Investigating marine microbial metabolism and diversity of Arctic ecosystems

David Colatriano

A Thesis
In the Department
of
Biology

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosphy (Biology) at
Concordia University
Montreal, Quebec, Canada

June 2019

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: **David Colatriano**

Entitled: **Investigating marine microbial metabolism and diversity of Arctic ecosystems**

and submitted in partial fulfillment of the requiremnets for the degree of

Doctor of Philosophy (Biology)

Complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

______________________________Chair
Dr. Natalie Phillips

______________________________Examiner
Dr. Dajana Vuckovic

______________________________Examiner
Dr. William Zerges

______________________________External to Program
Dr. Brandon Findlay

______________________________External Examiner
Dr. Lyle Whyte

______________________________Thesis Supervisor
Dr. David Walsh

Approved by ____
Dr. Robert Weladji, Gradute Program Director

Tuesday, August 27, 2019 ______________________________
Dr. André Roy, Dean
Faculty of Arts and Science

Abstract

Investigating marine microbial metabolism and diversity of Arctic ecosystems

David Colatriano, PhD

Concordia University, 2019

The world's oceans are essential for sustaining life on Earth and harbour a vast diversity of organisms. Marine microbes play crucial roles in global biogeochemical cycling and are at the base of marine food webs. Due to the technical difficulties associated with sampling northern marine systems, relatively little is known about the microbial community composition and metabolism of these ecosystems.

In this thesis work, the community composition and metabolism of northern marine ecosystems, including the Saint Lawrence Estuary, North Water and Canada Basin were described using meta-omic techniques. In the Saint Lawrence Estuary, differences in microbial community structure, metabolic lifestyles and carbon and nitrogen processing pathways were observed between the surface and deep waters. In the North Water, two distinct microbial communities with different taxonomic compositions and differing nutrient acquisition and resource allocation strategies were identified on either side of the polar mixed layer, and a third distinct community was described in the bottom waters. Functional and taxonomic analyses of the North Water polar mixed layer communities suggest a microbial community more typically associated with waters that undergo pulses of primary production on the Canadian side, while the community on the Greenland side was more typical of waters associated with a more steady level of primary production. In the Canada Basin, metagenomics was used to construct 360 Arctic Ocean metagenome assembled genomes. The analysis of six Chloroflexi MAGs revealed their potential for terrestrial derived aromatic compound degradation and that this metabolic capacity was acquired, at least in part, by lateral gene transfer from terrestrial organisms. To facilitate the meta-omic analyses performed in this thesis, a novel method to isolate microbial community DNA and proteins from the same environmental sample preserved in RNAlater was also developed.

This thesis not only describes the microbial community composition and metabolism of northern marine systems over a broad geographic range, but also adds to the growing metagenomic and metaproteomic resource-base that can be used to develop and test hypotheses about northern marine microbial systems. Additionally, this work has implications for our understanding of how climate change may affect northern marine ecosystems.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr. David Walsh for his support, guidance and patience throughout my thesis work. Your mentorship has not only allowed me to explore and experience parts of the world few get to see, but has allowed me to grow as a researcher and story-teller. I would also like to thank my committee members Dr. William Zerges and Dr. Dajana Vuckovic for their suggestions and valuable input during my studies.

I would like to thank my labmates and friends at Concordia for always being there to discuss or bounce ideas off of and for keeping my time working on this PhD fun.

Last but certainly not least, I would like to express my deepest gratitude to my family. I need to thank my parents who have supported me and fostered my sense of curiosity since the day I could form questions. Without your guidance and support I certainly would not be where I am today. I would like to thank my sister for her encouragement, and my amazing wife Karen for her unyielding support, encouragement and for keeping me grounded throughout this work.

Contribution of Authors

In Chapter 2 elemental and isotopic analyses were performed by Yves Gélinas, and metagenomic sequencing was performed by Etienne Yergeau. Arthi Ramachandran constructed the methanol dehydrogenase phylogeny. Data analysis, including metaproteomic analysis was performed by David Colatriano and David A. Walsh. David Colatriano and David A. Walsh, with contributions from Etienne Yergeau, Roxane Maranger and Yves Gélinas, wrote the manuscript.

