University of Southampton

Faculty of Natural and Environmental Sciences

School of Ocean and Earth Sciences

“Microbial Community Phylogeny and Function in Relation to Environment in the Norwegian Sea: A High-Throughput Community-Sequencing-Based Assessment”

By

Michael Carter-Gates

Thesis for the degree of Doctor of Philosophy

July 2019

Michael Carter-Gates:

Microbial Community Phylogeny and Function in Relation to Environment in the Norwegian Sea: A High-Throughput Community-Sequencing-Based Assessment

Supervised by:

Dr Declan Schroeder, Professor Thomas S. Bibby and Professor C. Mark Moore

University of Southampton

Abstract

Faculty of Natural and Environmental Sciences

School of Ocean and Earth Science

Doctor of Philosophy

"Microbial Community Phylogeny and Function in Relation to Environment in the Norwegian Sea: A High-Throughput Community-Sequencing-Based Assessment"

By Michael Carter-Gates

Significant changes to seasonal ice cover, stratification, and warming are altering the oceanic boundaries between Polar and Atlantic Water masses. The resultant increased mixing and intrusion of Atlantic Waters into the Arctic region is resulting in novel competition between extant microbial communities that drive biogeochemical cycles and underpin the food-web in these regions. However, it remains unclear how extant microbial communities will respond to these new opportunities and challenges.

This work aims to provide an insight into how the bacterial and microbial eukaryotic communities present across a transect in the Norwegian Sea may be impacted by predicted future environmental change to the Arctic region through the use of Next Generation Sequencing methodologies. It is revealed that the microbial communities of the region are being partitioned into distinct assemblages that correlate with gradients of temperature and salinity.

Analysis of the microbial communities from locations influenced by both Polar and Atlantic waters is used to indicate which components of the microbial communities will be selected for as these waters mix. The results of these analyses suggests the potential for the displacement of bacterial communities found at locations determined to be highly influenced by Polar Water, by bacterial communities from locations found to be primarily influenced by Atlantic Waters. This response appears consistent for all abundance fractions and constituent taxonomic groups within the bacterial community.

Analysis of the eukaryotic community suggests a more complex response whereby abundant eukaryotic cold water associated species could dominate over temperate associated species, and different eukaryotic lineages display contrasting responses.

Metatranscriptomes are generated for the eukaryotic community to determine the functional differences between the regional communities. Partitioning was observed which matched the gradient of Polar Water influence implying the presence of distinct genetic profiles between regional communities.

Each station is observed to feature different profiles of gene expression for genes related to key ecosystem process including primary production, nutrient cycling, biogeochemical cycles, the carbon cycle and metabolic processes. However, despite some differences in the expression of functional profiles, functionality is found to be largely conserved across regional communities, suggesting increase Atlantic Water influence within the sampled region may not result in large perturbations to ecosystem functionality, despite potential changes to community composition.

This study has significant implications for the vulnerability of polar associated community assemblages, which may become displaced under predicted increases of Atlantic mixing and warming within the Arctic region.

Contents

Abstract i

List of Figures viii

List of Tables x

Declaration of Authorship xi

Acknowledgements xiii

Abbreviations xiv

Chapter 1 1

1.1 What is Biodiversity? 1

1.2 The Importance of Measuring Biodiversity 1

1.3 Measuring Biodiversity 4

1.4 Microbial Biodiversity 5

1.5 Next Generation Sequencing Technologies 7

1.6 Sequencing Marine Systems 12

1.7 The Main Players in the Global Ocean 13

1.7.1 Haptophytes 13

1.7.2 Diatoms 15

1.7.3 Dinoflagellates 17

1.7.4 Marine Cyanobacteria 19

1.7.5 SAR11 21

1.8 Limitations of Existing Arctic Microbial Studies 23

1.9 Aims of Thesis 25

Chapter 2 29

2.1 Environmental Analysis 29

2.1.1 Study Location 29

2.1.2 Sea Surface Temperature and Circulation Maps 30

2.1.3 CTD Casts 30

2.2 Methods for Molecular Analysis - DNA 33

2.2.1 Total Community DNA Extractions 33

2.2.2 Probe Assays 33

2.2.3 Agarose Gel Electrophoresis 37

2.2.4 DNA Recovery from Agarose Gel Bands 38

2.2.5 DNA Quantification 39

2.3 Methods for Molecular Analysis - RNA 43

2.3.1 RNA Extraction 43

2.3.2 RNA Quantification 43

2.4 Bioinformatic Analysis of Amplified DNA Barcodes 45

2.4.1 Raw DNA Sequence Data 47

2.4.2 Pre-processing Quality Control of DNA Sequence Data 47

2.4.3 Bacteria OTU Assignment and Taxonomic Annotation 47

2.4.4 Eukaryote OTU Assignment and Taxonomic Annotation 48

2.4.5 Post-processing Quality Control of DNA Sequence Data 48

2.5 DNA Community Analysis and Generation of Diversity Outputs 49

2.5.1 Rarefaction Analysis 49

2.5.2 Generation of α-diversity Metrics 49

2.5.3 Generation of β-diversity Metrics 50

2.5.4 Community Composition Visualisation 50

2.5.5 OTU Distributions and Community Distinctness 50

2.5.6 Correlation with Environmental Factors 51

2.6 Bioinformatic Analysis of RNA Metatranscriptomes 51

2.6.1 Raw RNA Sequence Data 53

2.6.2 Pre-processing Quality Control of RNA Sequence Data 53

2.6.3 Transcriptome Assembly 53

2.6.4 Post-processing Quality Control of RNA Sequence Data 54

2.7 Metatranscriptomic Community Analysis 54

2.7.1 Community Composition 54

2.7.2 β-diversity of Functional Profiles 54

2.7.3 Correlation with Environmental Factors 55

2.7.4 Functional Annotation 55

2.7.5 Gene Expression Profiles 55

Chapter 3 57

3.1 Introduction 57

3.2 Results 61

3.2.1 Regional Grouping 61

3.2.2 Assignment of Station Groups 63

3.3 Discussion 68

3.4 Conclusion 71
Chapter 4 73
4.1 Introduction 73
4.2 Results 77
4.2.1 Sequence Analysis 77
4.2.2 Rarefaction Analysis 77
4.2.3 α-diversity 80
4.2.4 β-diversity 81
4.2.5 Community Composition 85
4.2.6 OTU Distribution and Community Distinctness 90
4.3 Discussion 97
4.3.1 Sequence Analysis and α-diversity 97
4.3.2 Environmental Correlation 98
4.3.3 Potential Responses of Rare Taxa 100
4.3.4 Potential Responses of Abundant Taxa 101
4.3.5 Compositional Differences Between Station Communities 102
4.3.6 Potential Winners and Losers 104
4.4 Conclusion 106
Chapter 5 109
5.1 Introduction 109
5.2 Results 113
5.2.1 Sequence Analysis 113
5.2.2 Rarefaction Analysis 113
5.2.3 α-diversity 116
5.2.4 β-diversity 117
5.2.5 Community Composition 122
5.2.6 OTU Distribution and Community Distinctness 127
5.3 Discussion 133
5.3.1 Sequence Analysis and α-diversity 133
5.3.2 Environmental Correlation 134
5.3.3 Potential Responses of Rare Taxa 135
5.3.4 Potential Responses of Abundant Taxa 136
5.3.5 Compositional Differences Between Station Communities 138
5.3.6 Potential Winners and Losers 139

5.4 Conclusion 141

Chapter 6 145

6.1 Introduction 145

6.2 Results 147

6.2.1 Sequence Analysis 147

6.2.2 Community Composition 152

6.2.3 β-diversity 154

6.2.4 Gene Expression Profiles 158

6.2.5 Gene Expression Profiles: Genes Expressed at HI Stations 160

6.2.6 Gene Expression Profiles: Genes Expressed at LI stations 161

6.2.7 Gene Expression Profiles: Genes Expressed at the MI station 162

6.2.8 Gene Expression Profiles: Most Abundant Genes 162

6.2.9 KEGG Analysis 164

6.2.10 Functional Expression Profiles 166

6.2.11 Expression of Photosynthetic Genes 169

6.3 Discussion 171

6.3.1 Sequence Analysis and Assembly 171

6.3.2 Community Composition 172

6.3.3 Regional Genetic Profiles 174

6.3.4 Predicted Changes to Functional Profiles 176

6.3.5 Defining Photosynthetic Strategy 180

6.4 Conclusion 182

Chapter 7 185

7.1 A First Insight into Predicted Biological Impacts of Arctic Atlantification 185

7.1.1 Main Findings 185

7.1.2 Potential Underlying Mechanisms of Community Partitioning 189

7.1.3 Potential Variability of Susceptibility to Atlantic Water Influence 190

7.1.4 Partitioning of the Rare Community 191

7.1.5 Predicted Future Changes to Physical Drivers of Partitioning 192

7.1.6 Additional Potential Functional Impacts of Alterations to Community Composition 193

7.1.7 Potential Changes to the Functionality of Additional Individual Taxonomic Groups 194

7.2 Limitations 195

7.3 Future Directions and Perspective 197

Appendix 201

Appendix 1. The Correlation of all Top 200 Most Abundant Bacterial OTUs with Each Environmental Factor 202

Appendix 2 The Relative Abundance of Selected Bacterial OTUs Between Stations. 215

Appendix 3. The Correlation of all Top 200 Most Abundant Eukaryotic OTUs with Each Environmental Factor 216

Appendix 4. The Identity of all Annotated Genes. 224

Appendix 5. Gene Expression Profiles of all Photosynthetic Genes. 229

Appendix 6. 16S Pre-processing Script 232

Appendix 7. 18S Pre-processing Script 234

Appendix 8. RNA Pre-processing Script 236

Appendix 9. Local SILVA Database Creation Script 239

Appendix 10. Swarm Rscript 241

Appendix 11. 18S OTU Clustering and Taxonomic Assignment Script 242

Appendix 12. 16S OTU Grouping and Taxonomic Assignment Script 244

Bibliography 246

List of Figures

Figure 1.1. Predicted future environmental conditions based upon RCP scenario projections. 2
Figure 1.2. The central dogma of biology. 8
Figure 1.3. Sanger sequencing gel electrophoretogram. 9
Figure 1.4. The method of 454 pyrosequencing. 11
Figure 1.5. The relative contribution of select taxonomic groups to total marine chlorophyll-a biomass. 14
Figure 1.6. Global distribution and diversity of marine Diatoms. 16
Figure 1.7. Dinoflagellate genomic richness as calculated from the Tara Oceans dataset. .. 19
Figure 1.8. The global distribution and abundance of *Prochlorococcus* and *Synechococcus*.21
Figure 1.9. Biogeography of SAR11 phylotypes. 22
Figure 2.1. Track of the RRS James Clarke Ross research Vessel during the JR271 cruise (01/06/12-02/07/12). 29
Figure 2.2. The pipeline followed for the processing of raw DNA sequence data through to data visualisation. 46
Figure 2.3. The pipeline followed for the processing of raw RNA sequence data through to data visualisation. 52
Figure 3.1. Surface temperature trends during 1979–2005 which show the temperature amplification in the Arctic region. 58
Figure 3.2. The decline of Arctic sea ice between 1979-2011. 59
Figure 3.3. Sea surface water temperature maps prior to and after the sampled period. ... 62
Figure 3.4. Vertical profiles of the environmental conditions at each station measured during CTD casts. 65
Figure 3.5. Temperature/salinity plot of water conditions at the sampled depth during the sampled period. 66
Figure 3.6. Sea surface water temperature maps at the time of sampling with the locations of the stations sampled. 67
Figure 3.7. The calculated relative contribution of Arctic water to each station. 68
Figure 3.8. Schematic view of the transition from Atlantic to Arctic associated waters. 69
Figure 4.1. Rarefaction analysis of bacterial sequence depth. 79
Figure 4.2. Bacterial β-diversity between stations. 82
Figure 4.3. Principle coordinate analysis of the Bray-Curtis dissimilarity matrix of bacterial OTUs between stations. 83
Figure 4.4. Regional partitioning of the representative major bacterial taxonomic groups. 84
Figure 4.5. Differences in bacterial community composition between stations. 88
Figure 4.6. Analysis of bacterial community distinctness. 91
Figure 4.7. The relative abundance of bacterial OTUs between stations. 93
Figure 4.8. Heat maps of bacterial OTUs deemed to display a station abundance bias. 96
Figure 5.1. Rarefaction analysis of eukaryotic sequence depth. 115
Figure 5.2. Eukaryotic β-diversity between stations. 118
Figure 5.3. Principle coordinate analysis of the Bray-Curtis dissimilarity matrix of eukaryotic OTUs between stations. 119

Figure 5.4. Regional partitioning of the representative major eukaryotic taxonomic groups. 121
Figure 5.5. Differences in eukaryotic community composition between stations. 125
Figure 5.6. Analysis of eukaryotic community distinctness. 128
Figure 5.7. The relative abundance of eukaryotic OTUs between stations. 130
Figure 5.8. Heat maps of eukaryotic OTUs showing strong station abundance bias. 132
Figure 6.1. Contig length distribution curve. 149
Figure 6.2. Sequence coverage of contigs. 150
Figure 6.3. The relative abundance of representative taxonomic groups of recovered genes. 153
Figure 6.4. The logFC of gene expression profiles between each pair of stations. 155
Figure 6.5. Analysis of genetic distinctness between stations. 156
Figure 6.6. Principle coordinate analysis of the Bray-Curtis dissimilarity matrix of genes between stations. 157
Figure 6.7. The relative expression of annotated genes by station. 159
Figure 6.8. The expression of the top 20 most abundant annotated genes. 163
Figure 6.9. KEGG functional pathway annotations for eukaryotic genes. 165
Figure 6.10. The relative expression of KEGG pathways within each station. 166
Figure 6.11. The relative expression of KEGG pathways at each station for each individual taxonomic group. 168
Figure 6.12. The relative expression of PSI, PSII and *Lhcf* photosynthetic genes. 170
Figure 6.13. Simplified schematic overview of photoacclimation strategies for marine phytoplankton. 181
Figure 7.1. The predicted bacterial community partitioning under increased Atlantic Water influence. 187
Figure 7.2. The predicted eukaryote community partitioning under increased Atlantic Water influence. 188

List of Tables

Table 2.1. The CTD casts from which environmental samples were taken. 32
Table 2.2. The primer combinations used for 18S V9 and 16S V5 probe assays. 34
Table 2.3. Primer sequences used for the probe assay. 35
Table 2.4. PCR cycle after which qPCR reactions were stopped. 37
Table 2.5. Quantities of ADB used during the Zymoclean Gel DNA recovery protocol. 39
Table 2.6. The quantity of DNA recovered for the 18S probe assay as determined by the dsDNA 12000 Series II assay. 41
Table 2.7. The quantity of DNA recovered for the 16S probe assay as determined by the dsDNA 12000 Series II assay. 42
Table 2.8. The highest quality DNA replicates chosen for sequencing. 42
Table 2.9. The quantity of RNA recovered from each sample as determined by the Eukaryote Total RNA Nano II assay. 45
Table 3.1. The environmental conditions at the sampled depth for which water was collected for sequence analysis. 64
Table 4.1. Sequence data metrics for the bacterial dataset. 78
Table 4.2. α-diversity of the bacterial communities recovered from each station. 80
Table 4.3. Pearson correlation of environmental factors to bacterial community variance. 83
Table 4.4. The number of bacterial OTUs showing correlations with each environmental factor. 94
Table 4.5. The number of bacterial OTUs deemed to display an abundance bias to each station. 95
Table 5.1. Sequence data metrics for the eukaryotic dataset. 114
Table 5.2. α-diversity of the eukaryotic communities recovered from each station. 116
Table 5.3. Pearson correlation of environmental factors to eukaryotic community variance. 119
Table 5.4. The number of eukaryote OTUs showing correlations with each environmental factor. 131
Table 5.5. The number of eukaryotic OTUs deemed to display an abundance bias to a certain station. 131
Table 6.1. Sequence data metrics for the metatranscriptomic dataset. 148
Table 6.2. The domain of origin of each identified contig. 149
Table 6.3. The number of RNA sequences recovered from each domain of life. 151
Table 6.4. The number of eukaryotic contigs by gene origin. 151
Table 6.5. Pearson correlation of environmental factors to metatranscriptomic community variance. 157

UNIVERSITY OF Southampton

Research Thesis: Declaration of Authorship

Print name:	Michael Carter-Gates

Title of thesis:	‘Microbial community phylogeny and function in relation to environment in the Norwegian Sea: A high-throughput community-sequencing-based assessment’,

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Either none of this work has been published before submission, or parts of this work have been published as: [please list references below]:

I would like to acknowledge the following contributions to this thesis:

Cecilia Balestreri - for the original collection and preservation of community RNA samples, and the collection, extraction and preservation of community DNA samples aboard the James Clarke Ross as part of the UK Ocean Acidification research program.

Sally Thorpe (British Antarctic Survey, Cambridge, UK) - for the generation of the SST maps used in Chapter 3.

Finlo Cottier (The Scottish Association for Marine Science, Argyll, UK) for providing a description of ocean currents in the sampled region which enabled the addition of these currents to Figure 3.6.

Ali Baylay (National Oceanographic Centre, Southampton, UK) - for the library preparation and Illumina sequencing of DNA barcode and metatranscriptomic datasets.

The crew of the RRS James Clarke Ross for the collection and provision of the environmental metadata used in Chapter 3. All metadata is available through the Natural Environmental Research Council Bristish Oceanographic Data Centre (BODC) stored under file name BODC_JR271_29082014.

Signature:	m Carter - Gates	Date:	28/07/2019

Acknowledgements

I would like to thank my current supervisors, Declan Schroeder, Thomas S. Bibby and C. Mark Moore for their help and advice throughout the duration of this project. I am grateful for the feedback and comments provided which have helped guide and structure the development of this thesis into what has become the largest body of work I have ever produced.

I would like to extend my thanks to Alison Baylay, Finlo Cottier, Sally Thorpe, Claire Jasper, Angela Ward and the crew of the RRS James Clarke Ross for all of their advice and contributions to the project. Additionally, I thank Cecilia Balestreri, Flavia Flaviani, Nathan Chrismas, Willy Wilson, Michael Cunliffe and Rowena Stern for their kind words and the time they have taken to listen to my ideas and provide constructive inputs.

I will be forever grateful for the help and support provided by my friends; Harriet Dale, Graham Epstein, Jack Dickenson, Dean Mckeown, Kim Bird, Friedrich Kleiner, and all those hosted at NOC. Thank you for making these 4 years of my life such an enjoyable experience.

Most of all I would like to thank Katie Vitiello, my mother Liz Carter, my Father Ken Gates and Pat Kelly for their love, support and belief in me. Without your support I don’t think I would have been able to finish.

Abbreviations

A ….. Adenine

ACE ….. Abundance based coverage estimator

ADB ….. Agarose dissolving buffer

ASCII ….. American Standard Code for Information Interchange

ASV ….. Amplicon sequence variant

ATP ….. Adenosine triphosphate

AVISO ….. Archiving, Validation and Interpretation of Satellite Oceanographic data

ß-ME ….. ß-mercaptoethanol

BLAST ….. Basic local alignment search tool

bp ….. Base pairs

C ….. Cytosine

C1 ….. One-carbon

cDNA ….. Complimentary deoxyribonucleic acid

CMIP ….. Coupled Model Intercomparison Project

CTD ….. Conductivity temperature depth

DCM ….. Deep chlorophyll maximum

DMS ….. Dimethyl sulphide

DNA ….. Deoxyribonucleic acid

DO_2 ….. Dissolved oxygen

DOM ….. Dissolved organic matter

dsDNA ….. Double stranded deoxyribonucleic acid

E. coli ….. *Escherichia coli*

EDTA ….. Ethylenediaminetetraacetic acid

G ….. Guanine

GISSTEMP ….. Goddard Institute for Space Studies Surface Temperature Analysis

GO ….. Gene ontology

GOS ….. Global Ocean Sampling

GUI ….. Graphic user interface

HCL ….. Hydrogen chloride

HI ….. High influence

HL ….. High light

ICoMM ….. International Census of Marine Microbes

KAAS ….. KEGG automatic annotation server

KEGG ….. Kyoto Encyclopaedia of Genes and Genomes

KO ….. Kyoto Encyclopaedia of Genes and Genomes Orthology

LI ….. Low influence

LL ….. Low light

LogFC ….. Log fold-change

MATOU ….. Marine Atlas of Tara Oceans Unigenes

MDM ….. Microbial dark matter

MDS ….. Multidimensional scaling

MI ….. Moderate influence

MMETSP….. Marine Microbial Eukaryote Transcriptome Sequencing Project

mRNA ….. Messenger ribonucleic acid

MYI ….. Multiyear ice

NADH ….. Nicotinamide adenine dinucleotide (reduced form)

NCBI ….. National Centre for Biotechnology Information

NGS ….. Next generation sequencing

NOC ….. National Oceanographic Centre

nt ….. Nucleotide

NTC ….. No template control

OSTIA ….. Operational Sea surface Temperature and Ice Analysis

OTU ….. Operational taxonomic unit

PBS ….. Phosphate buffered saline

PC1 ….. Principle coordinate 1

PC2 ….. Principle coordinate 2

PCOA ….. Principle coordinate analysis

PCR ….. Polymerase chain reaction

PEAR ….. Paired-end read merger

PI ….. Perennial ice

pm ….. Parts per million

Poly-A ….. Polyadenylated

qPCR ….. Quantitative polymerase chain reaction

PSI ….. Photosystem I

PSII ….. Photosystem II

PSU ….. Photosystem unit

RCPs ….. Representative concentration pathways

RNA ….. Ribonucleic acid

rRNA ….. Ribosomal ribonucleic acid

S.E ….. Standard error

Spp. ….. Species (plural)

SST ….. Sea surface temperature

SSU rRNA ….. Small subunit ribosomal ribonucleic acid

T ….. Thymine

TAE ….. Tris-acetate-EDTA

TMM ….. Trimmed mean of M values

tRNA ….. Transfer ribonucleic acid

U ….. Uracil

UV ….. Ultraviolet

SubsurVPAR ….. Subsurface vertical photosynthetically available radiation

Chapter 1

Introduction

1.1 What is Biodiversity?

The term ‘Biodiversity’ is used extensively in scientific literature. Yet a clearly defined, unequivocal and unanimously accepted definition does not exist[1]. Broadly speaking ‘Biodiversity’ is generally accepted to refer to the measure of the variation at the genetic, species and ecosystem level. This variation extends from the diversity of nucleic acid sequences both within and between species, to the differences in molecular and ecological states brought about by the actions and interactions of organisms within an ecosystem. These can be divided into sub-components of biodiversity which are often more commonly referred to as phylogenetic, structural and functional diversity[2,3]. Biodiversity can therefore be thought of as the total quantity of information needed to fully detail and recreate a defined ecological system. The need to define the spatial scales of biodiversity, which can have a significant effect on the interpretation of data, has resulted in three additional terms. Alpha, gamma and beta diversity are used to describe confirmed local diversity, regional diversity, and the variability of diversity between locals, respectively[4].

1.2 The Importance of Measuring Biodiversity

The assessment of current levels of biodiversity and its rate of change in the face of global climate change is becoming an ever increasing focus in current literature[5]. These changes are well documented and include rising global temperatures resultant from the increase of atmospheric carbon dioxide (CO_2)[6]. Under the Representative Concentration Pathways (RCPs), which are predicted greenhouse gas concentrations used to model future climate scenarios (Figure 1.1A), CO_2 may reach over 1000 pm by 2100 (Figure 1.1B) with significant impacts to global temperatures[7]. Much of this CO_2 is absorbed by the oceans which in turn reduces ocean pH levels, with a significant fall predicted by 2100[7] (Figure 1.1C). Additionally, human activity is fixing nitrogen at levels that exceed all natural sources. In 1860 30% (6 Tg year-1) of all fixed nitrogen entering the oceans was anthropogenically derived, by 2000 that

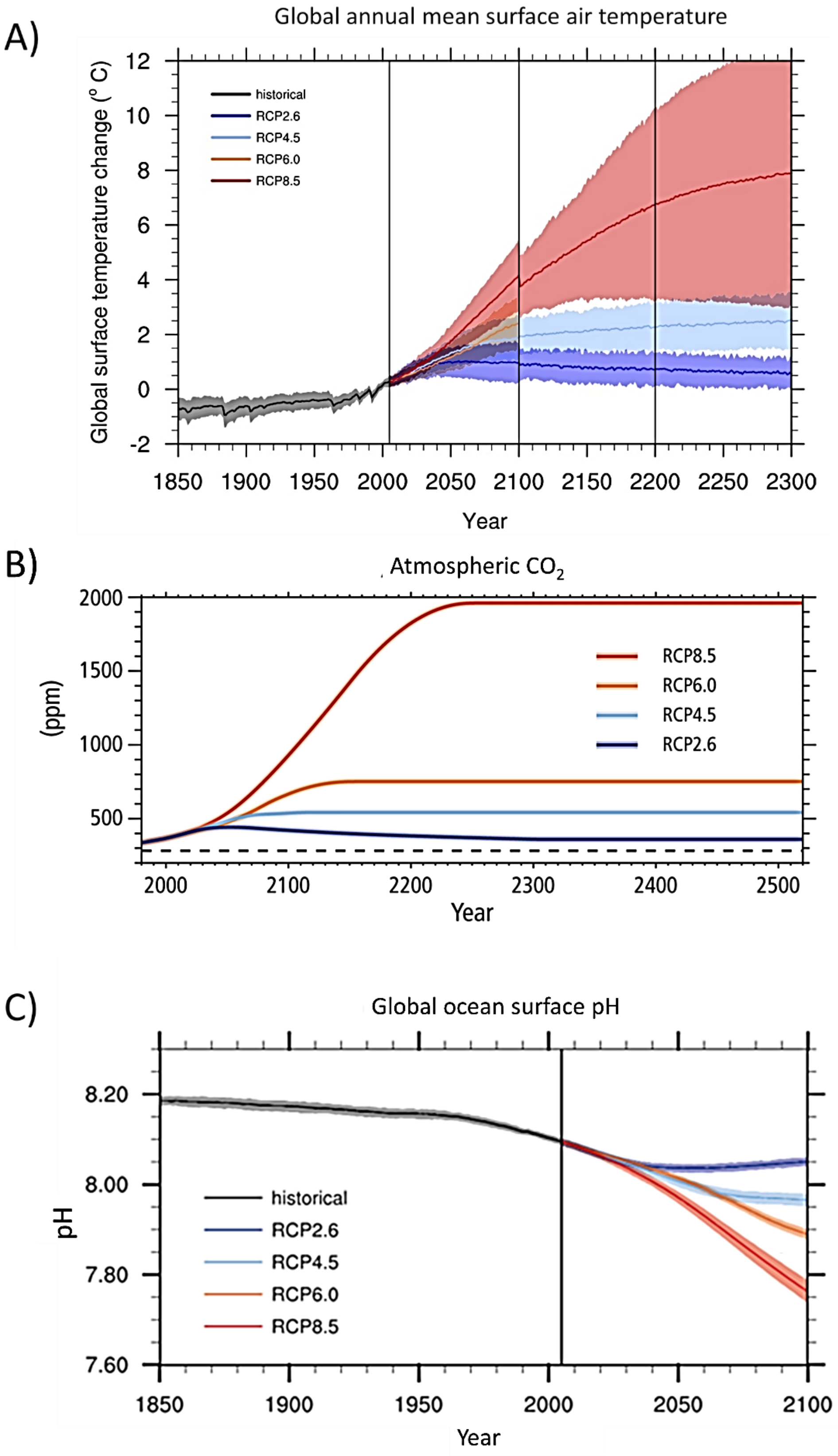


Figure 1.1. Predicted future environmental conditions based upon RCP scenario projections.

A) Global annual mean surface air temperature and RCP scenario projections, shading indicates 5-95% confidence range for each. B) Predicted global ocean surface pH and measure of uncertainty (shading) for the two end member scenarios of RCP2.6 (blue) and RCP8.5 (red). C) Projected CO_2 concentrations by 2300 for under various predicted scenarios. Figure adapted from [7].

proportion had increased to around 80% and the quantity increased to 54 Tg N year^{-1}. The increased quantity of fixed nitrogen in marine systems is altering marine ecosystems where primary production is typically limited by fixed nitrogen availability, leading to increased atmospheric carbon uptake, eutrophication and oxygen minimum zones[8]. These largely anthropogenically driven global changes are expected to have substantial effects on organisms and ecosystems with calls that their severity may warrant the division of a new geological time period popularly referred to as the "Anthropocene"[9]. Part of the evidence being put forward for this notion is that of current extinction rates which are occurring at a previously unprecedented rate[10], and with over a million species thought to be at risk of extinction[11]. As a result there is increasing pressure to conserve biological diversity, particularly in highly impacted regions. This, in part, has been fuelled over recent years by increasing public awareness which in turn has led to increased legislative pressure upon policy makers to address these concerns through better action and management strategies[12].

Such awareness and concern over biodiversity loss is clearly evident for terrestrial systems and large 'charismatic' macrofauna, which are often the focus of public funding campaigns, but there is far lower public concern for the vast majority of small microbial species that are of great importance to the global system. Indeed, marine phytoplankton, which are photoautotrophic primary producers, are responsible for an estimated 50% of the total global primary production, and fix an estimated 100 million tonnes of carbon per day[13]. The lower concern about the loss of microbial species diversity is suggested to be likely due to the lack of direct experience many members of the public have with microbial species, and a lack of awareness of the associated ecosystem services they provide[2]. Furthermore, the majority of the value attributed to these systems is present as non-tradable indirect use which is often outside of the public consciousness. The value of an indirect use is resultant from the functioning of a biological system or entity which conveys a useful benefit to human kind[2]. For example, many phytoplankton are not directly farmed for human consumption, but constitute a vital food source for many commercially important fish species, as well as for conservation target organisms which attract ecotourism, and carry out metabolic processes vital for the transformation of compounds in the ocean, including the evolution of oxygen[14]. Thus, indirect use value forms the largest total economic value in marine biodiversity and many species contribute to multiple types of value[15]. Such indirect values are irreplaceable due to being extremely challenging for humankind to replicate.

1.3 Measuring Biodiversity

An understanding of how to effectively measure biodiversity is required before suitable baseline measurements and effective conservation strategies can be developed. However, there is currently substantial debate as to how this should be achieved. A recent review of 136 papers by Bartkowski *et al*[16] highlights the current challenges to directly measuring biodiversity. One of these is that the term "Biodiversity" has been used to encompass everything from intra-species genetic variation to the spread of biomes across the planet[17]. Two of the most common such applications of the term are to describe species richness or habitat type[5]. However, these measures present limitations; habitat type reduces the complexity of a habitat, the species within it, their interactions, and other factors into a single overly simplified aspect, providing limited information. Species richness fails to meet commonly accepted definitions of biodiversity which state that measures of biodiversity are independent of species identity. In an attempt to combat these limitations studies have begun to promote a move towards a multiple trait orientated approach to measuring biodiversity, known as functional diversity[12,18].

Functional diversity is 'the value and range of those species and organismal traits that influence ecosystem functioning'[19]. It is used to describe the variance of traits that influence ecosystem properties and how a species responds to different environmental conditions[20]. It provides an informative measure of the current state of an ecosystem as it allows the mechanistic coupling of ecosystem functions to diversity[21], enabling greater accuracy in determining ecosystem responses to stressors. To describe functional diversity species are divided into different groups according to the similarity of their functional traits (often essential genes). While this form of grouping is necessary to allow analysis, it too presents a number of limitations[17], often related to methodologies[12,22–24], or due to the link between biodiversity and ecosystem functioning often being environmentally context dependant[25].

Despite methodological limitations there are a number of key patterns that have emerged from functional diversity studies, one of which is that species typically fall into one of two broad categories. These categories represent a compromise between the efficiency with which a species is able to exploit a particular environmental state and the range of states over which it is able to do so[26]. They are either "specialists" or "generalists". Specialists are characterised by exploiting a limited niche space, fulfilling narrow functional roles and showing a high level of success in a small number of habitats that are usually stable over a long period of time, or feature particular environmental conditions, such as those featuring extremes of temperature. Conversely, generalists contribute to broader functional roles and show lower peak levels of growth, but that remain more constant over a wider range of environmental conditions[27]. Generalists are therefore much more likely to become successfully established in novel habitats, or able to capitalise on environmental perturbations which negatively impact more specialised species, as is evidenced by a wealth of literature focusing on invasive species [28–30].

Increasing numbers of invasive species and a global decline of specialist species is a growing concern and is suggested to be resulting in a move towards a global functional homogenisation[26]. Many species are needed to maintain ecosystem functioning and a particular function is often highly impacted by one dominant specialist[22]. Such processes are tightly coupled with complex resource partitioning between specialist species, therefore the loss of specialist diversity opens up new niches for exploitation by generalists and reduces the efficiency by which these resources are used[31].

1.4 Microbial Biodiversity

While ‘generalist’ and ‘specialist’ groupings of species and a fear for global functional homogenisation are easy to attribute to macrofauna, microorganisms present more of a challenge. While there is no universal definition of a microorganism, they are typically taken to include bacteria, viruses, and single celled eukaryotes outside a scale directly observable with the human eye[32].

It remains open to debate whether microorganisms display similar diversity distribution patterns as macroorganisms in response to environmental factors due to the prevalence of two contrasting hypotheses. The first hypothesis is that of a global cosmopolitan distribution which follows the logic that the small size of microorganisms allows them to be distributed globally, driven primarily by random dispersal from a high number of natural forces. Microbial species often exist in populations of extremely high abundances which is suggested to maintain high levels of distribution making it unlikely that they will become locally extinct, resulting in ubiquitous global diversity[33,34]. Moreover, the ability of many microbial taxa to persist as dormant and resting stages is suggested to further act to ensure they maintain a persistent local presence[35]. Evidence is more prominent for microbial eukaryotes, partly due to greater ease of recovery, and it is thought that eukaryotic organisms within the size range of <2-10 mm display a cosmopolitan distribution[34], that is they are dispersed globally (whether metabolically dormant or active). The hypothesis is supported by examples of the same species of flagellated protozoan being isolated in Denmark fjords and Pacific hydrothermal vents[36] and similar patterns observed for a number of species including ciliates[33] and flagellates[36].

However, opposition to this hypothesis argue similar geographically restricted distribution patterns as seen for macroorganisms may be found for microorganisms, and will likely manifest over very large or very small spatial scales[37,38]. There is growing evidence that microbial communities display distinct distribution patterns as a result of local environmental factors[39,40], and studies such as the Darwin Project are attempting to visualise and model these patterns[41]. Indeed, environmental factors have been shown to influence community composition at global scales[42], and temperature and salinity are both commonly reported as key predictors of microbial community structure in marine systems[42–45]. It is suggested that the persistence of such conflicting viewpoints may be due to methodological differences between studies, for example, the temporal or spatial scales addressed may be too small to fully recover a local species pool. Smaller scale studies also tend to reveal strong environmental filtering of communities, whereas at larger scales geographical factors are found to exert the greatest influence[46]. Thus, microbial community structure appears to be determined by a combination of spatial and local environmental factors over different spatial scales[47].

The argument for biogeography of microbes becomes more complicated in the case of rare taxa. Rare taxa are often present at very low abundances, typically <0.01% of the community[48–50], and thus require a significant sampling effort to recover, therefore it is difficult to distinguish false negative recovery rates resultant from insufficient sampling effort from true absence. Furthermore, strong seasonal species turnover[51], sequencing errors from erroneous base calls[52], and the presence of dead and dormant cells[53] all affect the accuracy of rare species detection complicating their analysis. Despite the low abundance of constituent taxa, it is the rare biosphere which often holds the highest level of diversity and is most likely to contain novel metabolic pathways and genes[54]. Rare species have been suggested to be important players in maintaining key ecological functions in the face of environmental perturbations[35,55]. A number of studies hypothesise that the rare biosphere is likely to display a strong cosmopolitan distribution as a low abundance is thought to reduce loss by decreasing the chance of selective predation and the encounter rate with viruses[54]. However, there is a growing body of evidence that indicates this may not be the case. It has been suggested that communities from the rare biosphere are stable and moulded by similar processes to abundant individuals, implying that the rare biosphere is not just a random collection of taxa at the edge of their physiological tolerances[56]. Additionally, unique phylogenetic lineages have been discovered from what appear to be permanently rare taxa, and some rare taxa have been shown to display distinct biogeographies[56]. Confirmation that these taxa are metabolically active implies that rarity may constitute an evolutionary trait in some groups[49,56].

Resolving the opposing viewpoints of microbial distribution has been further driven by the advancement of the field into the application of molecular techniques. Such techniques include the use of genetic probes and Next Generation Sequencing (NGS) upon environmental samples. NGS has resolved morphologically identical organisms into a number of separate taxa, thus revealing a high number of cryptic species which previously wasn't possible with conventional microscopy based analyses[47]. This discovery has increased the already existing difficulty in applying a species concept to microbes, especially bacteria, and suggests the morphological differences historically used to assign taxonomy may be irrelevant. At current microbial species are designated into groups based upon arbitrary percentage similarities of DNA sequences[57]. Therefore, it may be simply that the correct level of taxonomic resolution required to define a species is yet to be determined for microbes[58]. Despite a few continuing challenges, the use of molecular techniques has led to a general acceptance amongst the majority of researchers that high levels of random dispersal by natural forces act to widely distribute microbial taxa, but local environmental conditions can select for particular community assemblages.

1.5 Next Generation Sequencing Technologies

The wider application of NGS techniques to microbial diversity surveys is partly the result of advances of sequencing technologies, which make use of the properties of DNA in their function. DNA is the material which encodes the information necessary for all life and is converted into proteins in a series of processes known as the central dogma of biology (Figure 1.2). These processes are common to all organisms and involve: 1) DNA-DNA during cell division; 2) DNA-RNA to transcribe DNA into the coding sequence for proteins; and 3) RNA-protein where the RNA is translated into a protein or protein sub-unit. The resulting proteins are then responsible for carrying

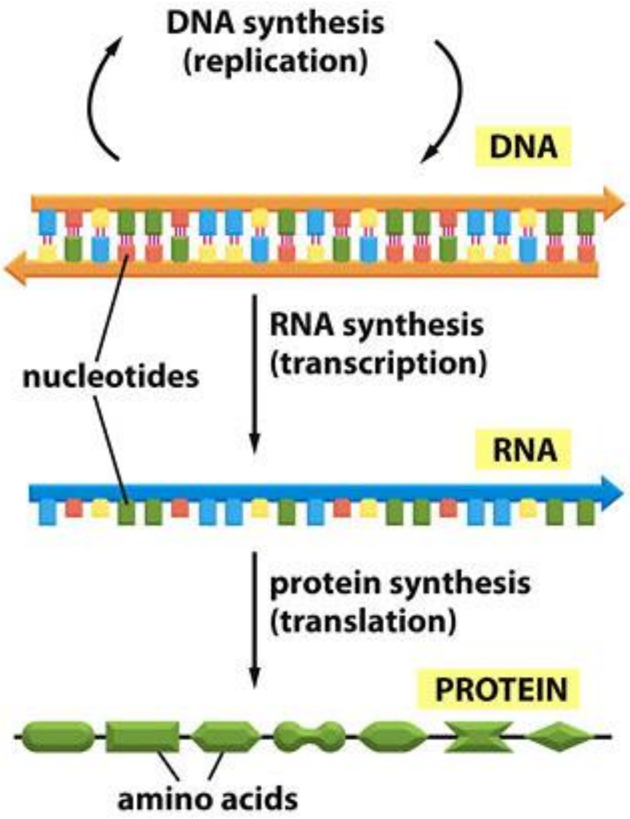


Figure 1.2. The central dogma of biology.

Shown is the process by which DNA is transcribed and translated into protein products. Replication of DNA occurs by helicase separating the double stranded DNA molecule and addition of new complimentary nucleotides before DNA ligase re-joins the phosphate backbones of the nucleotides to create a continuous chain. In transcription the DNA strand is again separated to form the coding strand and its compliment. During translation only the compliment strand is copied into messenger mRNA, and the base Uracil (U) replaces Thymine (T) in mRNA. Post-transcriptional modifications are then made to the mRNA (messenger RNA) before it moves to the cytoplasm to be read within ribosomes by tRNAs (transfer RNA), and translated into a polypeptide chain of amino acids which eventually undergoes conformational changes to form proteins. Figure from[60].

out all functional aspects within the organism, be it structural, hormonal or enzymatic[59]. DNA itself is made up of two complimentary strands of the nucleotide bases adenine (A), thymine (T), cytosine (C) and guanine (G), connected along a phosphate backbone. Each of these bases has a specific affinity for one other base, with A-T and C-G pairs being capable of joining together by hydrogen bonds. This specific affinity allows complimentary strands of DNA to bind together[61].

The exact sequence of nucleotides encoded by an organisms DNA can be determined through the process of DNA sequencing, and used to identify genes, their expression, and relatedness to other organisms. The first generation of sequencing, Sanger sequencing, was developed in 1977[62]. Sanger sequencing involves the fragmentation of DNA for cloning by inclusion in colonies of *Escherichia coli,* which are then extracted and amplified by multiple rounds of polymerase chain reaction

(PCR) to increase the quantity of DNA available. PCR comprises three stages, the first heat separates the DNA into two strands, the second allows the binding of free nucleotides to complimentary bases on each strand, and the final allows the annealing of two complimentary strands into a complete double stranded DNA molecule. During the final round of the PCR process in Sanger sequencing DNA bases labelled with a fluorescent tag (different for each of the four bases) are added which terminate the extension of the DNA strands. The resultant DNA stands are size separated by gel electrophoresis, with the smaller lighter fragments travelling further along the gel. The gel can then be exposed to UV light to fluoresce the tags and determine the sequence of nucleotides (Figure 1.3).

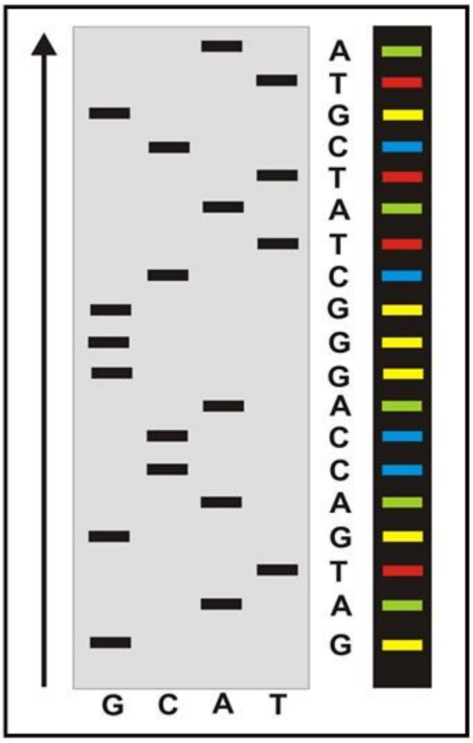


Figure 1.3. Sanger sequencing gel electrophoretogram.

In Sanger sequencing the sequence of nucleotides is determined based on the position of bands along a gradient. The bands separate out by weight, with the larger, heavier bands travelling a shorter distance along the gel, the bands are read sequentially and the fluorescent tag detected at each point recorded. This generates a list of bases from which the target sequence can be resolved. Figure from[63].

Sanger was the dominant form of sequencing for 4 decades, and is still in use today when the generation of long sequences (~800 bp) and relatively high accuracy of base calls is required. However, due to the need to clone DNA fragments within *E. coli* it is a time consuming method with throughput limited to ~800bp per run[64]. A new generation of sequencing technologies have been developed to address scientific questions requiring greater scale and throughput. For the purpose of this discussion only the two currently dominant platforms, Illumina and 454 pyrosequencing will be discussed. 454 pyrosequencing sequencing generates sequences of length 400-500bp by a 'sequence-by-synthesis' approach (Figure 1.4)[64], which works by ligating single stranded DNA fragments to adapters on specially prepared beads which are loaded into individual wells on a picotiter plate. A wash of free single nucleotide bases is added which bind to their complimentary bases on the DNA strands attached to the beads, releasing pyrophosphate in the process. ATP sulfurylase converts pyrophosphate to ATP, which in turn catalyses luciferin to oxyluciferin and emits light. This is repeated with multiple washes for each nucleotide base. The emissions are captured by camera to determine which base the emission corresponds to, and the size of the emission peak is proportional to the number of bases included. If no complimentary base is present the free nucleotides are degraded by apyrase. As the process continues the complimentary strand grows and the sequence is determined by the emission signal peaks generated. By utilising beads and picotiter wells during this method millions of strands of DNA can be sequenced in parallel, resulting in a possible throughput of ~500 Mb per run[64].

The Illumina sequencing platform quickly gained in popularity and now dominates the sequencing industry[65]. It is also based on a 'sequence-by-synthesis' approach and relies on dye incorporated into the nucleotide sequence to terminate the sequence elongation. A key difference from the 454 platform is that these terminators are reversible in the Illumina protocol, allowing polymerisation of the nucleotide chain to continue after fluorescence detection. The reads generated by this technique are typically much shorter, at around a few hundred base pairs, but the cost involved is much lower and very high levels of throughput, in the region of 35 Gb, are obtainable[64], far exceeding the other two techniques previously described (for a complete comparison of most commonly available sequencing platforms see[66]). The shorter reads produced by Illumina technology also have the benefit of reducing the likelihood of errors being incorporated into the sequenced DNA from a 1% probability associated with 454 platforms, to 0.1% for Ilumina[66].

Potential erroneous reads can be a significant problem associated with NGS technologies, and were a main area of criticism during their early development, however many of these can now be removed by applying filtering algorithms to the output sequences as part of an *in silico* post-processing pipeline. These are applied

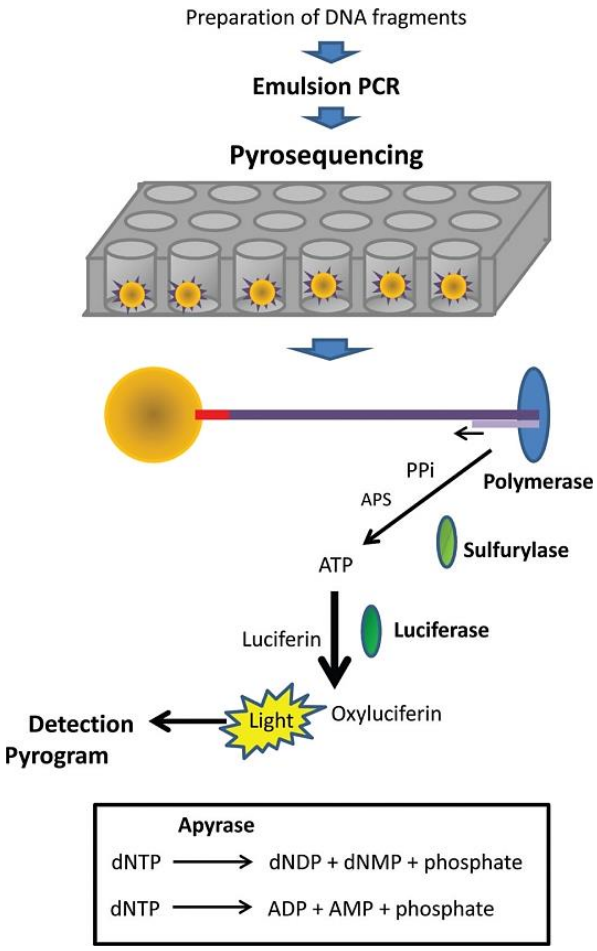


Figure 1.4. The method of 454 pyrosequencing.

Shown is each stage of the 454 pyrosequencing process, as well as the chemical reactions which occur during the binding of a free nucleotide base and result in the emission of light for the identification of which free nucleotide base bound to the DNA strand. Also shown are the enzymes involved at each stage of the process. Figure from[64].

to the plain text files generated by the NGS platforms, which are typically of FASTA format for 454 pyrosequencing or FASTQ format for Illumina platforms. Each of these file types contain a line of ASCII encoded quality scores, known as phred scores, for each base ranging from 0 (lowest) to 40 (highest) which represent the chances each base was called in error. These are calculated by the equation:

$$p = 10^{-Q/10}$$

Equation 1.1. Equation to calculate contig coverage.

Where p = probability of erroneous base call, and Q= phred quality score

Therefore, a phred score of 30 represents a 0.01% chance the base called was erroneous[67]. The phred scores can therefore be used to apply quality filtering to the sequences by using any one (or combination) of software tools available before being passed to later further downstream analysis.

1.6 Sequencing Marine Systems

The advancement of NGS technologies has enabled the deployment of large scale sampling surveys to help resolve global marine microbial diversity. The International Census of Marine Microbes (ICoMM) established in 2004 aimed to explore the changing distribution, diversity, and abundances of marine microbial species covering all three domains of life, and forms part of the Census of Marine Life program targeted at recording all marine life. Data collection was carried out over a 10 year span, including utilising NGS techniques, and featured projects specifically targeted at addressing the under sampling of Arctic microbial communities revealed in past studies through sampling of both surface and deep waters[57]. The data collected as part of the ICoMM has been reported on extensively and provided early insights into marine microbial community dynamics. The findings of these studies provided some of the first measures of the unexpected levels of diversity present within marine microbial communities[68], resolved regionally distinct community compositions[57], confirmed the rarity of marine cyanobacteria in Arctic marine environments[69], and revealed hydrographic controls on bacterial biogeographic patterns[70].

The Sorcerer II Global Ocean Sampling expedition (GOS) was launched in 2004 and constituted a two year sampling program of surface waters from the North Atlantic to the South Pacific over 32,000 miles that covered much of the world's major oceans (excluding the Arctic). The GOS aimed to explore the bacterial diversity of marine environments, to discover new genes of ecological importance, and better understand how these relate to ecosystem functioning by using whole-genome shotgun sequencing techniques. The 6.3 Gb GOS metagenomics datasets were the largest to ever be published within the public domain, composed of 7.7 million sequences, and revealed biogeographic patterns, an extensive array of some 6 million novel genes[71], and nearly doubled the number of known protein families[72] significantly expanding the understanding of marine microbial diversity.

Surveys such as GOS highlighted the extent of marine microbial diversity yet to be discovered, creating new enthusiasm for further surveys. One of the most prominent recent sampling efforts is that of the Tara Oceans expedition. Much like surveys before it the Tara expedition aimed to further the understanding of microbial diversity and the functional potential within all domains of life. Over 35,000 samples were collected from more than 210 ocean stations covering all major global oceans between 2009-2013. The Tara Oceans expedition primarily targeted the sun-lit upper layer of the ocean and generated 7.2Tb of raw sequence data, exceeding all previous sampling efforts. This data has been used to extensively expand current knowledge of the oceans, resolving environmental factors such as temperature as the key predictor of community structure in the open ocean, and again uncovered extensive novelty for over 80% of sequences generated during the expedition[42]. Further findings resultant from the Tara Oceans dataset include previously unknown diversity within heterotrophic eukaryotes[73], phytoplankton community level responses to iron[74], ubiquitous distributions of rare taxa[75] and previously overlooked contributions of Diatoms to the global carbon pump[76].

1.7 The Main Players in the Global Ocean

Assessing the diversity and functionality within marine systems is critical to understanding the global ecosystem functions provided by bacterial and eukaryotic microorganisms. These functions include catalysing biogeochemical cycles[77], carbon fixation by primary production[78], nutrient cycling[79], climate regulation through dimethyl sulphide (DMS) production[80] and organic matter export to depth[81]. However, different taxonomic groups feature different functionalities, and their relative contributions to certain functions varies. Some of the most abundant functionally important taxa in the global ocean includes Haptophytes, Diatoms, Dinoflagellates, Marine Cyanobacteria and the bacterial clade SAR11.

1.7.1 Haptophytes

Haptophytes are a group of marine phytoplankton composed of over 300 species, many of which exist in the pico- and nano- size fractions of marine planktonic communities. They are considered to be largely globally distributed throughout global marine systems and some freshwater systems[82]. Haptophytes constitute some of the most globally successful marine primary producers, estimated to comprise 30-50% of all global marine phytoplankton (Figure 1.5)[83], and two bloom forming taxa *Phaeocystis spp.* and *Emiliania huxleyi* are amongst the most globally abundant individuals. *Phaeocystis* is found to be globally distributed, its success is suggested to be related to featuring both colonial and flagellated cell morphologies which are able to reduce mortality to below that of competing taxa[84]. It typically appears in early bloom phases due to highly efficient growth under low light conditions, which also allow it to be competitive in light limited polar regions[84]. Different *Phaeocystis* species appear to feature different biogeography, with species such as *Phaeocystis pouchetii*

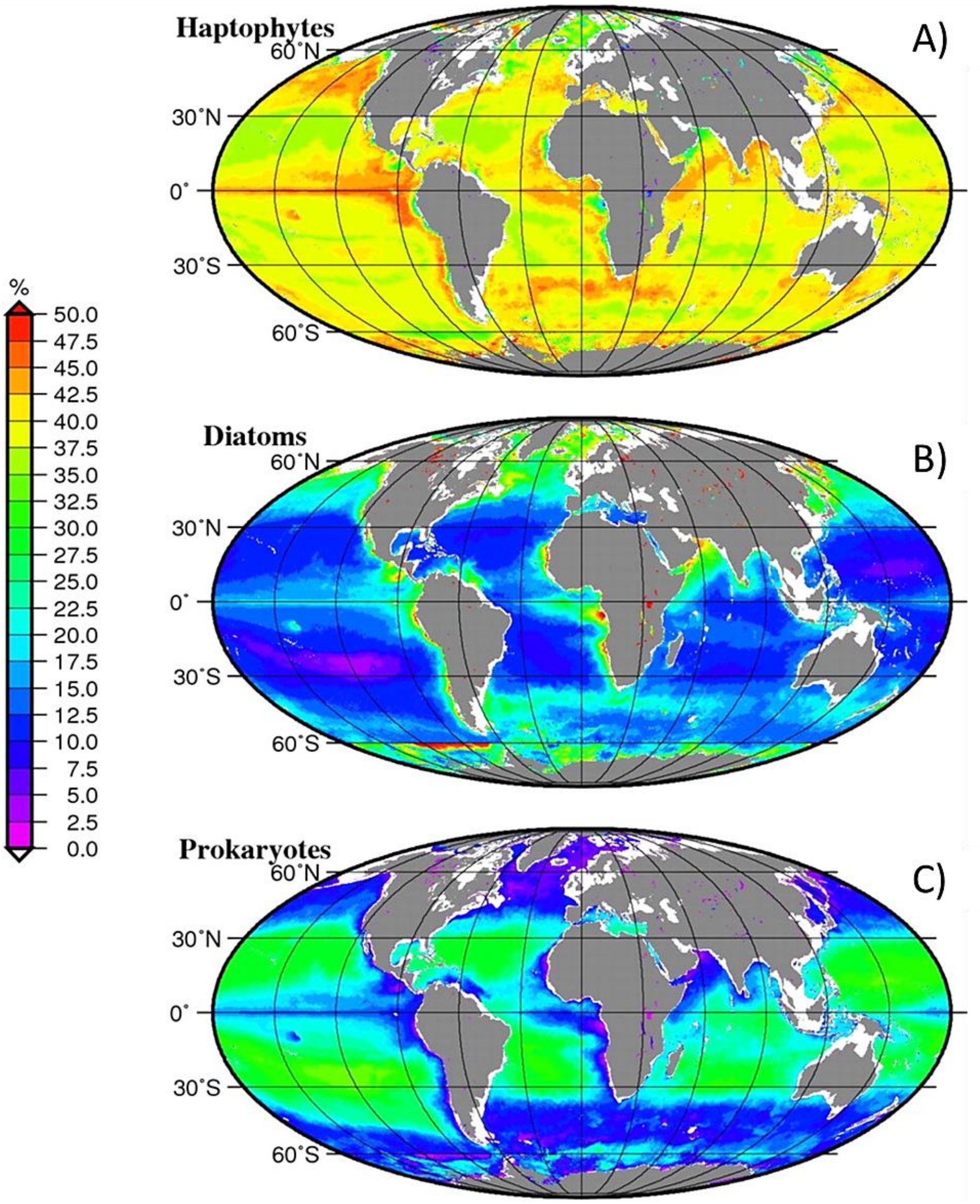


Figure 1.5. The relative contribution of select taxonomic groups to total marine chlorophyll-a biomass.

Shown is the relative percentage contribution of A) Haptophytes, B) Diatoms and C) photosynthetic bacteria to the total global chlorophyll-a biomass within the photic layer of the global oceans. Haptophytes are seen to be globally dominant over much of the global oceans at 30-50% total photosynthetic standing stock. Diatoms are seen to be distributed primary around coastal regions and Polar environments. Photosynthetic bacteria primarily occupy the open ocean around mid-latitudes with a near complete absence within Polar environments. Figure adapted from[83].

associated with cold waters and species such as *Phaeocystis globosa* associated with temperate waters[84]. Historically *E. huxleyi* was suggested to be absent from polar waters, but more recently it's northward expansion into the Arctic has been observed[44], and genetically distinct Arctic ecotypes have been resolved[85] implying a globally ubiquitous distribution.

Haptophytes are involved in a number of key ecosystem functions, being important contributors to marine primary production, but contribute a small fraction of ~10% carbon export to the deep ocean due to their typically smaller size and low sinking rates[86]. They are also key bacterivores, and have been demonstrated to be responsible for 27% of total bacterivory in marine systems[87]. Toxins released during blooms are detrimental to shellfish and zooplankton[88], and therefore are of ecological importance to marine food webs. Haptophytes produce significant quantities of dimethyl sulphide (DMS), a cloud condensation nuclei, which account for ~13% of the global DMS flux and therefore impacts climate regulation and the marine sulphur cycle[89]. Furthermore, members of the coccolithophore group within the Haptophytes, such as *E. huxleyi,* are characterised by their formation of extracellular calcareous "liths", and thus calcifying Haptophytes are key regulators of ocean calcium carbonate concentrations[86].

1.7.2 Diatoms

Diatoms are a bloom forming taxonomic group characterised by rigid silica frustules that surround the cell, and are capable of growth as filaments, chains or colonies. Diatoms are thought to have arisen by the endosymbiosis of a red algae by a eukaryotic heterotroph around 500 million years ago and differentiated as early as 250 million years ago[90]. They have become one of the most genetically diverse groups of phytoplankton present in the global ocean, with species richness estimated to lie between 12,000-30,000[91,92], some of which are capable of mixotrophy[93] whereby phototrophic and heterotrophic energy utilisation are used simultaneously[94]. Cell size is highly variable, being 2-6 x10^9 μm^3 dependent on species, with *Minutocellus spp.* noted as one of the smallest and *Ethmodiscus spp.* as one of the largest[95]. Diatoms are largely considered to be globally ubiquitous, but display particularly high abundances in nutrient rich polar regions, coastal environments and upwelling regions (Figure 1.5)[96]. The genus *Chaetoceros, Fragilariopsis* and *Thalassiosira* constitute the most globally abundant and genetically diverse Diatoms (Figure 1.6)[96]. However, biogeographic patterns have been observed within the Diatoms and certain

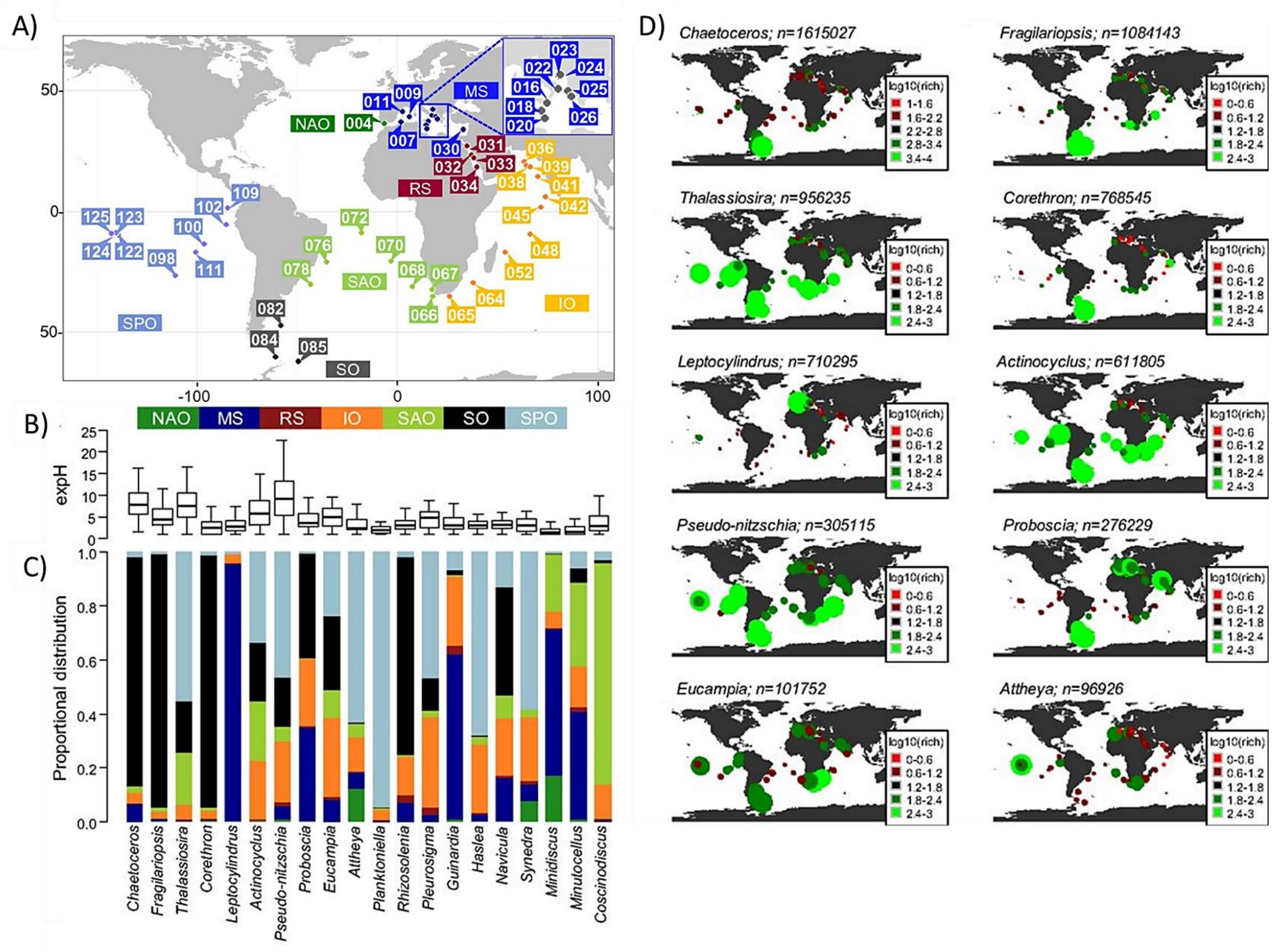


Figure 1.6. Global distribution and diversity of marine Diatoms.

A) The global stations sampled as part of the Tara Oceans dataset from which the distribution and diversity of marine Diatoms were inferred. Each station was sampled at the subsurface and deep chlorophyll maximum. Stations are labelled by ocean province IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic Ocean; RS, Red Sea; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean and colour coded as shown. B) The genetic diversity for the top 20 most globally abundant diatom genus inferred from the exponential Shannon Diversity Index (expH). C) The proportional distribution of sequences for each of the 20 genus in each ocean province as colour coded, which displays the proportional abundance of each genus in each ocean province. D) The global distribution of the top 10 most abundant genus, bubble sizes are scaled to the number of sequences recovered for each genus, as shown by *n*. Bubble colour represents genetic richness (red, low; green, high). Figure adapted from[96].

genetically distinct polar ecotypes, such as *Fragilariopsis cylindrus*, have been recovered[96,97].

The global success of Diatoms is attributed to factors including spore formation in response to unfavourable environmental conditions[95], protection from grazing afforded by silica frustules, and their ability to store nutrients in a large vacuole supported by the strong silica cell wall that affords them a competitive advantage in areas of high nutrient variability[98]. Iron concentrations are the limiting growth factor for most phytoplankton species in much of the global ocean, and Diatoms possess adaptions enabling them to be highly competitive in iron replete conditions of the open oceans. Diatoms store iron in ferritin proteins which can each contain 4500 iron atoms[99], they also possess modified photosynthetic complexes resulting in a reduced cellular concentrations of the high iron requiring photosystem I and cytochrome *b6f* molecules[100]. Diatoms may also replace the iron containing photosynthetic protein ferredoxin with iron free flavodoxin proteins, further reducing cellular iron demands, and which is often used as an environmental biomarker for iron stress[101].

Diatoms play key roles in vital ecosystem processes such as the silica cycle and are responsible for under saturation of silica in global surface oceans, their high sinking rates leads to the burial of 6.3×10^{12} mol Si year^{-1} [98]. They are also functionally important at the base of many marine food webs, and significantly contribute to the carbon cycle, estimated to be responsible for up to 40% of all marine global primary production[102]. Furthermore, due to their resistance to grazing, large average cell sizes, ballasting from silica frustules and high sinking rates diatoms contribute ~40% of global carbon export to depth[86].

1.7.3 Dinoflagellates

Dinoflagellates are a diverse group of phytoplankton found in all aquatic ecosystems characterised by the presence of two flagella. Some 2,000 extant species have been described, and cell sizes vary considerably from 5-2000 µm[103]. Dinoflagellates display significant biological complexity in their morphologies and life strategies, most species are autotrophic, but mixotrophic and heterotrophic species also exist. The majority of Dinoflagellates are free living, but symbionts and parasitic species account for ~7% of all Dinoflagellates[104]. Certain taxa can also form cysts when environmental conditions become unfavourable, enabling long term survival and distribution.

Karlodinium and *Gyrodinium* represent some of the most globally abundant representatives of Dinoflagellates, and are found in the pico-nano cell size fraction[105].

The ecological role of Dinoflagellates remains poorly understood, being primarily based on historic light microscopy observations, omitting the pico- and nanoplankton size classes which hold significant diversity, and are thought to be the most ecologically significant for marine Dinoflagellates[105]. Despite this, Dinoflagellates have been shown to constitute significant proportions of the eukaryotic community, on average representing ~47% of all phytoplankton diversity, and exhibit strikingly consistent taxonomic richness across the global ocean which remains largely understudied (Figure 1.7)[105]. Many Dinoflagellates produce toxins, including *Gymnodinium spp., Pyrodinium spp.* and *Alexandrium spp.*[106,107]*.* During blooms toxins can reach concentrations detrimental to higher trophic levels, as such the abundance and distributions of the aforementioned species are regularly monitored due to the health risk to humans. The production of toxins also impacts a wide collection of other aquatic life, including shellfish, and therefore have socioeconomic relevance[106]. Furthermore, Dinoflagellates play key roles in ecosystem functions, contributing up to 76% of total carbon production in some regions, with particular importance in polar environments[108]. Parasitic species such as *Blastodinium* also impact marine food webs due to infecting zooplankton communities, affecting productivity through reduced fitness and castration of host taxa[109].

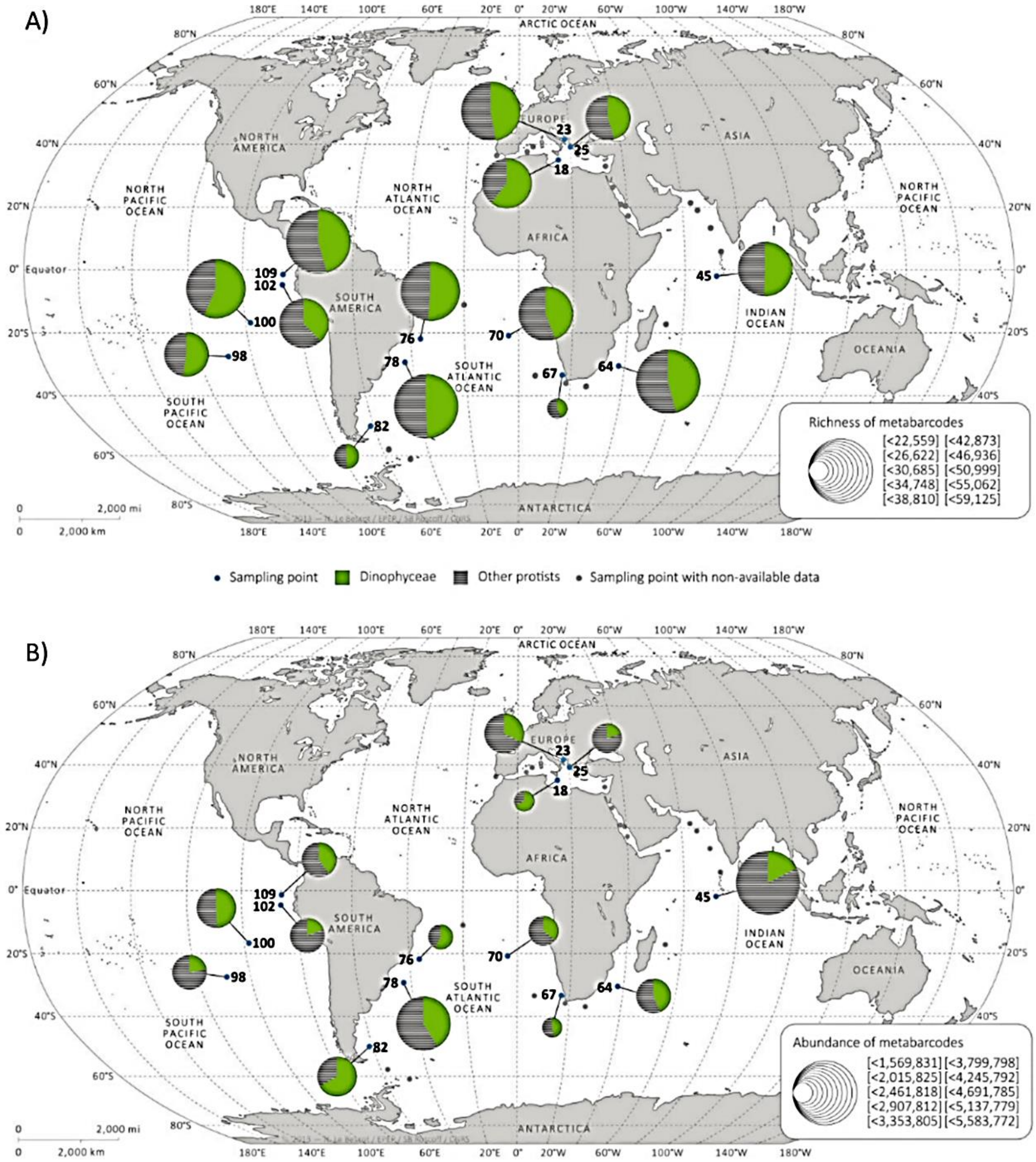


Figure 1.7. Dinoflagellate genomic richness as calculated from the Tara Oceans dataset.

Shown is the proportional richness (A) and abundance (B) of all Dinoflagellate metabarcode sequences (green) compared to the total microbial eukaryotic community (grey striped) from 33 stations sampled as part of the Tara Oceans dataset. Pie chart sizes are proportional to the total number of metabarcode sequences per sample, as shown by the key. Richness is highly consistent between all samples. Numbers represent sample site identity. Figure adapted from [105].

1.7.4 Marine Cyanobacteria

Marine cyanobacteria are primarily found at mid-latitudes in highly oligotrophic regions of the open ocean[83], but are often absent in polar regions and those containing sea ice[110] (Figure 1.5). Diversity of this group is thought to be limited, often dominated by just two of the most ecologically significant genus, *Synechococcus* and *Prochlorococcus*[111]. Combined these two taxa account for a predicted 25.2% of all global primary production[112] and are found ubiquitously in the global ocean. *Prochlorococcus* is present in waters down to a depth of 150m and shows geographic distributions between 40 °N and 40 °S, thought to be due to a temperature barrier of 15 °C exceeding its physiological limits (Figure 1.8). The geographic distribution of *Synechococcus* is wider than that of *Prochlorococcus*, suggested to be a result of a greater tolerance of temperatures down to 0°C, and greater competitiveness under high nutrient conditions, but appears less adapted to depth instead displaying preferences for surface waters[112]. Furthermore, it is thought that picoeukaryotic phytoplankton are more competitive than *Synechococcus* and *Prochlorococcus* under the high nutrient conditions associated with polar regions, partly explaining their absence[112].

Both *Synechococcus* and *Prochlorococcus* are often observed to co-occur, with *Prochlorococcus* frequently the more abundant[111], likely the result of its smaller size (0.5-0.7 μm compared to 0.8-1.5 μm for *Synechococcus*) or higher photosynthetic yields giving it a competitive advantage under nutrient replete conditions[113]. The small size, and associated large surface-to-volume ratio are key drivers of the global success of both *Synechococcus* and *Prochlorococcus* in highly oligotrophic open oceans, allowing efficient nutrient uptake at picomolar concentrations[113]. Significant microdiversity has been discovered with in *Prochlorococcus* strains that have significantly different light optima[114], typically termed LL (low light) and HL (high light) lineages. The diversity and efficiency of the photosynthetic apparatus of LL lineages are the drivers behind the success of this taxa at depths, where irradiance is 4 orders of magnitude below that at the surface[115]. Similar microdiversity has also been suggest for *Synechococcus,* and again is believed to promote adaption to specific ecological niches[116].

Marine cyanobacteria are involved in a number of key ecosystem functions such as forming an important food source for eukaryotic grazers in oligotrophic systems[87], the metabolism of diverse nitrogen sources (but *Prochlorococcus is* unable to utilise nitrate)[115], nutrient cycling and the carbon cycle, fixing in excess of 4 GT C year^{-1} globally[114].

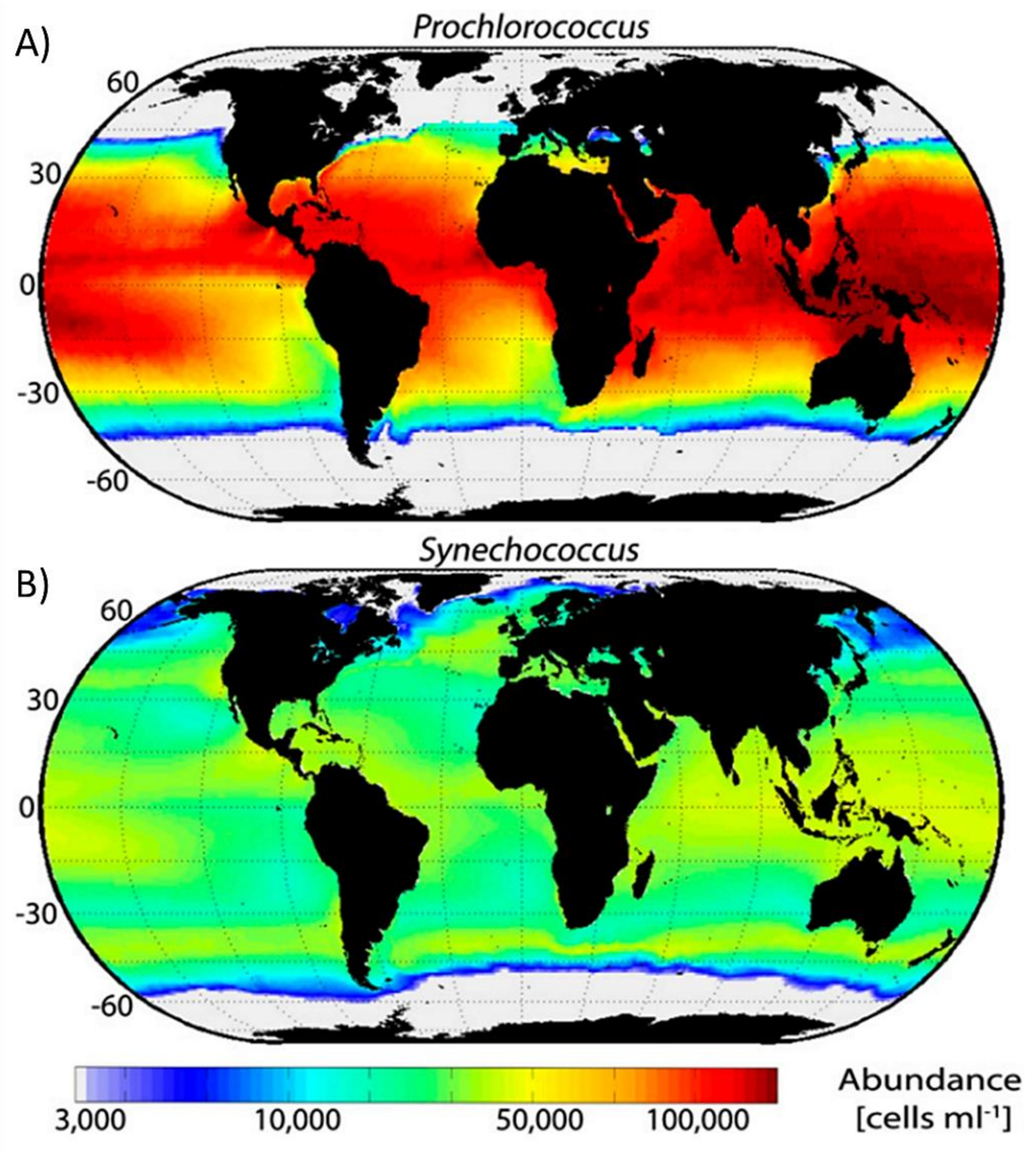


Figure 1.8. The global distribution and abundance of *Prochlorococcus* and *Synechococcus*.

(A) The distribution and abundance of *Prochlorococcus* and (B) the distribution and abundance of *Synechococcus*. Colour scale represents mean annual abundance of cells. White shows an absence of taxa. Data was taken from publically available and primary sources of flow cytometry data as detailed in [112]. Figure adapted from [112].

1.7.5 SAR11

SAR11 represented one of the first highly abundant but previously uncultivable groups of marine bacteria identified through the use of NGS studies of marine microbial diversity, eventually being bought into successful culture in 2002[117]. SAR11 is a clade of highly abundant carbon-oxidising bacteria estimated to represent ~25% of all marine plankton, yet are also found in some freshwater habitats. All members of the SAR11 clade currently identified are free-living, aerobic, chemoheterotrophs found primarily within surface waters[118]. SAR11 is of average size for a

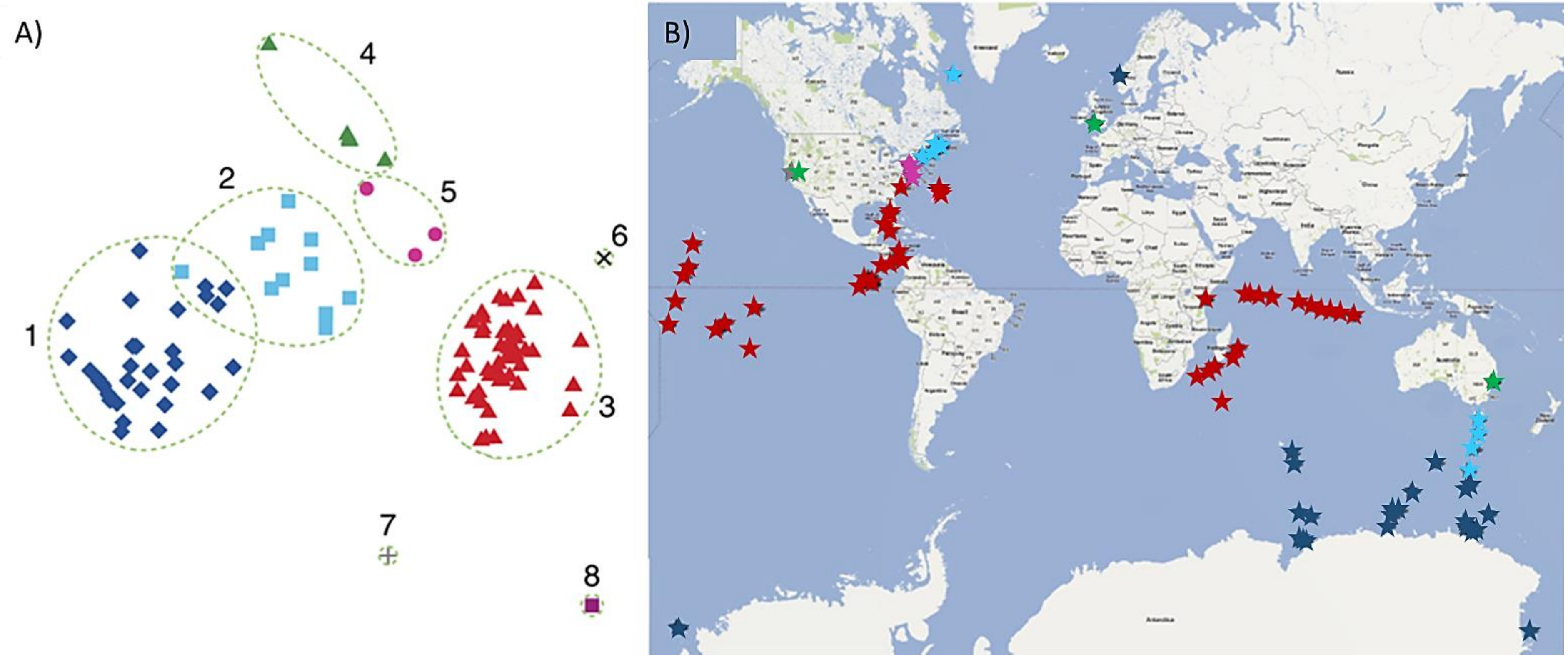


Figure 1.9. Biogeography of SAR11 phylotypes.

A) nMDS plot of Bray-Curtis similarities calculated using square root transformed SAR11 community composition and abundance data. 8 clusters can be seen based on >60% similarity of community composition as circled by green rings. Clusters are colour coded and numbered by group; (1) blue diamonds, samples from polar regions; (2) light blue squares, temperate regions; (3) red triangles, tropical regions; (4) green triangles, Western Channel Observatory (UK), Monterey Bay (USA) and Botany Bay (Australia); (5) fuchsia dots, Chesapeake Bay, Delaware Bay and nearby Nags Head; (6) black 'x', tropical Pacific Ocean and Sargasso Sea; (7) grey cross, Monterey Bay (only one phylotype); (8) pink square, Sargasso Sea (only one phylotype). B) Geographic locations of the clusters, colour coded as previously described. Figure adapted from [119].

bacterioplankton being 0.031–0.051 μm^3 [120], and features highly streamlined genomes suggested to be the result of strong selection pressure[121]. Despite being globally distributed SAR11 has been observed to display biogeographic patterns, including the existence of “cold water” and “warm water” adapted ecotypes with distinct latitudinal distributions (Figure 1.9)[119]. Ecotypes also exist which are highly abundant below the photic zone[118].

SAR11 are known to be highly functionally active within the marine environment, accounting for 30% of DMS assimilation[120], and 40% of glucose and ATP assimilation in some regions, and thus significantly contributes to global climate regulation and nutrient cycling[122]. It has also been shown to account for 50% of all amino acid metabolism in the North Atlantic, therefore substantially impacting phosphate cycling[120]. SAR11 also plays significant roles in the carbon cycle, specialising in the uptake of low-molecular weight carbon compounds[123], with different ecotypes featuring genes targeting different compounds that suggests niche differentiation within the SAR11 clade. The discovery of specialised C1 metabolism revealed previously unconsidered components of dissolved organic matter (DOM) in marine systems and likely affords SAR11 strains with a competitive advantage by providing access to a diverse array of compounds turned over at a high rate within ocean photic zones[118]. However, genomic variations between different strains have also been observed which feature different metabolic potentials limiting the breadth of compounds that can be utilised by some SAR11 representatives. For example some strains are unable to oxidise carbohydrates[124].

1.8 Limitations of Existing Arctic Microbial Studies

The biogeographic distributions of marine microbial species are often found to be related to their local environmental conditions. Climate driven changes to these factors are altering global environmental conditions, raising questions as to how the global distributions of marine microbes may be impacted. One area of particular concern is that of the Arctic region, and examples of the intrusion of Atlantic microbial species into the Arctic have already been recorded[125,126]. The potential impacts of such intrusions and distribution changes of endemic polar species are worrying as Arctic microbes provide vital ecosystem services and play key roles in a number of processes which have wide reaching impacts. Such processes include primary production, carbon sequestration, nutrient cycling[127,128] and the transformation and generation of marine dissolved organic matter, of which the Arctic receives 10% of the global total[129].

Despite increasing evidence and concern over the effects of Arctic environmental change, very little biological data exists on which accurate predictions and measurements can be made to track the ecological impacts. This is in part due to the majority of available reports concerning Arctic ecosystems focusing on the impacts of climate change on marine mammals and commercially important fish species, which is driven by the greater ease with which these organisms can be measured and both the publics' and policy makers' perceptions of significance[130]. While a few attempts have been made to explore the diversity of Arctic microbes many are limited by the methodologies and technologies applied[131–133]. Sea ice cover for much of the year inhibits access and greatly reduces the quantity of sampling that can be achieved in a single effort. As a result the majority of existing microbial assemblage data relates to the period of April to August when sea ice is lowest and access is easiest. Thus, current data is primarily limited to the duration of the spring bloom[130,134]. Furthermore, the scales addressed in many studies are often small, featuring few sampling stations separated by small, or very large distances. Therefore, little spatial data exists, particularly with regard to the Arctic Basin and Russian shelves[134]. A distinct lack of temporal data also greatly reduces the capacity to account for decadal variations in assemblage structure. The research community is aware of, and trying to combat these limitations. The first attempt to quantify the biodiversity of marine pelagic and sea-ice associated microbial eukaryotes occurred in 2011[130] by bringing together largely microscopy derived data from a vast array of sources spanning thesis publications, industry reports, published and unpublished research literature and government reports over a pan-Arctic scale. It revealed diversity to be much higher than was previously assumed, with 2,106 eukaryotic taxa recovered compared to a range of 114-863 in previous reports[130,133]. However, this first assessment suffered from much of the same limitations as those studies which preceded it, as much of the data analysed was limited to presence/absence records and lacked data which would allow the validation of reported taxa.

