATCC 12531	Eukaryota;Protists;Other Protists	GCA_000387505.2	NCBI
Pythium irregulare DAOM BR486	Eukaryota;Protists;Other Protists	GCA_000387425.2	NCBI
Achlya hypogyna	Eukaryota;Protists;Other Protists	GCA_002081595.1	NCBI
Thraustotheca clavata	Eukaryota;Protists;Other Protists	GCA_002081575.1	NCBI
Leishmania aethiopica L147	Eukaryota;Protists;Kinetoplasts	GCA_000444285.2	NCBI
Leishmania tropica L590	Eukaryota;Protists;Kinetoplasts	GCA_000410715.1	NCBI
Leishmania mexicana MHOM/GT/2001/U1103	Eukaryota;Protists;Kinetoplasts	GCA_000234665.4	NCBI
Spironucleus salmonicida	Eukaryota;Protists;Other Protists	GCA_000497125.1	NCBI
Stylonychia lemnae	Eukaryota;Protists;Other Protists	GCA_000751175.1	NCBI
Phytophthora capsici LT1534	Eukaryota;Protists;Other Protists	GCA_000325885.1	NCBI
Theileria orientalis strain Shintoku	Eukaryota;Protists;Apicomplexans	GCA_000740895.1	NCBI
Crithidia fasciculata	Eukaryota;Protists;Kinetoplasts	GCA_000331325.2	NCBI
Strigomonas culicis	Eukaryota;Protists;Kinetoplasts	GCA_000482145.1	NCBI
Babesia bigemina	Eukaryota;Protists;Apicomplexans	GCA_000981445.1	NCBI
Plasmodium inui San Antonio 1	Eukaryota;Protists;Apicomplexans	GCA_000524495.1	NCBI
Phytophthora lateralis MPF4	Eukaryota;Protists;Other Protists	GCA_000318465.2	NCBI
Phytophthora kernoviae	Eukaryota;Protists;Other Protists	GCA_000448265.2	NCBI
Aphanomyces astaci	Eukaryota;Protists;Other Protists	GCA_000520075.1	NCBI
Aphanomyces invadans	Eukaryota;Protists;Other Protists	GCA_000520115.1	NCBI
Pythium splendens	Eukaryota;Protists;Other Protists	GCA_006386115.1	NCBI
Phytophthora cambivora	Eukaryota;Protists;Other Protists	GCA_000443045.1	NCBI
Phytophthora cryptogea	Eukaryota;Protists;Other Protists	GCA_000468175.2	NCBI

Phytophthora pinifolia	Eukaryota;Protists;Other Protists	GCA_000500225.2	NCBI
Leishmania enriettii	Eukaryota;Protists;Kinetoplasts	GCA_000410755.2	NCBI
Plasmodium relictum	Eukaryota;Protists;Apicomplexans	GCA_900005765.1	NCBI
Naegleria fowleri	Eukaryota;Protists;Other Protists	GCA_000499105.1	NCBI
Angomonas desouzai	Eukaryota;Protists;Kinetoplasts	GCA_000482185.1	NCBI
Leishmania guyanensis	Eukaryota;Protists;Kinetoplasts	GCA_003664525.1	NCBI
Cryptosporidium meleagridis	Eukaryota;Protists;Apicomplexans	GCA_001593445.1	NCBI
Eimeria necatrix	Eukaryota;Protists;Apicomplexans	GCA_000499385.1	NCBI
Eimeria brunetti	Eukaryota;Protists;Apicomplexans	GCA_000499725.1	NCBI
Eimeria mitis	Eukaryota;Protists;Apicomplexans	GCA_000499745.1	NCBI
Eimeria praecox	Eukaryota;Protists;Apicomplexans	GCA_000499445.1	NCBI
Salpingoeca rosetta	Eukaryota;Protists;Other Protists	GCA_000188695.1	NCBI
Strigomonas galati	Eukaryota;Protists;Kinetoplasts	GCA_000482125.1	NCBI
Strigomonas oncopelti	Eukaryota;Protists;Kinetoplasts	GCA_000482165.1	NCBI
Herpetomonas muscarum	Eukaryota;Protists;Kinetoplasts	GCA_000482205.1	NCBI
Crithidia acanthocephali	Eukaryota;Protists;Kinetoplasts	GCA_000482105.1	NCBI
Leishmania turanica	Eukaryota;Protists;Kinetoplasts	GCA_000441995.1	NCBI
Leishmania gerbilli	Eukaryota;Protists;Kinetoplasts	GCA_000443025.1	NCBI
Albugo candida	Eukaryota;Protists;Other Protists	GCA_001306755.1	NCBI
Hyphochytrium catenoides	Eukaryota;Protists;Other Protists	GCA_900088475.1	NCBI
Leishmania sp. AIIMS/LM/SS/PKDL/LD-974	Eukaryota;Protists;Kinetoplasts	GCA_000981925.2	NCBI
Leishmania arabica	Eukaryota;Protists;Kinetoplasts	GCA_000410695.2	NCBI
Heterococcus sp. DN1	Eukaryota;Protists;Other Protists	GCA_000498555.1	NCBI
Plasmodium gaboni	Eukaryota;Protists;Apicomplexans	GCA_001602025.1	NCBI
Phytophthora fragariae	Eukaryota;Protists;Other Protists	GCA_000686205.4	NCBI
Phytophthora rubi	Eukaryota;Protists;Other Protists	GCA_000687305.2	NCBI
Trypanosoma grayi	Eukaryota;Protists;Kinetoplasts	GCA_000691245.1	NCBI
Plasmodium coatneyi	Eukaryota;Protists;Apicomplexans	GCA_001680005.1	NCBI
Paramecium sexaurelia	Eukaryota;Protists;Other Protists	GCA_000733375.1	NCBI
Paramecium biaurelia	Eukaryota;Protists;Other Protists	GCA_000733385.1	NCBI
Phytophthora pisi	Eukaryota;Protists;Other Protists	GCA_000751395.2	NCBI
Phytomonas sp. isolate EM1	Eukaryota;Protists;Kinetoplasts	GCA_000582765.1	NCBI
Pythium insidiosum	Eukaryota;Protists;Other Protists	GCA_001029375.1	NCBI
Cyclospora cayetanensis	Eukaryota;Protists;Apicomplexans	GCA_002999335.1	NCBI
Acytostelium subglobosum LB1	Eukaryota;Protists;Other Protists	GCA_000787575.2	NCBI
Phytopythium vexans	Eukaryota;Protists;Other Protists	GCA_003413675.1	NCBI
Schizochytrium sp. CCTCC M209059	Eukaryota;Protists;Other Protists	GCA_000818945.1	NCBI
Babesia divergens	Eukaryota;Protists;Apicomplexans	GCA_001077455.2	NCBI
Acanthamoeba polyphaga	Eukaryota;Protists;Other Protists	GCA_001567625.1	NCBI
Acanthamoeba royreba	Eukaryota;Protists;Other Protists	GCA_000826365.1	NCBI
Acanthamoeba rhysodes	Eukaryota;Protists;Other Protists	GCA_000826385.1	NCBI
Acanthamoeba divionensis	Eukaryota;Protists;Other Protists	GCA_000826405.1	NCBI
Acanthamoeba lugdunensis	Eukaryota;Protists;Other Protists	GCA_000826425.1	NCBI
Acanthamoeba quina	Eukaryota;Protists;Other Protists	GCA_000826445.1	NCBI
Acanthamoeba mauritaniensis	Eukaryota;Protists;Other Protists	GCA_000826465.1	NCBI
Acanthamoeba pearcei	Eukaryota;Protists;Other Protists	GCA_000826505.1	NCBI
Eimeria nieschulzi	Eukaryota;Protists;Apicomplexans	GCA_000826945.1	NCBI
Acanthamoeba lenticulata	Eukaryota;Protists;Other Protists	GCA_002179805.1	NCBI
Acanthamoeba healyi	Eukaryota;Protists;Other Protists	GCA_000826305.1	NCBI
Acanthamoeba palestinensis	Eukaryota;Protists;Other Protists	GCA_000826325.1	NCBI
Acanthamoeba astronyxis	Eukaryota;Protists;Other Protists	GCA_000826245.1	NCBI
Acanthamoeba culbertsoni	Eukaryota;Protists;Other Protists	GCA_000826265.1	NCBI
Cryptosporidium sp. chipmunk LX-2015	Eukaryota;Protists;Apicomplexans	GCA_000831705.1	NCBI
Lotmaria passim	Eukaryota;Protists;Kinetoplasts	GCA_000635995.1	NCBI
Plasmodiophora brassicae	Eukaryota;Protists;Other Protists	GCA_003833335.1	NCBI
Balamuthia mandrillaris	Eukaryota;Protists;Other Protists	GCA_001185145.1	NCBI
Perkinsela sp. CCAP 1560/4	Eukaryota;Protists;Kinetoplasts	GCA_001235845.1	NCBI
Urostyla sp. PUJRC_G1	Eukaryota;Protists;Other Protists	GCA_001272955.2	NCBI
Laurentiella sp. PUJRC_G5	Eukaryota;Protists;Other Protists	GCA_001272975.2	NCBI
Paraurostyla sp. PUJRC_G6	Eukaryota;Protists;Other Protists	GCA_001272965.2	NCBI
Tetmemena sp. SeJ-2015	Eukaryota;Protists;Other Protists	GCA_001273295.2	NCBI
Chrysochromulina sp. CCMP291	Eukaryota;Protists;Other Protists	GCA_001275005.1	NCBI
Leptomonas pyrrhocoris	Eukaryota;Protists;Kinetoplasts	GCA_001293395.1	NCBI
Leptomonas seymouri	Eukaryota;Protists;Kinetoplasts	GCA_001299535.1	NCBI
Phytophthora multivora	Eukaryota;Protists;Other Protists	GCA_001314345.1	NCBI
Phytophthora taxon totara	Eukaryota;Protists;Other Protists	GCA_001314375.1	NCBI
Phytophthora pluvialis	Eukaryota;Protists;Other Protists	GCA_001314425.1	NCBI
Phytophthora agathidicida	Eukaryota;Protists;Other Protists	GCA_001314435.1	NCBI
Leishmania peruviana	Eukaryota;Protists;Kinetoplasts	GCA_001403675.1	NCBI
Peronospora tabacina	Eukaryota;Protists;Other Protists	GCA_002099245.1	NCBI
Pseudocohnilembus persalinus	Eukaryota;Protists;Other Protists	GCA_001447515.1	NCBI
Trypanosoma equiperdum	Eukaryota;Protists;Kinetoplasts	GCA_001457755.2	NCBI
Aurantiochytrium sp. T66	Eukaryota;Protists;Other Protists	GCA_001462505.1	NCBI
Bodo saltans	Eukaryota;Protists;Kinetoplasts	GCA_001460835.1	NCBI
Phytophthora nicotianae	Eukaryota;Protists;Other Protists	GCA_003328465.1	NCBI