In Chapter 5 David Colatriano and David A. Walsh designed and carried out the metagenomic experiments. William J. Williams, Connie Lovejoy and Céline Guéguen designed the sampling strategy. William J. Williams provided oceanographic data. David A. Walsh collected samples. David Colatriano extracted the environmental genomic DNA and analyzed sequencing data. Patricia Tran contributed to the bioinformatic analyses. David Colatriano and David A. Walsh wrote the manuscript. All authors commented on the manuscript.

Table of Contents

List of figures

List of tables

List of supplementary figures

List of abbreviations

ABC	ATP-binding cassette
ABH	Arctic Basin halocline
AmtB	Ammonia channel protein
ATP	Adenosine triphosphate
BLAST	Basic local alignment search tool
BTM	Bottom
CB	Canada Basin
CDOM	Chromophoric dissolved organic matter
CID	Collision-induced dissociation
COG	Cluster of orthologous genes
CRAM	Carboxyl-rich alicyclic molecules
cRAP	Repository of Adventitious Proteins
CTD	Conductivity, Temperature and Density
DNA	Deoxyribonucleic acid
DOC	Dissolved organic carbon
DOM	Dissolved organic matter
EDTA	Ethylenediaminetetraacetic acid
EF-Tu	Elongation factor Tu
E-PML	Eastern Polar Mixed Layer
FDOM	Fluorescent dissolved organic matter
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
FMNO	(FM)/F420-dependent monooxygenase catalytic subunit
FT-ICR-MS	Fourier transform ion cyclotron resonance mass spectrometry
GSO	Sulfur-oxidizing Gamma-proteobacteria
HMW	High molecular weight
HPLC	High performance liquid chromatography
IMG	Integrated Microbial Genomes
KEGG	Kyoto Ecyclopedia of Genes and Genomes

LCA	Lowest common ancestor
LMW	Low molecular weight
LSLE	Lower Saint Lawrence Estuary
LTQ	Linear trap quadropole
MAG	Metagenome Assembled Genome
mDOM	Marine-derived dissolved organic matter
MG	Marine group
MRM	Multiple reaction monitoring
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
m/z	Mass to charge ratio
NC	Northern Canada
NG	Northern Greenland
NNI	Nearest-neighbor interchange
NOB	Nitrite oxidizing bacteria
NXR	Nitrite oxidoreductase
OGOR	2-oxogluturate:ferredoxin oxidoreductase
OM	Organic matter
OMZ	Oxygen minimum zone
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
PCoA	Principal coordinate analysis
PCR	Polymerase chain reaction
POC	Particulate organic carbon
POM	Particulate organic matter
PON	Particulate organic nitrogen
POR	Pyruvate:ferredoxin oxidoreductase
PQQ	Pyrroloquinoline quinone
PSM	Peptide spectra match
PSU	Practical salinity units

PTM	Post-translational modification
PTN	Particulate total nitrogen
RbcL	Ribulose bisphosphate carboxylase large subunit
RbcS	Ribulose bisphosphate carboxylase small subunit
rRNA	Ribosomal ribonucleic acid
rTCA	Reductive tricarboxylic acid cycle
RuBisCo	Ribulose bisphosphate carboxylase
SAG	Single-cell amplified genome
SC	Southern Canada
SCM	Subsurface chlorophyll maximum
SDS	Dodium dodecyl sulfate
SG	Southern Greenland
SLE	Saint Lawrence Estuary
SPM	Suspended particulate matter
SRM	Selected reaction monitoring
TBDT	TonB-dependent transporter
tDOM	Terrestrial-derived dissolved organic matter
TRAP	Tripartite ATP-independent periplasmic
Tr-COG	Taxonomically-resolved COG
UDOM	Ultrafiltered dissolved organic matter
USLE	Upper Saint Lawrence Estuary
WGC	West Greenland Current
W-PML	Western Polar Mixed Layer