Another barrier to measuring the diversity of Arctic microbes is that their small size complicates their identification. Molecular studies of bacterial diversity have met with a number of limitations, such as there is often debate and inconsistencies regarding threshold values for clustering similar sequences of bacteria into distinct groups of ecological or genetic significance[135]. By contrast, studies of microbial eukaryotes have historically relied on taxonomic classification using morphology to distinguish species by way of distinct physical structures, or in the case of many mobile flagellates by observing swimming behaviour[131]. Such methods are challenging, being dependent upon the ability of highly skilled taxonomists to distinguish often subtle morphological differences between taxa. The reliance on such techniques has favoured identifying species which are easy to analyse, e.g. those which grow well in culture, are abundant, are easily preserved, and exist at a scale easily viewed by light microscopy[131]. As a result plankton in the nano and pico size classes[136] were often excluded, but have recently been shown to dominate the biomass of Arctic primary producers[137]. Many Arctic microbial eukaryotes are also fragile, failing to survive even a few degrees warming from their natural temperatures[138], again leading to their likely omission from detection.

Molecular based techniques present an alternative approach towards identifying Arctic microbial diversity. Until recently these have remained few in number, primarily due to the cost limitations historically associated with sequencing, but as the cost of NGS has fallen[66,139] a corresponding increase in its application to identifying Arctic microbial assemblages has begun. This has allowed a detailed insight into the composition of microbial communities through sequencing environmental samples, enabling the detection of uncultivable taxa[140], nano and pico sized eukaryotic taxa[141], previously cryptic taxa (morphologically identical but genetically distinct) and rare taxa[49], to achieve a greater understanding of the Arctic microbiome as a whole. However NGS studies still feature limitations. Methodological differences such as choice of sequencing technology, the post-processing pipelines employed[142], and choice of similarity thresholds for clustering similar OTUs can limit the ability to robustly compare conclusions between studies. Additionally, sequencing errors may affect the accuracy of the microbial diversity reported, and can be especially problematic for analysis of rare taxa[52]. NGS environmental studies most commonly target the small subunit ribosomal RNA (SSU rRNA) 16S (prokaryotes) or 18S (eukaryotes) genes which feature both highly conserved and highly variable regions (V1-9)[143], but differences in the region selected for analysis can affect the recovered diversity and should be an important consideration during study design[144,145].

1.9 Aims of Thesis

The overall aim of the work presented in this thesis is to analyse the diversity and distribution of pelagic marine bacterial and eukaryotic communities across a transect of 5 stations in the Norwegian Sea using Illumina MiSeq sequencing technology to provide an insight into how these communities may be impacted by predicted future environmental change to the Arctic region, and how changes in environmental conditions may impact individual taxa or taxonomic groups.

The findings presented have significant implications for polar associated community assemblages under predicted increased Atlantic mixing and warming within the Arctic region.

Chapter 2 summarises the methodologies and techniques used throughout this thesis.

Chapter 3 aims to use sea surface temperature, circulation, and *in situ* environmental data to identify and classify the sampled stations into regional groups reflective of different quantities of Polar Water influence.

Chapter 4 aims to use 16S (V5 region) amplicon sequence data generated using universal primers and the Illumina MiSeq platform to examine potential partitioning of the bacterial community across the regional station groups, delineated by differing quantities of Polar Water influence as described in Chapter 3. The partitioning of different abundance fractions of the community, and constituent taxonomic groups is discussed. Correlations of the bacterial community to environmental factors and growth patterns of individual OTUs are explored.

The hypothesis for this chapter is that there will be partitioning observed for the bacterial community across the regional station groups delineated by differing quantities of Polar Water influence as described in Chapter 3.

Chapter 5 uses a similar approach as in Chapter 4, instead using 18S (V9) amplicon sequence data with the aim of revealing distinct partitioning of the microbial eukaryotic community across the regional station groups. Again, analyses of the partitioning within the different community abundance fractions and constituent taxonomic groups, correlations to environmental factors, and growth patterns of individual OTUs are explored.

The hypothesis for this chapter is that there will be partitioning observed for the eukaryotic community across the regional station groups delineated by differing quantities of Polar Water influence as described in Chapter 3.

Chapter 6 aims to determine if the functional profiles of the active eukaryotic community members differs across the regional station groups that were resolved to feature differing amounts of Polar Water influence by analysing metatranscriptomes derived from the whole microbial eukaryotic community. Patterns of gene expression, and key ecological functions based on KEGG (Kyoto Encyclopaedia of Genes and Genomes) functional pathway analyses across the sampled stations are discussed.

The hypothesis for this chapter is that there will be an observed difference in functional profiles between the eukaryotic communities across the regional station groups delineated by differing quantities of Polar Water influence as described in Chapter 3.

Chapter 7 synthesises the results from the previous chapters, provides speculative discussion on some of potential underlying mechanisms that may explain the results observed, discusses the limitations of this study, and provides suggestions of potential future directions for the field.

Chapter 2

Materials and Methods

2.1 Environmental Analysis

2.1.1 Study Location

Water samples were collected by Cecilia Balestreri from CTD casts taken as part of cruises for the UK Ocean Acidification research program aboard the RRS James Clarke Ross research Vessel during the JR271 cruise which ran from 1st June 2012 to 2nd July 2012 (see Figure 2.1 for cruise track). This program aimed to reduce uncertainties in the predictions of changing ocean carbonate chemistry and the response of marine organisms to such stressors. The data generated is used to inform and advise policy makers and managers of marine resources (http://www.oceanacidification.org.uk).

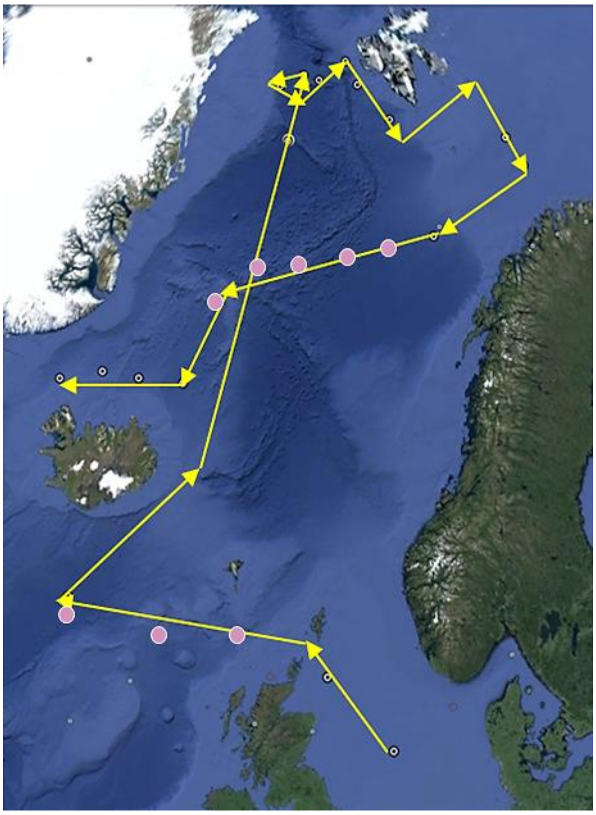


Figure 2.1. Track of the RRS James Clarke Ross research Vessel during the JR271 cruise (01/06/12-02/07/12). White points show stations sampled during the JR271 cruise and yellow arrows show the cruise path from the North Atlantic Ocean to the Norwegian Sea. Pink circles indicate the sampling points used during this study.

5 stations present in the Norwegian Sea were selected from those sampled during the JR271 cruise, these stations covered a transect across a natural temperature and salinity gradient, resolved to be resultant from varying levels of Polar Water influence, which acted to both freshen and cool the station waters (as described in Chapter 3). Three additional stations located in the North Atlantic Ocean were also sampled to provide further comparisons against a temperate community.

2.1.2 Sea Surface Temperature and Circulation Maps

Daily maps of absolute dynamic topography and sea surface temperature (SST) were created for the six month period prior to sampling by Sally Thorpe, Bristish Antarctic Survey in Matlab using the m_map package version 1.4j[146]. Absolute dynamic topography fields were calculated at 0.25 degree horizontal resolution from all remotely-sensed altimetry mission data available at a given time and referenced to the 20 year mean. The reference dataset used was the Archiving, Validation and Interpretation of Satellite Oceanographic data (AVISO)[147]. High resolution (0.05 degree) sea surface temperature data were obtained from the Operational Sea surface Temperature and Ice Analysis (OSTIA) system using both *in situ* and satellite data[148]. SST maps were used to examine the mesoscale circulation of the region during sampling.

2.1.3 CTD Casts

CTD casts were deployed at each sampling station to collect measurements of environmental data (Table 2.1). Casts were made using a standard Rosette CTD unit with either a stainless steel or titanium frame and equipped with the following sensors; Digiquartz temperature compensated pressure sensor, SeaBird-SBE 4C, SBE 3P, SBE 43, Chelsea MKIII Aquatracka fluorometer, WETLabs C-Star 25 cm path transmissometer, Biospherical QCD-905L PAR irradiance sensor, Tritech PA200 altimeter. These were used to determine the dissolved oxygen content (DO_2) [µmol l^{-1}], subsurface photosynthetically available radiation (SubsurVPAR) [µmol photons/m^2s^{-1}], pressure [dbar], density anomaly [kg/m^3], temperature [ºC], salinity and chlorophyll fluorescence [mg/m^3] directly on site at one metre intervals spanning from just below the sea surface to the sea floor. Sampled depth was calculated for all samples from the measured pressure and latitude. Nitrate, ammonium and phosphate measurements were obtained by running samples through a Skalar San+Segmented Flow Autoanalyser using colourimetric techniques[149].

Table 2.1. The CTD casts from which environmental samples were taken.

CTD casts were taken aboard the RRS James Clarke Ross research Vessel during the JR271 cruise (1st June 2012 to 2nd July 2012). Shown is the Station ID each cast was taken at from the cruise, the CTD cast ID, the oceanic region each cast was taken at, the longitude and latitude coordinates for each cast, the depth at which environmental water samples used in this study were collected at during each cast, and the date and time of CTD cast deployment.

Station	CTD cast	Original collection area	Latitude (DD.dddd°)	Longitude (DD.dddd°)	Sampled depth (m)	Date of deployment (dd/mm/yyyy)	Time of deployment (hh:mm:ss)
24	JR271 CTD 08	*North and North-West of Scotland (North Atlantic Ocean)*	60.1342	-6.7121	10.01	05/06/2012	07:03:00
25	JR271 CTD 10	*North and North-West of Scotland (North Atlantic Ocean)*	59.9710	-11.9751	19.02	06/06/2012	06:34:00
23	JR271 CTD 12	*South West of Iceland (North Atlantic Ocean)*	60.0014	-18.6702	10.01	07/06/2012	06:35:00
12	JR271 CTD 56	*Norwegian Sea*	71.7475	8.4428	19.00	26/06/2012	05:55:00
13	JR271 CTD 57	*Norwegian Sea*	71.7519	3.8717	20.00	26/06/2012	18:55:00
14	JR271 CTD 58	*Norwegian Sea*	71.7453	-1.2672	34.01	27/06/2012	05:58:00
15	JR271 CTD 59	*Norwegian Sea*	71.7517	-5.8638	25.01	27/06/2012	18:49:00
16	JR271 CTD 62	*Norwegian Sea*	70.5083	-10.1000	50.02	28/06/2012	19:18:00

2.2 Methods for Molecular Analysis - DNA

2.2.1 Total Community DNA Extractions

Total community DNA extractions were performed by Cecilia Balestreri. Water samples were collected during each CTD cast at each station at the deep chlorophyll maximum (DCM). Water was collected in Nalgene bottles previously washed with 1.5% HCl solution and rinsed three times with MilliQ water[150]. From each bottle 0.25-1.00 L of seawater was filtered by vacuum pump through a 0.45 µm polycarbonate membrane filter (PALL Corporation, Michigan, USA). Each filter was rinsed in a petri dish with 2 ml of phosphate buffered saline (PBS) solution and the resultant solution transferred to an Eppendorf tube. DNA was extracted from the PBS solution using Qiagen DNeasy Blood and Tissue kit protocol (QIAGEN, Valencia, CA, USA) before being frozen at -20 ºC for later laboratory analysis.

2.2.2 Probe Assays

Recovery of the eukaryotic community was achieved through a probe assay to isolate the V9 region of the 18S SSU rRNA gene for each extracted DNA sample. The target 18S V9 region spanned approximately 270bp, and was targeted using universal primer combinations (as shown in Table 2.2) to generate nucleotide sequences known as DNA barcodes or amplicons for taxonomic analysis across a broad spectrum of taxonomic groups[151]. Full primer information is displayed in Table 2.3.

The same approach was used to recover the bacterial community by isolating the V5 region of the 16S SSU rRNA gene[151]. The target region spanned approximately 450bp and was targeted using universal primer combination (as shown in Table 2.2) to generate DNA barcodes for taxonomic analysis. Full primer information is displayed in Table 2.3.

The generation of DNA barcodes was carried out in triplicate (technical replicates) with corresponding no-template controls (NTC) included for each sample. In some instances an additional fourth replicate was required to maximise the recovery of a sufficient quantity of DNA. The protocol was first carried out with the inclusion of Evagreen dye (Biotium, Fremont, CA, USA) to confirm the primer combinations selected successfully amplify the target region in each sample. The presence of the Evagreen dye enables the progress of the DNA amplification of each sample to be monitored, and the number of PCR reaction cycles needed to reach the cycle threshold of DNA amplification in the exponential phase to be determined[152]. The

Table 2.2. The primer combinations used for 18S V9 and 16S V5 probe assays.

Shown is the forward and reverse primer combinations used in the probe assays for each sample replicate for both the 18S (eukaryote) and 16S (bacteria) datasets.

	Eukaryote community		Bacterial community	
Sample	18S forward primers	18S reverse primer	16S forward primers	16S reverse primers
CTD08 Rep 1	1391F	EukB 2	515F	806R 1
CTD08 Rep 2	1391F	EukB 2	515F	806R 1
CTD08 Rep 3	1391F	EukB 2	515F	806R 1
CTD08 Rep 4	-	-	515F	806R 1
CTD10 Rep 1	1391F	EukB 9	515F	806R 4
CTD10 Rep 2	1391F	EukB 9	515F	806R 4
CTD10 Rep 3	1391F	EukB 9	515F	806R 4
CTD12 Rep 1	1391F	EukB 4	515F	806R 5
CTD12 Rep 2	1391F	EukB 4	515F	806R 5
CTD12 Rep 3	1391F	EukB 4	515F	806R 5
CTD12 Rep 4	-	-	515F	806R 5
CTD56 Rep 1	1391F	EukB 11	515F	806R 9
CTD56 Rep 2	1391F	EukB 11	515F	806R 9
CTD56 Rep 3	1391F	EukB 11	515F	806R 9
CTD57 Rep 1	1391F	EukB 12	515F	806R 10
CTD57 Rep 2	1391F	EukB 12	515F	806R 10
CTD57 Rep 3	1391F	EukB 12	515F	806R 10
CTD58 Rep 1	1391F	EukB 13	515F	806R 11
CTD58 Rep 2	1391F	EukB 13	515F	806R 11
CTD58 Rep 3	1391F	EukB 13	515F	806R 11
CTD58 Rep 4	-	-	515F	806R 11
CTD59 Rep 1	1391F	EukB 14	515F	806R 13
CTD59 Rep 2	1391F	EukB 14	515F	806R 13
CTD59 Rep 3	1391F	EukB 14	515F	806R 13
CTD62 Rep 1	1391F	EukB 20	515F	806R 15
CTD62 Rep 2	1391F	EukB 20	515F	806R 15
CTD62 Rep 3	1391F	EukB 20	515F	806R 15

Table 2.3. Primer sequences used for the probe assay.

The nucleotide sequences for each of the forward and reverse primers used during the probe assay, as detailed in Table 2.2, are shown. Full descriptions of each primer are displayed and delineated by '/'. Target region is the region of the 18S or 16S region to be amplified. Adapter is the adaptor sequence required to immobilise the sequence for amplification on the Illumina flow cell. Primer pad is a region to avoid primer-dimer formation. Primer linker is a sequence to prevent taxon specific PCR bias. Primer is the complimentary sequence to the target DNA barcode.

Target region	Primer name	Nucleotide sequence (Illumina adapter / primer pad / primer linker / primer)
18S V9	1391F	AATGATACGGCGACCACCGAGATCTACAC / TATGGTAATT / GT / GTACACACCGCCCGTC
18S V9	EukB 2	CAAGCAGAAGACGGCATACGAGAT / AGGACGCACTGT / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
18S V9	EukB 9	CAAGCAGAAGACGGCATACGAGAT / ACAGAGTCGGCT / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
18S V9	EukB 4	CAAGCAGAAGACGGCATACGAGAT / AACTCGTCGATG / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
18S V9	EukB 11	CAAGCAGAAGACGGCATACGAGAT / ACGGTGAGTGTC / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
18S V9	EukB 12	CAAGCAGAAGACGGCATACGAGAT / ACTCGATTCGAT / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
18S V9	EukB 13	CAAGCAGAAGACGGCATACGAGAT / AGACTGCGTACT / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
18S V9	EukB 14	CAAGCAGAAGACGGCATACGAGAT / AGCAGTCGCGAT / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
18S V9	EukB 20	CAAGCAGAAGACGGCATACGAGAT / AGACGTGCACTG / AGTCAGTCAG / CC / TGATCCTTCTGCAGGTTCACCTAC
16S V5	515F	AATGATACGGCGACCACCGAGATCTACAC / TATGGTAATT / GT / GTGCCAGCMGCCGCGGTAA
16S V5	806R 1	CAAGCAGAAGACGGCATACGAGAT / AACGCACGCTAG / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT
16S V5	806R 4	CAAGCAGAAGACGGCATACGAGAT / ACTCAGATACTC / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT
16S V5	806R 5	CAAGCAGAAGACGGCATACGAGAT / ACCAGACGATGC / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT
16S V5	806R 9	CAAGCAGAAGACGGCATACGAGAT / ACGGATCGTCAG / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT
16S V5	806R 10	CAAGCAGAAGACGGCATACGAGAT / AGCTGACTAGTC / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT
16S V5	806R 11	CAAGCAGAAGACGGCATACGAGAT / ACACTGTTCATG / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT
16S V5	806R 13	CAAGCAGAAGACGGCATACGAGAT / ACAGACCACTCA / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT
16S V5	806R 15	CAAGCAGAAGACGGCATACGAGAT / ACCAGCGACTAG / AGTCAGTCAG / CC / GGACTACHVGGGTWTCTAAT

protocol master mix was composed of 5 µl Colourless GoTaq Flexi Buffer (Promega, Madison, WI, USA), 1.5 µl $MgCl_2$ 25 mM, 2.5 µl PCR Nucleotide mix 10 mM, 1 µl Evagreen dye and 0.1 µl GoTaq DNA polymerase to 12.9 µl molecular grade water for each sample.

Reverse primers were diluted to 10 pmol/µl by adding 20 µl of each primer to 180 µl molecular grade water immediately before use. Forward primers were diluted to 10 pmol/ µl by adding 50 µl of each primer to 450 µl molecular grade water immediately before use. Diluted primers were mixed by vortex. 0.5 µl of both forward and reverse primers (10 pmol/µl) were added to each sample. 1 µl of the sample DNA was also added to give a final volume of 25 µl for each sample. Negative controls were created by adding 1 µl molecular grade water in place of DNA for each primer combination.

Quantitative polymerase chain reaction (qPCR) was used to amplify the DNA barcodes to ensure a sufficient quantity of DNA for use on the Illumina sequencing platform. Real-time qPCR was run on a Corbette Rotor-Gene 6000 with corresponding Rotor gene q Series software[152]. All sample tubes were loaded into the 36 qPCR rotor with each tube number matching to the corresponding number on the rotor. Machine settings were selected for a 3 step process with melt analysis, a rotor size of 36 tubes, confirmation of the attachment of the lock ring checked, and a final sample volume of 25 µl entered. The thermal profile was edited to reflect using an initial denaturation step of 94 ºC for three minutes, followed by up to 35 cycles of a three step qPCR involving denaturation at 94 ºC for 45 seconds, annealing at 50 ºC for 60 seconds and elongation at 72 ºC for 90 seconds. Each sample, and its corresponding negative control, were removed after the cycle which it was seen to exceed a fluorescence threshold of 80 to minimise the formation of artefacts such as chimeras during the plateau phase of the reaction (Table 2.4)[152]. Removed samples were stored in the absence of light until amplification of all samples in the run had completed.

Amplification of the 18S V9 region for samples CTD10 and CTD62 were found to require a 1:9 dilution with molecular grade water in order to amplify successfully.

Once all samples had completed they were loaded back into the qPCR machine for melt curve analysis using a ramp between 72-95 ºC in 0.5 ºC increments to confirm amplification by the primer combinations.

Table 2.4. PCR cycle after which qPCR reactions were stopped.

Shown is each probe assay sample and the cycle after which the qPCR reaction was stopped due to exceeding the required fluorescence threshold of 80. The corresponding control for each sample was also stopped at the same point.

Sample number	Number of PCR cycles before sample removal:	
	18S Samples	16S Samples
CTD08	25	27
CTD10	25	26
CTD12	21	24
CTD56	26	26
CTD57	22	22
CTD58	23	23
CTD59	24	24
CTD62	26	26

Once all samples had been successfully amplified, with controls remaining negative, the probe assay was repeated for each sample in triplicate (technical replicates) with 1µl molecular grade water in place of Evagreen dye, and amplified by qPCR carried out as previously described above[152]. Corresponding no-template controls (NTC) were also included for each sample which, as stated above, contained all reaction constituents except with molecular grade water in place of sample DNA.

2.2.3 Agarose Gel Electrophoresis

To confirm successful amplification of the target sequence region each PCR product was visualised using agarose gel electrophoresis[153]. A 1.5% agarose gel was created using 1.8 g of powdered agar dissolved in 120 ml of water. The solution was heated by microwave in 1 minute intervals, between which it was stirred, until all agar powder had fully dissolved. The solution was allowed to cool until warm. 1 µl of ethidium bromide was added to the solution and fully mixed. The gel solution was poured into a sealed electrophoresis tray, a well combe added, and the gel left to solidify. Once solidified the combe was removed and the gel submerged into an electrophoresis tank full of 1 x TAE running buffer (40 µl, Tris, 20 µl acetate, 1 mM EDTA) with an additional 1 µl of ethidium bromide. A HyperLadder 100bp (Bioline, London, UK) ladder was loaded into the gel to act as a standard molecular weight marker for visualising the fragment size of the sample. 2 µl of a marker solution (Orange G) was added to each PCR product and negative controls, which were then loaded into each well of the gel. The electrophoresis was run for 50 minutes at 110 V to separate the barcodes by size fraction and check for contamination[153]. For the 18S probe assay bands corresponding to the target barcode size of approximately 270bp, and for the 16S probe assay bands of approximately 450bp, were removed by razor blade under a UV transilluminator.

2.2.4 DNA Recovery from Agarose Gel Bands

DNA was recovered from the excised gel bands for each sample following the Zymoclean Gel DNA recovery protocol (Zymo Research, Irvine, CA, USA)[152]. The weights of the exercised gel fragments were determined (Table 2.5). 3 volumes of agarose dissolving buffer (ADB) were added for each microgram of gel fragment for each sample and incubated at 50 °C for 10 minutes until the gel had completely dissolved. Care was taken to ensure the samples didn't exceed 60 °C as this may result in denaturing of the sample DNA. The dissolved gel solution was loaded into a Zymo-spin column and placed into a collection tube for each sample then centrifuged at 13,000 rpm for 30 seconds. The flow-through was discarded by pipette to avoid contamination of the spin column rim through pouring. 200 µl of wash buffer was added to each spin column and centrifuged for 30 seconds at 13,000 rpm, the flow through was again discarded by pipette and this step repeated. Each spin column was transferred to a fresh Eppendorf tube and 10 µl molecular grade water added directly to the spin column filter for each sample before being centrifuged at 13,000 rpm for 30 seconds to elute the DNA.

Preliminary assessment of DNA yield (ng/µl) was performed on a NanoDrop 1000 spectrophotometer (Thermo Scientific, NanoDrop products, Wilmington, DE, USA)[154]. The instrument was first calibrated using 1 µl molecular grade water. 1 µl of each sample was then loaded onto the stage. All samples were deemed to display satisfactory yield.

Table 2.5. Quantities of ADB used during the Zymoclean Gel DNA recovery protocol.

Shown are the weights of the extracted gel fragments containing the target DNA barcodes for each sample replicate, and quantity of ADB added to each at a 3:1 ratio.

Sample	Gel fragment weight (g)		ADB volume added (µl)	
	18S	16S	18S	16S
CTD08 Rep 1	0.10	0.07	300	210
CTD08 Rep 2	0.12	0.09	360	270
CTD08 Rep 3	0.10	0.10	300	300
CTD08 Rep 4	-	0.13	-	390
CTD10 Rep 1	0.08	0.19	240	570
CTD10 Rep 2	0.10	0.09	300	270
CTD10 Rep 3	0.11	0.11	330	330
CTD12 Rep 1	0.11	0.08	330	240
CTD12 Rep 2	0.14	0.06	420	180
CTD12 Rep 3	0.08	0.08	240	240
CTD12 Rep 4	-	0.09	-	270
CTD56 Rep 1	0.13	0.06	390	180
CTD56 Rep 2	0.11	0.12	330	360
CTD56 Rep 3	0.09	0.07	270	210
CTD57 Rep 1	0.09	0.08	270	240
CTD57 Rep 2	0.13	0.09	390	270
CTD57 Rep 3	0.10	0.12	300	360
CTD58 Rep 1	0.08	0.13	240	390
CTD58 Rep 2	0.10	0.12	300	360
CTD58 Rep 3	0.14	0.15	420	450
CTD58 Rep 4	-	0.11	-	330
CTD59 Rep 1	0.07	0.12	210	360
CTD59 Rep 2	0.09	0.11	270	330
CTD59 Rep 3	0.08	0.13	240	390
CTD62 Rep 1	0.12	0.12	360	360
CTD62 Rep 2	0.10	0.13	300	390
CTD62 Rep 3	0.09	0.13	270	390

2.2.5 DNA Quantification

To provide an accurate qualification and quantification of the recovered DNA yield a dsDNA 12000 Series II assay (Agilent Technologies, Santa Clara, CA, USA) was performed using a corresponding Agilent DNA 1000 kit on an Agilent 2100 Bioanalyser (Agilent Technologies, Santa Clara, CA, USA)[42].

Both reaction reagents for the assay were allowed to acclimatise to room temperature in the absence of light. The blue DNA dye concentrate reagent was ensured to be thoroughly mixed by vortexing for 10 seconds, after which 25 µl was added to the red DNA gel matrix reagent. The mix was vortexed for 10 seconds before being transferred to a spin filter and centrifuged for 15 minutes at 6000 rpm. The elution was collected and the filter discarded.

A DNA chip was placed into the assay priming station and loaded with 9 µl of the reagent mix in the correct well (marked with a solid 'G'). The priming station plunger was set to 1 ml and the plunger latch set at its lowest setting. The plunger was fully depressed, secured into position by the latch and left for 60 seconds to allow the reagent mix to fill the DNA chip. The plunger was released and allowed to return to the 0.3 ml mark, before being slowly retracted back to the 1 ml mark.

9 µl of the reagent mix was added to the remaining 'G' wells. 5 µl of green DNA 12000 marker was added into each sample well and that for the standard ladder. 1 µl of standard ladder was added to the corresponding well. 1 µl of sample DNA was added to each sample well. 1 µl of molecular water was added in place of DNA for any unused sample wells. The chip was vortexed in an IKA vortex mixer for 1 minute at 2400 rpm. The chip was then loaded into the Bioanalyser and a dsDNA 12000 Series II assay run using the corresponding Agilent 2100 Expert Software. After each run the Bioanalyser was cleaned with an electrode cleaner chip filled with 350 µl molecular water and the electrode allowed to air dry. The quantity of DNA for each 18S (eukaryote) sample is shown in Table 2.6, and the quantity for each 16S (bacterial) sample is shown in Table 2.7.

Samples were deemed to be of a satisfactory quality with clear peaks observed at the target fragment length of approximately 270bp for the 18S (eukaryote) samples and approximately 450bp for the 16S (bacterial) samples on the Agilent assay graphical reports. Measurements of the quantity of DNA recovered (Table 2.6 & Table 2.7) were used to dilute samples to a final concentration of 4 nM/L for each sample and the highest quality replicate for each site selected (Table 2.8). The quality assignment of each replicate was based upon the strength of peaks generated during the Agilent dsDNA 12000 Series II assay and the total quantity of DNA recovered. Where similar quantities of DNA were recovered from two replicates the one with the narrowest peak observed during the Agilent assay was selected. 3 µl of the 4 nM/L solution of each of the chosen replicates was combined to give a pooled master mix. The master mix was sent for sequencing by Illumina MiSeq technology at The National Oceanography Centre in Southampton, England.

Table 2.6. The quantity of DNA recovered for the 18S probe assay as determined by the dsDNA 12000 Series II assay.

Shown is the total quantity of recovered DNA per sample, the concentration of the recovered DNA, and the amount of DNA and molecular water required to give a final DNA sample concentration of 4nM/L.

Sample	Recovered quantity of DNA (ng/µl)	Concentration of DNA (nM/L)	DNA required for dilution to 4Nm concentration (µl)	Molecular grade water required for dilution to 4 nM/L concentration (µl)
CTD08 Rep 1	8.88	50.00	1.60	18.40
CTD08 Rep 2	18.86	106.50	1.50	38.50
CTD08 Rep 3	26.98	153.40	1.56	58.44
CTD10 Rep 1	11.92	67.80	1.77	28.23
CTD10 Rep 2	26.06	149.30	1.61	58.39
CTD10 Rep 3	25.72	146.20	1.64	58.36
CTD12 Rep 1	4.95	28.30	1.41	8.59
CTD12 Rep 2	10.45	59.90	2.00	28.00
CTD12 Rep 3	6.27	36.10	2.22	17.78
CTD56 Rep 1	7.25	41.30	1.94	18.06
CTD56 Rep 2	10.62	60.50	1.98	28.02
CTD56 Rep 3	22.14	126.40	1.90	58.10
CTD57 Rep 1	18.70	107.30	1.86	48.14
CTD57 Rep 2	14.26	81.70	1.96	38.04
CTD57 Rep 3	20.26	116.70	1.71	48.29
CTD58 Rep 1	9.45	64.70	1.85	28.15
CTD58 Rep 2	9.76	67.40	1.78	28.22
CTD58 Rep 3	19.26	134.00	1.79	58.21
CTD59 Rep 1	5.77	40.10	2.00	18.00
CTD59 Rep 2	9.49	66.20	1.81	28.19
CTD59 Rep 3	13.33	92.70	1.73	38.27
CTD62 Rep 1	8.00	55.70	1.44	18.56
CTD62 Rep 2	12.69	88.50	1.81	38.19
CTD62 Rep 3	13.44	93.90	1.70	38.30

Table 2.7. The quantity of DNA recovered for the 16S probe assay as determined by the dsDNA 12000 Series II assay.

Shown is the total quantity of recovered DNA per sample, the concentration of the recovered DNA, and the amount of DNA and molecular water required to give a final DNA sample concentration of 4nM/L.

Sample	Recovered quantity of DNA (ng/µl)	Concentration of DNA (nM/L)	DNA required for dilution to 4Nm concentration (µl)	Molecular grade water required for dilution to 4 nM/L concentration (µl)
CTD08 Rep 1	3.40	12.70	3.15	6.85
CTD08 Rep 2	21.09	77.90	1.54	28.46
CTD08 Rep 3	20.27	74.90	1.60	28.40
CTD08 Rep 4	15.06	57.10	1.40	18.60
CTD10 Rep 1	8.71	32.40	1.23	8.77
CTD10 Rep 2	7.69	28.70	1.39	8.61
CTD10 Rep 3	5.23	19.60	2.04	7.96
CTD12 Rep 1	3.10	11.70	3.42	6.58
CTD12 Rep 2	11.31	42.50	1.88	18.12
CTD12 Rep 3	18.21	68.30	1.76	28.24
CTD12 Rep 4	5.98	22.60	1.77	8.23
CTD56 Rep 1	5.56	20.80	1.92	8.08
CTD56 Rep 2	24.71	93.20	1.72	38.28
CTD56 Rep 3	13.04	49.30	1.62	18.38
CTD57 Rep 1	6.04	22.60	1.77	8.23
CTD57 Rep 2	16.97	64.50	1.86	28.14
CTD57 Rep 3	29.64	112.50	1.78	48.22
CTD58 Rep 1	4.45	16.40	2.44	7.56
CTD58 Rep 2	15.93	59.50	2.02	27.98
CTD58 Rep 3	15.17	57.00	1.40	18.60
CTD58 Rep 4	26.09	99.00	1.62	38.38
CTD59 Rep 1	5.71	21.20	1.89	8.11
CTD59 Rep 2	3.98	15.00	2.67	7.33
CTD59 Rep 3	10.34	39.80	2.01	17.99
CTD62 Rep 1	13.68	51.40	1.56	18.44
CTD62 Rep 2	9.49	35.70	1.12	8.88
CTD62 Rep 3	3.59	13.60	2.94	7.06

Table 2.8. The highest quality DNA replicates chosen for sequencing.

Shown is each of the sample replicates for the 16S and 18S DNA barcoding chosen for Illumina MiSeq sequencing.

16S	18S
CTD08 Rep 2	CTD08 Rep 3
CTD10 Rep 1	CTD10 Rep 2
CTD12 Rep 3	CTD12 Rep 2
CTD56 Rep 2	CTD56 Rep 3
CTD57 Rep 3	CTD57 Rep 3
CTD58 Rep 4	CTD58 Rep 3
CTD59 Rep 3	CTD59 Rep 3
CTD62 Rep 1	CTD62 Rep 3

2.3 Methods for Molecular Analysis - RNA

2.3.1 RNA Extraction

RNA was extracted from water samples collected at each station during CTD casts for use in the generation of metatranscriptomes. Water samples were collected and processed for storage by Cecilia Balestreri from each station at the DCM in Nalgene bottles previously washed with 1.5% HCl solution and rinsed three times with MilliQ water[150]. From each bottle 0.25-1.00 L of seawater was filtered by vacuum pump through a 0.45 µm polycarbonate membrane filter (PALL Corporation, Michigan, USA). Each filter was stored in RNAlater and frozen at -80 °C for later analysis[155].

Once returned to the laboratory RNA was extracted. Filters were defrosted and cut into pieces approximately 5mm^3 in size using tweezers and scissors sterilised with 100% ethanol and an open flame. Filter fragments were then suspended in 1 ml of lysis buffer composed of 1 ml of RLT buffer and 10 µl β-ME under sterile conditions. Samples were vortexed and allowed to stand for 30 minutes, after which 1 ml of supernatant and 1 ml 70% ethanol were added to a fresh Eppendorf tube and the solution vortexed again. 700 µl of the supernatant ethanol mix was transferred into a spin column and centrifuged for 15 seconds, the supernatant was discarded. 700 µl RW1 buffer was added to the spin column and centrifuged for 15 seconds. The supernatant was again discarded. 500 µl RPE buffer was added to the spin column and centrifuged for 15 seconds before discarding the supernatant. 500 µl RPE buffer was added to the spin column, centrifuged for 2 minutes and the supernatant discarded. The column was placed into a fresh 1.5 ml collection tube. 30 µl RNA-free water was applied directly to the filter and centrifuged for 1 minute to elute the RNA. All samples were then stored at -80 °C and the filters discarded.

2.3.2 RNA Quantification

The quantity of RNA extracted was assessed using an Agilent RNA Nano Chip kit and corresponding RNA 6000 assay on an Agilent 2100 Bioanalyser[156] (Agilent Technologies, Santa Clara, CA, USA). Throughout the assay RNase-free microfuge tubes and pipettes were used, and the thawed RNA samples and the assay standard ladder kept on ice to avoid excessive denaturing.

Both reaction reagents for the assay were allowed to acclimatise to room temperature in the absence of light. 550 µl of the RNA 6000 Nano gel matrix (red reagent) was transferred into a spin filter and centrifuging at 4000 rpm for 10 minutes. 65 µl of the eluted gel was then transferred into RNase-free microfuge tubes.

The RNA 6000 Nano dye concentrate (blue) was vortexed for 10 seconds. 1 µl was added to the RNA 6000 Nano gel matrix. The resultant reagent mix was vortexed and centrifuged at 4000 rpm for 10 minutes.

An RNA Nano chip was placed into the assay priming station and loaded with 9 µl of the reagent mix to the corresponding well (marked with a solid 'G'). The priming station plunger was set to 1 ml and the plunger latch set at its top most position. The plunger was depressed until secured by the latch and left for 30 seconds to allow the reagent mix to fill the RNA chip. The plunger was released and allowed to return to the 0.3 ml mark before being slowly retracted back to the 1 ml mark.

9 µl of the reagent mix was added to the remaining 2 wells marked 'G'. 5 µl RNA 6000 Nano marker was added to all sample wells and the well marked for the ladder. 1 µl of the standard ladder was added to the corresponding well. 1 µl of sample RNA was added to each of the sample wells. 1 µl of RNA Nano marker was added in place of RNA in any unused wells. The chip was vortexed in an IKA vortex mixer for 1 minute at 2400 rpm before being loaded into the Bioanalyser and a Eukaryote Total RNA Nano II assay run using the corresponding Agilent 2100 Expert Software. After each run the Bioanalyser was cleaned with an electrode cleaner chip filled with 350 µl molecular water and the electrode allowed to air dry. The quantity of RNA recovered for each sample is shown in Table 2.9.

The extracted RNA was then sent for sequencing by Illumina MiSeq technology at The National Oceanography Centre in Southampton (NOC), England. At NOC RNA library prep was carried out by first using a Ribo-Zero Magnetic Kit (Illumina Inc. , San Diego, CA, USA) for rRNA depletion up until the ethanol precipitation of RNA, followed by the TruSeq stranded mRNA library preparation protocol (Illumina, San Diego, CA, USA).

Table 2.9. The quantity of RNA recovered from each sample as determined by the Eukaryote Total RNA Nano II assay.

Sample	Quantity of RNA (ng/µl)
CTD08	20.9
CTD10	39.5
CTD12	27.6
CTD56	47.5
CTD57	70.7
CTD58	31.6
CTD59	45.3
CTD62	80.7

2.4 Bioinformatic Analysis of Amplified DNA Barcodes

The processing of the Raw Illumina MiSeq sequences was carried out on the Biolinux 7 platform[157] using the custom pipeline as shown (Figure 2.2). This pipeline followed a workflow whereby raw sequence data returned from the Illumina MiSeq machine was first quality assessed before being run through a pre-processing step to improve the quality of the sequence dataset. The dataset was then passed through a processing stage whereby the sequences were clustered into OTUs and these OTUs annotated. After this the annotated OTUs were quality checked ready for community level analysis.

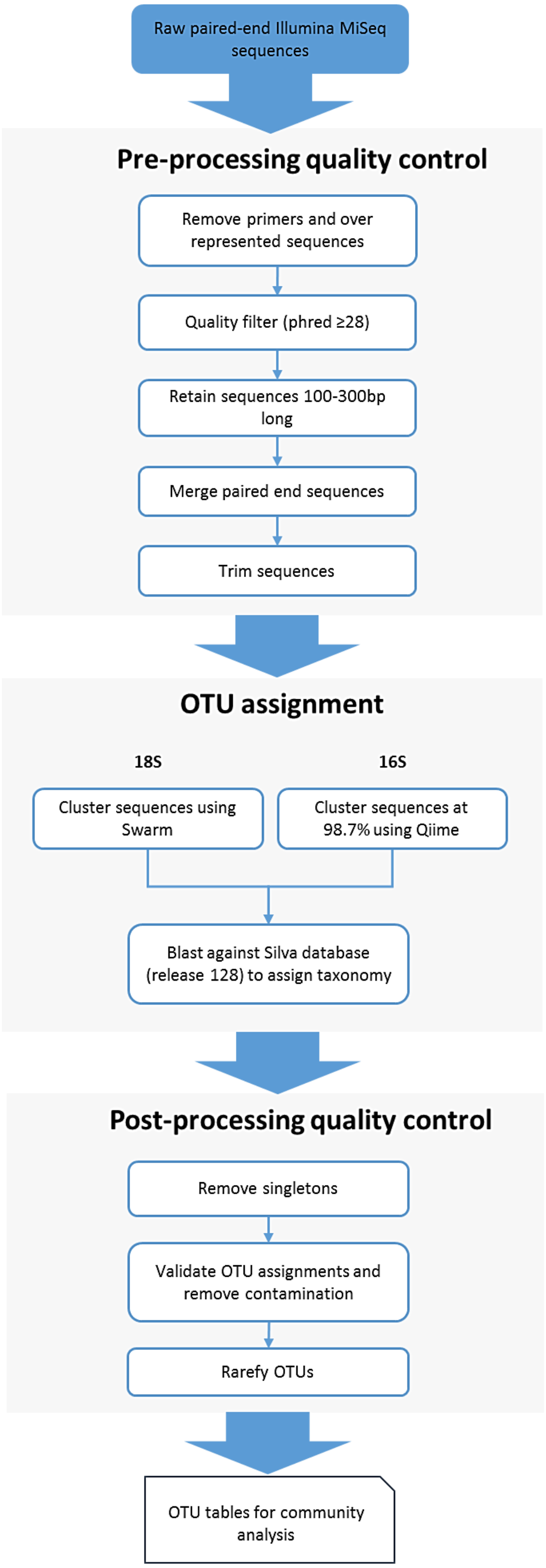

Raw paired-end Illumina MiSeq sequences
Pre-processing quality control
Remove primers and over represented sequences
Quality filter (phred ≥28)
Retain sequences 100-300bp long
Merge paired end sequences
Trim sequences
OTU assignment
18S
16S
Cluster sequences using Swarm
Cluster sequences at 98.7% using Qiime
Blast against Silva database (release 128) to assign taxonomy
Post-processing quality control
Remove singletons
Validate OTU assignments and remove contamination
Rarefy OTUs
OTU tables for community analysis

Figure 2.2. The pipeline followed for the processing of raw DNA sequence data through to data visualisation. Pre-processing included primer removal, sequence quality control, sequence trimming and merging paired-end sequences [158]. High quality sequence data generated from the pre-processing stage was used to cluster sequences into OTUs and generate taxonomic annotations. Post-processing quality control was used to omit potentially erroneous sequences and ensure comparability across different sequence depths from samples. Any singletons present in the annotated OTU table were discarded[158].

2.4.1 Raw DNA Sequence Data

A total of 3,741,702 paired-end sequences were generated from the Illumina MiSeq sequencing effort for the 16S barcode amplicons produced by the 16S probe assay, hence forth referred to as the 'bacterial dataset'.

7,533,558 paired-end sequences were generated from the Illumina MiSeq sequencing effort for the 18S barcode amplicons produced by the 18S probe assay, hence forth referred to as the 'eukaryotic dataset'.

All sequence files were initially quality assessed using FAST-QC[159] v0.11.3. Sequences were observed to have an average length of 251bp for all bacterial sequences, and 151bp for all eukaryote sequences with no sequences automatically flagged as poor quality according to FAST-QC default parameters for either dataset.

2.4.2 Pre-processing Quality Control of DNA Sequence Data

Pre-processing included primer removal, sequence quality control, sequence trimming and merging paired-end sequences to ensure low quality or potentially erroneous sequences were removed[158]. Any over represented and primer sequences present in the raw Illumina sequence datasets as flagged by FAST-QC were removed using Cutadapt v1.9.1[160]. Both forward and reverse sequences were quality filtered using PEAR v0.9.8[161] to retain only high quality sequences equal to or above a Phred score of 28, while simultaneously merging them. Sequences outside of 100-300 nucleotides long were removed with PEAR. Finally, all sequences were trimmed to a maximum length of 250bp for the bacterial dataset, and 270bp for the eukaryotic dataset, using R v3.3.0[162] to aid with alignment. These steps ensured poor quality data did not interfere with downstream processing.

The sequences retained after pre-processing quality control were deemed of satisfactory quality to be processed for OTU assignment and taxonomic annotation.

2.4.3 Bacteria OTU Assignment and Taxonomic Annotation

Qiime was used to cluster bacterial sequences into OTUs. First a mapping file was created to enable all sequences from all stations to be combined into a single FASTA file. All sequences were then clustered into OTUs based on a 98.7 % similarity threshold, suggested to be the most appropriate threshold for delineating bacterial sequences into OTUs analogous to species level characterisations[163]. Representative sequences for each OTU were selected based on the most abundant sequence for each OTU.

Taxonomy was assigned to each representative OTU using BLAST (Basic local alignment search tool) to search for similar sequences in the SILVA[164] database (release 128) with an e-value threshold of 10^{-8}. Lastly, the closest taxonomic annotation found for each representative sequence during BLAST was extracted and added to each corresponding OTU using R[162], resulting in a text file output of annotated unique OTUs present in each sample and their abundances.

2.4.4 Eukaryote OTU Assignment and Taxonomic Annotation

The Swarm cluster algorithm[165] was chosen as the amplicon clustering method for OTU assignment for the eukaryotic dataset. Swarm avoids a number of key limitations typically associated with other techniques, namely the use of arbitrary clustering thresholds globally applied to the entire dataset and input order induced bias. It creates robust OTUs that are more reflective of natural sequence similarities between taxa, and represents a more suitable choice for eukaryotic community sequence data than other clustering algorithms[165].

Swarm v2.1.6[165] was used to create a single dereplicated file for the entire study which contained only unique sequences, from this an amplicon contingency table of all unique OTUs in all samples was generated. All OTUs for each sample were then clustered using the Swarm v2.1.6[165] clustering algorithm with one ambiguous nucleotide allowed between OTUs.

Taxonomic annotations were determined for each OTU using Qiime v1.9.1[166] to BLAST each sequence against the SILVA[164] database (release 128) with an e-value threshold of 10^{-8}. R[162] was used to add the taxonomic annotation to each corresponding OTU and consolidate them based upon taxonomic assignment, resulting in a text file output of annotated unique OTUs present in each sample and their abundances.

2.4.5 Post-processing Quality Control of DNA Sequence Data

The high throughput of NGS technologies means that there is a risk of potentially erroneously base calls during the sequencing run. The risk of these is relatively low, and as such the number of sequences likely to be erroneous is small. In order to ensure such potentially erroneously sequences are not included in the final dataset post-processing quality control was applied to the OTU tables produced for the bacterial and eukaryotic datasets.

Any singletons, defined as OTUs identified by only one sequence across the entire study, where excluded from both the bacterial and eukaryotic datasets, respectively[158]. Within the bacterial dataset, sequences found to match to chloroplasts and mitochondria were removed as they are potentially the result of eukaryotic contaminants. Despite the SILVA[164] database being a curated dataset of taxonomic assignments the submissions are still prone to user error or incorrect formatting. As such the taxonomic assignment of each OTU in both the eukaryotic and bacterial datasets was manually validated and any necessary amendments made. Due to different sequencing numbers being recovered for each station the dataset was rarefied (subsampled) to the smallest number of sequences recovered at a single station to normalise the data and ensure comparability across stations[167]. This was achieved by using the *"rarefy"* function in the R package '*vegan*'[168].

2.5 DNA Community Analysis and Generation of Diversity Outputs

2.5.1 Rarefaction Analysis

In order to examine whether the Illumina MiSeq sequencing effort was sufficient and whether suitable sequence depth had been achieved rarefaction curves were constructed from the observed OTU richness and extrapolated to predicted OTU richness from the rarefied dataset using the R "iNEXT" package[169] in a custom script. For each dataset this was completed for each station, each station group, and the whole dataset.

2.5.2 Generation of α-diversity Metrics

To analyse the biological diversity recovered at each station α-diversity metrics calculated as part of the rarefaction analysis using the "iNEXT" R package[169] were recorded, these were also compared to the ACE diversity estimator calculated using the "estimateR" function in the R package '*EpiEstim*'[170] for each dataset using custom scripts.

2.5.3 Generation of β-diversity Metrics

To compare the β-diversity across the stations a Bray-Curtis dissimilarity matrix between OTUs recovered at each station were generated using the "vegdist" function in the R package '*vegan*'[168] and the resultant degree of dissimilarity between stations displayed as a dendrogram using custom scripts[49]. Additionally, this was carried out for OTUs representing each of the constituent major taxonomic groups for both the eukaryotic and bacterial datasets, and for the different abundance fractions of each community[49]. Abundance fractions were separated following thresholds commonly used in previous studies[48–50]. Fractions were designated as the abundant fraction (OTUs representing ≥1% of the total community), the intermediate fraction (0.01-1%) and rare fraction (≤0.01%).

2.5.4 Community Composition Visualisation

Krona[171] was used to visualise the relative abundance and composition of taxa at different hierarchal levels for both the eukaryotic and bacterial datasets at each station. Krona[171] charts were generated for each station by first creating an Excel template of the hierarchically separated taxonomic annotations for each OTU for all stations. 12 levels were used for the eukaryotic dataset and 6 levels for the bacterial dataset reflecting the depth of taxonomic levels available for each respective dataset in the SILVA[164] database. The Excel templates were then processed with the KronaTools script package[171] using custom bash scripts on the Biolinux 7 platform to generate interactive XML files in which different taxonomic levels could be navigated.

2.5.5 OTU Distributions and Community Distinctness

Venn diagrams were created using the *'VennDiagram'* R package[172] to visualise the distribution of shared and specific taxa by comparing the presence/absence of OTUs in order to determine how distinct the communities at different stations were. To determine whether the patterns of station specific OTUs were the result of low abundance OTUs Venn diagrams were also created for all OTUs with a T10 filter (OTUs represented by <10 sequences excluded) applied[152]. Additionally Venn diagrams were created for the top 200 most abundant OTUs to examine the distinctness of the most abundant members of the community.

To explore the distribution pattern of the most prominent community members the relative abundances, expressed as percentages, of the top 200 most abundant OTUs across each of the five transect stations were calculated and plotted using R[162] .

Heat maps of the relative abundance of OTUs across stations were generated for the whole community, top 200 most abundant OTUs, and major taxonomic groups for each of the bacterial and eukaryotic dataset using the “heatmap.2” function in the R package ‘*gplots*’[173]. To exclude any potential bias introduced through very low sequence number a T10 filter was applied.

2.5.6 Correlation with Environmental Factors

Principle coordinate analysis (PCoA) was used to quantify the dissimilarity of the eukaryotic and bacterial communities between stations explained by environmental factors. PCoA is a more suitable method than a Principle component analysis (PCA) for environmental data which features the absence of OTUs at some stations[174]. Pearson’s correlation was used to correlate *in situ* environmental factors from each station with the first and second eigenvalues generated by the PCoA from a Bray-Curtis dissimilarity matrix of each community. This was achieved using the “rcorr” function ” in the R package ‘*vegan*’[168].

The “rcorr” function ” in the R package ‘*vegan*’[168] was also used to explore the correlation of individual OTUs to specific *in situ* environmental factors by a Spearman’s rank correlation of the sequence abundance of each of the top 200 most abundant OTUs from both the bacterial and eukaryotic communities.

2.6 Bioinformatic Analysis of RNA Metatranscriptomes

Processing of the Raw Illumina MiSeq sequences recovered from the sequencing effort was carried out using the software package DNAStar 15[175] following the pipeline as shown in Figure 2.3.

The pipeline followed a workflow whereby raw Illumina MiSeq sequence data was first quality assessed, before entering a pre-processing stop to improve the quality of the sequenced dataset. The dataset was then passed through a processing stage whereby the sequences where assembled into transcripts and annotated. The resultant transcript annotations were then quality checked before diversity analyses.

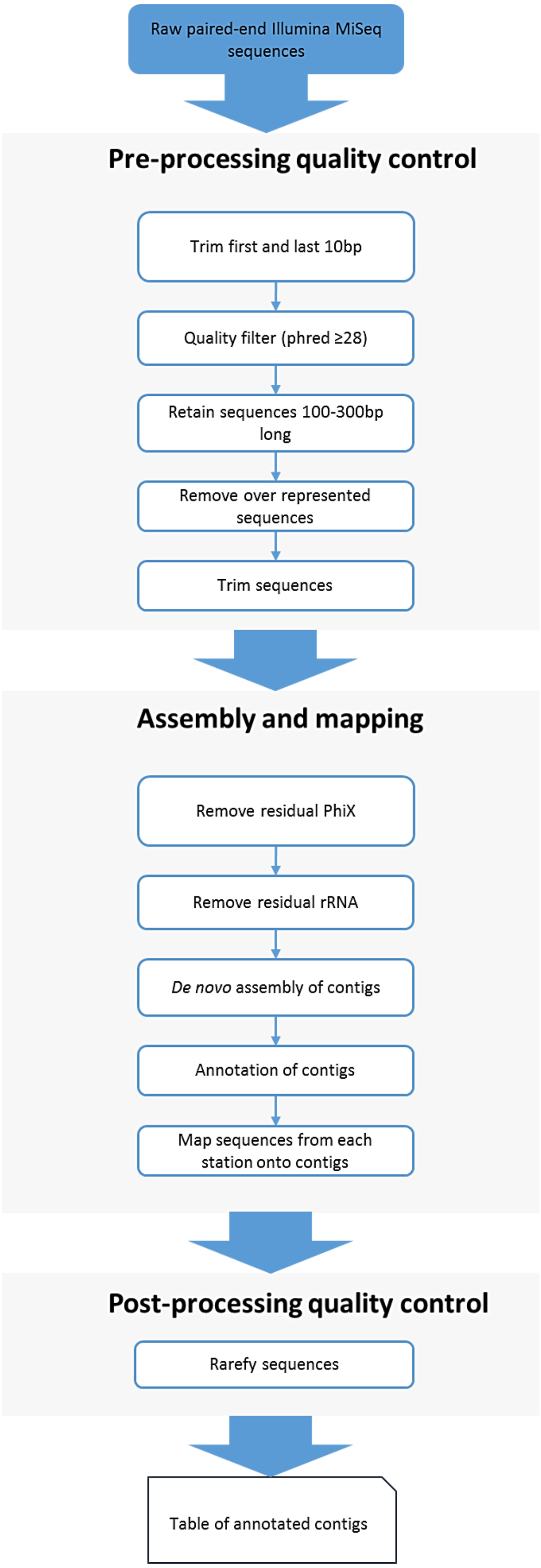

Raw paired-end Illumina MiSeq sequences
Pre-processing quality control
Trim first and last 10bp
Quality filter (phred ≥28)
Retain sequences 100-300bp long
Remove over represented sequences
Trim sequences
Assembly and mapping
Remove residual PhiX
Remove residual rRNA
De novo assembly of contigs
Annotation of contigs
Map sequences from each station onto contigs
Post-processing quality control
Rarefy sequences
Table of annotated contigs

Figure 2.3. The pipeline followed for the processing of raw RNA sequence data through to data visualisation.

High quality sequence data generated from the pre-processing stage was assembled into contigs. These were used as reference sequences for the mapping of sequences from each station as is standard protocol for novel transcripts[176]. Post-processing quality control was used to omit potentially erroneous annotations.

2.6.1 Raw RNA Sequence Data

A total of 18,199,192 raw pair end sequences were recovered from the Illumina MiSeq sequencing effort across all stations. All sequence files were initially quality assessed using FAST-QC[159] v0.11.3. Sequences had an average length of 257bp. No sequences were automatically flagged as being of poor quality according to FAST-QC default parameters.

2.6.2 Pre-processing Quality Control of RNA Sequence Data

For all sequence files the first and last 10 nucleotides from each sequence were removed as they may feature potential sequencing bias which could affect the quality of the transcriptome assembly[152]. Both forward and reverse sequences were quality filtered using PEAR v0.9.8[161] to retain only high quality sequences equal to or above a Phred score of 28. The recovered sequences were left unmerged as DNAStar is capable of managing paired-end sequences during assembly. Any over represented or primer sequences as flagged by FAST-QC[159] were removed using Cutadapt v1.9.1[160]. Sequences outside of 100-300bp long were removed. Finally, all sequences were trimmed to a maximum length of 250bp using R v3.3.0[162]. These steps ensured poor quality data did not interfere with downstream processing. 16,536,182 paired-end sequences were retained and available for further processing. These sequences were deemed of satisfactory quality for transcriptome assembly.

2.6.3 Transcriptome Assembly

The quality filtered data files output after pre-processing were loaded into the DNAStar platform[175] and a d*e novo* transcriptome assembly ran using the SeqMan NGen[175] application on the combined sample files[176]. Options were selected to filter out reads matching rRNA and PhiX contamination, an insert size of 400bp, and that the sequence technology used to generate the sequences was Illumina.

All three curated RefSeq reference databases included in the DNAStar platform (namely bacteria_r79_20161109, protozoa_r79_20161109 and viral_r79_20161109) were selected to allow the automated annotation of assembled contigs. All other options were left at their default values and the assembly started. The unassembled sequences from each station were then mapped against the assembled contigs using the SeqMan NGen[175] application, as is standard protocol for novel transcripts[176]. Options were selected to include both identified and novel contigs assembled during the de *novo* assembly, an insert size of 400bp, not to count variants and that the sequence technology used to generate the sequences was Illumina. The result was an output of each annotated contig and the number of sequences matching each contig for each station.

2.6.4 Post-processing Quality Control of RNA Sequence Data

Average coverage for the whole dataset was calculated using the formula shown in equation 2.1 to determine coverage for each contig, and then averaged for all contigs:

$$\boldsymbol{c} = \mathbf{R \times (L_R/L_C)}$$

Equation 2.1. Equation to calculate contig coverage.

Where R= contig read number, LR = average read length, LC = contig length.

All samples were rarefied to the same sequence number to remove potential sampling bias from different library sizes using the "rarefy" function in the R package '*vegan*'[168].

2.7 Metatranscriptomic Community Analysis

2.7.1 Community Composition

The relative abundance, as percentage, of representative taxonomic groups as annotated for each metatranscriptomic sequence was plotted using the R package '*gplots*'[173].

2.7.2 β-diversity of Functional Profiles

The β-diversity between functional profiles across the stations was visualised by plotting an MDS plot of the log fold change (logFC) in gene expression between all pairs of stations using the R Bioconductor package '*edgeR*'[177]. Bioconductor applies its own normalisation method, the default being trimmed-mean of M values (TMM), which accounts for different sequence depths between samples to ensure comparability, and therefore the dataset was not rarefied for this analysis. As stated in the '*edgeR*' software guide these normalisation steps are *edgeR* analysis specific and should not be used as generalised normalised pseudo-counts for analyses outside of the Bioconductor package[178], as such the dataset was rarefied to the smallest number of sequences at any one station for all other analyses.

Venn diagrams were created using the *'VennDiagram'* R package[172] to visualise the distribution of shared and specific genes in order to determine how distinct the transcriptomes at each station were.

2.7.3 Correlation with Environmental Factors

Correlation of the observed functional profiles with environmental data was explored by Principle coordinate analysis (PCoA), using the methods described in section 2.5.6.

2.7.4 Functional Annotation

Functional information for the identified genes was determined for all eukaryotic contigs by KEGG annotation as is common for environmental analyses[42,156]. KEGG analysis was performed using the publically available web based Kyoto Encyclopaedia of Genes and Genomes (KEGG) database[179]. The FASTA file of all eukaryotic contigs was uploaded to the KEGG Automatic Annotation Server (KAAS)[180] and the option to assign KEGG annotations using the single-bidirectional best hit method selected. This returned a list of contigs with assigned KEGG Orthology (KO) identifiers. The identifiers for each sequence were then categorised into functional groups according to KEGG BRITE hierarchy, excluding human disease, using the KEGG Mapper mapping tool. The count of KO identifiers belonging to each KEGG functional group, and the relative expression of KEGG annotated sequences at each station, were displayed using the R package '*gplots*'[173].

2.7.5 Gene Expression Profiles

Heat maps were generated of the relative expression, as percentage, of all KEGG functionally annotated sequences divided into their constituent representative taxonomic groups at each station using the R package *'heatmaply'*[168].

Similar analysis was performed to generate heat maps of the relative gene expression profiles across stations for all annotated genes, and cluster analysis performed using the R package *‘heatmaply’*[168]. The identity of annotated genes within each cluster was determined and recorded.

Gene expression profiles were constructed for the top 20 most abundant annotated genes using R[162] by plotting the normalised sequence count of each gene at each station.

Chapter 3

Regional Environmental Analysis

3.1 Introduction

The composition of local microbial assemblages are a result of the habitat in which they are found as environmental factors act to apply pressures and select for certain individuals most adapted to the current environmental conditions[181]. The Arctic Ocean presents a number of traits which pose both unique physiological challenges and novel niches that have sculpted the evolution of the regional microbial life. Such traits include seasonal fluctuations in ice cover, extreme annual irradiance patterns, an average water temperature of 0 ºC, strong halocline, and relatively high geographic isolation[110,127,150]. Yet many of these unique traits are under threat as the Arctic Ocean is currently experiencing significant environmental perturbations as a result of global climate change, which is particularly amplified in the Arctic[182] (Figure 3.1). The environmental processes which underpin this amplification are complex, but are hypothesised to include changes in snow/ice cover[183], ocean circulation[184], cloud cover[185], atmospheric forcing[186] and precipitation[187]. These changes are altering environmental conditions in Arctic regions, including by increasing average water temperatures by up to 0.8°C per decade regionally[182], altering regional hydrography[188], and reducing sea ice extent[189]. The rate of sea ice loss is being driven by a positive feedback mechanism, with the transition to a seasonal ice zone predicted before 2100[183]. Such a mechanism implies a degree of irreversibility to the environmental changes in the region[190].

The maximum warming observed appears most pronounced at the ocean surface and trends in surface water temperatures correlate with those of ice cover[191]. Ice cover forms a key component of the Arctic ecosystem affecting regional salinity and irradiance levels, and the amplified surface warming observed has been shown to be closely linked to the loss and thinning of ice cover[191]. This ice cover can be broadly distinguished as one of two types, Perennial Ice (PI) and Multiyear Ice (MYI)[190]. The extent of the loss of ice cover is reaching critical levels with record lows achieved repeatedly over the last decade. To highlight the severity of this trend the 2007 record low was 28% below the previous (2005) record and 37% below the climatological

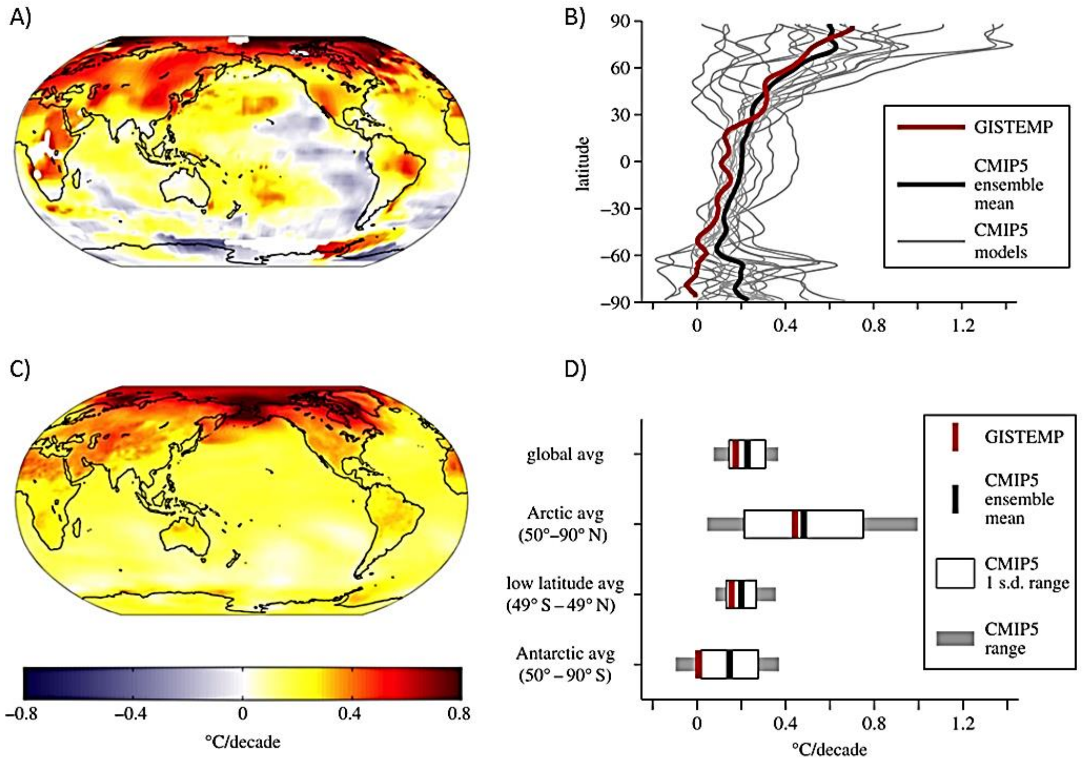


Figure 3.1. Surface temperature trends during 1979–2005 which show the temperature amplification in the Arctic region.

Shown are the surface temperature trends during 1979-2005 as calculated by the GISS Surface Temperature Analysis and CMIP5 models. A) The GISS Surface Temperature Analysis (GISTEMP), the extent of temperature amplification in the Arctic can be seen by the red shading. B) Zonal-mean surface temperature trends from GISTEMP (red line), CMIP5 ensemble mean (black line), individual CMIP5models (grey lines). C) An ensemble of CMIP5 models for surface temperature over the same time period which reflects similar findings. D) Surface temperature trends averaged over latitude bands (global, Arctic, low-to-mid latitudes and Antarctic), for GISTEMP (red line) and CMIP5 ensemble mean (black line); the white boxes show the CMIP5 ensemble ±1 s.d. range, and the grey boxes show the full CMIP5 ensemble range. Results clearly show increased temperature amplification in the Arctic latitude band. All trends are expressed in °C/decade. Figure adapted from[182].

average[192]. Shortly after, in 2012, the record was again broken by a reduction in ice cover to 49% that of the average for 1978-2010[193]. The rate of this loss is increasing, the 2007 low measured as a 8.3% ±0.6% reduction per decade but by 2012 had increased significantly, reaching 13.5% ±1.6% per decade for PI and 17.5% ±2.4% per decade for MYI (Figure 3.2)[193]. The increased rate of loss of MYI places PI at an increased vulnerability to loss during the summer melt period. Indeed, the most dramatic sea ice loss has been observed over the summer period[189]. The greatest concern however is that of decreasing winter ice cover as increasingly relatively warmer winter seasons are thought to be acting to drive a feedback mechanism which inhibits ice cover and thickness growth over the winter months. This mechanism is driven by the comparatively low surface albedo of open water fraction acting to increase the heat absorbed by the Arctic Ocean, which in turn acts to limit the annual

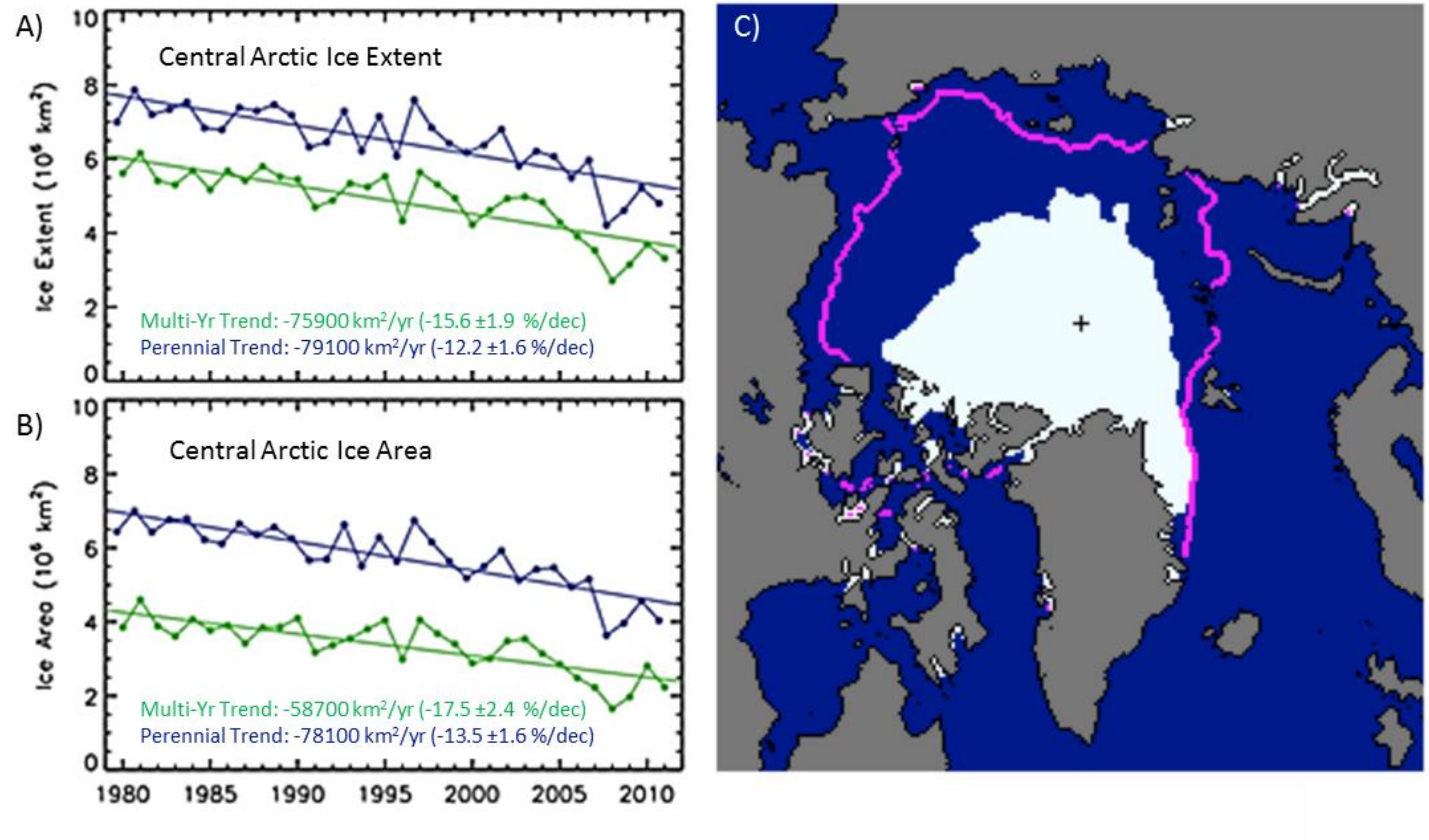


Figure 3.2. The decline of Arctic sea ice between 1979-2011.

The reduction in annual perennial and multiyear ice extent (A) and ice area (B) from 1979 to 2010. MYI values are averages of December–February values. Adapted from[192]. C) extent of Arctic sea ice cover (white) for September during the 2012 record low, representing a 49% decline below the median climatology for 1979–2011 (magenta line). The black cross is the geographic North Pole. Figure adapted from[193].

recovery of the ice cover and subsequently leads to greater summer sea ice decline[183]. However, recent work suggests that this feedback mechanism may play a subsidiary role to carbon dioxide forcing as the main driver of the temperature amplification and sea ice loss seen in the Arctic[186]. The extent of the irreversibility of ice-loss in the Arctic remains an open debate, but models predict a nearly ice free Arctic summer sometime this century. The exact timing of this event is unclear but 7 different models validated against observed trends place it between 2020-2060, with a mean date of 2042[193], and recent analyses report the extent of sea ice has reached a 66% reduction compared to the previous 6 decades as of 2018[189]. The consequences of such a loss of sea ice remain largely unknown but it has already been well documented that the level of primary production in Arctic waters has increased by up to 30% over the last 20 years as a result of climate driven reductions in the quantity and duration of seasonal ice cover[194].

There are other key processes undergoing change in the Arctic. Increased melt water from sea ice has led to concerns over the effect of increased stratification on long term nutrient reservoirs and nutrient cycling, reducing mixing with nutrient rich deep waters. A greater dependency upon recycled nutrient sources will have detrimental impacts on the endemic microbial assemblages, and consequently higher trophic levels[195]. Additionally, while wind circulation patterns have remained largely unchanged at present there have been an increasing number of summer cyclones over the last 50 years[196]. Alterations to wind patterns may lead to increased ocean turbulence across the Arctic basin margins and impact ocean stratification[197]. The direction of wind alterations is also a concern as changes in poleward atmospheric heat transport may enhance vertical warming[191]. By contrast a net increase in northerly winds could drive additional sea ice export via the Fram Strait increasing ice melt, although current satellite data shows only minor perturbations[190]. The increased level of warming across the Arctic also has important implications for meteorological factors. Water vapour is a powerful greenhouse gas and levels are expected to increase as surface warming and sea ice loss continues. Precipitation has already been demonstrated to have increased across the Arctic, and a transition from snow to rain dominated precipitation is expected during the 21st century[187].

The increased warming, loss of sea ice, and alterations to stratification are all resulting in what has been termed "Atlantification" of the Western Arctic region, whereby the characteristic temperatures and salinities observed for the Arctic Ocean are being diminished and the region transitioning into a common nutrient regime[198].

This transition is exasperated by increased inflow of Atlantic waters, which is largely attributed to the decline of sea ice reaching a level whereby fresh water supply is no longer able to sustain the stratified Arctic water layer in some regions[199]. Such Atlantification of the Arctic will have significant impacts to the biota of both the Atlantic and Arctic as a result of new competition between extant communities in these mixing regions. In this chapter the environmental conditions and hydrography of a transect of 5 stations in the Norwegian Sea at the inflow to the Arctic are analysed to provide environmental context for the microbial community analyses presented in Chapters 4, 5 and 6.

3.2 Results

3.2.1 Regional Grouping

Daily maps of sea surface temperature (SST) were created for the six month period prior to sampling to examine the mesoscale circulation of the sampled region. These maps allowed the identification of three regional groups across the sampled region. Maps for a 4 week period before, and 4 week period after sampling, at 2 week intervals are shown (Figure 3.3). The assigned regional groups reflected locations where waters were observed to experience continuous influence from Polar Water (squares), those which displayed periods of, but not constant Polar Water influence (triangles), and those with little Polar Water influence (circles), as shown by the variability in cold water (blue shading) extent during the period SST maps were generated.

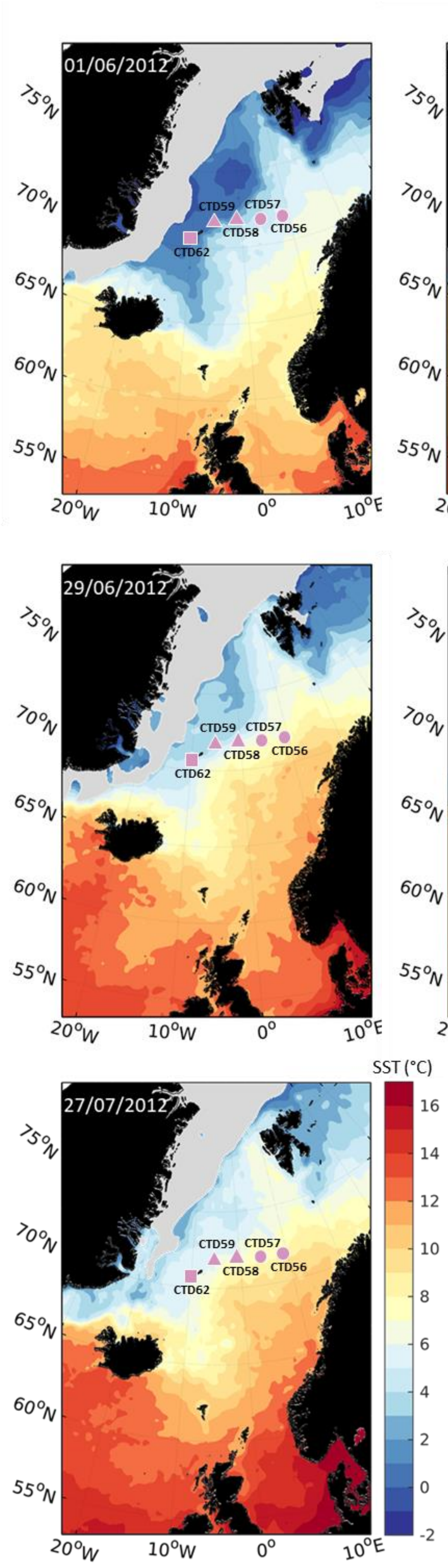


Figure 3.3. Sea surface water temperature maps prior to and after the sampled period.

Sea surface water temperature maps are shown with the locations of the stations sampled in the Norwegian Sea as part of UK Ocean Acidification research program during cruise JR271 (1st June 2012 to 2nd July 2012) for the 2 month period before and 2 month period after sampling, at 2 week intervals. Colour scale represents measured SST during the sampling period; grey shading indicates sea ice extent. Symbols represent the assigned regional group; square – constant influence from Polar Waters, triangle – intermittent periods of Polar Water influence, circle – little Polar Water influence.

3.2.2 Assignment of Station Groups

In situ measurements of the water column temperature and salinity at the DCM were taken during the sampled period (Table 3.1). These measurements were used to guide and validate the classification of the assigned regional groups from SST map analysis, and provide information about the station groupings during the sampled period. Measurements from 3 stations present in the North Atlantic were also included to act as a reference for Atlantic Water conditions. Temperature was observed to be lowest at CTD62, being 1.43°C, this was also the freshest station, featuring a salinity of 34.92. Salinity and temperature increased across the transect of stations towards those under lower Polar Water influence, with measurements at CTD56 and CTD57 being highest. The stations from the North Atlantic, namely CTD08, CTD10 and CTD12 were observed to exhibit the highest temperatures (>10°C), and greatest salinities.

Vertical profiles of the environmental characteristics were plotted to provide additional information throughout the water column at each station, and ensure the measurements of environmental factors at the depths water samples used for sequence analysis were taken fell within expected ranges (Figure 3.4). Stations CTD62 and CTD59 were observed to feature similar temperature and salinity profiles, with a thermocline of comparatively warmer fresher waters sat over colder waters of comparable freshness. In comparison, the stations CTD56 and CTD57 featured a more linear profile of warmer, more saline waters throughout their entire depth. Density profiles show similar trends due to being a function of temperature and salinity. The station CTD58 displayed a mixed profile of warm surface water similar to CTD56 and CTD57 which transitioned into cooler fresher waters similar to those observed at CTD59 and CTD62 at greater depths. The stations located in the North Atlantic, CTD08, CTD10 and CTD12 were warmer than the other stations, and featured comparatively linear temperature profiles similar to one another. However, the salinity profiles of the North Atlantic stations were varied, the surface waters at CTD12 displayed comparable freshness to CTD56 and CTD57. CTD08 was the most saline station, and the salinity at CTD10 was between CTD12 and CTD08.

From the measurements of *in situ* physical water characteristics at the sampled depth, vertical profiles, and SST maps the sampled stations were resolved to cover a natural temperature and salinity gradient as a result of differing amounts of Polar Water influence (Figure 3.5) and could be separated into four distinct groups, including the North Atlantic stations. The first group was that of stations under high Polar Water

Table 3.1. The environmental conditions at the sampled depth for which water was collected for sequence analysis.

Shown are the *in situ* values for environmental factors recorded during CTD casts for density, dissolved oxygen content, salinity, subsurface vertical photosynthetically available radiation, water temperature, ammonium, nitrate, silicate, phosphate and chlorophyll. Measurements were taken at the depth environmental water samples were collected for sequence analysis.

Station	Sampled depth (m)	Density	DO$_2$ (µmol/)]	Salinity	SubsurVPAR (µmol photons/m$^{2/}$s)	Temperature (°C)	Ammonium (nM)	Nitrate (nM)	Silicate (nM)	Phosphate (nM)	Chlorophyll (mg/m^3)
CTD08	10.01	27.215	307.4	35.43	79.11	10.437	931.50	6.53	4.27	0.45	0.58
CTD10	19.02	27.082	335.7	35.31	3.79	10.635	159.84	2.94	1.45	0.21	2.91
CTD12	10.01	27.036	314.3	35.16	37.97	10.225	69.78	6.06	1.68	0.40	1.32
CTD56	19.00	27.606	328.2	35.17	6.56	6.613	499.20	5.54	5.09	0.43	0.90
CTD57	20.00	27.614	329.0	35.19	5.87	6.644	284.86	5.92	4.57	0.42	0.73
CTD58	34.01	27.735	339.5	35.14	1.82	5.407	1036.47	8.97	5.27	0.68	1.26
CTD59	25.01	27.866	386.2	34.96	5.11	2.817	1101.83	7.19	2.27	0.59	1.74
CTD62	50.02	27.948	359.5	34.92	0.19	1.430	1204.59	7.93	3.16	0.64	5.28

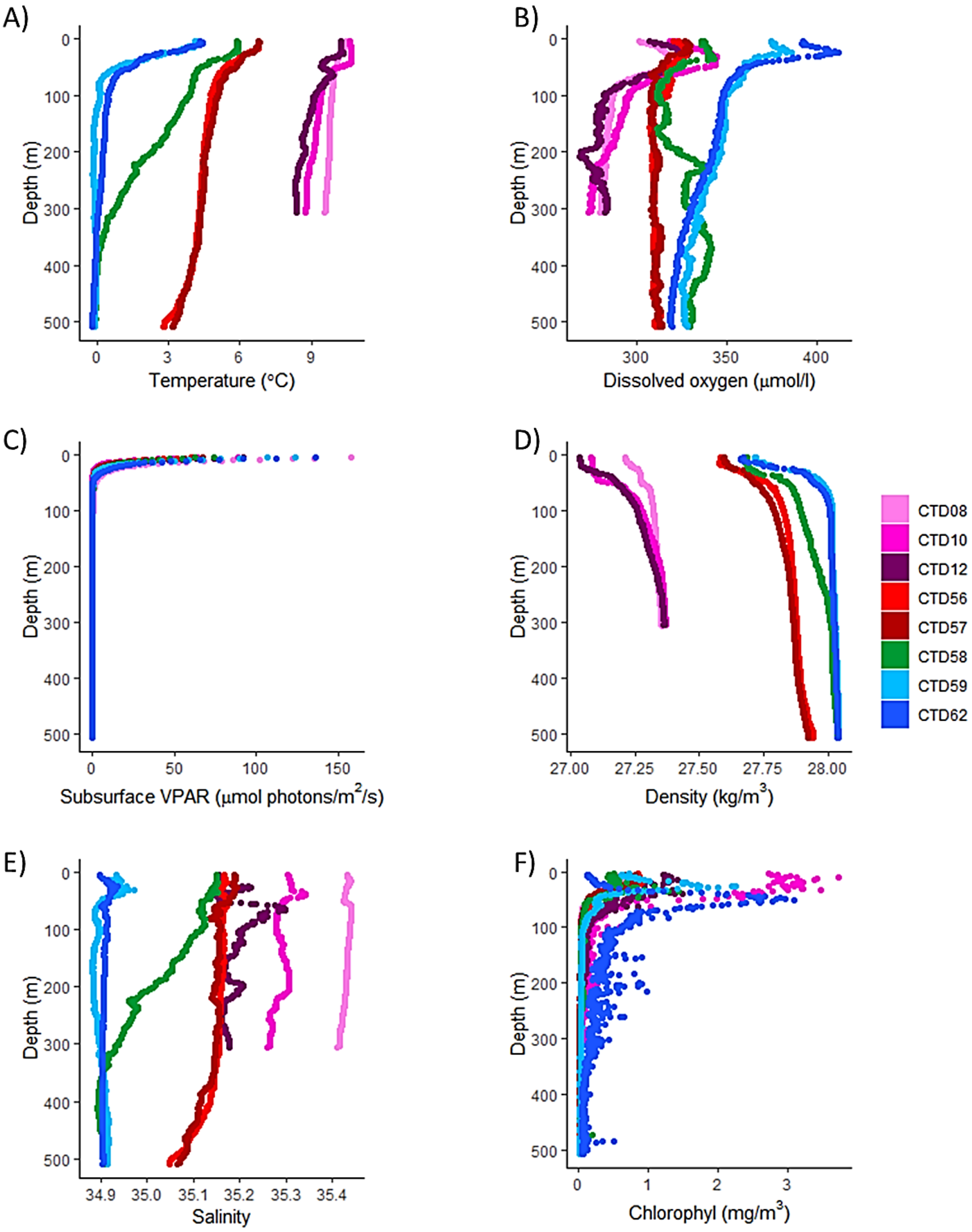


Figure 3.4. Vertical profiles of the environmental conditions at each station measured during CTD casts.

Shown are the vertical profiles for environmental conditions, including chlorophyll, taken at 1 meter intervals from the sea surface during the sample period when environmental water samples were collected for sequence analysis. Water temperature (A), dissolved oxygen (B), subsurface VPAR (C), density anomaly (D), salinity (E) and chlorophyll fluorescence (F). Stations are distinguished by colour as shown.