Plasmopara halstedii	Eukaryota;Protists;Other Protists	GCA_900000015.1	NCBI
Pythium oligandrum	Eukaryota;Protists;Other Protists	GCA_005966545.1	NCBI
Sphaeroforma sirkka	Eukaryota;Protists;Other Protists	GCA_001586965.3	NCBI
Cryptosporidium baileyi	Eukaryota;Protists;Apicomplexans	GCA_001593455.1	NCBI
Pilasporangium apinafurcum	Eukaryota;Protists;Other Protists	GCA_001600495.1	NCBI
Tieghemostelium lacteum	Eukaryota;Protists;Other Protists	GCA_001606155.1	NCBI
Haemoproteus tartakovskyi	Eukaryota;Protists;Apicomplexans	GCA_001625125.1	NCBI
Monocercomonoides sp. PA203	Eukaryota;Protists;Other Protists	GCA_001643675.1	NCBI
Prorocentrum minimum	Eukaryota;Protists;Other Protists	GCA_001652855.1	NCBI
Uroleptopsis citrina	Eukaryota;Protists;Other Protists	GCA_001653735.1	NCBI
Diplonema papillatum	Eukaryota;Protists;Other Protists	GCA_001655075.1	NCBI
Eukaryota sp. EH-2015	Eukaryota;Protists;Other Protists	GCA_001655205.1	NCBI
Plasmodium ovale	Eukaryota;Protists;Apicomplexans	GCA_900090025.2	NCBI
Plasmodium malariae	Eukaryota;Protists;Apicomplexans	GCA_900090045.1	NCBI
Halocafeteria seosinensis	Eukaryota;Protists;Other Protists	GCA_001687465.1	NCBI
Rhizaria sp. SCN 62-66	Eukaryota;Protists;Other Protists	GCA_001724265.1	NCBI
Fonticula-like sp. SCN 57-25	Eukaryota;Protists;Other Protists	GCA_001724245.1	NCBI
Paramoeba pemaquidensis	Eukaryota;Protists;Other Protists	GCA_002151225.1	NCBI
Phytomonas francai	Eukaryota;Protists;Kinetoplasts	GCA_001766655.1	NCBI
Cryptosporidium andersoni	Eukaryota;Protists;Apicomplexans	GCA_001865355.1	NCBI
Phytophthora x alni	Eukaryota;Protists;Other Protists	GCA_000439335.1	NCBI
Cryptosporidium ubiquitum	Eukaryota;Protists;Apicomplexans	GCA_001865345.1	NCBI
Tritrichomonas foetus	Eukaryota;Protists;Other Protists	GCA_001839685.1	NCBI
Moneuplotes crassus	Eukaryota;Protists;Other Protists	GCA_001880385.1	NCBI
Euplotes focardii	Eukaryota;Protists;Other Protists	GCA_001880345.1	NCBI
Plasmodium brasilianum	Eukaryota;Protists;Apicomplexans	GCA_001885115.2	NCBI
Sclerospora graminicola	Eukaryota;Protists;Other Protists	GCA_002933675.1	NCBI
Stramenopiles sp. TOSAG23-2	Eukaryota;Protists;Other Protists	GCA_900128395.1	NCBI
Stramenopiles sp. TOSAG23-6	Eukaryota;Protists;Other Protists	GCA_900128565.1	NCBI
Pythium periplocum	Eukaryota;Protists;Other Protists	GCA_001922765.1	NCBI
Symbiodinium microadriaticum	Eukaryota;Protists;Other Protists	GCA_001939145.1	NCBI
Stentor coeruleus	Eukaryota;Protists;Other Protists	GCA_001970955.1	NCBI
Spongospora subterranea	Eukaryota;Protists;Other Protists	GCA_900404475.1	NCBI
Creolimax fragrantissima	Eukaryota;Protists;Other Protists	GCA_002024145.1	NCBI
Acanthamoeba comandoni	Eukaryota;Protists;Other Protists	GCA_002025285.1	NCBI
Phytophthora cactorum	Eukaryota;Protists;Other Protists	GCA_003287315.1	NCBI
Protostelium mycophagum	Eukaryota;Protists;Other Protists	GCA_002081555.1	NCBI
Trypanosoma theileri	Eukaryota;Protists;Kinetoplasts	GCA_002087225.1	NCBI
Entodinium caudatum	Eukaryota;Protists;Other Protists	GCA_002087855.2	NCBI
Babesia sp. Xinjiang	Eukaryota;Protists;Apicomplexans	GCA_002095265.1	NCBI
Thraustochytrium sp. ATCC 26185	Eukaryota;Protists;Other Protists	GCA_002154235.1	NCBI
Plasmodium gonderi	Eukaryota;Protists;Apicomplexans	GCA_002157705.1	NCBI
Rostrostelium ellipticum	Eukaryota;Protists;Other Protists	GCA_900092235.1	NCBI
Synstelium polycarpum	Eukaryota;Protists;Other Protists	GCA_900092255.1	NCBI
Coremiostelium polycephalum	Eukaryota;Protists;Other Protists	GCA_900092265.1	NCBI
Cavenderia deminutiva	Eukaryota;Protists;Other Protists	GCA_900092275.1	NCBI
Acytostelium leptosomum	Eukaryota;Protists;Other Protists	GCA_900092245.1	NCBI
Phytophthora megakarya	Eukaryota;Protists;Other Protists	GCA_002215365.1	NCBI
Crithidia bombi	Eukaryota;Protists;Kinetoplasts	GCA_900240985.1	NCBI
Peronospora effusa	Eukaryota;Protists;Other Protists	GCA_003843895.1	NCBI
Phytophthora plurivora	Eukaryota;Protists;Other Protists	GCA_002247145.1	NCBI
Lagenidium giganteum	Eukaryota;Protists;Other Protists	GCA_002286825.1	NCBI
Phytophthora colocasiae	Eukaryota;Protists;Other Protists	GCA_002288995.1	NCBI
Eimeria falciformis	Eukaryota;Protists;Apicomplexans	GCA_002271815.1	NCBI
Besnoitia besnoiti	Eukaryota;Protists;Apicomplexans	GCA_002563875.1	NCBI
Cystoisospora suis	Eukaryota;Protists;Apicomplexans	GCA_002600585.1	NCBI
Ichthyophonus hoferi	Eukaryota;Protists;Other Protists	GCA_002751075.1	NCBI
Corallochytrium limacisporum	Eukaryota;Protists;Other Protists	GCA_002811645.1	NCBI
Ichthyosporea sp. XGB-2017a	Eukaryota;Protists;Other Protists	GCA_002811675.1	NCBI
Abeoforma whisleri	Eukaryota;Protists;Other Protists	GCA_002812265.1	NCBI
Pirum gemmata	Eukaryota;Protists;Other Protists	GCA_002812295.1	NCBI
Phytophthora litchii	Eukaryota;Protists;Other Protists	GCA_002812785.1	NCBI
Peronospora belbahrii	Eukaryota;Protists;Other Protists	GCA_002864105.1	NCBI
Chrysochromulina parva	Eukaryota;Protists;Other Protists	GCA_002887195.1	NCBI
Babesia ovata	Eukaryota;Protists;Apicomplexans	GCA_002897235.1	NCBI
Phytophthora palmivora var. palmivora	Eukaryota;Protists;Other Protists	GCA_002911725.1	NCBI
Paratrypanosoma confusum	Eukaryota;Protists;Kinetoplasts	GCA_002921335.1	NCBI
Heterostelium album PN500	Eukaryota;Protists;Other Protists	GCA_000004825.1	NCBI
Crithidia expoeki	Eukaryota;Protists;Kinetoplasts	GCA_900240875.1	NCBI
Paralagenidium karlingii	Eukaryota;Protists;Other Protists	GCA_002980425.1	NCBI
Planoprotostelium fungivorum	Eukaryota;Protists;Other Protists	GCA_003024175.1	NCBI
Aphanomyces stellatus	Eukaryota;Protists;Other Protists	GCA_900243725.1	NCBI
Aphanomyces euteiches	Eukaryota;Protists;Other Protists	GCA_900312765.1	NCBI
Naegleria lovaniensis	Eukaryota;Protists;Other Protists	GCA_003324165.1	NCBI
Globobulimina sp.	Eukaryota;Protists;Other Protists	GCA_003354225.1	NCBI