Chapter 1: Introduction

1.1 Marine microbial ecology

The world's oceans harbour an incredible amount of microbial life with an estimated 10^5 cells per mL in the surface waters and a predicted 3.6 x 10^{29} microbial cells in the oceans (Whitman *et al.*, 1998). Not only is microbial life abundant in marine systems, but it is also highly diverse (Sogin *et al.*, 2011) and plays essential roles in the biogeochemical cycles of the Earth, contributing greatly to global energy and nutrient cycling (Madsen, 2011). Examples include microbial participation in the Earth's carbon and nitrogen cycles, as well as oxygen (O_2) production (Arrigo, 2005; Benner, 1998; Kujawinski, 2011; Pomeroy *et al.*, 1998). The role of marine microbial ecologists is to understand how marine microorganisms interact with the complex patterns of physical, chemical and biological variation in marine systems, and to investigate drivers of evolution and diversification of the microorganisms living within these systems. However, marine microbial diversity and abundance varies with geographic location and time, and depends on many environmental factors including nutrient composition and concentration, temperature and oxygen concentration, to name a few (Fuhrman *et al.*, 2015; Giovannoni and Stingl, 2005; Pommier *et al.*, 2007). Current studies in marine microbiology are trying to answer questions such as: What is the community composition of particular ecosystems and what makes the community composition at one location different than at another? Are there microbial populations endemic to specific environments, and if so, what are the local adaptations to the environment that allow for their survival? What metabolic processes are most important to specific environments and what organisms perform these processes?

In the long term, the answers to these questions are used to better understand the dynamics of marine microbial communities; how they interact with each other and environmental variables, and how different environmental variables influence microbial community composition and metabolism. Given that global atmospheric warming is occurring, answering these questions will allow for a better understanding of how changes in the environment will impact microbial communities and the metabolic roles they play.

1.2 Culture-independent approaches in microbial ecology

There are many approaches available for characterizing microbial community structure and function. A powerful method to study the physiology of microbial cells is to isolate and cultivate strains from any given community to study their metabolic capabilities. Cultivated strains can provide complete genome sequences of high confidence, offering insights into their potential metabolisms (Giovannoni and Stingl, 2007). Although powerful, cultivation of environmental microbes is challenging. One of the main challenges of marine microbial culturing is designing artificial media to mimic natural seawater with an accurate reconstruction of the complex composition of dissolved organic carbon (DOC) and trace elements (Giovannoni and Stingl, 2007). Another challenge is resolving the complex community interactions that take place in the environment. Although many important clades of marine microbes now have cultured representatives (Giovannoni and Stingl, 2007; Martiny, 2019), several more remain uncultivated and are likely to remain that way without further advances in cultivation technology (Giovannoni and Stingl, 2005). Meta-omic technologies such as metagenomics, metatranscriptomics and metaproteomics can help in the understanding of *in situ* community structure and metabolism.