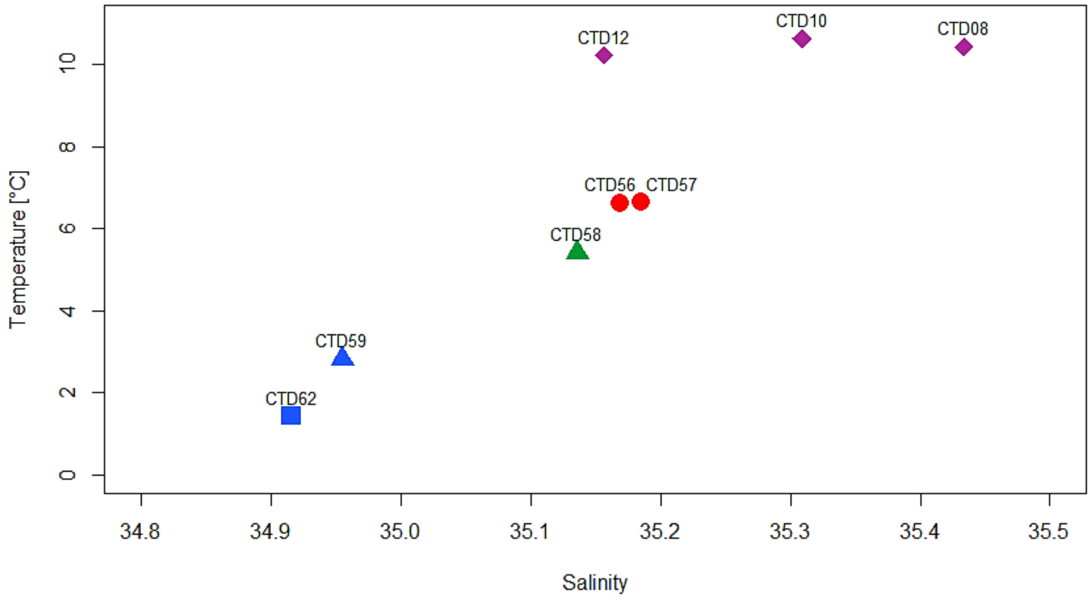


Figure 3.5. Temperature/salinity plot of water conditions at the sampled depth during the sampled period. Measurements were taken at the depth at which water samples were collected for sequence analysis. Stations separate along a gradient of temperature and salinity. Symbols represent previously assigned regional groupings based upon SST maps (Figure 3.3). Symbol colours represent the degree to which Polar Water influences the site determined from *in situ* environmental physical characteristics measured over the sampling period; blue – most highly influenced by Polar Water (HI), green- moderately influenced (MI), red- low influence (LI). Purple diamonds show stations sampled in the North Atlantic.

influence (blue) compared to the other stations, the second was of low influence (red) and third moderate influence (green) (Figure 3.6). The sample at station CTD62 (blue square) was assigned as under high influence (HI) due to featuring reduced temperatures and salinities closest to the <0°C and 34.5 cut offs typically used for defining Polar Waters[200], indicating mixing with Polar Waters (Table 3.1). Stations CTD56 and CTD57 (red triangles) were labelled as under low influence (LI) due to being located in a region SST maps showed experienced little mixing and featured water temperatures >6ºC during the sample period. The sample taken at station CTD58 (green triangle), was observed to be in a periodically mixed region, and featured temperatures and salinities between the HI and LI groups. CTD58 was therefore labelled as under moderate Polar Water influence (MI). SST maps showed that CTD59 (blue triangle) was also found in a region which experienced periods of mixing with Polar Waters (Figure 3.3), but *in situ* environmental data indicated that it was highly influenced (HI) at the time of sampling due to featuring temperature and salinity profiles similar to the HI stations (Figure 3.6, Table 3.1). For purposes of comparison, samples collected at stations CTD08, CTD10 & CTD12 present in the North Atlantic Ocean (purple diamonds) were also included, and featured SST >10ºC.

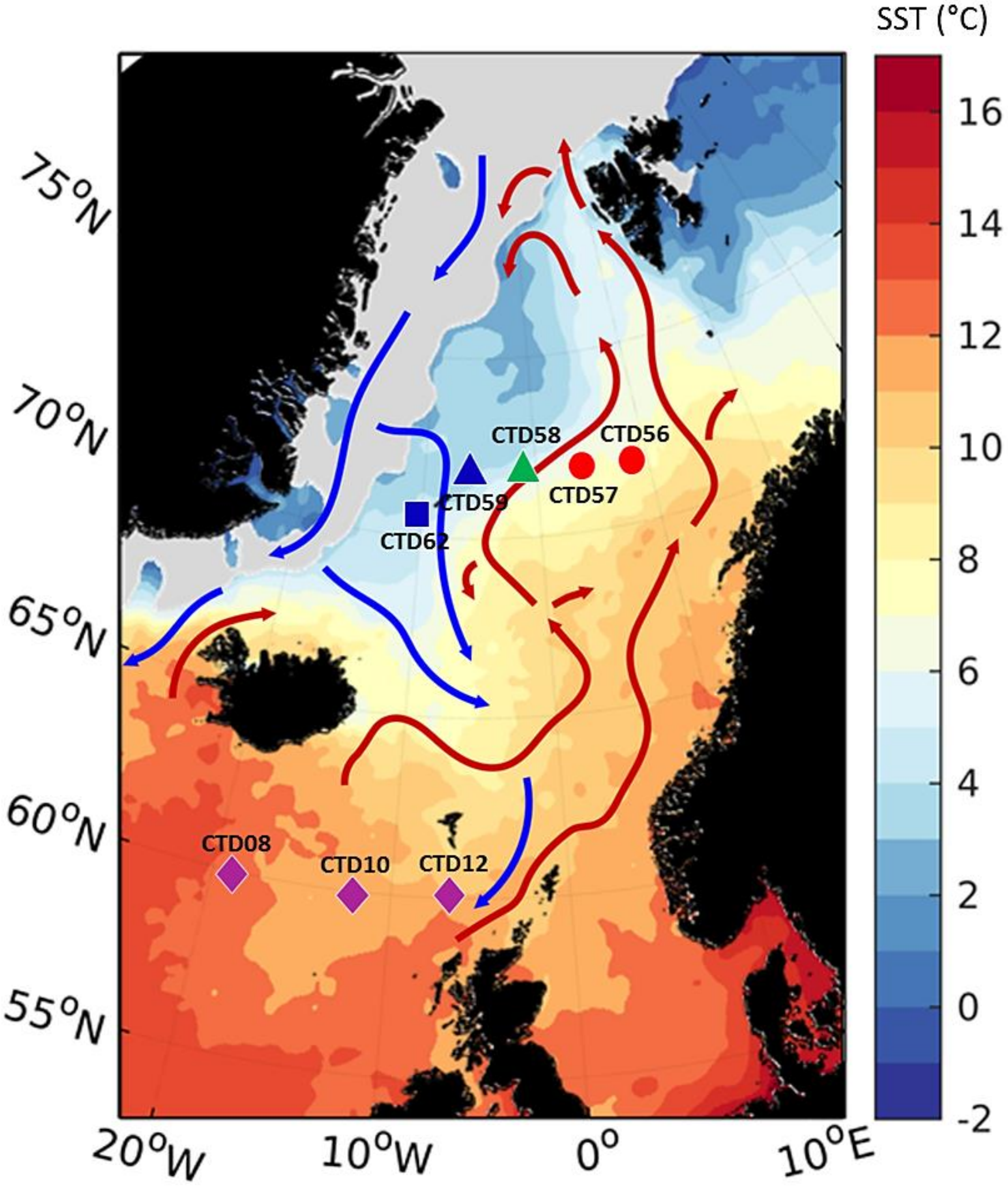


Figure 3.6. Sea surface water temperature maps at the time of sampling with the locations of the stations sampled.

Samples were collected in the Norwegian Sea as part of UK Ocean Acidification research program during cruise JR271 (1st June 2012 to 2nd July 2012). Colour scale represents SST during the sampling period. Descriptions of the currents were provided by Finlo Cottier and are illustrated by coloured arrows, blue- Polar Water currents, red – Atlantic Water currents. Symbols represent the assigned regional group determined from daily SST maps over a 6 month period prior to sampling; square – constant influence from Polar Waters, triangle – intermittent periods of Polar Water influence, circle – little Polar Water influence. Symbol colours represent the degree to which Polar Water influences the site determined from *in situ* environmental physical characteristics measured over the sampling period; blue – most highly influenced by Polar Water (HI), green- moderately influenced (MI), red- low influence (LI). Purple diamonds show stations sampled in the North Atlantic.

The contribution of Polar Water to each station was calculated (Figure 3.7), wherein Polar Water was defined as featuring a salinity ≤34.5[200]. Stations CTD62 and CTD59 featured the greatest influence from Polar Waters, being 81.7% and 75.6% respectively. CTD56 and CTD57 were primarily Atlantic in character, featuring 39.3% and 41.8% contributions of Polar Water, respectively. CTD58 was calculated to feature near half Polar Water (47.0%). From solely salinity data, CTD12 appeared to feature a large amount of Polar Water influence (43.7%) which was primarily the result of fresh water input from Polar currents (Figure 3.6). Indeed, vertical salinity profiles do appear to show a subsurface layer of fresher water (Figure 3.4). However, vertical profiles of the remaining environmental characteristics at CTD12 were most similar to the North Atlantic stations CTD08 and CTD10, and water temperature was >10ºC, implying it is a temperate station.

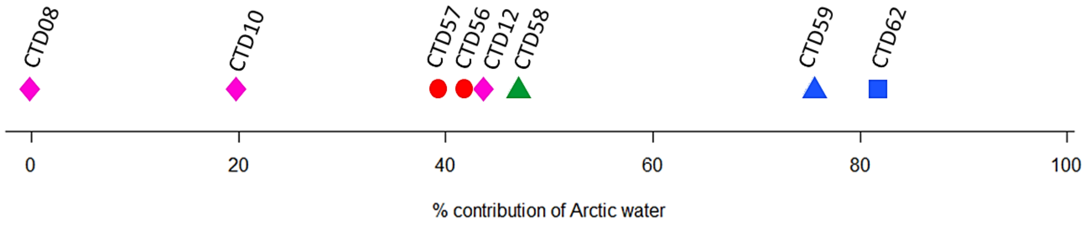


Figure 3.7. The calculated relative contribution of Arctic water to each station.

Shown is the calculated percentage contribution of Polar Water to each station. Polar end members were defined as featuring a salinity ≤34.5. Atlantic end members were defined as the most saline observed measurement taken during CTD casts, 35.4 from CTD08. Symbol colours represent the degree to which Polar Water influences the site determined from *in situ* environmental physical characteristics measured over the sampling period; blue – most highly influenced by Polar Water (HI), green- moderately influenced (MI), red- low influence (LI). Purple diamonds show stations sampled in the North Atlantic.

3.3 Discussion

The Arctic Ocean has been demonstrated to be warming considerably over recent decades, reaching levels of up to 0.8 °C per decade[182] in some regions. The western side of the Arctic that is fed by Atlantic Waters is suggested to be experiencing the most significant increase, with a calculated heat input of 39 ± 9 MJ m^{-2} yr^{-1} (1.2 ± 0.3 W m^{-2})[199]. The warming of this western region is concerning as it represents one of the gateways to the Arctic Ocean, and has historically been divided into two separate ecosystems defined by distinct climate regimes (Figure 3.8). The first is one of cold, fresh Polar Water typically associated with an Arctic domain, and the other is composed of warmer saline water belonging to the Atlantic domain. Similar distinct

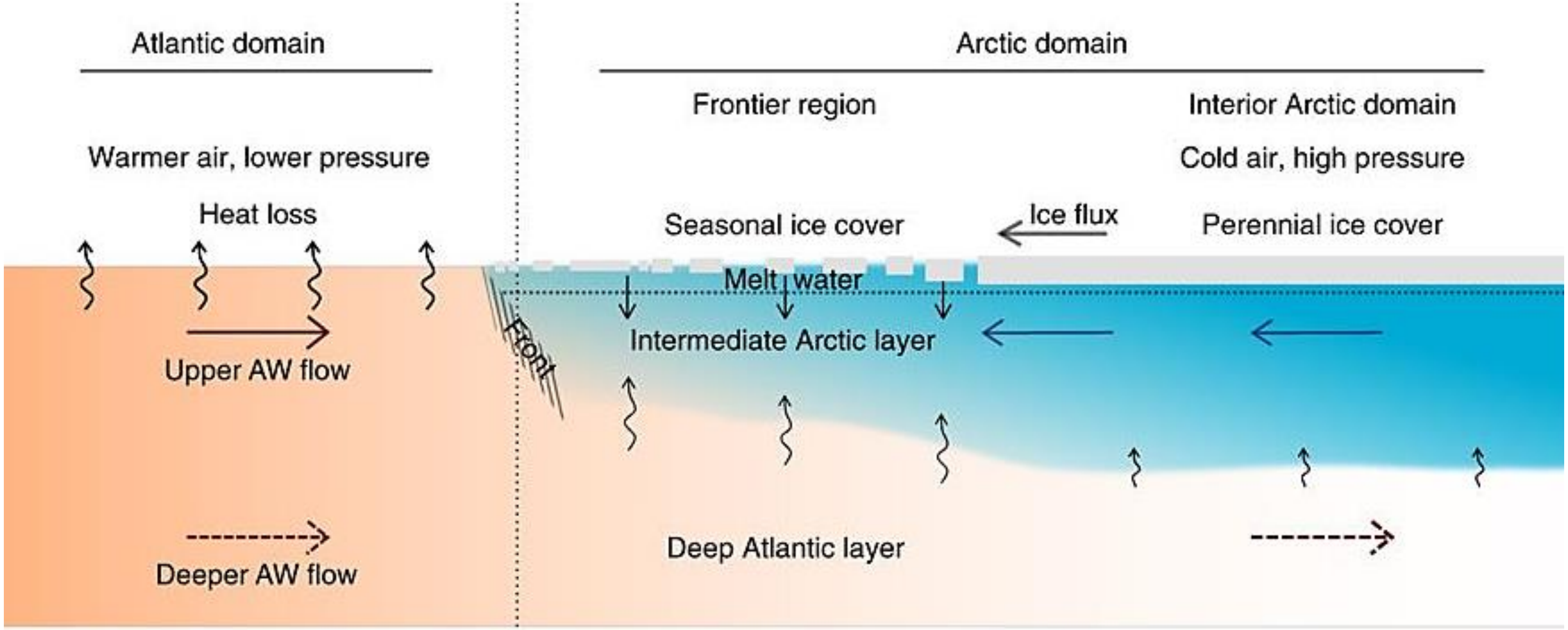


Figure 3.8. Schematic view of the transition from Atlantic to Arctic associated waters.

Warm, saline Atlantic Water is shown on the left (orange) occupying the entire water column and transitions into colder, fresher, stratified waters in the Arctic domain consisting of cold fresh Arctic Water (blue) over a deep Atlantic layer. Horizontal arrows denote transports and vertical arrows denote vertical fluxes, arrow size indicates the size of fluxes. The Atlantic domain features significant heat loss to the atmosphere over winter, whereas the Arctic experiences upward fluxes of warm saline water from the underlying Atlantic layer, which is greatest in the frontier region between both domains. Sea ice extent (grey) covers much of the interior Arctic domain and provides an important input of cool, fresher water into the frontier region to maintain stratification, and resist upwelling from deep Atlantic Waters. Figure adapted from[199].

climate regimes were resolved in Figure 3.3, with the variability in front position observed over the sampled period likely the result of annual cycles, confirming the sampled region covers a boundary between Arctic and Atlantic climate domains.

Within this boundary region between Atlantic and Arctic domains, the maintenance of Polar Water conditions within the Arctic domain is dependent upon the import of sea ice from further north to provide fresh, cool water to the water column, the melting of which supports the stratification of Arctic Water over the top of deeper Atlantic Water[201]. The conditions sampled at the HI stations (Table 3.1) suggest that these stations were reflective of stratified Polar Waters associated with typical Arctic domain conditions. Vertical profiles of temperature and salinity further support this conclusion (Figure 3.4), demonstrating the presence of cooler, fresher waters (Figure 3.5), close to the <0°C and 34.5 cut offs typically used for classifying Polar waters[200]. In comparison the LI stations were observed to feature warmer, more saline waters typically Atlantic in character, supporting that the LI stations were located in the Atlantic domain. These environmental differences enabled the stations to be classified into distinct regional groups (Figure 3.6), agreeing with descriptions in the literature of distinct regional ecosystem domains (Figure 3.8)[199].

It is suggested that the presence of sea ice within the Arctic domain is required to maintain the stratification of Polar Waters. However, the levels of sea ice are rapidly declining at a rate of up to 17.5% per decade[193] and total sea ice expanse has fallen by 66% over the past 6 decades[189]. These changes are weakening the stratification of Polar Waters, affecting their ability to resist the upwelling of warm saline Atlantic water, leading to the reported warming[199] and an increase in the expanse of Atlantic Water within the region[199], termed “Atlantification” of the Arctic. Through such Atlantification it is suggested that Arctic region will transition into a new steady state, with altered hydrography, including reduced nutrient supply, and altered biogeochemistry[198]. Despite these suggestions quantification and details of specific changes remain largely unknown. A complete transition of the entire region to an Atlantic-like domain has occurred in the past, as inferred from paleoclimate records[202], and another transition is currently underway as sea-loss is set to continue, with a completely sea-ice free Arctic predicted to occur by the middle of the 21st century[193]. Such a transition would have significant ecological consequences. Indeed, ecosystem impacts have already been documented including up to a 30% increase of annual primary production[194], increased zooplankton productivity, changes to phytoplankton size classes[203], and resultant northward migration of predatory fish[204] within regions typically considered as Arctic Waters. Other impacts include the northward range shift of a number of taxa following Atlantic Waters featuring suitable conditions[106,205], and the intrusion of Atlantic phytoplankton species into the Arctic region[206]. The alterations to the oceanic boundaries between Atlantic and Arctic waters, and the resultant range extensions and intrusion of novel taxa, is therefore introducing new competition between the extant microbial communities within these regions. However, it is unclear how the composition of communties in these regions will respond, and what the ecological consequences of such mixing will be.

3.4 Conclusion

The sampled region was resolved to contain four station groups (including those sampled in the North Atlantic) that featured distinct environmental conditions. Stations covered both temperate waters typically Atlantic in character, and Polar Waters, that matched descriptions in current literature. Each station group was delineated by different amounts of Polar Water influence (Figure 3.7) acting to alter local water column salinity and temperature. These groups are hereafter referred to as LI (low influence), MI (moderate influence) and HI (high influence) in further analyses. The sampled transect of stations, which includes the MI station placed within the boundary between both Atlantic and Polar regions (Figure3.6) where the water profiles were calculated to feature nearly half (47%) Polar Water, presents an opportunity to gain an insight into which community members will be "winners or loosers" under the new pressures of mixing between Atlantic and Polar associated taxa. The following chapters will attempt to address this knowledge gap.

.

Chapter 4

Bacterial Community Gradients Along a Transect in the Norwegian Sea, in the Context of Climate Change

4.1 Introduction

Bacterial communities underpin a number of vital biogeochemical and biological cycles such as carbon[42], nitrogen and phosphorous[120] by assimilating and transforming organic matter utilising hydrolytic enzymes[207]. Therefore, understanding the effects of environmental change, including associated alterations to temperature, stratification and nutrient content/composition upon these communities is crucial to be able to extrapolate and predict the resultant impacts to bacterial communities and the ecosystem services they provide.

Unlike for Eukaryotes, it is not possible to use conventional light microscopy to visually taxonomically detail bacterial communities. As a result, it has only been with the advent and advancement of molecular biological techniques that it has been possible to study the biogeography and distribution of marine bacterioplankton[208]. It is now widely accepted that as a domain bacteria are globally distributed throughout all terrestrial and marine habitats. Global sampling studies seem to suggest that the majority of the high level taxonomic groups of bacterioplankton also display cosmopolitan distribution, and there is wide spread evidence for a cosmopolitan distribution at the class level within members of the Cyanobacteria, Flavobacteriia, Betaproteobacteria and Actinobacteria[209]. Global surveys, such as the Tara Ocean expedition, which attempt to catalogue the global microbiome also suggest these classes are ubiquitous and that Alphaproteobacteria are the dominant taxonomic group within marine bacterioplankton, followed by Cyanobacteria or Gammaproteobacteria (depending on location)[42]. However, the majority of current studies of marine bacterial diversity are largely based on sampling from temperate locations, meaning those relating to Arctic water masses are lacking[210]. Given the unique environmental conditions of the Arctic environment, a key question is raised as to whether Arctic bacterial diversity is different to that of temperate waters. Indeed, there are examples of bacterial communities showing distinct associations with environmental factors in cold water marine habitats[211]. Possibly the most significant distinction between open water bacterial communities in Arctic Waters compared to those of temperate waters is the lack of the photoheterotrophic cyanobacteria *Synechococcus* and *Prochlorococcus*[68]. These cyanobacteria are globally ubiquitous, and are vital players in global carbon fixation being responsible for a predicted 25% of all globally primary production, but are excluded from Arctic waters due to temperature boundaries exceeding their physiological limits[112]. Instead their functional role is fulfilled by picoeukaryotes[212].

Early studies of Arctic bacterial communities illustrated an unexpectedly high level of diversity and recovered many taxa which appeared to differ from those isolated from other marine environments[129,208]. The majority of these early studies primarily investigated sea-ice communities, and demonstrated endemism or bipolar distributions between the Arctic and Antarctic in selected taxa such as *Shewanella frigidimara*[213]. Representatives of the Flavobacteriia, including *Polaribacter, Psychroflexus*, *Flavobacterium,* and the Gammaproteobacteria *Glaciecola* and *Colwellia* have been observed to dominate sea-ice bacterial communities[110]. These taxa feature highly efficient pathways for the degradation of organic matter, including polyunsaturated fatty acids, and thus form an important functional component of the Arctic food web[110]. Additionally, representatives of Betaproteobacteria display high abundances in later summer sea ice communities when sea-ice melt freshens surrounding waters due to being highly competitive within the low nutrient conditions of sea ice[213]. The presence of high quantities of perennial sea ice has led to gas vacuolated bacteria such as *Polaribacter irgensii* forming important components of the Arctic microbiome, but are scarce in other locations[208].

Following the work studying sea-ice associated communities greater focus is now being given to open water bacterial communities, yet they remain poorly understood[129]. Within these communities Alphaproteobacteria are observed to typically dominate Arctic waters[129], as they do in temperate marine systems globally[42]. Common representatives of Alphaproteobacteria include SAR406, SAR324 and *Nitrospina*[208], that are known to contribute to sulphur oxidation, carbon fixation, short-chain fatty acids metabolism and nitrite reduction[214–216]. Gammaproteobacteria are also found to be highly abundant in Arctic Waters at similar proportions to the Alphaproteobacteria[132], with representative taxa including *Thiomicrospira, Oceanospirillum, Alteromonas, and Pseudomonas*[208]. Additionally, there have been some reports of Bacteroidetes being dominant across the Greenland Sea[39]. Verrucomicrobia and Actinobacteria also commonly appear within the abundant fraction of Arctic bacterial communities but feature comparatively low overall abundances (1-3% total community abundance)[217]. During molecular surveys of Arctic bacterial communities a high number of novel taxa with no prior known representatives are frequently reported highlighting the extent of marine bacterial diversity still to be uncovered. A large proportion of these novel taxa have been shown to be most closely related to members of the Alphaproteobacteria SAR11 group, which is found to be ubiquitous in the global Ocean surface layer[208].

Annual patterns of seasonal variation have been demonstrated within bacterial communities[133], including members of the Alphaproteobacteria, Betaproteobacteria, Flavobacteriia and Acidimicrobiia, which manifest as changes in abundance[218]. Within the Arctic summer period Gammaproteobacteria and Flavobacteriia dominate surface communities[219], including *Polaribacter spp.*[208]*.* Alphaproteobacteria, such as *Roseobacter spp.,* Cytophaga and Deltaproteobacteria have also been observed to display the highest annual abundances during the summer months[208]. However, other studies suggest an absence of Deltaproteobacteria over summer[219]. These discrepancies may be the result of upwelling events or differences in sample depth as Deltaproteobacteria are more typically associated with deep waters[70,207], possibly due to being competitive under conditions of low light availability[219]. The seasonal fluctuations of Alphaproteobacteria, Gammaproteobacteria and Flavobacteriia groups described have been shown to primarily be driven by changes in water currents and temperature[218,219]. Additionally, there is evidence these bacterial community dynamics are linked to strong seasonal solar cycles, matching similar responses by eukaryotic phytoplankton which show peak growth over the summer months and implying biological links exist between certain members of both bacterial and eukaryotic communities[219,220].

Over winter, bacterial communities have been demonstrated to become more phylogenetically and functionally diverse, with a transition to a greater reliance on chemolithoautrophy[207]. Total abundances of Gammaproteobacteria and Flavobacteriia have been shown to reduce, yet they remain the dominant taxonomic groups, and compositional shifts are observed including the Gammaproteobacteria *Alteromonas* and *Oceanospirillum* displaying peak abundance over the winter period[219]. Deltaproteobacteria, such as nitrite oxidising *Nitrospinaceae,* also become more abundant during the polar winter, while within the Alphaproteobacteria taxa such as *Rhodospirillaceae, Rhodobacteraceae* and SAR202 become most abundant at this time of year[219].

A key question is raised as whether the presence of taxa adapted to the harsh Arctic habitat places the Arctic ecosystem at risk from climate driven environmental forcing, and to what degree it is susceptible. Different taxonomic groups of bacteria are likely to be affected by environmental perturbations in different ways, and by different amounts. It is suggested that the decreasing levels of ice-cover[183] is likely to lead to enhanced bacterial production through increased mixing and transfer of nutrients into the upper water column[129]. The thinning of ice, and move towards a higher seasonal to perennial sea ice ratio[192] will favour under-ice and sea-ice associated phytoplankton blooms, phototrophic bacterial taxa and those which show position associations to bloom forming phytoplankton[221]. Although, if enhanced melt water increases stratification, primary production may in fact be negatively affected, but could benefit halotolerant freshwater taxa affecting their range and distribution[129]. Despite these theories real world data is sparse, and the degree of adaptability of existing taxa is unknown. It has been presumed that some bacterial taxa, especially those which appear to be rare in any particular location, may be able to respond to environmental changes through changes in their abundance[143]. Rare taxa are believed to represent a large pool of genetic diversity which acts as a buffer to environmental disturbance, acting to provide functional redundancy to protect ecosystem services[222]. Additionally, it has been shown that some highly prevalent taxa also show a high degree of genetic diversity at highly resolved inter-OTU levels, and thus are able to maintain highly abundant in response to changes in environmental stressors[58]. However, the Arctic appears to possess unique taxa, and have been shown to feature low levels of genetic diversity compared to other locations[211], and thus may not follow conventional bacterial diversity patterns, instead featuring reduced functional redundancy and increasing the susceptibility of these communities to climate change.

This chapter details the application of Illumina MiSeq technology to analyse the diversity of marine bacterial communities across a transect of five stations in the Norwegian Sea, which were resolved to cover a gradient of Polar Water influence (Chapter 3). This enabled assessment of the susceptibility of these communities to increased Atlantic influence within the Arctic by examining community partitioning and correlations to environmental factors. The findings presented in this chapter have significant implications for the vulnerability of polar associated bacterial community assemblages under the current climate change trends.

4.2 Results

4.2.1 Sequence Analysis

3,741,702 raw paired-end sequences were recovered from the Illumina MiSeq sequencing effort (Table 4.1). The number of raw paired-end sequences recovered from each individual station varied from 143,986 (CTD57) to 242,780 (CTD12) with an average of 233,856 sequences per station. During pre-processing quality control 14,249 (0.4%) sequences were lost, leaving a total of 1,856,602 merged pair-end sequences for taxonomic assignment by BLAST. The SILVA database was chosen for BLAST analysis as it represents a quality checked and curated sequence repository which provides accurate and standardised up to date taxonomic information, and has become widely utilised in high-throughput analyses[164,223]. After BLAST 24,202 sequences were unassigned, featuring no matches to known sequences in the SILVA database. 39,182 sequences, representing 48.2% of the dataset, were identified as belonging to singletons, defined as OTUs represented by only one sequence across the entire sampling effort and thus were removed from the analysis due to the risk of being the result of erroneous base calls during sequencing. 871,112 sequences were found to match to chloroplasts and 23,767 to mitochondria. These are potentially the result of eukaryotic contaminants and so were removed from further analysis. 922,541 merged pair-end sequences were retained after pre-processing, representing 49.7% of the initial raw sequencing effort. These varied by station from 38,088 (CTD12) to 279,869 (CTD59) with an average of 115,318 sequences per station.

4.2.2 Rarefaction Analysis

An estimated 74.7% ±0.8 saturation was achieved. Rarefaction curves did not reach full saturation but did begin to approach a plateau for the full dataset (Figure 4.1A), each station group (Figure 4.1B), and each individual station (Figure 4.1C). Extrapolated OTU richness curves (dashed lines) also indicated under saturation.

Table 4.1. Sequence data metrics for the bacterial dataset.

Shown are the raw Illumina sequence file names generated by the Illumina MiSeq pair-end sequencing effort. Files are named by representative CTD sample with 16S donating the target gene during the probe assay, R1 donating files generated from the forward (5'-3) sequence run, and R2 donating those generated by the reverse (3'-5') sequence run. R1 and R2 were merged as part of the processing pipeline. Information for processing stages after the merger is stated in the R1 filename row. Raw sequence number shows the total number of sequences as returned by the Illumina MiSeq sequencing run before any processing. Sequences remaining after quality control gives the final sequence number for each sample file after merger of the R1 and R2 sequence runs and completion of the quality control processing stage of bioinformatics pipeline (Section 2.4.5) to improve sequence quality and that were used for analysis. Also shown are the number and percentage of initial sequence number lost during this step, and the minimum, maximum and average sequence lengths after pre-processing. Unassigned sequences lists the number of sequences for each sample file in the bacterial dataset for which no taxonomic annotation could be found, these are concatenated from annotations of "none" and "no blast hit". Number of singletons lists the number of sequences representing singletons for each sample. The number of sequences identified as matching chloroplasts and those matching mitochondrial sequences are also shown.

File name	Raw sequence number	Sequences remaining after quality control	Number of sequences lost	Percent of sequences lost (%)	Minimum sequence length (bp)	Maximum sequence length (bp)	Average sequence length (bp)	Unassigned sequences	Number of singletons	Sequences matching chloroplasts	Sequences matching mitochondria
CTD08-16S_R1	174,441	172,860	1,581	0.91	112	250	250	1,684	37,136	37,136	2,160
CTD08-16S_R2	174,441										
CTD10-16S_R1	301,031	299,083	1,948	0.65	101	250	250	2,209	187,640	187,640	1,183
CTD10-16S_R2	301,031										
CTD12-16S_R1	121,390	120,736	654	0.54	127	250	250	1,156	78,280	78,280	1,806
CTD12-16S_R2	121,390										
CTD56-16S_R1	223,377	221,597	1,780	0.8	103	250	250	4,225	114,188	114,188	6,012
CTD56-16S_R2	223,377										
CTD57-16S_R1	71,993	71,467	526	0.73	105	250	250	1,259	27,038	27,038	402
CTD57-16S_R2	71,993										
CTD58-16S_R1	131,929	130,863	1,066	0.81	125	250	250	4,356	32,341	32,341	894
CTD58-16S_R2	131,929										
CTD59-16S_R1	606,210	601,723	4,487	0.74	103	250	250	8,242	300,447	300,447	9,705
CTD59-16S_R2	606,210										
CTD62-16S_R1	240,480	238,273	2,207	0.92	100	250	250	1,071	94,042	94,042	1,605
CTD62-16S_R2	240,480										

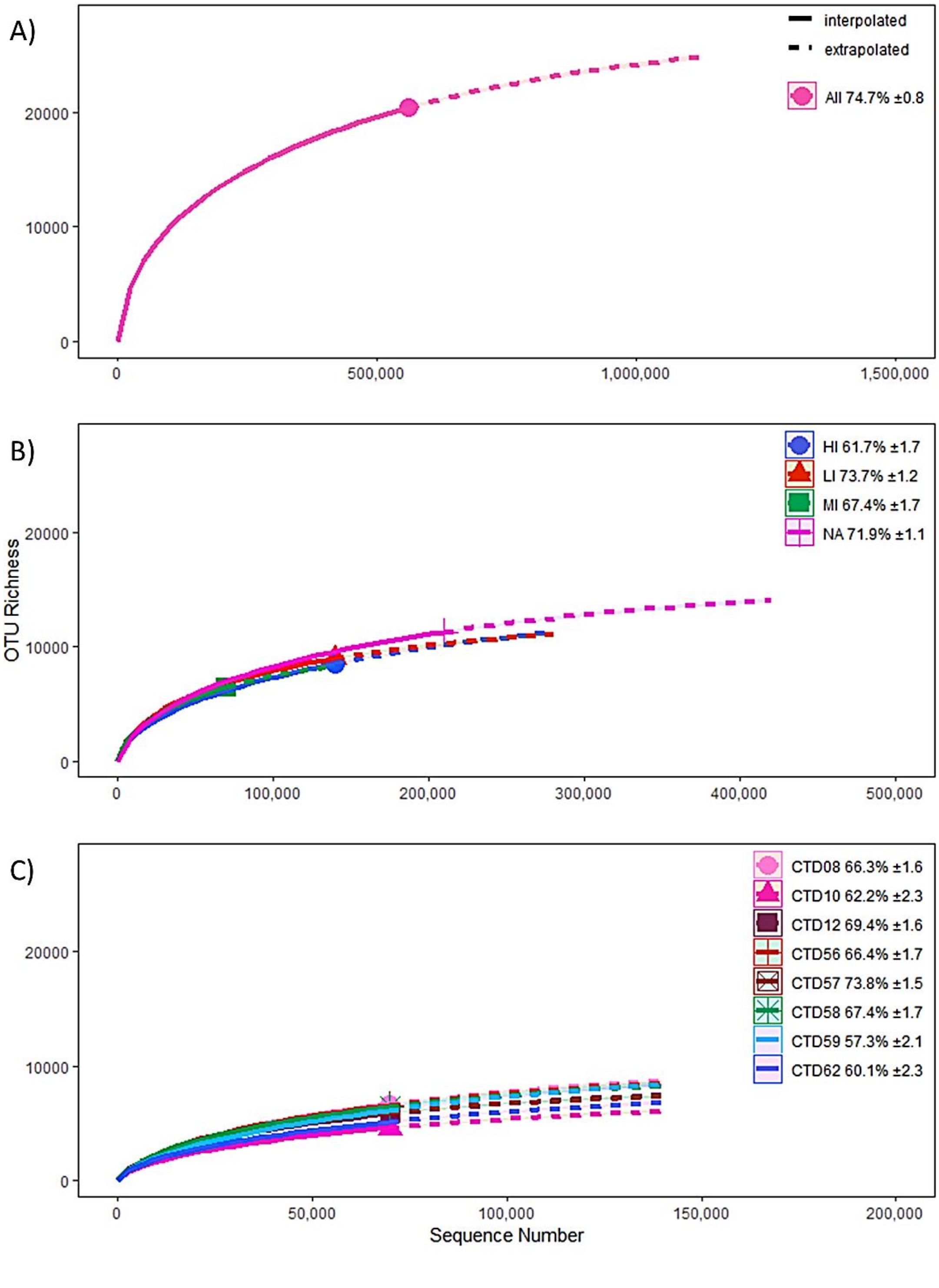


Figure 4.1. Rarefaction analysis of bacterial sequence depth.

Rarefaction (solid line) and extrapolation (dashed line) of OTU richness for the whole rarefied bacterial dataset (A), each station group (B) and each individual stations (C). Legend displays each group for which rarefaction was calculated, the % saturation of the recovered OTU number of the total number of estimated OTU number, and standard error for each group. Stations are labelled by the extent of Polar Water influence; blue – high Polar Water influenced stations (HI), green – moderately Polar Water influenced stations (MI), red – little Polar Water influence stations (LI), purple – Stations sampled in the North Atlantic.

4.2.3 α-diversity

A total of 10,272 OTUs were recovered within the transect from the rarefied bacterial dataset. Of these 687 were deemed to be novel, with no prior representation in the SILVA database (release 128). Within the transect the highest number of OTUs was recovered from station CTD56 (4,556), and the lowest from station CTD62, being 3,630 (Table 4.2).

Analysis of the α-diversity metrics generated for the bacterial dataset revealed that the Shannon diversity index, a common metric used to characterise species diversity in a sample taking into account both species abundance and evenness[224], was highest at CTD56 (Table 4.2), and lower at both HI stations, CTD59 and CTD62. The identification of a lower Shannon diversity index at the HI stations implies that the bacterial communities present at these stations were composed of only a few highly abundant taxa, with the majority of taxa being rare. In contrast, a higher Shannon index observed at CTD56 implies a more even distribution of constituent OTUs, with a higher number of more abundant taxa. The estimated OTU richness output by the "iNEXT" R package[169] and ACE (abundance-based coverage estimator) are metrics used to estimate the total OTU richness present in a sample based upon the observed number of OTUs and taking into account the number of rare OTUs[169,225]. The output of these was lowest for CTD57 and highest for CTD59, contrasting with the observed OTU richness.

Table 4.2. α-diversity of the bacterial communities recovered from each station.

Shown is the number of observed and the calculated estimated number of OTUs recovered from each station, including standard error. Also shown are the Shannon diversity and Simpson diversity metrics. Station groups reflect those determined during the regional assignment of stations in Chapter 3; HI – high Polar Water influenced stations, MI – moderately Polar Water influenced stations, LI – little Polar Water influence stations, NA – Stations sampled in the North Atlantic.

Diversity metric	CTD56	CTD57	CTD58	CTD59	CTD62	CTD08	CTD10	CTD12
Station group	LI	LI	MI	HI	HI	NA	NA	NA
Observed OTU richness	4,556	4,094	4,482	4,324	3,630	4,279	2,979	3,430
Estimated OTU richness (iNEXT)	6,684	5,750	7,095	7,708	5,772	6,856	4,433	4,647
iNEXT estimated S.E	167.72	117.37	159.90	227.26	193.63	165.99	167.54	137.03
Estimated OTU richness (ACE)	6,949	5,969	7,449	8,025	5,991	7,218	4,458	4,940
Shannon diversity	464.2	384.4	433.2	308.2	248.4	359.3	225.5	267.0
Simpson diversity	83.3	68.5	65.8	39.4	41.4	57.8	48.5	45.7

4.2.4 β-diversity

The partitioning of the community at the MI station, resolved to feature a near equal mix of Atlantic and Arctic waters, in relation to the LI and HI communities reveals which community will become dominant under increased Atlantic Water influence within the region. β-diversity analysis revealed the whole bacterial community structure across the transect partitioned into two main clusters, which were ~60% dissimilar to one another (Figure 4.2A). The left most cluster comprised the LI stations CTD56 and CTD57 which were ~35% dissimilar, as well the MI station CTD58 which was ~40% dissimilar to the LI stations. The second cluster comprised the HI stations (CTD59 and CTD62) which were ~40% dissimilar to each other, and ~65% dissimilar to the LI group. The arrangement of the clustering follows the hydrography described in Chapter 3, with the communities at the two most environmentally dissimilar stations (CTD56 and CTD62) being most dissimilar. Similar clustering was observed for the most abundant members of the community (OTUs ≥1% of the community) (Figure 4.2C), those at intermediate abundances (OTUs 0.01-1%) (Figure 4.2D) and those at rare abundances (<0.01%) (Figure 4.2E). The partitioning of the MI station with the LI stations suggests that temperate water associated taxa are more competitive under mixed conditions and may displace cold water associated taxa under increased Atlantic Water influence in the region. Validation of this clustering by the addition of the stations CTD08, CTD10 and CTD12 sampled from the North Atlantic maintained the existing clusters, with the North Atlantic stations most similar to the LI stations, and most dissimilar to the HI stations (Figure 4.2B).

Principle coordinate analysis (PCoA) revealed a total of 79.2% of the variance in the bacterial community could be explained by environmental variables (Figure 4.3). 63.6% was explained by PC1 and 15.6% by PC2. Pearson correlation of *in situ* environmental data collected at each station with PC1 and PC2 revealed temperature and salinity as the greatest explanatory factors (Table 4.3) (p<0.01), silicate was also deemed a significant factor (p<0.05). No factors significantly correlated with PC2.

Analysis of the representative major bacterial taxonomic groups revealed that the Betaproteobacteria, Acidimicrobiia, Alphaproteobacteria, Deltaproteobacteria, Gammaproteobacteria, Flavobacteriia, Verricomicrobia and Firmicutes displayed strong regional partitioning (similarity of 56-73% for the LI and 4-72% for the HI groups) (Figure 4.4A). For all groups the MI stations were observed to be most similar to the LI stations, suggesting that temperate associated taxa within these groups may displace cold water associated taxa under increased Atlantic Water influence in the

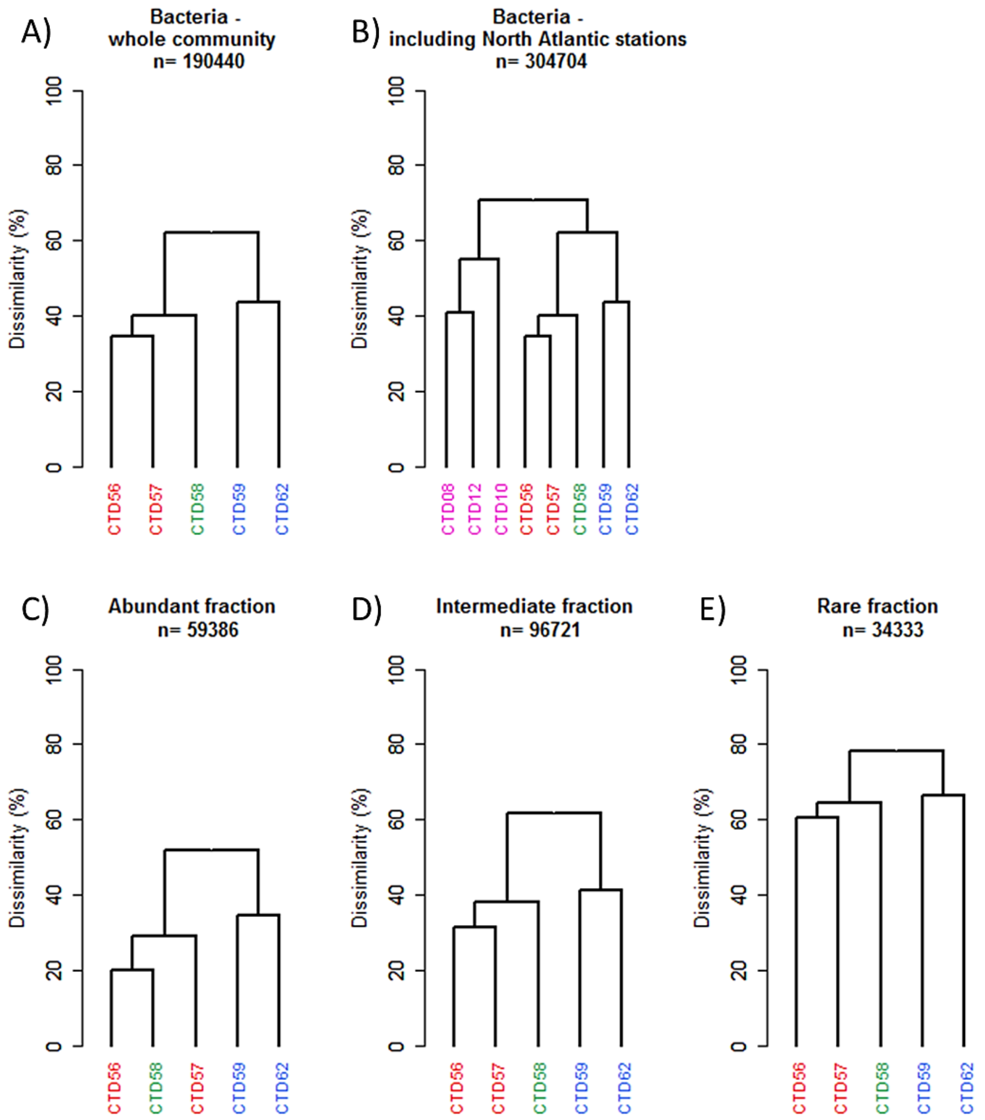


Figure 4.2. Bacterial β-diversity between stations.

β-diversity between stations was calculated by generation of Bray-Curtis dissimilarity matrixes and the degree of dissimilarity displayed by a dendrogram. Shown is the degree of dissimilarity between stations for the whole community (A), and including stations from the North Atlantic (B). Also shown is the degree of dissimilarity for separate abundance fractions of the community, the abundant fraction (OTUs representing ≥1% of the community) (C), the intermediate abundance fraction (0.01-1%) (D), and the rare fraction (≤0.01%) of the community (E). Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI, purple – stations sampled in the North Atlantic.

region. No regional partitioning was observed for the Cytophagia or Epsilonprotobacteria (Figure 4.4B). This may be explained by the low number of sequences present for each of these taxonomic groups, and thus it may be that a greater sequence depth is required in order to provide sufficient information to reveal any ecological patterns. Analysis of the Sphingobacteriia was not possible due to poor OTU recovery.

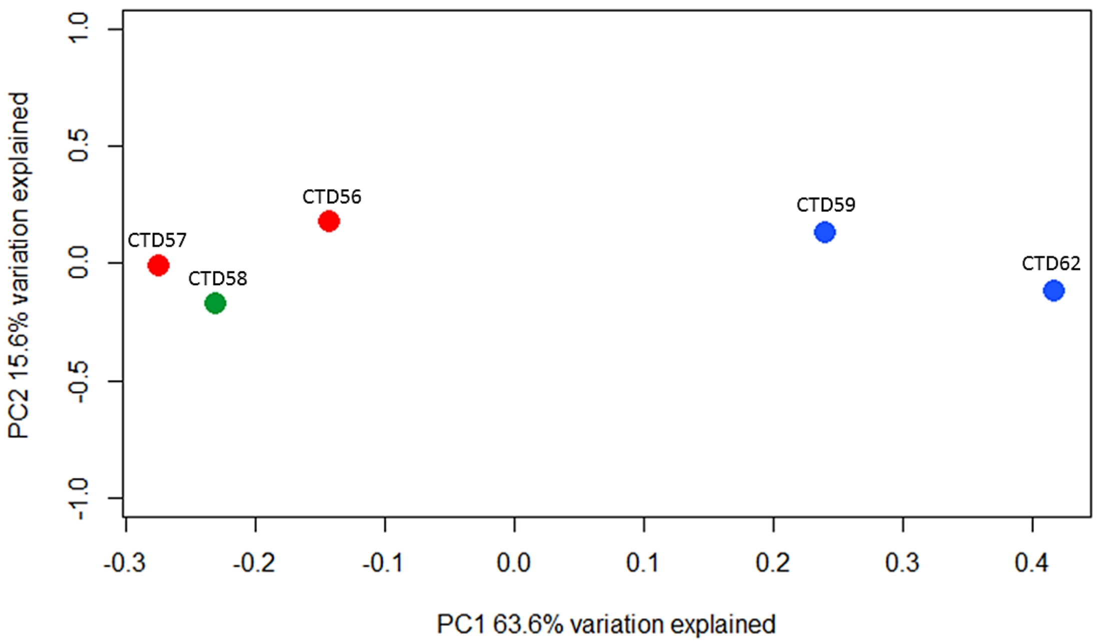


Figure 4.3. Principle coordinate analysis of the Bray-Curtis dissimilarity matrix of bacterial OTUs between stations.

79.2% of the variance in the bacterial community could be explained by environmental variables. Stations primarily separated along PC1, the proportion of variance in the bacterial dataset explained by each coordinate is shown. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

Table 4.3. Pearson correlation of environmental factors to bacterial community variance.

The correlation of each *in situ* environmental factor to the first (PC1) and second (PC2) coordinates of the PCoA is shown, R represents the correlation coefficient. Negative R values indicate negative correlations, and positive R values indicate positive correlations. The significance of each correlation is shown by P, which represents the p-value. * donates *p*≤0.05, ** donates *p*≤0.01.

Environmental factor	**PC1**		**PC2**	
	R	**P**	**R**	**P**
DO_2	0.79	0.11	0.13	0.83
Salinity	-0.96	**0.01	0.14	0.82
SubSurVPAR	-0.52	0.36	0.76	0.13
Temperature	-0.96	**0.01	0.24	0.70
Ammonium	0.68	0.20	-0.55	0.33
Nitrate	0.31	0.62	-0.70	0.19
Silicate	-0.90	*0.04	-0.33	0.59
Phosphate	0.46	0.43	-0.68	0.21

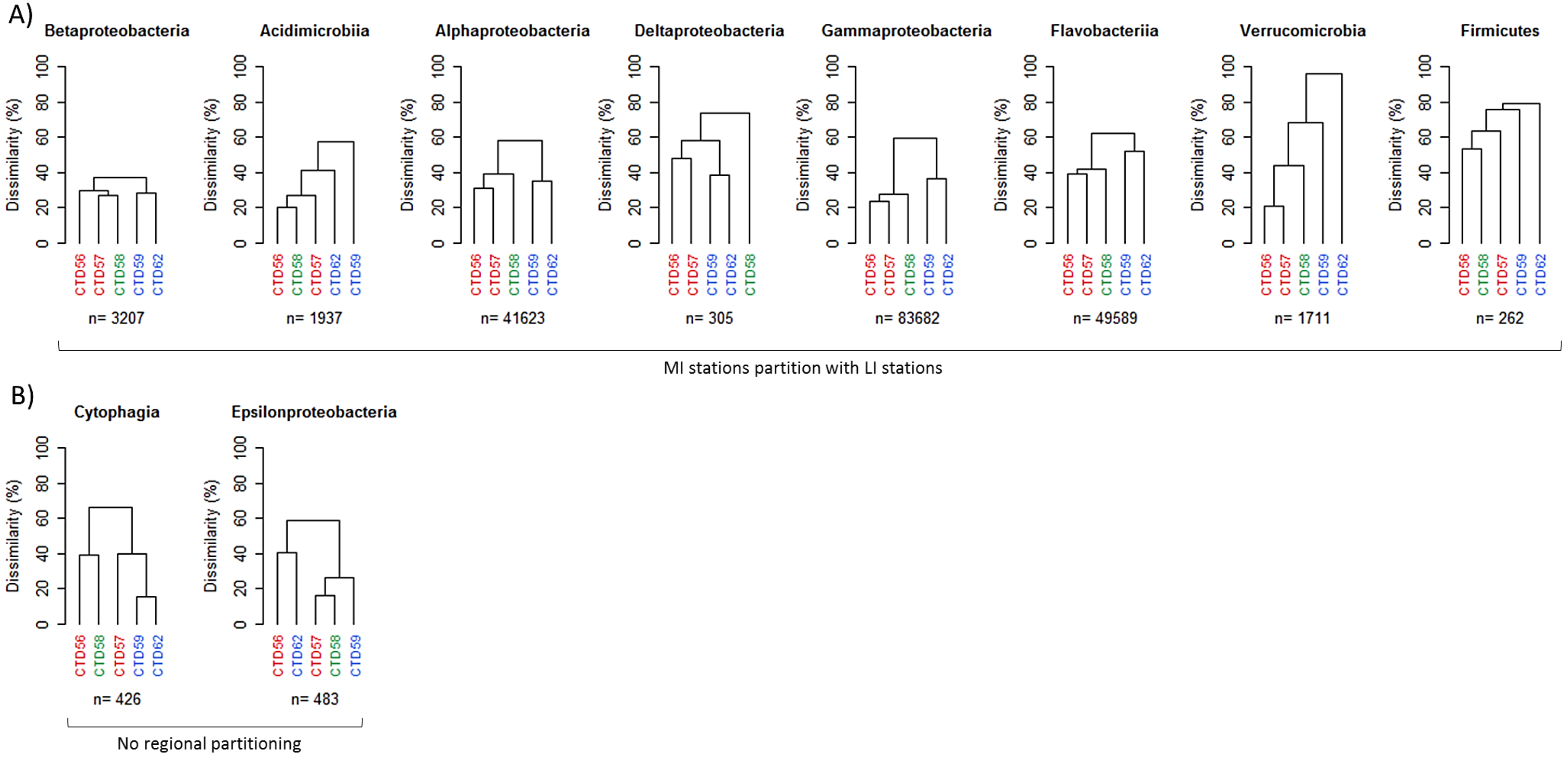
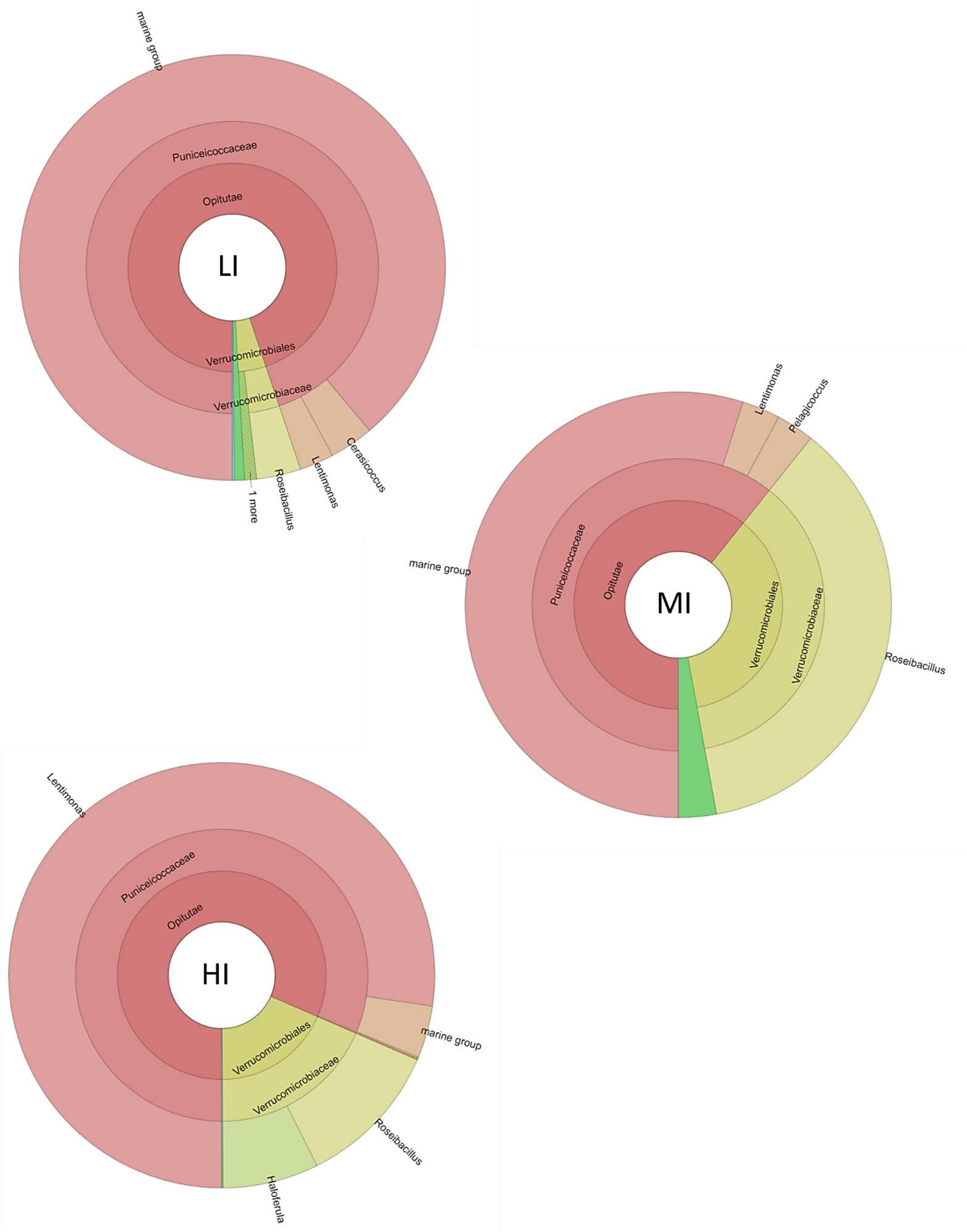


Figure 4.4. Regional partitioning of the representative major bacterial taxonomic groups.

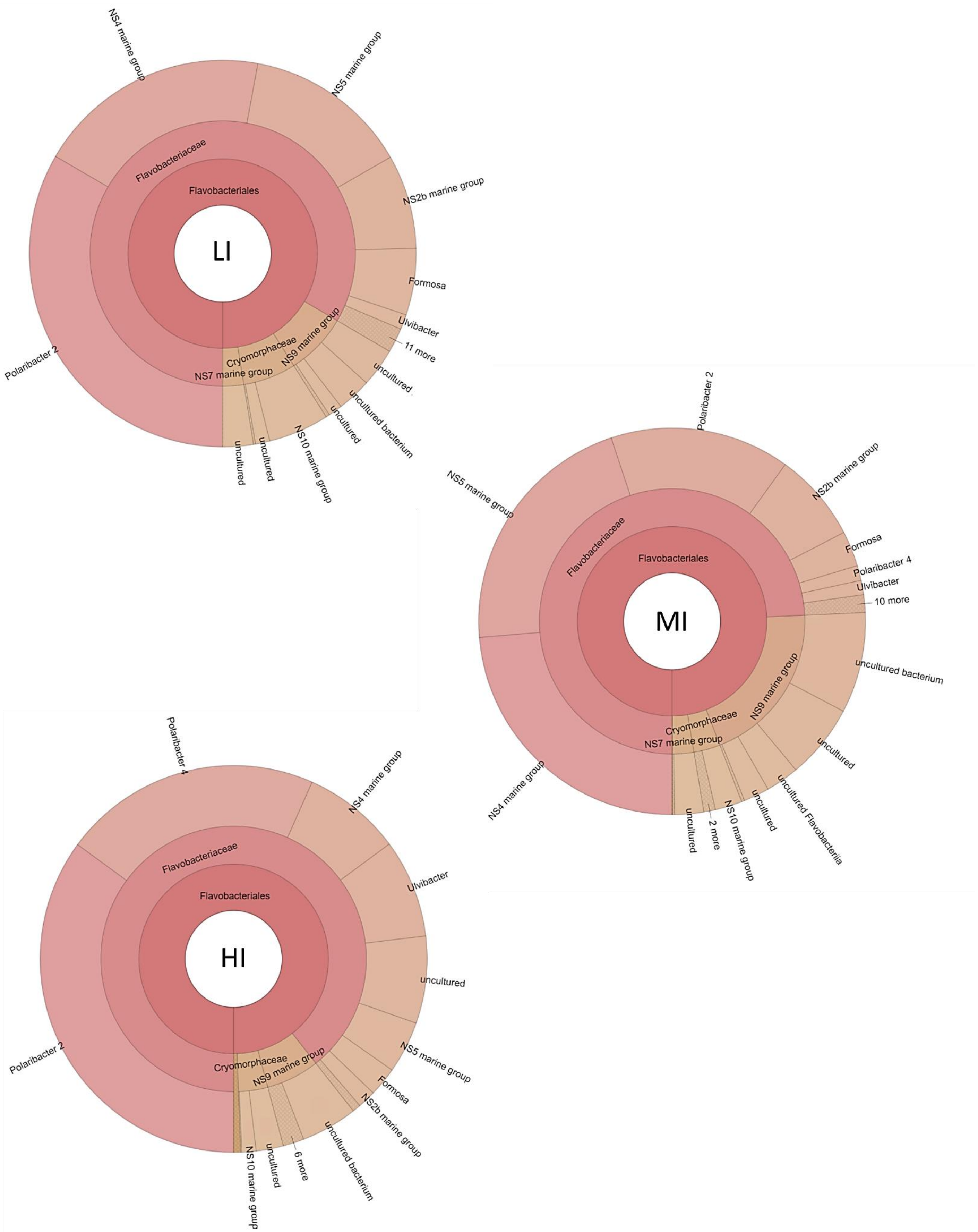
Regional partitioning of the bacterial community divided into representative major taxonomic groups was based upon dendrograms of the Bray-Curtis dissimilarity matrix between stations. Shown are those groups which A) show regional partitioning which matched the regional assignment of stations as described in Chapter 3. All taxonomic groups partitioned with the MI station most similar to the LI stations. Taxonomic groups are ordered by similarity of the LI to HI stations from most to least similar from left to right. B) Taxonomic groups which displayed no clear regional clustering, and so aren't ordered by similarity. Stations are coloured based upon the extent of Polar Water influence determined to be present at each station; red- LI, green- MI, blue- HI. The number of sequences comprising each taxonomic group is shown (n).

4.2.5 Community Composition

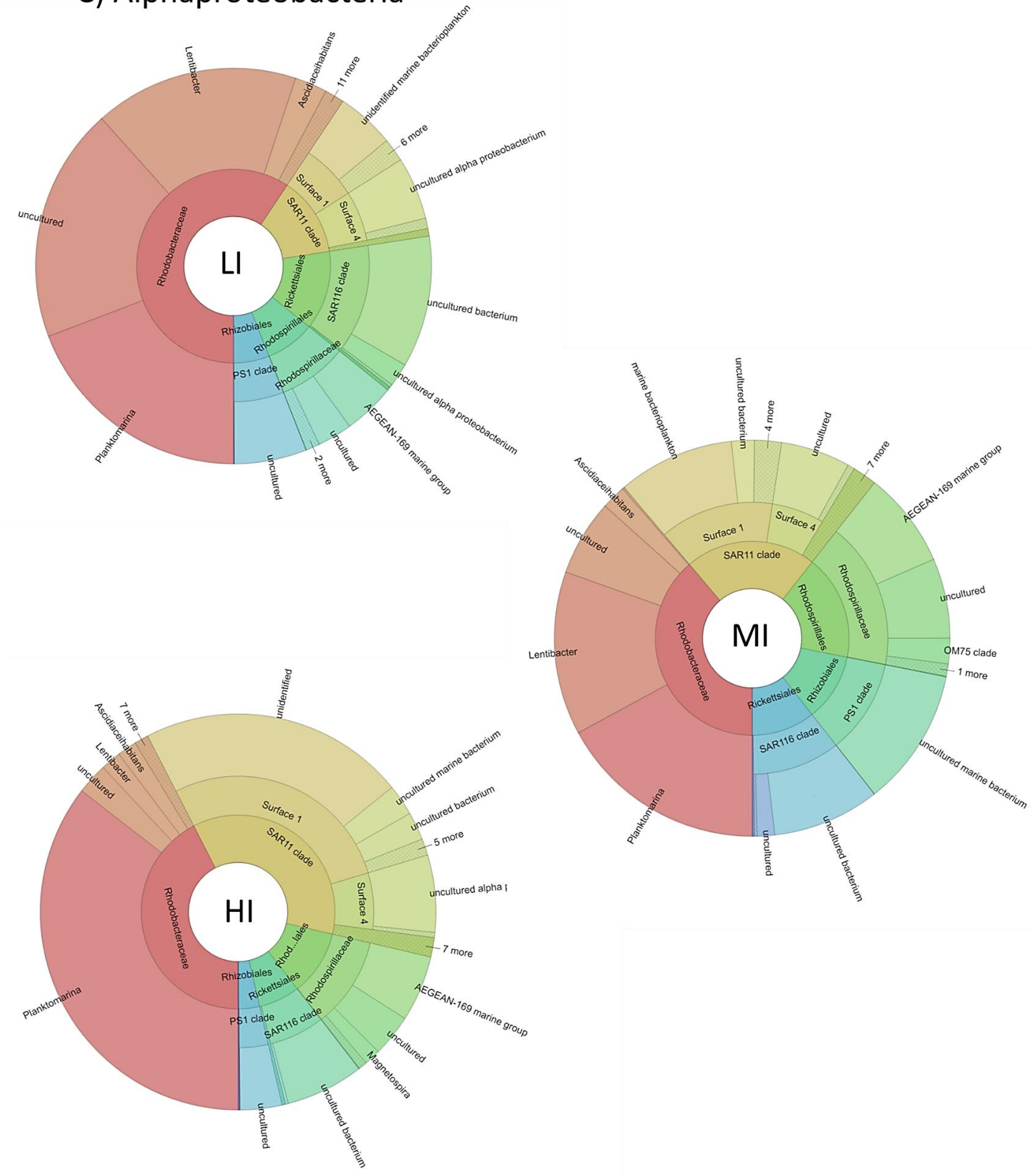


(A) Verrucomicrobia

B) Flavobacteriia



C) Alphaproteobacteria



D) Gammaproteobacteria

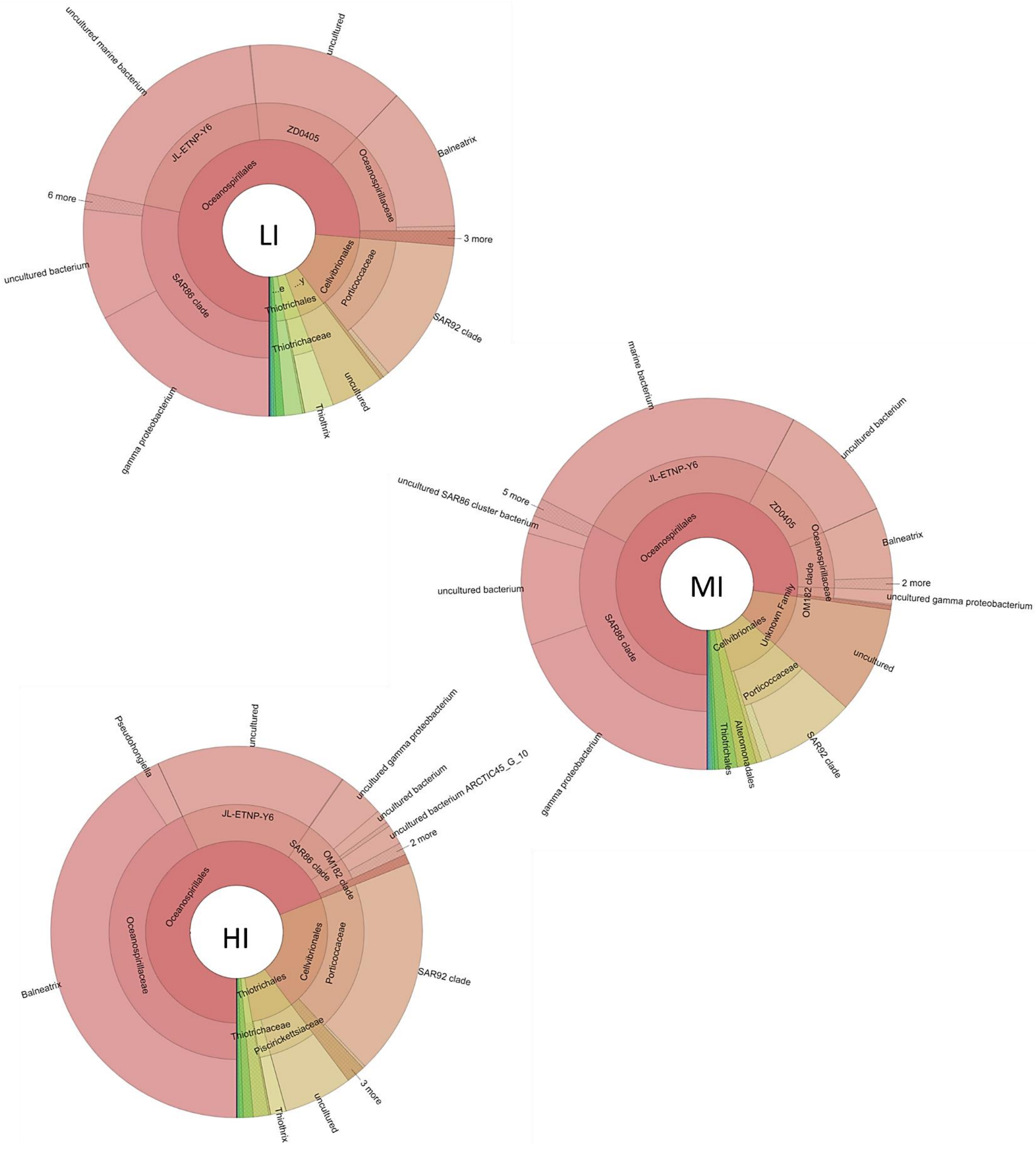


Figure 4.5. Differences in bacterial community composition between stations.

Shown are select examples to illustrate the difference in community compositon between stations under high Polar Water influence (HI), those under moderate Polar Water influence (MI), and those under low Polar Water influence (LI). Data is displayed for (A) the Verrucomicrobia, (B) the Flavobacteriia, (C) the Alphaproteobacteria, and (D) the Gammaproteobacteria.

Analysis of the community composition was undertaken using Krona[171]. Proteobacteria were observed to dominate all stations across the transect in terms of relative abundance, dominance was most pronounced within the MI station group (84%), and lowest within the HI group (64%). The LI group featured similar levels of abundance to the HI group, being 69%. The Flavobacteriia displayed a different pattern of abundance, being highest in the HI group (34%), lowest in the MI group (12%) and 27% in the LI group. The Actinobacteria constituted 2% of the community in the MI group, and ≤1% in both the LI and HI groups. The Verrucomicrobia constituted 1% within the HI group. No other taxonomic groups at the same taxonomic level were observed to be present at ≥1% of the total community within each station group.

Exploring each of the aforementioned taxonomic groups at greater taxonomic depth revealed that the Actinobacteria was dominated by *Candidatus Actinomarina*, which constituted <1% of the total community abundance in all except the HI group. The Verrucomicrobia displayed a similar trend, with only *Lentimonas* forming ≥1% of the total community composition at the HI group, no other taxonomic group was observed to exceed 1% of the community in either of the other two station groups. Despite this, some notable compositional differences were observed. The Puniceicoccaceae marine group constituted nearly the entirety of the Verrucomicrobia within the LI group, but featured reduced dominance within the MI group, resulting in a roughly equal split between *Roseibacillus* and the Puniceicoccaceae marine group (Figure 4.5A). *Lentimonas* dominated the Verrucomicrobia within the HI group, with *Roseibacillus* and the Puniceicoccaceae marine group present at reduced levels.

Within the Flavobacteriia *Polaribacter 4* was most highly abundant at the HI group (7%), but had low abundance at the other two groups (<1%) (Figure 4.5B). *Polaribacter 2* dominated the HI group (12%), and LI group (9%) but was observed to be much less abundant in the MI group (<2%). The HI group also featured elevated levels of *Ulvibacter* (3%) which appeared less abundant at the other station groups (<1%). By contrast the NS4 marine group was more abundant in the LI group (5%) compared to the other two (3%). The observed pattern for the NS5 marine group was similar, being most abundant in the LI group at 4%, falling to 3% and 1% for the MI and HI groups respectively. Lastly the NS10 marine group was identified to constitute 1% of the LI group, but <1% in the other two station groups.

Exploring the Proteobacteria in greater detail revealed similar levels of abundance of the three main constituent classes across all station groups. Betaproteobacteria were present at 1-2% across all groups and were heavily dominated by representatives of the OM43 clade. Alphaproteobacteria were observed to be present at 24%, 23% and 19% of the bacterial community for the LI, MI and HI groups respectively. Within this class, Rhodospirillaceae was present at similar abundances for all station groups (2-4%) (Figure 4.5C). The Planktomarina constituted from 4-7% of each group, and was the most abundant taxa within the Alphaproteobacteria for both the LI and MI group. The most abundant Alphaproteobacteria taxa within the HI group was SAR11 surface 1 (7%), which featured reduced levels in the MI (5%) and LI group (3%). *Lentibacter* was most abundant within the LI group (4%), falling to 3% for the MI group and <1% for the HI group. SAR116 clade also reduced in abundance from the LI group (3%), falling to 1% in the HI group. Abundance of the Alphaproteobacteria PS1 clade peaked at the MI group (3%), being ≤1% for both other station groups. SAR11 surface clade 4 constituted 1% across all station groups. A relatively abundant uncultured Alphaproteobacteria representative constituted 5% of the total bacterial community within the LI group, but was recovered at <1% elsewhere. Lastly, the Gammaproteobacteria constituted 41-42% of the community in both LI and HI groups, 55% of the MI group community, and featured some of the most abundant individual taxa. *Balneatrix* heavily dominated the HI group at 17%, but decreased in abundance across in the MI group (3%) and LI group (5%) (Figure 4.5D). A similar trend was observed for a member of the SAR92 clade (Porticoccaceae) which featured an abundance of 8% in the HI group but <5% for the other two groups, and Piscirickettsiaceae (2% HI group and <1% for the other station groups). By contrast the SAR86 clade featured the highest abundance within the LI group (12%) and MI group (18%), but had low abundance in the HI group (2%). A similar trend was observed for ZD0405 (6% LI, 6% MI and <1% HI), and the Thiotrichaceae (1% LI, <1% HI and MI). Lastly JL-ETNP-Y6 constituted 14% of the total community abundance within the MI group, 9% in the LI group, and was lowest in the HI group at 7%.

4.2.6 OTU Distribution and Community Distinctness

Venn diagrams were created to visualise the community distinctness and OTU distributions of the bacterial community between stations. Analysis of the whole community revealed a high number of station specific OTUs, found only at individual stations, for all stations (Figure 4.6A). CTD56 featured the highest number of station

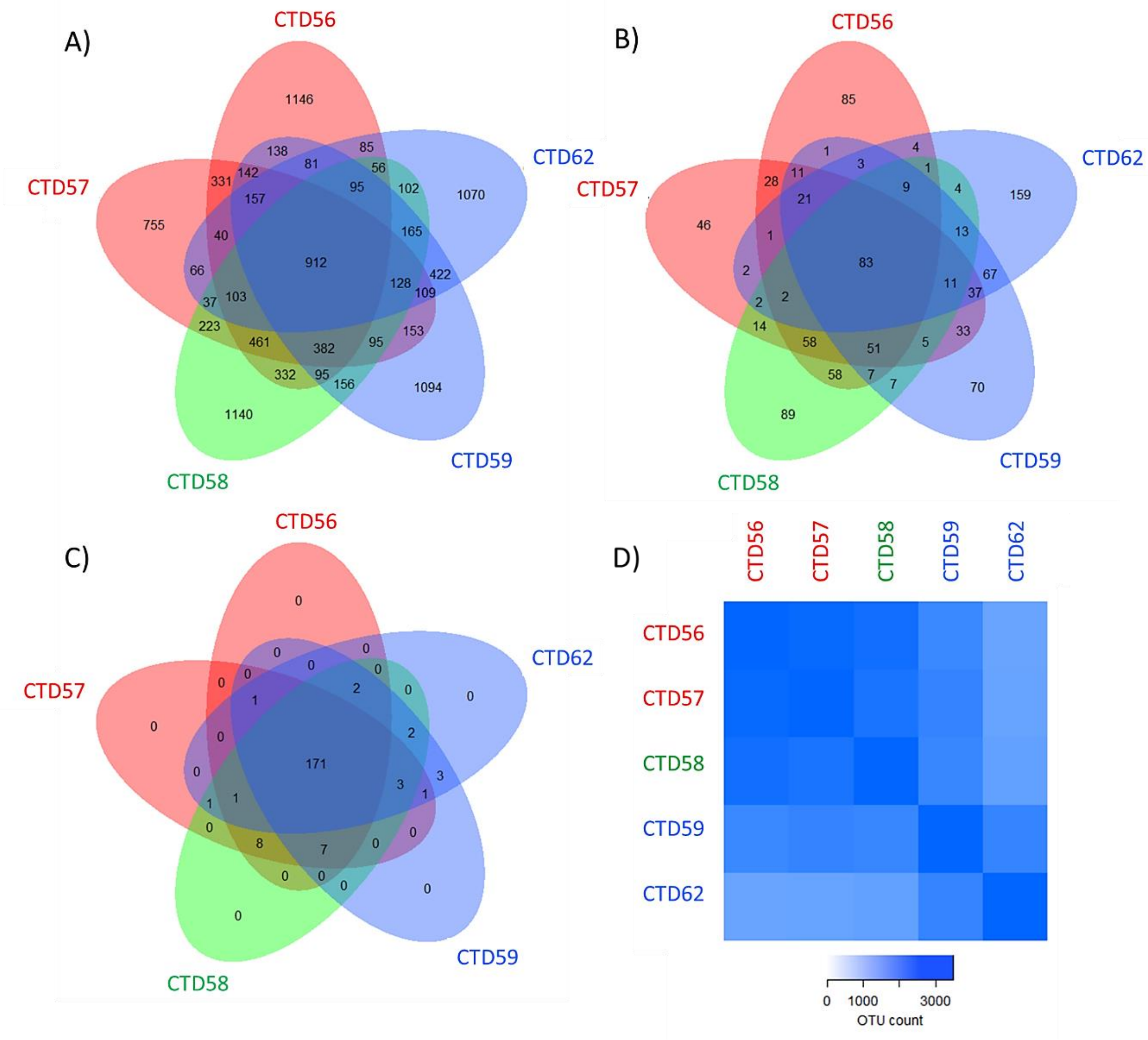


Figure 4.6. Analysis of bacterial community distinctness.

Shown are the shared and specific OTUs found within the transect. Numbers represent the count of OTUs. OTU counts in station ellipses which do not overlap with any other represent the number of OTUs specific to that station, whereas those that do overlap represent the number of OTUs found across those stations for which ellipses overlap. Community distinctness is shown for the whole community (A), as well as with those OTUs represented by ≥10 sequences at any one station (B), and just the top 200 most abundant OTUs (C). Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI. Also shown is the number of OTUs shared between each pair of stations (D), with the number of OTUs represented by the colour scale, lighter shades indicate fewer OTUs.

specific OTUs (1146), but similar numbers were recovered at all other stations except CTD57 which featured a notably lower number (755). A core community of 912 OTUs was observed at all stations. CTD62 featured the most isolated community, sharing comparatively few OTUs with other stations, except the other HI station CTD59.

Repeating the Venn diagram analysis with only those OTUs represented by ≥10 sequences resulted in a dramatic drop in the number of OTUs showing that the vast majority of OTUs recovered were rare, represented by only a small number of sequences (Figure 4.6B). Despite this, the broad patterns discussed still remained, implying they were not artefacts from OTUs of low sequence depth. Each station featured a pool of unique OTUs with CTD57 still featuring the lowest number, and CTD62 shared the highest number of OTUs with its neighbouring station CTD59 than did any other station and their neighbour. A core community shared between all stations was still observed and constituted a similar fraction of the community as it did when OTUs represented by <10 sequences were included.

Analysis of just the top 200 most abundant OTUs (representing OTUs constituting 67% of the bacterial community) revealed that the majority were found across all stations as part of the core community (Figure 4.6C). 7 of these abundant OTUs were observed to be present across all but CTD62, 8 across all except CTD59 and CTD62 and 3 were unique to just CTD59 and CTD62.

For all stations the number of OTUs shared with other stations decreased as the distance between them across the transect increased (Figure 4.6D). Furthermore, station groups formed two compositionally similar clusters, the two HI stations were observed to be compositionally most similar, and were most distinct from the LI and MI stations which were compositionally similar.

To explore the community structure of the top 200 most abundant OTUs present at all stations their relative abundance across stations was plotted. At the level of OTU some trends can be seen, which manifest as different patterns of abundance. OTUs were observed to either be present across all stations at similar abundances (Figure4.6A), or display a strong abundance at either the LI stations (CTD56 and CTD57) (Figure 4.7B), the MI station (CTD58) (Figure 4.7C), or the HI stations (CTD59 and CTD62) (Figure 4.7D). Correlation of these OTUs with environmental parameters revealed the existence of significant correlations with temperature and salinity (*p*≤0.05) for 49 of the 200 OTUs (Table 4.4). Correlations were positive for OTUs showing higher abundances at LI stations (R>0.9), and negative for those at HI stations(R<-0.9). The highest number of correlations were seen with silicate (full Spearman’s rank correlation data available in Appendix 1).

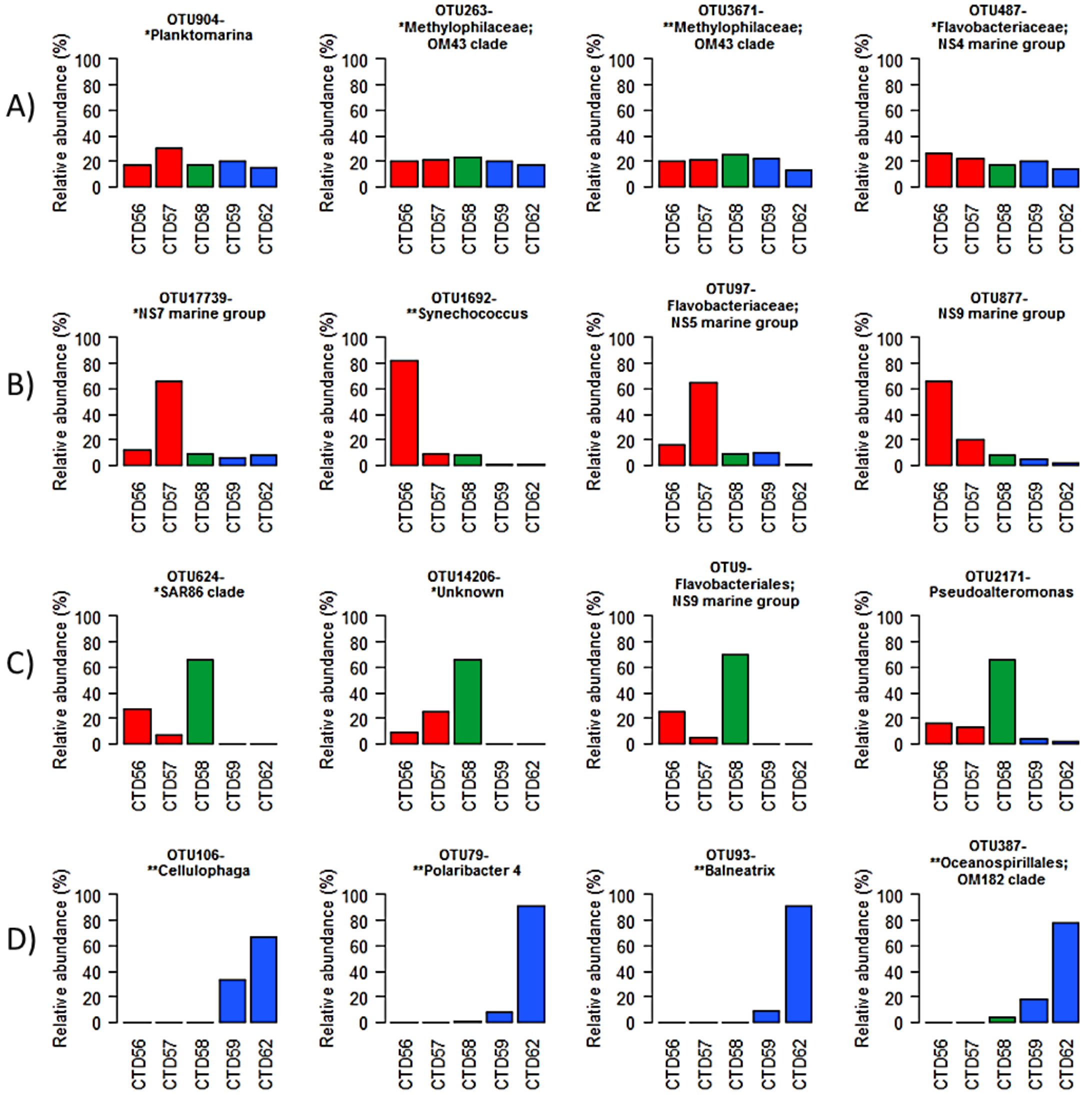


Figure 4.7. The relative abundance of bacterial OTUs between stations.

OTUs were selected from the top 200 most abundant in the bacterial dataset. OTUs show different distribution patterns which help explain the community assemblage structure and can be grouped into those which are present across all stations at relatively similar levels (A), show strong abundance bias for primarily LI stations (B), primarily MI stations (C), or primarily HI stations (D). ** indicates a significant association with temperature as revealed by Spearman's rank correlation with environmental conditions measured at the sampled depth for which water samples were taken. * indicates a significant association with other environmental variables. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

Table 4.4. The number of bacterial OTUs showing correlations with each environmental factor.

In situ environmental measurements were taken from the depth at which environmental water samples were collected for sequence analysis during CTD casts for each station. Shown is the number of OTUs that displayed a significant (p≤0.05) correlation with each of the listed environmental variables. OTUs tested were taken from the top 200 most abundant OTUs.

Environmental factor	Number of correlated OTUs (p<0.05)
DO_2	47
Salinity	49
SubSurVPAR	30
Temperature	49
Ammonium	49
Nitrate	7
Silicate	57
Phosphate	6

Heat maps of abundance based distribution patterns were generated to allow how OTUs which displayed an abundance pattern are distributed across each station to be further distinguished. OTUs with an abundance ≥60% at a particular station were deemed to display an abundance based "station bias" to that station, henceforth are referred to as "station biased" OTUs. 391 OTUs (Table 4.5) were observed to display an abundance bias (red) to certain stations indicating preferential conditions at individual stations (Figure 4.8), the highest number of which were observed at CTD62, being 159.

Heat maps were plotted for each major bacterial taxonomic group to visualise the distribution of station biased OTUs within each constituent taxonomic group (Figure 4.8). Total OTU number varied between taxonomic groups, but still provided a snapshot of the relative distribution of station biased OTUs for each group. The Flavobacteriia featured 151 station biased (red) OTUs (Table 4.5), 101 of which were identified at the HI station CTD62 (Figure 4.8). An absence (black) of these OTUs at CTD56, CTD57 and CTD58 was also observed. The Alphaproteobacteria, Acidimicrobia and Gammaproteobacteria featured 71, 3 and 94 station biased OTUs (Table 4.6), respectively. Of these OTUs the highest numbers were identified at CTD56 (32), CTD57 (2) and CTD58 (37) for the Alphaproteobacteria, Acidimicrobia and Gammaproteobacteria respectively. An absence of OTUs at HI stations was again observed for these three taxonomic groups (Figure 4.8). No representative OTUs of Betaproteobacteria were deemed to show an abundance bias to any stations. All other taxonomic groups featured few OTUs making it difficult to draw conclusions likely to be reflective of the group.

Table 4.5. The number of bacterial OTUs deemed to display an abundance bias to each station.

OTUs that featured ≥60% relative abundance at a particular station were deemed to display an abundance bias to that station. Shown are the counts of OTUs deemed to display an abundance bias to a particular station for the whole community, and for each constituent taxonomic group. The total number of OTUs deemed to display an abundance bias within each taxonomic group are also shown.