Hondaea fermentalgiana	Eukaryota;Protists;Other Protists	GCA_002897355.1	NCBI
Kipferlia bialata	Eukaryota;Protists;Other Protists	GCA_003568945.1	NCBI
Goniomonas avonlea	Eukaryota;Protists;Other Protists	GCA_003573635.1	NCBI
Plasmopara obducens	Eukaryota;Protists;Other Protists	GCA_003640625.1	NCBI
Leishmania lainsoni	Eukaryota;Protists;Kinetoplasts	GCA_003664395.1	NCBI
Trypanosomatidae sp. JR-2017a	Eukaryota;Protists;Kinetoplasts	GCA_003671325.1	NCBI
Heterostelium multicystogenum	Eukaryota;Protists;Other Protists	GCA_003667245.1	NCBI
Speleostelium caveatum	Eukaryota;Protists;Other Protists	GCA_003667305.1	NCBI
Plasmopara muralis	Eukaryota;Protists;Other Protists	GCA_003676415.1	NCBI
Nothophytophthora sp. Chile5	Eukaryota;Protists;Other Protists	GCA_001712635.2	NCBI
Polymyxa betae	Eukaryota;Protists;Other Protists	GCA_003693705.1	NCBI
Trypanosoma conorhini	Eukaryota;Protists;Kinetoplasts	GCA_003719485.1	NCBI
Pythium guiyangense	Eukaryota;Protists;Other Protists	GCA_003730235.1	NCBI
Pseudoperonospora humuli	Eukaryota;Protists;Other Protists	GCA_003991265.1	NCBI
Breviolum minutum Mf 1.05b.01	Eukaryota;Protists;Other Protists	GCA_000507305.1	NCBI
Nannochloropsis salina CCMP537	Eukaryota;Protists;Other Protists	GCA_004335465.1	NCBI
Globisporangium ultimum DAOM BR144	Eukaryota;Protists;Other Protists	GCA_000143045.1	NCBI
Cryptosporidium cuniculus	Eukaryota;Protists;Apicomplexans	GCA_004337835.1	NCBI
Cryptosporidium viatorum	Eukaryota;Protists;Apicomplexans	GCA_004337795.1	NCBI
Perkinsus sp. BL_2016	Eukaryota;Protists;Other Protists	GCA_004369235.1	NCBI
Nephromyces sp. ex Molgula occidentalis	Eukaryota;Protists;Apicomplexans	GCA_004523865.1	NCBI
Hydrurus foetidus	Eukaryota;Protists;Other Protists	GCA_900617105.1	NCBI
Aurantiochytrium acetophilum	Eukaryota;Protists;Other Protists	GCA_004332575.1	NCBI
Amoebophrya sp. AT5.2	Eukaryota;Protists;Other Protists	GCA_005223375.1	NCBI
Halteria grandinella	Eukaryota;Protists;Other Protists	GCA_006369765.1	NCBI
Stentor roeselii	Eukaryota;Protists;Other Protists	GCA_006503475.1	NCBI
Diophrys appendiculata	Eukaryota;Protists;Other Protists	GCA_006510565.1	NCBI
Pseudokeronopsis carnea	Eukaryota;Protists;Other Protists	GCA_006510595.1	NCBI
Giardia muris	Eukaryota;Protists;Other Protists	GCA_006247105.1	NCBI
Stramenopiles sp. TOSAG23-3	Eukaryota;Protists;Other Protists	GCA_900128585.1	NCBI
Aurantiochytrium sp. KH105	Eukaryota;Protists;Other Protists	GCA_003116975.1	NCBI
Cryptosporidium sp. 37763	Eukaryota;Protists;Apicomplexans	GCA_004936735.1	NCBI
Schizochytrium sp. TIO01	Eukaryota;Protists;Other Protists	GCA_004764695.1	NCBI
Phytomonas sp. isolate Hart1	Eukaryota;Protists;Kinetoplasts	GCA_000982615.1	NCBI
Leishmania sp. MAR LEM2494	Eukaryota;Protists;Kinetoplasts	GCA_000409445.2	NCBI
Plasmodium sp. DRC-Itaito	Eukaryota;Protists;Apicomplexans	GCA_900240055.1	NCBI
Symbiodinium sp. clade C Y103	Eukaryota;Protists;Other Protists	GCA_003297045.1	NCBI
Blastocystis sp. subtype 2	Eukaryota;Protists;Other Protists	GCA_000963365.1	NCBI
Cryptosporidium hominis	Eukaryota;Protists;Apicomplexans	GCA_000006425.1	NCBI
Blastocystis sp. subtype 3	Eukaryota;Protists;Other Protists	GCA_000963385.1	NCBI
Plasmodium sp. gorilla clade G1	Eukaryota;Protists;Apicomplexans	GCA_900095595.1	NCBI
Stramenopiles sp. TOSAG41-1	Eukaryota;Protists;Other Protists	GCA_900128575.1	NCBI
Plasmodium sp. gorilla clade G3	Eukaryota;Protists;Apicomplexans	GCA_900097035.1	NCBI
Blastocystis sp. subtype 6	Eukaryota;Protists;Other Protists	GCA_000963415.1	NCBI
Blastocystis sp. subtype 8	Eukaryota;Protists;Other Protists	GCA_000963455.1	NCBI
Blastocystis sp. subtype 9	Eukaryota;Protists;Other Protists	GCA_000963465.1	NCBI
Blastocystis sp. ATCC 50177/Nand II	Eukaryota;Protists;Other Protists	GCA_001651215.1	NCBI
Chlamydomonas reinhardtii	Eukaryota;Plants;Green Algae	GCA_000002595.2	NCBI
Ostreococcus tauri	Eukaryota;Plants;Green Algae	GCA_000214015.2	NCBI
Ostreococcus lucimarinus CCE9901	Eukaryota;Plants;Green Algae	GCA_000092065.1	NCBI
Volvox carteri f. nagariensis	Eukaryota;Plants;Green Algae	GCA_000143455.1	NCBI
Micromonas pusilla CCMP1545	Eukaryota;Plants;Green Algae	GCA_000151265.1	NCBI
Chlorella variabilis	Eukaryota;Plants;Green Algae	GCA_000147415.1	NCBI
Chlorella vulgaris	Eukaryota;Plants;Green Algae	GCA_008119945.1	NCBI
Micromonas sp. ASP10-01a	Eukaryota;Plants;Green Algae	GCA_001430725.1	NCBI
Coccomyxa subellipsoidea C-169	Eukaryota;Plants;Green Algae	GCA_000258705.1	NCBI
Botryococcus braunii	Eukaryota;Plants;Green Algae	GCA_002005505.1	NCBI
Prototheca wickerhamii	Eukaryota;Plants;Green Algae	GCA_003255715.1	NCBI
Parachlorella kessleri	Eukaryota;Plants;Green Algae	GCA_001598975.1	NCBI
Dunaliella salina	Eukaryota;Plants;Green Algae	GCA_002284615.1	NCBI
Tetradesmus obliquus	Eukaryota;Plants;Green Algae	GCA_900108755.1	NCBI
Bathycoccus prasinos	Eukaryota;Plants;Green Algae	GCA_002220235.1	NCBI
Helicosporidium sp. ATCC 50920	Eukaryota;Plants;Green Algae	GCA_000690575.1	NCBI
Nannochloris sp. RS	Eukaryota;Plants;Green Algae	GCA_004335565.1	NCBI
Ulva prolifera	Eukaryota;Plants;Green Algae	GCA_004138255.1	NCBI
Auxenochlorella pyrenoidosa	Eukaryota;Plants;Green Algae	GCA_001430745.1	NCBI
Chlamydomonas sp. WS7	Eukaryota;Plants;Green Algae	GCA_004335715.1	NCBI
Gonium pectorale	Eukaryota;Plants;Green Algae	GCA_001584585.1	NCBI
Auxenochlorella protothecoides	Eukaryota;Plants;Green Algae	GCA_000733215.1	NCBI
Chlorella sorokiniana	Eukaryota;Plants;Green Algae	GCA_003130725.1	NCBI
Mychonastes homosphaera	Eukaryota;Plants;Green Algae	GCA_009193075.1	NCBI
Coccomyxa sp. LA000219	Eukaryota;Plants;Green Algae	GCA_000812005.1	NCBI
Trebouxia gelatinosa	Eukaryota;Plants;Green Algae	GCA_000818905.1	NCBI
Picochlorum sp. SENEW3	Eukaryota;Plants;Green Algae	GCA_000876415.1	NCBI
Monoraphidium neglectum	Eukaryota;Plants;Green Algae	GCA_000611645.1	NCBI