Taxonomic group	CTD56	CTD57	CTD58	CTD59	CTD62	Total
Whole community	75	49	72	36	159	391
Acidimicrobiia	0	2	0	1	0	3
Alphaproteobacteria	32	13	5	8	13	71
Betaproteobacteria	0	1	0	0	0	1
Cytophagia	0	0	1	0	0	1
Deltaproteobacteria	0	0	2	0	2	4
Epsilonproteobacteria	2	0	0	0	0	2
Firmicutes	1	3	0	0	0	4
Flavobacteriia	13	23	6	8	101	151
Gammaproteobacteria	9	5	37	9	34	94

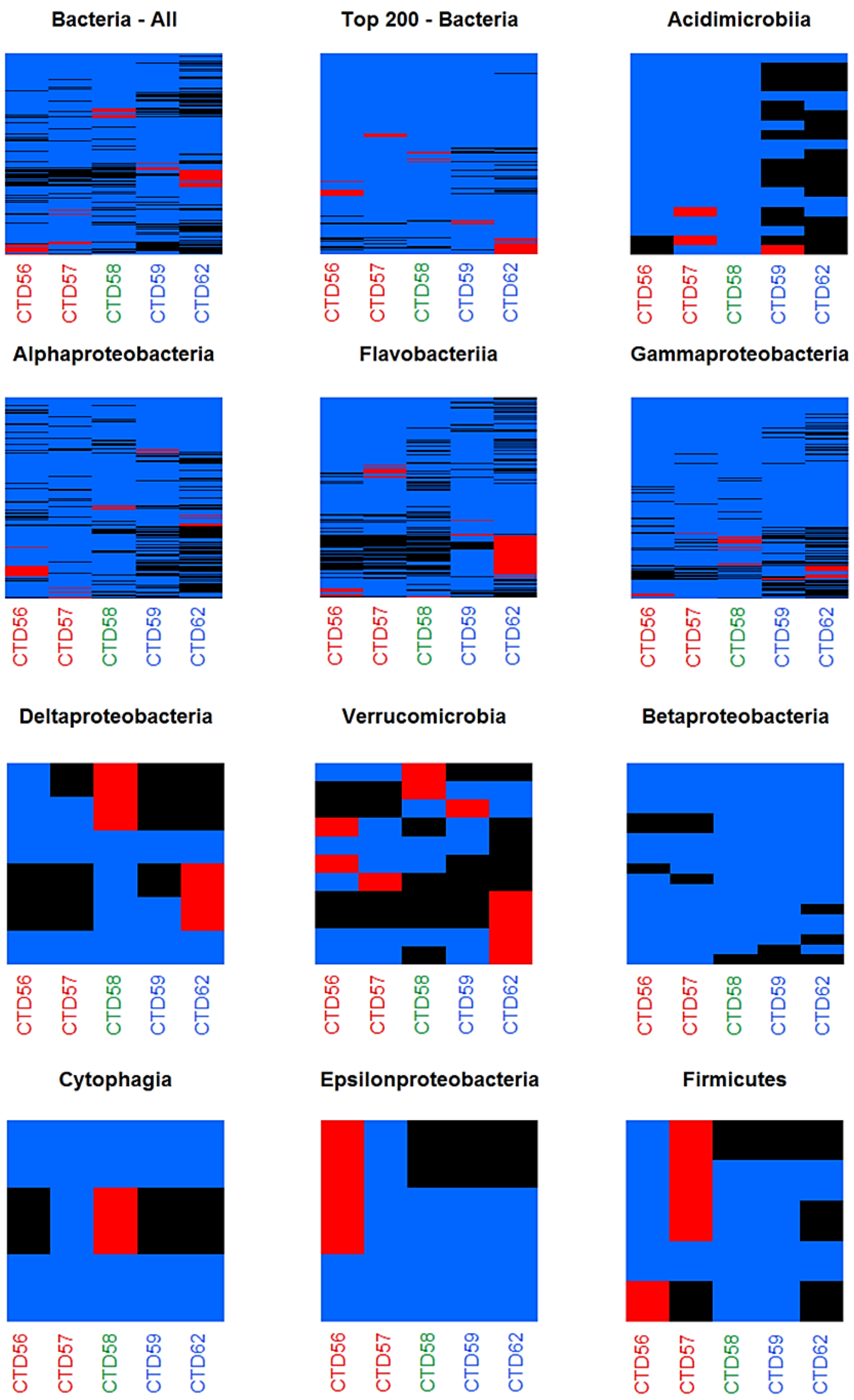


Figure 4.8. Heat maps of bacterial OTUs deemed to display a station abundance bias.

Shown is the distribution of OTUs deemed to display a station abundance bias within the whole bacterial dataset, the top 200 most abundant OTUs and for OTUs representing the major bacterial taxonomic groups. Black represents OTU absence at a particular station, blue represents OTUs that displayed 0-60% abundance at each station, and red represents OTUs that displayed ≥60% abundance at one station and thus were deemed to be display an abundance bias to that particular station. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

4.3 Discussion

This chapter explored the bacterial diversity present across a transect in the Norwegian Sea to provide an insight into the potential susceptibility of such communities to perturbations in the face of current environmental change. The sampled transect covered a gradient which followed the transition of water masses displaying varying amounts of influence from Polar Water (as discussed in Chapter 3).

4.3.1 Sequence Analysis and α-diversity

Analysis of sequence depth (Figure 4.1) revealed that saturation was not reached for either the full bacterial dataset (Figure 4.1A), each station group (Figure 4.1B) or at each individual station (Figure 4.1C). Under saturation is common in environmental datasets as the number of raw sequences needed to generate a complete inventory of a community is prohibitively high due to the presence of extremely rare taxa[129]. The sequence quality filtering methodology applied to the bacterial dataset resulted in a retention of 49.7% of the raw sequence data, however, as the excluded sequences were likely the result of sequencing errors (in the form of singletons) or eukaryotic contaminants their removal is justified.

In total 10,272 bacterial OTUs were recovered from the transect (Table 4.2). Sequences were assigned to OTUs based on a more stringent similarity value of 98.7% than conventionally used in other studies, which typically use a 97% threshold, but that is suggested to be more suitable for resolving genetic uniqueness of bacterial sequences at equivalent to the species level[163]. Even with this high stringency comparatively few novel bacterial OTUs were identified that had no prior database representation, which may be an indication of the progress recent sampling efforts are making in the mission to sequence the global microbiome, with much of the most abundant community members now represented in online databases. A recent update on the status of the bacterial census reports >95% of bacterial sequences targeting the 16S rRNA gene have now been observed more than once, and that the rate of novel OTU discovery has been declining since 2007[226] supporting this suggestion.

The lowest number of novel OTUs were observed at CTD62, which also featured the lowest number of recovered OTUs, Shannon index, and Simpson index when compared with the other stations (the latter excluding CTD59) (Table 4.2). This indicates low biological diversity at CTD62 relative to the LI and MI stations, and that the community is composed of a greater number of low abundance taxa, featuring a restricted number of abundant community members. Additionally, the HI stations were seen to be compositionally most similar, and distinct from the LI and MI stations (Figure 4.6D). Bacterial communities in polar marine habitats face different selection pressures than those inhabiting temperate marine waters, such as being exposed to perennially low water temperatures, seasonal fluctuations of ice-cover and irradiance, and reduced salinity levels[110,127,150]. These environmental conditions are potentially inhospitable to a number of taxa and place selective pressure upon them, favouring those more competitive under such pressures. Therefore, the aforementioned pressures likely select for different assemblages between stations with different environmental conditions. Indeed, bipolar distributions have been reported for polar regions[211] and it has been suggested that polar environmental conditions select for assemblages composed of a greater number of specialist species which are adapted to cooler and fresher environmental conditions[27].

The differing community evenness observed between the HI and LI stations also explains the discrepancy between the predicted and observed OTU richness (Table 4.2). The metrics produced by the iNEXT package and ACE are used to estimate the total OTU richness present in a sample based upon the observed number of OTUs, taking into account the number of rare OTUs. However, the estimated OTU richness may be over inflated if the community features a greater level of evenness with a comparatively low number of rarer taxa, as evidenced by the low Simpson index for CTD59 (Table 4.2). This likely explains why CTD59 had the highest predicted OTU richness, despite CTD56 having the highest observed OTU richness and highest Shannon index.

4.3.2 Environmental Correlation

The degree of biogeography exhibited by marine bacterioplankton is under ongoing debate. On one hand it is thought that the small size and estimated abundances of 10^4-10^6 cells per ml allow marine bacteria to be distributed globally by natural forces, resulting in a ubiquitous presence of many taxa at any location[42]. It is also suggested that selection by environmental factors and historical events may result in regional differences in community structure[32,33]. Data presented in this chapter revealed the presence of community partitioning across the sampled transect (Figure 4.2A), which strongly correlated with the environmental factors temperature and salinity (p<0.01)

(Table 4.3) and is further supported by the correlation of individual OTUs with environmental factors (Table 4.4). A recent analysis derived from the Tara Oceans dataset which found that temperature was the main driver behind global epipelagic microbial community structure supporting these findings[42]. Arctic communities are currently excluded from the Tara analyses, but other studies have reported that polar bacterial communities are distinctly different from those found elsewhere and that environmental factors are likely drivers of Arctic bacterial biogeography[211]. Work by Lucas *et al*[218] reached a similar conclusion, showing temperature as a major niche-defining factor for bacterioplankton communities which contained examples of taxa also found in the analysis presented in this chapter (such as SAR86, SAR92, NS4 & NS5 marine groups, *Polaribacter* and ZD0405 clade). That both temperature and salinity were observed to significantly correlate with the bacterial community is likely partly the result of the nature of the sampled region, as Polar Water influence acts to both freshen and cool (Chapter 3).

The HI stations, CTD62 and CTD59, were observed to be more isolated in terms of OTU distribution (Figure 4.6D), and featured the greatest number of shared taxa (Figure 4.6A). Mixing of the water profiles and the contained microbial taxa between these two stations partly explains the observed similarity. However, it is likely that the high number of shared taxa are also the result of similar environmental conditions between the HI stations (Figure 3.1) resulting in similar selection pressures, distinct from other stations, acting to select for a distinct community composition. Indeed, responses of bacterial communities to abiotic factors including temperature and salinity are well documented in current literature[227,228], with changes in community composition shown to occur sharply over both spatial and temporal gradients[229,230]. Therefore, similar selection pressures between the HI stations would result in similarities between their respective bacterial communities.

The partitioning of the MI station community in relation to the LI and HI communities provides a potential means to predict which community may become dominant under increased Atlantic Water influence within the region. Partitioning that mirrored the regional assignment of station groups based on the measured physical environmental data (as described in Chapter 3) was present at all abundance fractions of the bacterial community (Figure 4.2C,D,E). In all cases the MI station community was most similar to that of the LI stations. Similar partitioning was also observed for the majority of the major bacterial taxonomic groups (Figure 4.4A), namely the Alphaproteobacteria, Betaproteobacteria, Deltaproteobacteria, Gammaproteobacteria, Flavobacteriia, Verricomicrobia, Firmicutes and

Acidimicrobiia. In all instances the MI station partitioned with the LI stations, suggesting that warm adapted temperate taxa may replace cold water adapted taxa under increased Atlantic Water influence in the sampled region. While the ecological consequences of changes to these taxonomic groups remain largely unknown, the 3 most abundant groups within the bacterial dataset, namely the Alphaproteobacteria, Flavobacteriia and Gammaproteobacteria are frequently reported as the most dominant taxa in the global ocean[40,129,152]. As high levels of bacterial abundance are often the result of active community members, capable of high turnover rates under favourable conditions[37], the impacts of a deviation from these conditions has the potential to be substantial. It is worth noting that such turnover rates also generate high levels of genetic variation which may confer adaptions to mitigate the impacts of environmental change, but measuring the rates of such diversification within the environment *in situ* is difficult[37].

4.3.3 Potential Responses of Rare Taxa

As is common in NGS studies, the majority of OTUs were observed to belong to the rare fraction of the community, with comparatively few OTUs represented by >10 sequences (Figure 4.6B). The rare component of the microbial community represents the vast majority of the biological diversity in the oceans[54]. The purpose of rarity in bacteria is still unclear, it has been suggested that rarity may be an active survival strategy to evade predation and viral lysis[53]. It has also been proposed that rarity may be a passive artefact resultant from ubiquitous dispersal of taxa, with taxa persisting as rare due to being unable to reach high abundances due to unfavourable conditions[232]. However, a study by Galand *et al*[49] that analysed rare bacterial taxa present across the Arctic Ocean suggested that rare bacterial taxa display biogeography, and that environmental conditions could be an explanatory factor. The analysis of the rare fraction of the community (those constituting <0.01% of the total community abundance), and those within the intermediate abundance faction (0.01-1%) presented in this chapter support this notion (Figure 4.2D, 4.2E), implying rare bacterial taxa are also potentially susceptible to environmental perturbations.

Little is known about the ecological or functional role of members of the rare community[49]. It is suggested that these taxa are capable of acting as a seedbank, responding to environmental perturbations if conditions become more favourable by increasing in abundance and providing a level of functional redundancy to maintain ecosystem services[233]. Frequently these taxa are suggested to be metabolically inactive or feature very slow growth rates[54], but there is evidence that rare species play a key role in the resilience of an ecosystem by significantly supporting ecosystem functions with low functional redundancy[222] and can display high levels of metabolic activity to disproportionally contribute to certain ecosystem functions[234]. Therefore, the loss of such taxa would have implications both upon the functioning of an ecosystem or it's resilience to future change, and are suggested as the most vulnerable taxa to local extinctions[222].

4.3.4 Potential Responses of Abundant Taxa

The majority of abundant members of the community were observed to be cosmopolitan, being recovered from all stations (Figure 4.6C). Given the degree of difference between the environmental conditions measured at different stations it could be that these taxa represent generalist species tolerant of a range of environmental conditions[27]. However, such cosmopolitan distributions may be in part be driven by the random dispersal of locally highly abundant taxa[34,235]. It has also been shown that bacterial taxa are capable of entering periods of metabolic dormancy in order to persist in the community despite unfavourable conditions by interpreting environmental cues such as crowding, nutrient limitation or temperature stress[35], which would still be recovered by the DNA barcoding methods applied in this chapter. However, examples of taxa which were able to maintain consistently high levels of abundance across all stations were observed suggesting the presence of metabolically active taxa (Figure 4.7A). Taxa that displayed high levels of abundance at only certain station groups were also observed (Figure 4.7A, Table 4.5), which suggests a growth response of viable living cells in reaction to favourable conditions. These findings are validated by select examples of taxa known to display biogeographic patterns. For example, representatives of *Polaribacter*, *Cellulophaga* and members of the OM182 clade are known to be associated with sea-ice, and for which distinct Arctic ecotypes have been identified[208,236,237], displayed the highest abundances at the HI station CTD62 (Figure 4.7D). Additionally representatives of *Balneatrix*, which is known to be associated with fresh or brackish waters and has been shown to exhibit biogeographic patterns linked to temperature[228,238], featured the highest relative abundances at the HI station CTD62. Such distribution patterns are further supported by significant negative correlations with temperature and salinity for these OTUs, as would be expected for cold water associated taxa. The presence of these distribution patterns was not found to be a common feature within the bacterial community as only 31 of the top 200 most abundant OTUs displayed an abundance bias for a certain station group, and the total number of 391, represented a much reduced fraction of the community compared to the eukaryotic dataset (as described in Chapter 5). The lower recovery of bacterial OTUs displaying abundance biases to certain stations might be due to the use of 16S amplicon sequencing which may not be able to resolve bacterial taxa into units with the greatest ecological significance. Inter-OTU diversity may better resolve ecological patterns as has been suggested in previous studies[58,239]. Despite this, visual trends of the top 200 most abundant OTUs resolved many which exhibited a graduated distribution across stations (Appendix 2). Therefore, it is possible that the observation of bacterial distribution patterns are dampened by the greater ease of the their passive dispersal compared to larger eukaryotes[240], or that environmental forcing has a greater impact at the community level than at the level of individual OTUs within bacterioplankton communities.

The association of individual OTUs with environmental factors revealed a high number of significant correlations with salinity, temperature and ammonium (49 OTUs for each factor), but the greatest number was seen for silicate (57) (Table 4.4) and which was resolved as a significant explanatory factor (p<0.05) of the partitioning of the whole community (Table 4.3). A large proportion of the OTUs displaying significant correlations with silicate were resolved to be representatives of *Polaribacter*, *Balneatrix* and SAR86 clade. For each of these taxa the association appeared non-random as the direction of the association was consistent across multiple OTUs, being negative for *Polaribacter* and *Balneatrix* (R>0.87), and positive for SAR86 clade (R<0.9). The reason for the correlations of these taxa with silicate was not clear, but may be the result of associations with Diatoms. *Polaribacter* belongs to the Flavobacteriia, and *Balneatrix* and SAR86 the Gammaproteobacteria, which are known to associate with diatoms[241].

4.3.5 Compositional Differences Between Station Communities

Compositional differences were observed at broad levels of taxonomic resolution under different environmental conditions on which speculations can be based as to how these taxa may react to climate driven environmental change. Analysis of the community composition revealed that all station groups were dominated by Proteobacteria, which is frequently reported to be the dominant marine bacterial group globally[39,42]. The highest abundance of Proteobacteria were recovered from the

MI group (80% of the whole community), and lowest within the HI group (63%). Flavobacteriia were also observed to be highly abundant, reaching the highest abundance in the HI group, agreeing with previous reports of Flavobacteriia abundance in Polar Waters[39,132]. The remainder of the community constituted representatives of Actinobacteria (<2% of the community) and Verrucomicrobia (1%), both of which are commonly reported as minor constituents of marine bacterioplankton[45,242]. No other groups were present at an abundance of ≥1% of the total community at each group.

At greater taxonomic resolution clear differences in community structure between the HI and LI station groups were observed. The Flavobacteriia in the HI group were mainly composed of *Polaribacter. Polaribacter 4* consituted 7% of total community, but was much less abundant in the other two groups (<1%). By contrast, *Polaribacter 2* was observed to be highly abundant in the HI and LI groups (12% and 9% respectively), but not in the MI group (2%). *Polaribacter* species were first isolated from sea ice in polar locations[237] but have since been recovered from temperate locations[243,244]. The genus is suggested to include distinct ecotypes adapted to different thermal profiles, with some taxa exclusively psychrophilic[237] or restricted to polar waters[245]. The findings presented in this chapter appear to indicate the possibility of two distinct ecotypes. *Polaribacter 4* appeared to be cold adapted based on high abundance at HI stations, whereas *Polaribacter 2* maintained more consistent abundance at LI and HI station groups, which implies it may be a cosmopolitan or generalist species tolerant of a wider range of environmental conditions[27]. The HI group featured elevated levels of *Ulvibacter* (3%, compared to <1% at other stations). Representatives of this genus were also first isolated from a polar environment[246], and those identified in the bacterial dataset appear cold adapted based upon their abundance profiles. By contrast the NS4 marine group was more abundant in the LI group (5%) compared to the other two (3%). Representatives of the NS5 marine group were similar, being most abundant in the LI group at 4%, falling to 3% and 1% for the MI and HI groups respectively. However these differences are marginal, and NS marine groups are frequently recovered from a variety of marine habitats[218,247].

Of the Proteobacteria, the Betaproteobacteria were consistent in their abundance (1-2%), and heavily dominated by the OM43 clade across all station groups. The Alphaproteobacteria were most abundant in the LI group (24%, falling to 19% in the HI group). Alphaproteobacteria are a highly diverse group and are reported to feature high abundances in pelagic water systems, often being the dominate Proteobacteria group in Arctic waters[129], with SAR11 reported as a main representative[40,248]. SAR11 is a globally successful clade thought to play a significant role in nutrient and carbon cycling[249]. A number of studies report the existence of a number of SAR11 subclades which display different preferences for environmental conditions[250,251]. Representatives of the SAR11 clade (7%) and *Planktomarina* (7%) dominated the HI group but the LI group featured reduced levels of SAR11 (3%) instead being dominated by *Planktomarina* (5%), *Lentibacter* (4%) and an uncultured Rhodobacteraceae (5%) which are widely reported in temperate waters[252]. Representatives of the SAR116 clade were also observed to be more abundant in the LI group (3%) compared to the HI group (1%). These taxa have been associated with coastal water profiles[248], and are thought to be significant contributors to dimethyl sulphide (DMS) production[253]. Therefore, the differences in community composition observed between station groups agree with reports from the literature.

Gammaproteobacteria were present across all station groups at roughly the same abundance, but showed a clear difference in composition. *Balneatrix* heavily dominated the HI group at 17%, but decreased in abundance across the MI (3%) and LI groups (5%). *Balneatrix* is a known potential human pathogen, originally isolated from contaminated spar water[238], and is thought to be a freshwater taxa[254]. It has recently been recovered from North Pacific surface waters influenced by melting Arctic sea ice[228]. It's abundance in the HI group is therefore in line with its previously reported habitat range. SAR86 clade is another taxa which has been shown to display temperature driven distributions with what appear to be cold associated ecotypes[255], and was most abundant within the LI group (12%) and MI group (18%), but had low abundance in the HI group (2%), suggesting that the OTUs recovered for representative were not cold-adapted and unable to flourish at the HI stations. This may also be true for ZD0405, which has been shown to display reduced abundance in response to reduced salinities and temperatures[256], and was observed to feature a similar abundance trend (6% LI, 6% MI and <1% HI), as did representatives of Thiotrichaceae (1% LI, <1% HI) and representatives of JL-ETNP-Y6 (14% MI, 9% LI and 7% HI).

4.3.6 Potential Winners and Losers

From the observed community partitioning, taxonomic differences observed between station groups, and validation gained from the literature as described in section 4.3.5, it is possible to make some general inferences as to how these communities may be influenced under future climate driven "Atlantification" of Arctic waters.

Betaproteobacteria consisted a small proportion of the total community and was dominated by OM43 clade representatives at all stations. As such the impact of changes to this group may be minor. Based upon the observed differences in composition across the stations featuring different amounts of Polar Water influence, it is suggested that reduced abundances of the uncultured strain that was dominant in the HI group is likely under increased Atlantic mixing. This will potentially be offset by an increase of another uncultured representative that was recovered.

Within the Alphaproteobacteria the *Planktomarina* may experience a reduction in abundance at the HI stations. A similar reduction may occur for the SAR11 surface 1 clade. *Planktomarina* plays a vital role in global carbon and sulphur cycles, as well as forming important relationships with animals and seaweeds[257]. SAR11 is an important contributor to the global carbon cycle, metabolising low-molecular-weight compounds, including amino acids, poly-amines and one-carbon (C1) compounds[118]. The metabolism of C1 compounds is reported to be a specialisation of the SAR11 bacterial group. As such, the predicted community changes to these two taxonomic groups could impact important biogeochemical cycles and organisms at higher trophic levels.

The abundance of SAR116 may increase in the Arctic region under increased Atlantic influence, and it is often reported as a dominant member of temperate bacterioplankton communities[247]. SAR116 has been suggested to show positive associations with algal blooms, which given reports of increasing algal biomass in Arctic waters may further compound abundance changes for this taxa[195]. *Lentibacter* is also predicted to increase but little information is available for this genus. The potential changes to SAR11 surface clade 4, PS1 clade and Rhodospirillaceae are unclear, it may be that their abundances will remain largely unchanged.

The findings presented in this chapter suggest a potential decline of *Polaribacter 4* and *Ulvibacter*, and potential increases of NS4, NS5 and NS10 marine groups are likely amongst the Flavobacteriia. Similar trends are suggested for the Gammaproteobacteria. The abundance of Piscirickettsiaceae and *Balneatrix* are likely to fall, with corresponding increases of ZD0405 representatives, Thiotrichaceae and SAR86. Despite little being known about the metabolic and functional abilities of SAR86, it is widely believed to be ubiquitous in marine pelagic communities, and like SAR11 features specialised chemotrophic metabolism of lipids and polysaccharides, thought to be an adaption to avoid resource competition[255]. Therefore changes to the abundance of this taxa may impact local nutrient cycling.

The Verrucomicrobia constituted a small proportion of the bacterial community but large compositional differences across the transect were observed. Thus, the potential for a reduction of the abundance of *Lentimonas* is predicted under further Atlantic influence, and corresponding increase of an uncultured representative of the Puniceicoccaceae marine group.

4.4 Conclusion

The findings presented in this chapter have demonstrated that the bacterial community present across a gradient of Polar Water influence in the Norwegian Sea displayed specific distribution patterns, which correlated with environmental factors (salinity and temperature, p<0.01). These distribution patterns remained when findings were validated against stations collected from the North Atlantic (Figure 4.2B), and agreed with expected distributions based upon previous reports of environmental sequencing efforts and cultivation experiments, with taxa displaying clear preferences for certain stations which are likely a result of the different environmental conditions found across station groups. Temperature is believed to be a significant explanatory factor of the bacterial community structure at multiple levels, existing at different abundance thresholds, within major bacterial taxonomic groups, for the bacterial community as a whole, and for certain individual OTUs.

Increasing ocean temperatures and mixing of Atlantic waters into the Arctic region are likely to result in changes to the composition of the extant bacterial communities. The data presented resolved distinct community partitioning which matched the regional assignment of station groups based on physical environmental data, and suggests that temperate associated communities from the most heavily Atlantic Water influenced stations may displace cold water associated communities found at highly Polar Water influenced stations under increased Atlantic Water influence in the sampled region. Furthermore, by exploring current literature for taxa known to display temperature driven distribution patterns, those which feature cold-water adapted ecotypes, and which have restricted geographic ranges, examples of individual taxa which are likely to be particularly susceptible to such environmental changes have been discussed. Such findings raise important concerns for the future of bacterial communities and the maintenance of existing ecosystem services in the face of ongoing environmental change to the region. However, select examples of well characterised bacterial taxa with known temperature responses are few, and as such detailed predictions quantifying the impacts upon ecosystem services resultant from such changes remain challenging. Given the high global abundance of marine bacterial communities and the rate at which community shifts can occur the impacts could be significant.

The analysis in this chapter focused upon the bacterial community, and it is unknown whether the eukaryotic community displays similar partitioning, associations with environmental factors, and whether local eukaryote community assemblages are at a similar risk of displacement under predicted increased Atlantic mixing and warming within the Arctic region.

Chapter 5

Eukaryote Community Gradients Along a Transect in the Norwegian Sea, in the Context of Climate Change

5.1 Introduction

Some of the first attempts to use molecular techniques to explore Arctic microbial communities have revealed highly diverse communities of microbial eukaryotes that include representatives from all major phytoplankton groups[57] and that feature a high degree of uniquely endemic taxa[258]. Throughout most of the year small picophytoplankton cells in the size range 0.2-2.0 µm dominate Arctic microbial eukaryotic communities due to being more competitive in the oligotrophic conditions of the Arctic Ocean[259]. Their large surface-area to volume ratios help prevent sinking and enables effective nutrient uptake[141]. In addition to significantly contributing to primary production in the Arctic[131], picoeukaryotes are also thought to account for an important fraction of carbon export to depth[260]. Some of the main picoeukaryote taxa have been reported to include Chlorophytes, Haptophytes, and parasitic Dinoflagellates (Syndiniales)[141]. The Chlorophyte *Micromonas pusilla* isolated from the Arctic is believed to represent a cold adapted endemic ecotype[261] and is found ubiquitously across the Arctic as one of the most abundant individuals. It is thought to fulfil the same functional niche as the picocyanobacteria *Prochlorococcus* and *Synechococcus* in rest of the global ocean[57]. Picoeukayotic Haptophytes such as *Phaeocystis spp.* are also amongst the most abundant key players in Arctic Waters[141]*.* Additionally*,* a number of larger species are highly abundant within Arctic microbial eukaryote communities including representatives of the Dinoflagellate *Gyrodinium spp.*[262]*,* and the Stramenopile *Fragilariopsis cylindrus*[150], another species which features a cold adapted Arctic endemic ecotype[97].

Seasonal fluctuations and patterns of succession have been observed within Arctic microbial eukaryote communities. During the spring and summer phytoplankton cells dominate the community. During this time bloom forming Chlorophyta such as *Micromonas,* and representatives of the Haptophytes (especially *Phaeocystis)* become particularly abundant, as do Dinoflagellate *Gyrodinium spp.,* but which remain abundant year round[262]. During early spring, before the late summer ice melt, Diatom species such as *Fragilariopsis cylindrus*, *Nitzschia frigida* and *Melosira arctica* form dense aggregates under sea ice, dominating ice-associated communities, but sink to the sea floor after sea-ice melt and thus constitute a small fraction of summer pelagic communities[110]. Seasonal dynamics are also present within other less abundant groups, for example Ciliophora display peak abundance over the spring/summer period[262].

During the polar winter distinct community composition changes have been observed. Phytoplankton become much rarer, likely only maintaining a presence in the community through dormancy or as resting spores[262]. However, both *Micromonas* and *Phaeocystis* have been widely detected throughout the Arctic during the polar winter, abet at lower abundances[263]. In the case of *Micromonas* it is possible that it is utilising mixotrophic feeding strategies such as phagotrophy as a winter survival strategy[264]. Indeed, mixotrophy forms an important winter survival strategy for microbial eukaryotes in Polar Waters, and the lack of phototrophic taxa is mirrored by an increased abundance of mixotrophic taxa[262]. The winter community is typically dominated by Dinoflagellates, which are found to be more abundant at greater depths, likely due to high sinking rates. Specifically representatives of *Gymnodinium*, *Woloszynskia* and *Gyrodinium* that feed upon other algal species during the winter are common[258]. In addition to being numerically dominant, Dinoflagellates display the highest level of genetic diversity over the winter months, featuring a large frequency of undescribed species within the core phylotypes (those persistently present amongst locations)[258]. By contrast other taxonomic groups show poor diversity. Representatives of the Cryptophyta and Picozoa increase in abundance over the winter months, with representatives of the Picozoa and parasitic Syndiniales (members of the Dinoflagellates) becoming particularly abundant[262]. These Syndiniales potentially infect the abundant mixotrophic Dinoflagellate community, but their high relative abundance is likely biased, in part, by low winter light levels inhibiting the growth of photosynthetic species[258,265]. Rhizaria are weakly represented year round, but are present within the winter community and include relatives of the algal predator *Cryothecomonas*[150]. Ciliophora, Chlorophyta and Amoebozoa are also rare over winter[258,262].

Regional differences in community composition exist across the Arctic that are suggested to be a result of differences in temperature, salinity and water mass influence[141]. Diatoms have been shown to be both rare[137] and abundant in Arctic waters[266] between contrasting molecular studies, and abundance has been observed to decrease with increasing lattitudes[137]. Differences in sampling locations and timing between studies likely explain the discrepancies observed. Other factors including bloom timings[267], nutrient upwelling suppression events during the sampling period in some studies[137], and severe light limitation acting to decrease primary production or an increase in grazing pressure on diatoms from heterotrophic taxa may also be contributing factors. The latter is supported by the detection of typical diatoms predators such as *Pirsonia*, *Cryothecomonas* and very high abundances of the dinoflagellate *Gyrodinium spirale* during winter[258]. However it is generally accepted that diatom abundance is usually high in the Arctic, particularly during summer, reducing further north into ice covered regions and during the polar winter, likely due to limited light availability[137,268]. Additionally, Diatoms are observed to feature high abundances throughout the Western Arctic regions which are fed by silicic acid rich Pacific Water, required for Diatom growth. By contrast low Diatom abundance is observed in regions fed by silicic acid deplete Atlantic Water[150]. Similar biogeography has been observed for Rhizaria, such as *Cryothecomonas,* which have been reported in Atlantic fed Arctic regions but not detected within Pacific fed regions[150]. As latitude decreases phytoplankton taxa including Chlorophytes (such as *Micromonas*) and Haptophytes (such as *Phaeocystis*) appear to make more significant contributions to microbial eukaryotic diversity, particularly in Atlantic fed waters[150,269], likely explained by a reduction in light limitation for these phototrophic taxa at lower latitudes. Stramenopile abundance remains low with evidence of complete absence in areas fed by Atlantic waters, but are suggested as important members of the rare community fraction[259].

Largely anthropogenically driven environmental perturbations are expected to have substantial implications upon Arctic microbial community compositions, with impacts to key ecosystem services. These implications have been suggested to include increasing levels of primary production[194], significant pelagic community assemblage shifts[217] and the import of invasive zooplankton which alter grazing pressures[126]. Significant changes in community structure have already been observed as a result of increased warming and sea ice loss. Small phytoplankton size classes (<2 μm) have been shown to be becoming more abundant in Arctic waters such as the Canadian basin[203], and it is expected that reductions in organism cell size will become common as temperatures warm further[270]. Abundances of Stramenopiles (mainly Diatoms), heterotrophic flagellates, and representatives of the Rhizaria such as *Cryothecomonas,* have fallen significantly over the last two decades. By contrast Ciliates such as *Strombidium*, *Novistrombidium*, *Codonellopsis* and *Pseudotontonia* appear to have increased in abundance. Haptophytes species such as *Phaeocystis* and *Chrysochromulina* have also significantly increased in Arctic Waters[217]. In addition to changes in native community structure the increasing intrusion of oceanic species into Arctic waters has been documented. These do not form resting spores capable of surviving the polar winter and so remain summer visitors[125]. Such species include *Emiliania huxleyi*, which is considered to be one of the major calcifiers in the global ocean. Both *in situ* collections and satellite images of high reflectance waters characteristic of *E. huxley*i blooms show this species to have significantly extended its range poleward as a result of increasingly ice-free summer conditions and rising temperatures[44]. *E. huxley*i now forms a regular part of the Arctic summer marine microbial community[44], but at the cost of Diatom abundance diminishing substantially during blooms[125]. Interestingly, the poleward ranges extension of *E. huxley*i is being led by only a few distinct morphotypes that appear best adapted to the changing polar habitat[44]. Further community level changes are likely as the extent of warming and sea ice loss in the Arctic increases. However, the interpretation of such reports over large spatial scales should be met with caution as different classes of phytoplankton may respond to changing environmental factors differently[137]. What is clearer is that the ecological effect of such changes will likely be significant. The timing of the spring bloom is expected to change which could decouple primary production patterns from annual animal cycles resulting in major shifts to food webs, impacts to higher trophic levels and biochemical processes, which may increase selection pressures even further[131]. Additionally, warming induced reductions of average community cell size will likely further impact food web structure, and result in lower carbon export[203]. The continued expansion of oceanic species into Polar Waters will compound these effects, impacting grazing communities and water biochemistry[44].

Current literature has begun to categorise Arctic microbial eukaryotic diversity, but few attempts have been made to predict how the structures of such communities may change in future. Temperate ocean currents, such as those in the North Atlantic, feed into the Arctic and carry with them temperate taxa which have the potential to become established in Arctic waters under favourable environmental conditions. Given the accelerating rate of environmental change currently occurring in the Arctic Ocean, attention needs to be given to directly comparing the communities found in temperate regions to those in Arctic waters using molecular techniques to achieve a better understanding of the susceptibility of native Arctic communities to increased “Atlantification” of Arctic Water, and predict how extant microbial eukaryotic communities may change.

This chapter reports on the application of Illumina MiSeq technology to analyse marine microbial eukaryote communities across five stations in the Norwegian Sea, which were resolved to cover a gradient of Polar Water influence (Chapter 3). This enables assessment of the susceptibility of these communities to predicted alterations of these factors under increased Atlantic Water influence within the Arctic by examining community partitioning and correlations to environmental factors. The findings presented in this chapter have significant implications for the impact of current climate change trends upon eukaryotic community assemblages.

5.2 Results

5.2.1 Sequence Analysis

7,533,558 raw paired-end sequences were recovered from the Illumina MiSeq sequencing effort (Table 5.1). The number of raw paired-end sequences recovered from individual stations varied from 572,698 (CTD56) to 1,438,424 (CTD10), with an average of 941,694 sequences per station. 155,662 (2.1%) sequences were lost during the pre-processing stage, resulting in a total of 3,611,117 merged paired-end sequences that were available for taxonomic assignment by BLAST. 26,367 sequences failed to match any know sequences in the SILVA database, and 26,699 were found to belong to singletons, the latter were removed from further analysis as they are potentially the result of sequence errors[158]. 3,584,418 (95.2%) merged paired-end sequences were retained after pre-processing and constituted the final processed dataset. Sequence number varied by station from 263,844 (CTD58) to 693,780 (CTD10), with an average of 448,052 merged paired-end sequences per station.

5.2.2 Rarefaction Analysis

Rarefaction curves did not indicate that sampling had reached saturation but did begin to plateau, as shown by the solid line (Figure 5.1). Extrapolated OTU richness curves (dashed lines) resolved that the estimated number of additional OTUs that would be recovered from twice the sampling effort is small, indicting sufficient sampling depth within the dataset and that the majority of eukaryotic OTUs had been recovered. Similar trends were observed for the whole dataset (Figure 5.1A), each station group (Figure 5.1B) and each individual station (Figure 5.1C).

Table 5.1. Sequence data metrics for the eukaryotic dataset.

Shown are the raw Illumina sequence file names generated by the Illumina MiSeq pair-end sequencing effort. Files are named by representative CTD sample with 18S donating the target gene during the probe assay, R1 donating files generated from the forward (5'-3) sequence run, and R2 donating those generated by the reverse (3'-5') sequence run. R1 and R2 were merged as part of the processing pipeline. Information for processing stages after the merger is stated in the R1 filename row. Raw sequence number shows the total number of sequences as returned by the Illumina MiSeq sequencing run before any processing. Sequences remaining after quality control gives the final sequence number for each sample file after merger of the R1 and R2 sequence runs and completion of the quality control processing stage of bioinformatics pipeline (Section 2.4.5) to improve sequence quality and that were used for analysis. Also shown are the number and percentage of initial sequence number lost during this step, and the minimum, maximum and average sequence lengths after pre-processing. Unassigned sequences lists the number of sequences for each sample file in the bacterial dataset for which no taxonomic annotation could be found, these are concatenated from annotations of "none" and "no blast hit". Number of singletons lists the number of sequences representing singletons for each sample.

File name	Raw sequence number	Sequences remaining after quality control	Number of reads lost	Percent reads lost (%)	Minimum sequence length (bp)	Maximum sequence length (bp)	Average sequence length (bp)	Unassigned sequences	Number of singletons
CTD08-18S_R1	496,311	477,926	18,385	3.7	100	270	127	1,687	3,328
CTD08-18S_R2	496,311								
CTD10-18S_R1	719,212	699,042	20,170	2.8	100	270	128	14,533	5,262
CTD10-18S_R2	719,212								
CTD12-18S_R1	518,172	501,325	16,847	3.25	100	270	127	741	3,574
CTD12-18S_R2	518,172								
CTD56-18S_R1	286,349	275,158	11,191	3.91	100	270	127	552	2,323
CTD56-18S_R2	286,349								
CTD57-18S_R1	340,516	329,189	11,327	3.33	100	270	127	459	1,901
CTD57-18S_R2	340,516								
CTD58-18S_R1	308,221	266,227	41,994	13.62	100	270	126	5,122	2,383
CTD58-18S_R2	308,221								
CTD59-18S_R1	585,360	565,767	19,593	3.35	100	270	127	1,778	4,138
CTD59-18S_R2	585,360								
CTD62-18S_R1	512,638	496,483	16,155	3.15	101	270	127	1,495	3,790
CTD62-18S_R2	512,638								

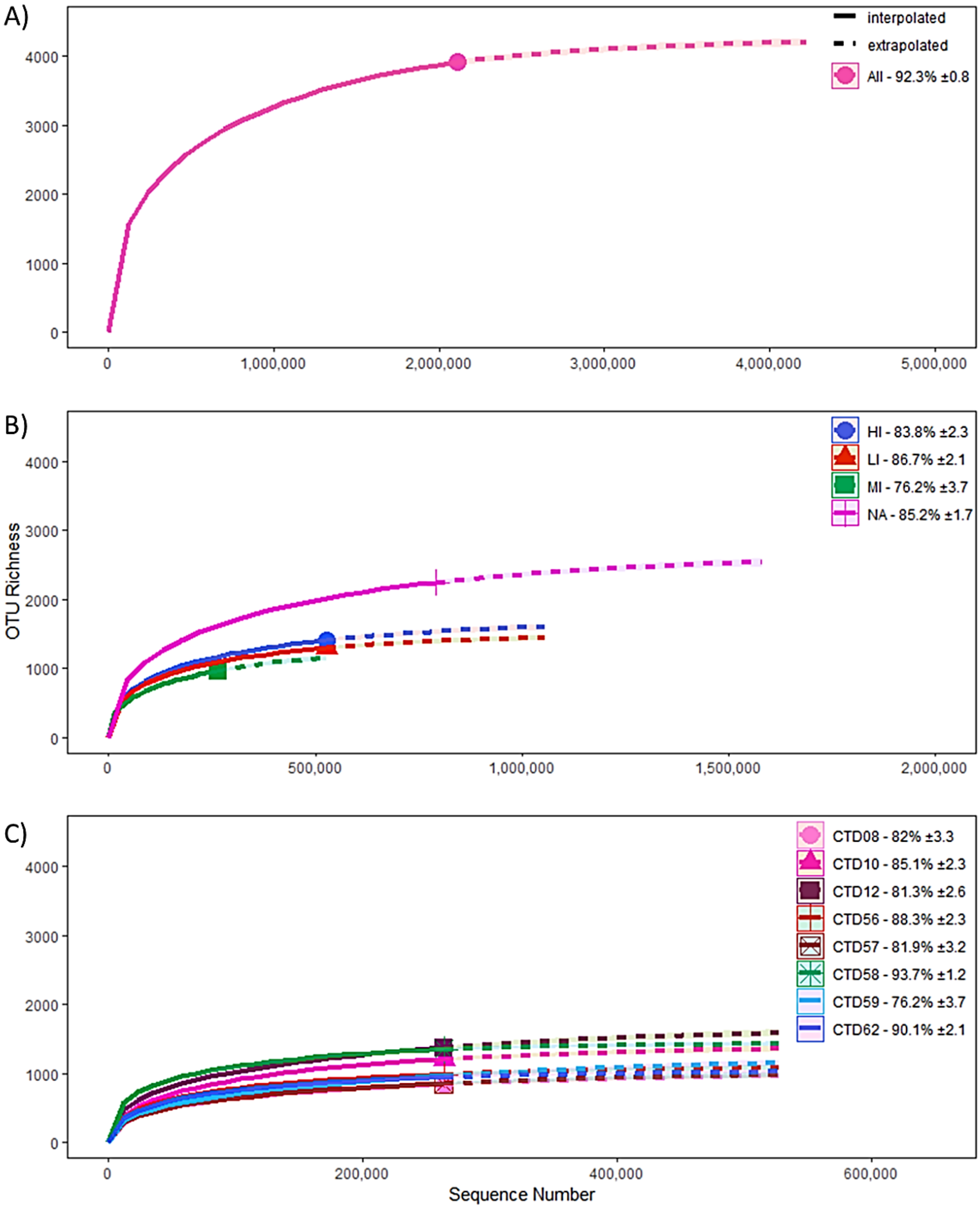


Figure 5.1. Rarefaction analysis of eukaryotic sequence depth.

Rarefaction (solid line) and extrapolation (dashed line) of OTU richness for the whole rarefied bacterial dataset (A), each station group (B) and each individual stations (C). Legend displays each group for which rarefaction was calculated, the % saturation of the recovered OTU number of the total number of estimated OTU number, and standard error for each group. Stations are labelled by the extent of Polar Water influence; blue – high Polar Water influenced stations (HI), green – moderately Polar Water influenced stations (MI), red – little Polar Water influence stations (LI), purple – Stations sampled in the North Atlantic.

5.2.3 α-diversity

A total of 2,558 unique OTUs were recovered across the transect from the rarefied dataset. Taxonomic identities could be assigned to 2,394 of these, meaning 164 novel OTUs with no previous representatives in the SILVA database (release 128) were also identified. Within the transect the highest number of OTUs were recovered from station CTD58 (1,344) which featured over half of the total number of OTUs present in the transect (Table 5.2). The lowest number of observed OTUs recovered from the transect were from CTD57, at just 894.

Analysis of the α-diversity metrics revealed that the Shannon diversity index was highest at the MI station, CTD58 (Table 5.2). Therefore, CTD58 featured OTUs with more even levels of abundance, as was validated by visual checks of the OTU tables. By contrast CTD57 featured the lowest level of evenness and thus contained more highly dominant taxa at the time of sampling.

The estimated OTU richness output by the "iNEXT" R package[169] and ACE revealed that CTD58 featured the highest estimated level of diversity and CTD57 the lowest. Across all stations the estimated figures are in general agreement with the actual observed OTU richness (Table 5.2), further supporting that the sequence depth was sufficient to recover the majority of OTUs from each station and that the diversity reported is reflective of the community.

Table 5.2. α-diversity of the eukaryotic communities recovered from each station.

Shown is the number of observed and calculated estimated number of OTUs recovered from each station, and standard error. Also shown are the Shannon diversity and Simpson diversity metrics. Stations groups reflect those determined during the regional assignment of stations in Chapter 3; HI – high Polar Water influenced stations, MI – moderately Polar Water influenced stations, LI – little Polar Water influence stations, NA – Stations sampled in the North Atlantic.

Diversity metric	CTD56	CTD57	CTD58	CTD59	CTD62	CTD08	CTD10	CTD12
Station group	LI	LI	MI	HI	HI	NA	NA	NA
Observed OTU richness	983	851	1,344	963	940	840	1,201	1,370
Estimated OTU richness (iNEXT)	1,113	1,039	1,434	1,263	1,043	1,024	1,411	1,685
iNEXT estimated S.E	25.26	33.69	17.26	46.58	21.43	34.11	32.93	43.52
Estimated OTU richness (ACE)	1,115	1,065	1,482	1,317	1,042	1,029	1,425	1,731
Shannon diversity	38.7	21.7	76.9	26.3	28.4	10.8	30.3	51.2
Simpson diversity	13.1	9.7	22.8	10.2	9.2	4	11.2	17.5

5.2.4 β-diversity

The partitioning of the MI station community in relation to the LI and HI communities reveals which community will become dominant under increased Atlantic Water influence within the sampled region. β-diversity analysis of the community structure (Figure 5.2) revealed that the whole eukaryotic community structure can be seen to partition into two distinct clusters (Figure 5.2A). The two LI stations (CTD56 and CTD57) are ~55% similar to each other, and are most dissimilar to the HI station (CTD62) found at the extreme right (only ~25% similarity to the other stations). The MI station was more similar to the HI station CTD59. By contrast the fractions of the community present at intermediate abundances (OTUs representing 0.01-1% of the community) (Figure 5.2E) and rare abundances (>0.01% of the community) (Figure 5.2F) partition with the MI station most similar to the LI stations.

The eukaryotic community partitioning was influenced by a number of highly abundant OTUs annotated as copepods, which had a strong influence on the structure of the abundant community members, here defined as OTUs constituting ≥1% of the community (Figure 5.2D). These copepod OTUs are likely the result of debris rather than intact whole organisms, as a result these taxa were removed from the dataset, resulting in a greater similarity of CTD59 to CTD62 (27% to 33%) (Figure 5.2C) as would be expected based on the similarity of environmental characteristics (Chapter 3). Hereon in only the eukaryotic dataset with copepods excluded is used in further analysis.

Validation of the observed partitioning by addition of stations CTD08, CTD10 and CTD12 collected in the North Atlantic maintained the existing partitioning, with these station being most similar to the LI stations, and most dissimilar to the HI stations, however, similarity across all the LI stations reduced to around 35% (Figure 5.2B).

Principle coordinate analysis (PCoA) revealed a total of 72.1% of the variance in the eukaryotic community was explained the first two coordinates (Figure 5.3), with 45.2% explained by PC1 and 26.9% by PC2. Pearson correlation of environmental data with PC1 and PC2 revealed temperature as the greatest explanatory environmental factor (p<0.05) (Table 5.3). Ammonium and phosphate were the strongest nutrient factors (p<0.05). No factors significantly correlated with PC2.

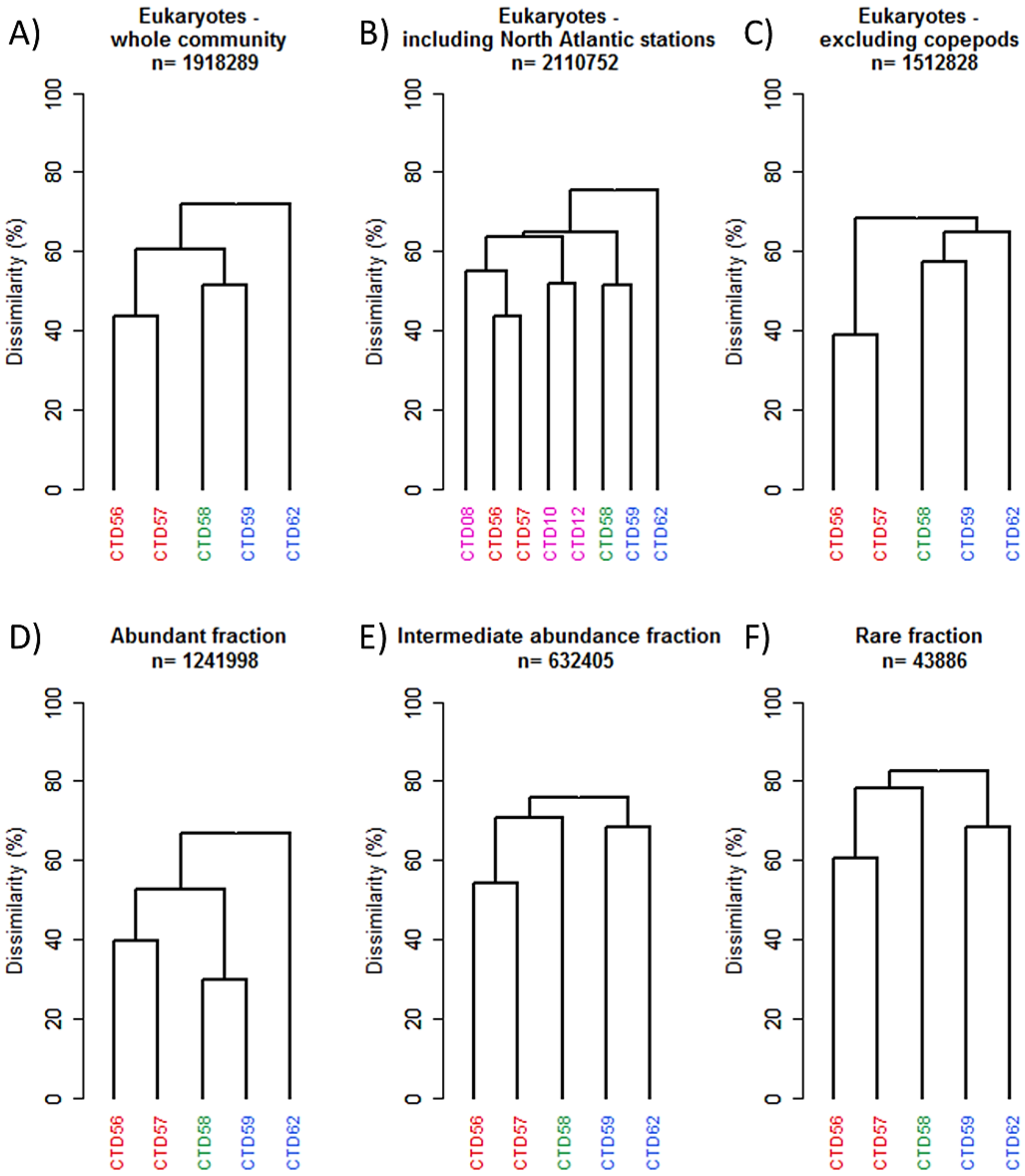


Figure 5.2. Eukaryotic β-diversity between stations.

β-diversity between stations was calculated by generation of Bray-Curtis dissimilarity matrixes and the degree of dissimilarity displayed by a dendrogram. Shown is the degree of dissimilarity between stations for the whole community (A), including stations from the North Atlantic (B) and the whole community with copepods excluded (C). Also shown is the degree of dissimilarity for separate abundance fractions of the community, the abundant fraction (OTUs representing ≥1% of the community) (D), the intermediate abundance fraction (0.01-1%) (E) and the rare fraction (≤0.01%) of the community (F). Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI, purple- stations sampled in the North Atlantic.

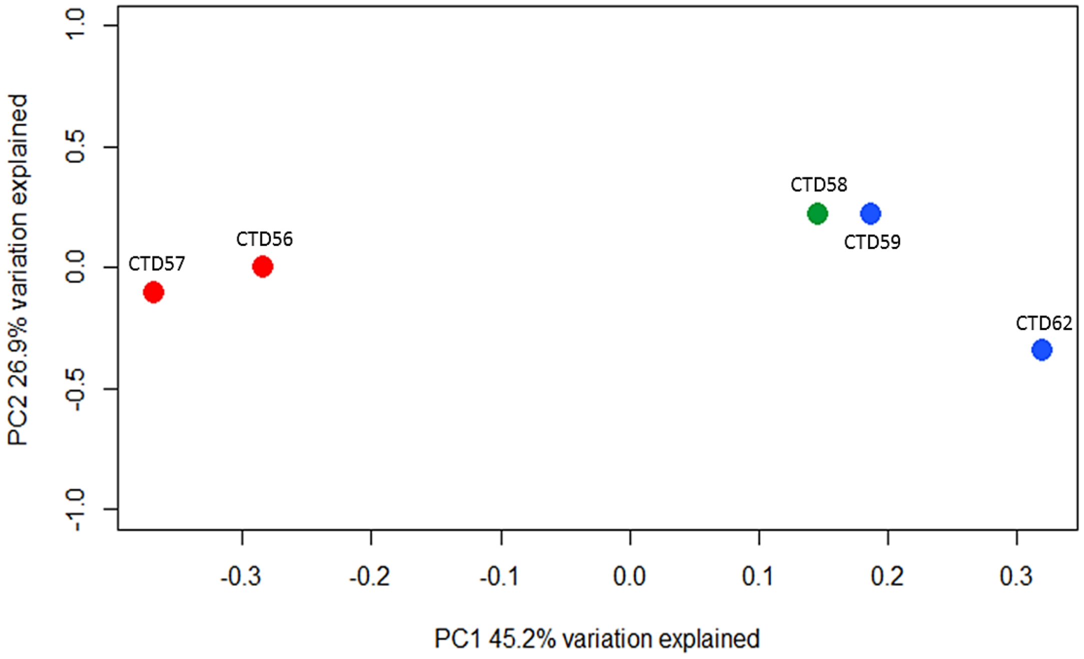


Figure 5.3. Principle coordinate analysis of the Bray-Curtis dissimilarity matrix of eukaryotic OTUs between stations.

72.1% of the variance in the eukaryotic community was explained by environmental variables. Stations primarily separated along PC1, the proportion of variance in the eukaryotic dataset explained by each coordinate is shown. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

Table 5.3. Pearson correlation of environmental factors to eukaryotic community variance.

The correlation of each *in situ* environmental factor to the first (PC1) and second (PC2) coordinate of the PCoA is shown, R represents the correlation coefficient. Negative R values indicate negative correlations, and positive R values indicate positive correlations. The significance of each correlation is shown by P, which represents the calculated p-value, * donates *p*≤0.05, ** donates *p*≤0.01.

Environmental factor	**PC1**		**PC2**	
	R	**P**	**R**	**P**
DO_2	0.74	0.15	0.19	0.76
Salinity	-0.84	0.08	0.27	0.66
SubSurVPAR	-0.83	0.08	0.31	0.61
Temperature	-0.89	*0.05	0.31	0.62
Ammonium	0.96	*0.01	0.02	0.97
Nitrate	0.84	0.07	0.23	0.72
Silicate	-0.61	0.28	0.07	0.91
Phosphate	0.93	*0.02	0.16	0.80

Analysis of the major eukaryotic taxonomic groups revealed that eight taxonomic groups; namely the Charophyta, Chlorophyta, Opisthokonta, Excavata, Dinoflagellata, Cryptophyta, Apicomplexa and Protalveolata displayed strong regional partitioning (similarity of 20-93% for the LI and 40-99% for the HI groups) (Figure 5.4).

Within four of these groups; the Charophyta, Chlorophyta, Opisthokonta and Excavata, the MI stations were most similarly to the LI stations implying that cold water associated species within the eukaryotic communities found at highly Polar Water influenced regions could potentially be displaced under predicted increases of Atlantic mixing and warming within the Arctic region (Figure 5.4A). For the remaining groups that displayed regional partitioning; the Dinoflagellata, Cryptophyta, Apicomplexa and Protalveolata, the MI station was most similar to the HI stations, which implies that cold water associated taxa within these taxonomic groups may dominate over warm adapted temperate communities under increased Atlantic Water influence.

The Rhizaria, Stramenopiles, Picozoa, Ciliophora, Haptophyta and Centrohelida displayed no clear regional partitioning (Figure 5.4B). Differences in the community structure of Amoebozoa between stations were not found due to featuring a low number of sequences and a complete absence of any representative sequences at CTD59 preventing analysis. Analysis was also not possible for Rhodophyta due to the recovery of only two representative OTUs.

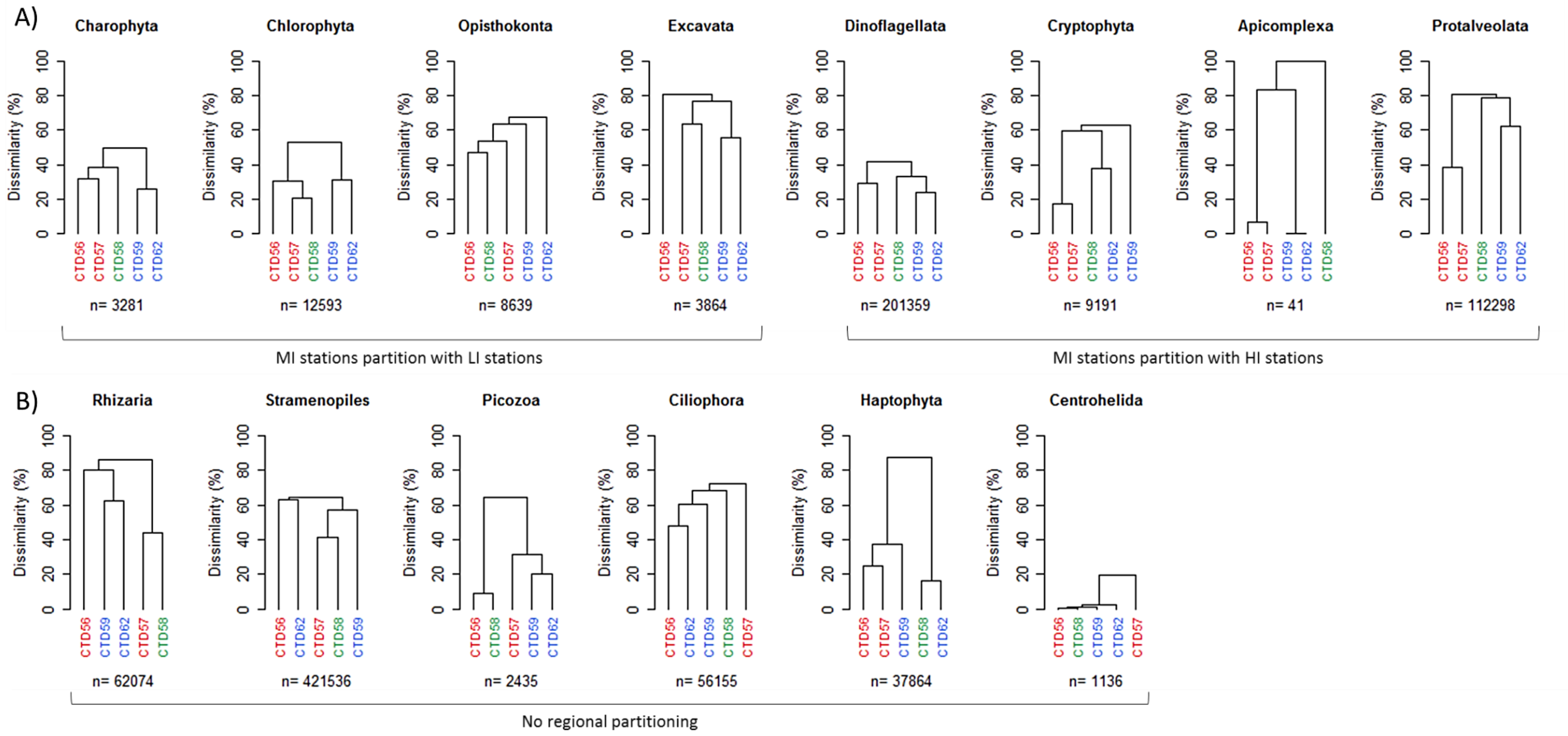
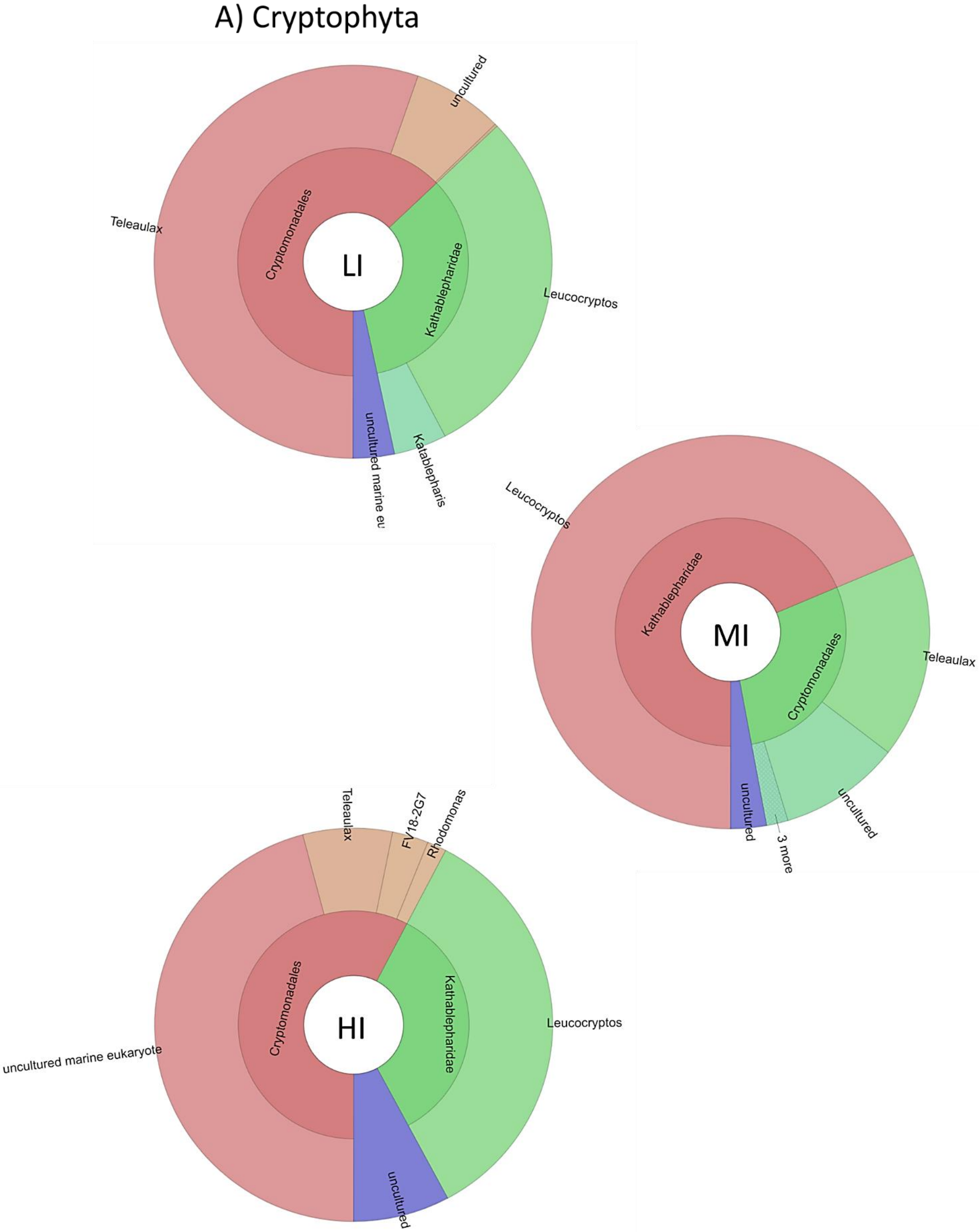


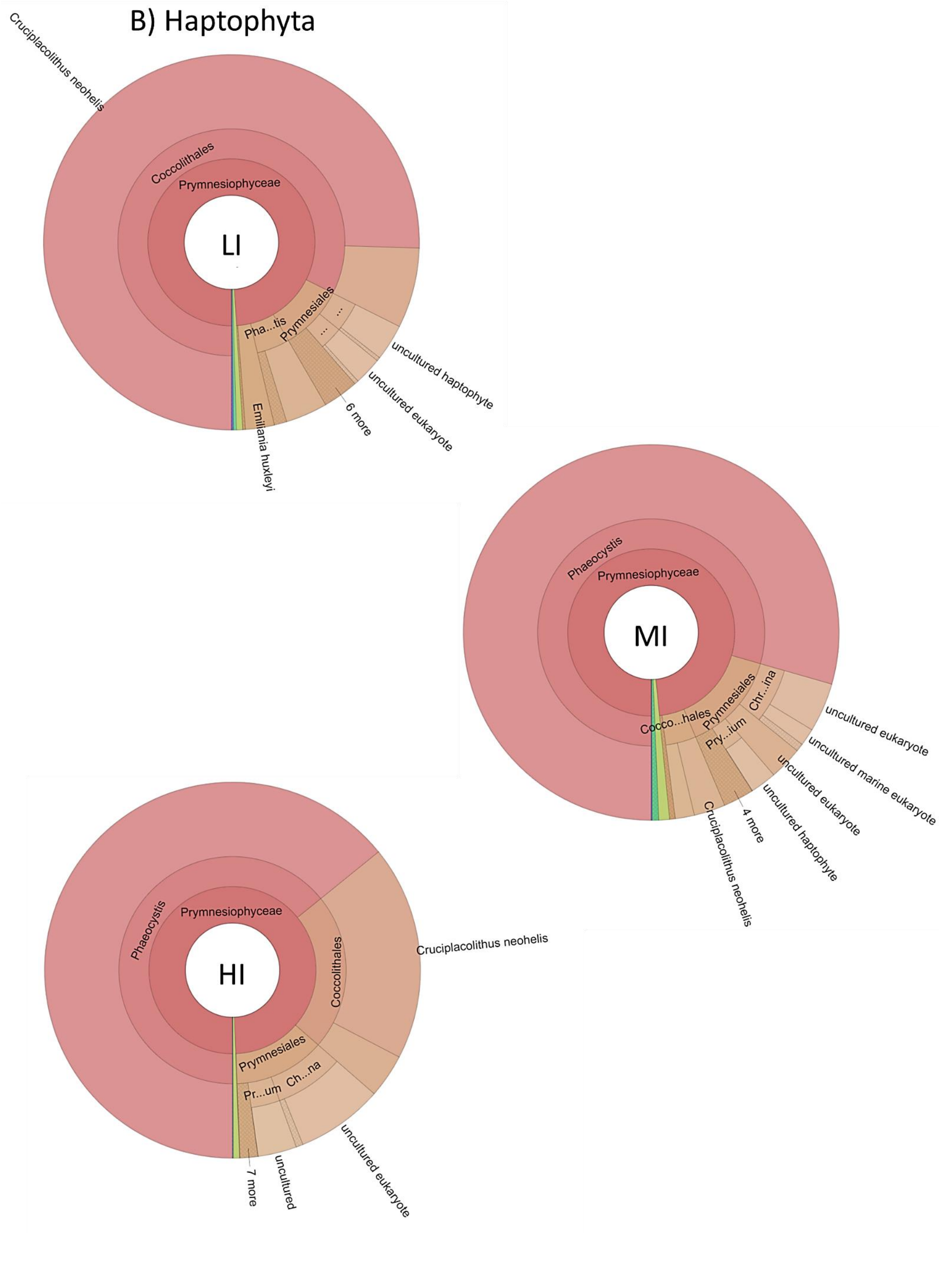
Figure 5.4. Regional partitioning of the representative major eukaryotic taxonomic groups.

Regional partitioning of the eukaryotic community divided into representative major taxonomic groups was based upon dendrograms of the Bray-Curtis dissimilarity matrix between stations. Shown are those groups which A) show regional partitioning which matched the regional assignment of stations as described in Chapter 3. Taxonomic groups displayed contrasting partitioning and are separated into those for which the MI station was most similar to the LI stations, and those where the MI station was most similar to the HI stations. Taxonomic groups are ordered by similarity of the LI to HI stations from most to least similar from left to right. B) Taxonomic groups which displayed no clear regional clustering, and so aren't ordered by similarity. Stations are coloured based upon the extent of Polar Water influence determined to be present at each station; red- LI, green- MI, blue- HI. The number of sequences comprising each taxonomic group is shown (n).

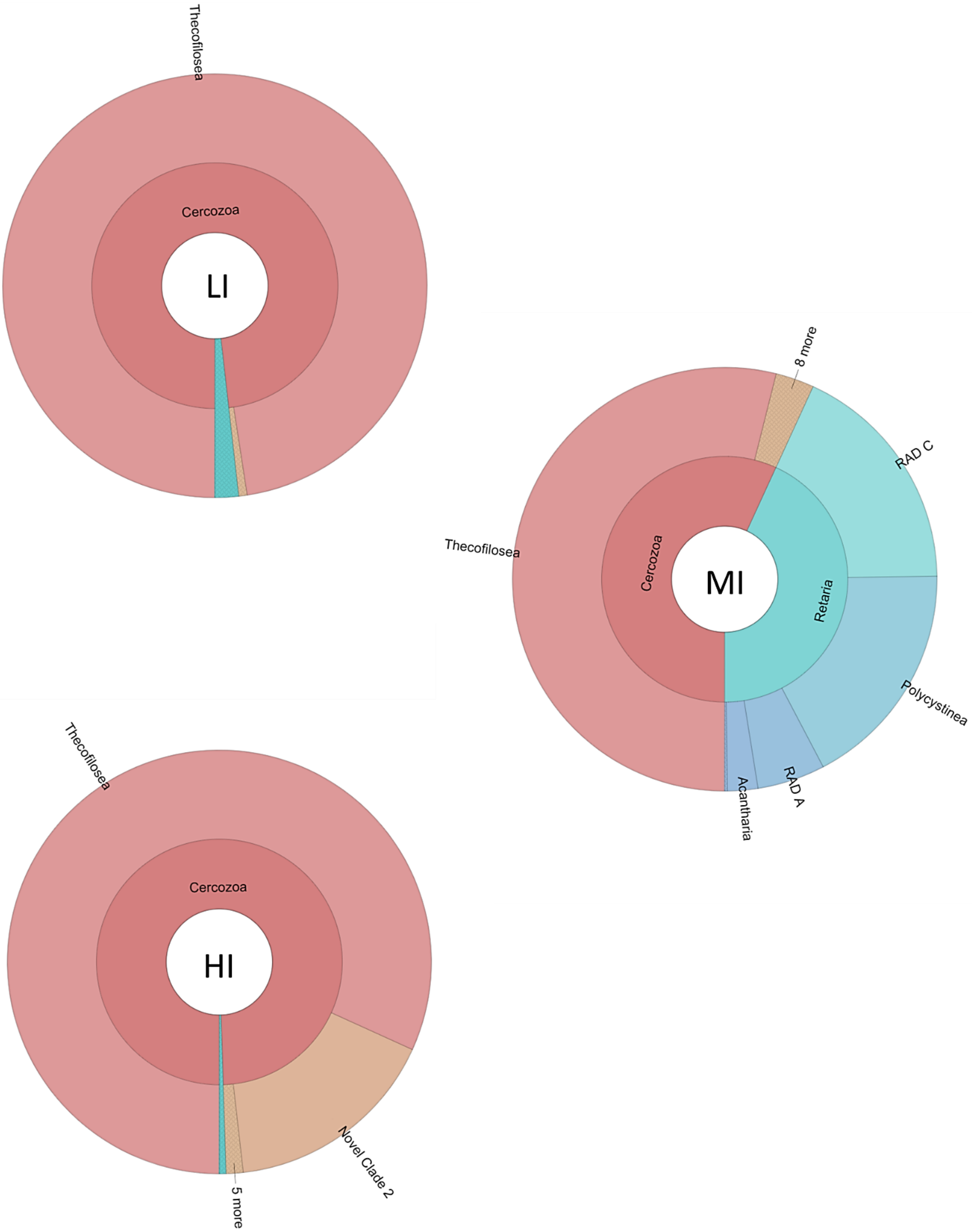
5.2.5 Community Composition



A) Cryptophyta


B) Haptophyta

C) Rhizaria



Thecofilosea
Cercozoa
LI
8 more
RAD C
Thecofilosea
Cercozoa
MI
Retaria
Polycystinea
RAD A
Acantharia
Thecofilosea
Cercozoa
HI
Novel Clade 2
5 more

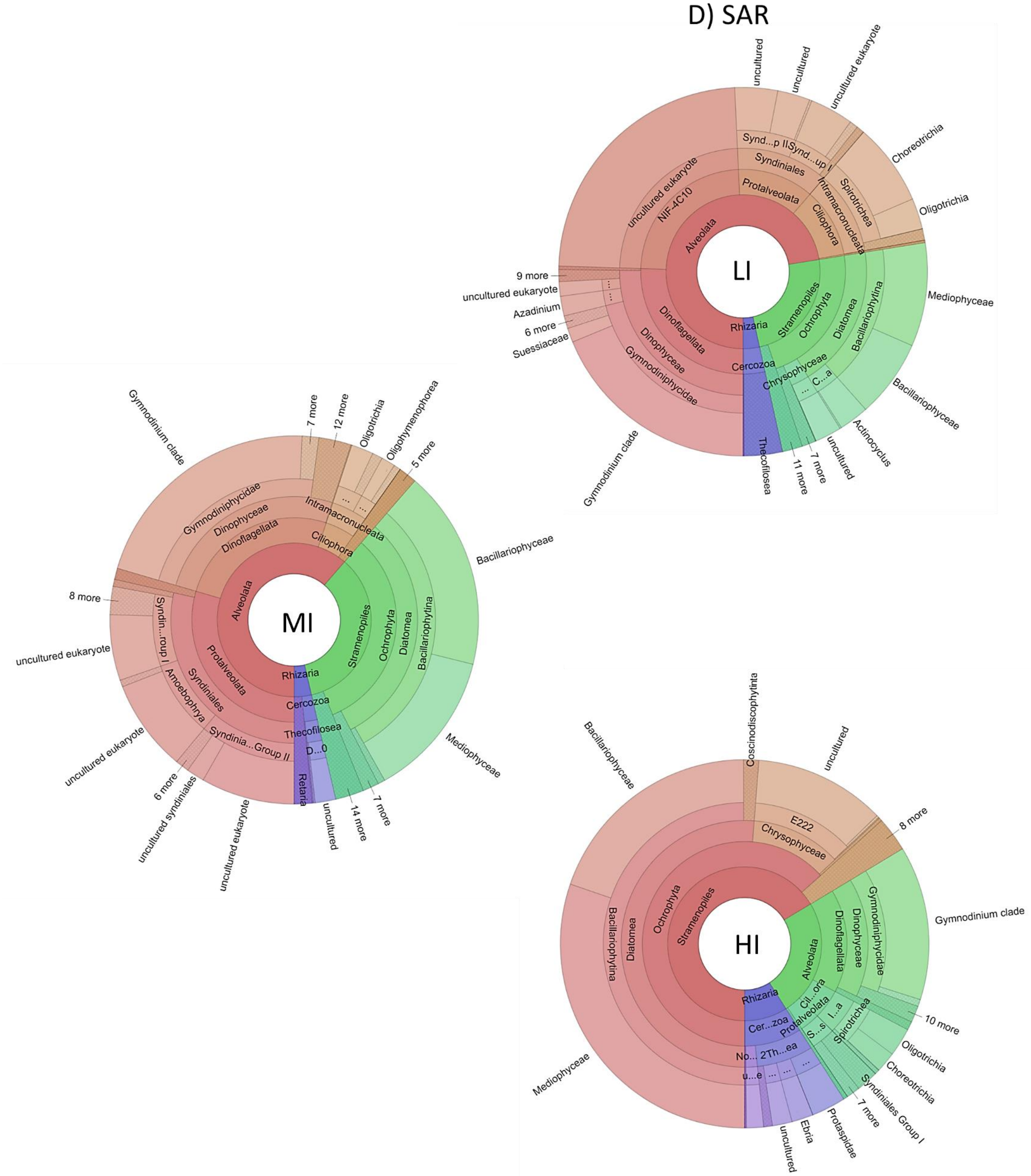


Figure 5.5. Differences in eukaryotic community composition between stations.

Shown are select examples to illustrate the difference in community compositon between stations under high Polar Water influence (HI), those under moderate Polar Water influence (MI), and those under low Polar Water influence (LI). Data is displayed for (A) the Cryptophyta, (B) the Haptophyta, (C) the Rhizaria, and (D) the SAR group.

Analysis of the community composition undertaken using Krona[171], revealed that the SAR group (Stramenopiles, Alveolates and Rhizaria) dominated the eukaryote community in terms of relative abundance. Dominance of SAR increased across the transect from 87% of the total community abundance within the LI group, to 94% in the HI group. Haptophytes displayed minimal variations in abundance between station groups, falling from 7% to 3% across the LI to HI station groups respectively. Archaeplastida constituted 2% of both the LI and MI communities, and 1% of the HI community. The Opisthokonta constituted ≤1% across all groups. Cryptophyta were present at 2% at the LI group, falling to <1% elsewhere. No other taxonomic groups at the same level were present at >1%.

Exploring these taxonomic groups at greater taxonomic depth revealed a number of compositional differences. The Cryptophyta featured approximately an equal compositional split between *Teleaulax* and *Leucocryptos* in the LI group, which changed to *Leucocryptos* and an unclassified Cryptophyta in the HI group (Figure 5.5A).

Despite a minor overall abundance change, the Haptophyta also displayed a change in the composition of constituent taxa across station groups. The LI group was primarily composed of *Cruciplacolithus neohelis* (5% total community abundance), which changed to *Phaeocystis* (2%) in the HI group (Figure 5.5B).

Exploring the SAR group in greater detail revealed other trends. The abundance of the Rhizaria remained low in both the LI and MI groups (3%), but peaked in the HI group of 8%. Thecofilosea was the main constituent of the Rhizaria (7%) across HI and LI groups, but the MI group displayed elevated levels of members of the Retaria (1%) (Figure 5.5C). The Alveolate and Stramenopiles featured different distribution patterns, the LI group was dominated by Alveolata (63%, falling to 23% at HI stations), and Stramenopiles dominated the HI group (21% in LI, rising to 62% in the HI group). Despite the aforementioned abundance changes of the Alveolata group the abundance of constituent taxa remained fairly constant. The presence of an unknown Alveolata noted as NIF-4C10 (21%) in the LI group and increase in Syndinales (26%) in the MI group, both of which were low in the HI group (≤3%), and increase of *Gymnodinium* in the HI group were the main explanation of the differences observed (Figure 5.5D). The observed differences in Stramenopile abundance across station groups were resolved be the result of an increased abundance of Diatoms from 16% in the LI group to 48% in the HI group, with much of this the result of an increase in

Mediophyceae (8% LI to 28% HI) (Figure 5.5D). The composition of the remaining representatives of Stramenopiles remained similar across the station groups.

5.2.6 OTU Distribution and Community Distinctness

Community distinctness and OTU distributions were compared by way of Venn diagrams. Analysis of the whole community revealed that each station had a number of station specific OTUs which were found only at each individual station (Figure 5.6A). CTD58, which displayed the highest level of recovered OTUs during α-diversity analysis, featured the highest number of station specific OTUs (445), and CTD57 the lowest number (139). The HI station CTD62 was compositionally the most distinct as it shared comparatively few OTUs with other stations. Analysis was repeated for all OTUs represented by ≥10 sequences at any one station (Figure 5.6B). Results indicated the presence of similar trends, suggesting that the station specific OTUs found only at each station were not simply rare OTUs represented by only a few sequences which may have possibly avoided detection at other stations, and that in fact each station harboured a diverse pool of unique taxa. The number of station specific OTUs at CTD58 remained much higher than at the other stations.

All stations were observed to feature a core community of OTUs common to all stations (Figure 5.6A). Analysis of the shared and specific OTU for the top 200 most abundant OTUs, (representing OTUs constituting 94% of the total community) revealed that the core community was mainly composed of the most abundant OTUs, and that very few abundant OTUs were found to be unique to individual stations (Figure 5.6C). The highest number of shared taxa outside of the core community for the top 200 most abundant OTUs was found to be shared between stations CTD56, CDT57, CTD58 and CTD59 further highlighting the uniqueness of the taxa at the HI station CTD62. Interestingly the only three OTUs in the top 200 most abundant found to be unique to a particular station were found to be present at CTD62, and all belonged to members of the SAR group, namely *Azadinium spinosum*, an uncultured representative of Syndiniales Group II, and an uncultured member of the *Chaetoceros* genus.

All stations shared fewer OTUs with stations at a greater distance, with two compositionally similar groups apparent either side of CTD58 (Figure 5.6D).

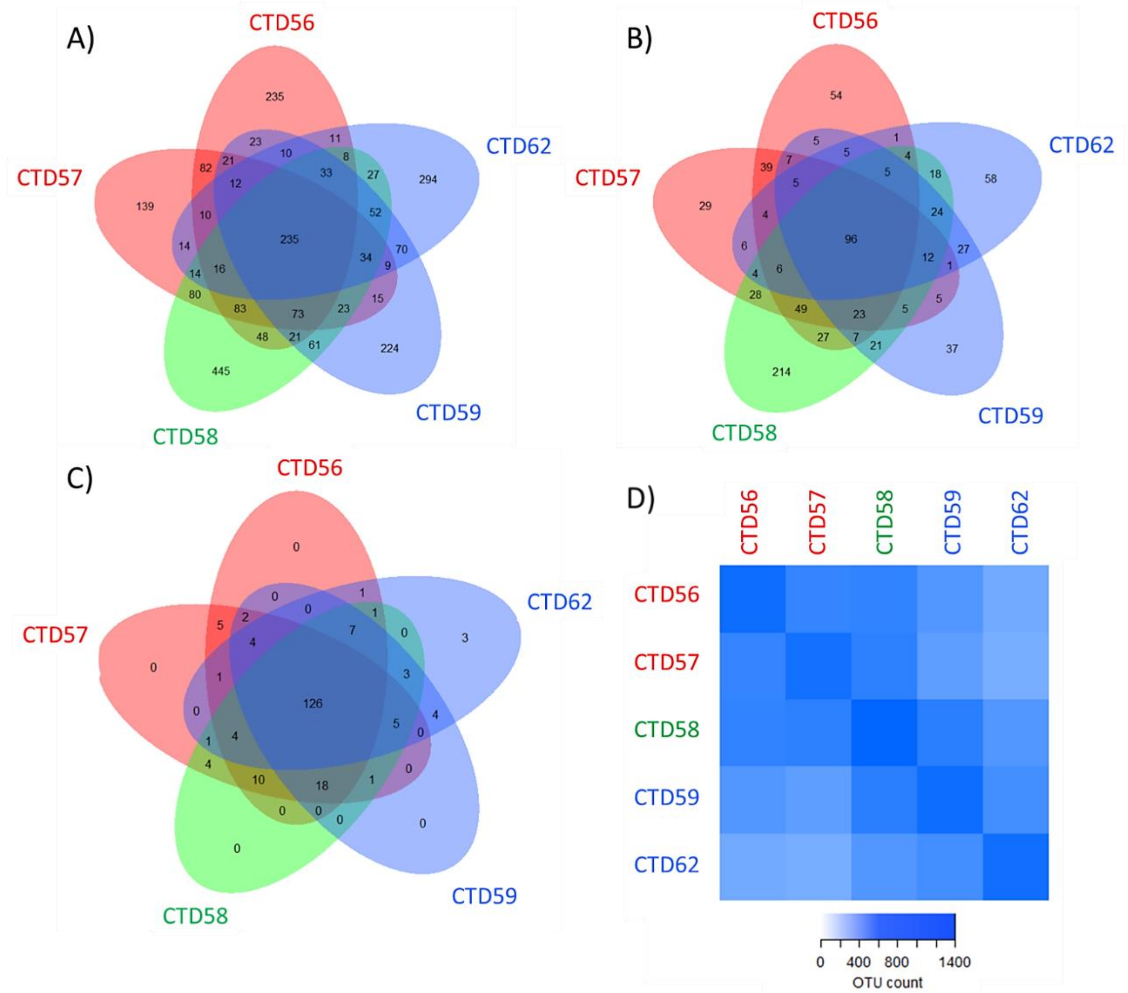


Figure 5.6. Analysis of eukaryotic community distinctness.

Shown are the shared and specific OTUs found within the transect. Numbers represent the count of OTUs. OTU counts in station ellipses which do not overlap with any other represent the number of OTUs specific to that station, whereas those that do overlap represent the number of OTUs found across those stations for which ellipses overlap. Community distinctness is shown for the whole community (A), as well as with those OTUs represented by >10 sequences at any one station (B), and the top 200 most abundant OTUs (C). Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI. Also shown is the number of OTUs shared between each pair of stations (D), with the number of OTUs represented by the colour scale, lighter shades indicate fewer OTUs.

To explore the distribution patterns of the top 200 most abundant OTUs in greater detail their relative abundance across stations was plotted. Clear abundance based distribution patterns were observed, OTUs were present across all stations at similar abundances (Figure 5.7A), or displayed a strong abundance bias at either the LI stations (CTD56 and CTD57) (Figure 5.7B), the MI station (CTD58) (Figure 5.7C), or the HI stations (CTD59 and CTD62) (Figure 5.7D).

OTUs displaying the abundance based distribution patterns described were found to show correlations with a number of environmental variables (*p*<0.05). Positive correlations with temperature were observed for OTUs showing greatest abundance at LI stations (R<0.9), and negative for OTUs showing greatest abundance for HI stations (R>-0.9). However, within the top 200 most abundant OTUs the highest number of correlations were seen with SubsurVPAR (Table 5.4) (full Spearman's rank correlation data available in Appendix 3).

Heat maps of abundance based distribution patterns were generated to visualise the distribution of OTUs which displayed an abundance pattern, referred to as "station biased" OTUs, across stations. Across the whole eukaryotic community 604 OTUs (Table 5.5) were observed to display an abundance bias (red), 314 of which were observed at CTD58 (Figure 5.8). Additionally, a high number of these OTUs were absent (black) at other stations. A number of OTUs were also seen to be absent at all stations except CTD62, similar patterns of absence were seen within the top 200 most abundant OTUs with patterns of absence observed at CTD56, CTD57, CTD58 and CTD59, as well as absence at both CTD62 and CTD59.

Heat maps were also plotted for eukaryotic OTUs grouped into their representative major taxonomic groups (Figure 5.8). Total OTU number varied between groups, but still provided a snapshot of the distribution patterns of station biased OTUs for each of the major taxonomic groups. 189 of the 223 Protalveolata station biased OTUs were observed at CTD58 (Table 5.5), with a number absent at all other stations (black) (Figure 5.8). Within the Charophyta and Excavata most station biased OTUs were observed at CTD58, being 12 and 6 respectively. The Chlorophyta, Haptophyta and Cryptophyta displayed the most station biased OTUs at CTD56, being 21, 13 and 6 respectively. The Rhizaria featured 20, and the Stramenopiles 31 station biased OTUs at CTD62. Representatives of the Ciliophora, Opisthokonta and Dinoflagellata featured a similar numbers of station biased OTUs across all stations. All other taxonomic groups featured few station biased OTUs.

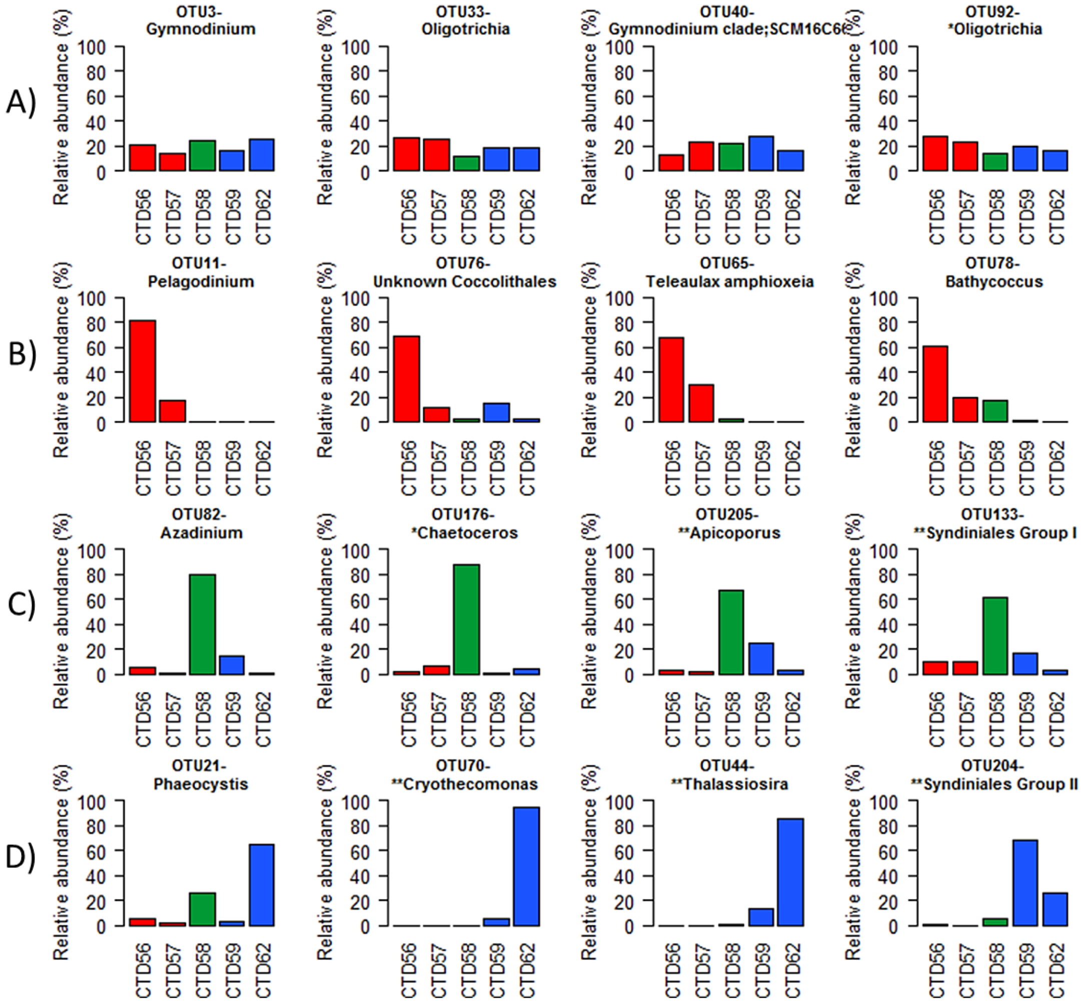


Figure 5.7. The relative abundance of eukaryotic OTUs between stations.

OTUs were selected from the top 200 most abundant in the eukaryotic dataset. OTUs show different distribution patterns which help explain the community assemblage structure and can be grouped into those which are present across all stations at relatively similar levels (A), show strong abundance bias for primarily LI stations (B), primarily MI stations (C), or primarily HI stations (D). Stations are coloured based upon the extent of Polar Water influence determined to be present at each station; red- LI, green- MI, blue- HI. ** indicates a significant association with temperature as revealed by Spearman's rank correlation with environmental conditions measured at the sampled depth for which water samples were taken. * Indicates a significant association with other environmental or chemical variables.

Table 5.4. The number of eukaryote OTUs showing correlations with each environmental factor.

In situ environmental measurements were taken from the depth at which environmental water samples were collected for sequence analysis during CTD casts for each station. Shown is the number of OTUs that displayed a significant ($p \leq 0.05$) correlation with each of the listed environmental variables.

Environmental factor	Number of correlated OTUs (p<0.05)
DO_2	32
Salinity	28
SubsurVPAR	38
Temperature	28
Ammonium	28
Nitrate	27
Silicate	24
Phosphate	21

Table 5.5. The number of eukaryotic OTUs deemed to display an abundance bias to a certain station.

OTUs that featured ≥60% relative abundance at a particular station were deemed to display an abundance bias to that station. Shown are the counts of OTUs deemed to display an abundance bias to a particular station for the whole community, and for each constituent taxonomic group. The total number of OTUs deemed to display an abundance bias within each taxonomic group are also shown.

	CTD56	CTD57	CTD58	CTD59	CTD62	Total
Whole community	101	47	314	60	82	604
Apicomplexa	0	0	1	0	0	1
Centrohelida	0	0	0	0	0	0
Charophyta	0	0	12	0	0	12
Chlorophyta	21	4	1	1	0	27
Ciliophora	5	14	12	8	9	48
Cryptophyta	6	0	0	2	0	8
Dinoflagellata	13	7	28	7	4	59
Excavata	0	0	6	0	3	9
Haptophyta	13	2	0	2	2	19
Opisthokonta	5	2	4	1	5	17
Picozoa	0	0	2	1	0	3
Protalveolata	14	6	189	12	2	223
Rhizaria	7	3	11	5	20	46
Stramenopiles	10	3	22	19	31	85

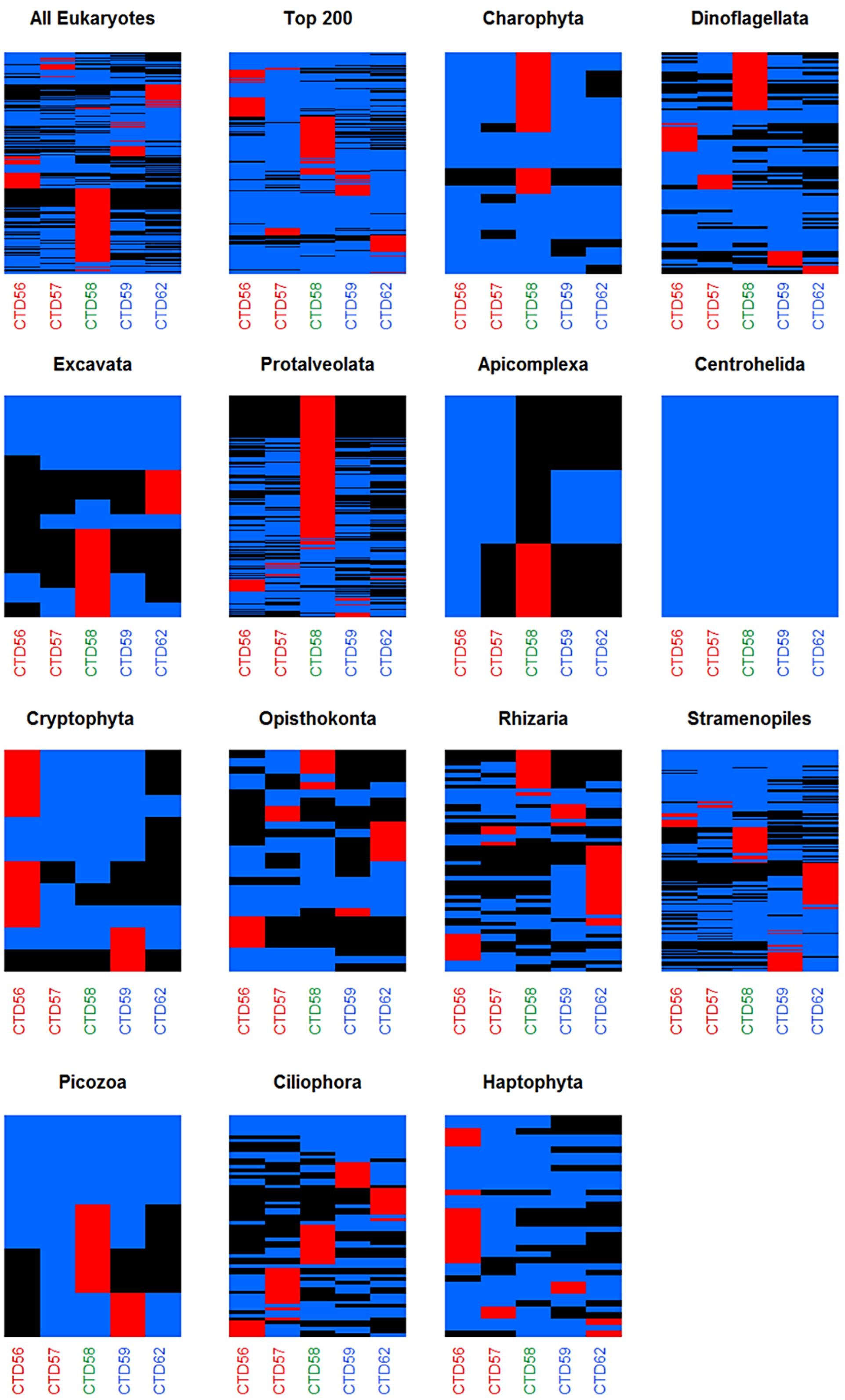


Figure 5.8. Heat maps of eukaryotic OTUs showing strong station abundance bias.

Shown is the distribution of OTUs deemed to display a station abundance bias within the whole eukaryotic dataset, the top 200 most abundant OTUs and for OTUs representing the major eukaryotic taxonomic groups. Black represents OTU absence at a particular station, blue represents OTUs that displayed 0-60% abundance at each station, and red represents OTUs that displayed ≥60% abundance at one station and thus were deemed to be display an abundance bias to that particular station. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

5.3 Discussion

This Chapter has explored the eukaryotic diversity present across a transect of five stations in the Norwegian Sea in order to provide insights into the potential susceptibility of pelagic marine microbial eukaryote assemblages to perturbations in the face of current environmental change. The sampled transect covered a gradient which followed the transition of water masses displaying varying amounts of influence from Polar Water.

5.3.1 Sequence Analysis and α-diversity

In total 2,558 OTUs were identified from the eukaryotic dataset (Table 5.2). Only 164 novel eukaryotic OTUs were identified that featured no prior representation in the SILVA database. The low number of novel sequences recovered contrasts with previous studies[42,271]. The primer combinations used were selected to recover a broad range of known taxonomic groups to reveal the dynamics of the entire eukaryotic community, but may still potentially have omitted some taxa, and so could be a contributing factor to low novel sequence recovery. Other contributing factors may include the smaller scale addressed compared to other global sequencing efforts, or the impact such large studies have had in increasing the current catalogue of diversity available in genomic databases[272].

Full saturation of environmental sequence data is rare in such ecological datasets due to the sampling effort required to extremely rare taxa[14]. However, sequence depth analysis (Figure 5.1) revealed near saturation was reached for the whole dataset, within each station group, and for each individual station. The near agreement of a number of diversity metrics to the actual number of OTUs recovered further supports that undersaturation of the eukaryotic dataset is unlikely (Table 5.2). Thus, strong confidence can be placed that the sampling effort was sufficient, and that the level of diversity recovered is reflective of the region.

The highest number of OTUs were recovered from the MI station, CTD58 (Table 5.2). This is unlikely to be solely the result of direct water mixing between stations as CTD58 also featured the highest number of unique station specific OTUs not found at any other stations (Figure 5.6). Shannon diversity was calculated to be highest at CTD58 and lowest at CTD57 (Table 5.2), indicating presence of a more even spread of abundances across OTUs relative to the other stations in the transect, and visual confirmation supported that CTD58 featured fewer highly abundant OTUs than the other stations. By contrast CTD57 featured the lowest level of evenness and thus contained more highly dominant taxa at the time of sampling. This likely explains the differences between the predicted OTU richness for both the iNEXT and ACE estimators at CTD57 and CTD58, which may overestimate OTU richness in communities featuring greater evenness.

5.3.2 Environmental Correlation

The factors driving the assemblage of microbial eukaryotes into distinct communities is still under debate[32,47,273,274]. However, the findings of an increasing number of molecular surveys support that microbial eukaryotes display some degree of dispersal limitation, which when coupled with environmental selection can give rise to geographically distinct populations[275]. Such distribution patterns have been reported for a number of microbial eukaryotic taxa that manifest either as differing patterns of abundance, as in the case of typically widely cosmopolitan taxa[137], or the existence of geographically restricted ecotypes as found for certain Ciliates[276] and Diatoms[97]. Despite a small number of examples focusing on individual taxa, studies pertaining to whole community assemblage structure in response to environmental conditions are lacking. The work presented in this chapter provides an insight into such patterns by supporting current evidence that environmental factors play a key role in the spatial variation reported for microbial taxa across the whole community[277]. Temperature (Table 5.3) was resolved to be the greatest explanatory factor of the eukaryotic community structure, resulting in distinct community partitioning between stations (Figure 5.2). The eukaryotic community structure was influenced by abundant representatives of Copepods (Figure 5.2D), likely resultant from fragments that passed through filtering, removal of which increased community similarity between stations (Figure 5.2C).

Phosphate is often suggested to be a growth limiting nutrient within phytoplankton communities, especially those within fresh water systems[278], but phosphate limitation of marine communities has also been demonstrated[279]. Nitrogen is typically considered to be the main growth limiting nutrient of marine phytoplankton[8], which may be preferentially uptaken in the form of ammonium rather than nitrate[280]. Therefore, that both ammonium and phosphate were also determined to be significant correlates with the observed community partitioning is largely expected (Table 5.3).

As discussed in section 4.3.2, the partitioning of the MI station in relation to the HI and LI stations provides an insight into which community may become dominant under increased Atlantic Water influence within the sampled region. The partitioning observed for the eukaryotic community was more complex than that observed for the bacterial community. Partitioning of the MI station with the LI stations suggested that, like the bacterial community, temperate water associated community members within the rare and intermediate eukaryotic fractions will potentially displace cold water associated communities found at the HI stations (Figure 5.2E and 5.2F) under increased Atlantic Water influence in the sampled region. Similar partitioning was observed for the major taxonomic groups Charophyta, Chlorophyta, Opisthokonta and Excavata (Figure 5.4A), suggesting that temperate water associated taxa may displace cold water associated taxa within these groups. However, contrasting partitioning was observed which implies that cold water associated community members within the abundant eukaryotic fraction (Figure 5.2D), and within the Dinoflagellata, Cryptophyta, Apicomplexa and Protalveolata may dominate over temperate associated taxa under conditions of mixing with Atlantic Waters (Figure 5.4A).

5.3.3 Potential Responses of Rare Taxa

It has previously been assumed that extreme environments harbour genetically distinct community assemblages consisting of highly divergent taxa, and that taxa of temperate origin constitute a more homogenous mix of similar taxa[97]. However, results indicated that even seemingly similar locales separated by relatively small geographic distances may contain a highly varied selection of unique taxa which displayed station specific presence/absence patterns (Figure 5.6). The majority of station specific taxa were recovered from the rare community, and their arrangement may be shaped by environmental processes rather than just being a random collection of taxa at the limit of their physiological tolerance[56]. Rare taxa are particularly at risk from local extinction, typically being the first members of a community to disappear due to lacking the high level of abundance required to drive high levels of random dispersal required to maintain a persistent presence within the community[222]. Instead rare taxa are suggested to display greater geographic isolation, placing the unique genetic heritage of members of the rare fraction at risk of loss[56].

The maintenance of key ecosystem functions such as oxygen evolution, carbon sequestration, degradation of organic matter and primary productivity is dependent upon the functional attributes of particular taxa. It has recently been demonstrated that rare taxa are more likely to feature distinct functional traits, which are not present in more abundant counterparts, and fulfil unique functional roles that cannot be provided by more common members of the community[222]. The potential displacement of rare taxa from HI stations (Figure 5.2F) would therefore have a significant effect on ecosystem functioning at both local and regional scales. The severity of this loss could be substantial as it has been shown that despite their low abundances certain rare taxa can have considerable metabolic actively, even exceeding that of more abundant counterparts, and so can disproportionally contribute to the maintenance of certain ecosystem functions[234] *(and references therein)*.

In cases where functional trait redundancy is present, rare taxa are hypothesised to act as a seed bank which aids in ecosystem resilience and recovery during environmental perturbations. These seed banks can exist as a large number of genetically similar variants which are adapted to slightly different environmental conditions. As environmental perturbations occur changes in the abundance of these variants occurs without an overall effect on ecosystem function. Indeed, compensatory reactions of rare phytoplankton taxa to environmental stressors have been evidenced to maintain core ecosystem functions under experimental conditions[281]. However, the window of tolerance for these different variants is likely to be small and unable to cope with larger environmental changes[55], leading to impacts such as regime shifts in taxonomic composition as evidenced in the Arctic following a record Sea Ice low in 2007[217]. Therefore, the loss of rare taxa could substantially increase the vulnerability of ecosystem services and their ability to recover after environmental fluctuations. The high number of station specific OTUs found at CTD58 (Figure 5.6A) may be such a seed bank, formed from opportunistic taxa ready to take periodic advantage of potential environmental variability at the MI station.

5.3.4 Potential Responses of Abundant Taxa

The majority of the 200 most abundant taxa in the eukaryotic dataset were observed to belong to a “core community” of OTUs found at all stations (Figure 5.6A). Yet, these still featured distinct correlations with environmental variables (Table 5.4), and appeared to display distinct abundance patterns with a preference for certain stations (Figure 5.7). The degree of these patterns is striking, with OTUs displaying high relative abundances for a certain station group and comparatively low relative abundances for the others. The abundance patterns appeared most pronounced for OTUs showing high abundances at the HI group (Figure 5.7D), which also featured the most distinct community assemblages (Figure 5.6D). The presence of such abundance patterns suggests that abundant members of the community are also susceptible to environmental change, particularly as all significant correlations with temperature for those OTUs most abundant at the HI stations were negative, and those most abundant at the LI stations were positive. The discovery of such correlations raises important concerns for the susceptibility of these OTUs under predicted future climate change[282], which has already been demonstrated to be resulting in measurable perturbations in polar communities[203,217]. Some correlations were observed between OTUs and multiple environmental factors, such as temperature and salinity, likely due to the simultaneous freshening and cooling resultant from increased Polar Water influence (Chapter 3). The highest number of correlations within the top 200 most abundant OTUs were observed for subsurface VPAR, but is unsurprising as the majority of these OTUs were either autotrophic taxa such as *Emiliania huxleyi, Azadinium, Teleaulax, Chaetoceros* and *Thalassiosira*, or heterotrophic predators or parasites known to associate with phytoplankton blooms, e.g. *Katablepharis*[283], *Leucocryptos*[284] and Syndiniales[285].

The presence of strong abundances of individual OTUs at certain stations suggests that individual taxa may have specific adaptions to a narrow range of environmental conditions. The extent and occurrence of such specificity in environmental communities remains largely unknown, but the data presented in this chapter highlights examples found within most of the major taxonomic groups (Figure 5.8). A number of studies have highlighted similar findings in the form of 'ecotypes' for particular taxa, for example, the cold adapted Diatom *Fragilariopsis cylindrus* has been shown to feature highly divergent alleles which display differential expression in response to environmental stresses which have enabled it to be particularly successful in Polar Waters[97]. A similar mechanism may be what is enabling the northward range extension of *Emiliania huxleyi*, especially given reports that only certain morphotypes are undergoing this move[44] and are reported to display sub-polar biogeographies[85]. More recently, highly resolved analysis has revealed ecologically significant variants within the clustered sequences of defined OTUs, which display differential patterns of succession and abundance in all three domains of life. These variants were most prevalent within temporally ubiquitous abundant taxa[58]. Such findings raises important questions about the ecological resilience of taxa, the resilience of the ecosystem services they provide, and the degree to which different variants are able to maintain ecosystem service levels under changing environmental stressors. They imply that environmental effects impact communities at fine scale of taxonomic resolution that is often overlooked, highlighting the importance revealing patterns and environmental correlations at highly resolved taxonomic levels.

5.3.5 Compositional Differences Between Station Communities

Compositional differences were resolved for the eukaryotic community at higher taxonomic levels between different station groups. Reduced abundance of Dinoflagellata and an uncultured Alveolata in the HI group relative to the LI group was observed, coupled with a corresponding increase in Diatom abundance. While such trends are typical of the respective water profiles, it has been shown that Diatoms display some of the strongest responses to warming[126], largely through changes in compositional variation rather than total abundance, and the timing of blooms across multiple taxonomic levels is changing under global climate change[286]. Changes in the magnitude and timing of bloom forming taxa will impact biochemical processes in the region. For example, Haptophytes are one of the most globally significant and intensely studied taxa, capable of forming blooms extending for many hundreds of kilometres. While overall levels of Haptophytes recovered in the eukaryotic dataset were low, this may be the result of sample timing[287]. Despite this, clear differences in the community composition between the LI and HI groups were observed, including a replacement of the Coccolithales *Cruciplacolithus neohelis* with *Phaeocystis,* for each group respectively. Coccolithales have a distinct functional role with significant impacts upon biogeochemical processes due to the formation of calcium carbonate liths on their outer surface[98]. The formation of these liths impacts the calcium cycle by removing carbonate ions from the water which reduces the alkalinity of surrounding waters, and in turn affects the ability of the surrounding water to uptake atmospheric CO_2. Additionally, the periodic shedding and sinking of liths is an important constituent of total carbon export[81,288]. *Phaeocystis* does not form calcium carbonate liths and so is unable to fulfil the same functional role or contribute to biogeochemical processes in the same way. Diatoms too make significant contributions to global primary production, the silica cycle[289] and carbon export in subsurface waters[290,291]. Haptophytes show a dramatic increase in photosynthetic rate with CO_2 increase which is set to rise under predicted future climate change and would lead to a displacement of the Diatom dominated assemblage found at HI stations. This suggestion is supported by previous experimental[292], and environmental studies during periods of reduced sea ice cover[217]. Such displacements will likely be amplified as environmental change continues to increase[293], which would act to further disfavour Diatoms for such smaller phytoplankton with lower rates of carbon and nutrient export. The consistent northward shift of Atlantic oceanic species, and their blooming within Arctic waters in relation to positive temperature anomalies is well documented, with temperate associated species displacing those of Arctic origin[125,126]. As the Arctic Oceans continue to be impacted by climate change Atlantic taxa will encroach further northward which has been demonstrated to reduce primary production and alter local community assemblage structure[294]. These affects may be particularly pronounced as changes in thermal stratification mediated by increased sea ice melt and increases to the growing season continue[295]. Other examples of such range extensions have been documented across various microbial eukaryotic taxa[126,287] and the number of additional examples is likely to increase in future. The changes in phytoplankton assemblages suggested would have further knock on effects across the Arctic food web, as they would likely result in changes to grazing assemblages featuring taxa which are sensitive to species composition or display size based feeding selectivity[296]. If Arctic waters do see a reduction in the size class of dominant phytoplankton then a move away from the comparatively large species of zooplankton which currently dominant Polar Waters, towards smaller zooplankton is likely. The resulting assemblage would be a less productive community negatively impacting further trophic levels[98].

5.3.6 Potential Winners and Losers

From the compositional differences between station groups highlighted throughout this chapter, and with the support of existing literature it is possible to make some general predictions as to how Polar associated eukaryotic communities may change under future climate driven warming and increased Atlantic Water influence in the sampled region.

Within the Excavata the HI group may experience increased abundance of the SCM38C39 marine group, and a corresponding reduction of uncultured *Neobodo* representatives. Data suggests that representatives of *Jakobida* may retain their current fraction of the community. SCM38C39 marine group appears to be an environmental isolate with little information available making commenting on this taxa further difficult. *Neobodo spp.* are commonly found in freshwater, and marine environments featuring low salinities, and have been suggested to display biogeographic distrubtions[297]. The patterns observed for this taxa therefore agree with expectations based on limited current literature.

The Chlorophyta displayed a clear difference in composition, with a potential transition from a *Micromonas* to *Bathycoccus* dominated community likely. Endemic Arctic *Micromonas* ecotypes have been reported to dominate late summer communities, and are suggested to be unable to survive warming conditions[261]. By contrast *Bathycoccus* is stated to be globally cosmopolitan[298].

A potential shift from *Phaeocystis* to *Cruciplacolithus neohelis* within the Haptophytes is also predicted, as well as a potential increase in total Haptophyta abundance, which has been previously been documented to be occurring in Arctic Waters[217].

The eukaryotic dataset was heavily dominated by members of the SAR group (which includes Stramenopiles, Rhizaria and Dinoflagellata), therefore some of most extensive predicted changes to the community are a potential increase in the abundance of uncultured Alveolata representative NIF-4C10, which was one of the most abundant OTUs recovered from the eukaryotic dataset. A potential decline in Stramenopile abundance in highly Polar Water influenced regions is also predicted, specifically for of members of the genus *Thalassiosira*, *Chaetoceros*, *Rhaphoneis* and *Fragilariopsis* which have been reported as dominant in Arctic waters[137]. Negative correlations with salinity and ice coverage have previously been demonstrated for Stramenopiles in Arctic waters, and decreased abundance has been observed after the 2007 record ice minimum[217] further supporting the suggested compositional changes.

Within the Dinoflagellata it is predicted that *Gynodinium* abundance may fall, and a number of uncultured *Gymnodinium* clade representative and *Azadinium* could increase. *Gynodinium* is already found in abundance in Arctic waters[17], but is also suggested to feature extensive habitat ranges and be highly competitive[299]. *Azadinium* is typically a temperate group, but more recently genetically distinct sub-Arctic variants have been identified[300].

The Protalveolata displayed clear compositional differences between the LI and HI groups, with the potential for a transition from a Group 1 to a Group 2 dominated community predicted. Furthermore, the diversity of both Groups 1 and 2 could reduce. Reductions of *Amoebophrya* are implied, but as this group contains parasitic taxa host responses may have a more significant impact on abundance[301].

5.4 Conclusion

The findings presented in this chapter have demonstrated that the microbial eukaryotic communities present within the Norwegian Sea display specific distribution patterns which correlate with environmental factors, with temperature found to be the most significant physical factor ($p \leq 0.05$). Associations with environmental factors were seen even at the level of individual OTUs, which showed clear preferences for certain station groups through abundance patterns.

The observed partitioning of the eukaryotic community was more complex than that of the bacterial community discussed in Chapter 4, and differed between abundance fractions and constituent major taxonomic groups. Partitioning suggests that temperate water associated community members may displace cold water associated members within the intermediate and rare abundance fractions of the community and for the Charophyta, Chlorophyta, Opisthokonta and Excavata under increased Atlantic Water influence in the sampled region. However, contrasting partitioning was observed which suggested that cold water associated taxa within the abundant community fraction, and within the taxonomic groups Dinoflagellata, Cryptophyta, Apicomplexa and Protalveolata may dominate over temperate associated communities under increased Atlantic Water influence in the region.

The findings presented in this chapter raise important concerns about the future of these communities under the consequences of climate change and highlight that they are potentially more susceptible to restructuring than previously thought. Stations under high Polar Water influence were observed to feature the most isolated taxa and OTUs with the strongest abundance patterns, thus it is this region which will likely undergo the most dramatic community shift. A number of experimental and field examples allowed reasonable speculations of compositional changes communities may experience to be made as supported by current literature. The community changes discussed in this chapter focus on a select few groups of comparatively low taxonomic resolution and for which key functional roles are known. However, similar impacts will likely be felt across a much wider range of taxonomic groups, at much finer taxonomic resolution, and throughout more complex webs of ecological interactions. As the identity of many environmental taxa and their functional roles remain largely unknown it is difficult to qualify and quantify the impacts environmental perturbations will have upon them. However, what is clear from existing examples of better studied taxa is that these impacts are likely to have wide reaching ecological effects, thus future monitoring and assessment of the Arctic region is imperative to validate the speculations suggested.

DNA barcoding techniques have been used to analyse a broad spectrum of taxonomic groups during the last two chapters, they have been a primary tool of molecular scientists exploring community composition and dynamics for a number of decades[62]. The strength of NGS methods lies in the ease of their recovery of a broad range of taxonomic groups through the use of universal primers covering both highly conserved and highly variable genetic regions[302], known as amplicon barcodes. Comparison of these isolated regions to a database of previously reported sequences held online can be used to quickly and easily determine the taxonomic identity of the organisms the barcodes were derived from. However, DNA barcoding analysis presents a number of limitations. The use of different genetic regions for the recovery of DNA barcodes can affect the recovered diversity or reported composition of communities[144,145]. DNA analysis will also detect free environmental DNA, or that of dead or metabolically inactive organisms, leading to false positive reports of taxonomic distributions and potentially misleading ecological significance[53]. However, the most significant limitation is that DNA analysis is not able to resolve information about the functionality of taxa. Therefore, DNA barcoding techniques can be thought of as a powerful tool to report on the presence of a particular organism or gene, but that is ill suited to the determination of functional information.

To confirm the functional viability of the communities previously sampled, and to resolve the potential long term ecosystem effects they may have, the recovery of functional information is required. Revealing the functionality of a community, and of the organisms within it allows the mechanistic understanding of the biochemical actions they are performing in their environment, and how they contribute to its maintenance[21]. Thus, functional profiling provides vital information about the functionality of ecosystem services within a region and is of particular importance within regions subject to environmental change.

An effective tool for determining community level functionality is that of metatranscriptomics. This technique involves the sequencing of the entire transcriptome from the whole community and is achieved through the amplification and sequencing of mRNA. As detailed in the central dogma of biology, described in section 1.5, mRNA represents a transcribed copy of the DNA compliment strand which is read by tRNAs and translated into polypeptide chains of amino acids to eventually form proteins (Figure 1.2). Sequencing of mRNA thus targets those genes which are actively being transcribed, allowing the collection of functional information[303]. Historically metatranscriptomics was confined to microarray analysis or cDNA clone libraries which were limited, by design, to certain known sequences in target organisms and incurred possible abundance biases[271]. The development and advancement of high-throughput NGS technologies has enabled the use of transcriptomic methods to recover the functionality of marine communities by comparisons against sequences stored in public databases, and characterisation of genes from uncultured organisms[71].

Chapter 6

Metatranscriptomic Analysis of Microbial Community Gradients along a Transect in the Norwegian Sea, in the Context of Climate Change

6.1 Introduction

The quantification of the abundance of functionally expressed genes and their relationship to environmental conditions is often a key aim of molecular studies in order to determine the functionality of microbial communities[304]. Such studies are common for terrestrial soil communities[305] and functional profiling in the human gut[306]. More recently, a number of studies have begun to address questions within the marine environment. A series of papers by A. Marchetti aimed to characterise the response of Diatoms to iron limitation, revealing multiple functional roles of the *ferritin* gene between different phylogenies, including high iron storage capacity and elevated expression in *Pseudo-nitzschia granii*[99], and the role of Proteorhodopsins in Diatoms potentially for additional ATP synthesis to promote survival under iron replete conditions[307]. On a community scale, exploration of bloom transcriptome dynamics has revealed potentially novel mechanisms for growth efficiency under carbon-limited conditions, as well as alterations to cellular surface molecules to alter adhesion, and that phylogeny may predict ecological roles across “boom” and “bust” bloom phases[308]. Other studies have revealed unique metabolic responses to simulated blooms of Haptophytes, Dinoflagellates and Diatoms that drive community dynamics, and that have been suggested to feature a dependence on physical and biogeochemical forcing that are susceptible to impacts under a changing climate[309]. The use of transcriptomics has enabled insights into how marine organisms cope with the variability of environmental conditions through changes in phytoplankton community structure and gene expression patterns, with different taxonomic groups observed to display different strategies that imply differing response timescales[74]. Additionally, diverse microbial groups may coordinate gene expression as a result of specific environmental cues to enable coupling of metabolic activity between species[310], providing information on functional networks of microbes acting together.

Despite the findings of such studies, the appropriate application and interpretation of transcriptomic methods within the context of quantifiable environmental factors such as nutrient concentrations, hydrographic processes, seasonal cycles and biological events such as blooms is still challenging. However, achieving a functional understanding of microbial communities is important in the context of climate change to elucidate the functional impacts it will have. This is particularly important at higher latitudes which can be seen as an early indicator for global change in the ocean[191] and are experiencing change above the global average rate[182], including range shifts of a number of taxonomic groups[125] and the displacement of endemic species[217].

Metatranscriptomic studies have revealed significant differences in the distribution and expression of genes within environmentally distinct habitats. Physical water characteristics have been suggested as significant descriptors of these differences[311]. These findings are supported by studies of well described taxa featuring known ecotypes, confirming differential expression of genes as an adaption to ecological drivers in colder environments[97]. Such gene expression profiles are likely to result in differences in the ecosystem functions carried out by members of the community. Examples of functional changes have already been evidenced, such as alterations to phytoplankton size classes[203] and increasing photosynthetic productivity in the Arctic[295]. While increasing photosynthetic productivity has so far been attributed to the decline in ice extent, there is evidence that warming directly impacts the resource allocation of phytoplankton metabolism through increased investment in photosynthetic pathways, with knock on impacts to key biogeochemical cycles, including nitrogen and phosphate cycling[311]. Despite these suggestions, the consequences of changes to ecosystem functions in marine systems remain largely unknown, and more studies considering a wide breadth of functional impacts are required.

In chapters 4 and 5 the microbial communities analysed were shown to display distinct partitioning across a transect of Polar Water influence in the Norwegian Sea. From the analysis presented and supporting literature it was possible to make suggestions as to how these communities may change under further predicted future environmental change within the region, but it was not possible to comment on the functional activity of the community. Here a shotgun metatranscriptomic approach is applied in an attempt to resolve metabolically active members of the eukaryotic community, and determine the functional profile of these communities in relation to the observed environmental gradient described in Chapter 3.

6.2 Results

6.2.1 Sequence Analysis

18,199,192 raw pair-end sequences were recovered from the Illumina MiSeq sequencing effort across all stations (Figure 3.2), including those collected from the North Atlantic (Table 6.1). The number of raw sequences recovered varied between stations from 1,972,220 at CTD62 to 2,722,240 at CTD57, with an average recovery of 2,274,900 sequences per station. A total of 9.1% of sequences were lost as a result of the quality filtering during the pre-processing stage as described in 2.6.2, during which paired-end sequences were left unmerged. 16,536,182 pair-end sequences were retained and available for further processing, ranging from 1,781,566 (CTD12) to 2,437,570 (CTD57), with an average of 2,067,022 per station.

The identity of genes represented by the recovered sequences can be found by annotating them directly, but reference genes or genomes of the target organism are required for accurate annotations limiting the use of this method for environmental samples[312]. *In silico* analysis represents a more suitable approach for taxonomically diverse environmental sequence data by assembling them into an overlapping series of sequences known as contigs (Figure 2.3) to effectively elongate them. Annotation was then applied to these contigs. DNAStar took 14,316,868 pair-end sequences from the entire sequencing effort for the assembly, and successfully assembled 1,492,778 sequences into 958 contigs which had an average length of 1322nt (range of 163-10561nt, Figure 6.1). The majority of contigs were identified as of eukaryotic origin, with few viral or bacterial contigs (Table 6.2). 267 of the 958 contigs were identified as novel, meaning no matches to existing sequences could be found (Figure 6.1).

Table 6.1. Sequence data metrics for the metatranscriptomic dataset.

Shown are the raw Illumina sequence file names generated by the Illumina MiSeq pair-end sequencing effort. Files are named by representative CTD sample (Figure 3.2), with R1 donating files generated from the forward (5'-3) sequence run, and R2 donating those generated by the reverse (3'-5') sequence run. Raw sequence number shows the total number of sequences as returned by the Illumina MiSeq sequencing run before any processing. Sequences remaining after quality control gives the final sequence number for each sample file completion of the quality control processing stage of the bioinformatics pipeline to improve sequence quality and that were used for further analysis. Also shown are the number and percentage of initial sequence number lost during this step, the minimum, maximum and average sequence lengths after pre-processing. Mapped sequences lists the number of sequences for each sample file in the dataset for which DNAStar was able to map onto the assembled contigs. Unmapped sequences lists the number of sequences for each sample file in the dataset for which DNAStar was unable to map onto the assembled contigs.

File name	Raw read number	Sequences remaining quality filter	Number of sequences lost	Sequences lost (%)	Mapped sequences	Unmapped sequences
CTD08_RNA_R1_001	1,021,779	943,086	78,693	7.7	265,709	1,620,532
CTD08_RNA_R2_001	1,021,779	943,086	78,693	7.7		
CTD10_RNA_R1_001	1,150,195	1,000,711	149,484	13.0	327,058	1,674,386
CTD10_RNA_R2_001	1,150,195	1,000,711	149,484	13.0		
CTD12_RNA_R1_001	1,000,107	890,783	109,324	10.9	292,269	1,489,324
CTD12_RNA_R2_001	1,000,107	890,783	109,324	10.9		
CTD56_RNA_R1_001	1,131,522	980,630	150,892	13.3	240,830	1,720,490
CTD56_RNA_R2_001	1,131,522	980,630	150,892	13.3		
CTD57_RNA_R1_001	1,361,120	1,218,785	142,335	10.5	820,065	1,617,574
CTD57_RNA_R2_001	1,361,120	1,218,785	142,335	10.5		
CTD58_RNA_R1_001	1,175,686	1,096,956	78,730	6.7	883,038	1,310,908
CTD58_RNA_R2_001	1,175,686	1,096,956	78,730	6.7		
CTD59_RNA_R1_001	1,273,077	1,206,809	66,268	5.2	529,488	1,884,185
CTD59_RNA_R2_001	1,273,077	1,206,809	66,268	5.2		
CTD62_RNA_R1_001	986,110	930,331	55,779	5.7	489,341	1,371,350
CTD62_RNA_R2_001	986,110	930,331	55,779	5.7		

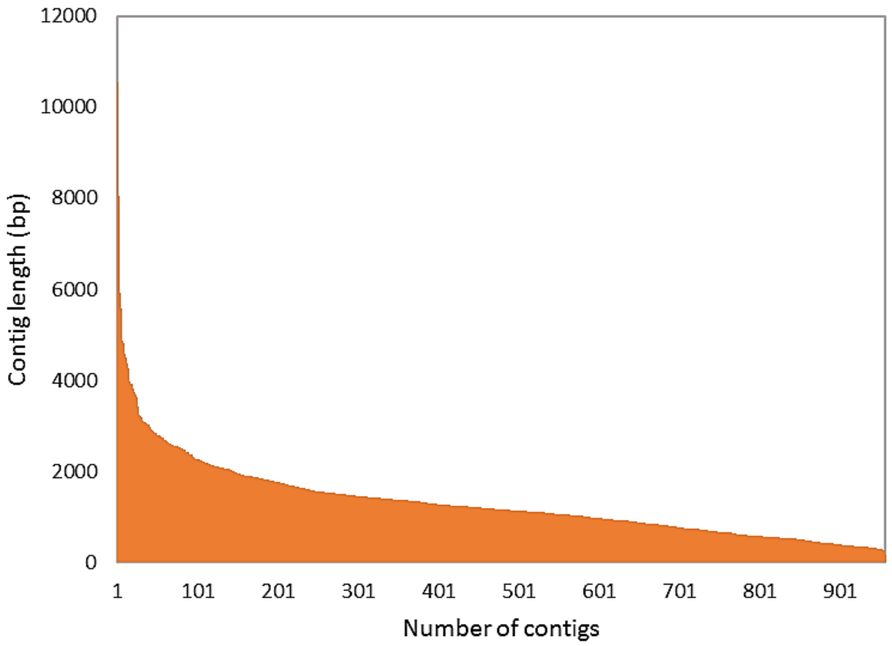


Figure 6.1. Contig length distribution curve.

Shown is the distribution of the contigs lengths for all 958 assembled contigs. X axis shows the number of contigs, the base pair length of which are shown on the Y axis.

Table 6.2. The domain of origin of each identified contig.

Shown is the number of contigs which were annotated as representing each of the three domains of life, as well as those for which no match to existing sequences, and thus no representative domain could be found, and were therefore annotated as novel.

Domain	Contig number
Viral	43
Bacterial	39
Eukaryotes	609
Novel	267
Total	**958**

A total of 3,847,798 sequences were successfully mapped onto the reference contigs, representing 26.9% of the dataset, ranging from 240,830 at CTD56 to 883,038 at CTD58 (Table 6.1). Average coverage of the mapped contigs for the whole dataset was calculated to be 649.36x (Figure 6.2A), meaning that on average each base had been sequenced 649.36 times. Average coverage was skewed to some degree by a few contigs with very large coverage values, but only 6 contigs had coverage less than 20.00x. Average coverage was calculated to be 699.04x for contigs identified as of eukaryotic origin (Figure 6.2B), and 370.26x for novel contigs (Figure 6.2C).

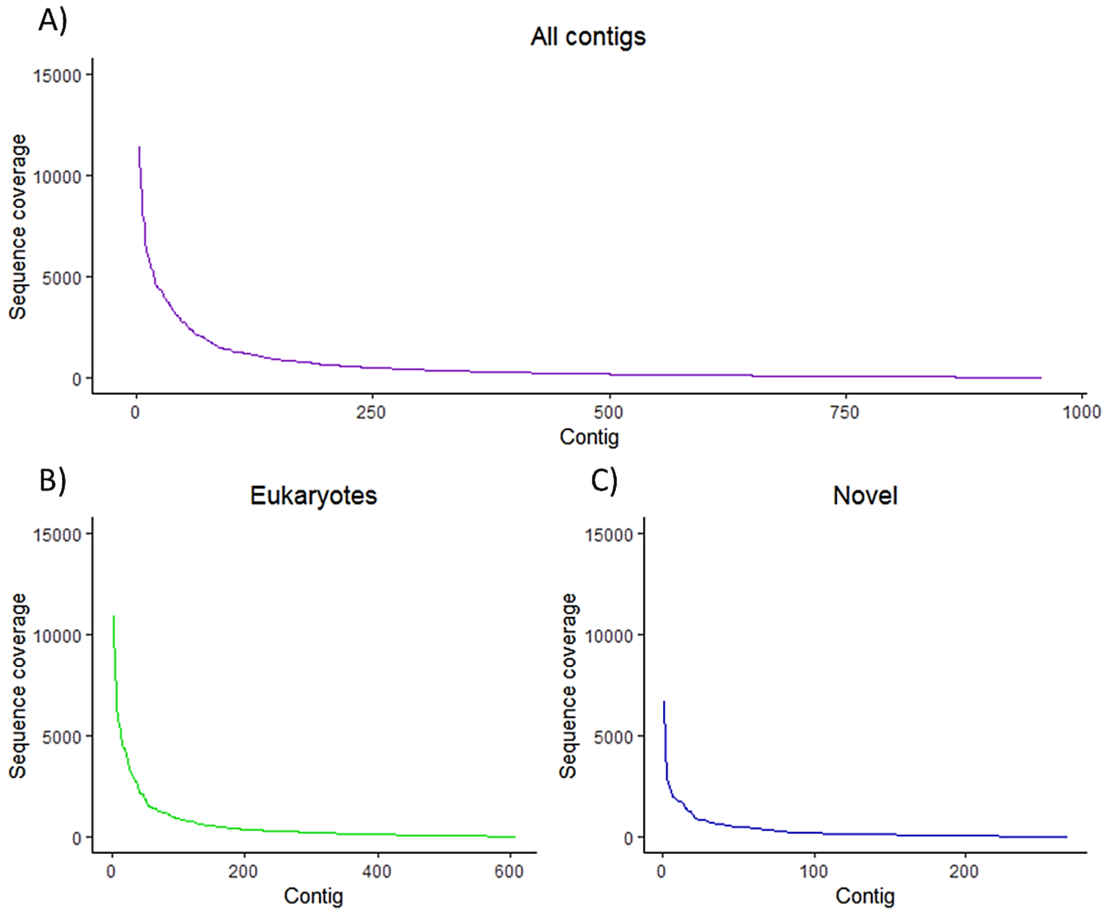


Figure 6.2. Sequence coverage of contigs.

(A) The sequence coverage of all contigs assembled by DNAStar. An average coverage of 649.36x was calculated for the whole dataset. Also shown is the contig coverage for subsets of the data, specifically contigs annotated as being of eukaryotic origin (B) and those annotated as novel (C).

74.1% of the mapped sequences matched known eukaryotic sequences, 3.5% bacterial sequences, 10.3% viral sequences, and 12.1% were annotated as novel (Table 6.3). CTD62 displayed the greatest proportion of novel matches, CTD57 and CTD58 featured elevated proportions of eukaryotic matches compared to the other stations, but reduced matches to all other groups. CTD56 featured elevated levels of viral sequences relative to other stations.

During RNA preparation oligo-dT primers, which select for polyadenylated (poly-A) tails, were used as part of the cDNA synthesis step. Poly-A tails are added to nuclear encoded mRNA in eukaryotes to act as a stabiliser and protect eukaryotic mRNA against degradation, however they have been shown to promote degradation in bacterial mRNA[313]. As such, this method specifically selects for eukaryotic derived

Table 6.3. The number of RNA sequences recovered from each domain of life.

Shown is the number of sequences from each station determined to originate from each of the three domains of life, namely eukaryotes, bacteria and viruses. Also shown are the number of sequences determined to originate from contigs annotated as novel.

Station	Eukaryotic sequences	Bacterial sequences	Viral sequences	Novel sequences
CTD-08	121,980	13,479	72,975	37,678
CTD-10	182,095	24,071	39,882	51,035
CTD-12	170,284	14,570	41,926	41,460
CTD-56	119,599	9,399	43,651	47,361
CTD-57	432,564	15,440	25,484	45,444
CTD-58	597,777	24,989	17,586	31,868
CTD-59	301,684	17,312	73,632	45,337
CTD-62	199,876	11,729	69,083	100,317

sequences and any recovered bacterial RNA may be degraded affecting the accuracy of taxonomic assignment, therefore all phage and bacterial sequences were excluded from further analysis. Novel sequences of unknown origin may also represent affected taxonomic groups and so were also excluded further analysis. 607 contigs annotated as belonging to eukaryotic organisms remained for further analysis (Table 6.4).

Table 6.4. The number of eukaryotic contigs by gene origin.

Shown are the number of eukaryotic contigs and how they were annotated during assembly. Annotated genes represents those which could be successfully assigned a gene name. Unnamed genes with descriptions represents genes which were not successfully annotated with a known gene name but were annotated with a description regarding their function or identity. Genes annotated as unnamed with no descriptions represent genes which were assembled but for which no gene name or any descriptive information regarding their function or identity was found.

Contig origin	Gene number
Chloroplast	7
Plastids	12
rRNA	143
mRNA	76
Hypothetical	255
Annotated genes	78
Unnamed genes with descriptions	20
Unnamed with no descriptions	16

Despite rRNA depletion using a Ribo-Zero Magnetic Kit (Illumina Inc., San Diego, CA, USA) during library preparation (section 2.3.2) 143 of the 607 eukaryotic contigs were annotated as genes originating from rRNA (Table 6.4). Ribo-Zero is highly efficient rRNA removal procedure, but complete removal of rRNA sequences is unlikely[314]. As rRNA is a constituent of ribosomes involved in the translation process rather than encoding genomic information for gene expression sequences annotated as originating from rRNA were removed from further analysis. 7 genes were annotated as chloroplast derived, and 12 genes as plastid derived. As the promotion of organelle derived mRNA degradation by the presence of poly-A tails is a possibility[315], these were also excluded.

255 of the remaining contigs were annotated as hypothetical genes. 16 assembled genes were annotated as unnamed and featured no description, these are likely the result of submissions of environmental isolates to the RefSeq database files for which little information exists. 78 genes were successfully annotated with gene names. Additionally, 20 genes were annotated which lacked a gene name, but for which annotated descriptions were present allowing some insight into their function. 76 genes were annotated as mRNA genes and also lacked a gene name, but did feature annotated descriptions allowing an insight into their function. 2 genes were removed due to featuring low absolute sequence count (<10 sequences). The filtered metatranscriptomic dataset was thus composed of 443 eukaryotic genes.

6.2.2 Community Composition

The relative abundance of taxonomic groups for all sequences were plotted to show the distribution of community members within the metatranscriptomic dataset across stations (Figure 6.3A). Stramenopiles constituted the most abundant taxonomic group at all stations (27-55%) except the LI station CTD57 (23%), where Rhodophyta was dominant (49%). Apicomplexa were also present at high abundance at CTD56 (24%). Other patterns of abundance were observed, genes annotated as Opisthokonta (11%), Amoeboza (2%), Excavata (5%), and Protalveolata (3%) were highest at the LI station CTD56, and reduced in abundance across the rest of the transect. Genes annotated as Apusozoa displayed an opposite trend, increasing in abundance across the transect from the LI (<1%) to HI (1% at CTD62) stations. Haptophyta increased in abundance from the LI station CTD56 (5%) to HI stations with a peak at CTD62 (13%) and slightly reduced abundance at CTD59 (3%) and CTD57 (3%), and genes annotated as originating from Ciliophora and Cryptophyta featured peak abundance at CTD57 (3% and 8%) respectively.

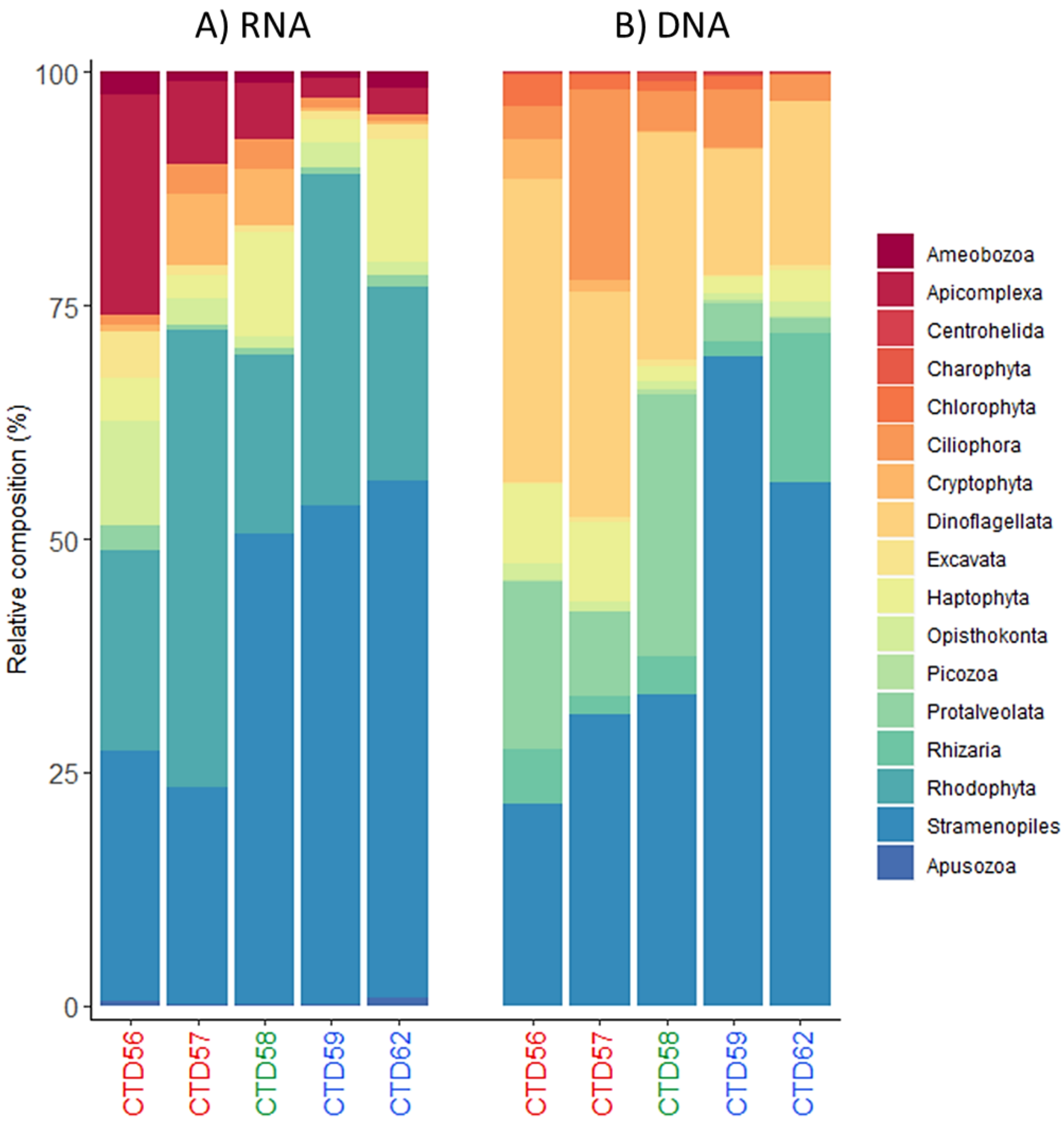


Figure 6.3. The relative abundance of representative taxonomic groups of recovered genes.

(A) The proportional abundance of major taxonomic groups at each station for all genes recovered during transcriptomic analysis which featured a taxonomic annotation, and (B) the taxonomic abundance of the eukaryotic community determined from DNA barcoding analysis as described in Chapter 5. All samples were normalised to the same sequence number per station. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

The proportion of sequences constituting each taxonomic group within the community displayed some discrepancies between the DNA and transcriptome datasets. Stramenopile sequences recovered in the transcriptome displayed a higher abundance at CTD58 (51%) than observed in the DNA sequence data (32%), however for the remaining stations the proportion of Strameopiles constituting the whole community was broadly similar between datasets (Figure 6.3B). The Rhodophyta, Opisthokonta, Ameoboza, Cryptophyta and Apicomplexa constituted a larger fraction of the transcriptome than observed in the DNA sequence data. The Opisthokonta and Apicomplexa displayed the most pronounced differences at CTD56 (11% vs 1%, and 24% vs <1% for each taxonomic group respectively between the transcriptome and DNA datasets), whereas the Rhodophyta featured much greater expression (19-49%) across all stations in the transcriptome dataset than the DNA dataset (<1% across all stations). By contrast the Protalveolata and Ciliophora constituted a smaller fraction of the community in the transcriptomic dataset. The Ciliophora mainly featured reduced abundance at CTD57 (3% vs 16% for the transcriptome and DNA datasets respectively), but the abundance of Protalveolata sequences was lower across all stations with the greatest differences observed at CTD58 (1% vs 27%) and CTD56 (3% vs 14%).

Compared to the DNA analysis (Figure 6.3B) no representatives of the groups Centrohelida, Charophyta, Chlorophyta, Dinoflagellata, or Picozoa were recovered from the transcriptomic dataset.

6.2.3 β-diversity

Multidimensional scaling (MDS) plots of the LogFC of genes within the transcriptomic samples resolved that the functional diversity separated into three groups that reflected the regional assignment of stations based on the physical environmental data as described in Chapter 3, similar to partitioning observed in Chapter 5. CTD62 and CTD59 form one group at the bottom right of the plot, separating from the second group comprised of CTD56 and CTD57 along dimensions 1 and 2. CTD58 forms the third group, separating from the HI stations primarily along dimension 1, and LI stations primarily along dimension 2. The observed grouping implies each regional station group featured distinct gene expression profiles, and therefore potentially distinct functional profiles. This may be the result of differences in gene composition between stations or differential expression of similar genes found across all stations. The gene expression profile of the MI station was distinct, but can be seen to be most similar to those of the LI stations, due to being within the closest proximity (Figure 6.4A). The existing grouping remained when stations in the North Atlantic were included, providing a further comparison against a temperate community (Figure 6.4B). The samples collected from the North Atlantic were most similar to the LI station CTD56.

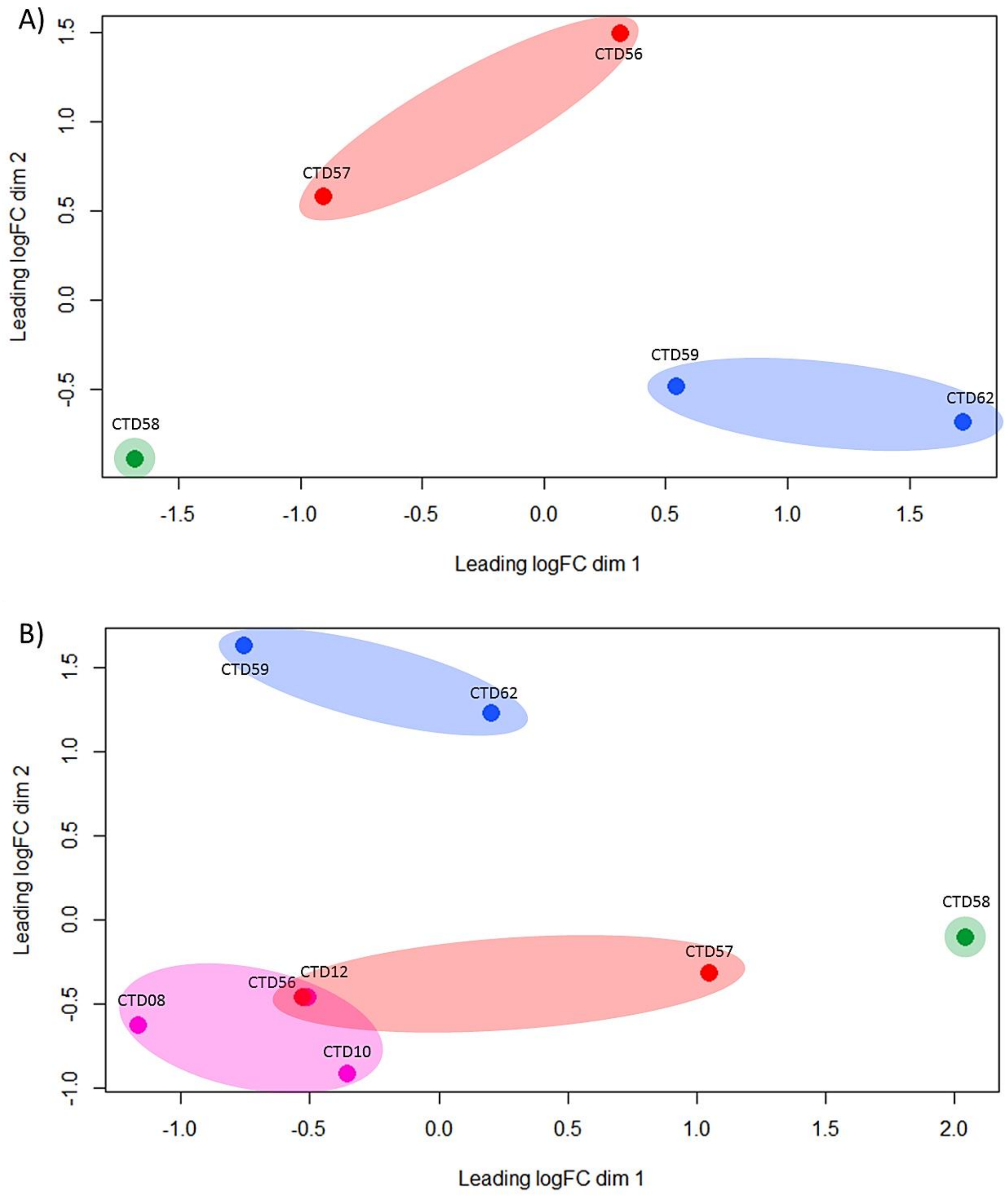


Figure 6.4. The logFC of gene expression profiles between each pair of stations.

Leading log-fold-change is the average of the largest absolute log-fold-change between each sample pair. Shown is the LogFC for the transect stations (A), as well as including the stations sampled in the North Atlantic (B). Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI, purple- stations sampled in the North Atlantic. Stations are grouped into three groups reflecting the regional groupings of the stations assigned during Chapter 3. Stations are normalised by library size using trimmed mean of M-values as detailed in section 2.7.2.

Venn diagrams resolved the observed partitioning not to be the result of distinct gene compositions at individual stations (Figure 6.5), and therefore likely the result of differential expression of compositionally similar profiles of genes found across all stations.

Principal coordinate analysis revealed a total of 68.7% of the β-diversity between stations was explained by the first two coordinates (Figure 6.6), with 38.9% explained by PC1 and 29.8% explained by PC2. Pearson correlation of environmental data with PC1 and PC2 revealed salinity as the greatest explanatory factors of the variance observed within the dataset (Table 6.5) (*p*=0.06). Nitrate significantly correlated with PC2 (*p*<0.05).

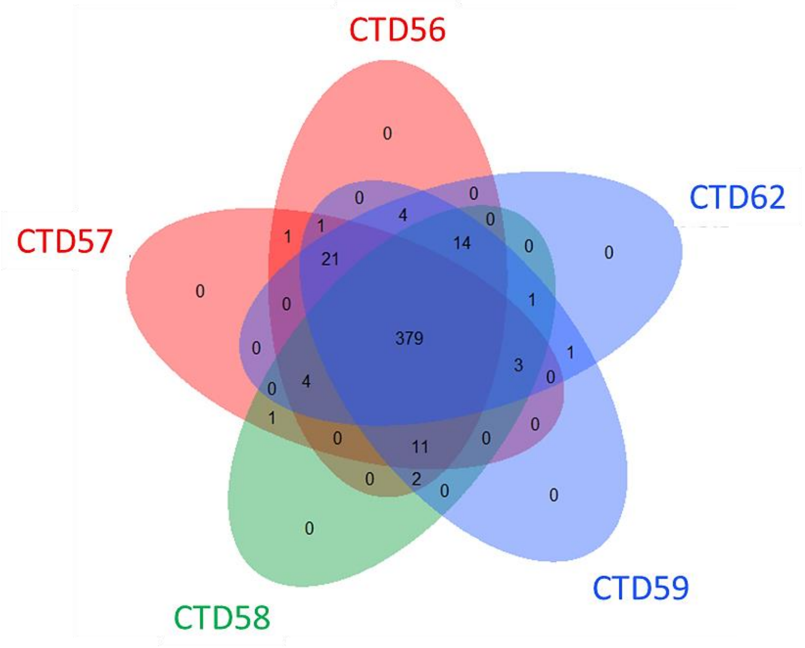


Figure 6.5. Analysis of genetic distinctness between stations.

Shown are the shared and specific genes found within the transect. Numbers represent the count of genes. Gene counts in station ellipses which do not overlap with any other represent the number of genes specific to that station, whereas those that do overlap represent the number of genes found across those stations for which ellipses overlap. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

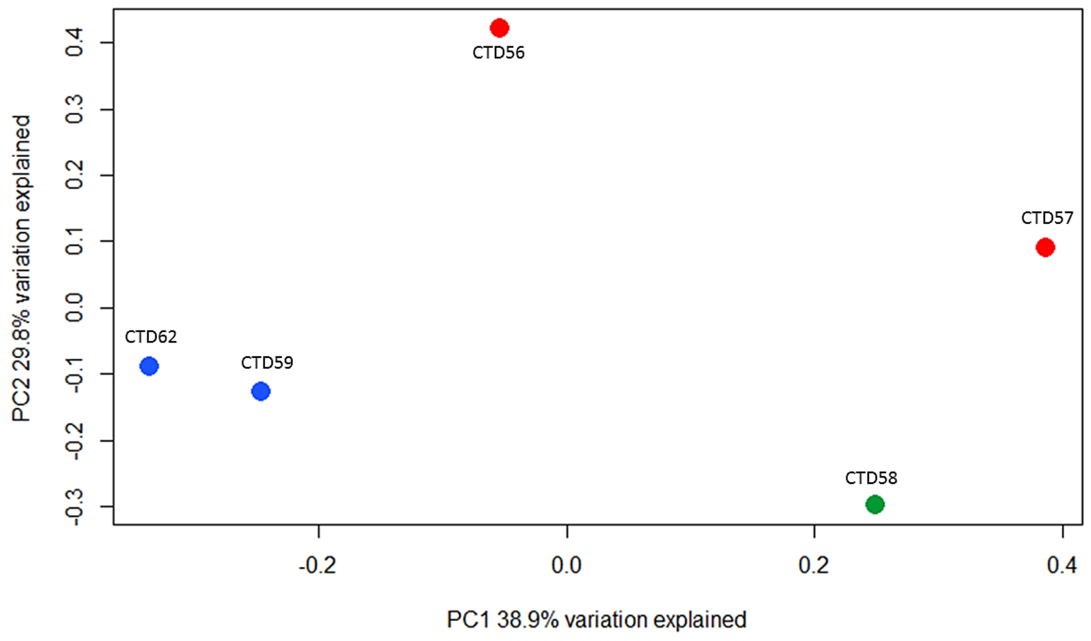


Figure 6.6. Principle coordinate analysis of the Bray-Curtis dissimilarity matrix of genes between stations.

68.7% of the variability of gene expression profiles between stations was explained by environmental factors. Stations primarily separated along PC1, the percentage of variance in the dataset explained by each axis is shown. Stations are coloured according to the degree of Polar Water influence, red- LI, green- MI, blue- HI. Stations were normalised to the same sequence number per station.

Table 6.5. Pearson correlation of environmental factors to metatranscriptomic community variance.

The correlation of each *in situ* environmental factor to the first (PC1) and second (PC2) coordinate of the PCoA is shown, R represents the correlation coefficient. Negative R values indicate negative correlations, and positive R values indicate positive correlations. The significance of each correlation is shown by P, which represents the calculated p-value, * donates $p \leq 0.05$.

	PC1		**PC2**	
	R	**P**	**R**	**P**
DO_2	-0.71	0.18	-0.50	0.39
Salinity	0.86	0.06	0.44	0.46
SubSurVPAR	0.31	0.62	0.70	0.19
Temperature	0.82	0.09	0.51	0.38
Ammonium	-0.66	0.23	-0.72	0.17
Nitrate	-0.12	0.85	-0.91	*0.03
Silicate	0.73	0.16	0.30	0.62
Phosphate	-0.34	0.57	-0.87	0.06

6.2.4 Gene Expression Profiles

The relative expression (as percentage expression) for each rarefied annotated gene across the transect was plotted as a heat map (Figure 6.7) to provide highly resolved information about the expression of individual genes within the community that might explain the transcriptome partitioning observed in section 6.2.3 Also included were genes annotated as originating from mRNA, and those lacking a gene name but which were annotated with a description because the annotated description contained information about their potential function. Heat maps resolved genes into 9 clusters. Cluster 1 (composed of 6 genes) was most highly expressed at CTD57. Cluster 2 (15 genes) and 3 (29 genes) were most highly expressed at CTD62, but cluster 2 also displayed expression at CTD59. Cluster 6 (27 genes) was most highly expressed at CTD56. Cluster 9 (8 genes) was most highly expressed at the MI station, CTD58. Cluster 4 (19 genes) and 5 (33 genes) displayed the highest expression at both HI stations. Cluster 7 (29 genes) featured more consistent levels of expression across all stations, with slightly higher expression at CTD56 and CTD59. Cluster 8 (8 genes) was most highly expressed at CTD58 and CTD59. The observed clustering implies distinct subsets of gene expression at individual stations, namely those within clusters 1, 2, 3 and 6, agreeing with the partitioning observed in section 6.2.3.

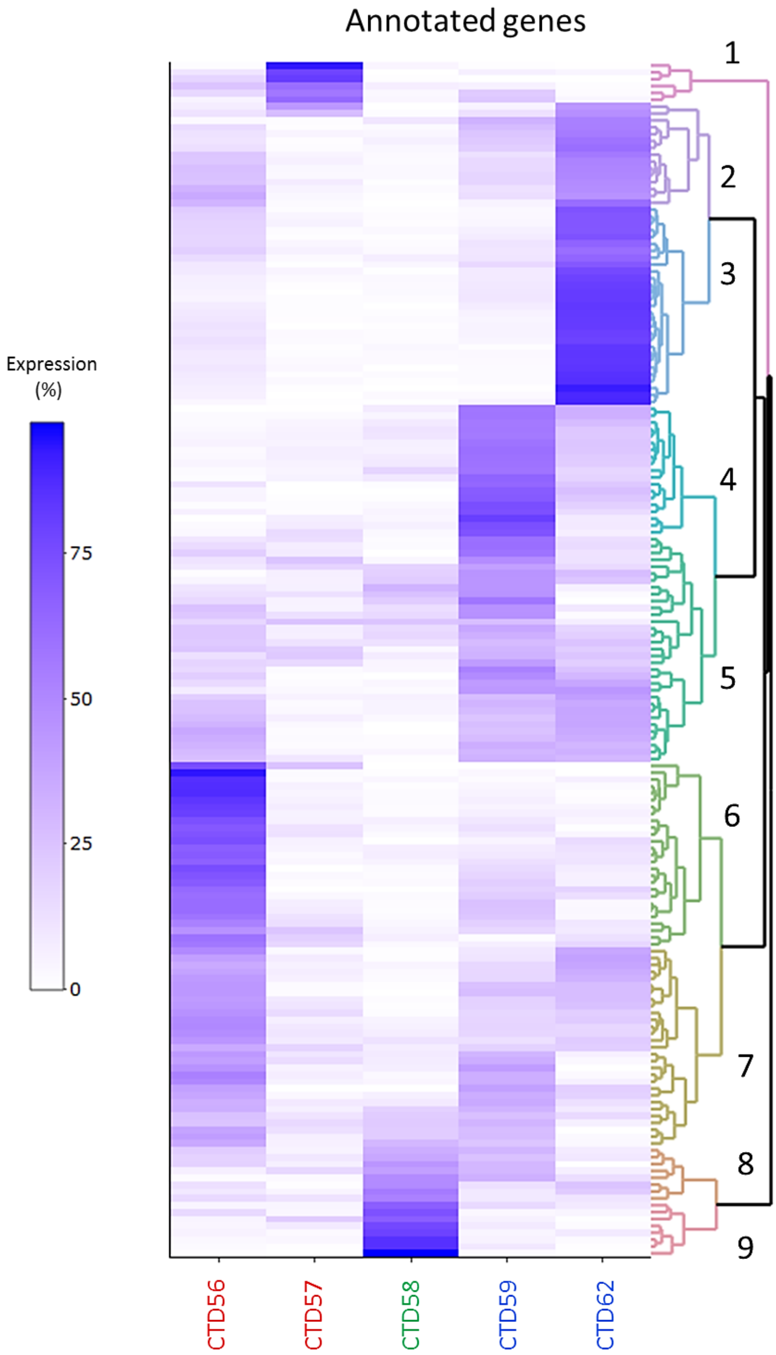


Figure 6.7. The relative expression of annotated genes by station.

Shown is the percentage gene expression across the transect for those genes successfully annotated with known gene names. Also included are unnamed genes with known descriptions and genes annotated as mRNA as the annotated descriptions provide information regarding their function and identity. All genes were taken from the rarefied dataset which was normalised to the same sequence number per station. Colour scale represents the percentage expression of each gene at each station. Numbers represent the identity of distinct clusters. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

6.2.5 Gene Expression Profiles: Genes Expressed at HI Stations

The genes present in cluster 4 and 5 (Figure 6.7) were most highly expressed at the HI stations. The identities of all genes found in each cluster are shown in Appendix 4. Within cluster 4 and 5 photosynthetic genes were identified, namely; *bchl, Gasu_40100, Gasu_40450, Gasu_40560, petB, PSAF, PSBC* and *PsbV.* Both clusters featured homologues of *Gasu_40100*, *Gasu_40450, petB* and *PSBc.* Gene homologues are those that were annotated with the same gene name during assembly due to having the closest match to the same database entry, but that were assembled as separate genes, likely derived from taxa for which a specific representative sequences were not available.

Some structural genes were identified such as *Act2* which encodes actin, *TUA* and *NCLIV_058890* which encode α-tubulins, and *TUB_2* which encodes β-tubulin[316]. Additionally, 10 unnamed genes which were annotated as encoding structural proteins such as tubulin and actin. *H4_1*, *hsp70_4, hsp70A* and *hsp90_1* were present and encode heat shock proteins which may be expressed in response hypoxia, physical trauma[317], temperature[318] and salinity stress[319]. However, they are multifunctional proteins that are also involved in protein assembly, secretion[319] and degradation[318]. Also present in cluster 4 was *GAPD1,* a gene involved in glucose metabolism, but that has also been shown to have a role in endocytosis, DNA repair and apoptosis[320].

The genes present in cluster 2 and 3 (Figure 6.7) were most highly expressed at CTD62. Homologues of the previously mentioned genes; *GAPD1, Gasu_40100, petB, PSAF, petB, PSBC, PsbV, TUB2, bchl, HSP90_1* and *NCLIV_058890* were identified, as well as further tubulin genes. Other photosynthesis genes *Lhcf22, Lhcf28, Lhcf29, Lhcf34_2, Lhcf61, Lhcf62* and *Lhcf67* were identified. *Lhcf* genes encode light harvesting complexes[321,322], and upregulation is thought to promote photoprotection of newly produced photo-reaction centres by way of regulating and modifying light harvesting antennae[322,323], or to promote light harvesting under lower light conditions[322].

Some genes related to metabolic processes other than photosynthesis were most highly expressed at HI stations, such as *ACA1_097000,* which plays a role in the metabolism of non-optimal carbon sources[324]. The *ANT1* gene encodes an inner membrane channel in mitochondria to transport ADP into the mitochondria and ATP back out into the cytoplasm of the cell[320]. Other metabolic related genes identified included an unnamed gene annotated as an ATP synthase, an S-adenosyl-L-homocysteine hydrolase, and what is thought to be a dTDP-glucose 4,6-dehydratase, although the annotation appears incomplete.

In summary, the HI stations featured high expression of mostly photosynthetic genes, implying that photosynthesis was a dominant function. The recovery of a number of heat shock genes suggests that members of the community might be experiencing stress, but may have been induced by sampling or be the result of cellular maintenance and regulatory processes[318,319]. The recovery of genes annotated as structural tubulins and actins at the HI stations may also represent a response to stress. Conclusions regarding the gene expression observed at the HI stations are largely mirrored by genes recovered in cluster 2 and 3, which were most highly expressed at the HI station CTD62. It may be that these genes represent similar homologues with slightly different induction limits that weren't found at CTD59. Photosynthetic genes *Gasu_40450, petB* and *PSBV* were recovered across multiple clusters, implying some conserved functionality across the gradient of Polar Water influence, but different profiles of expression.

6.2.6 Gene Expression Profiles: Genes Expressed at LI stations

Genes present in cluster 6 (Figure 6.7) were most highly expressed at CTD56. Within this cluster homologues of photosynthetic genes previously identified in clusters 4 and 5 were again recovered (*Gasu_40100,* two *Gasu_40450* homologues*,* two *Gasu_40630* homologues*, petB* and four *PSBC* homologues*).* A number of other *Gasu* genes were identified, namely *Gasu_40150, Gasu_40260, Gasu_40270* and *Gasu_40760. Gasu_40260* is an ATPase encoding gene[325], and as such behaves similarly to the previously described *Gasu_40630* gene identified in cluster 3*. Gasu_40760* encodes ribulose-bisphosphate carboxylase (*RUBISCO*)[320] that catalyses the carboxylation of ribulose-1,5-bisphosphate.

PetF_1 was recovered and encodes a ferredoxin, an iron containing gene involved in photosynthesis[101]. Additionally, *GapC1* was most highly expressed at the LI station CTD56, it is a cytosolic glyceraldehyde-3-phosphate dehydrogenase involved in the generation of ATP and NADH for cellular energy[326], suggested to be upregulated during heat-shock and other abiotic stressors[327,328].

Cluster 1 was mostly highly expressed at the other LI station CTD57. All genes present in cluster 1 were unnamed, making commenting on them further difficult. However two were annotated as actin genes, and one as an elongation factor gene.

In summary, the LI stations featured high expression of genes related to the light reactions of photosystem II photosynthesis, and also a *RUBISCO* gene related to carbon fixation, suggesting photosynthesis as a key process at the LI stations. Some genes involved in cellular structure and translation were also identified, as would be expected from a community featuring actively metabolising cells.

6.2.7 Gene Expression Profiles: Genes Expressed at the MI station

The genes present in cluster 8 and cluster 9 (Figure 6.7) were most highly expressed at the MI station, CTD58. With the exception of an *ubiquitin* gene, all the genes identified in cluster 8 and 9 were unnamed. One unnamed gene was annotated as a potential *helicase* in cluster 8, but the remaining genes in both cluster 8 and 9 were annotated as antisense rRNA, which are suggested to play a role in gene expression regulation[329].

6.2.8 Gene Expression Profiles: Most Abundant Genes

Many of the genes already discussed in the clustering analysis presented in section 6.2.5-6.2.7 constituted the most highly expressed annotated genes (Figure 6.8). 12 were present in clusters 4 and 5, which were expressed at the highest levels at the HI stations, 2 from cluster 2, and 3 from cluster 6. The majority of the most abundant genes were observed to be involved in photosynthesis and included; *PSBC, petB, PsbV*, *bchl* and various *Gasu* genes, including two *RUBISCO* encoding *Gasu_40760* homologues. Within the most abundant genes recovered were homologues of the same genes, for which different profiles of expression were observed. For example, one homologue of *Gasu_40100* was most highly expressed at the HI station CTD62, while another was most highly expressed at the LI station CTD56. A similar trend was also observed for *PSBC* homologues, three of which show the highest expression at the HI stations, one at the LI station CTD56, and one at CTD56 and both HI stations but with low expression at the CTD57 and CTD58. These results further support photosynthesis and carbon fixation is a key process within the communities at each station group. The top 20 most abundant genes were identified from the two most abundant taxonomic groups, the Rhodophyta and Stramenopiles (Figure 6.3A).

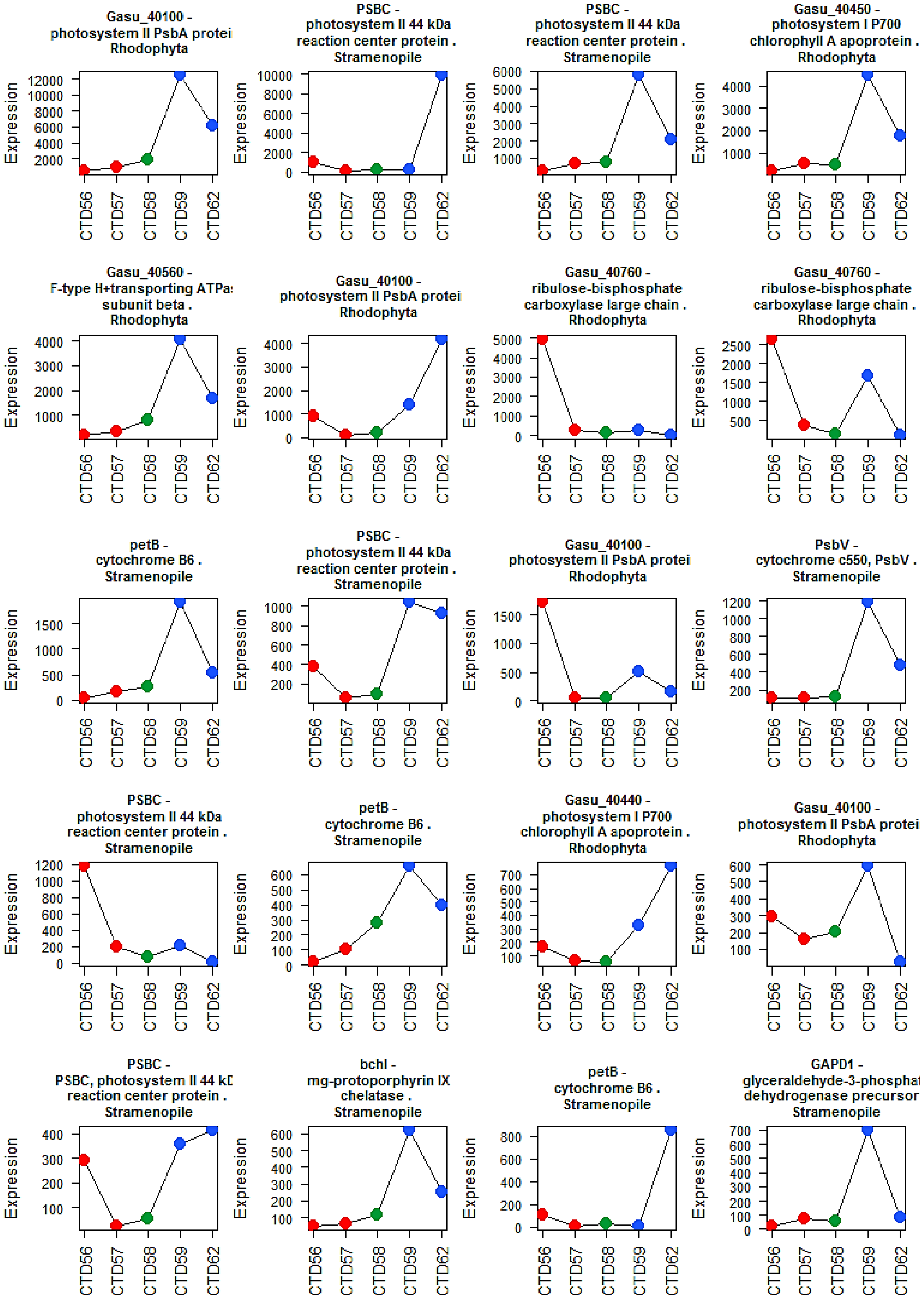


Figure 6.8. The expression of the top 20 most abundant annotated genes.

Shown are the expression profiles of the top 20 most abundant annotated genes recovered from the transcriptome in descending order of abundance from left to right. Titles show the annotated gene name, description, and the representative major taxonomic group the gene was annotated as being most similar to. Some descriptions have been simplified due to space constraints. Expression values are total sequence count per gene which had been normalised by library size. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

6.2.9 KEGG Analysis

Each station group was observed to display distinct partitioning (Figure 6.4), which was resolved to be the result of different gene expression profiles between stations (Figure 6.7). However, many of these genes featured annotations which suggested similar functions. KEGG analysis was performed to generate information about the functional pathways annotated genes were involved in to address whether the different gene expression profiles resulted in different functional potentials between each station group.

The KAAS (KEGG Automatic Annotation Server)[180] applies annotations directly to assembled contig sequences, as such it was prudent to include all 443 eukaryotic genes in the analysis. However, not all genes were annotated with KEGG annotations and so some functions may have been missed in the following KEGG analysis. KEGG annotation was successfully applied to 118 of the 443 eukaryotic genes, which matched 75 unique KO identifiers. KO identifiers were categorised into functional pathways according to BRITE functional hierarchy excluding human disease. The top 5 most highly represented pathways were ‘energy metabolism’, ‘translation’, ‘transport and catabolism’, ‘cell growth and death’ and ‘cellular community – eukaryotes’ (Figure 6.9). Low numbers of unique KO identifiers were present within most groups implying high levels of genetic redundancy. Results match expectations for a diverse and metabolically active community with various mechanisms for metabolism and growth.

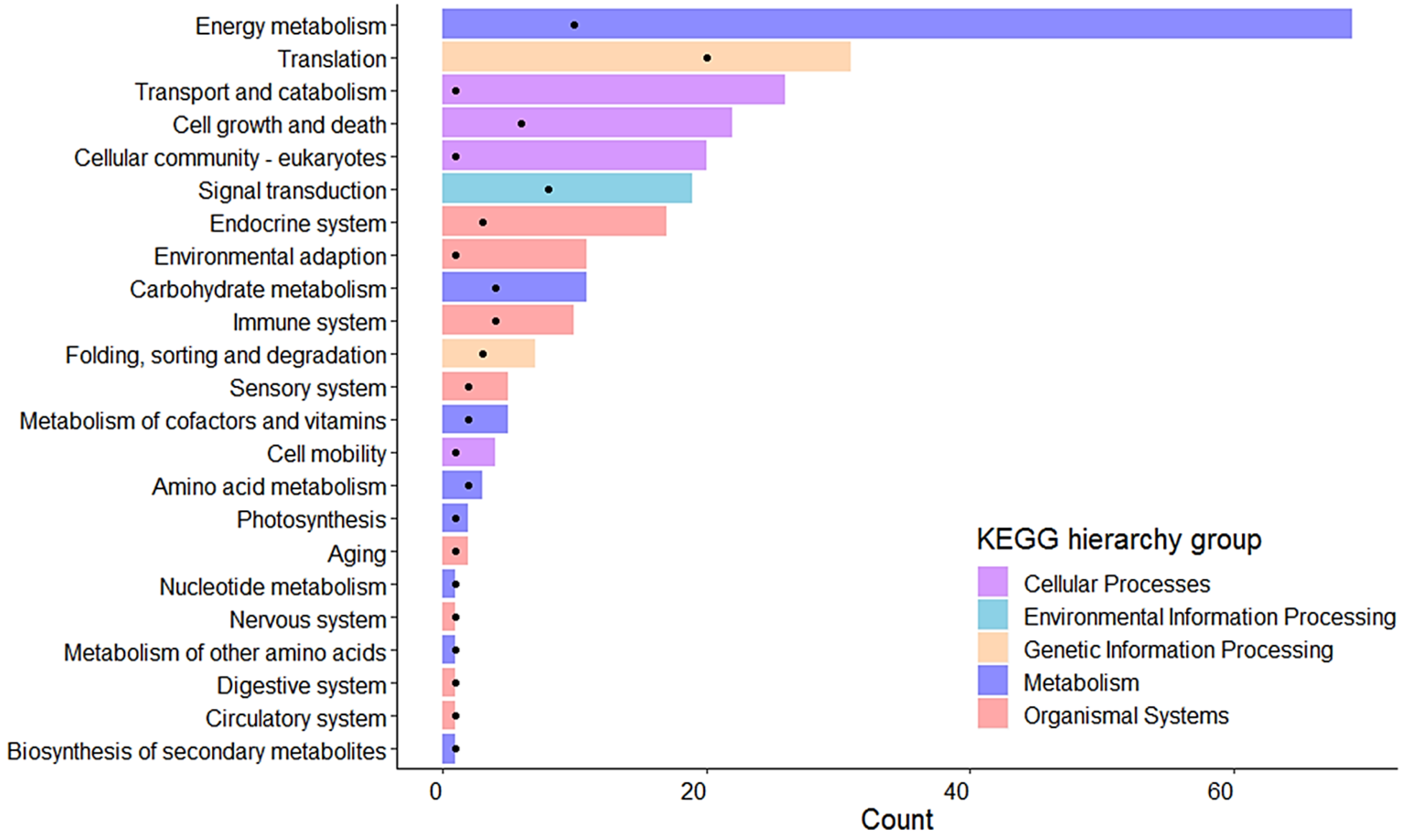


Figure 6.9. KEGG functional pathway annotations for eukaryotic genes.