Cymbomonas tetramitiformis	Eukaryota;Plants;Green Algae	GCA_001247695.1	NCBI
Micromonas commoda	Eukaryota;Plants;Green Algae	GCA_000090985.2	NCBI
Coelastrella sp. M60	Eukaryota;Plants;Green Algae	GCA_001630525.1	NCBI
Chlamydomonas applanata	Eukaryota;Plants;Green Algae	GCA_001662365.1	NCBI
Chlamydomonas asymmetrica	Eukaryota;Plants;Green Algae	GCA_001662385.1	NCBI
Edaphochlamys debaryana	Eukaryota;Plants;Green Algae	GCA_001662405.1	NCBI
Chlamydomonas sphaeroides	Eukaryota;Plants;Green Algae	GCA_001662425.1	NCBI
Bathycoccus sp. TOSAG39-1	Eukaryota;Plants;Green Algae	GCA_900128745.1	NCBI
Yamagishiella unicocca	Eukaryota;Plants;Green Algae	GCA_003116995.1	NCBI
Trebouxia sp. TZW2008	Eukaryota;Plants;Green Algae	GCA_002118135.1	NCBI
Micractinium conductrix	Eukaryota;Plants;Green Algae	GCA_002245815.2	NCBI
Scenedesmus quadricauda	Eukaryota;Plants;Green Algae	GCA_002317545.1	NCBI
Chlamydomonas eustigma	Eukaryota;Plants;Green Algae	GCA_002335675.1	NCBI
Prototheca stagnorum	Eukaryota;Plants;Green Algae	GCA_002794665.1	NCBI
Monoraphidium sp. 549	Eukaryota;Plants;Green Algae	GCA_002814315.1	NCBI
Prototheca cutis	Eukaryota;Plants;Green Algae	GCA_002897115.1	NCBI
Tetrabaena socialis	Eukaryota;Plants;Green Algae	GCA_002891735.1	NCBI
Chlorella sp. A99	Eukaryota;Plants;Green Algae	GCA_003063905.1	NCBI
Haematococcus lacustris	Eukaryota;Plants;Green Algae	GCA_003970955.1	NCBI
Eudorina sp. 2006-703-Eu-15	Eukaryota;Plants;Green Algae	GCA_003117195.1	NCBI
Raphidocelis subcapitata	Eukaryota;Plants;Green Algae	GCA_003203535.1	NCBI
Trebouxiophyceae sp. KSI-1	Eukaryota;Plants;Green Algae	GCA_003568905.1	NCBI
Ulva mutabilis	Eukaryota;Plants;Green Algae	GCA_900538255.1	NCBI
Picocystis sp. ML	Eukaryota;Plants;Green Algae	GCA_003665715.1	NCBI
Mamiellophyceae sp. 2017MT	Eukaryota;Plants;Green Algae	GCA_004115355.1	NCBI
Haematococcus sp. NG2	Eukaryota;Plants;Green Algae	GCA_004335575.1	NCBI
Chloroidium sp. CF	Eukaryota;Plants;Green Algae	GCA_004335625.1	NCBI
Dunaliella sp. M2	Eukaryota;Plants;Green Algae	GCA_004335885.1	NCBI
Chloromonas sp. AAM2	Eukaryota;Plants;Green Algae	GCA_004335635.1	NCBI
Characiochloris sp. AAM3	Eukaryota;Plants;Green Algae	GCA_004335845.1	NCBI
Scenedesmus sp. ARA3	Eukaryota;Plants;Green Algae	GCA_004335835.1	NCBI
Scenedesmus vacuolatus	Eukaryota;Plants;Green Algae	GCA_004764505.1	NCBI
Chloropicon primus	Eukaryota;Plants;Green Algae	GCA_007859695.1	NCBI
Tetraselmis striata	Eukaryota;Plants;Green Algae	GCA_006384855.1	NCBI
Desmodesmus armatus	Eukaryota;Plants;Green Algae	GCA_007449985.1	NCBI
Chlorophyta sp.	Eukaryota;Plants;Green Algae	GCA_007760615.1	NCBI
Messastrum gracile	Eukaryota;Plants;Green Algae	GCA_008037345.1	NCBI
Prototheca bovis	Eukaryota;Plants;Green Algae	GCA_003612995.1	NCBI
Prototheca ciferrii	Eukaryota;Plants;Green Algae	GCA_003613005.1	NCBI
Scenedesmus sp. ARA	Eukaryota;Plants;Green Algae	GCA_004335915.1	NCBI
Dunaliella sp. WIN1	Eukaryota;Plants;Green Algae	GCA_004335645.1	NCBI
Chloroidium sp. JM	Eukaryota;Plants;Green Algae	GCA_004335615.1	NCBI
Chlorella sp. ArM0029B	Eukaryota;Plants;Green Algae	GCA_002896455.3	NCBI
Trebouxia sp. A1-2	Eukaryota;Plants;Green Algae	GCA_008636185.1	NCBI
Coelastrella sp. UTEX B 3026	Eukaryota;Plants;Green Algae	GCA_002588565.1	NCBI
Picochlorum sp. ’soloecismus’	Eukaryota;Plants;Green Algae	GCA_002818215.1	NCBI
Coccomyxa sp. SUA001	Eukaryota;Plants;Green Algae	GCA_001244535.1	NCBI
Chlamydomonas sp. WS3	Eukaryota;Plants;Green Algae	GCA_004335755.1	NCBI
Nannochloris sp. X1	Eukaryota;Plants;Green Algae	GCA_004335555.1	NCBI
Chlamydomonas sp. 3222	Eukaryota;Plants;Green Algae	GCA_004335795.1	NCBI
Chlorella sp. KRBP	Eukaryota;Plants;Green Algae	GCA_004335735.1	NCBI
Dunaliella sp. YS1	Eukaryota;Plants;Green Algae	GCA_004335685.1	NCBI
Dunaliella sp. RO	Eukaryota;Plants;Green Algae	GCA_004335775.1	NCBI
Chlorella sp. Dachan	Eukaryota;Plants;Green Algae	GCA_006782975.1	NCBI
Chlamydomonas sp. AIC	Eukaryota;Plants;Green Algae	GCA_004335895.1	NCBI
Chlamydomonas sp. 3112	Eukaryota;Plants;Green Algae	GCA_004335865.1	NCBI
Thalassiosira pseudonana CCMP1335	Eukaryota;Protists;Other Protists	Project ID 16452	JGI
Phaeodactylum tricornutum CCAP 10551	Eukaryota;Protists;Other Protists	Project ID 16244	JGI
Fragilariopsis cylindrus	Eukaryota;Protists;Other Protists	Project ID 16035	JGI

Table A.5 Reference taxa included in eukaryotic phylogenomic tree (Figure 4.12). Taken from NCBI and JGI, identifiers as assembly accession or project ID respectively.