Annotations represent level 2 KEGG pathway annotation terminology and are categorised according to BRITE hierarchy functional groups, excluding human diseases, as colour coded. Bar length represent the total count of KEGG annotated genes per term. Points represent the number of unique KO identifiers per level 2 annotated term.

6.2.10 Functional Expression Profiles

Plotting the relative functional expression of KEGG functional pathways within each station provides an overview of the differences between the functional profiles, and the determination of the dominant functions at each station. Results revealed functional pathways related to metabolic processes were the most dominant across all stations, with ‘energy metabolism’ being the most dominant pathway (46-71%) (Figure 6.10).

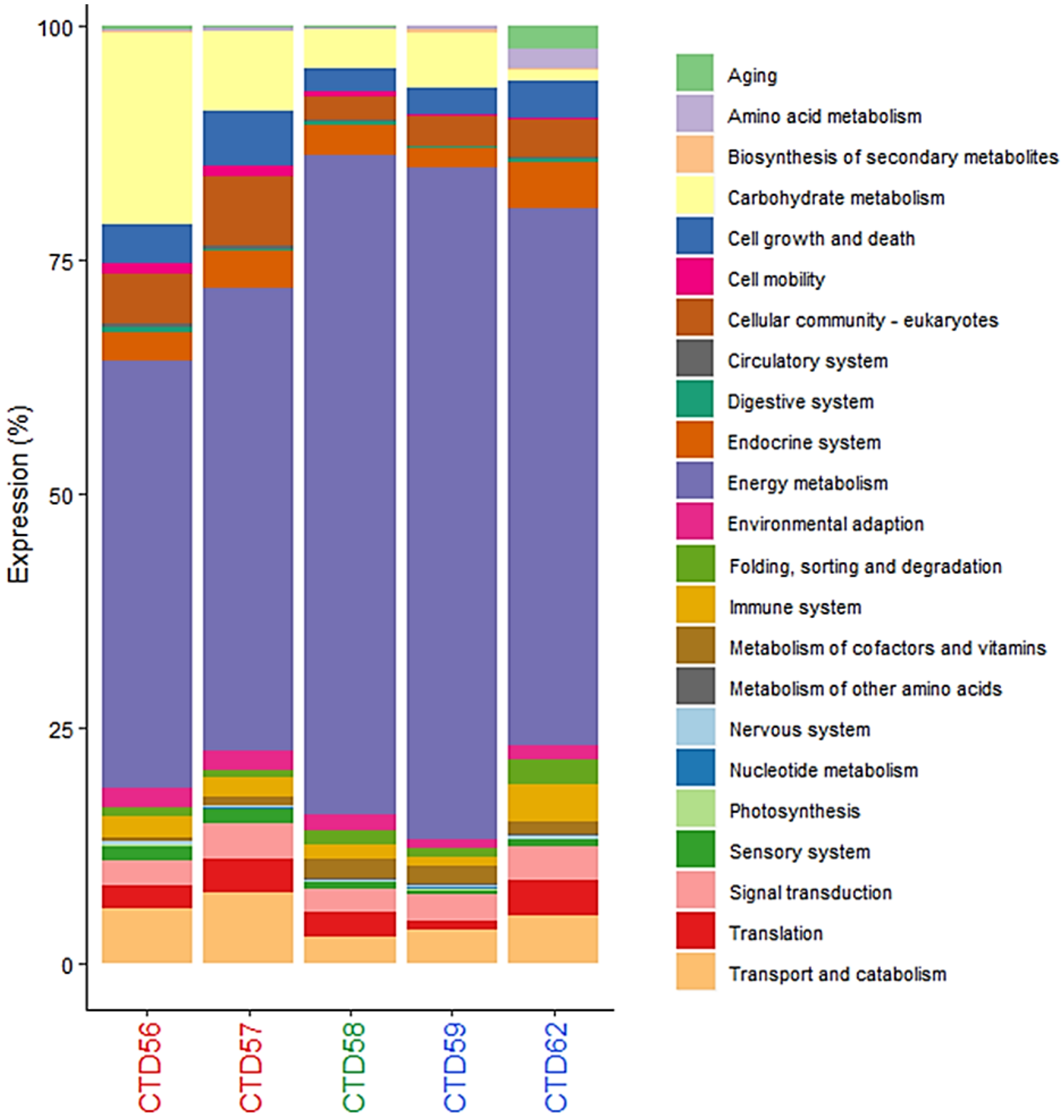


Figure 6.10. The relative expression of KEGG pathways within each station.

The relative expression of all sequences with KEGG annotation at each station, collapsed into their respective functional pathways as colour coded. All samples were taken from the rarefied dataset which was normalised to the same sequence number per station. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

A clear pattern of expression was observed for ‘carbohydrate metabolism’, which was most dominant at the LI stations (20% and 8% for CTD56 and CTD57 respectively) and reduced across the transect to the lowest value at CTD62 (1%). The majority of other pathways featured fairly consistent expression across all stations but some minor differences were observed. ‘Cell growth and death’, ‘cell mobility’, ‘cellular community - eukaryotes’, ‘environmental adaption’, ‘metabolism of other amino acids’, ‘nucleotide metabolism’, ‘sensory system’ and ‘transport and catabolism’ were all slightly more highly expressed at the LI stations. ‘Biosynthesis of secondary metabolites’ was most highly expressed at the HI stations, and ‘metabolism of cofactors and vitamins’ at the MI station. All other KEGG pathways displayed no clear patterns of expression.

Examining the functional expression of KEGG pathways by taxonomic group allows a greater insight into the functional profiles of constituent community members. Multiple KEGG pathways were observed to be present across the majority of taxonomic groups (Figure 6.11). Few functional hierarchy groups were active for the Ameoboza, Apicomplexa and Rhodophyta. Haptophyta and Stramenopiles were the best represented taxonomic groups. Different functional expression profiles were observed within individual taxonomic groups across stations. For example the Amoebozoa displayed a high expression of the KEGG pathways ‘translation’ at CTD62, with comparatively low expression at other stations. This contrasts with the Excavata for which ‘translation’ was highest at the LI stations (Figure 6.11). The Haptophyta displayed high levels of expression for the majority of functional pathways at CTD62, the Stramenopiles at both HI stations, and the Opisthokonta at LI stations. These different functional profiles imply different taxonomic groups featured different levels of functional potential at different station groups, potentially due to the different environmental conditions between station groups. The expression profile of the majority of KEGG functional pathways for the Apusozoa were highest at either end of the transect, with reduced expression levels in between, and the lowest levels of expression observed at CTD58. The Haptophyta exhibited a similar trend for the majority of functional pathways within the KEGG hierarchy group ‘organismal systems’. However, constituent pathways associated with ‘ageing’, and ‘immune system’, and those contained in the remaining hierarchy groups, namely ‘cellular processes’, ‘environmental information processing’, ‘genetic information processing’ and ‘metabolism’ displayed the highest levels of expression at CTD62, with reduced expression elsewhere. The expression profiles for the ‘carbohydrate metabolism’ pathway differed between the Haptophyta, Rhodophyta and Stramenopiles. The

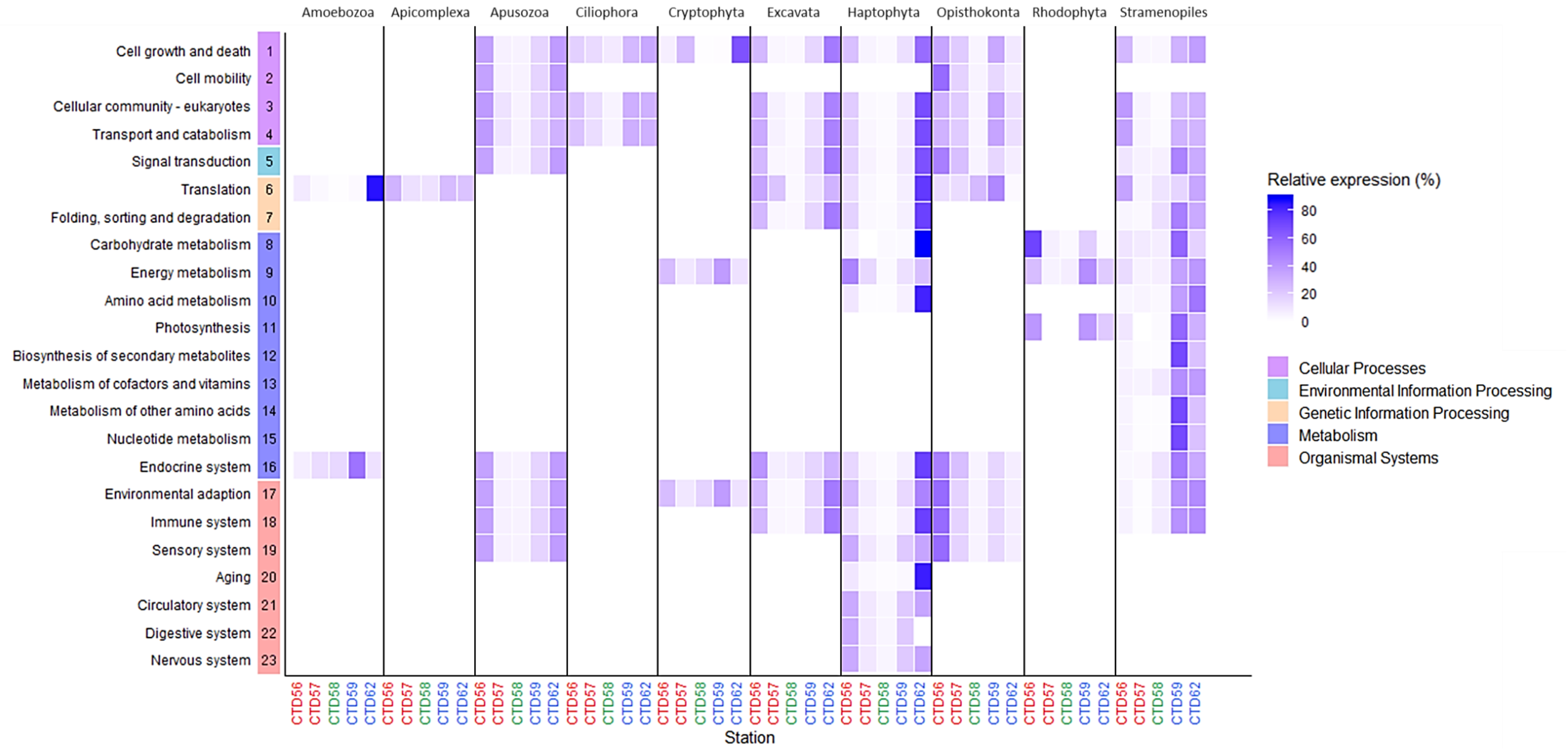


Figure 6.11. The relative expression of KEGG pathways at each station for each individual taxonomic group.

Shown is the relative percentage expression of sequences annotated with KEGG pathways for each taxonomic group across the transect. Annotations represent level 2 KEGG annotation terminology and are categorised according to BRITE functional hierarchy, excluding human diseases, as colour coded. Colour scale represents the percentage expression of sequences for each KEGG pathway at each station so that the expression of each pathway may be compared across stations. All samples were taken from the rarefied dataset which was normalised to the same sequence number per station. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.

Haptophyta displayed high expression at the HI station CTD62, the Stramenopiles at both HI stations, and the Rhodophyta at CTD56. The functional profiles of ‘energy metabolism’ also differed between the Cryptophyta, Haptophytes, Rhodophyta and Stramenopiles, being fairly consistent across stations for the Cryptophyta, high at LI stations for the Haptophyta, and high at the HI stations for the Rhodophyta and Stramenopiles. The differences in the expression profiles of these pathways may reflect the utilisation of different energy sources of different organisms under different environmental conditions. Stramenopiles also displayed high levels of certain metabolic pathways at CTD59, specifically ‘biosynthesis of secondary metabolites’, ‘metabolism of other amino acids’ and ‘nucleotide metabolism’.

The expression profile of pathways contained the hierarchy group ‘organismal systems’, which includes ‘environmental adaption’ were most elevated at the HI stations for the Apusozoa, Cryptophyta, Excavata, Haptophyta and Stramenopiles. However, the Opisthokonta displayed the highest expression at the LI stations. These profiles may reflect differing responses to environmental stressors between taxonomic groups and agree with the observed trends for ‘cell growth and death’.

6.2.11 Expression of Photosynthetic Genes

The expression of photosynthetic genes dominated the metatranscriptomes of both the LI and HI stations, however some differences were observed between stations. The LI station featured a greater number of genes (10) annotated as belonging to photosystem II (PSII) (Figure 6.12A) than compared to those annotated as belonging to photosystem I (PSI) (Figure 6.12B). Additionally, all of the 8 light harvesting *Lhcf* genes identified were most highly expressed at HI station CTD62 (Figure 6.12C).

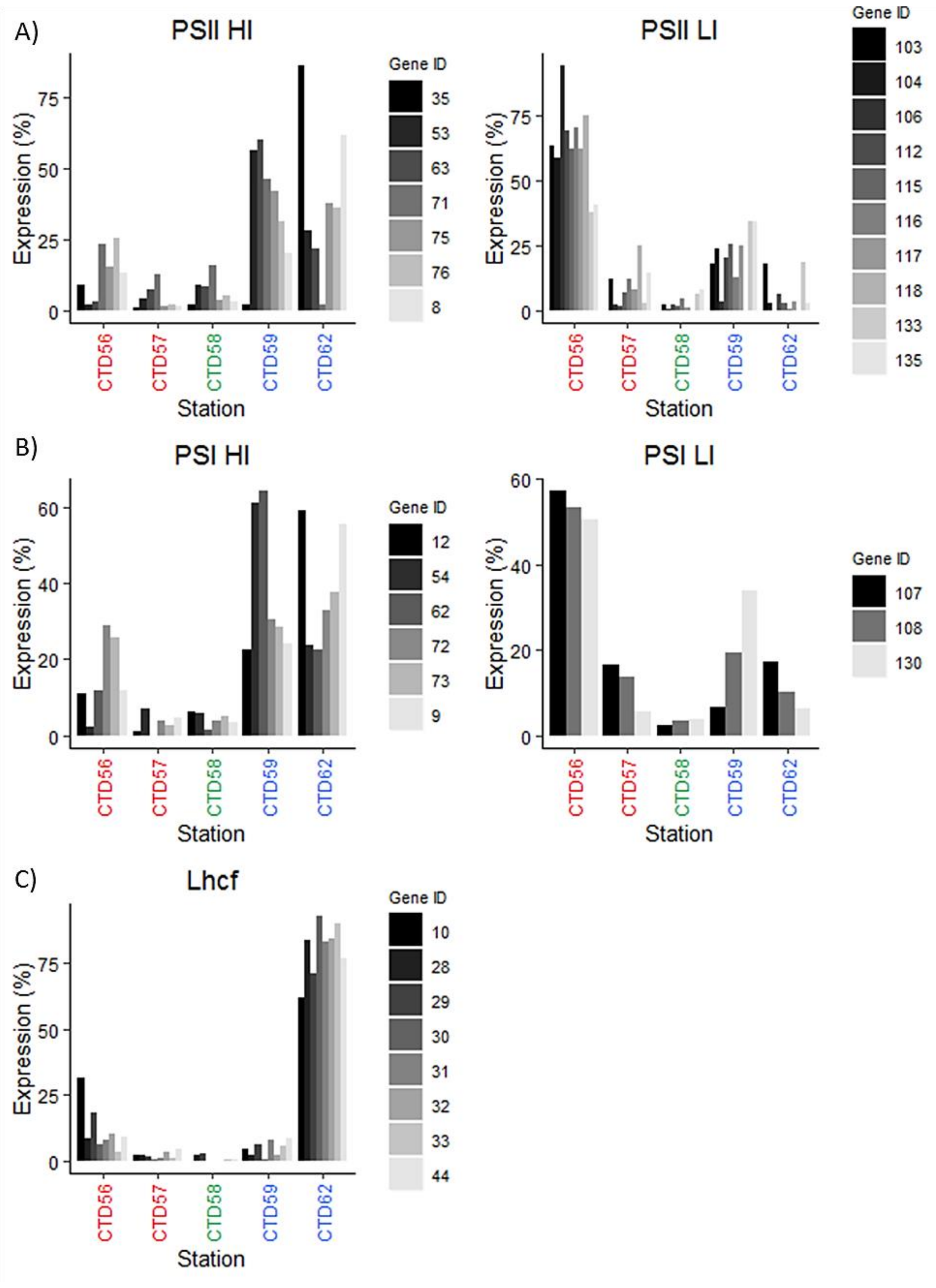


Figure 6.12. The relative expression of PSI, PSII and *Lhcf* photosynthetic genes.

Shown is the relative expression of normalised genes annotated as being involved in PSII (A), PSI (B), and *Lhcf* (C) genes, grouped by those which show the highest expression at the HI (left) and LI (right) stations. The relative expression of genes as a percentage is shown on the y axes. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI. The ID number of each gene is shown and corresponds to those displayed in Appendix 4, where full gene annotation information can be found.

6.3 Discussion

The sampled stations were assigned into one of three regional groups reflective of different degrees of Polar Water influence (as discussed in Chapter 3). The analysis presented in Chapter 5 revealed complex partitioning of the eukaryotic community across these regional groups, with contrasting partitioning observed between different abundance fractions and constituent taxonomic groups. The observed partitioning implies a susceptibility of these communities to perturbations under predicted increased Atlantic Water influence in the region. Such susceptibility is likely to result in changes to the composition and abundance of community members with knock on effects to ecosystem service functions[24]. However, the analysis presented in previous chapters was derived from DNA barcoding sequencing and thus were constrained to only reporting the genetic abundance of taxa, and was not able to resolve functional information. The findings presented in this chapter attempt to compliment the findings presented in Chapter 5 by applying a metatranscriptomic approach, utilising the Illumina MiSeq platform, to recover the diversity of actively expressed eukaryotic transcripts across the sampled region and predict the direction of potential functional changes under further Atlantic Water influence in the region.

6.3.1 Sequence Analysis and Assembly

1.5 million of the 14.3 million high quality processed pair-end sequences available for use in the transcript assembly (Table 6.1) were successfully assembled into 958 contigs (Table 6.2) of mostly eukaryotic origin (Table 6.3, 6.4), with satisfactory length (Figure 6.1). 3.8 million sequences where mapped back onto these contigs, as is standard protocol for estimating the expression of novel transcripts[176]. The fraction of the sequences assembled was lower than seen for large environmental sequencing efforts such as the GOS[71] and Tara Oceans expedition[42]. The quality filtering steps used during pre-processing were stringent, but resulted in the loss of only <10% of sequences (Table 6.1), meaning low quality sequence data was not a primary factor in effecting the assembly. Differences between the bioinformatics pipelines used, the assembled contig lengths, the quantity of available sequences, and sequencing technologies applied complicates comparisons between studies. For example, the Tara Oceans dataset consists of 137.5 million sequences, representing a far greater sequence depth than available in this study, and making it likely that a greater number of contigs would be successfully assembled. Additionally, the GOS assembled 3.1 million contigs from 7.7 million sequences, but featured only 15% of sequences with over 1x coverage[71], meaning a significant proportion of contigs within the GOS were represented by very few sequences. This contrasts with the present study which featured an average coverage of 649.36x (Figure 6.2). Studies of environmental metatranscriptomes of phytoplankton communities report mapping 25.6-51.2% of sequences to assembled contigs[156] which is comparable to the 26.9% (3.8 million sequences) successfully mapped onto the assembled contigs in this study (Table 6.1). Furthermore, the fraction of assembled sequences was close to figures reported for some of the most commonly used software packages when *de novo* assembly methods are applied to high-complexity sequence data[330]. The number of assembled and mapped sequences presented in this chapter is thus within expectations for a *de novo* assembly of highly diverse environmental samples.

It is possible that the success of the assembly may have been affected by the internal algorithm DNAStar utilises or the settings used during the assembly stage. However, DNAStar was chosen due to its ability to generate greater average contig lengths, reduced number of erroneous contigs and greater assembly accuracy compared to other assemblers[331,332]. The factor which likely had the greatest impact upon the assembly success was that metagenomic assembly tools are still in development, and at current find accurately assembling such complex environmental sequence data *de novo,* in the absence of reference genomes, challenging[333]. Global sampling surveys report similar limitations, able to assembly only small fractions of the total available sequence data into contigs. For example, only 116.8 million sequences were assembled from 16.5 terrabases of raw data during creation of the MATOU (Marine Atlas of Tara Ocean Unigenes) database derived from the Tara Oceans dataset[334].

6.3.2 Community Composition

Examining the community composition across all stations revealed that all recovered major eukaryote taxonomic groups were present at all stations (Figure 6.3A). The most abundant taxonomic group was the Stramenopiles, reflecting previous observations from the DNA barcoding analysis performed in Chapter 5 (Figure 6.3B), and implies that the Stramenopiles were the most abundant and functionally active members of the community. The distribution of the Stramenopiles is known to show abundance correlations with environmental variables[335,336], and their distribution across the sampled region also broadly matched the pattern observed in the DNA analyses, with relatively low abundance at the LI stations, and high abundance at the

HI stations. The notable exception was that within Chapter 5 the Stramenopiles at the MI station displayed similar abundance to those at the LI stations (Figure 5.3B), whereas in the transcriptomic dataset the MI station displayed abundances similar to the HI stations (Figure 6.3A). Differences in the proportion of sequences constituting each taxonomic group within the DNA and transcriptome datasets existed for most taxonomic groups. The reason for this is unclear, but could be due to differences in biological activity between each dataset, potentially from compositional differences between recovered taxa, or related to technical differences between the methods used. Unfortunately, due to the use of DNA barcoding in Chapter 5 but whole genome metatranscriptomics in the present chapter, directly comparing DNA:RNA may not be robust, and database limitations prevented more resolved taxonomic analysis of the transcriptomic dataset.

Representatives of Apicomplexa were the most highly recovered taxonomic group at the LI station CTD56, and reduced in abundance across the transect (Figure 6.3A). The Apicomplexa are known to display preferences for warmer temperatures[337], therefore the observed distribution matches expectations. The other LI station, CTD57, also featured comparatively high abundance of Apicomplexa, but was dominated by Rhodophyta, a diverse group of red algae.

Apusozoa were identified within the transcriptomic dataset, but lacked representation within the DNA dataset (Chapter 5). This is a small group suggested as a sister group to the Opisthokonta and its current phylogenetic placement remains under discussion[338]. It’s omission is likely the result of the absence of this group from the SILVA database at the time of writing[164], and thus inability for a BLAST match to be found, rather than absence of Apusozoa sequences in the dataset used in Chapter 5.

When compared to the dataset presented in Chapter 5 (Figure 6.3B) no representatives of the groups Centrohelida, Charophyta, Chlorophyta, Dinoflagellata, or Picozoa were recovered from the metatranscriptome. Metatranscriptomics is an expensive molecular technique, and its application to environmental samples is still in its infancy. Furthermore, eukaryotic genomes are large and significantly more complex to analyse compared to those of prokaryotes[339]. As such marine organisms are weakly represented in online databases due to a prioritisation of those derived from taxa of commercial importance, such as those which are pathogenic to humans or to important aquaculture species[339]. Additionally, many microbial marine environmental taxa are resistant to cultivation[340] which poses considerable challenges to their characterisation. Therefore, the lack of recovery of representatives for some major taxonomic groups is unsurprising. In the case of the Picozoa and Centrohelida no database representation was found. The lack of recovery of representatives of Charophyta, Chlorophyta and Dinoflagellata could be due to a lack of related marine sequences despite these groups being represented in the RefSeq database.

That marine organisms are weakly represented in online databases may also explain the low sequence similarity of recovered genes to those present in online databases, which were often <90%, implying that the sequences recovered were environmental isolates which lacked database representation. Therefore, caution has been taken to only discuss the taxonomy of sequences at a high level to ensure the stated taxonomic classifications are robust. The novel contigs assembled from the transcriptomic dataset (Table 6.2) also likely belong to taxa lacking database representation. Recently, attempts have been made to address the lack of representation of marine eukaryotes in online sequence databases, such as through the generation of dedicated marine microbial eukaryote databases including the MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project)[339] and MATOU[334]. The sequencing surveys which have generated these databases have massively increased the quantity of available marine eukaryotic reference sequences for analysis, and provide select reference databases focused specifically at the marine habitat. Indeed, the MATOU represents the largest catalogue of eukaryotic transcripts from any single biome, but the majority of sequences still lack annotation[334], implying that while a suitable foundation now exists for exploring marine eukaryotic diversity, much of the natural environmental diversity still remains uncharacterised. However, the use of such databases may have led to a higher number of successful annotations.

6.3.3 Regional Genetic Profiles

Distinct genetic profiles for each station group were observed (Figure 6.4A). The genetic profiles of the constituent stations within each of the LI and HI station groups were observed to be most similar to one another. The MI station was seen to partition closest to the LI stations (Figure 6.4A) and therefore featured a profile most similar to the LI stations, yet was still distinct. When stations collected from the North Atlantic were included (Figure 6.4B) the profiles maintained the observed partitioning, supporting that the communities at each station group were genetically distinct, which was resolved not to be due compositional differences of genes between stations

(Figure 6.5). 68.7% of the variance observed within the genetic profiles of communities sampled from the stations present within the Norwegian Sea was resolved to be the result of environmental factors (Figure 6.6), comparable to the 72.1% resolved in Chapter 5.

More resolved analysis revealed different expression profiles of individual genes between stations that likely explain the observed partitioning of the genetic profiles between regional station groups (Figure 6.7). However, each station group was observed to feature similar genes, including examples of gene homologues. Genes involved in photosynthesis and organism structuring were particularly common in the LI and HI station groups. Indeed, genes involved in photosynthesis were observed to be some of the most dominant within the transcriptome, and homologous examples were present which displayed different levels of expression at different stations (Figure 6.8), supporting the possibility of similar functionality between stations despite differences in the expression profiles of individual genes. Previous studies have reported different gene expression profiles[341] and well as community composition across environmental gradients[227,229]. Additionally, select psychrophilic taxa have been reported as genetically distinct, such as a Polar *Fragilariopsis cylindrus*[97] ecotype, a psychrophilic *Micromonas* Arctic ecotype[261], cold-adapted Chlorophyta[342], and unique polar *Emiliania huxleyi* genotypes[85]. The distinct genetic profiles observed at different stations therefore match expectations based on evidence from current literature, but no homologues of genes identified in the aforementioned studies were identified in the transcriptome.

Salinity was calculated to be the most significant environmental variable that correlated with the observed partitioning of genetic profiles across the transect at *p*=0.06, just below the typically accepted 95% confidence interval (Table 6.5). Different genetic profiles have been observed within microbial eukaryotic communities over salinity gradients[343,344], including at the entrance to the Arctic Ocean[141]. Analysis of the microbial eukaryotic community presented in Chapter 5 resolved temperature to be the most significant environmental factor affecting community partitioning (*p*=0.05), but a correlation with salinity which approached the traditional threshold of statistical significance (*p*=0.08) was also observed, and salinity was significantly correlated (*p*=0.01) with the partitioning of the bacterial community observed in Chapter 4. Distinguishing between the impacts of abiotic factors can be challenging as they rarely act in isolation, and within the sampled region temperature and salinity measurements both reduced as the influence of cooler and fresher Polar Water increased (Chapter 3) which may have impacted the correlations observed. It is also possible that there were complex interactions at play which have profound impacts upon the gene expression profiles between stations beyond the environmental factors measured, or solely environmental factors. The complexity of the sequence data, the number of assembled genes, or limited number of sampled stations may also be contributing factors to the observed statistical power.

As mentioned in section 6.3.2, Stramenopiles were the most abundant taxonomic group recovered within the transcriptome, and were primarily annotated as Diatoms. Diatoms are capable of both the uptake and release of nitrate, and feature potentially lower temperature adapted nitrate reduction enzymes to promote a competitive advantage under nitrate rich environments[345]. Furthermore, differential gene expression of Diatom transcriptomes has been observed under nitrate limitation[346]. Such factors may explain the observed correlation of nitrate with the partitioning of genetic profiles (Table 6.5).

6.3.4 Predicted Changes to Functional Profiles

KEGG analysis provided an insight into the functional processes operating within the microbial communities sampled across the sampled region. High numbers of pathways related to ‘energy metabolism’, ‘translation’, ‘transport and catabolism’ and ‘cell growth and death’ (Figure 6.9) support that processes important to metabolism and cellular survival/maintenance were dominant within the community. It is important to note that while it initially appears that the KEGG pathway ‘photosynthesis’ is under represented, this is in fact an artefact of the KEGG database formatting, which classifies many photosynthetic KO identifiers under the ‘energy metabolism’ pathway when level 2 KO annotations are considered. If broader KO annotation levels had been considered annotated terms would have become too general (e.g. ‘energy metabolism’ would have become ‘metabolism’) and resulted in the loss of meaningful functional information, whereas at more resolved levels relevant functional information would often be displaced by specific protein identity information, again limiting the recovery of meaningful functional information.

Low numbers of unique KO identifiers found in each group implied a high degree of functional redundancy within a highly diverse community (Figure 6.9). Such redundancy may act to buffer functional changes within the community as a result of environmental perturbations[17]. However, as KEGG classification contains pathways which share the same genes some KO identifiers were likely assigned to similar functional motifs. Therefore, redundancy is an inherent feature of the KEGG classification methodology and may be over inflated. The sharing of similar genes within different pathways also explains why KO identifiers relating to ‘circulatory system’, ‘nervous system’ and ‘aging’ were recovered from marine microbial eukaryotic community samples. For example, one KO identifier was annotated as part of the ‘nervous system’ pathway, but examining higher level KEGG annotation revealed this to represent calmodulin, a multifunctional calcium binding protein highly conserved across eukaryotes[347]. The recovery of KO identifiers associated with ‘circulatory system’, ‘digestive system’ and ‘aging’ pathways were also due to the presence of highly conserved multifunctional genes.

In Chapter 5 the partitioning of the rare and intermediate fraction of the eukaryotic community revealed that the community at the MI station was most similar to the LI communities, implying a likely displacement of those taxa at the HI stations under increased Atlantic Water influence within the sampled region. Similar partitioning was observed within the transcriptome (Figure 6.4), suggesting that such a displacement would potentially impact the genetic profiles expressed by the communities. Examining the relative distribution of sequences with KEGG annotation between stations revealed that KEGG functional pathways associated with ‘energy metabolism’, which included photosynthetic pathways, were dominant across all stations (Figure 6.10). Pathways associated with ‘energy metabolism’ were slightly reduced at the LI stations compared to the MI and HI stations, and LI stations featured elevated levels of ‘carbohydrate metabolism’ which suggests the potential for minor alterations to the profiles of energy production between station groups under increased Atlantic Water influence. Pathways associated with ‘metabolism of other amino acids’ and ‘nucleotide metabolism’ were also most highly expressed at LI stations, further supporting the suggested potential changes to profiles of energy production. Despite these differences the majority of KEGG pathways were relatively conserved between station groups, agreeing with the previous suggestion of functional redundancy revealed by low unique KO identifier number (Figure 6.9). Thus data implies that while each station group is genetically distinct (Figure 6.7), that the functional potential between them is largely conserved. However, KEGG analysis reports on a subsample of genes within the transcriptome for which KEGG annotations could be found, and it is possible that some functional differences exist across the transect for genes lacking KEGG annotation which are omitted in this analysis. Furthermore, the activity levels of functionally similar genes can vary between taxa, such as iron utilisation efficiency between diatom species[348], or carbon dioxide-fixation efficiency of *RUBISCO* genes[349], and thus minor changes in gene expression of specific taxa, or compositional changes of constituent taxa with different gene activity/efficiency levels may impact local ecosystem functionality.

Examining the expression of KEGG pathways by taxonomic group allows an insight into the functional profiles of constituent community members. A low number of functional pathways were recovered for the Ameoboza, but is unsurprising as this groups was shown to constitute a small fraction of the community (Figure 6.3A). However, a low number of functional pathways were also recovered for both the Apicomplexa and Rhodophyta which were recovered at greater abundances, the latter being the second most abundant taxonomic group. The low number of KEGG functional pathways recovered for the Apicomplexa (Figure 6.11) was likely due to a lack of represented sequences in the KEGG database. The majority of Apicomplexa sequences were annotated as hypothetical genes during assembly, further supporting this conclusion. A similar limitation likely existed for the Rhodophyta which featured a number of genes lacking KEGG annotation. However, some annotated genes lacked KEGG annotation despite being annotated as homologues of genes which were successfully annotated. The reason for this remains unclear as the gene identities of both annotated and unannotated homologues were similar, but may still have been different enough to result discrepancies due to thresholds used by the KAAS annotation parameters. Considering only functional annotations successfully recovered by the KAAS prevents the reporting of false positive functional annotations, and so it was prudent not to manually add functional annotations to unannotated genes for which annotated homologues were recovered.

By considering the partitioning of each taxonomic group observed in Chapter 5 (Figure 5.6) it is possible to comment on predicted impacts to the functional profiles displayed by individual taxonomic groups. The Excavata and Opisthokonta partitioned with the MI samples clustered most similarly to the LI stations, implying the HI associated communities will become displaced under increased Atlantic Water influence. Such changes are likely to impact the functional profile of the Excavata through reducing the expression of genes related to most recovered KEGG pathways, including ‘cell growth and death’, ‘transport and catabolism’, ‘protein folding, sorting and degradation’ and ‘environmental adaption’, resulting in reduced functionality of the Excavata (Figure 6.11). Excavata are diverse group constituting the basal lineage of flagellates an include both free-living and symbionts[350], but the functionality of marine Excavata remains poorly understood. Within the Excavata the pathways ‘translation’ and ‘endocrine system’ may experience increased expression (Figure

6.11). It has previously been shown that phytoplankton communities secrete endocrine disrupting chemicals[351], therefore the suggested changes to the ‘endocrine system’ expression profile may have consequences for the regional marine wildlife such as promoting reduced fertility, feminization, reproductive organ anomalies, and changes in sexual behaviour[352].

Data suggested that some members of the Opisthokonta may also experience increased expression of ‘endocrine system’ pathways (Figure 6.11), with similar functional impacts as described for the Excavata. Data predicts increased expression for ‘signal transduction’, ‘cell mobility’, and ‘environmental adaption’ pathways. Decreases are predicted for the remaining pathways, which included ‘translation’, ‘transport and catabolism’ and ‘cell growth and death’ implying the functionality of Opisthokonta may be reduced under increased Atlantic Water influence in the region. The α-tubulin gene *TUA* and a β-tubulin gene *TUB_2* were identified as being involved in the latter two pathways, and form the basis for the formation of microtubules. Disruption of the expression of these genes has been shown to impact intracellular transport, secretion pathways and cell division in phytoplankton taxa, slowing or stopping cellular growth[353], further supporting the suggested reduction in functionality of the Opisthokonta. The identified Opisthokonta sequences were identified as most closely matching choanoflagellates, which display functionally important traits such as bacteriovory and detritivory. Thus, they contribute to carbon and nutrient cycling within the microbial food web, and couple bacterial and microbial eukaryotic communities[354]. The predicted reduction of functionality may therefore likely have broad impacts across the microbial community and upon biogeochemical cycling within the region.

Within Chapter 5, partitioning was observed for the Apicomplexa and Cryptophyta that suggested the likely displacement of temperate water associated communities with those from cold water associated communities. Data suggested that some members of the Cryptophyta may experience increased expression of ‘cell growth and death’, ‘energy metabolism’ and ‘environmental adaption’ pathways under predicted future environmental change to the region (Figure 6.11). The Cryptophyta are a group of phytoplankton highly competitive under nutrient replete conditions, and that form blooms depending on environmental with temperature[355]. Therefore, the predicted alterations to Cryptophyte functionality under future warming associated with increased Atlantic Water influence within the region is likely to impact Cryptophyte bloom timing and nutrient cycling, with the potential for negative impacts to taxa less competitive under nutrient replete conditions. The Apicomplexa are a parasitic group of eukaryotes[355], alterations of the abundance and distribution of this group would have important implications for the health of higher trophic organisms. However, few functional pathways were recovered for the Apicomplexa and no clear expression pattern across the transect was observed, meaning the functional profile of Apicomplexa may remain unchanged under increased Atlantic Water influence (Figure 6.11).

No other taxonomic group that was resolved to display regional clustering within Chapter 5 (Figure 5.6) was recovered in the transcriptomic dataset (Figure 6.3), meaning it is not possible to comment on the potential functional impacts of these groups. Furthermore, the remaining taxonomic groups that were recovered within the transcriptome displayed no clear regional partitioning during previous analysis (Figure 5.6B), making commenting on possible changes to their functional profiles difficult. This difficulty is compounded due to the existence of the contrasting partitioning observed between different eukaryotic taxonomic groups in Chapter 5.

6.3.5 Defining Photosynthetic Strategy

Both HI and LI stations were dominated by genes related to photosynthesis which constituted 77% of the total eukaryotic transcriptomic sequences. Analysis of these genes can provide insight into the photosynthetic strategy of the cells or communities, which in turn may provide an insight into the how cells are utilising light energy and their potential nutrient status.

Photosystem units (PSU) are the cellular apparatus that allow phytoplankton to generate energy from light. They are composed of light harvesting antenna, and a photosynthetic electron transport (PET) chain which includes photosystem I (PSI) and photosystem II (PSII) subcomponents. At the HI stations similar numbers of genes annotated as belonging to PSI and PSII were expressed (Figure 6.12). However, at the LI station more genes annotated as belonging to PSII were highly expressed (10) compared to those from PSI (3). Such a difference implies potentially functionally different photosynthetic processes and strategies between the communities at HI and LI stations. Under low light conditions phytoplankton can invest in an acclimation

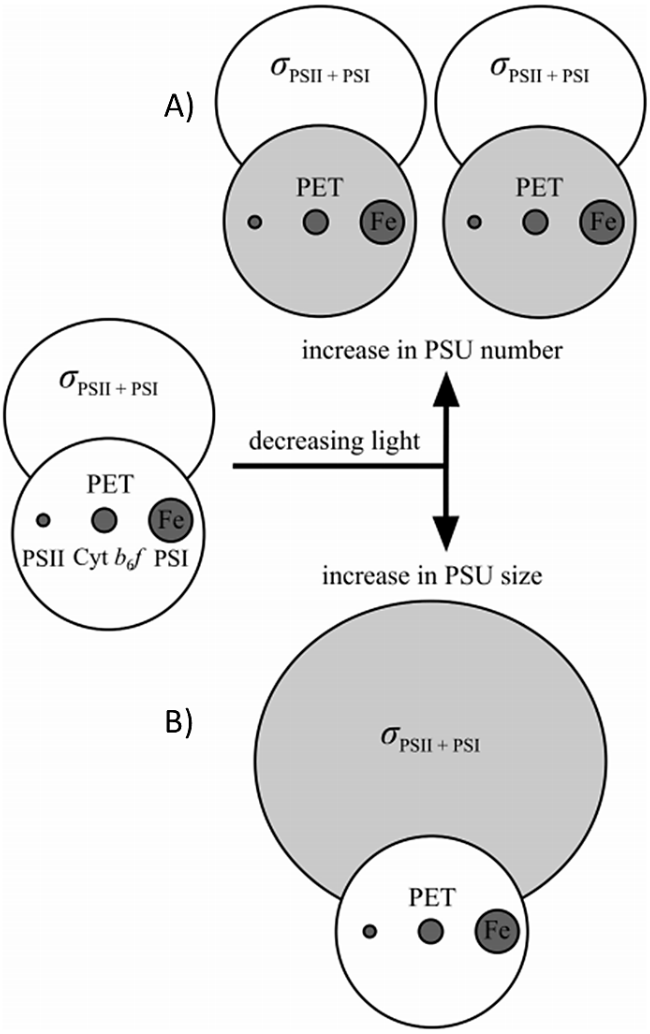


Figure 6.13. Simplified schematic overview of photoacclimation strategies for marine phytoplankton. Acclimation to low light may occur by two strategies, either A) an increase in the total number of cellular photosynthetic units (PSU), or B) and increase in the size of light harvesting apparatus. Increased PSU number increases the cellular concentrations of the iron (Fe) containing complexes of the photosynthetic electron transport (PET) chain. Dark circle size depicts relative iron content for both photosystem I (PSI), cytochrome b_6f complex and photosystem II (PSII). σPSII + σPSI represents the functional absorption cross section for light. Figure adapted from [356].

strategy by increasing the total number of photosynthetic units (Figure 6.13A), however due to the high iron demands of PSI this strategy increases the cellular iron requirements[356]. Iron plays a vital role in phytoplankton growth and primary production, being required for the synthesis of chlorophyll, respiration, and transport proteins, with low levels shown to result in low primary production and alterations to the community assemblage structure of open ocean communities[348]. *PetF_1*, a gene encoding a ferredoxin[101], was most highly expressed at the LI stations and functions as an electron accepter in metabolic reactions including PSI in photosynthesis (Figure 6.13). Low expression of *PetF_1* at the HI stations could suggest that iron limitation could be a contributing factor to the reduced expression of iron containing PSI and PSII genes observed at the HI stations (Figure 6.12).

To avoid increasing cellular iron requirements phytoplankton may deploy an alternative strategy and invest in a greater number of light harvesting apparatus per PSU, instead of total PSU number[356] (Figure 6.13B). All of the 8 light harvesting *Lhcf* genes identified were most highly expressed at HI station CTD62 (Figure 6.12), suggesting the HI community is potentially displaying an adaptive response to the observed lower light levels (Table 3.1) by investing in this strategy to promote light harvesting under lower light conditions[322], and that iron limitation could be an active selection pressure on the community. The phytoplankton communities of the North Atlantic are known to display iron limited growth patterns[357], but little is known about the iron status of the Norwegian Sea. Under iron replete conditions certain phytoplankton may substitute ferredoxin for flavodoxin which does not require iron, and the expression of flavodoxin can be used as a biomarker of iron stress conditions[348]. No flavodoxin genes were identified in the metatranscriptome implying that iron limitation wasn’t present, but a lack of detection does not necessarily indicate a true lack of presence within the community.

Phytoplankton are key for a large proportion of iron export to deep water, driven by sinking phytoplankton and cellular debris[348]. Therefore, changes in expression of genes related to iron uptake and utilisation will likely have community wide impacts upon photosynthetic processes and biogeochemical cycling, and are of ecological importance. However, physical environmental factors will likely have an impact on iron availability as melt water from sea ice is known to act as a source of iron in Polar Waters, being the dominant source during the spring and summer seasons[358], but is decreasing due to diminishing ice extent[183]. Further study is needed to explore the affects changes to phytoplankton composition, gene expression, and environmental change will have upon regional iron cycling and photosynthetic processes.

6.4 Conclusion

This chapter aimed to explore the metatranscriptome of microbial eukaryote communities across a transect of five sampling stations in the Norwegian Sea which covered a gradient of Polar Water influence. Multidimensional scaling of the LogFC of gene expression profiles of transcripts recovered from each station revealed partitioning that matched the gradient of Polar Water influence across the sampled region that correlated with environmental variables approaching the traditional threshold for statistical significance. This partitioning was resolved to be the result of different gene expression profiles of similar gene compositions between stations, implying each station group was genetically distinct. Data suggested that predicted increased Atlantic Water influence in the region may result in the displacement of gene profiles expressed by cold water associated communities by those expressed by warm water associated communities.

Some functional differences were observed between station groups which suggested a possible increase of metabolic processes related to the KEGG pathways ‘carbohydrate metabolism’, ‘metabolism of other amino acids’ and ‘nucleotide metabolism’ under increased Atlantic Water influence. Additionally, differences in the expression profiles of individual genes support the suggested alterations to metabolic processes, including photosynthesis. Furthermore, certain members of the Excavata and Opisthokonta will likely experience reduced functionality, and functionality of certain members of the Cryptophyta will increase under increased Atlantic influence.

Contrasting functional changes between and within individual taxonomic groups, and of individual genes involved in key function processes, implies that the functional response of members of the eukaryotic community to increased Atlantic influence may be complex. Such changes may impact key ecosystem services including primary production, nutrient cycling, biogeochemical cycles and the carbon cycle. However, overall functionality was observed to be largely conserved between station communities, implying perturbations to overall community functionality between station groups may be minimal.

The findings presented in this chapter have important implications for the ecosystem functions performed by the microbial eukaryotic communities present in the sampled region under climate driven warming, and supports the analysis performed in Chapter 5. The identification of a full suite of KEGG pathways was not recovered for any taxonomic group. However, an inability to identify the presence of a functional pathway does not mean it is not occurring, as limitations within the assembly, sequence depth, current taxonomy databases and the KEGG annotation database will all restrict the number of functional annotations that can be made. Therefore, the possibility that the eukaryotic communities are carrying out additional diverse arrays of functions which remain undetected cannot be excluded, and continued study of the Arctic region is encouraged.

Chapter 7

Synthesis

7.1 A First Insight into Predicted Biological Impacts of Arctic Atlantification

Changes to the global marine habitat as a result of climate change are well documented[182,188,189], and are significantly amplified in the Arctic[191]. The reported rapid loss of sea ice[189] is acting to weaken stratification, diminishing the boundaries between the temperate Atlantic and cold Arctic climate waters[199], enabling increased intrusion of Atlantic Waters into Polar Waters[198,199]. However, the responses of extant microbial communities within these mixing regions to such environmental change and the introduction of novel competition remain largely unknown. The work presented in this thesis contributes to the development and progression of the field by beginning to address this knowledge gap. The community partitioning presented in Chapter 4 and Chapter 5 represent novel findings, which in addition to the observed functional profiles presented in Chapter 6 provides one of the first insights into the potential response of extant microbial communities to increasing Atlantic Water influence in the the Arctic region and resultant impacts to ecosystem functionality.

7.1.1 Main Findings

The research presented in this thesis aimed to explore the extant marine microbial communities present across a gradient of Polar Water influence to assess whether such communities may be impacted by predicted increased Atlantic Water influence and environmental perturbations within the Arctic region. This was accomplished by achieving the aims of each chapter presented in this thesis, as summarised below with the main findings of each chapter:

Chapter 3 - Analysis of the environmental conditions within the sampled region were completed that allowed the classification of the sampled stations into regional groups reflective of high, moderate, and low amounts of Polar Water influence which acted to freshen and cool local water conditions at each station, achieving the aim of this chapter.

Chapter 4 - The aim of this chapter was to examine whether the bacterial communities present across a transect of stations featuring different amounts of Polar Water influence displayed partitioning to provide an insight into how these communities may be impacted by predicted future environmental change in the Arctic region. This aim was achieved as partitioning of the bacterial community was observed at all abundance fractions, thus the hypothesis for Chapter 4 (as stated in section 1.9) can be accepted. Data suggests that, under increased Atlantic mixing and warming of the Arctic region, cold water associated bacterial taxa will likely be displaced by temperate water associated taxa (Figure 7.1).

Chapter 5 - The aim of this chapter was to examine whether the microbial eukaryotic communities present across a transect of stations featuring different amounts of Polar Water influence displayed partitioning to provide an insight into how these communities will be impacted by predicted future environmental change to the Arctic region. This aim was achieved as partitioning was observed within the microbial eukaryotic community, thus the hypothesis for this Chapter (as stated in section 1.9) can be accepted. However, in contrast to the bacterial community the partitioning of the microbial eukaryotic community was observed to be more complex and implies the potential for different responses to be displayed by different abundance fractions of the eukaryotic community (Figure 7.2).

Chapter 6 – The aim of this chapter was to determine if the functional potential of the active microbial eukaryotic community members differed across the regional station groups that were resolved to feature differing amounts of Polar Water influence. This aim was achieved as analysis of metatranscriptomes derived from the whole microbial eukaryotic community resolved different gene expression profiles to be present at different station groups which resulted in the expression of different KEGG functional profiles, thus the hypothesis for this Chapter (as stated in section 1.9) can be accepted.

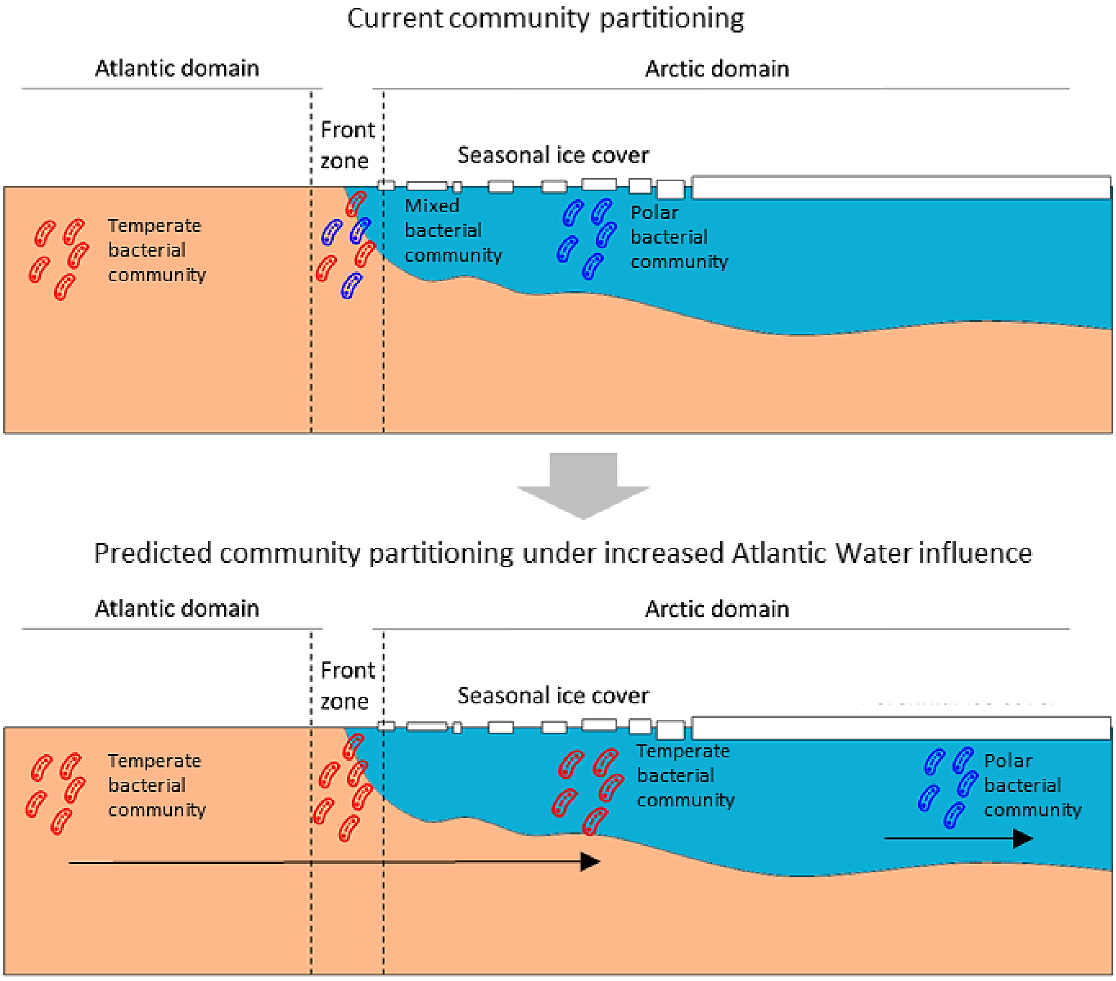

Current community partitioning
Atlantic domain
Arctic domain
Front zone
Seasonal ice cover
Temperate bacterial community
Mixed bacterial community
Polar bacterial community
Predicted community partitioning under increased Atlantic Water influence
Atlantic domain
Arctic domain
Front zone
Seasonal ice cover
Temperate bacterial community
Temperate bacterial community
Polar bacterial community

Figure 7.1. The predicted bacterial community partitioning under increased Atlantic Water influence.

Shown is the current bacterial community partitioning and the predicted displacement of Polar Water associated bacterial communities located in the Arctic domain by temperate water associated communities within the Atlantic domain under increased Atlantic Water influence. Arrows show the direction of displacement. Communities are colour coded; red – temperate water associated, blue – Polar Water associated. Water masses are colour coded; orange – temperate waters within the Atlantic domain, blue – Polar Waters within the Arctic domain.

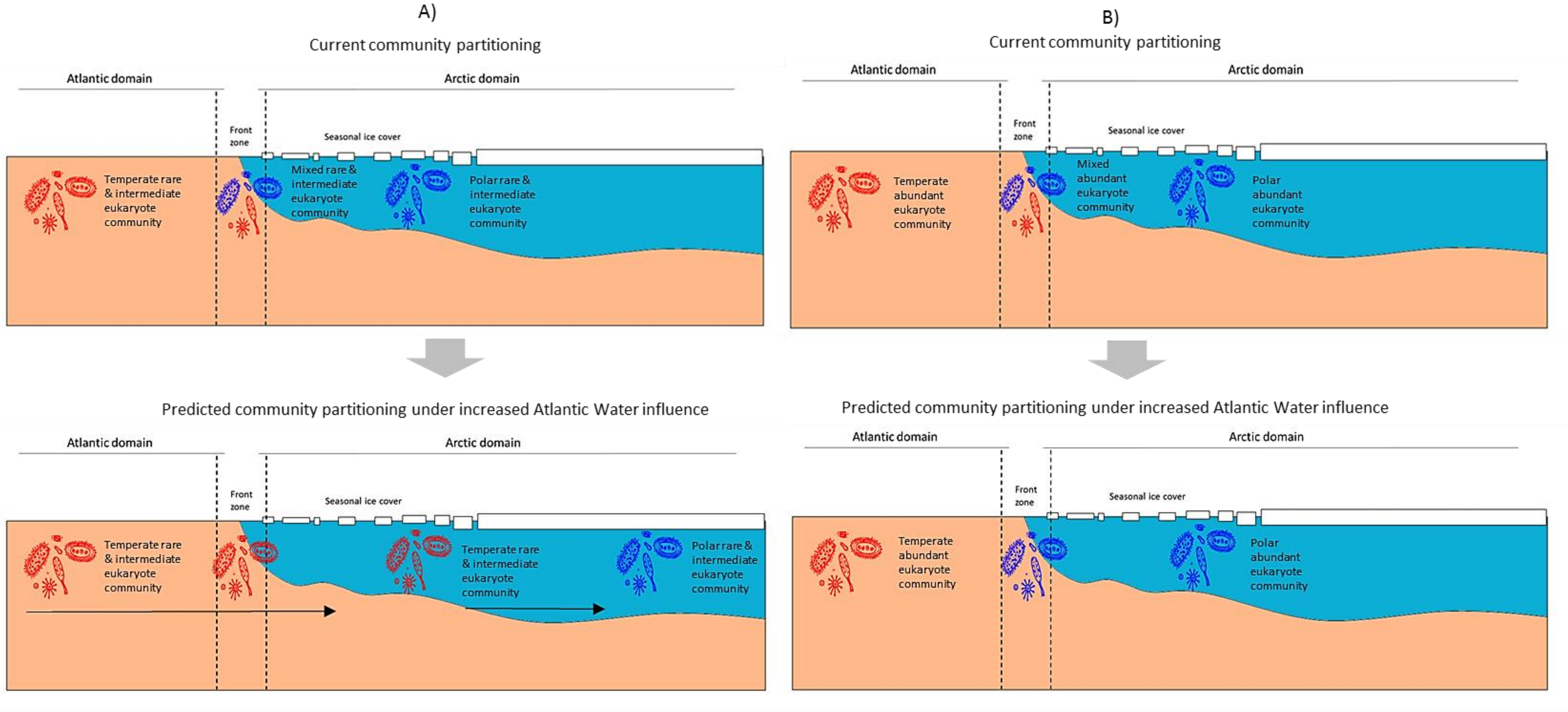

A)
Current community partitioning
Atlantic domain
Arctic domain
Front zone
Seasonal ice cover
Temperate rare & intermediate eukaryote community
Mixed rare & intermediate eukaryote community
Polar rare & intermediate eukaryote community
B)
Current community partitioning
Atlantic domain
Arctic domain
Front zone
Seasonal ice cover
Temperate abundant eukaryote community
Mixed abundant eukaryote community
Polar abundant eukaryote community
Predicted community partitioning under increased Atlantic Water influence
Atlantic domain
Arctic domain
Front zone
Seasonal ice cover
Temperate rare & intermediate eukaryote community
Temperate rare & intermediate eukaryote community
Polar rare & intermediate eukaryote community
Predicted community partitioning under increased Atlantic Water influence
Atlantic domain
Arctic domain
Front zone
Seasonal ice cover
Temperate abundant eukaryote community
Polar abundant eukaryote community

Figure 7.2. The predicted eukaryote community partitioning under increased Atlantic Water influence.

Shown is the current eukaryotic community partitioning and the predicted partitioning under increased Atlantic Water influence for A) the rare and intermediate abundance fractions, and B) the abundant fraction of the community. Arrows show the direction of displacement. Communities are colour coded; red – temperate water associated, blue – Polar Water associated. Water masses are colour coded; orange – temperate waters within the Atlantic domain, blue – Polar Waters within the Arctic domain.

7.1.2 Potential Underlying Mechanisms of Community Partitioning

The bacterial community present at the MI station was observed to be most similar to those found at the LI stations, and the same trend was observed for a number of constituent taxonomic groups. The consistency of bacterial partitioning observed across the majority of taxonomic groups may be due to a number of factors. A more stringent sequence similarity threshold was applied than conventionally used in current studies, meant to be analogous to species level relationships between taxa[163], but despite the threshold being chosen based on guidance from the literature[163] such thresholds are still arbitrary values. These thresholds are unlikely to be universally applicable across all bacterial groups[359]. For example, some strains are reported to display over 99.5% sequence similarity, yet belong to different species[360]. Additionally, sub-species level, and intra OTU genomic variations have been demonstrated to feature distinct patterns of succession and are suggested to be of ecological significance[58,361]. As such, it is possible that bacterial ecotypes existed within the dataset which displayed unique ecologically significant patterns which could not be resolved by a sequence similarity threshold of 98.7%. However, analysis at the finest scale of resolution commonly applied in NGS studies, the OTU, did resolve distribution patterns of specific OTUs that were consistent with those observed at higher taxonomic levels, and examples of OTUs displaying station specific growth responses were present in all taxonomic groups except the Betaproteobacteria. It is possible that inherent features of bacterial taxa such as high levels of genetic diversity, high mutation rates, high community turnover rates and broad metabolic potential result in similar structuring and partitioning across different taxonomic groups. Reports within the literature that abundant bacterial classes have been reported as cosmopolitan[209], and that distributions of major taxonomic groups appear analogous to one another[39] provides some evidence in support of this proposition.

The observed partitioning of the bacterial and eukaryotic communities is analogous to taxonomic biogeography, whereby taxa show specific geographical ranges associated with favourable conditions. Phytoplankton communities are known to display biogeographic distributions between ocean habitats as a result of taxon specific traits and biological interactions[42,49,150], and the geographical range of a number of phytoplankton taxa have been demonstrated to be experiencing northward range shifts in response to climate change[125,126,206]. One explanatory mechanism of these poleward expansions is that of thermal niche width, which suggests communities follow habitats close to their thermal optima, that under raising global temperatures moves those in the northern hemisphere towards high latitudes[362]. Interestingly, it is also suggested that the niche of polar taxa remains constant, instead resulting is a reduction in habitat range and susceptibility to loss or displacement. Such a mechanism may underpin the partitioning observed in cases where the MI communities were most similar to LI communities, suggesting a displacement of HI associated taxa[362]. However it does not explain why some eukaryotic HI communities were predicted to dominate under increased Atlantic mixing. It might be that these communities feature a greater composition of specialised taxa, able to outcompete more generalist Atlantic associated taxa under specific conditions[27]. Indeed, the HI stations were dominated by Diatoms which are known to be particularly well adapted to Polar Waters[96], and included representatives of *Chaetoceros, Fragilariopsis* and *Thalassiosira,* some of which are known to contain cold adapted ecotypes[97,363]*.* Unfortunately, the majority of Diatom OTUs recovered were of uncultured isolates, and so it is unclear whether they are specifically associated with cold waters.

7.1.3 Potential Variability of Susceptibility to Atlantic Water Influence

In addition to addressing how marine microbial communities may be impacted by predicted future environmental changes to the Arctic region, comparison of the results from the bacterial and eukaryotic analyses allows novel suggestions to be made as to the relative potential suspectibility of each community to such change.

Differences between the similarity of the LI and HI stations were observed between the bacterial and eukaryotic communities, being 60% dissimilar for the whole bacterial community (Figure 4.2A) and 70% for the whole eukaryotic community (excluding Copepods) (Figure 4.2C). Similar trends were observed for the rare and intermediate abundance fractions, with the bacterial dataset being 65% and 75% dissimilar for the rare and intermediate abundance fractions respectively (Figure 4.2D and 4.2E), and 75% and 85% for the eukaryotic rare and intermediate abundance fractions respectively (Figure 5.2E and 5.2F). The observed greater dissimilarity between the eukaryotic communities at the LI and HI stations suggests that they feature greater genetic distance between station groups, possibly due to a stronger response to environmental selection pressures, and thus the eukaryotic community may be more susceptible to future increases of Atlantic Water influence than the bacterial community. Indeed, greater responses to environmental perturbations within the Arctic has been reported for eukaryotes[217].

Differences were observed between the structure and composition of the bacterial and eukaryotic communities which further suggest that the response of each community to increased Atlantic Water influence may differ. The bacterial community was seen to feature a high number station specific OTUs (Figure 4.6), the bacterial communities at the HI stations were compositionally the most distinct (Figure 4.6D), and featured the lowest Shannon index implying the presence of just a few abundant bacterial taxa but many rare taxa (Table 4.2). Therefore, the bacterial community may be more susceptible to the loss of local OTUs from each station under increased Atlantic Water influence as their environmental niche is displaced. The decline of bacterial diversity, particularly of rare OTUs, observed during environmental perturbations[217] support this conclusion. In contrast, the eukaryotic community appeared less compositionally distinct between station groups (Figure 5.6) and a greater number of OTUs were observed which were present at all stations but that displayed changes in abundance between station groups (Figure 5.7). This data suggests that eukaryotic OTUs may experience a lower level of local loss compared to the bacterial community, instead displaying changes in OTU abundance between station groups. Time series data would be needed to confirm these speculations.

7.1.4 Partitioning of the Rare Community

The rare microbial eukaryotic community fraction is poorly explored in current literature in comparison to the rare bacteria fraction, and information relating to rare microbial eukaryotes from Arctic waters is especially limited. The findings presented in this thesis provide a valuable insight into responses of these taxa to environmental change. Studies of temperate coastal waters have previously demonstrated that the rare fraction of microbial communities may display biogeography likely the result of environmental factors, or similar structuring compared to abundant fraction[53,56]. However, the detection of similar biogeographic structuring between rare and abundant community fractions, but likely contrasting responses to environmental change, within this study appears to represent a novel observation.

The contrasting partitioning observed between the abundant and rare fractions of the eukaryotic community might have interesting ecological consequences. Under increased Atlantic influence data suggests that rare eukaryotic taxa from the LI stations will displace those from HI stations, but that the abundant eukaryotic HI communities will dominate, resulting in the potential formation of a novel community composition in the region. The rare fraction of the community is often suggested to contain low abundance opportunistic taxa that are capable of increasing in abundance under perturbations to environmental conditions in order to maintain important ecological functions[55]. Therefore, if a transition to a community containing novel rare taxa occurs at the HI stations, and they are functionally different to those that have been displaced, then alterations to ecosystem functionality and ecosystem service resilience are possible.

7.1.5 Predicted Future Changes to Physical Drivers of Partitioning

Environmental factors were seen to explain the majority of the community variation observed across the sampled region for both bacterial and eukaryotic datasets, and individual OTUs were observed to display correlations with environmental variables. Under predicted future climate models regional environmental factors are expected to undergo further changes in the Arctic. As of 2018 sea ice has declined to 66% below the 1981-2010 long-term average extent, representing a loss of 340,000 square miles of ice cover[189], with impacts to regional salinity and ice albedo feedback mechanisms further enhancing open water warming[183]. Atmospheric forcing is thought to be one of the main factors driving these changes, and global CO_2 levels are expected to continue to rise under projected RCPs, resulting in further increases in global surface temperature[7]. Therefore, similar community partitioning to that observed is likely to become common across larger regional scales within the Arctic, and the genetic distance observed between regional station groups is likely to become impacted further as regional environmental conditions experience increased perturbations. Speculating whether the genetic distance between regional station groups will increase further or decrease is difficult, and would require long term monitoring to resolve. On the one hand increased climate driven warming of intruding temperate waters, but cooling of Polar Waters from increased melt water resultant from further sea ice loss may increase the dissimilarity of regional conditions, and thus genetic distance between respective inhabiting microbial communities. However, the loss of sea ice extent may result in a reduced capacity for local freshening and increased regional warming of polar habitats that may diminish the distinctness of environmental conditions between regional conditions over a longer time scale, or larger spatial scale, and eventually result in increased genetic similarity between respective communities[198].

7.1.6 Additional Potential Functional Impacts of Alterations to Community Composition

Some of the largest compositional differences between stations were present for the Dinoflagellata and Stramenopiles, the latter of which was primarily represented by Diatoms. Diatoms feature an external silica frustule which is suggested to provide protection from copepod grazing, and to tube-feeding Dinoflagellates that feed by piercing the cell of the prey to such out it's contents[364]. Diatom and Dinoflagellate cell sizes vary significantly by species[95,103] and grazers may display feeding selectivity, targeting prey within certain size classes[365]. Additionally, certain taxa are known to display taxonomic based feeding selectivity, which may result in community wide impacts such as increased abundance of toxic species[366] or enhanced bloom formation of *Phaeocystis* due preferential feeding on competitor species[367]*.* The Dinoflagellate *Gymnodinium* has a size range of 5-200 μm[368], and *Azadinium* 10-20 μm[369], both were predicted to increase in abundance. By contrast *Micromonas, Phaeocystis* and *Chaetoceros we*re predicted to decrease, and are much smaller, displaying sizes of ~2 μm[370], 4-8 μm[371] and 2-45 μm[368] respectively, the latter of which also exudes a gelatinous matrix as a grazing defence[84]. *Thalassiosira* may be 2-186 μm[372] and is also predicted to decrease in abundance. While representatives of these genus were environmental isolates and could not be resolved to species level preventing exact sizes to be determined, the potential size range displayed for each genus makes it likely that size class differences between HI and LI station communities existed. Thus, the predicted changes to community composition may affect regional food webs through alterations to community size class linked feeding regimes or specific taxon based feeding selectivity, with impacts upon higher trophic levels[98].

The predicted decline of the abundance of certain taxa under increased Atlantic Water influence raises concerns over the loss of potentially "specialist", or endemic taxa from the HI stations. OTUs were identified in Chapter 4 and 5 from genus known to feature distinct Arctic ecotypes, including *Micromonas, Fragilariopsis*[97,261], and *Polaribacter*[245]. Furthermore, potentially specialist OTUs displaying strong abundance peaks, or that were specific only to certain station groups were observed, possibly reflecting OTUs specialised to particular environmental niches. Additionally, during transcriptomic analyses genes were identified which suggested specialist functionality within the community. Maintaining a high diversity of specialist species has been shown to provide a greater level of functional redundancy to buffer and resist environmental disturbance than that of a community composed of generalists[26], with many species needed to maintain ecosystem functioning, and a particular function often highly impacted by one dominant specialist[22]. Therefore, alterations to the abundance of, or loss of, potentially “specialist” taxa may impact local ecosystem functionality. For example, the predicted reduction of functionality of taxa at the HI stations suggests the expression of *ACA1_097000* could potentially decline, resulting in reduced capacity for specialist carbon source utilisation[324] with impacts to local carbon cycling.

7.1.7 Potential Changes to the Functionality of Additional Individual Taxonomic Groups

The partitioning observed in Chapter 5 was used to guide suggestions of potential changes to the functionality of individual taxonomic groups. However, a number of taxonomic groups that were resolved to display regional clustering within Chapter 5 (Figure 5.6) were not recovered in the transcriptomic dataset (Figure 6.3), meaning commenting on the potential direction of functional changes for the remaining groups may not be robust. However, some speculations can be made with the support of current literature. The Haptophyta were resolved to be one of the most functionally diverse taxonomic groups recovered, displaying activity in 17 of the 23 KEGG functional pathways identified in the community (Figure 6.11). The highest levels of functional expression were observed to be expressed at the HI station CTD62 for the majority of functional pathways, except for energy metabolism. The Haptophyta include species that are highly competitive in Polar Waters[84], including potentially distinct polar ecotypes[44,85]. Database limitations restricted the taxonomic resolution that it was possible to obtain for the recovered Haptophytes sequences, but the higher expression of KEGG pathways at the HI stations observed (Figure 6.10) may be the result of such polar ecotypes or cold adapted species. As Atlantic ecotypes continue their poleward expansion[206] such polar associated ecotypes might be lost, leading to reduced functionality of the majority of the recovered Haptophyte KEGG pathways, with impacts to biogeochemical cycling, primary production and bloom dynamics.

Similar trends were observed for the Stramenopiles, with the majority of KEGG functional pathways also expressed at the HI stations (Figure 6.11). Again, database limitations restricted the taxonomic resolution that was possible, but suggested the presence of Diatoms within the community. The abundance of Stramenopiles is known to be declining as the environmental conditions within Arctic Waters are perturbed[217], and Diatoms have been observed to display one of the greatest response to warming within phytoplankton communities[286]. Therefore, it is reasonable to speculate that the functionality of Stramenopiles within the sampled region may decrease under increased Atlantic Water influence, with potential impacts to ecosystem services including global primary production, the silica cycle[289] and carbon export[290,291].

7.2 Limitations

The present study featured one sampled time point, as a result it provides a snapshot of the microbial communities at a specific point in time and was unable to resolve temporal patterns of species responses to increased Atlantic Water influence. If time series data had been available it would have been possible to validate the suggested compositional changes on an annual basis, and quantify the extent of abundance changes or localised loss of select taxa from each regional station group. Furthermore, correlations of microbial communities with environmental factors such as temperature and salinity were observed, but as these factors display seasonal patterns[373], and microbial communities are known to display seasonal patterns of succession[218,262] (as highlighted in sections 4.1 and 5.1), there may be inter-annual or intra-annual responses which it has not been possible to resolve.

The analyses performed in this thesis were subject to certain constraints. The most significant of which was the lack of biological replicate samples. Replicate samples were not available for each CTD cast, therefore it was not possible to check and control for compositional differences in samples as a result of potential sampling bias. A lack of replicate samples also limited the ability to perform differential gene expression analysis across the sampled region in Chapter 6. To perform differential expression analysis samples need to be grouped, with each group specified programmatically within the experimental design, each sample is then normalised based on library size and the significance of differences in gene expression between each sample group determined[178]. Without replicates, each sample would be the sole representative of its own group, leading to a high false positive rate of differentially expressed genes.

Limitations also existed within Chapter 4 and 5. Care was taken to select primers designed to be universal in their recovery of taxonomic groups but some taxa may still have potentially been omitted as a result of the chosen primer combinations[359]. For example, highly divergent taxa have been identified which feature unique 16S rRNA sequences which remain undetectable by conventional sequencing approaches[374]. This unknown fraction of the community is often referred to as "microbial dark matter" (MDM), and is suggested may represent a significant fraction of the community, up to 15% of the bacterial domain[374]. Indeed, it has been suggested MDM could represent the majority of microbial diversity[375]. Therefore, current NGS sequencing techniques may in fact only be reporting upon a small fraction of true environmental communities, limiting the relevance of current findings, and excluding potentially ecologically significant taxa.

Limitations within currently available online sequence databases may have impacted the diversity it was possible to recover. A lack of representation of marine environmental samples was particularly limiting during transcriptome analysis, and should be a key focus of future research. Furthermore, submissions to online sequence databases originate from the research community, often with little curation, and thus are subject to user input error or inclusion of incorrect sequences[376]. These limitations are compounded as environmentally derived samples may be uploaded with little or conflicting taxonomic annotations. Furthermore, taxonomic classifications are dynamic, and frequently the position of different groups comes under revision leading to ambiguous positioning within the hierarchies used within these databases. The extent of the discussion surrounding correct taxonomic placement for taxa is highlighted by the debate regarding the number of domains of life, the most basic level of taxonomic classification, that should exist[377,378]. Throughout this study care has been taken to best to mitigate these issues through the manual validation of assigned OTU taxonomy against current literature, by selecting unique taxonomic groups at analogous levels within the SILVA database, and selecting those groups with known and well documented ecological significance. For example, Haptophyta were long considered part of the Stramenopiles, but this has recently been contested[379], and currently the Haptophyta have been placed as an unclassified group in online databases pending further analysis. As this is a globally significant group with important biogeochemical and ecosystem impacts[380], it was prudent to include it as one of the major taxonomic groups for analysis despite a lack of defined taxonomic positioning. Despite such care, it is still possible that some inaccuracies may have remained.

Another possibility is that there are still formally recognised taxonomic groups omitted from the databases used for taxonomic annotation. Results presented in chapter 6 demonstrated representative genes from the Apusozoa were recovered. However, no representative sequences of Apusozoa were recovered during prior eukaryotic analysis (Chapter 5). Manual validation confirmed that the SILVA database does not contain any representative sequences for this group at the time of writing, preventing their detection, and it is possible that additional taxonomic groups may have been excluded from the analysis due to a similar lack of representation in online databases.

Full saturation was not reached during the rarefaction analysis for either the bacterial or eukaryotic datasets, and thus it is possible that some rare taxa have been omitted from the analyses. Despite this, the analyses performed are still informative for discussing the abundance and distributions of the taxa recovered. Improvements to sequence depth, assembly, and sequence quality will always lead to improved detection of taxa and more robust analyses, but often time and financial constraints mean such improvements aren't possible and in most cases (excluding insufficient sequence quality due to technical errors) aren't necessary.

7.3 Future Directions and Perspective

The effect of temperature upon marine phytoplankton communities is often a key research focus, especially within studies pertaining to large geographic scales, but regional changes in nutrient concentrations, circulation and hydrography are all also likely to have profound impacts to community composition and functionality. This is especially important when considering local or regional scale communities, as local factors which operate over smaller spatial scales, such as organic matter input from melt water[381] and local nutrient regimes, are likely to have significant impacts upon community structure. Additionally, microbial communities display seasonal changes in community composition[133,218,262,263] and symbiotic relationships between constituent taxa. Therefore, it would be valuable to consider the impacts of a broader range of factors upon microbial communities[382], and over longer temporal scales to further resolve the impacts of future environmental change in the Arctic.

16S and 18S rRNA gene diversity and activity are powerful tools in identifying the taxa present, and their abundance, within environmental microbial communities. However, they are limited in their ability to recover functional information and abundance differences may be observed depending on the targeted gene regions[144], limiting comparability between studies. Furthermore, such techniques recover dead and dormant cells which may be of little functional relevance[53]. Metagenomic and metatranscritpomic techniques represent more powerful techniques to uncover the functional potential within microbial communities and determine ecologically significant patterns in the face of environmental perturbations. However, these are limited by the lack of characterisation of the majority of identified genes in online databases, limiting the information that can be obtained and the potential power of these methods. Addressing these limitations through the isolation, culture, and study of functional traits from environmental taxa is key to identifying functional traits and linking them to ecological processes.

While the techniques used to assign sequences into OTUs in this thesis followed commonly used methods or justified cut off values[163], and phylogenetic composition is thought to be a greater determinant of ecological function than variation in OTUs[383], the diversity and complexity of environmental niches marine microbial communities experience mean it is possible that the functionality assigned to these communities may be affected by the taxonomic resolution used to analyse them. Amplicon sequence variants (ASVs) are genetically different sequences within OTU similarity thresholds that can be resolved to single nucleotide differences. ASVs have already been shown to highly prevalent in phytoplankton communities, suggested to be of ecological relevance to resilience, niche differentiation or substrate usage[58]. Therefore, the potential power of ASVs in resolving ecologically important ecotypes is significant[384], and would allow the differentiation and monitoring of microbial variants under environmental perturbations to better resolve potential resultant impacts to ecosystem functionality.

Appendix

Appendix 1. The Correlation of all Top 200 Most Abundant Bacterial OTUs with Each Environmental Factor.

Environmental data used was measured from the depth at which water samples were taken at each station during CTD casts. R_s represents the correlation coefficient, values ≥0.9 represent a strong positive association, values ≤-0.9 represent a strong negative association. P represents the p-value for each correlation, values ≤0.05 are considered statistically significant. The taxonomic annotation of each OTUID is displayed below.