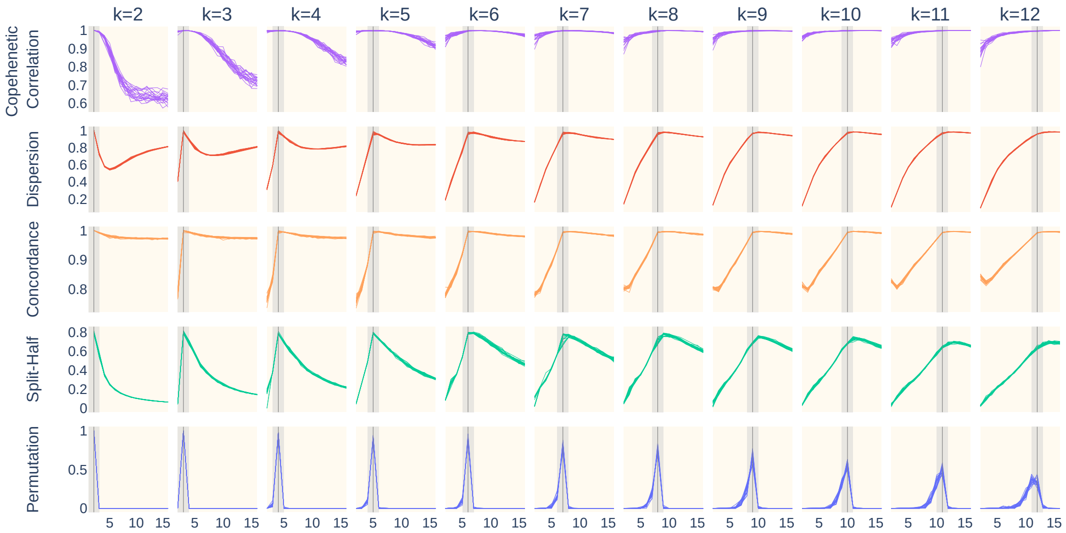
Appendix B

Appendices for Chapter 5

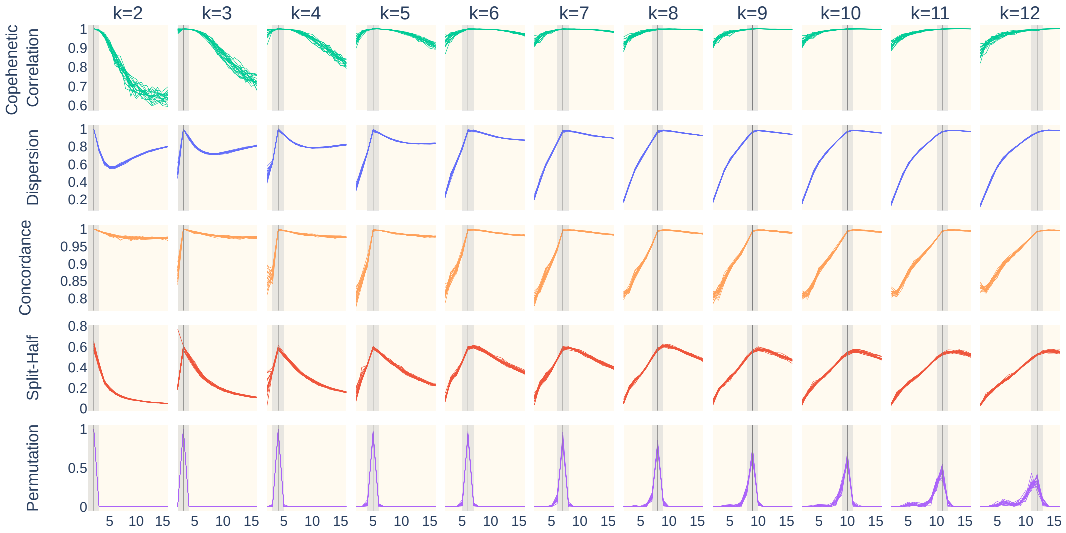
B.1 Model Selection Criteria Plotted for Synthetic Data

This appendix gives rank selection criteria over values of *k* for all synthetic data we generated (Table 5.1, Section 5.3.6). Each column is a different true latent rank, each row a different rank selection method. The vertical grey line indicates the true latent rank, with the grey band showing ∓ 1.

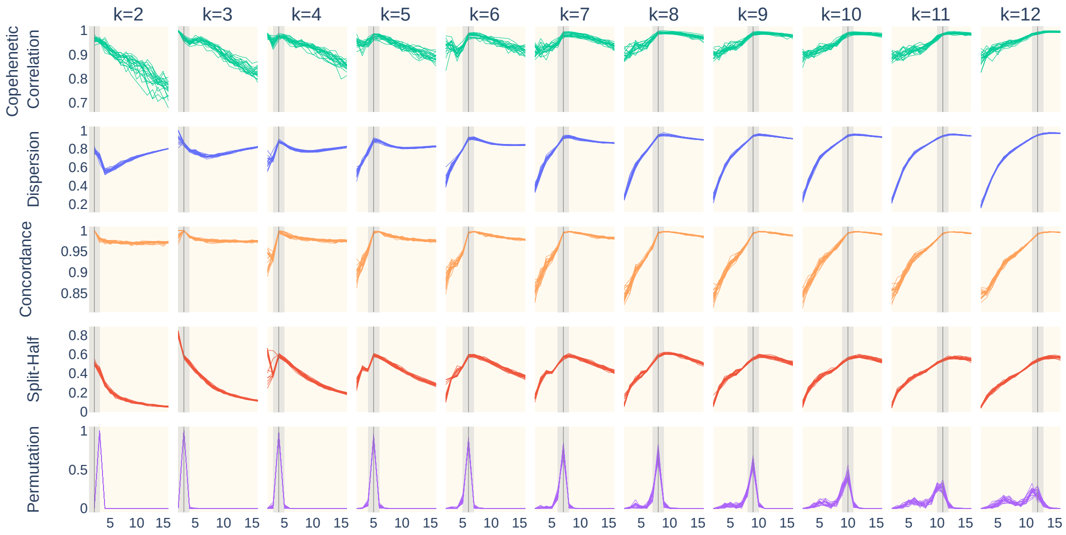
Model selection plots for discrete_lownoise



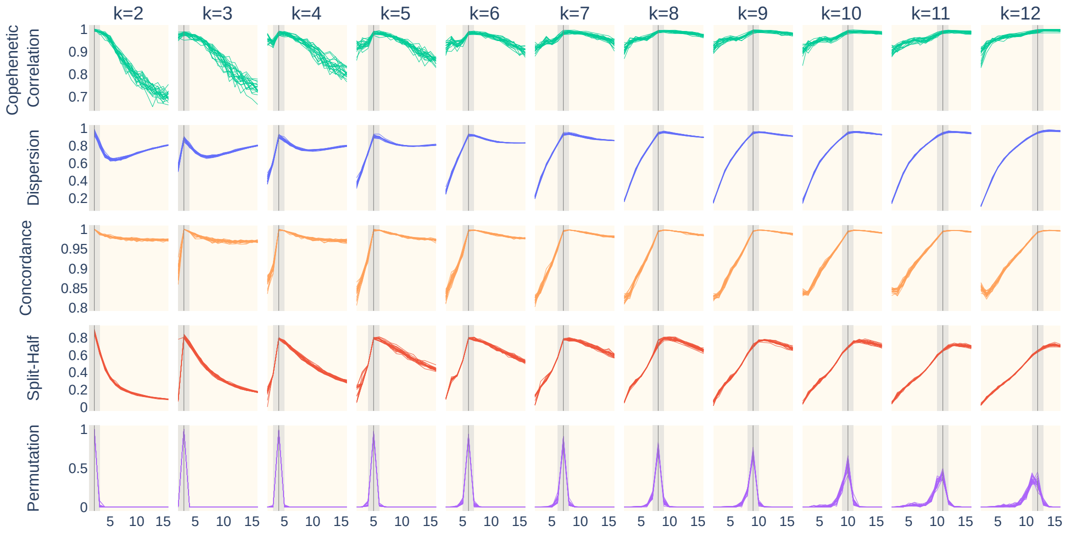
Model selection plots for lo_f_lownoise



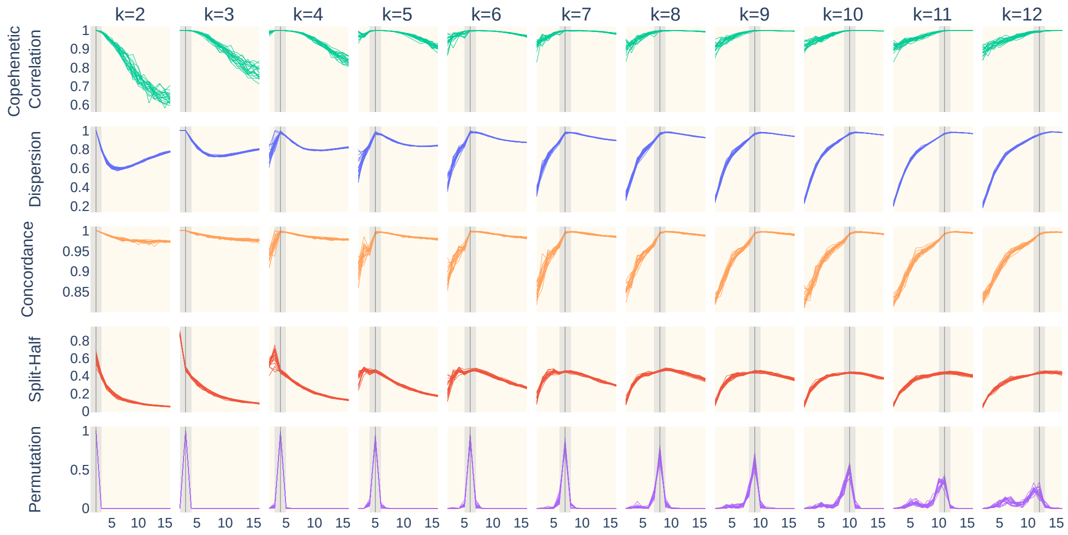
Model selection plots for lo_sf_lownoise