OTUID	DO_2		Salinity		SubSurVPAR		Temperature		Ammonium		Nitrate		Silicate		Phosphate	
	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P
OTU106	0.60	0.28	-0.90	0.04	-0.50	0.39	-0.90	0.04	0.90	0.04	0.20	0.75	-0.60	0.28	0.60	0.28
OTU118	0.60	0.28	-0.90	0.04	-0.50	0.39	-0.90	0.04	0.90	0.04	0.20	0.75	-0.60	0.28	0.67	0.22
OTU14206	-0.56	0.32	0.67	0.22	0.21	0.74	0.67	0.22	-0.67	0.22	0.15	0.80	0.87	0.05	0.60	0.28
OTU155	0.80	0.10	-1.00	0.00	-0.80	0.10	-1.00	0.00	1.00	0.00	0.60	0.28	-0.50	0.39	-0.70	0.19
OTU1635	0.72	0.17	-0.67	0.22	-0.56	0.32	-0.67	0.22	0.67	0.22	0.21	0.74	-0.87	0.05	-0.70	0.19
OTU1692	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU1698	0.78	0.12	-0.89	0.04	-0.67	0.22	-0.89	0.04	0.89	0.04	0.34	0.58	-0.78	0.12	-0.90	0.04
OTU17739	-0.90	0.04	0.90	0.04	0.60	0.28	0.90	0.04	-0.90	0.04	-0.50	0.39	0.60	0.28	-0.70	0.19
OTU2171	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.60	0.28
OTU387	0.87	0.05	-0.97	0.00	-0.87	0.05	-0.97	0.00	0.97	0.00	0.67	0.22	-0.56	0.32	0.70	0.19
OTU4	-0.70	0.19	0.80	0.10	1.00	0.00	0.80	0.10	-0.80	0.10	-0.90	0.04	0.20	0.75	-0.60	0.28
OTU44	-0.50	0.39	0.70	0.19	0.80	0.10	0.70	0.19	-0.70	0.19	-0.90	0.04	-0.20	0.75	0.60	0.28
OTU4674	0.97	0.00	-0.87	0.05	-0.67	0.22	-0.87	0.05	0.87	0.05	0.56	0.32	-0.67	0.22	-0.60	0.28
OTU49	0.97	0.00	-0.87	0.05	-0.67	0.22	-0.87	0.05	0.87	0.05	0.56	0.32	-0.67	0.22	-0.60	0.28
OTU602	-0.97	0.00	0.87	0.05	0.82	0.09	0.87	0.05	-0.87	0.05	-0.67	0.22	0.67	0.22	-0.70	0.19
OTU624	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	0.60	0.28
OTU696	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28

OTU783	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	-0.40	0.50
OTU79	0.87	0.05	-0.97	0.00	-0.87	0.05	-0.97	0.00	0.97	0.00	0.67	0.22	-0.56	0.32	0.60	0.28
OTU7977	0.78	0.12	-0.89	0.04	-0.67	0.22	-0.89	0.04	0.89	0.04	0.34	0.58	-0.78	0.12	0.67	0.22
OTU820	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	-0.90	0.04
OTU877	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	0.60	0.28
OTU9	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	-0.60	0.28
OTU93	0.72	0.17	-0.97	0.00	-0.67	0.22	-0.97	0.00	0.97	0.00	0.41	0.49	-0.56	0.32	0.70	0.19
OTU97	-0.60	0.28	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.80	0.10	0.10	0.87	0.60	0.28
OTU1009	0.70	0.19	-0.50	0.39	-0.20	0.75	-0.50	0.39	0.50	0.39	-0.10	0.87	-1.00	0.00	-0.60	0.28
OTU1025	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.40	0.50
OTU10411	0.90	0.04	-0.90	0.04	-0.60	0.28	-0.90	0.04	0.90	0.04	0.50	0.39	-0.60	0.28	0.60	0.28
OTU1056	0.90	0.04	-0.60	0.28	-0.50	0.39	-0.60	0.28	0.60	0.28	0.30	0.62	-0.90	0.04	0.60	0.28
OTU109	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.60	0.28
OTU110	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.34	0.58
OTU112	0.90	0.04	-0.90	0.04	-0.60	0.28	-0.90	0.04	0.90	0.04	0.50	0.39	-0.60	0.28	0.34	0.58
OTU114	0.50	0.39	-0.10	0.87	0.10	0.87	-0.10	0.87	0.10	0.87	-0.30	0.62	-0.90	0.04	0.60	0.28
OTU1202	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	-0.50	0.39
OTU127	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	0.10	0.87
OTU129	-0.80	0.10	1.00	0.00	0.80	0.10	1.00	0.00	-1.00	0.00	-0.60	0.28	0.50	0.39	-0.20	0.75
OTU131	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.20	0.75
OTU1317	-0.72	0.17	0.67	0.22	0.56	0.32	0.67	0.22	-0.67	0.22	-0.21	0.74	0.87	0.05	-0.20	0.75
OTU1322	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.80	0.10
OTU1327	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.20	0.75
OTU135	0.70	0.19	-0.50	0.39	-0.20	0.75	-0.50	0.39	0.50	0.39	-0.10	0.87	-1.00	0.00	-0.30	0.62
OTU1376	-0.40	0.50	0.50	0.39	0.90	0.04	0.50	0.39	-0.50	0.39	-0.80	0.10	0.00	1.00	0.00	1.00
OTU1441	-0.60	0.28	0.90	0.04	0.50	0.39	0.90	0.04	-0.90	0.04	-0.20	0.75	0.60	0.28	0.10	0.87
OTU1464	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	0.00	1.00

OTU1467	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	-0.10	0.87
OTU1477	0.50	0.39	-0.70	0.19	-0.80	0.10	-0.70	0.19	0.70	0.19	0.90	0.04	0.20	0.75	-0.80	0.10
OTU1496	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	0.15	0.80
OTU1575	0.90	0.04	-0.60	0.28	-0.50	0.39	-0.60	0.28	0.60	0.28	0.30	0.62	-0.90	0.04	-0.30	0.62
OTU1621	0.90	0.04	-0.60	0.28	-0.50	0.39	-0.60	0.28	0.60	0.28	0.30	0.62	-0.90	0.04	0.10	0.87
OTU1623	0.70	0.19	-0.80	0.10	-1.00	0.00	-0.80	0.10	0.80	0.10	0.90	0.04	-0.20	0.75	-0.80	0.10
OTU170	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.00	1.00
OTU1731	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	-0.56	0.32
OTU1747	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.50	0.39
OTU1771	-0.87	0.05	0.67	0.22	0.62	0.27	0.67	0.22	-0.67	0.22	-0.36	0.55	0.87	0.05	0.10	0.87
OTU18	0.50	0.39	-0.10	0.87	0.10	0.87	-0.10	0.87	0.10	0.87	-0.30	0.62	-0.90	0.04	0.82	0.09
OTU1809	-0.20	0.75	0.70	0.19	0.30	0.62	0.70	0.19	-0.70	0.19	0.00	1.00	0.30	0.62	-0.90	0.04
OTU181	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.60	0.28
OTU183	0.30	0.62	-0.30	0.62	-0.80	0.10	-0.30	0.62	0.30	0.62	0.90	0.04	0.30	0.62	0.30	0.62
OTU185	0.90	0.04	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.70	0.19	-0.40	0.50	0.70	0.19
OTU1937	-0.60	0.28	0.60	0.28	0.10	0.87	0.60	0.28	-0.60	0.28	0.20	0.75	0.90	0.04	-1.00	0.00
OTU194	0.80	0.10	-1.00	0.00	-0.80	0.10	-1.00	0.00	1.00	0.00	0.60	0.28	-0.50	0.39	0.10	0.87
OTU1949	0.60	0.28	-0.60	0.28	-0.10	0.87	-0.60	0.28	0.60	0.28	-0.20	0.75	-0.90	0.04	0.15	0.80
OTU2088	-0.72	0.17	0.97	0.00	0.67	0.22	0.97	0.00	-0.97	0.00	-0.41	0.49	0.56	0.32	-0.30	0.62
OTU21	0.70	0.19	-0.50	0.39	-0.20	0.75	-0.50	0.39	0.50	0.39	-0.10	0.87	-1.00	0.00	-0.60	0.28
OTU2100	0.90	0.04	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.70	0.19	-0.40	0.50	0.10	0.87
OTU2111	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.30	0.62
OTU218	-0.60	0.28	0.60	0.28	0.10	0.87	0.60	0.28	-0.60	0.28	0.20	0.75	0.90	0.04	-0.30	0.62
OTU2191	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.05	0.93
OTU220	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.56	0.32
OTU232	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	0.20	0.75
OTU233	0.80	0.10	-1.00	0.00	-0.80	0.10	-1.00	0.00	1.00	0.00	0.60	0.28	-0.50	0.39	0.30	0.62

OTU235	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	0.10	0.87
OTU2430	0.70	0.19	-0.50	0.39	-0.20	0.75	-0.50	0.39	0.50	0.39	-0.10	0.87	-1.00	0.00	0.60	0.28
OTU25	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.50	0.39
OTU269	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	0.05	0.93
OTU2723	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.15	0.80
OTU2765	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	-0.56	0.32
OTU278	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.00	1.00
OTU2816	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.50	0.39
OTU287	0.70	0.19	-0.50	0.39	-0.20	0.75	-0.50	0.39	0.50	0.39	-0.10	0.87	-1.00	0.00	0.10	0.87
OTU289	0.50	0.39	-0.70	0.19	-0.80	0.10	-0.70	0.19	0.70	0.19	0.90	0.04	0.20	0.75	-0.50	0.39
OTU2913	0.40	0.50	-0.50	0.39	-0.90	0.04	-0.50	0.39	0.50	0.39	0.80	0.10	0.00	1.00	0.10	0.87
OTU30	-0.97	0.00	0.87	0.05	0.82	0.09	0.87	0.05	-0.87	0.05	-0.67	0.22	0.67	0.22	0.10	0.87
OTU31	0.50	0.39	-0.10	0.87	0.10	0.87	-0.10	0.87	0.10	0.87	-0.30	0.62	-0.90	0.04	0.60	0.28
OTU314	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU343	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.50	0.39
OTU347	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	0.70	0.19
OTU355	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.41	0.49
OTU36	-0.80	0.10	1.00	0.00	0.80	0.10	1.00	0.00	-1.00	0.00	-0.60	0.28	0.50	0.39	0.30	0.62
OTU3600	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	0.00	1.00
OTU3671	0.20	0.75	0.20	0.75	0.00	1.00	0.20	0.75	-0.20	0.75	0.40	0.50	0.30	0.62	1.00	0.00
OTU368	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	-0.50	0.39
OTU369	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	0.10	0.87
OTU38	-0.70	0.19	0.50	0.39	0.20	0.75	0.50	0.39	-0.50	0.39	0.10	0.87	1.00	0.00	0.10	0.87
OTU4077	0.90	0.04	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.70	0.19	-0.40	0.50	0.00	1.00
OTU41	-0.87	0.05	0.67	0.22	0.62	0.27	0.67	0.22	-0.67	0.22	-0.36	0.55	0.87	0.05	0.05	0.93
OTU413	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.50	0.39
OTU416	0.50	0.39	-0.10	0.87	0.10	0.87	-0.10	0.87	0.10	0.87	-0.30	0.62	-0.90	0.04	0.10	0.87

OTU426	-1.00	0.00	0.80	0.10	0.70	0.19	0.80	0.10	-0.80	0.10	-0.60	0.28	0.70	0.19	-0.50	0.39
OTU4289	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.30	0.62
OTU444	0.50	0.39	-0.10	0.87	0.10	0.87	-0.10	0.87	0.10	0.87	-0.30	0.62	-0.90	0.04	-0.15	0.80
OTU446	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	0.30	0.62
OTU469	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.00	1.00
OTU47	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.15	0.80
OTU470	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	-0.60	0.28
OTU48	0.70	0.19	-0.50	0.39	-0.20	0.75	-0.50	0.39	0.50	0.39	-0.10	0.87	-1.00	0.00	0.70	0.19
OTU487	-0.70	0.19	0.80	0.10	1.00	0.00	0.80	0.10	-0.80	0.10	-0.90	0.04	0.20	0.75	-0.80	0.10
OTU50	-0.60	0.28	0.90	0.04	0.50	0.39	0.90	0.04	-0.90	0.04	-0.20	0.75	0.60	0.28	-0.20	0.75
OTU500	-0.80	0.10	1.00	0.00	0.80	0.10	1.00	0.00	-1.00	0.00	-0.60	0.28	0.50	0.39	-0.30	0.62
OTU539	-0.60	0.28	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.80	0.10	0.10	0.87	-0.50	0.39
OTU546	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.20	0.75
OTU553	-0.50	0.39	0.10	0.87	-0.10	0.87	0.10	0.87	-0.10	0.87	0.30	0.62	0.90	0.04	-0.70	0.19
OTU5535	0.20	0.75	0.20	0.75	0.50	0.39	0.20	0.75	-0.20	0.75	-0.60	0.28	-0.70	0.19	1.00	0.00
OTU564	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.70	0.19
OTU57	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.30	0.62
OTU579	-0.60	0.28	0.60	0.28	0.10	0.87	0.60	0.28	-0.60	0.28	0.20	0.75	0.90	0.04	-0.20	0.75
OTU603	0.78	0.12	-0.89	0.04	-0.67	0.22	-0.89	0.04	0.89	0.04	0.34	0.58	-0.78	0.12	-0.10	0.87
OTU61	0.90	0.04	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.70	0.19	-0.40	0.50	0.15	0.80
OTU617	0.50	0.39	-0.10	0.87	0.10	0.87	-0.10	0.87	0.10	0.87	-0.30	0.62	-0.90	0.04	-0.20	0.75
OTU6172	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU6245	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.15	0.80
OTU6754	0.87	0.05	-0.67	0.22	-0.62	0.27	-0.67	0.22	0.67	0.22	0.36	0.55	-0.87	0.05	-0.15	0.80
OTU7108	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	-0.50	0.39
OTU730	0.80	0.10	-1.00	0.00	-0.80	0.10	-1.00	0.00	1.00	0.00	0.60	0.28	-0.50	0.39	0.10	0.87
OTU733	-0.80	0.10	1.00	0.00	0.80	0.10	1.00	0.00	-1.00	0.00	-0.60	0.28	0.50	0.39	-0.20	0.75

OTU737	-0.80	0.10	1.00	0.00	0.80	0.10	1.00	0.00	-1.00	0.00	-0.60	0.28	0.50	0.39	-0.50	0.39
OTU75	-0.40	0.50	0.50	0.39	0.90	0.04	0.50	0.39	-0.50	0.39	-0.80	0.10	0.00	1.00	0.10	0.87
OTU753	0.87	0.05	-0.82	0.09	-0.82	0.09	-0.82	0.09	0.82	0.09	0.56	0.32	-0.72	0.17	-0.05	0.93
OTU82	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.60	0.28
OTU87	-0.50	0.39	0.10	0.87	-0.10	0.87	0.10	0.87	-0.10	0.87	0.30	0.62	0.90	0.04	0.70	0.19
OTU931	0.82	0.09	-0.56	0.32	-0.36	0.55	-0.56	0.32	0.56	0.32	0.10	0.87	-0.97	0.00	0.56	0.32

OTUID	**Taxonomy**
OTU368	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Porticoccaceae; SAR92 clade; uncultured bacterium
OTU79	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 4; uncultured Flavobacteriia bacterium
OTU278	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 1; unidentified marine bacterioplankton
OTU129	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS4 marine group; uncultured organism
OTU36	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; uncultured; uncultured alpha proteobacterium
OTU93	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU546	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Porticoccaceae; SAR92 clade; uncultured bacterium
OTU539	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS2b marine group; uncultured organism
OTU155	Bacteria; Proteobacteria; Gammaproteobacteria; Thiotrichales; Piscirickettsiaceae; uncultured; uncultured bacterium
OTU500	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR116 clade; uncultured bacterium
OTU106	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; uncultured; uncultured Cellulophaga sp.
OTU235	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; uncultured; uncultured organism
OTU820	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured bacterium
OTU194	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured bacterium
OTU730	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Pseudohongiella; uncultured gamma proteobacterium
OTU232	Bacteria; Proteobacteria; Gammaproteobacteria; Thiotrichales; Thiotrichaceae; Thiothrix; uncultured bacterium

OTU220	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS4 marine group; uncultured organism
OTU269	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured bacterium
OTU1464	Bacteria; Verrucomicrobia; Opitutae; Puniceicoccales; Puniceicoccaceae; marine group; uncultured bacterium
OTU733	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Ascidiaceihabitans; uncultured bacterium
OTU737	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS4 marine group; uncultured bacterium
OTU2191	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 1; uncultured marine bacterium
OTU877	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured bacterium
OTU1441	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR116 clade; uncultured alpha proteobacterium
OTU170	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU387	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; OM182 clade; uncultured bacterium ARCTIC45_G_10
OTU97	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS5 marine group; uncultured bacterium
OTU1202	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured marine bacterium
OTU17739	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS7 marine group; uncultured bacterium
OTU233	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Halieaceae; OM60(NOR5) clade; uncultured bacterium
OTU10411	Bacteria; Proteobacteria; Gammaproteobacteria; Thiotrichales; Thiotrichaceae; Thiothrix; uncultured marine bacterium
OTU1692	Bacteria; Cyanobacteria; Cyanobacteria; SubsectionI; FamilyI; Synechococcus; uncultured bacterium
OTU118	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Cryomorphaceae; uncultured; uncultured bacterium
OTU783	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; uncultured alpha proteobacterium
OTU50	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured bacterium
OTU112	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; Magnetospira; uncultured bacterium
OTU1467	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured Bacteroidetes bacterium
OTU355	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 1; unidentified marine bacterioplankton
OTU7977	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Cryomorphaceae; uncultured; uncultured bacterium
OTU603	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured bacterium
OTU696	No blast hit
OTU1747	Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Colwelliaceae; Colwellia; uncultured bacterium
OTU6245	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 1; unidentified marine bacterioplankton

OTU2088	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1698	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 4; uncultured Flavobacteriia bacterium
OTU18	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU51	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU135	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU181	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured gamma proteobacterium
OTU15	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; ZD0405; uncultured bacterium
OTU267	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Lentibacter; uncultured bacterium
OTU48	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Planktomarina; uncultured bacterium
OTU57	Bacteria; Proteobacteria; Alphaproteobacteria; Rhizobiales; PS1 clade; uncultured marine bacterium
OTU904	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Planktomarina; uncultured bacterium
OTU253	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS5 marine group; uncultured bacterium
OTU6	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Ulvibacter; uncultured Bacteroidetes bacterium
OTU618	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 4; uncultured alpha proteobacterium
OTU1621	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Planktomarina; uncultured marine bacterium
OTU642	Bacteria; Proteobacteria; Gammaproteobacteria; Gammaproteobacteria Incertae Sedis; Unknown Family; uncultured; uncultured gamma proteobacterium
OTU413	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured bacterium
OTU311	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; ZD0405; uncultured bacterium
OTU2852	No blast hit
OTU1809	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Planktomarina; uncultured Roseobacter sp.
OTU215	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Formosa; uncultured bacterium
OTU1496	Bacteria; Proteobacteria; Gammaproteobacteria; Gammaproteobacteria Incertae Sedis; Unknown Family; uncultured; uncultured proteobacterium
OTU110	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured bacterium
OTU316	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Cryomorphaceae; NS10 marine group; uncultured Bacteroidetes bacterium
OTU1177	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS5 marine group; uncultured bacterium
OTU446	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR116 clade; uncultured bacterium
OTU218	Bacteria; Actinobacteria; Acidimicrobiia; Acidimicrobiales; OM1 clade; Candidatus Actinomarina; uncultured bacterium

OTU1257	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS4 marine group; uncultured bacterium
OTU2262	None
OTU602	No blast hit
OTU553	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Porticoccaceae; SAR92 clade; uncultured gamma proteobacterium
OTU347	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured bacterium
OTU1327	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; uncultured; uncultured bacterium
OTU420	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; OM182 clade; uncultured gamma proteobacterium
OTU3058	Bacteria; Proteobacteria; Gammaproteobacteria; KI89A clade; uncultured bacterium
OTU1376	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured Bacteroidetes/Chlorobi group bacterium
OTU199	Bacteria; Verrucomicrobia; Opitutae; Puniceicoccales; Puniceicoccaceae; Lentimonas; uncultured Verrucomicrobia bacterium
OTU1818	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured bacterium
OTU70	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured marine bacterium
OTU1132	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; uncultured; uncultured bacterium
OTU44	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured Bacteroidetes bacterium
OTU1056	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Porticoccaceae; SAR92 clade; uncultured bacterium
OTU624	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured SAR86 cluster bacterium
OTU127	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured bacterium
OTU2816	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured bacterium
OTU422	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; marine metagenome
OTU75	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Cryomorphaceae; NS10 marine group; uncultured bacterium
OTU564	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; OM75 clade; uncultured bacterium
OTU1051	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured bacterium
OTU2111	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Porticoccaceae; Porticoccus; uncultured bacterium
OTU784	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 4; uncultured alpha proteobacterium
OTU4439	None
OTU263	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured OM43 clade bacterium
OTU49	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; uncultured; uncultured Flavobacteriia bacterium

OTU1623	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 1; uncultured bacterium
OTU753	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS5 marine group; uncultured bacterium
OTU314	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured gamma proteobacterium
OTU38	Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Colwelliaceae; Colwellia; uncultured bacterium
OTU698	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 1; unidentified marine bacterioplankton
OTU82	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; marine metagenome
OTU4077	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS5 marine group; uncultured Flavobacteriia bacterium
OTU131	Bacteria; Actinobacteria; Acidimicrobiia; Acidimicrobiales; Sva0996 marine group; uncultured actinobacterium
OTU3600	Bacteria; Proteobacteria; Gammaproteobacteria; Thiotrichales; Thiotrichaceae; Thiothrix; uncultured marine bacterium
OTU219	Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales; Campylobacteraceae; Arcobacter; uncultured bacterium
OTU2285	Bacteria; Proteobacteria; Epsilonproteobacteria; Campylobacterales; Campylobacteraceae; Arcobacter; uncultured bacterium
OTU245	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; uncultured bacterium
OTU59	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured organism
OTU251	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured Flavobacteriia bacterium
OTU2100	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured organism
OTU369	Bacteria; Bacteroidetes; Cytophagia; Cytophagales; Flammeovirgaceae; Marinoscillum; uncultured Cytophagia bacterium
OTU487	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS4 marine group; uncultured bacterium
OTU1635	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 4; uncultured Flavobacteriia bacterium
OTU1771	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured Bacteroidetes bacterium
OTU30	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Cryomorphaceae; uncultured; uncultured Sphingobacterium sp. EB080_L08E11
OTU346	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Cryomorphaceae; uncultured; marine metagenome
OTU87	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; marine metagenome
OTU17	Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Alteromonadaceae; Paraglaciecola; uncultured Antarctic sea ice bacterium
OTU25	Bacteria; Bacteroidetes; Cytophagia; Cytophagales; Flammeovirgaceae; uncultured; uncultured bacterium
OTU120	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 4; uncultured bacterium
OTU343	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; uncultured bacterium
OTU1575	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium

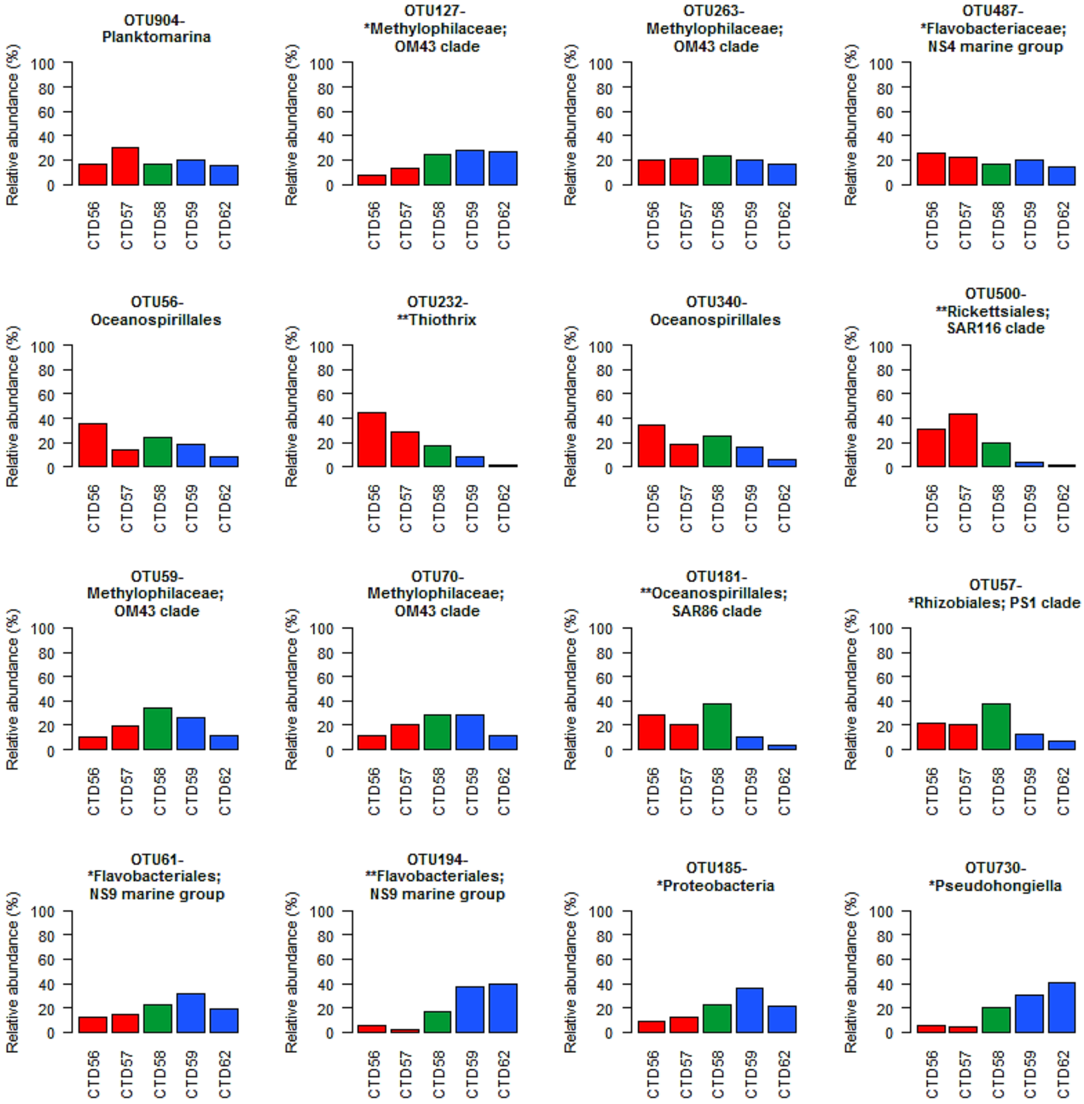
OTU9	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured Bacteroidetes bacterium
OTU1754	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; uncultured bacterium
OTU185	Bacteria; Proteobacteria; AEGEAN-245; uncultured gamma proteobacterium
OTU6172	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured SAR86 cluster bacterium
OTU9320	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured bacterium
OTU1731	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; NS5 marine group; uncultured bacterium
OTU61	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS9 marine group; uncultured bacterium
OTU183	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 2; unidentified marine bacterioplankton
OTU31	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU15732	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured organism
OTU1341	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1320	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU289	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured marine bacterium
OTU617	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU47	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured Oceanospirillales bacterium
OTU1025	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured gamma proteobacterium
OTU4289	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU109	Bacteria; Proteobacteria; Gammaproteobacteria; Xanthomonadales; JTB255 marine benthic group; uncultured gamma proteobacterium
OTU731	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1476	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU14206	None
OTU426	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Porticoccaceae; SAR92 clade; uncultured SAR92 cluster bacterium
OTU2171	Bacteria; Proteobacteria; Gammaproteobacteria; Alteromonadales; Pseudoalteromonadaceae; Pseudoalteromonas; uncultured bacterium
OTU7108	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured gamma proteobacterium
OTU2723	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU416	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU3671	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium

OTU2430	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU41	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SUP05 cluster; uncultured gamma proteobacterium
OTU962	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU56	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU470	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured gamma proteobacterium
OTU168	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Sulfitobacter; uncultured bacterium
OTU2276	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; uncultured; uncultured bacterium
OTU4	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; uncultured; uncultured marine bacterium
OTU287	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU40643	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS7 marine group; uncultured bacterium
OTU2235	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU10412	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; NS7 marine group; uncultured bacterium
OTU37	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Porticoccaceae; SAR92 clade; uncultured bacterium
OTU1322	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; SAR86 clade; uncultured gamma proteobacterium
OTU5535	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU1477	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; OM43 clade; uncultured bacterium
OTU340	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU830	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU1965	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1009	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU349	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1949	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU6754	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU627	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1937	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; Magnetospira; uncultured bacterium
OTU62	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1219	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 1; uncultured bacterium

OTU2765	None
OTU579	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1824	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU1350	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU196	Bacteria; Proteobacteria; Gammaproteobacteria; Cellvibrionales; Cellvibrionaceae; uncultured; uncultured organism
OTU21	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU444	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU469	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU2505	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU547	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU114	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium
OTU395	None
OTU1317	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1174	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU2913	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; AEGEAN-169 marine group; uncultured bacterium
OTU2611	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU1603	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; uncultured; uncultured bacterium
OTU2678	Bacteria; Proteobacteria; Alphaproteobacteria; SAR11 clade; Surface 1; Candidatus Pelagibacter; uncultured SAR11 cluster alpha proteobacterium
OTU931	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; Oceanospirillaceae; Balneatrix; uncultured gamma proteobacterium
OTU4674	Bacteria; Verrucomicrobia; Verrucomicrobiae; Verrucomicrobiales; Verrucomicrobiaceae; Roseibacillus; uncultured bacterium
OTU2650	Bacteria; Proteobacteria; Gammaproteobacteria; Oceanospirillales; JL-ETNP-Y6; uncultured marine bacterium
OTU929	Bacteria; Bacteroidetes; Flavobacteriia; Flavobacteriales; Flavobacteriaceae; Polaribacter 2; uncultured bacterium

Appendix 2 The Relative Abundance of Selected Bacterial OTUs Between Stations.

OTUs were selected from the top 200 most abundant for the bacterial dataset. OTUs show more graduated distribution patterns which help explain the community assemblage structure, but can still be seen to broadly reflect those which are present across all stations (1), found primarily in LI stations (2), found primarily in the MI stations (3), or are found primarily within the HI station (4). Stations are coloured based upon the degree of Polar Water influence determined to be present at each station; red - LI, green- MI, blue- HI. ** indicates a significant correlation with temperature. * indicates a significant correlation with other environmental variables.



Appendix 3. The Correlation of all Top 200 Most Abundant Eukaryotic OTUs with Each Environmental Factor.

Environmental data used was measured from the depth at which water samples were taken at each station during CTD casts. R_s represents the correlation coefficient, values ≥0.9 represent a strong positive association, values ≤-0.9 represent a strong negative association. P represents the p-value for each correlation, values ≤0.05 are considered statistically significant. The taxonomic annotation corresponding to each OTUID is displayed below.

OTU ID	DO_2		Salinity		SubSurVPAR		Temperature		Ammonium		Nitrate		Silicate		Phosphate	
	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P	R_s	P
OTU118	0.78	0.12	-0.89	0.04	-0.67	0.22	-0.89	0.04	0.89	0.04	0.34	0.58	-0.78	0.12	0.34	0.58
OTU147	0.90	0.04	-0.90	0.04	-0.90	0.04	-0.90	0.04	0.90	0.04	0.70	0.19	-0.60	0.28	0.60	0.28
OTU153	0.70	0.19	-0.50	0.39	-0.20	0.75	-0.50	0.39	0.50	0.39	-0.10	0.87	-1.00	0.00	-0.20	0.75
OTU157	0.89	0.04	-0.78	0.12	-0.45	0.45	-0.78	0.12	0.78	0.12	0.22	0.72	-0.89	0.04	0.22	0.72
OTU204	0.90	0.04	-0.90	0.04	-0.60	0.28	-0.90	0.04	0.90	0.04	0.50	0.39	-0.60	0.28	0.60	0.28
OTU264	0.89	0.04	-0.78	0.12	-0.45	0.45	-0.78	0.12	0.78	0.12	0.22	0.72	-0.89	0.04	0.22	0.72
OTU29	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	0.50	0.39
OTU300	0.87	0.05	-0.97	0.00	-0.87	0.05	-0.97	0.00	0.97	0.00	0.67	0.22	-0.56	0.32	0.67	0.22
OTU44	0.80	0.10	-1.00	0.00	-0.80	0.10	-1.00	0.00	1.00	0.00	0.60	0.28	-0.50	0.39	0.70	0.19
OTU46	0.90	0.04	-0.90	0.04	-0.60	0.28	-0.90	0.04	0.90	0.04	0.50	0.39	-0.60	0.28	0.60	0.28
OTU49	0.87	0.05	-0.97	0.00	-0.87	0.05	-0.97	0.00	0.97	0.00	0.67	0.22	-0.56	0.32	0.67	0.22
OTU70	0.60	0.28	-0.90	0.04	-0.50	0.39	-0.90	0.04	0.90	0.04	0.20	0.75	-0.60	0.28	0.40	0.50
OTU95	0.78	0.12	-0.89	0.04	-0.67	0.22	-0.89	0.04	0.89	0.04	0.34	0.58	-0.78	0.12	0.34	0.58
OTU113	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU130	0.60	0.28	-0.60	0.28	-0.90	0.04	-0.60	0.28	0.60	0.28	1.00	0.00	0.10	0.87	0.90	0.04
OTU134	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU135	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU137	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU151	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87

OTU183	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU185	0.50	0.39	-0.70	0.19	-0.80	0.10	-0.70	0.19	0.70	0.19	0.90	0.04	0.20	0.75	1.00	0.00
OTU19	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU20	-0.60	0.28	0.60	0.28	0.10	0.87	0.60	0.28	-0.60	0.28	0.20	0.75	0.90	0.04	0.10	0.87
OTU205	0.60	0.28	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.80	0.10	0.10	0.87	0.90	0.04
OTU218	0.60	0.28	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.80	0.10	0.10	0.87	0.90	0.04
OTU224	0.56	0.32	-0.67	0.22	-0.87	0.05	-0.67	0.22	0.67	0.22	0.97	0.00	0.15	0.80	0.97	0.00
OTU239	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	0.15	0.80
OTU244	0.60	0.28	-0.60	0.28	-0.90	0.04	-0.60	0.28	0.60	0.28	1.00	0.00	0.10	0.87	0.90	0.04
OTU262	0.70	0.19	-0.50	0.39	-0.70	0.19	-0.50	0.39	0.50	0.39	0.90	0.04	0.00	1.00	0.80	0.10
OTU274	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	0.15	0.80
OTU299	0.60	0.28	-0.60	0.28	-0.90	0.04	-0.60	0.28	0.60	0.28	1.00	0.00	0.10	0.87	0.90	0.04
OTU328	0.60	0.28	-0.60	0.28	-0.90	0.04	-0.60	0.28	0.60	0.28	1.00	0.00	0.10	0.87	0.90	0.04
OTU340	-0.45	0.45	0.22	0.72	0.11	0.86	0.22	0.72	-0.22	0.72	0.22	0.72	0.89	0.04	0.45	0.45
OTU342	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU362	0.10	0.87	-0.50	0.39	-0.60	0.28	-0.50	0.39	0.50	0.39	0.70	0.19	0.50	0.39	0.90	0.04
OTU77	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU79	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU94	-0.67	0.22	0.56	0.32	0.31	0.61	0.56	0.32	-0.56	0.32	0.05	0.93	0.97	0.00	0.15	0.80
OTU11	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU111	-1.00	0.00	0.80	0.10	0.70	0.19	0.80	0.10	-0.80	0.10	-0.60	0.28	0.70	0.19	-0.50	0.39
OTU123	0.70	0.19	-0.50	0.39	-0.70	0.19	-0.50	0.39	0.50	0.39	0.90	0.04	0.00	1.00	0.80	0.10
OTU140	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU141	-0.67	0.22	0.72	0.17	0.97	0.00	0.72	0.17	-0.72	0.17	-0.97	0.00	0.05	0.93	-0.87	0.05
OTU158	-0.97	0.00	0.87	0.05	0.82	0.09	0.87	0.05	-0.87	0.05	-0.67	0.22	0.67	0.22	-0.56	0.32
OTU160	-0.89	0.04	0.78	0.12	0.89	0.04	0.78	0.12	-0.78	0.12	-0.89	0.04	0.34	0.58	-0.78	0.12
OTU167	-0.89	0.04	0.78	0.12	0.89	0.04	0.78	0.12	-0.78	0.12	-0.89	0.04	0.34	0.58	-0.78	0.12

OTU209	-0.78	0.12	0.89	0.04	0.78	0.12	0.89	0.04	-0.89	0.04	-0.78	0.12	0.22	0.72	-0.89	0.04
OTU211	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU22	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU25	-0.67	0.22	0.56	0.32	0.82	0.09	0.56	0.32	-0.56	0.32	-0.97	0.00	-0.05	0.93	-0.87	0.05
OTU254	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU269	-0.89	0.04	0.78	0.12	0.89	0.04	0.78	0.12	-0.78	0.12	-0.89	0.04	0.34	0.58	-0.78	0.12
OTU270	-0.97	0.00	0.87	0.05	0.82	0.09	0.87	0.05	-0.87	0.05	-0.67	0.22	0.67	0.22	-0.56	0.32
OTU35	-0.40	0.50	0.50	0.39	0.90	0.04	0.50	0.39	-0.50	0.39	-0.80	0.10	0.00	1.00	-0.60	0.28
OTU55	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU62	-0.70	0.19	0.80	0.10	1.00	0.00	0.80	0.10	-0.80	0.10	-0.90	0.04	0.20	0.75	-0.80	0.10
OTU64	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU65	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU69	-0.70	0.19	0.80	0.10	1.00	0.00	0.80	0.10	-0.80	0.10	-0.90	0.04	0.20	0.75	-0.80	0.10
OTU71	-0.70	0.19	0.80	0.10	1.00	0.00	0.80	0.10	-0.80	0.10	-0.90	0.04	0.20	0.75	-0.80	0.10
OTU72	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU78	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU99	-0.50	0.39	0.70	0.19	0.80	0.10	0.70	0.19	-0.70	0.19	-0.90	0.04	-0.20	0.75	-1.00	0.00
OTU101	-0.60	0.28	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.80	0.10	0.10	0.87	-0.90	0.04
OTU106	-0.40	0.50	0.50	0.39	0.90	0.04	0.50	0.39	-0.50	0.39	-0.80	0.10	0.00	1.00	-0.60	0.28
OTU110	-0.70	0.19	0.80	0.10	1.00	0.00	0.80	0.10	-0.80	0.10	-0.90	0.04	0.20	0.75	-0.80	0.10
OTU114	0.80	0.10	-1.00	0.00	-0.80	0.10	-1.00	0.00	1.00	0.00	0.60	0.28	-0.50	0.39	0.70	0.19
OTU117	0.60	0.28	-0.60	0.28	-0.10	0.87	-0.60	0.28	0.60	0.28	-0.20	0.75	-0.90	0.04	-0.10	0.87
OTU119	0.70	0.19	-0.50	0.39	-0.70	0.19	-0.50	0.39	0.50	0.39	0.90	0.04	0.00	1.00	0.80	0.10
OTU12	0.90	0.04	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.70	0.19	-0.40	0.50	0.60	0.28
OTU125	-0.80	0.10	1.00	0.00	0.80	0.10	1.00	0.00	-1.00	0.00	-0.60	0.28	0.50	0.39	-0.70	0.19
OTU143	-0.60	0.28	0.60	0.28	0.90	0.04	0.60	0.28	-0.60	0.28	-1.00	0.00	-0.10	0.87	-0.90	0.04
OTU148	0.80	0.10	-0.70	0.19	-0.70	0.19	-0.70	0.19	0.70	0.19	0.40	0.50	-0.80	0.10	0.20	0.75

OTU16	-0.60	0.28	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.80	0.10	0.10	0.87	-0.90	0.04
OTU164	-0.89	0.04	0.78	0.12	0.89	0.04	0.78	0.12	-0.78	0.12	-0.89	0.04	0.34	0.58	-0.78	0.12
OTU172	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU173	-0.90	0.04	0.60	0.28	0.50	0.39	0.60	0.28	-0.60	0.28	-0.30	0.62	0.90	0.04	-0.10	0.87
OTU186	1.00	0.00	-0.80	0.10	-0.70	0.19	-0.80	0.10	0.80	0.10	0.60	0.28	-0.70	0.19	0.50	0.39
OTU200	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU208	0.67	0.22	-0.87	0.05	-0.97	0.00	-0.87	0.05	0.87	0.05	0.87	0.05	-0.15	0.80	0.87	0.05
OTU231	-0.40	0.50	0.50	0.39	0.90	0.04	0.50	0.39	-0.50	0.39	-0.80	0.10	0.00	1.00	-0.60	0.28
OTU245	0.90	0.04	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.70	0.19	-0.40	0.50	0.60	0.28
OTU27	-0.50	0.39	0.70	0.19	0.80	0.10	0.70	0.19	-0.70	0.19	-0.90	0.04	-0.20	0.75	-1.00	0.00
OTU289	-0.10	0.87	0.50	0.39	0.60	0.28	0.50	0.39	-0.50	0.39	-0.70	0.19	-0.50	0.39	-0.90	0.04
OTU30	-0.70	0.19	0.80	0.10	1.00	0.00	0.80	0.10	-0.80	0.10	-0.90	0.04	0.20	0.75	-0.80	0.10
OTU31	0.90	0.04	-0.60	0.28	-0.60	0.28	-0.60	0.28	0.60	0.28	0.70	0.19	-0.40	0.50	0.60	0.28
OTU310	-0.60	0.28	0.60	0.28	0.90	0.04	0.60	0.28	-0.60	0.28	-1.00	0.00	-0.10	0.87	-0.90	0.04
OTU32	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU33	-0.60	0.28	0.60	0.28	0.90	0.04	0.60	0.28	-0.60	0.28	-1.00	0.00	-0.10	0.87	-0.90	0.04
OTU36	0.50	0.39	-0.70	0.19	-0.80	0.10	-0.70	0.19	0.70	0.19	0.90	0.04	0.20	0.75	1.00	0.00
OTU38	-0.80	0.10	1.00	0.00	0.80	0.10	1.00	0.00	-1.00	0.00	-0.60	0.28	0.50	0.39	-0.70	0.19
OTU45	0.40	0.50	-0.50	0.39	-0.90	0.04	-0.50	0.39	0.50	0.39	0.80	0.10	0.00	1.00	0.60	0.28
OTU5	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU73	-0.90	0.04	0.90	0.04	0.90	0.04	0.90	0.04	-0.90	0.04	-0.70	0.19	0.60	0.28	-0.60	0.28
OTU8	-1.00	0.00	0.80	0.10	0.70	0.19	0.80	0.10	-0.80	0.10	-0.60	0.28	0.70	0.19	-0.50	0.39
OTU87	-0.60	0.28	0.60	0.28	0.40	0.50	0.60	0.28	-0.60	0.28	0.00	1.00	0.90	0.04	0.10	0.87
OTU91	-0.87	0.05	0.97	0.00	0.72	0.17	0.97	0.00	-0.97	0.00	-0.56	0.32	0.56	0.32	-0.67	0.22
OTU92	-0.60	0.28	0.60	0.28	0.90	0.04	0.60	0.28	-0.60	0.28	-1.00	0.00	-0.10	0.87	-0.90	0.04
OTU97	-0.10	0.87	0.50	0.39	0.60	0.28	0.50	0.39	-0.50	0.39	-0.70	0.19	-0.50	0.39	-0.90	0.04

OTU ID	Taxonomic annotation
OTU118	Eukaryota;SAR;Rhizaria;Cercozoa;Thecofilosea;WHOI-LI1-14;uncultured cercozoan
OTU147	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Mediophyceae;Thalassiosira;uncultured eukaryote
OTU153	Eukaryota;SAR;Alveolata;Ciliophora;Intramacronucleata;Spirotrichea;Choreotrichia;Salpingella;uncultured tintinnid ciliate
OTU157	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Coscinodiscophytina;Rhizosolenids;Proboscia;Proboscia alata
OTU204	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU264	Eukaryota;SAR;Stramenopiles;Ochrophyta;Chrysophyceae;E222;uncultured marine eukaryote
OTU29	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Mediophyceae;Chaetoceros;uncultured marine eukaryote
OTU300	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Coscinodiscophytina;Rhizosolenids;Leptocylindrus;Leptocylindrus hargravesii
OTU44	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Mediophyceae;Thalassiosira;Thalassiosira aestivalis
OTU46	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Mediophyceae;Chaetoceros;uncultured marine eukaryote
OTU49	Eukaryota;SAR;Rhizaria;Cercozoa;Thecofilosea;Cryomonadida;Protaspidae;Cryothecomonas;uncultured eukaryote
OTU70	Eukaryota;SAR;Rhizaria;Cercozoa;Thecofilosea;Cryomonadida;Protaspidae;Cryothecomonas;uncultured marine eukaryote
OTU95	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Bacillariophyceae;uncultured marine eukaryote
OTU113	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group I;uncultured eukaryote
OTU130	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Mediophyceae;Chaetoceros;uncultured stramenopile
OTU134	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured syndiniales
OTU135	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group I;uncultured eukaryote
OTU137	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU151	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Amoebophrya;uncultured eukaryote
OTU183	No blast hit
OTU185	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Amoebophrya;uncultured eukaryote
OTU19	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;FV18-2D9;uncultured marine eukaryote
OTU20	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;Woloszynskia;uncultured eukaryote
OTU205	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Incertae Sedis;Apicoporus;uncultured eukaryote
OTU218	Eukaryota;SAR;Stramenopiles;MAST-1;MAST-1B;uncultured eukaryote
OTU224	Eukaryota;SAR;Alveolata;Ciliophora;Intramacronucleata;Spirotrichea;Oligotrichia;Strombidium;uncultured alveolate

OTU239	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU244	No blast hit
OTU262	Eukaryota;SAR;Alveolata;SCM37C52;uncultured eukaryote
OTU274	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group I;uncultured eukaryote
OTU299	Eukaryota;Excavata;Discoba;Jakobida;uncultured eukaryote
OTU328	No blast hit
OTU340	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured alveolate
OTU342	Eukaryota;Excavata;Discoba;Discicristata;Euglenozoa;Diplonemea;SCM38C39 marine group;uncultured eukaryote
OTU362	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU77	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU79	Eukaryota;Picozoa;uncultured marine eukaryote
OTU94	Eukaryota;SAR;Alveolata;Protalveolata;Incertae Sedis;Ellobiopsidae;Ellobiopsis;uncultured eukaryote
OTU11	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Suessiaceae;Pelagodinium;uncultured Gymnodinium
OTU111	Eukaryota;SAR;Stramenopiles;Ochrophyta;Bolidomonas;uncultured stramenopile
OTU123	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group I;uncultured Rhizaria
OTU140	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Incertae Sedis;Azadinium;uncultured eukaryote
OTU141	Eukaryota;SAR;Rhizaria;Cercozoa;Thecofilosea;WHOI-LI1-14;uncultured marine eukaryote
OTU158	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Incertae Sedis;Azadinium;Azadinium polongum
OTU160	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Amoebophrya;Amoebophrya sp. ex Prorocentrum minimum
OTU167	Eukaryota;Cryptophyta;Kathablepharidae;Katablepharis;uncultured marine picoeukaryote
OTU209	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU211	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU22	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU25	Eukaryota;SAR;Rhizaria;Cercozoa;Thecofilosea;Ebriacea;Ebria;uncultured eukaryote
OTU254	Eukaryota;Haptophyta;Prymnesiophyceae;Isochrysidales;Emiliania;Emiliania huxleyi CCMP1516
OTU269	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Prorocentrales;Prorocentrum;Prorocentrum gracile
OTU270	Eukaryota;SAR;Stramenopiles;MAST-8;MAST-8D;uncultured eukaryote

OTU35	Eukaryota;SAR;Rhizaria;Cercozoa;Thecofilosea;Cryomonadida;Protaspidae;Cryothecomonas;uncultured eukaryote
OTU55	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group I;uncultured eukaryote
OTU62	Eukaryota;Cryptophyta;Kathablepharidae;Leucocryptos;uncultured eukaryote
OTU64	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Incertae Sedis;Azadinium;Azadinium spinosum
OTU65	Eukaryota;Cryptophyta;Cryptomonadales;Teleaulax;Teleaulax amphioxeia
OTU69	Eukaryota;Cryptophyta;Cryptomonadales;Teleaulax;uncultured phytoplankton
OTU71	Eukaryota;Haptophyta;Prymnesiophyceae;Prymnesiales;Chrysochromulina;uncultured eukaryote
OTU72	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured marine picoeukaryote
OTU78	Eukaryota;Archaeplastida;Chloroplastida;Chlorophyta;Mamiellophyceae;Mamiellales;Bathycoccus;uncultured marine picoeukaryote
OTU99	Eukaryota;SAR;Alveolata;Ciliophora;Intramacronucleata;Spirotrichea;Choreotrichia;Codonellopsis;uncultured tintinnid ciliate
OTU101	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Amphidinium;uncultured marine eukaryote
OTU106	Eukaryota;Haptophyta;Prymnesiophyceae;Prymnesiales;Prymnesium;uncultured haptophyte
OTU110	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Mediophyceae;Thalassiosira;Thalassiosira oceanica
OTU114	Eukaryota;SAR;Stramenopiles;MAST-7;MAST-7A;uncultured marine eukaryote
OTU117	Eukaryota;SAR;Alveolata;Ciliophora;Intramacronucleata;Conthreep;Nassophorea;uncultured;uncultured ciliate
OTU119	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Kareniaceae;Karlodinium;uncultured eukaryote
OTU12	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Bacillariophyceae;Pseudo-nitzschia;uncultured marine eukaryote
OTU125	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote
OTU143	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;Gymnodinium;uncultured marine eukaryote
OTU148	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Incertae Sedis;Azadinium;uncultured eukaryote
OTU16	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;Gymnodinium;uncultured marine eukaryote
OTU164	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Peridiniphycidae;Peridiniales;Peridiniopsis;uncultured marine eukaryote
OTU172	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Dinophysiales;uncultured eukaryote
OTU173	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;Gymnodinium;uncultured eukaryote
OTU186	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured marine Syndiniales
OTU200	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group I;uncultured alveolate
OTU208	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured eukaryote

OTU231	Eukaryota;Cryptophyta;Cryptomonadales;uncultured marine eukaryote
OTU245	Eukaryota;Archaeplastida;Chloroplastida;Charophyta;Phragmoplastophyta;Streptophyta;Embryophyta;Tracheophyta;Spermatophyta;Magnoliophyta;Solanales;Solanum;Solanum lycopersicum (tomato)
OTU27	Eukaryota;Haptophyta;Prymnesiophyceae;Coccolithales;Cruciplacolithus;Cruciplacolithus neohelis
OTU289	Eukaryota;SAR;Alveolata;Dinoflagellata;SCM28C5;uncultured eukaryote
OTU30	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Coscinodiscophytina;Actinocyclus;Actinocyclus curvatulus
OTU31	Eukaryota;SAR;Stramenopiles;Ochrophyta;Diatomea;Bacillariophytina;Mediophyceae;Chaetoceros;uncultured eukaryote
OTU310	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;Gymnodinium;Gymnodinium sp. GSSW10
OTU32	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;uncultured eukaryote
OTU33	Eukaryota;SAR;Alveolata;Ciliophora;Intramacronucleata;Spirotrichea;Oligotrichia;uncultured eukaryote
OTU36	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Peridiniphycidae;Peridiniales;Heterocapsa;uncultured marine eukaryote
OTU38	Eukaryota;SAR;Stramenopiles;Ochrophyta;Chrysophyceae;E222;uncultured marine picoeukaryote
OTU45	Eukaryota;SAR;Rhizaria;Cercozoa;Thecofilosea;DSGM-50;uncultured eukaryote
OTU5	Eukaryota;SAR;Alveolata;NIF-4C10;uncultured eukaryote
OTU73	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;Gymnodiniphycidae;Gymnodinium clade;uncultured eukaryote
OTU8	Eukaryota;SAR;Alveolata;Dinoflagellata;Dinophyceae;SL163A10;uncultured eukaryote
OTU87	Eukaryota;SAR;Stramenopiles;MAST-1;MAST-1C;uncultured stramenopile
OTU91	Eukaryota;SAR;Alveolata;Protalveolata;Syndiniales;Syndiniales Group II;uncultured marine picoeukaryote
OTU92	Eukaryota;SAR;Alveolata;Ciliophora;Intramacronucleata;Spirotrichea;Oligotrichia;uncultured;uncultured marine eukaryote
OTU97	Eukaryota;SAR;Alveolata;Ciliophora;Intramacronucleata;Spirotrichea;Choreotrichia;uncultured;uncultured eukaryote

Appendix 4. The Identity of all Annotated Genes.

Shown is the name of each annotated gene (where present), the cluster to which it belonged and as shown in Figure 6.7, the representative major taxonomic group the gene was annotated as most similar to, and the description the gene was annotated with during assembly. Genes have been given a unique identifier number (ID) to delineate those with same names.

ID	Cluster	Gene Name	Taxonomic group	Annotation
1	1	Unknown	Excavata	MHOM/GT/2001/U1103 elongation factor 1-alpha partial mRNA
2	1	Unknown	Opisthokonta	ATCC 30864 actin mRNA
3	1	Unknown	Opisthokonta	non-muscle actin mRNA
4	1	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
5	1	Unknown	Stramenopile	mRNA
6	1	Unknown	Apicomplexa	3D7 conserved protein, unknown function (MAL13P1.455) mRNA
7	2	GAPD1	Stramenopile	CCMP1335 glyceraldehyde-3-phosphate dehydrogenase precursor
8	2	Gasu_40100	Rhodophyta	[pt] photosystem II PsbA protein
9	2	Gasu_40440	Rhodophyta	[pt] photosystem I P700 chlorophyll A apoprotein
10	2	Lhcf22	Haptophyta	CCMP1516 light harvesting protein
11	2	petB	Stramenopile	CCMP1335 cytochrome B6
12	2	PSAF	Stramenopile	CCMP1335 PSAF, photosystem I reaction center subunit
13	2	NCLIV_058890	Apicomplexa	Liverpool tubulin alpha chain
14	2	Unknown	Stramenopile	CCAP 1055/1 histone H4 isoform 1b (H4-1b)
15	2	Unknown	Stramenopile	CCMP1335 translation elongation factor alpha
16	2	Unknown	Haptophyta	CCMP1516 putative histone H3 partial mRNA
17	2	Unknown	Haptophyta	CCMP1516 putative histone H2B mRNA
18	2	Unknown	Haptophyta	CCMP1516 putative histone H2B mRNA
19	2	Unknown	Stramenopile	tubulin alpha-2 chain mRNA
20	2	Unknown	Excavata	beta-tubulin partial mRNA
21	2	Unknown	Excavata	cytosolic HSP90-like protein partial mRNA
22	3	ACA1_097000	Amoebozoa	str. Neff ribosomal family s4e, putative
23	3	ANT1	Stramenopile	CCMP1335 adenine nucleotide translocator; ATP/ADP translocase
24	3	bchI	Stramenopile	CCMP1335 mg-protoporphyrin IX chelatase
25	3	EFG1	Haptophyta	CCMP1516 putative EF-1 alpha/Tu like protein
26	3	Gasu_40630	Rhodophyta	[pt] F-type H+-transporting ATPase subunit alpha
27	3	HSP90_1	Stramenopile	CCMP1335 HSP90 family member
28	3	Lhcf28	Haptophyta	CCMP1516 light harvesting protein
29	3	Lhcf29	Haptophyta	CCMP1516 light harvesting protein
30	3	Lhcf34_2	Haptophyta	CCMP1516 light harvesting protein
31	3	Lhcf61	Haptophyta	CCMP1516 light harvesting protein
32	3	Lhcf62	Haptophyta	CCMP1516 light harvesting protein

33	3	Lhcf67	Haptophyta	CCMP1516 light harvesting protein
34	3	petB	Stramenopile	CCMP1335 cytochrome B6
35	3	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
36	3	PsbV	Stramenopile	CCMP1335 cytochrome c550, PsbV
37	3	TUB2	Stramenopile	CCMP1335 tubulin beta
38	3	Unknown	Stramenopile	CCMP1335 ABC cassette-containing protein
39	3	Unknown	Haptophyta	CCMP1516 ATP synthase F1 subunit alpha mRNA
40	3	Unknown	Haptophyta	CCMP1516 60S ribosomal protein L10a partial mRNA
41	3	Unknown	Haptophyta	CCMP1516 putative mitochondrial carrier protein mRNA
42	3	Unknown	Haptophyta	CCMP1516 alpha-tubulin partial mRNA
43	3	Unknown	Haptophyta	CCMP1516 receptor of activated protein kinase C partial mRNA
44	3	Lhcf67	Haptophyta	CCMP1516 light harvesting protein
45	3	Unknown	Haptophyta	CCMP1516 putative histone H4 mRNA
46	3	Unknown	Haptophyta	CCMP1516 S-adenosyl-L-homocysteine hydrolase mRNA
47	3	Unknown	Haptophyta	CCMP1516 dTDP-glucose 4
48	3	Unknown	Haptophyta	CCMP1516 60S ribosomal protein L3 partial mRNA
49	3	Unknown	Excavata	gambiense DAL972 elongation factor 1-alpha
50	4	Act2	Stramenopile	CCAP 1055/1 actin/actin like protein
51	4	bchI	Stramenopile	CCMP1335 mg-protoporphyrin IX chelatase
52	4	GAPD1	Stramenopile	CCMP1335 glyceraldehyde-3-phosphate dehydrogenase precursor
53	4	Gasu_40100	Rhodophyta	[pt] photosystem II PsbA protein
54	4	Gasu_40450	Rhodophyta	[pt] photosystem I P700 chlorophyll A apoprotein
55	4	Gasu_40560	Rhodophyta	[pt] F-type H+-transporting ATPase subunit beta
56	4	H4_1	Stramenopile	CCMP1335 histone H4
57	4	hsp70_4	Stramenopile	CCMP1335 heat shock protein/chaperone
58	4	HSP70A	Stramenopile	CCAP 1055/1 protein heat shock protein Hsp70
59	4	HSP90_1	Stramenopile	CCMP1335 HSP90 family member
60	4	hUbi	Stramenopile	CCAP 1055/1 ubiquitin extension protein 4
61	4	petB	Stramenopile	CCMP1335 cytochrome B6
62	4	PSAF	Stramenopile	CCMP1335 PSAF, photosystem I reaction center subunit
63	4	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
64	4	PsbV	Stramenopile	CCMP1335 cytochrome c550, PsbV
65	4	PsbV	Stramenopile	CCMP1335 cytochrome c550, PsbV
66	4	Unknown	Stramenopile	CCMP1335 histone H2A
67	4	Unknown	Alveolata	ATCC 50983 alpha-tubulin, putative
68	4	Unknown	Stramenopile	VS20 elongation factor 1-alpha 1 mRNA
69	5	bchI	Stramenopile	CCMP1335 mg-protoporphyrin IX chelatase

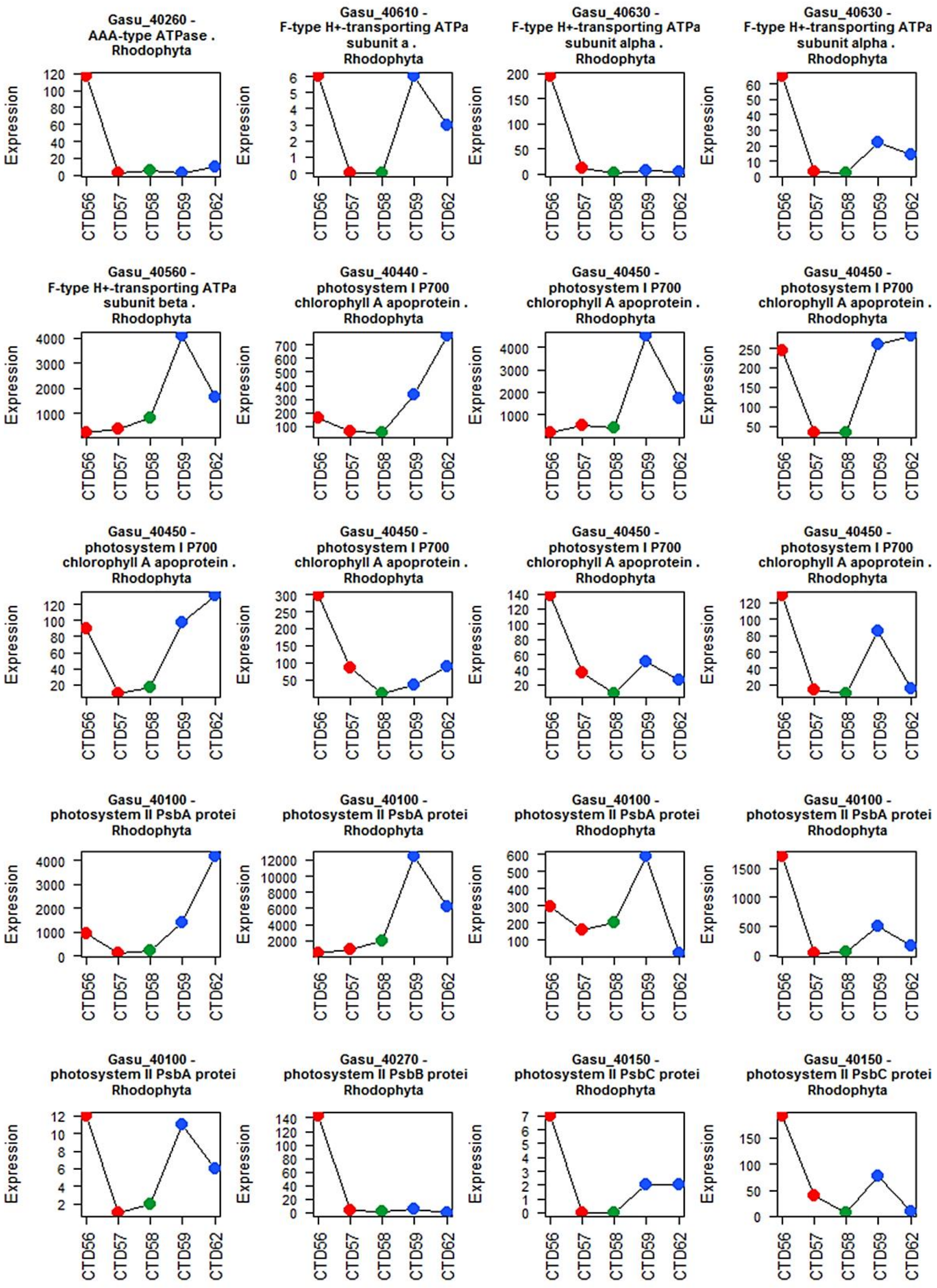
70	5	BEWA_006510	Apicomplexa	histone H3, putative
71	5	Gasu_40100	Rhodophyta	[pt] photosystem II PsbA protein
72	5	Gasu_40450	Rhodophyta	[pt] photosystem I P700 chlorophyll A apoprotein
73	5	Gasu_40450	Rhodophyta	[pt] photosystem I P700 chlorophyll A apoprotein
74	5	petB	Stramenopile	CCMP1335 cytochrome B6
75	5	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
76	5	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
77	5	TUA	Opisthokonta	MX1 alpha tubulin MONBRDRAFT_17634
78	5	TUB_2	Opisthokonta	MX1 beta tubulin MONBRDRAFT_44314
79	5	Unknown	Ciliophora	SB210 tubulin/FtsZ family
80	5	NCLIV_058890	Apicomplexa	Liverpool tubulin alpha chain
81	5	Unknown	Ciliophora	SB210 tubulin/FtsZ family
82	5	Unknown	Ciliophora	SB210 tubulin partial mRNA
83	5	Unknown	Ciliophora	SB210 tubulin partial mRNA
84	5	Unknown	Stramenopile	CCAP 1055/1 histone H4 isoform 1b (H4-1b)
85	5	Unknown	Stramenopile	CCMP1335 histone H2B
86	5	Unknown	Stramenopile	CCMP1335 thiamine biosynthesis protein
87	5	Unknown	Opisthokonta	ATCC 30864 ribosomal protein L10 mRNA
88	5	Unknown	Opisthokonta	ATCC 30864 actin mRNA
89	5	Unknown	Opisthokonta	tubulin beta chain mRNA
90	5	Unknown	Opisthokonta	non-muscle actin mRNA
91	5	Unknown	Haptophyta	CCMP1516 alpha-tubulin partial mRNA
92	5	Unknown	Stramenopile	histone mRNA
93	5	Unknown	Stramenopile	tubulin alpha chain mRNA
94	5	Unknown	Apicomplexa	elongation factor 1-alpha (EF-1-ALPHA)
95	5	Unknown	Stramenopile	centrin partial mRNA
96	5	Unknown	Stramenopile	tubulin beta chain mRNA
97	5	Unknown	Stramenopile	histone H3 partial mRNA
98	5	Unknown	Stramenopile	polyubiquitin partial mRNA
99	5	Unknown	Apicomplexa	core histone H2A/H2B/H3/H4 family protein
100	5	Unknown	Opisthokonta	JP610 histone H4 mRNA
101	5	Unknown	Stramenopile	. ST4 60S ribosomal protein L23 mRNA
102	6	GapC1	Stramenopile	CCAP 1055/1 glyceraldehyde-3-phosphate dehydrogenase precursor
103	6	Gasu_40150	Rhodophyta	[pt] photosystem II PsbC protein
104	6	Gasu_40150	Rhodophyta	[pt] photosystem II PsbC protein
105	6	Gasu_40260	Rhodophyta	[pt] AAA-type ATPase
106	6	Gasu_40270	Rhodophyta	[pt] photosystem II PsbB protein
107	6	Gasu_40450	Rhodophyta	[pt] photosystem I P700 chlorophyll A apoprotein
108	6	Gasu_40450	Rhodophyta	[pt] photosystem I P700 chlorophyll A apoprotein
109	6	Gasu_40630	Rhodophyta	[pt] F-type H+-transporting ATPase subunit alpha

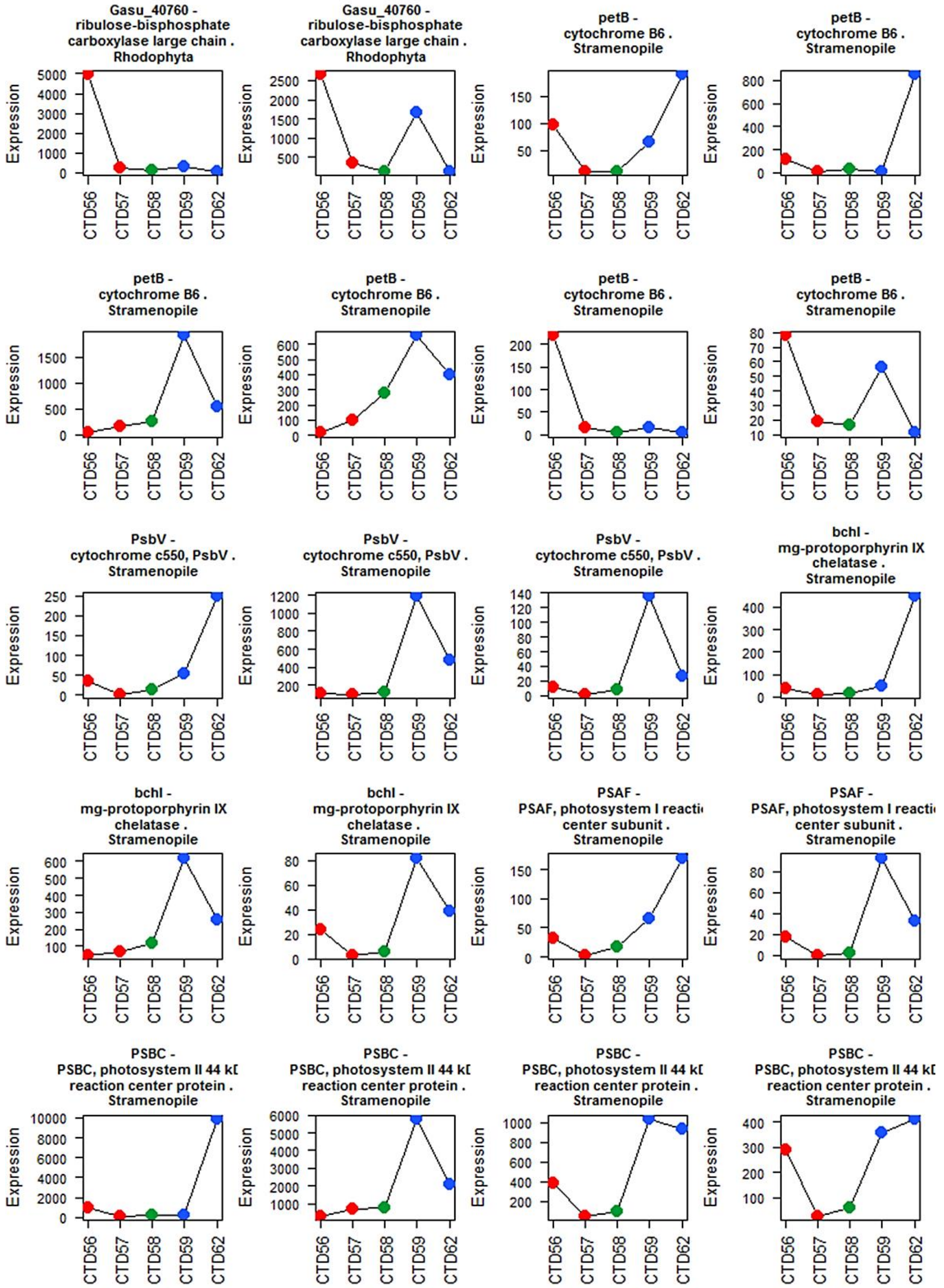
110	6	Gasu_40630	Rhodophyta	[pt] F-type H+-transporting ATPase subunit alpha
111	6	Gasu_40760	Rhodophyta	[pt] ribulose-bisphosphate carboxylase large chain
112	6	Gasu_40100	Rhodophyta	[pt] photosystem II PsbA protein
113	6	petB	Stramenopile	CCMP1335 cytochrome B6
114	6	petF_1	Stramenopile	CCMP1335 ferredoxin
115	6	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
116	6	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
117	6	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
118	6	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
119	6	Unknown	Amoebozoa	HM-1:IMSS midasin, putative
120	6	LMJF_36_3530	Excavata	strain Friedlin putative polyubiquitin
121	6	Unknown	Excavata	MHOM/GT/2001/U1103 elongation factor 1-alpha partial mRNA
122	6	Unknown	Opisthokonta	non-muscle actin mRNA
123	6	Unknown	Stramenopile	VS20 elongation factor 1-alpha 1 mRNA
124	6	Unknown	Stramenopile	tubulin alpha-2 chain mRNA
125	6	Unknown	Stramenopile	tubulin beta chain partial mRNA
126	6	Unknown	Stramenopile	centrin partial mRNA
127	6	Unknown	Stramenopile	Saprolegnia parasitica CBS 223.65 caltractin mRNA
128	6	Unknown	Apusozoa	ATCC 50062 tubulin beta chain partial mRNA
129	7	ACA1_387860	Amoebozoa	str. Neff high molecular weight heat shock protein
130	7	Gasu_40450	Rhodophyta	[pt] photosystem I P700 chlorophyll A apoprotein
131	7	Gasu_40610	Rhodophyta	[pt] F-type H+-transporting ATPase subunit a
132	7	Gasu_40760	Rhodophyta	[pt] ribulose-bisphosphate carboxylase large chain
133	7	Gasu_40100	Rhodophyta	[pt] photosystem II PsbA protein
134	7	petB	Stramenopile	CCMP1335 cytochrome B6
135	7	PSBC	Stramenopile	CCMP1335 PSBC, photosystem II 44 kDa reaction center protein
136	7	Unknown	Excavata	MHOM/BR/75/M2904 putative polyubiquitin partial mRNA
137	7	Unknown	Excavata	beta-tubulin
138	7	Unknown	Excavata	beta-tubulin
139	7	Unknown	Excavata	alpha-tubulin
140	7	Unknown	Alveolata	ATCC 50983 elongation factor 1-alpha, putative
141	7	Unknown	Alveolata	ATCC 50983 60 kDa glycoprotein, putative
142	7	Unknown	Opisthokonta	tubulin beta chain partial mRNA
143	7	Unknown	Opisthokonta	non-muscle actin mRNA
144	7	Unknown	Haptophyta	CCMP1516 putative histone H3 partial mRNA

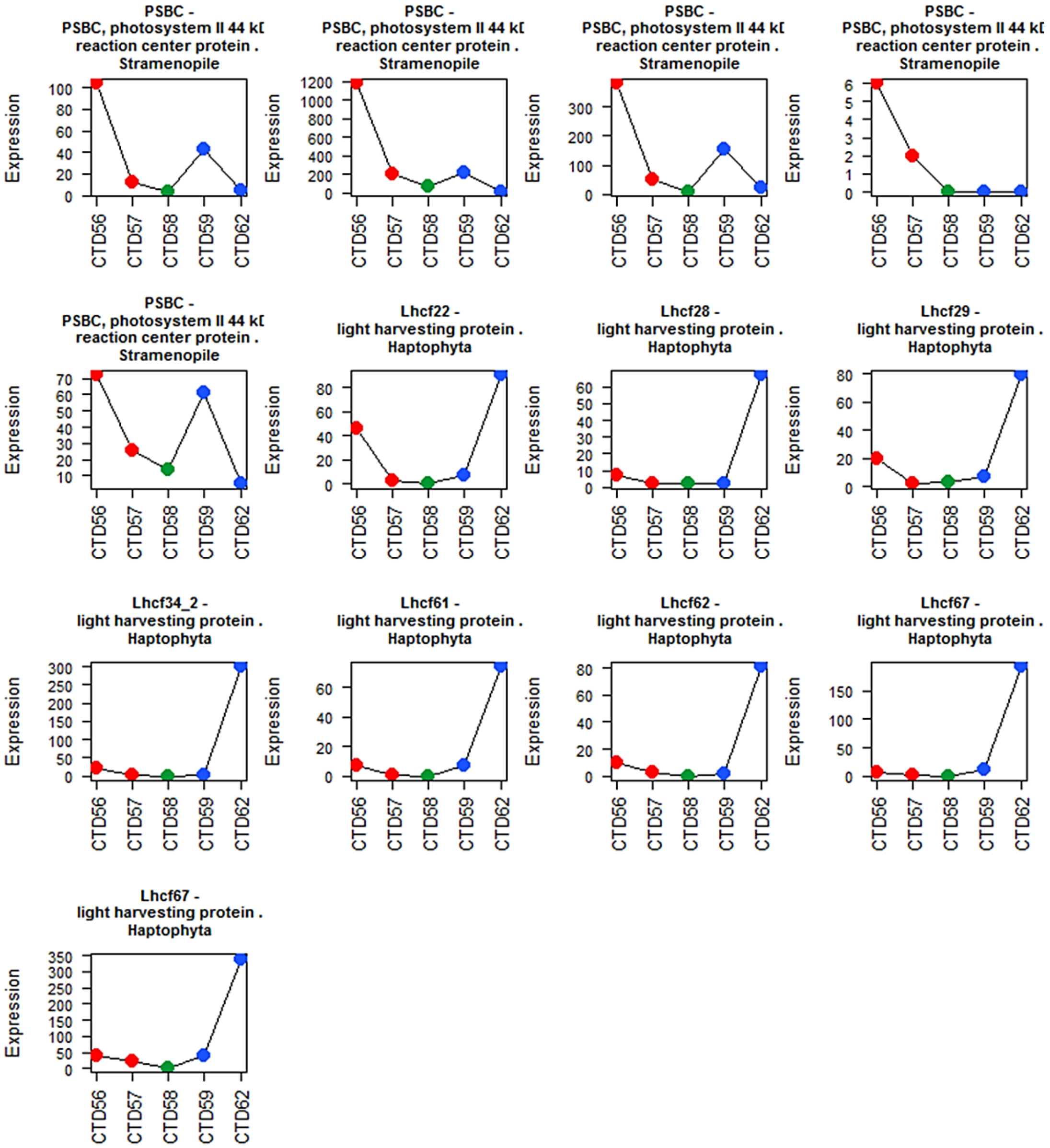
145	7	Unknown	Haptophyta	CCMP1516 alpha-tubulin partial mRNA
146	7	Unknown	Stramenopile	tubulin beta chain mRNA
147	7	Unknown	Stramenopile	histone partial mRNA
148	7	Unknown	Stramenopile	elongation factor 1-alpha partial mRNA
149	7	Unknown	Apicomplexa	ribosomal protein RPL10A partial mRNA
150	7	Unknown	Stramenopile	tubulin alpha-2 chain mRNA
151	7	Unknown	Stramenopile	tubulin beta chain partial mRNA
152	7	Unknown	Stramenopile	tubulin alpha-2 chain mRNA
153	7	Unknown	Stramenopile	histone partial mRNA
154	7	Unknown	Apusozoa	ATCC 50062 actin-3 partial mRNA
155	7	Unknown	Opisthokonta	alpha-tubulin mRNA
156	7	Unknown	Opisthokonta	JP610 histone H2A mRNA
157	7	Unknown	Opisthokonta	JP610 histone H2B type 1-A mRNA
158	8	PKH_131070	Apicomplexa	strain H ubiquitin
159	8	Unknown	Apicomplexa	3D7 conserved protein, unknown function (MAL13P1.455) mRNA
160	8	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
161	8	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
162	8	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
163	8	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
164	8	Unknown	Stramenopile	40S ribosomal protein S2 mRNA
165	8	Unknown	Stramenopile	INRA-310 ATP-dependent RNA helicase eIF4A partial mRNA
166	9	Unknown	Apicomplexa	RN66 senescence-associated protein, putative
167	9	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
168	9	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
169	9	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
170	9	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
171	9	Unknown	Stramenopile	mRNA
172	9	Unknown	Stramenopile	mRNA
173	9	Unknown	Stramenopile	CCMP526 transcript antisense to ribosomal rna protein mRNA
174		Unknown	Apusozoa	ATCC 50062 tubulin beta chain partial mRNA

Appendix 5. Gene Expression Profiles of all Photosynthetic Genes.

Shown are the expression profiles of all annotated genes related to photosynthesis that were. Titles show the annotated gene name, description, and the representative major taxonomic group the gene was annotated as being most similar to. Some descriptions have been simplified due to space constraints. Expression values are total sequence count per gene normalised by library size. Stations are coloured according to the degree of Polar Water influence; red- LI, green- MI, blue- HI.




Gasu_40760 - ribulose-bisphosphate carboxylase large chain . Rhodophyta
Expression
0 1000 2000 3000 4000 5000
Gasu_40760 - ribulose-bisphosphate carboxylase large chain . Rhodophyta
500 1000 1500 2000 2500
petB - cytochrome B6 . Stramenopile
50 100 150
petB - cytochrome B6 . Stramenopile
0 200 400 600 800
petB - cytochrome B6 . Stramenopile
0 500 1000 1500
petB - cytochrome B6 . Stramenopile
0 100 200 300 400 500 600
petB - cytochrome B6 . Stramenopile
0 50 100 150 200
petB - cytochrome B6 . Stramenopile
10 20 30 40 50 60 70 80
PsbV - cytochrome c550, PsbV . Stramenopile
0 50 100 150 200 250
PsbV - cytochrome c550, PsbV . Stramenopile
200 400 600 800 1000 1200
PsbV - cytochrome c550, PsbV . Stramenopile
0 20 40 60 80 100 120 140
bchI - mg-protoporphyrin IX chelatase . Stramenopile
0 100 200 300 400
bchI - mg-protoporphyrin IX chelatase . Stramenopile
100 200 300 400 500 600
bchI - mg-protoporphyrin IX chelatase . Stramenopile
0 20 40 60 80
PSAF - PSAF, photosystem I reacti center subunit . Stramenopile
0 50 100 150
PSAF - PSAF, photosystem I reacti center subunit . Stramenopile
0 20 40 60 80
PSBC - PSBC, photosystem II 44 kI reaction center protein . Stramenopile
0 2000 4000 6000 8000 10000
PSBC - PSBC, photosystem II 44 kI reaction center protein . Stramenopile
1000 2000 3000 4000 5000 6000
PSBC - PSBC, photosystem II 44 kI reaction center protein . Stramenopile
200 400 600 800 1000
PSBC - PSBC, photosystem II 44 kI reaction center protein . Stramenopile
100 200 300 400
CTD56 CTD57 CTD58 CTD59 CTD62


PSBC -
PSBC, photosystem II 44 kI
reaction center protein .
Stramenopile
Expression
0
20
40
60
80
100
CTD56
CTD57
CTD58
CTD59
CTD62
PSBC -
PSBC, photosystem II 44 kI
reaction center protein .
Stramenopile
Expression
0
200
400
600
800
1000
1200
CTD56
CTD57
CTD58
CTD59
CTD62
PSBC -
PSBC, photosystem II 44 kI
reaction center protein .
Stramenopile
Expression
0
100
200
300
CTD56
CTD57
CTD58
CTD59
CTD62
PSBC -
PSBC, photosystem II 44 kI
reaction center protein .
Stramenopile
Expression
0
1
2
3
4
5
6
CTD56
CTD57
CTD58
CTD59
CTD62
PSBC -
PSBC, photosystem II 44 kI
reaction center protein .
Stramenopile
Expression
10
20
30
40
50
60
70
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf22 -
light harvesting protein .
Haptophyta
Expression
0
20
40
60
80
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf28 -
light harvesting protein .
Haptophyta
Expression
0
10
20
30
40
50
60
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf29 -
light harvesting protein .
Haptophyta
Expression
0
20
40
60
80
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf34_2 -
light harvesting protein .
Haptophyta
Expression
0
50
100
150
200
250
300
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf61 -
light harvesting protein .
Haptophyta
Expression
0
20
40
60
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf62 -
light harvesting protein .
Haptophyta
Expression
0
20
40
60
80
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf67 -
light harvesting protein .
Haptophyta
Expression
0
50
100
150
CTD56
CTD57
CTD58
CTD59
CTD62
Lhcf67 -
light harvesting protein .
Haptophyta
Expression
0
50
100
150
200
250
300
350
CTD56
CTD57
CTD58
CTD59
CTD62

Appendix 6. 16S Pre-processing Script

```
#First move into the folder which contains all the raw sequence data for all stations
#Quality check all data files
        mkdir QC_report_RAW

#Create a folder for the output data files and move them into this folder. Clean up files not required.
        fastqc *.fastq
        rm *.zip
        mv *.html QC_report_RAW/

#Quality filter each data file to a phred score of 28 and merge paired end sequences. Remove sequences
shorter than 300bp.
        pear -f CTD-8-16S_S14_L001_R1_001.fastq -r CTD-8-16S_S14_L001_R2_001.fastq -o CTD-8-
16S_S14_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-10-16S_S15_L001_R1_001.fastq -r CTD-10-16S_S15_L001_R2_001.fastq -o CTD-10-
16S_S15_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-12-16S_S16_L001_R1_001.fastq -r CTD-12-16S_S16_L001_R2_001.fastq -o CTD-12-
16S_S16_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-56-16S_S21_L001_R1_001.fastq -r CTD-56-16S_S21_L001_R2_001.fastq -o CTD-56-
16S_S21_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-57-16S_S22_L001_R1_001.fastq -r CTD-57-16S_S22_L001_R2_001.fastq -o CTD-57-
16S_S22_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-58-16S_S23_L001_R1_001.fastq -r CTD-58-16S_S23_L001_R2_001.fastq -o CTD-58-
16S_S23_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-59-16S_S24_L001_R1_001.fastq -r CTD-59-16S_S24_L001_R2_001.fastq -o CTD-59-
16S_S24_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-62-16S_S25_L001_R1_001.fastq -r CTD-62-16S_S25_L001_R2_001.fastq -o CTD-62-
16S_S25_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10

#Move the quality filtered files to their own directory and rerun the fastqc quality check on them. Remove any
uneeded files.
        cd ..
        mkdir Quality_filtered
        mv RAW/*_merged.fastq.* Quality_filtered/
        cd Quality_filtered
        mkdir QC_report_processed
        fastqc *_merged.fastq.assembled.fastq
        rm *.zip
        mv *.html QC_report_processed/

#Convert the quality filtered fastq files to fasta format.
        for f in *_merged.fastq.assembled.fastq
                {
                fastq_to_fasta -Q33 -n -i "$f" -o
"${f/_merged.fastq.assembled.fastq/}_merged_assembled.fasta"
                }

#Open R to trim the sequences to 250bp.
        R --vanilla

#Write the function for trimming.
        trim.sequences.uniform.length<-function(input.fasta, output.fasta, seq.length)
                {
                cut.command<- paste("cut -c1-", seq.length, " ", input.fasta, " > ", output.fasta, sep = "")
                system(cut.command)
                }

#Call the function to perform the trimming.
        trim.sequences.uniform.length("CTD-8-16S_S14_L001_merged_assembled.fasta", "CTD-8-
16S_S14_L001_ready.fasta", 250)
```

```
        trim.sequences.uniform.length("CTD-10-16S_S15_L001_merged_assembled.fasta", "CTD-10-16S_S15_L001_ready.fasta", 250)
        trim.sequences.uniform.length("CTD-12-16S_S16_L001_merged_assembled.fasta", "CTD-12-16S_S16_L001_ready.fasta", 250)
        trim.sequences.uniform.length("CTD-56-16S_S21_L001_merged_assembled.fasta", "CTD-56-16S_S21_L001_ready.fasta", 250)
        trim.sequences.uniform.length("CTD-57-16S_S22_L001_merged_assembled.fasta", "CTD-57-16S_S22_L001_ready.fasta", 250)
        trim.sequences.uniform.length("CTD-58-16S_S23_L001_merged_assembled.fasta", "CTD-58-16S_S23_L001_ready.fasta", 250)
        trim.sequences.uniform.length("CTD-59-16S_S24_L001_merged_assembled.fasta", "CTD-59-16S_S24_L001_ready.fasta", 250)
        trim.sequences.uniform.length("CTD-62-16S_S25_L001_merged_assembled.fasta", "CTD-62-16S_S25_L001_ready.fasta", 250)
        q()

#Make a folder for, and move the trimmed files.
        cd ..
        mkdir Ready
        mv Quality_filtered/*_ready.fasta Ready/

#Check  the sequence length of the trimmed files.
        cd Ready
        for f in *_ready.fasta
                {
                perl /home/shared/PreprocessingFunctionsScripts/ExecutableScripts/fastaNamesSizes.pl "$f"
                }
```

Appendix 7. 18S Pre-processing Script

```
#First move into the folder which contains all the raw sequence data for all stations
#Quality check all data files
        fastqc *.fastq
#Create a folder for the output data files and move them into this folder. Clean up files not required.
        mkdir QC_report_RAW
        rm *.zip
        mv *.html QC_report_RAW/

#Quality filter each data file to a phred score of 28 and merge paired end sequences. Remove sequences shorter than 300bp.
        pear -f CTD-8-18S_S1_L001_R1_001.fastq -r CTD-8-18S_S1_L001_R2_001.fastq -o CTD-8-18S_S1_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-10-18S_S2_L001_R1_001.fastq -r CTD-10-18S_S2_L001_R2_001.fastq -o CTD-10-18S_S2_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-12-18S_S3_L001_R1_001.fastq -r CTD-12-18S_S3_L001_R2_001.fastq -o CTD-12-18S_S3_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-56-18S_S7_L001_R1_001.fastq -r CTD-56-18S_S7_L001_R2_001.fastq -o CTD-56-18S_S7_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-57-18S_S8_L001_R1_001.fastq -r CTD-57-18S_S8_L001_R2_001.fastq -o CTD-57-18S_S8_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-58-18S_S9_L001_R1_001.fastq -r CTD-58-18S_S9_L001_R2_001.fastq -o CTD-58-18S_S9_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-59-18S_S10_L001_R1_001.fastq -r CTD-59-18S_S10_L001_R2_001.fastq -o CTD-59-18S_S10_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10
        pear -f CTD-62-18S_S11_L001_R1_001.fastq -r CTD-62-18S_S11_L001_R2_001.fastq -o CTD-62-18S_S11_L001_merged.fastq -q 28 -m 300 -n 100  -p 0.0001 -j 10

#Move the quality filtered files to their own directory and rerun the fastqc quality check on them. Remove any uneeded files.
        cd ..
        mkdir Quality_filtered
        mv RAW/*_merged.fastq.* Quality_filtered/
        cd Quality_filtered
        mkdir QC_report_processed
        fastqc *_merged.fastq.assembled.fastq
        rm *.zip
        mv *.html QC_report_processed/

#Convert the quality filtered fastq files to fasta format.
        for f in *_merged.fastq.assembled.fastq
                {
                fastq_to_fasta -Q33 -n -i "$f" -o "${f/_merged.fastq.assembled.fastq/}_merged_assembled.fasta"
                }

#Open R to trim the sequences to 270bp.
        R --vanilla
#Write the function for trimming.
        trim.sequences.uniform.length<-function(input.fasta, output.fasta, seq.length)
                {
                cut.command<- paste("cut -c1-", seq.length, " ", input.fasta, " > ", output.fasta, sep = "")
                system(cut.command)
                }

#Call the function to perform the trimming.
        trim.sequences.uniform.length("CTD-8-18S_S1_L001_merged_assembled.fasta", "CTD-8-18S_S1_L001_ready.fasta", 270)
        trim.sequences.uniform.length("CTD-10-18S_S2_L001_merged_assembled.fasta", "CTD-10-18S_S2_L001_ready.fasta", 270)
        trim.sequences.uniform.length("CTD-12-18S_S3_L001_merged_assembled.fasta", "CTD-12-
18S_S3_L001_ready.fasta", 270)
        trim.sequences.uniform.length("CTD-56-18S_S7_L001_merged_assembled.fasta", "CTD-56-
18S_S8_L001_ready.fasta", 270)
        trim.sequences.uniform.length("CTD-57-18S_S8_L001_merged_assembled.fasta", "CTD-57-
18S_S9_L001_ready.fasta", 270)
        trim.sequences.uniform.length("CTD-58-18S_S9_L001_merged_assembled.fasta", "CTD-58-
18S_S10_L001_ready.fasta", 270)
        trim.sequences.uniform.length("CTD-59-18S_S10_L001_merged_assembled.fasta", "CTD-59-
18S_S11_L001_ready.fasta", 270)
        trim.sequences.uniform.length("CTD-62-18S_S11_L001_merged_assembled.fasta", "CTD-62-
18S_S12_L001_ready.fasta", 270)
        q()

#Make a folder for, and move the trimmed files.
        cd ..
        mkdir Ready
        mv Quality_filtered/*_ready.fasta Ready/

#Check  the sequence length of the trimmed files.
        cd Ready
        for f in *_ready.fasta
                {
                perl /home/shared/PreprocessingFunctionsScripts/ExecutableScripts/fastaNamesSizes.pl
"$f"
                }
```

Appendix 8. RNA Pre-processing Script

```
#First move into the folder which contains all the raw sequence data for all stations.
#Quality check all data files.
        mkdir QC_report_RAW

#Create a folder for the output data files and move them into this folder. Clean up files not required.
                fastqc *.fastq
        rm *.zip
        mv *.html QC_report_RAW/

#Trim first and last 10 bases as fastqc showed these to be of lower quality.
        for f in ctd*
                {
                fastx_trimmer -f 11 -l 291 -i "$f" -o "${f/_001.fastq/}"_001_trimmed.fastq
                }
        done

#Create a folder for the trimmed data files and move them into this folder.
        cd ..
        mkdir Trimmed
        mv RAW/*_001_trimmed.fastq Trimmed/

#Quality check the trimmed data files.
        fastqc Trimmed/*_001_trimmed.fastq

#Create a folder for the output data files and move them into this folder. Clean up files not required.
        mkdir Trimmed/QC_report_Trimmed
        rm Trimmed/*.zip
        mv Trimmed/*.html Trimmed/QC_report_Trimmed


#Quality filter the trimmed data files at a phred score of 28.
        for f in Trimmed/ctd*
                {
                cutadapt -q 28 -o "${f/_001_trimmed.fastq/}"_001_trimmed_filtered.fastq "$f"
                }
        done

#Create a folder for the output data files and move them into this folder.
        mkdir Quality_filtered_trimmed
        mv Trimmed/*_001_trimmed_filtered.fastq Quality_filtered_trimmed/

#Quality check the quality filtered data files.
        fastqc Quality_filtered_trimmed/*_001_trimmed_filtered.fastq

#Create a folder for the output data files and move them into this folder. Clean up files not required.
        mkdir Quality_filtered_trimmed/QC_report_quality_filtered_trimmed
        rm Quality_filtered_trimmed/*.zip
        mv Quality_filtered_trimmed/*.html Quality_filtered_trimmed/QC_report_quality_filtered_trimmed



#Remove primers and over represented sequences flagged by FAST-QC
        cd Quality_filtered_trimmed
        cutadapt -a GCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAA -a
GCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAAAA -a
AAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAAGAA -o
ctd10_S1_L001_R1_001_trimmed_filtered_primers.fastq ctd10_S1_L001_R1_001_trimmed_filtered.fastq
        cutadapt -a GCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAAAAA -a
GCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATTAAAAAAAAAAAA -o
ctd10_S1_L001_R2_001_trimmed_filtered_primers.fastq ctd10_S1_L001_R2_001_trimmed_filtered.fastq
```

```
            cutadapt -a AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -a
AACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAA -a
ACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAACAAC -o
ctd56_S6_L001_R1_001_trimmed_filtered_primers.fastq ctd56_S6_L001_R1_001_trimmed_filtered.fastq
            cutadapt -a GTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGTTGT -o
ctd56_S6_L001_R2_001_trimmed_filtered_primers.fastq ctd56_S6_L001_R2_001_trimmed_filtered.fastq
            cutadapt -a AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -o
ctd57_S4_L001_R2_001_trimmed_filtered_primers.fastq ctd57_S4_L001_R2_001_trimmed_filtered.fastq
            cutadapt -a TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT -o
ctd59_S3_L001_R1_001_trimmed_filtered_primers.fastq ctd59_S3_L001_R1_001_trimmed_filtered.fastq
            cutadapt -a AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -o
ctd59_S3_L001_R2_001_trimmed_filtered_primers.fastq ctd59_S3_L001_R2_001_trimmed_filtered.fastq
            cutadapt -a AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -a
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT -o
ctd62_S10_L001_R1_001_trimmed_filtered_primers.fastq ctd62_S10_L001_R1_001_trimmed_filtered.fastq
            cutadapt -a AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -o
ctd62_S10_L001_R2_001_trimmed_filtered_primers.fastq ctd62_S10_L001_R2_001_trimmed_filtered.fastq

#Create a folder for the output data files and move them into this folder.
            cd ..
            mkdir Quality_filtered_trimmed_primers
            mv Quality_filtered_trimmed/*primers.fastq Quality_filtered_trimmed_primers/

#Format the file names of files which were not reported to contain primers and over represented sequences so
all files are the same. Move these into a common folder.
            reads=(8 12 58)
            for f in "${reads[@]}"
                        {
                        cp Quality_filtered_trimmed/ctd"$f"* Quality_filtered_trimmed_primers/
                        }
            cp Quality_filtered_trimmed/ctd57_S4_L001_R1_001_trimmed_filtered.fastq
Quality_filtered_trimmed_primers/ctd57_S4_L001_R1_001_trimmed_filtered.fastq

#Quality check the quality filtered data files.
            fastqc Quality_filtered_trimmed_primers/*.fastq

#Create a folder for the output data files and move them into this folder. Clean up files not required.
            mkdir Quality_filtered_trimmed_primers/QC_report_filtered_trimmed_primers
            rm Quality_filtered_trimmed_primers/*.zip
            mv Quality_filtered_trimmed_primers/*.html
Quality_filtered_trimmed_primers/QC_report_filtered_trimmed_primers


###Remove sequences shorter than 100bp
            cd Quality_filtered_trimmed_primers
            reads=(8 10 12 56 57 58 59 62 64)
            for f in "${reads[@]}"
                        {
                        java -jar /home/mikgat/Downloads/Trimmomatic-0.36/trimmomatic-0.36.jar PE
ctd"$f"*_R1_*.fastq ctd"$f"*_R2_*.fastq ctd"$f"_R1_output_paired.fastq ctd"$f"_R1_output_unpaired.fastq
ctd"$f"_R2_output_paired.fastq ctd"$f"_R2_output_unpaired.fastq MINLEN:100
                        }

#Create a folder for the output data files and move them into this folder.
            cd ..
            mkdir Quality_filtered_trimmed_primers_100bp
            mv Quality_filtered_trimmed_primers/ctd*output* Quality_filtered_trimmed_primers_100bp/

#Quality check the quality filtered data files.
            fastqc Quality_filtered_trimmed_primers_100bp/ctd*_output_paired.fastq

#Create a folder for the output data files and move them into this folder. Clean up files not required.
```

```
            mkdir Quality_filtered_trimmed_primers_100bp/QC_report_filtered_trimmed_primers_100bp
            rm Quality_filtered_trimmed_primers_100bp/*.zip
            mv Quality_filtered_trimmed_primers_100bp/*.html
Quality_filtered_trimmed_primers_100bp/QC_report_filtered_trimmed_primers_100bp


#Convert the quality filtered FASTQ files to fasta format.
            cd Quality_filtered_trimmed_primers_100bp
            for f in ctd*_output_paired.fastq
                        {
                        fastq_to_fasta -Q33 -n -i "$f" -o "${f/.fastq/}_ready.fasta"
            }

#Create a folder for the output data files and move them into this folder.
            cd ..
            mkdir Ready_trimmed
            mv Quality_filtered_trimmed_primers_100bp/*.fasta Ready_trimmed/
```