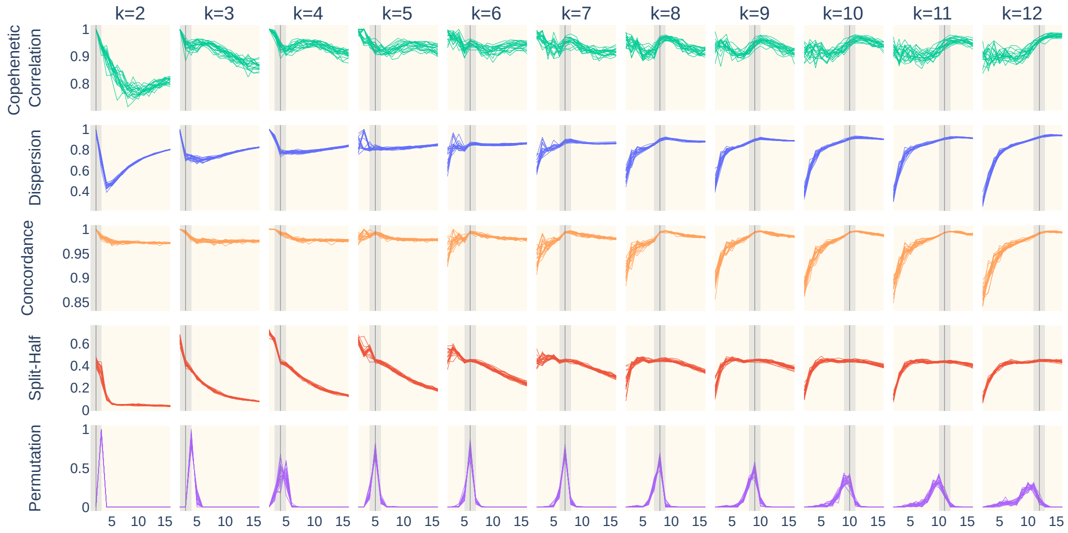
Model selection plots for lo_s_lownoise

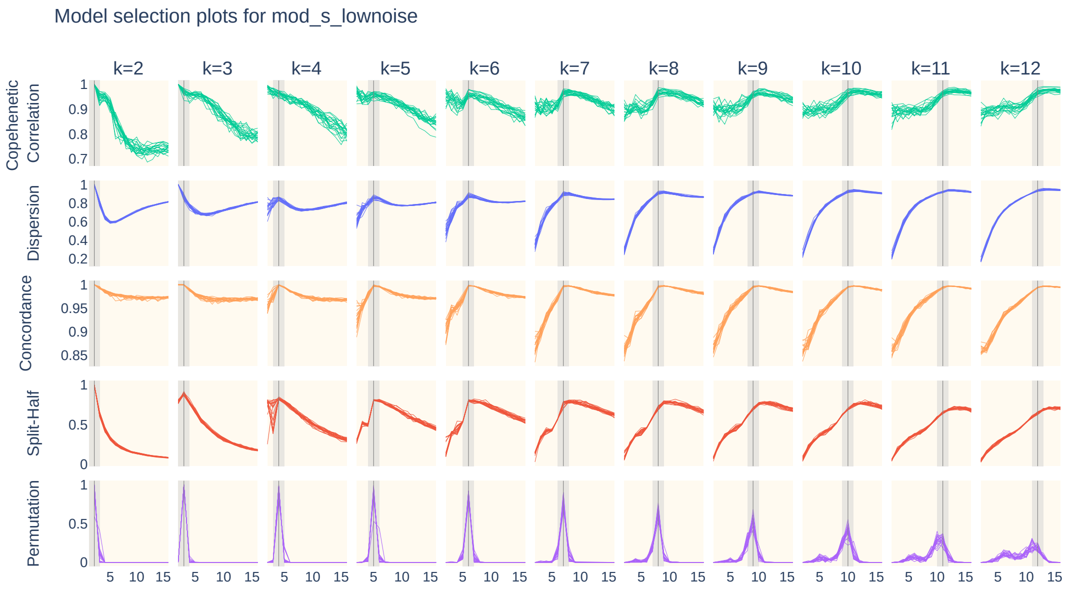


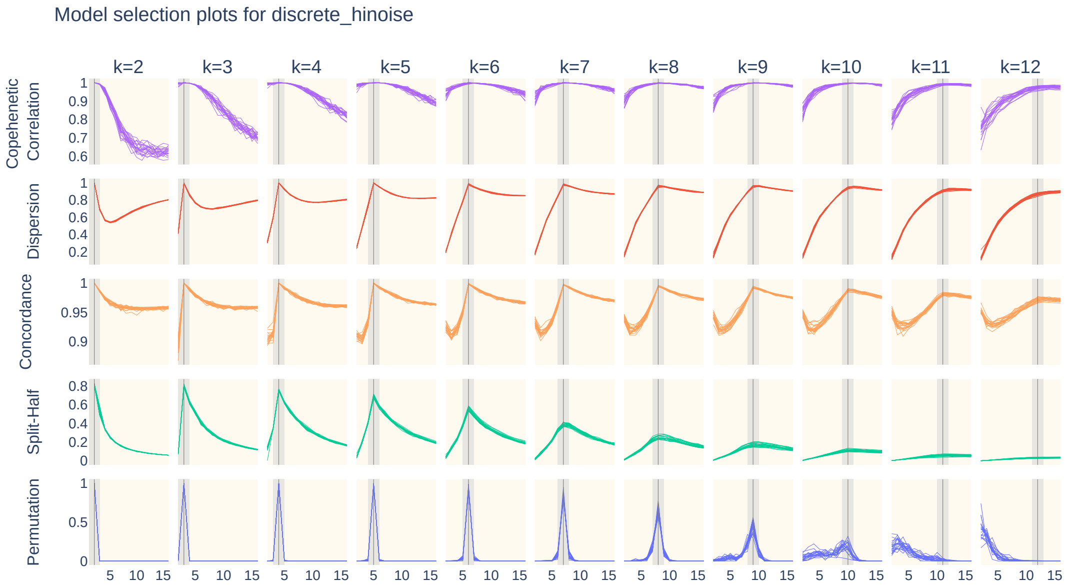
Model selection plots for mod_f_lownoise



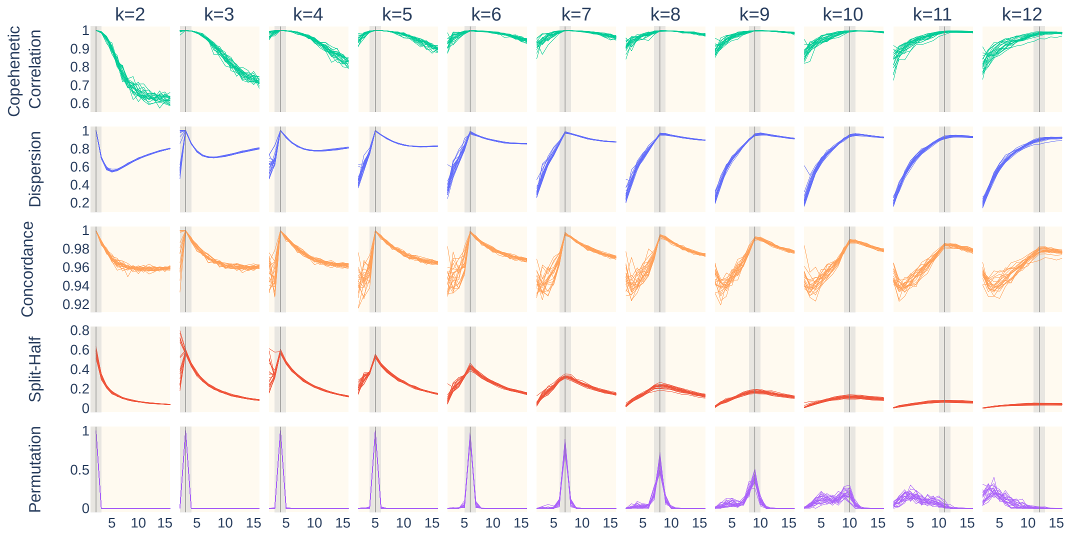
Model selection plots for mod_sf_lownoise



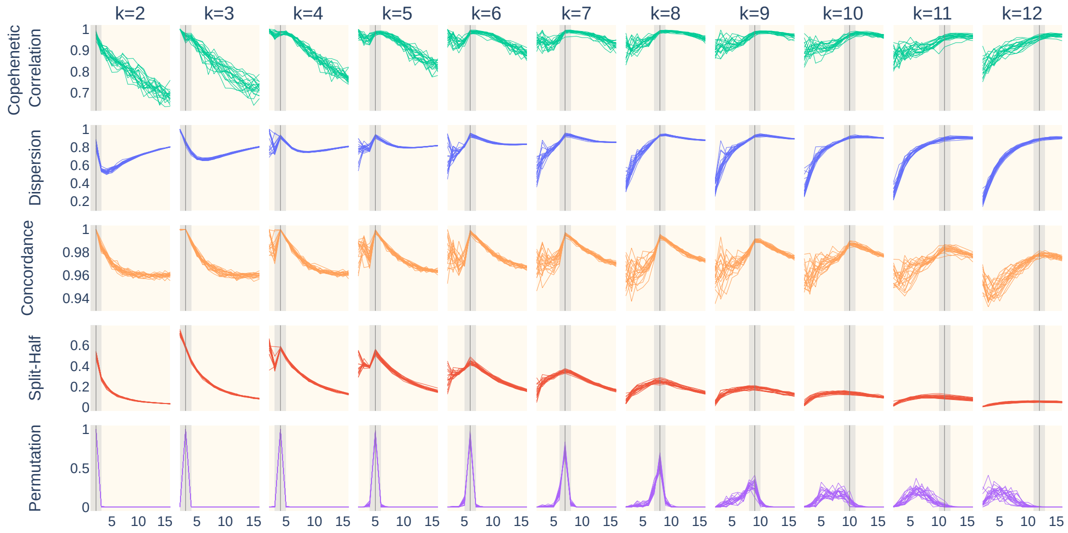

Model selection plots for mod_s_lownoise
k=2
k=3
k=4
k=5
k=6
k=7
k=8
k=9
k=10
k=11
k=12
Copehenetic Correlation
Dispersion
Concordance
Split-Half
Permutation
1
0.9
0.8
0.7
1
0.8
0.6
0.4
0.2
1
0.95
0.9
0.85
1
0.5
0
1
0.5
0
5
10
15

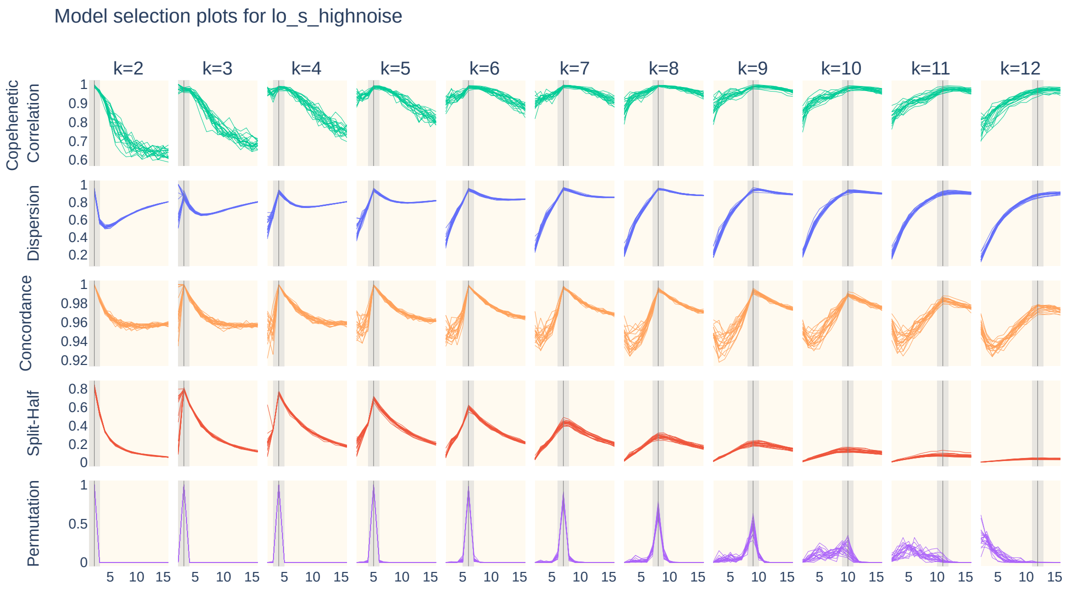

Model selection plots for discrete_hinoise
k=2
k=3
k=4
k=5
k=6
k=7
k=8
k=9
k=10
k=11
k=12
Copehenetic Correlation
Dispersion
Concordance
Split-Half
Permutation
1
0.9
0.8
0.7
0.6
1
0.8
0.6
0.4
0.2
1
0.95
0.9
0.8
0.6
0.4
0.2
0
1
0.5
0
5
10
15

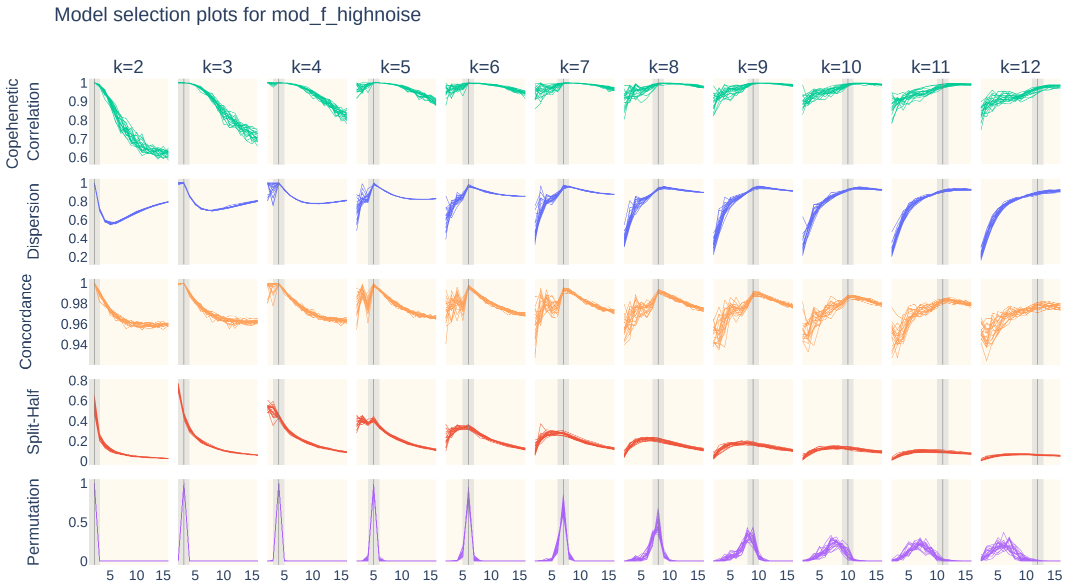
Model selection plots for lo_f_highnoise