Appendix 9. Local SILVA Database Creation Script

```
#Change into the SILVA directory
        cd /home/shared/silva/

#Make a new directory for the latest database release
        mkdir release_128
        cd release_128

#Download the latest release
        wget https://www.arb-silva.de/fileadmin/silva_databases/release_128/Exports/SILVA_128_SSURef_Nr99_tax_silva_trunc.fasta.gz
        gzip -d SILVA_128_SSURef_Nr99_tax_silva_trunc.fasta.gz

#Prepare downloaded fasta file
        python /home/shared/silva/FormattingScripts/formatSilva.py /home/shared/silva/release_128/SILVA_128_SSURef_Nr99_tax_silva_trunc.fasta /home/shared/silva/release_128/SILVA_128_SSURef_Nr99_taxonomy_mapping_for_Qiime.txt /home/shared/silva/release_128/SILVA_128_SSURef_Nr99_Sequences_for_Qiime.fasta

#Separate the sequences into eukaryotes, bacteria and archaea
        grep Eukaryota SILVA_128_SSURef_Nr99_taxonomy_mapping_for_Qiime.txt | cut -f1 > AccessionEukaryota.txt
        grep Bacteria SILVA_128_SSURef_Nr99_taxonomy_mapping_for_Qiime.txt | cut -f1 > AccessionBacteria.txt
        grep Archaea SILVA_128_SSURef_Nr99_taxonomy_mapping_for_Qiime.txt | cut -f1 > AccessionArchaea.txt

#Extract the sequences and accession numbers from the text files above. Make sure the python file being called has been edited to reflect the lastest version number of SILVA.
        python /home/mikgat/PhD/16S/ExtractAccessionIDsFromFasta.py

#Extract the corresponding sequence entries for the mapping files.
        grep 'Bacteria' SILVA_128_SSURef_Nr99_taxonomy_mapping_for_Qiime.txt > Bacteria_Mapping.txt
        grep 'Archaea' SILVA_128_SSURef_Nr99_taxonomy_mapping_for_Qiime.txt > Archaea_Mapping.txt
        grep 'Eukaryota' SILVA_128_SSURef_Nr99_taxonomy_mapping_for_Qiime.txt > Eukaryota_Mapping.txt

#Check fasta and mapping files have the same number of lines.
        grep -c '>' *.fasta
        wc -l *_Mapping.txt

#Trim the sequences to just the region of interest (V5 for bacteria and V9 for eukaryotes).
        cutadapt -g GCCGCGGTAA -e 0.15 -O 7 -m 250 -o SILVA_128BacteriaV5.fasta Bacteria.fasta
        cutadapt -g GCCGCGGTAA -e 0.15 -O 7 -m 250 -o SILVA_128ArchaeaV5.fasta Archaea.fasta
        cutadapt -g ACCGCCCGTC -e 0.15 -O 7 -m 270 -o SILVA_128EukaryotaV9.fasta Eukaryota.fasta

#Get IDs of only sequences contained in output from this trimming.
        grep '>' SILVA_128BacteriaV5.fasta | tr -d '>' > AccessionIDsBacteriaV5.txt
        grep '>' SILVA_128ArchaeaV5.fasta | tr -d '>' > AccessionIDsArchaeaV5.txt
        grep '>' SILVA_128EukaryotaV9.fasta | tr -d '>' > AccessionIDsEukaryotaV9.txt

#Output a mapping file of just the accession IDs and V5 and V9 markers
        join -1 1 -2 1 <(sort AccessionIDsBacteriaV5.txt) <(sort Bacteria_Mapping.txt) > MappingBacteriaV5.txt
        join -1 1 -2 1 <(sort AccessionIDsArchaeaV5.txt) <(sort Archaea_Mapping.txt) > MappingArchaeaV5.txt
        join -1 1 -2 1 <(sort AccessionIDsEukaryotaV9.txt) <(sort Eukaryota_Mapping.txt) > MappingEukaryotaV9.txt

#Collate Archaea and Bacterial files.
        cat SILVA_128BacteriaV5.fasta SILVA_128ArchaeaV5.fasta > Silva_128_V5.fasta
        cat MappingBacteriaV5.txt MappingArchaeaV5.txt > Silva_128_V5_Mapping.txt
```

```
#Rename the Eukaryote file to match the prokaryote format
        mv SILVA_128EukaryotaV9.fasta Silva_128_V9.fasta
        mv MappingEukaryotaV9.txt Silva_128_V9_Mapping.txt

#Trim V5 and V9 sequences to lengths corresponding to my read lengths.
cat Silva_128_V5.fasta | cut -c 1-250 > Silva_128_V5_200_bases.fasta
cat Silva_128_V9.fasta | cut -c 1-270 > Silva_128_V9_270_bases.fasta

#make the db
makeblastdb -in Silva_128_V5_250_bases.fasta -input_type fasta -dbtype nucl -title SILVA_128_V5_250 -parse_seqids -out SILVA_128_V5_250
makeblastdb -in Silva_128_V9_270_bases.fasta -input_type fasta -dbtype nucl -title SILVA_128_V9_270 -parse_seqids -out SILVA_128_V9_270
```

Appendix 10. Swarm Rscript

```
#!/usr/bin/Rscript
        swarm <- read.table ("swarm_order.txt", header = F, as.is = T)
        blast <- read.table ("ParallelBlastTaxonomy128/taxonomy_file_to_create_labelled_otu_tab.txt",
header = F, as.is = T, sep = "\t", quote = "")
        row.names(swarm) <- swarm[,1]
        row.names(blast) <- blast[,1]
        blast <- blast[swarm[,1],]
        write.table(blast, col.names = F, row.names = F, quote = F, sep = "\t", file =
"ParallelBlastTaxonomy128/RepSeqTaxonomyInSwarmOrderedForAddingOTUContingencyTable.txt")

#Match amplicon IDs to OTU table and strip _numeric from amplicon column
        blast <-
read.table("ParallelBlastTaxonomy128/RepSeqTaxonomyInSwarmOrderedForAddingOTUContingencyTable.txt",
header = F, as.is = T, sep = "\t", quote = "")

#change the names of the headers (V1, V2)
        names(blast) <- c("amplicon", "taxonomy")
        temp <- strsplit(blast[,1], split = "_", fixed = T)

#define the amplicon column without the appended read count
        amplicon<-rep("", length(temp))
        for(i in 1:length(amplicon))
                {
                amplicon[i]<- temp[[i]][1]
                }
        blast[,1]<-amplicon

#write the table to a file
        write.table(blast, quote = F, sep = "\t", col.names = names(blast), row.names = F, file
="ParallelBlastTaxonomy128/RepSeqTaxonomyInSwarmOrderedForAddingOTUContingencyTable_final.txt")
        q()
```

Appendix 11. 18S OTU Clustering and Taxonomic Assignment Script

```
#Set the current directory as a variable
        r="$PWD"

#Make new folders for each of the pre-processed sample files and copy the relevant data file into them.
        stations=(8 10 12 56 57 58 59 62)
                for s in ${stations[@]}
                        do
                                mkdir Data"$s"
                                cp Preprocessing/Ready/CTD-"$s"-18S*_ready.fasta Data"$s"/
                done

#Make a new folder to store the dereplicated files generated by Swarm.
        mkdir DereplicatedFastaFiles

#Loop through each sample file and prepare it for dereplication.
        for f in Data*
                do
                        for d in "$f"/*_ready.fasta;
                                do
                                        echo -e "\nProcessing $d file...";
                                        bash
/home/shared/PreprocessingFunctionsScripts/ExecutableScripts/prepare_data_swarm.sh "$d"
"${d/_ready.fasta/}"_prepared.fasta;
                                        echo -e "\nFile $d : processing complete";
                                done

                        for e in "$f"/*_prepared.fasta;
                                do
                                        echo -e "\nMoving file $e...";
                                        cp "$e" DereplicatedFastaFiles/
                                        echo -e "\nDone.";
                                done
                        done

#Create a dereplicated file which contains only unique sequences found across all repeats. This must be done
from the folder containing only dereplicated files.
        cd DereplicatedFastaFiles/
        echo -e "\nDereplicating entire study...";
        /home/shared/PreprocessingFunctionsScripts/ExecutableScripts/dereplicate_whole_study_for_swar
m.sh
        echo -e "\nDone.";

#Make a folder ready to run Swarm and move the dereplicated file into it.
        mkdir Swarm
        cp all_samples.fa Swarm/

#Create an amplicon contingency table of all OTUs in all sample files.
        echo -e "\nCreating amplicon contingency table...";
        python
/home/shared/PreprocessingFunctionsScripts/ExecutableScripts/amplicon_contingency_table.py
*_prepared.fasta > Swarm/100_ID_OTU_table.csv
        echo -e "\nDone.";
        cd ..

#Run Swarm.
        echo -e "\nRunning Swarm...";
```

```
		swarm -d 1 -f -t 4 -w DereplicatedFastaFiles/Swarm/fastidious_seed_sequence.fasta -s
DereplicatedFastaFiles/Swarm/fastidious.amplicon.stats -o DereplicatedFastaFiles/Swarm/fastidious.output -l
DereplicatedFastaFiles/Swarm/fastidious.logfile DereplicatedFastaFiles/Swarm/all_samples.fa

#Create an OTU table for the swarm cluster (from within the folder containing the swarm outputs).
		cd DereplicatedFastaFiles/Swarm
		bash
/home/shared/PreprocessingFunctionsScripts/ExecutableScripts/make_OTU_table_post_swarm.sh
fastidious.amplicon.stats fastidious.output 100_ID_OTU_table.csv OTU_contingency_table.csv
		echo -e "\nDone.";

#Set the default path for the database to be used for taxonomic assignment of OTUs (see "make_silva_db"
script for database generation details).
		echo -e "\nExporting database...";
		export BLASTDB=/home/shared/silva/release_128/
		echo -e "\nDatabase ready.";


#Making a directory to store the result and then assign taxonomy to the OTUs using qiime.
		mkdir ParallelBlastTaxonomy128
		echo -e "\nAssigning taxonomy...";
		qiime parallel_assign_taxonomy_blast.py -i fastidious_seed_sequence.fasta -t
/home/shared/silva/release_128/Eukaryota_Mapping.txt -b
/home/shared/silva/release_128/SILVA_128_V9_270 -o ParallelBlastTaxonomy128/ -U
/usr/lib/qiime/bin/start_parallel_jobs.py -B /usr/share/ncbi/data -e 0.00000001 -O 10

#Extract the required columns to label the OTU table with taxonomic assignments and match IDs to give an
annotated table of OTUs and their abundance at each station. See "Rscript_for_swarm(18S)" for details of the
matching process.
		cat ParallelBlastTaxonomy128/fastidious_seed_sequence_tax_assignments.txt | cut -f1,2 >
ParallelBlastTaxonomy128/taxonomy_file_to_create_labelled_otu_tab.txt
		grep '>' fastidious_seed_sequence.fasta | tr -d '>' > swarm_order.txt
		echo -e "\nProcessing taxonomy files for dataset $f...";

		"$r"/Rscript_for_swarm.r

		join OTU_contingency_table.csv
ParallelBlastTaxonomy128/RepSeqTaxonomyInSwarmOrderedForAddingOTUContingencyTable_final.txt --
header -t $'\t' -1 2 -2 1 | cut -f2-50,52 > Taxonomy_OTU_Contingency_table_all.txt
		echo -e "\nDone.";

#At this point we now have an annotated table of OTUs and their abundance at each station as a text file. This
was exported for community analysis. Code below formats a copy of this table for the generation of Krona
charts after OTUs had been rarefied.
#Rename the OTU table header to OTU_ID so it is compatible with classic biom formatting.
		sed -i 's/OTU/OTU_ID/g' Taxonomy_OTU_Contingency_table_all.txt
		echo -e "\nProcessing of data set done."

#Convert the OTU table to biom format
biom convert -i Taxonomy_OTU_Contingency_table_all.txt -o Taxonomy_OTU_Contingency_table_all.biom --
table-type 'OTU table' --to-hdf5 --process-obs-metadata taxonomy

#Collapse OTUs based on their taxonomic level using qiime ready for the generation of Krona charts.
		echo -e "\nCollapsing OTUs based on taxonomy for dataset $f..."
		qiime summarize_taxa.py -i Taxonomy_OTU_Contingency_table_all.biom -a -o
./Taxonomy_OTU_Contingency_table_all/ -L 12
		echo -e "\nDone.";


#Run Krona to generate interactive html charts
		ktImportText Taxonomy_OTU_Contingency_table_all_L12.txt -o
Taxonomy_OTU_Contingency_table_all_final_report_L12.html
```

Appendix 12. 16S OTU Grouping and Taxonomic Assignment Script

```
#Set the current directory as a variable.
    r="$PWD"

#Make a new directory to store the data files during processing.
    mkdir Processed_data

    for f in Preprocessing/Ready
        do
            echo -e "\nProcessing data set $f..."

#Move each pre-processed sample file into the new directory.
            for d in "$f"/*_ready.fasta
                do
                    echo -e "\nMoving file $d...";
                    cp "$d" Processed_data/
                    echo -e "\nDone.";
                done
        done

#Merge all sample files to create a mapping file ready for use in qiime. All sequences will be automatically
prefixed with their sample ID.
    add_qiime_labels.py -m mapping_file.txt -i Processed_data/ -c File_name -o Processed_data/

#Change into the new directory.
    cd Processed_data/

#Bin OTUs at 98.7% similarity.
    mkdir OTUs
    pick_otus.py -i combined_seqs.fasta -m cdhit -o OTUs/ -n 250 -M 2000 -s 0.987 --threads 10

#Pick representative sequences of each OTUs.
    pick_rep_set.py -i OTUs/*_otus.txt -f combined_seqs.fasta -m most_abundant -o
OTUs/Representative_sequences.fasta -l OTUs/Representative_logfile.txt

#Set the default path for the database to be used for taxonomic assignment of OTUs (see "make_silva_db"
script for database generation details).
    echo -e "\nExporting database...";
    export BLASTDB=/home/shared/silva/release_128/
    echo -e "\nDatabase ready.";

#Assign taxonomy to the binned OTUs.
    parallel_assign_taxonomy_blast.py -i OTUs/Representative_sequences.fasta -t
/home/shared/silva/release_128/Bacteria_Mapping.txt -b
/home/shared/silva/release_128/SILVA_128_V5_250 -o ParallelBlastTaxonomy128/ -U
/usr/lib/qiime/bin/start_parallel_jobs.py -B /usr/share/ncbi/data -e 0.00000001 -O 10

#Extract the required columns for taxonomic annotation of OTUs.
    cat ParallelBlastTaxonomy128/fastidious_seed_sequence_tax_assignments.txt | cut -f1,2 >
ParallelBlastTaxonomy128/taxonomy_file_to_create_labelled_otu_tab.txt

#Add the taxonomic information to the OTU table
    make_otu_table.py -i combined_seqs_otus.txt -o ParallelBlastTaxonomy128/biom/otu_table.biom -t
ParallelBlastTaxonomy128/taxonomy_file_to_create_labelled_otu_tab.txt

#Convert output from biom to txt files
    biom convert -i ParallelBlastTaxonomy128/biom/otu_table.biom -o
ParallelBlastTaxonomy128/biom/otu_table.txt --table-type "OTU Table" --to-tsv --header-key taxonomy
```

#At this point we now have an annotated table of OTUs and their abundance at each station as a text file. This was exported for community analysis. Code below detailed the generation of Krona charts after OTUs had been rarefied.

#Collapse OTUs based on their taxonomic level using qiime ready for the generation of Krona charts.

```
        summarize_taxa.py -i ParallelBlastTaxonomy128/biom/otu_table.biom -L 7 -a -o
ParallelBlastTaxonomy128/biom/collapsed
```

#Run Krona to generate interactive html charts

```
        mkdir ParallelBlastTaxonomy128/Krona_charts
        cd ParallelBlastTaxonomy128/Krona_charts
        for f in krona_data_*.txt
                do
                        echo -e "\nGenerating visualisation of $f file..."
                        ktImportText "$f" -o "${f/.txt/}"_final_report_L7.html
                        echo -e "\nDone.";
                done
```

Bibliography

1. Swingland, I. R. Biodiversity, Definition of. *Encyclopedia of Biodiversity* **1**, 399–410 (20100).

2. Farnsworth, K. D., Adenuga, A. H. & Groot, R. S. De. The complexity of biodiversity : A biological perspective on economic valuation. *Ecological Economics* **120**, 350–354 (2015).

3. Farnsworth, K., Nelson, J. & Gershenson, C. Living is Information Processing: From Molecules to Global Systems. *Acta Biotheoretica* **61**, 203–222 (2013).

4. Falgueras, J. *et al.* SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC bioinformatics* **11**, 1–12 (2010).

5. Bartkowski, B., Lienhoop, N. & Hansjurgens. Capturing the complexity of biodiversity: A critical review of economic valuation studies of biological diversity. *Ecological Economics* **113**, 1–14 (2015).

6. Smith, S. J., Edmonds, J., Hartin, C. A., Mundra, A. & Calvin, K. Near-term acceleration in the rate of temperature change. *Nature Climate Change* **5**, 333–336 (2015).

7. Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, P. M. M. *IPCC, 2013: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (2013).

8. Duce, R. a *et al.* Impacts of atmospheric anthropogenic nitrogen on the open ocean. *Science* **320**, 893–897 (2008).

9. Lewis, S. L. & Maslin, M. a. Defining the Anthropocene. *Nature* **519**, 171–180 (2015).

10. Ceballos, G., Ehrlich, P. R. & Dirzo, R. Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. *PNAS* **114**, 6089–6096 (2017).

11. Dziba, L., Erpul, G., Fazel, A., Fischer, M. & Hernández, A. M. *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. (2019).

12. Thompson, R. & Starzomski, B. M. What does biodiversity actually do? A review for managers and policy makers. *Biodiversity and Conservation* **16**, 1359–1378 (2007).

13. Petrou, K. *et al.* Southern Ocean phytoplankton physiology in a changing climate. *Journal of Plant Physiology* **203**, 135–150 (2016).

14. Lovejoy, C. & Potvin, M. Microbial eukaryotic distribution in a dynamic Beaufort Sea and the Arctic Ocean. *Journal of Plankton Research* **33**, 431–444 (2011).

15. de Groot, R. *et al.* Global estimates of the value of ecosystems and their services in monetary units. *Ecosystem Services* **1**, 50–61 (2012).

16. Bartkowski, B. *et al.* Capturing the complexity of biodiversity: A critical review of economic valuation studies of biological diversity. *Ecological Economics* **113**, 1–14 (2015).

17. Hooper, D. U., Chapin III, F. S. & Ewel, J. J. Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. *Ecological Monographs* **75**, 3–35 (2005).

18. Reiss, J., Bridle, J. R., Montoya, J. M. & Woodward, G. Emerging horizons in biodiversity and ecosystem functioning research. *Trends in Ecology and Evolution* **24**, 505–514 (2009).

19. Maira, L., Laureto, O., Vinicius, M., Soares, D. & Samia, M. Functional diversity : an overview of its history and applicability. *Natureza & Conservação* **13**, 112–116 (2015).

20. Snelgrove, P. V. R., Thrush, S. F., Wall, D. H. & Norkko, A. Real world biodiversity–ecosystem functioning: a seafloor perspective. *Trends in Ecology & Evolution* **29**, 398–405 (2014).

21. Petchey, O. L. & Gaston, K. J. Functional diversity: back to basics and looking forward. *Ecology Letters* **9**, 741–758 (2006).

22. Duffy, J. E. Why biodiversity is important to the functioning of real-world ecosystems. *Frontiers in Ecology and the Environment* **7**, 437–444 (2009).

23. Bulling, M. T., White, P. C. L., Raffaelli, D. G. & Pierce, G. J. Using model systems to address the biodiversity-ecosystem functioning process. *Marine Ecology Progress Series* **311**, 295–309 (2006).

24. Schwartz, M. W. *et al.* Linking biodiversity to ecosystem function: implications for conservation ecology. *Oecologia* **122**, 297–305 (2000).

25. Hicks, N. *et al.* Impact of biodiversity-climate futures on primary production and metabolism in a model benthic estuarine system. *BMC Ecology* **11**, 7 (2011).

26. Clavel, J., Julliard, R. & Devictor, V. Worldwide decline of specialist species: Toward a global functional homogenization? *Frontiers in Ecology and the Environment* **9**, 222–228 (2011).

27. Richmond, C. E., Breitburg, D. L. & Rose, K. A. The role of environmental generalist species in ecosystem function. *Ecological Modelling* **188**, 279–295 (2005).

28. Jeschke, J. M. & Strayer, D. L. Invasion success of vertebrates in Europe and North America. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 7198–202 (2005).

29. Leidenberger, S. *et al.* Evaluating the potential of ecological niche modelling as a component in marine non-indigenous species risk assessments. *Marine Pollution Bulletin* **97**, 470–487 (2015).

30. Lowry, E. *et al.* Biological invasions: A field synopsis, systematic review, and database of the literature. *Ecology and Evolution* **3**, 182–196 (2013).

31. Cadotte, M. W., Carscadden, K. & Mirotchnick, N. Beyond species: Functional diversity and the maintenance of ecological processes and services. *Journal of Applied Ecology* **48**, 1079–1087 (2011).

32. Jennifer B. Hughes Martiny, B. J. M. B. *et al.* Microbial biogeography: putting microorganisms on the map. *Nature reviews. Microbiology* **4**, 102–112 (2006).

33. Finlay, B. J., Esteban, G. F., Olmo, J. L. & Tyler, P. A. Global distribution of free-living microbial species. *Ecography* **22**, 138–144 (1999).

34. Finlay, B. J. Global dispersal of free-living microbial eukaryote species. *Science* **296**, 1061–1063 (2002).

35. Jones, S. E. & Lennon, J. T. Dormancy contributes to the maintenance of microbial diversity. *PNAS* **107**, 5881–5886 (2010).

36. Atkins, M. S., Teske, A. P. & Anderson, O. R. A survey of flagellate diversity at four deep-sea hydrothermal vents in the Eastern Pacific Ocean using structural and molecular approaches. *Journal of Eukaryotic Microbiology* **47**, 400–411 (2000).

37. Horner-Devine, M. C., Carney, K. M. & Bohannan, B. J. M. An ecological perspective on bacterial biodiversity. *Proceedings of the Royal Society B: Biological Sciences* **271**, 113–122 (2004).

38. Meier, S. & Soininen, J. Phytoplankton metacommunity structure in subarctic rock pools. *Aquatic Microbial Ecology* **73**, 81–91 (2014).

39. Pommier, T. *et al.* Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology* **16**, 867–880 (2007).

40. Zinger, L. *et al.* Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS ONE* **6**, 1–11 (2011).

41. Follows, M. J., Dutkiewicz, S., Grant, S. & Chisholm, S. Emergent Biogeography of Microbial communities in a model ocean. *Science* **315**, 1843–1847 (2007).

42. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

43. Beaugrand, G., Reid, P. C., Lindley, J. A., Edwards, M. & Beaugrand, G. Reorganization of North Atlantic Marine Copepod Biodiversity and Climate. *Science* **296**, 1692–1694 (2002).

44. Winter, A., Henderiks, J., Beaufort, L., Rickaby, R. E. M. & Brown, C. W. Poleward expansion of the coccolithophore Emiliania huxleyi. *Journal of Plankton Research* **36**, 316–325 (2014).

45. Barberán, A. & Casamayor, E. O. Global phylogenetic community structure and β-diversity patterns in surface bacterioplankton metacommunities. *Aquatic Microbial Ecology* **59**, 1–10 (2010).

46. Jyrkänkallio‐ Mikkola, J. *et al.* Disentangling multi‐ scale environmental effects on stream microbial communities.pdf. *Journal of Biogeography* **44**, 1512–1523 (2017).

47. Bass, David and Boenigk, J., Bass, D., Boenigk, J. & Bass, David and Boenigk, J. Everything is everywhere: A twenty-first century de-/reconstruction with respect to protists. in *Biogeography of Microscopic Organisms* (ed. Diego Fontaneto) 88–110 (Cambridge University Press, 2011). doi:10.1017/CBO9780511974878.007

48. Pedros-Alio, C. The Rare Bacterial Biosphere. *Annual Review of Marine Science* **4**, 449–466 (2012).

49. Galand, P. E., Casamayor, E. O., Kirchman, D. L. & Lovejoy, C. Ecology of the rare microbial biosphere of the Arctic Ocean. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 22427–22432 (2009).

50. Mangot, J. *et al.* Short-term dynamics of diversity patterns: evidence of continual reassembly within lacustrine small eukaryotes. *Environmental Microbiology* **15**, 1745–1758 (2013).

51. Nolte, V. *et al.* Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity. *Molecular Ecology* **19**, 2908–2915 (2010).

52. Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12**, 118–123 (2010).

53. Logares, R., Mangot, J. F. & Massana, R. Rarity in aquatic microbes: Placing protists on the map. *Research in Microbiology* **166**, 831–841 (2015).

54. Pedros-Alio, C. Marine microbial diversity: can it be determined? *Trends in Microbiology* **14**, 257–263 (2006).

55. Caron, D. A. & Countway, P. D. Hypotheses on the role of the protistan rare

biosphere in a changing world. *Aquatic Microbial Ecology* **57**, 227–238 (2009).

56. Logares, R. *et al.* Patterns of rare and abundant marine microbial eukaryotes. *Current Biology* **24**, 813–821 (2014).

57. Lovejoy, C. Picoplankton diversity in the Arctic Ocean and surrounding seas. *Marine Biodiversity* **41**, 5–12 (2011).

58. Needham, D. M., Sachdeva, R. & Fuhrman, J. A. Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *The ISME Journal* **11**, 1614–1629 (2017).

59. CRICK, F. Central Dogma of Molecular Biology. *Nature* **227**, 561 (1970).

60. Alberts, B. *et al.* From DNA to Protein: How Cells Read the Genome. in *Essential Cell Biology* (ed. Morales, M.) 248–249 (New York: Garland Science, 2010).

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

62. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463–5467 (1977).

63. Curiocity. Sanger sequencing gel electrophoretogram. (2019). Available at: http://explorecuriocity.org/Explore/ArticleId/2027/sanger-sequencing-2027.aspx. (Accessed: 21st January 2019)

64. Siqueira, J. F., Fouad, A. F. & Rôças, I. N. Pyrosequencing as a tool for better understanding of human microbiomes. *Journal of Oral Microbiology* **4**, 10743 (2012).

65. Quail, M. A. *et al.* A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

66. Goodwin, S., Mcpherson, J. D. & Mccombie, W. R. Coming of age: ten years of next- generation sequencing technologies. *Nature Reviews Genetics* **17**, 333–351 (2016).

67. Illumina Inc. *Quality Scores for Next-Generation Sequencing: Assessing*

sequencing accuracy using Phred quality scoring. (2000).

68. Pedrós-Alió, C., Potvin, M. & Lovejoy, C. Diversity of planktonic microorganisms in the Arctic Ocean. *Progress in Oceanography* **139**, 233–243 (2015).

69. Waleron, M., Waleron, K., Vincent, W. F. & Wilmotte, A. Allochthonous inputs of riverine picocyanobacteria to coastal waters in the Arctic Ocean. *FEMS Microbiology Ecology* **59**, 356–365 (2007).

70. Galand, P. E., Potvin, M., Casamayor, E. O. & Lovejoy, C. Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *The ISME Journal* **4**, 564–576 (2009).

71. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling Expedition : Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biology* **5**, 398–431 (2007).

72. Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling Expedition : Expanding the Universe of Protein Families. *PLoS Biology* **5**, e16 (2007).

73. Engelen, S. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605-1-1261605–11 (2015).

74. Caputi, L. *et al.* Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. *Global Biogeochemical Cycles* **33**, 1–29 (2019).

75. Ser-giacomi, E. *et al.* Ubiquitous abundance distribution of non-dominant plankton across the global ocean. *Nature Ecology & Evolution* **2**, 1243–1249 (2018).

76. Leblanc, K. *et al.* Nanoplanktonic diatoms are globally overlooked but play a role in spring blooms and carbon export. *Nature Communications* **9**, 1–12 (2018).

77. Falkowski, P. G., Fenchel, T. & Delong, E. F. The Microbial Engines That Drive Earth 's Biogeochemical Cycles. *Science* **320**, 1034–1039 (2008).

78. Field, C. B., Behrenfeld, M. J. & Randerson, J. T. Primary Production of the Biosphere : Integrating Terrestrial and Oceanic Components. *Science* **281**,

237–241 (1998).

79. Tremblay, J.-éric *et al.* Global and regional drivers of nutrient supply , primary production and CO 2 drawdown in the changing Arctic Ocean. *Progress in Oceanography* **139**, 171–196 (2015).

80. Charlson, R. J., Lovelockt, J. E., Andreae, M. & Warren, S. G. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).

81. Passow, U. & Carlson, C. A. The biological pump in a high CO2 world. *Marine Ecology Progress Series* **470**, 249–271 (2012).

82. Supraha, L. Haptophyte Diversity and Vertical Distribution Explored by 18S and 28S Ribosomal RNA Gene Metabarcoding and Scanning Electron Microscopy. *Eukaryotic Microbiology* **64**, 514–532 (2017).

83. Liu, H. *et al.* Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *PNAS* **106**, 12803–12808 (2009).

84. Verity, P. G. *et al.* Current understanding of Phaeocystis ecology and biogeochemistry, and perspectives for future research. *Biogeochemistry* **83**, 311–330 (2007).

85. Balestreri, C., Rickaby, R. E. M., Brownlee, C. & Schroeder, D. C. Genetic variability of the microalga Emiliania huxleyi (Haptophyta): a temporal and geographical study. *PhD thesis* (University of Oxford, 2016).

86. Jin, X. *et al.* Diagnosing the contributions of phytoplankton functional groups to the production and export of particulate organic carbon, CaCO3, and opal from global nutrient and alkalinity distributions. *Global Biogeochemical Cycles* **20**, 1–17 (2006).

87. Unrein, F., Gasol, J. M., Not, F., Forn, I. & Massana, R. Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. *The ISME Journal* **8**, 164–176 (2014).

88. Medlin, L. & Zingone, A. A taxonomic review of the genus Phaeocystis. *Biogeochemistry* **83**, 3–18 (2007).

89. Wang, S., Elliott, S., Maltrud, M. & Cameron-Smith, P. Influence of explicit

Phaeocystis parameterizations on the global distribution of marine dimethyl sulphide. *Journal of Geophysical Research: Biogeosciences* **120**, 2158–2177 (2015).

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

91. Mann, D. G. & Vanormelingen, P. An Inordinate Fondness? The Number, Distributions, and Origins of Diatom Species. *Eukaryotic Microbiology* **60**, 414–420 (2013).

92. Guiry, M. D. How many species of algae are there? *Journal of Phycology* **1063**, 1057–1063 (2012).

93. Villanova, V. *et al.* Investigating mixotrophic metabolism in the model diatom Phaeodactylum tricornutum. *Philosophical Transactions of the Royal Society B* **372**, 20160404 (2017).

94. Caron, D. A. Mixotrophy stirs up our understanding of marine food webs. *PNAS* **113**, 2806–2808 (2016).

95. Hasle, G. R., Syvertsen, E. E., Steidinger, K. A. & Tangen, K. Marine diatoms. in *Identifying Marine Diatoms and Dinoflagellates* (ed. Carmelo, T. R.) 5–385 (Academic Press, 1996).

96. Malviya, S. *et al.* Insights into global diatom distribution and diversity in the world's ocean. *Proceedings of the National Academy of Sciences* **113**, E1516–E1525 (2016).

97. Mock, T. *et al.* Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus. *Nature* **541**, 536–540 (2017).

98. Litchman, E. *et al.* Global biogeochemical impacts of phytoplankton: A trait-based perspective. *Journal of Ecology* **103**, 1384–1396 (2015).

99. Cohen, N. R. *et al.* Iron storage capacities and associated ferritin gene expression among marine diatoms. *Limnology and Oceanography* **63**, 1677–1691 (2018).

100. Natf, S., Strzepek, R. F. & Harrison, P. J. Photosynthetic architecture differs in coastal and oceanic diatoms. *Letters to Nature* **403**, 689–692 (2004).

101. Lommer, M. *et al.* Recent transfer of an iron-regulated gene from the plastid to the nuclear genome in an oceanic diatom adapted to chronic iron limitation. *BMC Genomics* **718**, 1–13 (2010).

102. Sugie, K. & Suzuki, K. Size of Dominant Diatom Species Can Alter Their Evenness. *PLoS ONE* **10**, 1–12 (2015).

103. Taylor, F. J. R., Hoppenrath, M. & Saldarriaga, J. F. Dinoflagellate diversity and distribution. in *Protist Diversity and Geographical Distribution* (eds. Foissner, W. & Hawksworth, D. L.) 173–184 (Springer Netherlands, 2009). doi:10.1007/978-90-481-2801-3_13

104. Gómez, F. A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates (Dinoflagellata, Alveolata). *Systematics and Biodiversity* **10**, 267–275 (2012).

105. Bescot, N. Le *et al.* Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environmental Microbiology* **2**, 609–626 (2016).

106. Trainer, V. L. *et al.* Pseudo-nitzschia physiological ecology, phylogeny, toxicity, monitoring and impacts on ecosystem health. *Harmful Algae* **14**, 271–300 (2012).

107. Anderson, D. M. *et al.* The globally distributed genus Alexandrium : Multifaceted roles in marine ecosystems and impacts on human health. *Harmful Algae* **14**, 10–35 (2012).

108. Zingone, A., SarnoRaffaele, D. & Marino, S. The importance and distinctiveness of small- sized phytoplankton in the Magellan Straits. *Polar Biology* **34**, 1269–1284 (2010).

109. Skovgaard, A., Karpov, S. A. & Guillou, L. The parasitic dinoflagellates Blastodinium spp. Inhabiting the gut of marine, Planktonic copepods: Morphology, ecology, and unrecognized species diversity. *Frontiers in Microbiology* **3**, 1–22 (2012).

110. Boetius, A., Anesio, A. M., Deming, J. W., Mikucki, J. & Rapp, J. Z. Microbial ecology of the cryosphere : sea ice and glacial habitats. *Nature Reviews Microbiology* **13**, 677–690 (2014).

111. Partensky, F., Blanchot, J. & Vaulot, D. Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters a review. *Bulletin de l'Institut Océanographique - Special issue: Marine cyanobacteria* **19**, 457–476 (1999).

112. Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Global distributions of marine Cyanobacteria* **110**, 9824–9829 (2013).

113. Partensky, F. & Hess, W. R. Prochlorococcus , a Marine Photosynthetic Prokaryote of Global Significance. *American Society for Microbiology* **63**, 106–127 (1999).

114. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology* **13**, 13–27 (2014).

115. García-ferna, J. M., Marsac, N. T. De & Diez, J. Streamlined Regulation and Gene Loss as Adaptive Mechanisms in Prochlorococcus for Optimized Nitrogen Utilization in Oligotrophic Environments. *Microbiology and Molecular Biology Reviews* **68**, 630–638 (2004).

116. Haverkamp, T. H. A., Schouten, D., Doeleman, M. & Wollenzien, U. Colorful microdiversity of Synechococcus strains (picocyanobacteria) isolated from the Baltic Sea. *The ISME Journal* **3**, 397–408 (2009).

117. Connon, S. A. & Vergin, K. L. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Letters to Nature* **418**, 630–633 (2002).

118. Giovannoni, S. J. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Annual Review of Marine Science* **9**, 231–255 (2017).

119. Brown, M. V *et al.* Global biogeography of SAR11 marine bacteria. *Molecular Systems Biology* **8**, 1–13 (2012).

120. Malmstrom, R. R., Kiene, R. P., Cottrell, M. T. & Kirchman, D. L. Contribution of SAR11 Bacteria to Dissolved Dimethylsulfoniopropionate and Amino Acid Uptake in the North Atlantic Ocean. *Applied and Environmental Microbiology* **70**, 4129–4135 (2004).

121. Giovannoni, S. J. *et al.* Genome Streamlining in a Cosmopolitan Oceanic

Bacterium. *Science* **309**, 1242–1246 (2005).

122. Laghdass, M., Catala, P., Caparros, J., Oriol, L. & Lebaron, P. High Contribution of SAR11 to Microbial Activity in the North West Mediterranean Sea. *Microbial Ecology* **63**, 324–333 (2012).

123. Sun, J. *et al.* One Carbon Metabolism in SAR11 Pelagic Marine Bacteria. *PLoS ONE* **6**, e23973 (2011).

124. Smith, D. P. & Giovannoni, S. J. The presence of the glycolysis operon in SAR11 genomes is positively correlated with. *Environmental Microbiology* **12**, 490–500 (2010).

125. Hegseth, E. N. & Sundfjord, A. Intrusion and blooming of Atlantic phytoplankton species in the high Arctic. *Journal of Marine Systems* **74**, 108–119 (2008).

126. Barton, A. D., Irwin, A. J., Finkel, Z. V & Stock, C. A. Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *PNAS* **113**, 2964–2969 (2016).

127. Matrai, P. a. *et al.* Synthesis of primary production in the Arctic Ocean: I. Surface waters, 1954-2007. *Progress in Oceanography* **110**, 93–106 (2013).

128. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).

129. Boeuf, D., Humily, F. & Jeanthon, C. Diversity of arctic pelagic bacteria with an emphasis on photoheterotrophs: A review. *Biogeosciences* **11**, 3309–3322 (2014).

130. Poulin, M. *et al.* The pan-Arctic biodiversity of marine pelagic and sea-ice unicellular eukaryotes: a first-attempt assessment. *Marine Biodiversity* **41**, 13–28 (2011).

131. Lovejoy, C. Changing views of Arctic protists (marine microbial eukaryotes) in a changing Arctic. *Acta Protozoologica* **53**, 91–100 (2014).

132. Malmstrom, R. R., Straza, T. R. A., Cottrell, M. T. & Kirchman, D. L. Diversity, abundance, and biomass production of bacterial groups in the western Arctic Ocean. *Aquatic Microbial Ecology* **47**, 45–55 (2007).

133. Boeuf, D., Humily, F. & Jeanthon, C. Diversity of Arctic Pelagic Prokaryotes with an emphasis on photoheterotrophic bacteria: a review. *Biogeosciences Discussions* **11**, 2419–2455 (2014).

134. Wassmann, P., Duarte, C. M., Agusti, S. & Serj, M. K. Footprints of climate change in the Arctic marine ecosystem. *Global Change Biology* **17**, 1235–1249 (2011).

135. Rossi-tamisier, M., Benamar, S. & Raoult, D. Cautionary tale of using 16S rRNA gene sequence similarity values in identification of human-associated bacterial species. *International Journal of Systematic and Evolutionary Microbiology* **65**, 1929–1934 (2015).

136. Grattepanche, J.-D., Santoferrara, L. F., McManus, G. B. & Katz, L. a. Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends in Microbiology* **22**, 432–437 (2014).

137. Zhang, F., He, J., Lin, L. & Jin, H. Dominance of picophytoplankton in the newly open surface water of the central Arctic Ocean. *Polar Biology* **38**, 1081–1089 (2015).

138. Gast, R. J., Dennett, M. R. & Caron, D. a. Characterization of protistan assemblages in the Ross Sea, Antarctica, by denaturing gradient gel electrophoresis. *Applied and environmental microbiology* **70**, 2028–37 (2004).

139. Kircher, M. & Kelso, J. High-throughput DNA sequencing - Concepts and limitations. *BioEssays* **32**, 524–536 (2010).

140. Monier, A. *et al.* Upper Arctic Ocean water masses harbor distinct communities of heterotrophic flagellates. *Biogeosciences* **10**, 4273–4286 (2013).

141. Wolf, C., Metfies, K., Kilias, E. S. & Eva-maria, N. Picoeukaryote Plankton Composition off West Spitsbergen at the Entrance to the Arctic Ocean. *Eukaryotic Microbiology* **61**, 569–579 (2014).

142. Zhan, A., Xiong, W., He, S. & MacIsaac, H. J. Influence of artifact removal on rare species recovery in natural complex communities using high-throughput sequencing. *PLoS ONE* **9**, e96928 (2014).

143. Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology Microbiology* **13**, 217–229 (2015).

144. Lebret, K. *et al.* Choice of molecular barcode will affect species prevalence but not bacterial community composition. *Marine Genomics* **29**, 39–43 (2016).

145. Dunthorn, M., Klier, J., Bunge, J. & Stoeck, T. Comparing the Hyper-Variable V4 and V9 Regions of the Small Subunit rDNA for Assessment of Ciliate Environmental Diversity. *The Journal of Eukaryotic Microbiology* **59**, 185–187 (2012).

146. Pawlowicz, R. M_Map: A mapping package for MATLAB. (2018).

147. Ducet, N., Traon, P. Y. Le & Reverdin, G. Global high-resolution mapping of ocean circulation from TOPEX/Poseidon and ERS-1 and -2. *Journal of Geophysical Research* **105**, 19477–19498 (2000).

148. Donlon, C. J. *et al.* The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sensing of Environment* **116**, 140–158 (2012).

149. Kirkwood, D. S. Nutrients: a practical note on their determination in seawater. in *ICES* (1996).

150. Lovejoy, C., Massana, R. & Pedro, C. Diversity and Distribution of Marine Microbial Eukaryotes in the Arctic Ocean and Adjacent Seas. *Applied and Environmental Microbiology* **72**, 3085–3095 (2006).

151. Hadziavdic, K. *et al.* Characterization of the 18S rRNA Gene for Designing Universal Eukaryote Specific Primers. *PLoS ONE* **9**, e87624 (2014).

152. Flaviani, F. *et al.* A pelagic microbiome (Viruses to protists) from a small cup of seawater. *Viruses* **9**, (2017).

153. Hii, Y. S., Law, A. T., Shazili, N. a M., Abdul-Rashid, M. K. & Lee, C. W. Biodegradation of Tapis blended crude oil in marine sediment by a consortium of symbiotic bacteria. *International Biodeterioration and Biodegradation* **63**, 142–150 (2009).

154. Yang, C. *et al.* Illumina sequencing-based analysis of free-living bacterial

community dynamics during an Akashiwo sanguine bloom in Xiamen sea, China. *Scientific Reports* **5**, 1–11 (2015).

155. Sergio, B. *et al.* Transcriptome analyses to investigate symbiotic relationships between marine protists. *Frontiers in Microbiology* **6**, 1–14 (2015).

156. Lampe, R. H. *et al.* Divergent gene expression among phytoplankton taxa in response to upwelling. *Environmental Microbiology* **20**, 3069–3082 (2018).

157. Field, D., Tiwari, B., Booth, T., Houten, S., Swan, D., Bertrand, N. and Thurston, M. Open Software for biologists: from famine to feast. *Nature Biotechnology* **24**, 801–803 (2006).

158. Salazar, G. *et al.* Global diversity and biogeography of deep-sea pelagic prokaryotes. *The ISME Journal* **10**, 596–608 (2016).

159. Andrews, S. FastQC: A quality control tool for high throughput sequence data. (2014).

160. Marcel, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. (2011).

161. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: A fast and accurate Illumina Paired-End read merger. *Bioinformatics* **30**, 614–620 (2014).

162. R Development Core Team. R: a language and environment for statistical computing. (2018).

163. Stackebrandt, Erko. Ebers, J. Taxonomic parameters revisited: tarnished gold standards. *Microbiology today* **8**, 6–9 (2006).

164. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research* **41**, 590–596 (2013).

165. Mahé, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014).

166. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).

167. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 1–18 (2017).

168. Oksanen, J. *et al.* vegan: Community Ecology Package. (2019).

169. Chao, A. & Colwell, R. K. Rarefaction and extrapolation with Hill numbers : A framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84**, 45–67 (2014).

170. Cori, A., Ferguson, N. M., Fraser, C. & Cauchemez, S. A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics. *American Journal of Epidemiology* **178**, 1505–1512 (2013).

171. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385 (2011).

172. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35 (2011).

173. R. Warnes, G. *et al.* gplots: Various R Programming Tools for Plotting Data. (2019).

174. Legendre, P. & Gallagher, E. D. Ecologically meaningful transformations for ordination of species data. *Oecologia* **129**, 271–280 (2001).

175. Burland, T. G. DNASTAR's Lasergene sequence analysis software. *Methods in Molecular Biology* **132**, 71–91 (2000).

176. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome biology* **17**, 1–19 (2016).

177. Robinson, M., McCarthy, D. & Smyth, G. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

178. Yunshun, C., McCarthy, D., Ritchie, M., Robinson, M. & Gordon, S. edgeR: differential expression analysis of digital gene expression data User's Guide. (2018).

179. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**, 29–34 (1999).

180. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**, 182–185 (2007).

181. Irwin, A. J., Finkel, Z. V, Müller-karger, F. E. & Troccoli, L. Phytoplankton adapt to changing ocean environments. *PNAS* **112**, 5762–5766 (2015).

182. Marshall, J. *et al.* The ocean's role in polar climate change: asymmetric Arctic and Antarctic responses to greenhouse gas and ozone forcing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **372**, 20130040 (2014).

183. Kashiwase, H., Ohshima, K. I., Nihashi, S. & Eicken, H. Evidence for ice-ocean albedo feedback in the Arctic Ocean shifting to a seasonal ice zone. *Scientific Reports* **7**, 1–10 (2017).

184. Chylek, P., Folland, C. K., Lesins, G., Dubey, M. K. & Wang, M. Arctic air temperature change amplification and the Atlantic Multidecadal Oscillation. *Geophysical Research Letters* **36**, 2–6 (2009).

185. Abe, M., Nozawa, T., Ogura, T. & Takata, K. Effect of retreating sea ice on Arctic cloud cover in simulated recent global warming. *Atmospheric Chemistry and Physics* **16**, 14343–14356 (2016).

186. Stuecker, M. F., Bitz, C. M., Armour, K. C., Proistosescu, C. & Kang, S. M. Polar amplification dominated by local forcing and feedbacks. *Nature Climate Change* **8**, 1076–1082 (2018).

187. Bintanja, R. & Andry, O. Towards a rain-dominated Arctic. *Nature Climate Change* **7**, 263–268 (2017).

188. Haine, T. W. N. *et al.* Arctic freshwater export: Status, mechanisms, and prospects. *Global and Planetary Change* **125**, 13–35 (2015).

189. Kwok, R. Arctic sea ice thickness, volume, and multiyear ice coverage: losses and and coupled variability (1958 – 2018). *Environmental Pollution* **13**, 105005 (2018).

190. Polyakov, I. V., Walsh, J. E. & Kwok, R. Recent Changes of Arctic Multiyear Sea Ice Coverage and the Likely Causes. *Bulletin of the American Meteorological Society* **93**, 145–151 (2012).

191. Screen, J. A. & Simmonds, I. The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature* **464**, 1334–1337 (2010).

192. Comiso, J. C. Large Decadal Decline of the Arctic Multiyear Ice Cover. *Journal of Climate* **25**, 1176–1193 (2012).

193. Overland, J. E. & Wang, M. When will the summer Arctic be nearly sea ice free? *Geophysical Research Letters* **40**, 2097–2101 (2013).

194. Arrigo, K. R. & Dijken, G. L. Van. Continued increases in Arctic Ocean primary production. *Progress in Oceanography* **136**, 60–70 (2015).

195. Arrigo, K. R. & Van Dijken, G. L. Secular trends in Arctic Ocean net primary production. *Journal of Geophysical Research: Oceans* **116**, 1–15 (2011).

196. Williams, W. J. & Bacon, S. Wind-driven mixing at intermediate depths in an ice-free Arctic Ocean. *Geophysical Research Letters* **43**, 9749–9756 (2016).

197. Williams, T. J. *et al.* A metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula coastal surface waters. *The ISME journal* **6**, 1883–900 (2012).

198. Yool, A., Popova, E. E. & Coward, A. C. Future change in ocean productivity: Is the Arctic the new Atlantic. *Journal of Geophysical Research: Oceans* **120**, 7771–7790 (2015).

199. Lind, S., Ingvaldsen, R. B. & Furevik, T. Arctic warming hotspot in the northern Barents Sea linked to declining sea-ice import. *Nature Climate Change* **8**, 634–639 (2018).

200. Loeng, H. Features of the physical oceanographic conditions of the Barents Sea. *Polar Research* **10**, 5–18 (1991).

201. Lind, S., Ingvaldsen, R. B. & Furevik, T. Arctic layer salinity controls heat loss from deep Atlantic layer in seasonally ice-covered areas of the Barents Sea. *Geophysical Research Letters* **43**, 5233–5242 (2016).

202. Dokken, T. M., Nisancioglu, K. H., Li, C., Battisti, D. S. & Kissel, C. Dansgaard-Oeschger cycles Interactions between ocean and sea ice intrinsic to the Nordic seas. *Paleoceanography and Paleoclimatology* **28**, 491–502 (2013).

203. Li, W. K. W., McLaughlin, F. a, Lovejoy, C. & Carmack, E. C. Smallest algae thrive as the Arctic Ocean freshens. *Science* **326**, 539 (2009).

204. Dalpadado, P. *et al.* Productivity in the Barents Sea - Response to Recent Climate Variability. *PLoS ONE* **9**, e95273 (2014).

205. Beaugrand, G. Reorganization of North Atlantic Marine Copepod Biodiversity and Climate. *Science* **296**, 1692–1694 (2002).

206. Neukermans, G., Oziel, L. & Babin, M. Increased intrusion of warming Atlantic water leads to rapid expansion of temperate phytoplankton in the Arctic. *Global Change Biology* **24**, 2545–2553 (2018).

207. Bunse, C. & Pinhassi, J. Marine Bacterioplankton Seasonal Succession Dynamics. *Trends in Microbiology* **25**, 494–505 (2017).

208. Bano, N. & Hollibaugh, J. T. Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. *Applied and Environmental Microbiology* **68**, 505–518 (2002).

209. Ramette, A. & Tiedje, J. M. Biogeography: An emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. *Microbial Ecology* **53**, 197–207 (2007).

210. Balmonte, J. P., Teske, A. & Arnosti, C. Structure and function of high Arctic pelagic, particle-associated and benthic bacterial communities. *Environmental Microbiology* **20**, 2941–2954 (2018).

211. Sul, W. J., Oliver, T. A., Ducklow, H. W., Amaral-Zettler, L. A. & Sogin, M. L. Marine bacteria exhibit a bipolar distribution. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 2342–2347 (2013).

212. Sørensen, N., Daugbjerg, N. & Gabrielsen, T. M. Molecular diversity and temporal variation of picoeukaryotes in two Arctic fjords, Svalbard. *Polar Biology* **35**, 519–533 (2012).

213. Mock, T. & Thomas, D. N. Recent advances in sea-ice microbiology. *Environmental Microbiology* **7**, 605–619 (2005).

214. Sheik, C. S., Jain, S. & Dick, G. J. Metabolic flexibility of enigmatic SAR324

revealed through metagenomics and metatranscriptomics. *Environmental microbiology* **16**, 304–317 (2014).

215. Wright, J. J. *et al.* Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. *The ISME journal* **8**, 455–468 (2014).

216. Lücker, S., Nowka, B., Rattei, T., Spieck, E. & Daims, H. The genome of Nitrospina gracilis illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Frontiers in Microbiology* **4**, 1–19 (2013).

217. Comeau, A. M. *et al.* Arctic ocean microbial community structure before and after the 2007 record sea ice minimum. *PLoS ONE* **6**, e27492 (2011).

218. Lucas, J. *et al.* Annual dynamics of North Sea bacterioplankton: seasonal variability superimposes short-term variation. *FEMS Microbiology Ecology* **91**, 1–11 (2015).

219. Wilson, B. *et al.* Changes in Marine Prokaryote Composition with Season and Depth Over an Arctic Polar Year. *Frontiers in Marine Science* **4**, 1–17 (2017).

220. Sherr, E. B., Sherr, B. F., Wheeler, P. A. & Thompson, K. Temporal and spatial variation in stocks of autotrophic and heterotrophic microbes in the upper water column of the central Arctic Ocean. *Deep-Sea Research Part I: Oceanographic Research Papers* **50**, 557–571 (2003).

221. Zhang, J. *et al.* Modeling the impact of declining sea ice on the Arctic marine planktonic ecosystem. *Journal of Geophysical Research* **115**, 1–24 (2010).

222. Mouillot, D. *et al.* Rare Species Support Vulnerable Functions in High-Diversity Ecosystems. *PLoS Biology* **11**, e1001569. (2013).

223. Pruesse, E. *et al.* SILVA : a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188–7196 (2007).

224. Hill, M. O. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology1* **54**, 427–432 (1973).

225. Kim, B. *et al.* Deciphering Diversity Indices for a Better Understanding of Microbial Communities. *Journal of microbial biotechnology* **27**, 2089–2093 (2017).

226. Schloss, P. D., Girard, R. A., Martin, T., Edwards, J. & Thrash, J. C. Status of the Archaeal and Bacterial Census: an Update. *American Society for Microbiology* **7**, e00201-16 (2016).

227. Herlemann, D. P. R. *et al.* Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal* **5**, 1571–1579 (2011).

228. Han, D. *et al.* Bacterial communities of surface mixed layer in the pacific sector of the western Arctic Ocean during sea-ice melting. *PLoS ONE* **9**, e86887 (2014).

229. Bouvier, T. C. & Giorgio, P. A. Compositional changes in free-living bacterial communities along a salinity gradient in two temperate estuaries. *Limnology and Oceanography* **47**, 453–470 (2002).

230. Schmidt, S. K. *et al.* Biogeochemical consequences of rapid microbial turnover and seasonal succession in soil. *Ecology* **88**, 1379–1385 (2007).

231. Pernthaler, J. Predation on prokaryotes in the water column and its ecological implications. *Nature Reviews Microbiology* **3**, 537–546 (2005).

232. Aanderud, Z. T., Jones, S. E., Fierer, N. & Lennon, J. T. Resuscitation of the rare biosphere contributes to pulses of ecosystem activity. *Frontiers in Microbiology* **6**, 1–11 (2015).

233. Lennon, J. T. & Jones, S. E. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nature Reviews Microbiology* **9**, 119–130 (2011).

234. Jousset, A. *et al.* Where less may be more : how the rare biosphere pulls ecosystems strings. *The ISME Journal* **11**, 853–862 (2017).

235. Bland J. Finlay, G. F. E. Ubiquitous Dispersal of Free-Living Microorganisms. in *Microbial Diversity and Bioprospecting.* (ed. Bull, A.) 216–224 (ASM Press, 2004).

236. Hollibaugh, J., Lovejoy, C. & Murray, A. Microbiology in Polar Oceans. *Oceanography* **20**, 140–145 (2007).

237. Gosink, J. J., Woese, C. R. & Staley, J. T. Polaribacter gen. nov., with three

new species, P. irgensii sp. nov., P. franzmannii sp. nov. and P. filamentus sp. nov., gas vacuolate polar marine bacteria of the Cytophaga-Flavobacterium-Bacteroides group and reclassification of ’Flectobacillus glomer. *International Journal of Systematic Bacteriology* **48**, 223–235 (1998).

238. Dauga, C. *et al.* Balneatrix alpica gen. nov., sp. nov., a bacterium associated with pneumonia and meningitis in a spa therapy centre. *Research in Microbiology* **144**, 35–46 (1993).

239. Bergmann, G. T. *et al.* The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biology Biochemistry* **43**, 1450–1455 (2012).

240. Wilkinson, D. M., Koumoutsaris, S., Mitchell, E. A. D. & Bey, I. Modelling the effect of size on the aerial dispersal of microorganisms. *Journal of Biogeography* **39**, 89–97 (2012).

241. Amin, S. A., Parker, M. S. & Armbrust, E. V. Interactions between Diatoms and Bacteria. *Microbiology and molecular biology reviews* **76**, 667–684 (2012).

242. Freitas, S. *et al.* Global distribution and diversity of marine Verrucomicrobia. *The ISME Journal* **6**, 1499–1505 (2012).

243. Yoon, J., Kang, S. & Oh, T. Polaribacter dokdonensis sp . nov ., isolated from seawater. *International Journal of Systematic and Evolutionary Microbiology* **56**, 1251–1255 (2006).

244. Fukui, Y. *et al.* Polaribacter porphyrae sp . nov ., isolated from the red alga Porphyra yezoensis , and emended descriptions of the genus Polaribacter and two Polaribacter species. *International Journal of Systematic and Evolutionary Microbiology* **63**, 1665–1672 (2013).

245. Abell, G. C. J. & Bowman, J. P. Ecological and biogeographic relationships of class Flavobacteria in the Southern Ocean. *FEMS Microbiology Ecology* **51**, 265–277 (2005).

246. Choi, T., Lee, H. K., Lee, K. & Cho, J. Ulvibacter antarcticus sp . nov ., isolated from Antarctic coastal seawater. *International Journal of Systematic and Evolutionary Microbiology* **57**, 2922–2925 (2017).

247. Yang, C. *et al.* Illumina sequencing-based analysis of free-living bacterial community dynamics during an Akashiwo sanguine bloom in Xiamen sea, China. *Scientific reports* **5**, 1–11 (2015).

248. Biers, E. J., Sun, S. & Howard, E. C. Prokaryotic genomes and diversity in surface ocean waters: Interrogating the global ocean sampling metagenome. *Applied and Environmental Microbiology* **75**, 2221–2229 (2009).

249. Connon, S. A. *et al.* SAR11 clade dominates ocean surface bacterioplankton communities. *Letters to Nature* **420**, 806–810 (2002).

250. Carlson, C. A. *et al.* Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *The ISME Journal* **3**, 283–295 (2009).

251. Heinrich, F., Shalchian-tabrizi, K., Bertilsson, S., Logares, R. & Bra, J. Infrequent Transitions between Saline and Fresh Waters in One of the Most Abundant Microbial Lineages (SAR11). *Molecular Biology Evolution* **27**, 347–357 (2017).

252. Voget, S. *et al.* Adaptation of an abundant Roseobacter RCA organism to pelagic systems revealed by genomic and transcriptomic analyses. *The ISME Journal* **9**, 371–384 (2014).

253. Choi, D. H., Park, K., An, S. M., Lee, K. & Cho, J. Pyrosequencing Revealed SAR116 Clade as Dominant dddP -Containing Bacteria in Oligotrophic NW Pacific Ocean. *PLoS ONE* **10**, 1–13 (2015).

254. *Bergey's Manual of Systematic Bacteriology Volume 2: The Proteobacteria, Part B: The Gammaproteobacteria*. (Springer US, 2005). doi:10.1007/0-387-28022-7

255. Dupont, C. L. *et al.* Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *The ISME journal* **6**, 1186–99 (2012).

256. Lucas, J. *et al.* Short-term dynamics of north sea bacterioplankton-dissolved organic matter coherence on molecular level. *Frontiers in Microbiology* **7**, 1–14 (2016).

257. Zhang, Y., Sun, Y., Jiao, N., Stepanauskas, R. & Luo, H. Ecological genomics of the uncultivated marine Roseobacter lineage CHAB-I-5. *Applied*

and Environmental Microbiology **82**, 2100–2111 (2016).

258. Bachy, C., López-García, P., Vereshchaka, A. & Moreira, D. Diversity and vertical distribution of microbial eukaryotes in the snow, sea ice and seawater near the North Pole at the end of the polar night. *Frontiers in Microbiology* **2**, 1–12 (2011).

259. Kilias, E., Kattner, G., Wolf, C., Frickenhaus, S. & Metfies, K. A molecular survey of protist diversity through the central Arctic Ocean. *Polar Biology* **37**, 1271–1287 (2014).

260. Richardson, T. L. & Jackson, G. A. Small Phytoplankton and Carbon Export from the Surface Ocean. *Science* **315**, 838–841 (2007).

261. Lovejoy, C. *et al.* Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in arctic seas. *Journal of Phycology* **43**, 78–89 (2007).

262. Marquardt, M., Vader, A., Stübner, E. I., Reigstad, M. & Gabrielsen, T. M. Strong Seasonality of Marine Microbial Eukaryotes in a High-Arctic Fjord (Isfjorden, in West Spitsbergen, Norway). *Applied and Environmental Microbiology* **82**, 1868–1880 (2016).

263. Vader, A., Marquardt, M., Meshram, A. & Gabrielsen, T. M. Key Arctic phototrophs are widespread in the polar night. *Polar Biology* **38**, 13–21 (2014).

264. Mckie-krisberg, Z. M. & Sanders, R. W. Phagotrophy by the picoeukaryotic green alga Micromonas : implications for Arctic Oceans. *The ISME journal* **8**, 1953–1961 (2014).

265. Wolf, C., Kilias, E. & Metfies, K. Protists in the polar regions: comparing occurrence in the Arctic and Southern oceans using pyrosequencing. *Polar Research* **34**, 1–8 (2015).

266. Blais, M. *et al.* Contrasting interannual changes in phytoplankton productivity and community structure in the coastal Canadian Arctic Ocean. *Limnology and Oceanography* **62**, 2480–2497 (2017).

267. Olli, K. *et al.* The fate of production in the central Arctic Ocean - top-down regulation by zooplankton expatriates? *Progress in Oceanography* **72**, 84–113 (2007).

268. Eddie, B., Juhl, A., Krembs, C., Baysinger, C. & Neuer, S. Effect of environmental variables on eukaryotic microbial community structure of land-fast Arctic sea ice. *Environmental Microbiology* **12**, 797–809 (2010).

269. Belevich, T. A. *et al.* Metagenomic Analyses of White Sea Picoalgae: First Data. *Biochemistry* **80**, 1514–1521 (2015).

270. Moran, X. A. G. *et al.* Increasing importance of small phytoplankton in a warmer ocean. *Global Change Biology* **16**, 1137–1144 (2010).

271. Gilbert, J. A. *et al.* Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *PLoS ONE* **3**, e3042 (2008).

272. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data* **2**, 150023 (2015).

273. Hedlund, B. P. & Staley, J. T. Microbial Endemism and Biogeography. in *Microbial Diversity and Bioprospecting* (ed. Bull, A.) 225–231 (ASM Press, 2004).

274. O'Malley, M. A. The nineteenth century roots of 'everything is everywhere'. *Nature Reviews Microbiology* **5**, 647–651 (2007).

275. Grossmann, L. *et al.* Protistan community analysis: key findings of a large-scale molecular sampling. *The ISME Journal* **10**, 2269–2279 (2016).

276. Di Giuseppe, G., Dini, F., Vallesi, A. & Luporini, P. Genetic relationships in bipolar species of the protist ciliate, Euplotes. *Hydrobiologia* **761**, 71–83 (2015).

277. Rodríguez-Martínez, R. *et al.* Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. *The ISME Journal* **7**, 1531–1543 (2013).

278. Ly, J., Philippart, C. J. M. & Kromkamp, J. C. Phosphorus limitation during a phytoplankton spring bloom in the western Dutch Wadden Sea. *Journal of Sea Research* **88**, 109–120 (2014).

279. Elser, J. J. *et al.* Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecology*

Letters **10**, 1135–1142 (2007).

280. Dortch, Q. The interaction between ammonium and nitrate uptake in phytoplankton. *Marine Ecology Progress Series* **61**, 183–201 (1990).

281. Flöder, S., Jaschinski, S., Wells, G. & Burns, C. W. Dominance and compensatory growth in phytoplankton communities under salinity stress. *Journal of Experimental Marine Biology and Ecology* **395**, 223–231 (2010).

282. IPCC; Matthew Collins, R. K. *et al.* Long-term Climate Change: Projections, Commitments and Irreversibility. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* 1029–1136 (2013). doi:10.1017/CBO9781107415324.024

283. Kwon, J. E. *et al.* Newly discovered role of the heterotrophic nanoflagellate Katablepharis japonica, a predator of toxic or harmful dinoflagellates and raphidophytes. *Harmful Algae* **68**, 224–239 (2017).

284. Vørs, N. Ultrastructure and autecology of the marine, heterotrophic flagellate Leucocryptos marina (Braarud) Butcher 1967 (Katablepharidaceae/Kathablepharidae), with a discussion of the genera Leucocryptos and Katablepharis/Kathablepharis. *European Journal of Protistology* **28**, 369–389 (1992).

285. Jephcott, T. G. *et al.* Ecological impacts of parasitic chytrids, syndiniales and perkinsids on populations of marine photosynthetic dinoflagellates. *Fungal Ecology* **19**, 47–58 (2016).

286. Edwards, M. & Richardson, A. J. Impact of climate change on marine pelagic phenology and trophic mismatch. *Nature* **363**, 2002–2005 (2004).

287. Dylmer, C. V., Giraudeau, J., Hanquiez, V. & Husum, K. The coccolithophores Emiliania huxleyi and Coccolithus pelagicus: Extant populations from the Norwegian-Iceland Seas and Fram Strait. *Deep-Sea Research Part I: Oceanographic Research Papers* **98**, 1–9 (2015).

288. Jin, X., Gruber, N., Dunne, J. P., Sarmiento, J. L. & Armstrong, R. A. Diagnosing the contributions of phytoplankton functional groups to the production and export of particulate organic carbon, CaCO3, and opal from

global nutrient and alkalinity distributions. *Global Biogeochemical Cycles* **20**, 1–17 (2006).

289. Rocha, P. J. T. and C. L. D. La. The World Ocean Silica Cycle. *Annual review of Marine Sciene* **5**, 477–501 (2013).

290. Smetacek, V. Diatoms and the Ocean Carbon Cycle. *Protis* **150**, 25–32 (1999).

291. Spaulding, S. a *et al.* Diatoms as indicators of environmental change in Antarctic and subantarctic freshwaters. in *The Diatoms: Applications for the Environmental and Earth Sciences* (eds. Stoermer, J., Smol, P. & Eugene, F.) 267–283 (Cambridge University Press, 2010). doi:http://dx.doi.org/10.1017/CBO9780511763175.015

292. Riebesell, U. Effects of CO2 Enrichment on Marine Phytoplankton. *Journal of Oceanography* **60**, 719–729 (2004).

293. Koç, N., Miettinen, A. & Stickley, C. E. DIATOM RECORDS | North Atlantic and Arctic. *Encyclopedia of Quaternary Science* **1**, 562–570 (2013).

294. Winder, M. & Sommer, U. Phytoplankton response to a changing climate. *Hydrobiologia* **698**, 5–16 (2012).

295. Pabi, S., van Dijken, G. L. & Arrigo, K. R. Primary production in the Arctic Ocean, 1998–2006. *Journal of Geophysical Research* **113**, C08005 (2008).

296. Fuchs, H. L. & Franks, P. J. S. Plankton community properties determined by nutrients and size-selective feeding. *Marine Ecology Progress Series* **413**, 1–15 (2010).

297. Heyden, S. Von Der, Cavalier-smith, T. & Cavalier-smith, T. Culturing and environmental DNA sequencing uncover hidden kinetoplastid biodiversity and a major marine clade within ancestrally freshwater Neobodo designis. *International Journal of Systematic and Evolutionary Microbiology* **55**, 2605–2621 (2005).

298. Vannier, T. *et al.* Survey of the green picoalga Bathycoccus genomes in the global ocean. *Scientific Reports* **6**, 1–11 (2016).

299. Okolodkov, Y. B. & Okolodkov, Y. B. Species range types of recent marine

dinoflagellates recorded from the Arctic Species range types of recent marine dinoflagellates recorded from the Arctic. *Biogeography of Arctic Dinoflagellates* **38**, 162–169 (1999).

300. Botanik, S. First records of Amphidoma languida and Azadinium dexteroporum (Amphidomataceae , Dinophyceae) from the Irminger Sea off Iceland. *Marine Biodiversity Records* **8**, 1–11 (2015).

301. Staay, S. Y. M. Der *et al.* Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001).

302. Genes, D. N. A., Kress, W. J. & Erickson, D. L. DNA barcodes: Genes, genomics, and bioinformatics. *PNAS* **105**, 2761–2762 (2008).

303. Aguiar-pulido, V. *et al.* Approaches for Microbiome Analysis. *Evolutionary Bioinformatics* **12**, 5–16 (2016).

304. Toseland, A., Moxon, S., Mock, T. & Moulton, V. Metatranscriptomes from diverse microbial communities: assessment of data reduction techniques for rigorous annotation. *BMC Genomics* **15**, 1–7 (2014).

305. Kopecky, J. *et al.* Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *The ISME Journal* **6**, 248–258 (2012).

306. Pérez-cobas, A. E. *et al.* Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. *Gut Microbiota* **62**, 1591–1601 (2013).

307. Marchetti, A., Catlett, D., Hopkinson, B. M., Ellis, K. & Cassar, N. Marine diatom proteorhodopsins and their potential role in coping with low iron availability. *The ISME Journal* **9**, 2745–2748 (2015).

308. Gifford, S. M., Sharma, S., Rinta-kanto, J. M. & Moran, M. A. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. *The ISME Journal* **5**, 461–472 (2010).

309. Alexander, H. *et al.* Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *PNAS* **112**, 5972–5979 (2015).

310. Ottesen, E. A. *et al.* Pattern and synchrony of gene expression among sympatric marine microbial populations. *PNAS* **110**, 488–497 (2013).

311. Toseland, A. *et al.* The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nature Climate Change* **3**, 979–984 (2013).

312. Mangul, S. *et al.* Transcriptome assembly and quantification from Ion Torrent RNA-Seq data. *BMC genomics* **15**, S7 (2014).

313. Frias-lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *PNAS* **105**, 3805–3810 (2008).

314. Petrova, O. E., Garcia-alcalde, F., Zampaloni, C. & Sauer, K. Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Scientific Reports* **7**, 1–15 (2017).

315. Komine, Y., Kikis, E., Schuster, G. & Stern, D. Evidence for in vivo modulation of chloroplast RNA stability by 3' -UTR homopolymeric tails in Chlamydomonas reinhardtii. *PNAS* **99**, 4085–4090 (2002).

316. Pastor-fernández, I. *et al.* The tandemly repeated NTPase (NTPDase) from Neospora caninum is a canonical dense granule protein whose RNA expression , protein secretion and phosphorylation coincides with the tachyzoite egress. *Parasites & Vectors* **9**, 1–15 (2016).

317. Mayer, M. P. & Bukau, B. Cellular and Molecular Life Sciences Hsp70 chaperones : Cellular functions and molecular mechanism. *Cellular and Molecular Life Sciences* **62**, 670–684 (2005).

318. Itskovich, V. B., Shigarova, A. M., Glyzina, O. Y., Kaluzhnaya, O. V & Borovskii, G. B. Heat shock protein 70 (Hsp70) response to elevated temperatures in the endemic Baikal sponge Lubomirskia baicalensis. *Ecological Indicators* **88**, 1–7 (2018).

319. Schill, R. O., Steinbrück, G. H. B. & Köhler, H. Stress gene (hsp70) sequences and quantitative expression in Milnesium tardigradum (Tardigrada) during active and cryptobiotic stages. *The Journal of Experimental Biology* **207**, 1607–1613 (2004).

320. Consortium, T. U. UniProt : the universal protein knowledgebase. **45**, 158–169 (2018).

321. Eppard, M. & Rhiel, E. Investigations on Gene Copy Number , Introns and Chromosomal Arrangement of Genes Encoding the Fucoxanthin Chlorophyll a / c -Binding Proteins of the Centric Diatom Cyclotella cryptica. *Protist* **151**, 27–39 (2000).

322. Lefebvre, S. C. *et al.* Characterization And Expression Analysis Of The Lhcf Gene Family In Emiliania huxleyi (Haptophyta) Reveals Differential Responses To Light And CO2. *Journal of Phycology* **134**, 123–134 (2010).

323. Oeltjen, A., Krumbein, W. E. & Rhiel, E. Investigations on Transcript Sizes , Steady State mRNA Concentrations and Diurnal Expression of Genes Encoding Fucoxanthin Chlorophyll a / c Light Harvesting Polypeptides in the Centric Diatom Cyclotella cryptica. *Plant Biology* **4**, 250–257 (2002).

324. Garcia-gimeno, M. A. & Struhl, K. Aca1 and Aca2 , ATF / CREB Activators in Saccharomyces cerevisiae , Are Important for Carbon Source Utilization but Not the Response to Stress. *Molecular and Cellular Biology* **20**, 4340–4349 (2000).

325. Erzberger, J. P. & Berger, J. M. Evolutionary Relationships and Structural Mechanisms of AAA + Proteins. *Annual review of Biophysiology and Biomolecular Structure* **35**, 93–114 (2006).

326. Henry, E., Fung, N., Liu, J., Drakakaki, G. & Coaker, G. Beyond Glycolysis: GAPDHs Are Multi-functional Enzymes Involved in Regulation of ROS, Autophagy, and Plant Immune Responses. *PLOS Genetics* **11**, e1005199 (2015).

327. Yang, Y., Kwon, H., Peng, H. & Shih, M. Stress Responses and Metabolic Regulation of Clyceraldehyde-3-Phosphate Dehydrogenase in Arabidopsis. *Plant Physiology* **101**, 209–216 (1993).

328. Alipanah, L., Rohloff, J., Winge, P., Bones, A. M. & Brembu, T. Whole-cell response to nitrogen deprivation in the diatom Phaeodactylum tricornutum. *Journal of Experimental Botany* **66**, 6281–6296 (2015).

329. Axmann, I. M., Hess, W. R. & Wilde, A. An internal antisense RNA regulates expression of the photosynthesis gene isiA. *PNAS* **103**, 7054–7058 (2006).

330. Forouzan, E., Shariati, P., Sadat, M., Maleki, M. & Karkhane, A. A. Practical

evaluation of 11 de novo assemblers in metagenome assembly. *Journal of Microbiological Methods* **151**, 99–105 (2018).

331. Chacon, J. & Cuajungco, M. P. Comparative De Novo Transcriptome Assembly of Notophthalmus viridescens RNA-seq Data using Two Commercial Software Programs. *Californian journal of health promotion* **16**, 46–53 (2018).

332. Kumar, S. & Blaxter, M. L. Comparing de novo assemblers for 454 transcriptome data. *BMC genomics* **571**, 1–12 (2010).

333. Thomas, T., Gilbert, J. & Meyer, F. Metagenomics - a guide from sampling to data analysis. *Microbial Informatics and Experimentation* **2**, 3 (2012).

334. Carradec, Q. A global ocean atlas of eukaryotic genes. *Nature communications* **9**, 1–13 (2018).

335. Zhang, H., Huang, X., Huang, L., Bao, F. & Xiong, S. Science of the Total Environment Microeukaryotic biogeography in the typical subtropical coastal waters with multiple environmental gradients. *Science of the Total Environment* **635**, 618–628 (2018).

336. Thaler, M. & Lovejoy, C. Environmental selection of marine stramenopile clades in the Arctic Ocean and coastal waters. *Polar Biology* **37**, 347–357 (2013).

337. Chu, F.-L. E. & Greene, K. H. Effect of temperature and salinity on in vitro culture of the oyster pathogen, Perkinsus marinus (Apicomplexa: Perkinsea). *Journal of Invertebrate Pathology* **53**, 260–268 (1989).

338. Cavalier-Smith, T. & Chao, E. E. Phylogeny and Evolution of Apusomonadida (Protozoa: Apusozoa): New Genera and Species. *Protist* **161**, 549–576 (2010).

339. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology* **12**, e1001889 (2014).

340. McManus, G. B. & Katz, L. A. Molecular and morphological methods for identifying plankton: what makes a successful marriage? *Journal of Plankton*

Research **31**, 1119–1129 (2009).

341. Keygene, N. V. Expression profiling and local adaptation of Boechera holboellii populations for water use efficiency across a naturally occurring water stress gradient. *Molecular Ecology* **15**, 1229–1237 (2006).

342. Bischoff, B. & Wiencke, C. Temperature ecotypes and biogeography of Acrosiphoniales (Chlorophyta) with Arctic- Antarctic disjunct and Arctic / cold-temperature distributions. *European Journal of Phycology* **30**, 19–27 (1995).

343. Chai, C., Jiang, T., Cen, J., Ge, W. & Lu, S. Phytoplankton pigments and functional community structure in relation to environmental factors in the Pearl River Estuary. *Oceanologia* **58**, 201–211 (2016).

344. Gong, W. *et al.* Eukaryotic phytoplankton community spatiotemporal dynamics as identified through gene expression within a eutrophic estuary. *Environmental Microbiology* **20**, 1095–1111 (2018).

345. Lomas, M. W. & Glibert, P. M. Comparison of nitrate uptake, storage, and reduction in marine diatoms and flagellates. *Journal of Phycology* **913**, 903–913 (2000).

346. Bender, S. J., Durkin, C. A., Berthiaume, C. T., Morales, R. L. & Armbrust, E. V. Transcriptional responses of three model diatoms to nitrate limitation of growth. *Frontiers in Marine Science* **1**, 1–15 (2014).

347. Stevens, F. C. Calmodulin: an introduction. *Canadian Journal of Biochemistry and Cell Biology* **61**, 906–910 (1983).

348. Street, J. H., Commission, C. C., Paytan, A. & Cruz, S. *Iron, Phytoplankton Growth, and the Carbon Cycle*. (2005). doi:10.1201/9780824751999.ch7

349. Valegård, K. *et al.* Structural and functional analyses of Rubisco from arctic diatom species reveal unusual posttranslational modifications. *The Journal of Biological Chemistry* **293**, 13033–13043 (2018).

350. Hampl, V. *et al.* Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "“ supergroups ”". *PNAS* **106**, 3859–3864 (2009).

351. Gong, Y. *et al.* Phytoplankton blooms : An overlooked marine source of

natural endocrine disrupting chemicals. *Ecotoxicology and Environmental Safety* **107**, 126–132 (2014).

352. Jia, Y. *et al.* Cyanobacterial blooms act as sink and source of endocrine disruptors in the third largest freshwater lake in China. *Environmental Pollution* **245**, 408–418 (2019).

353. Durak, G. M., Brownlee, C. & Wheeler, G. L. The role of the cytoskeleton in biomineralisation in haptophyte algae. *Scientific Reports* **7**, 1–12 (2017).

354. King, N. *et al.* The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature* **451**, 783–788 (2008).

355. Winder, M., Berger, S. A. & Lewandowska, A. Spring phenological responses of marine and freshwater plankton to changing temperature and light conditions. *Marine Biology* **159**, 2491–2501 (2012).

356. Strzepek, R. F., Hunter, K. A., Frew, R. D., Harrison, P. J. & Boyd, P. W. Iron-light interactions differ in Southern Ocean phytoplankton. *Limnology and Oceanography* **57**, 1182–1200 (2012).

357. Achterberg, E. P. *et al.* Iron Biogeochemistry in the High Latitude North Atlantic Ocean. *Scientific Reports* **8**, 1283 (2018).

358. Deming, J. W. & Miller, L. A. Iron in sea ice : Review and new insights. *Elementa: Science of the Anthropocene* **4**, 1–19 (2016).

359. Baker, G. C., Smith, J. J. & Cowan, D. A. Review and re-analysis of domain-specific 16S primers. *Journal of Microbiological Methods* **55**, 541–555 (2003).

360. Fox, G. E., Wisotzkey, J. D. & Jurtshuk, P. How Close Is Close: 16s rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. *International Journal of Systematic Bacteriology* **42**, 166–170 (2019).

361. Staley, J. T. The bacterial species dilemma and the genomic – phylogenetic species concept. *Philosophical Transactions of the Royal Society B* **361**, 1899–1909 (2006).

362. Thomas, M. K., Kremer, C. T., Klausmeier, C. A. & Litchman, E. A Global Pattern of Thermal Adaptation in Marine Phytoplankton. *Science* **338**, 1085–1089 (2012).

363. Gogorev, R. M., Samsonov, N. I., Гогорев, Р. М. & Самсонов, Н. И. The genus Chaetoceros (Bacillariophyta) in Arctic and Antarctic. *Novosti Sistematiki Nizshikh Rastenii* **50**, 56–111 (2016).

364. Pančić, M., Torres, R. R., Almeda, R. & Kiørboe, T. Silicified cell walls as a defensive trait in diatoms. *Proceedings of the Royal Society B: Biological Sciences* **286**, 20190184 (2019).

365. Jang, M.-C., Shin, K., Lee, T. & Noh, I. Feeding selectivity of calanoid copepods on phytoplankton in Jangmok Bay, south coast of Korea. *Ocean Science Journal* **45**, 101–111 (2010).

366. Ger, K. A., Naus-Wiezer, S., De Meester, L. & Lürling, M. Zooplankton grazing selectivity regulates herbivory and dominance of toxic phytoplankton over multiple prey generations. *Limnology and Oceanography* **64**, 1214–1227 (2019).

367. Hansen, F. C., Reckermann, M., Breteler, W. C. M. K. & Riegman, R. Phaeocystis blooming enhanced by copepod predation on protozoa - Evidence from incubation experiments. *Marine Ecology Progress Series* **102**, 51–58 (1993).

368. Hasle, G. R. & Syvertsen, E. E. Marine Diatoms. in *Identifying Marine Phytoplankton* (ed. Tomas, C. R.) 5–385 (Academic Press, 1997). doi:https://doi.org/10.1016/B978-012693018-4/50004-5.

369. Percopo, I. *et al.* A new potentially toxic Azadinium species (Dinophyceae) from the Mediterranean Sea, A. dexteroporum sp. nov. *Journal of Phycology* **49**, 950–966 (2013).

370. Worden, A. Z. *et al.* Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. *Science* **324**, 268–272 (2009).

371. Throndsen, J. The Planktonic Marine Flagellates. in *Identifying Marine Phytoplankton* (ed. Tomas, C. R.) 591–729 (Academic Press, 1997). doi:https://doi.org/10.1016/B978-0-12-693018-4.X5000-9

372. Borowitzka, M. A. Biology of Microalgae. in *Microalgae in Health and Disease Prevention* (eds. Levine, I. & Fleurence, J.) 23–72 (Academic Press, 2018).

doi:https://doi.org/10.1016/B978-0-12-811405-6.00003-7.

373. Johnson, G. C., Schmidtko, S. & Lyman, J. M. Relative contributions of temperature and salinity to seasonal mixed layer density changes and horizontal density gradients. *Journal of Geophysical Research* **117**, 1–13 (2012).

374. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208 (2015).

375. Solden, L., Lloyd, K. & Wrighton, K. The bright side of microbial dark matter : lessons learned from the uncultivated majority. *Current Opinion in Microbiology* **31**, 217–226 (2016).

376. Ashelford, K. E., Chuzhanova, N. a., Fry, J. C., Jones, A. J. & Weightman, A. J. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology* **71**, 7724–7736 (2005).

377. Williams, T. A. & Embley, T. M. Archaeal “ Dark Matter ” and the Origin of Eukaryotes. *Genome Biology Evolution* **6**, 474–481 (2014).

378. Woese, C. R., Kandlert, O. & Wheelis, M. L. 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* **87**, 4576–4579 (1990).

379. Burki, F., Okamoto, N. & Keeling, P. J. The evolutionary history of haptophytes and cryptophytes : phylogenomic evidence for separate origins. *Proceedings. Biological sciences* **279**, 2246–2254 (2012).

380. Egge, E. S. *et al.* Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing. *Molecular Ecology* **24**, 3026–3042 (2015).

381. Underwood, G. J. C. *et al.* Organic matter from Arctic sea-ice loss alters bacterial community structure and function. *Nature Climate Change* **9**, 170–176 (2019).

382. Pen, O., Boyd, P. W. & Hutchins, D. A. Understanding the responses of ocean biota to a complex matrix of cumulative anthropogenic change. *Marine*

Ecology Progress Series **470**, 125–135 (2012).

383. Tanaseichuk, O., Borneman, J. & Jiang, T. Phylogeny-based classification of microbial communities. *Bioinformatics* **30**, 449–456 (2014).

384. Callahan, B. J., McMurdie, P. J. & Holmes, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal* **11**, 2639–2643 (2017).