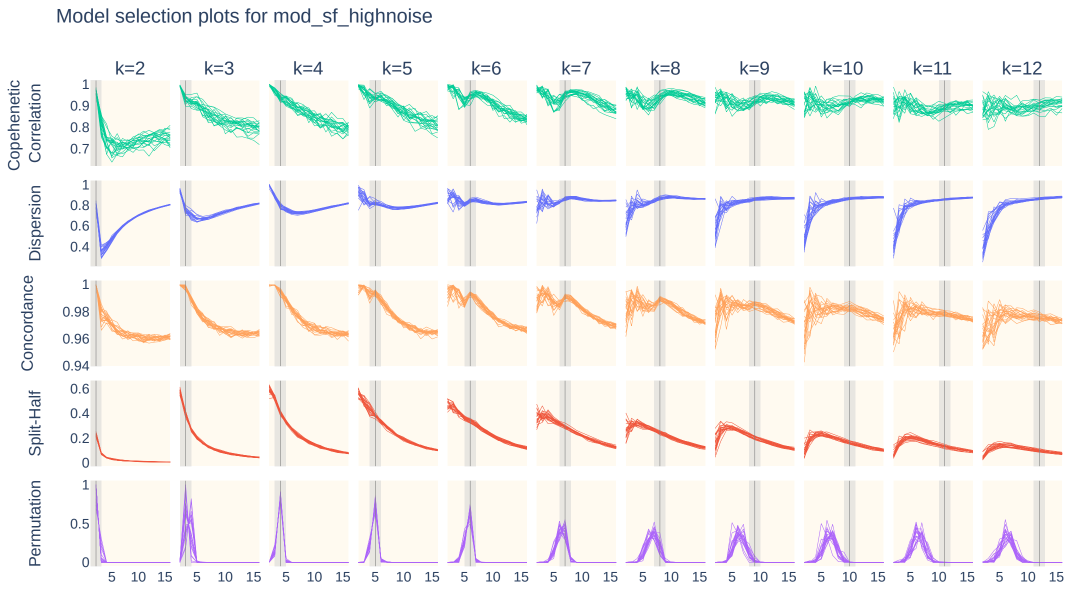


Model selection plots for lo_sf_highnoise

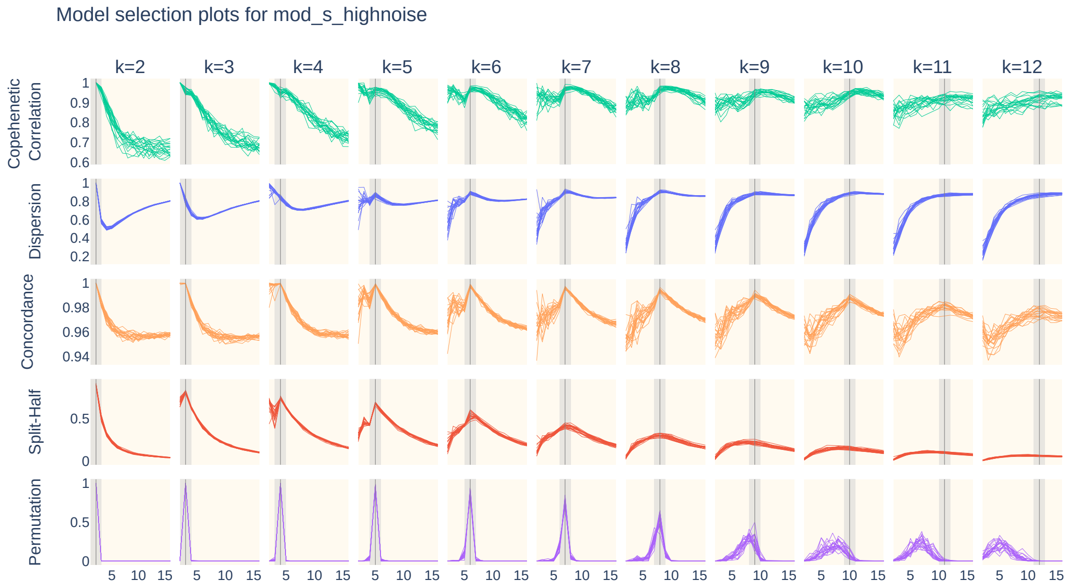



Model selection plots for lo_s_highnoise


Model selection plots for mod_f_highnoise



Model selection plots for mod_sf_highnoise



Model selection plots for mod_s_highnoise

B.2 MOSAiC Derived Simulated Community Composition

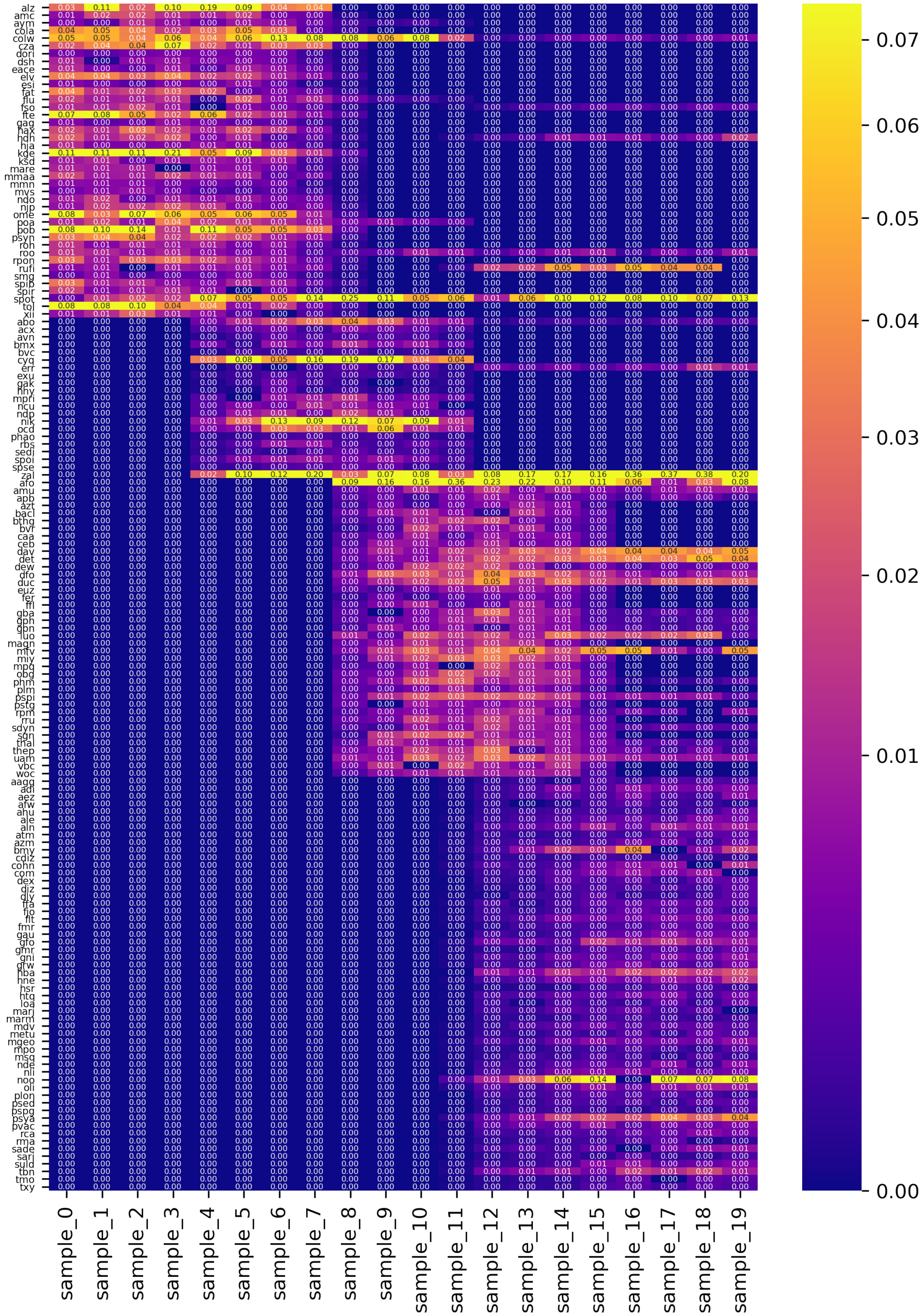


Fig. B.1 Relative abundance of genomes in MOSAiC derived simulated community. Vertical axis is genome, horizontal is sample. Genomes given as KEGG organism code.

B.3 Waiwera River Estuary Gene Details

Gene	Group	Subgroup
rubisco_form_I	Carbon fixation	Rubisco
rubisco_form_II	Carbon fixation	Rubisco
rubisco_form_III	Carbon fixation	Rubisco
aclA	Carbon fixation	reverse TCA cycle
aclB	Carbon fixation	reverse TCA cycle
codhC	Carbon fixation	wood_lungdahl_pathway
codhD	Carbon fixation	wood_lungdahl_pathway
codh_catalytic	Carbon fixation	wood_lungdahl_pathway
amoA_AOA	Nitrogen	Ammonia oxidation
amoA_AOB	Nitrogen	Ammonia oxidation
amoA_AOB_like	Nitrogen	Ammonia oxidation
amoA_comammox	Nitrogen	Complete ammonia oxidization
cphA	Nitrogen	Cyanophycin metabolism
cphB	Nitrogen	Cyanophycin metabolism
nrfA	Nitrogen	DNRA
nrfH	Nitrogen	DNRA
nifA_Mo	Nitrogen	N_fixation
nifB_Mo	Nitrogen	N_fixation
nifH	Nitrogen	N_fixation
napA	Nitrogen	nitrate_reduction
napB	Nitrogen	nitrate_reduction
narG	Nitrogen	nitrate_reduction
narH	Nitrogen	nitrate_reduction
ndma	Nitrogen	nitrate_reduction
norB	Nitrogen	nitric_oxide_reduction
norC	Nitrogen	nitric_oxide_reduction
nxrA	Nitrogen	nitrite_oxidation
nxrB	Nitrogen	nitrite_oxidation
nirB	Nitrogen	nitrite_reduction
nirD	Nitrogen	nitrite_reduction
nirK	Nitrogen	nitrite_reduction
nirS	Nitrogen	nitrite_reduction
nosD	Nitrogen	nitrous_oxide_reduction

nosZ	Nitrogen	nitrous_oxide_reduction
mnhB	Osmoregulation	Na+_H+ antiporter Mnh
mnhE	Osmoregulation	Na+_H+ antiporter Mnh
mnhF	Osmoregulation	Na+_H+ antiporter Mnh
mnhG	Osmoregulation	Na+_H+ antiporter Mnh
NhaA	Osmoregulation	Na+_H+ antiporter Nha
NhaB	Osmoregulation	Na+_H+ antiporter Nha
TrkA	Osmoregulation	Potassium transport_TrkA
kdpA	Osmoregulation	Potassium transport_kdp
kdpB	Osmoregulation	Potassium transport_kdp
kdpC	Osmoregulation	Potassium transport_kdp
kup	Osmoregulation	Potassium transport_kup
nqrF	Osmoregulation	Sodium-translocating_dehydrogenases_nqrF
OpuAC/proX	Osmoregulation	glycine betaine transport
bcct	Osmoregulation	glycine betaine transport
proV	Osmoregulation	glycine betaine transport
aphA	Phosphorus	Acid Phosphatase
phoN	Phosphorus	Acid Phosphatase
phoA	Phosphorus	Alkaline Phosphatase
phoD	Phosphorus	Alkaline Phosphatase
phoB	Phosphorus	PhoR PhoB Phosphate Regulon
phoR	Phosphorus	PhoR PhoB Phosphate Regulon
phoU	Phosphorus	PhoU phosphate regulon
pit	Phosphorus	Phosphate Inorganic Transporter
pstA	Phosphorus	Phosphate Specific Transport System
pstB	Phosphorus	Phosphate Specific Transport System
pstC	Phosphorus	Phosphate Specific Transport System
pstS	Phosphorus	Phosphate Specific Transport System
pufL	Photosynthesis	Anoxygenic photosynthesis
pufM	Photosynthesis	Anoxygenic photosynthesis
psaA_psaB protein	Photosynthesis	Photosystem I
psaF	Photosynthesis	Photosystem I
psaL	Photosynthesis	Photosystem I
PII	Photosynthesis	Photosystem II
psbA	Photosynthesis	Photosystem II
psbI	Photosynthesis	Photosystem II

psbZ	Photosynthesis	Photosystem II
aprA	Sulfur	Sulfate_reduction
cysN	Sulfur	Sulfate_reduction
sat	Sulfur	Sulfate_reduction
sqr	Sulfur	Sulfide_oxidation
asrB	Sulfur	Sulfite_reduction
dsrA	Sulfur	Sulfite_reduction
dsrB	Sulfur	Sulfite_reduction
dsrD	Sulfur	Sulfite_reduction
rdsrA	Sulfur	Sulfur_oxidation
rdsrB	Sulfur	Sulfur_oxidation
soxB	Sulfur	Thiosulfate_oxidation
soxC	Sulfur	Thiosulfate_oxidation
soxY	Sulfur	Thiosulfate_oxidation

Table B.1 Grouping of genes in the Waiwera River Estuary case study (Section 5.5.2) [1].