1.2.1 rRNA gene sequence analysis

The first attempts to overcome challenges associated with cultivation-dependent strategies involved the use of nucleic acid-based cloning and sequencing of universal phylogenetic markers like the small-subunit ribosomal RNA (rRNA) genes from mixed microbial communities and subsequent phylogenetic analysis. This was the technique used to construct the universal tree of life in 1987 (Woese, 1987). Since then rRNA gene sequence analysis has enabled microbial ecologists to identify the phylogenetic identity and relative abundance of microbial community constituents, improving our understanding of the microbial world. For example, rRNA gene analysis was used to reveal that Archaea were not strictly extremophile organisms, but could instead be found from depths of 100 m - 500 m in the Pacific Ocean (Fuhrman *et al.*, 1992) and in the coastal surface waters of North America (Delong, 1992). Further rRNA gene analysis (Massana *et al.*, 1997; Karner *et al.*, 2001) suggested that these Archaea might represent one of the most abundant cell type in the ocean. Ribosomal RNA gene analysis also led to the identification of a novel clade of Alpha-proteobacteria, the SAR11 cluster, as a significant component of the oligotrophic community in the Sargasso Sea (Giovannoni *et al.*, 1990). Based on these findings and others identifying SAR11 throughout the oceans, fluorescence *in situ* hybridization (FISH) analysis of Sargasso Sea and Oregon coastal waters revealed that SAR11 represented one of the most abundant and successful organisms in marine systems (Morris *et al.*, 2002). Ribosomal RNA gene sequencing has also been used for comparative studies of microbial community composition along spatial and temporal scales, allowing microbial ecologists to explore links between microbial community composition and differing environmental factors, for example in the Baltic Sea (Herlemann *et al.*, 2011) or the North Water (Joli *et al.*, 2018). Short term temporal studies have shown the dynamic nature of microbial communities, even during short time scales of 24 hours (Joli *et al.*, 2018). Longer term time-series studies have been used to link environmental parameters to bacterioplankton variability and assemblages of bacteria associated with different seasons (El-Swais *et al.*, 2015) and to track the succession of bacterial taxa in response to phytoplankton blooms (Teeling *et al.*, 2012; El-Swais *et al.*, 2015). Longer time-series studies have also been used to capture the microbial community response to long-term environmental changes including changes in microbial community composition in the Beaufort Sea over an eight year period (Comeau *et al.*, 2011), or changes in the phytoplankton community over a nine year period in the same system (Li *et al.*, 2013).

Initially, when clone libraries were used for rRNA gene analysis, rare organisms that may contribute significantly to metabolic processes in the community, were scarcely identified. With the advent of next generation sequencing technologies, it became apparent that there was ~ 1-2 orders of magnitude more diversity than previously thought and that although a relatively small number of different populations dominated all samples, thousands of low-abundance populations accounted for most of the observed phylogenetic diversity (Sogin *et al.*, 2006). Uncovering the ecological roles and contributions that members of this "rare biosphere" play in an ecosystem is important for a complete understanding of ecosystem function and how community composition and functions might change with a changing environment. Unfortunately the limitations of rRNA gene analysis prevent this. Although the small-subunit rRNA genes are recognized as phylogenetic markers, they do not divulge any information about the metabolism of the community.

1.2.2 Metagenomics

Whole genome shotgun metagenomics is the analysis of the genetic content of an assemblage of organisms enabling the identification of protein coding genes as well as housekeeping genes like the 16S rRNA genes. Therefore metagenomics can provide information about community taxonomy, relative abundance, metabolic diversity (DeLong, 2002; Temperton and Giovannoni, 2012), as well as genomic characteristics like nucleotide content and codon utilization. In the past, metagenomic analysis was performed by constructing metagenomic libraries with isolated high-quality microbial DNA from environmental samples. This DNA was then digested and cloned into a plasmid, fosmid or cosmid for sequencing and assembly into contigs (ex. (Stein *et al.*, 1996; Vergin *et al.*, 1998; Beja *et al.*, 2000; Venter *et al.*, 2004; Gilbert and Dupont, 2011)). Advances in sequencing technologies has now enabled researchers to bypass the cloning step and directly sequence isolated community DNA (Tully *et al.*, 2017; Pesant *et al.*, 2015). Metagenomic assemblies can then be functionally annotated and the contigs can be binned, or grouped together, into metagenome assembled genomes (MAGs) based on genomic characteristics like GC content, tetranucleotide frequency and/or coverage (Alneberg *et al.*, 2014; Kang *et al.*, 2015; Strous *et al.*, 2012).

By reconstructing the genomes of environmental organisms from metagenomic data, a better understanding of the metabolic potential of environmental bacteria and archaea that are resistant to laboratory cultivation is gained. For instance, a landmark study using shotgun metagenomic sequencing of a low-complexity acid mine drainage acidophilic biofilm identified the microbial community composition of the system and reconstructed the genomes of the five dominant members (Tyson *et al.*, 2004). Metabolic reconstruction of these MAGs revealed that a low abundance member, *Leptospirillum* group III was the sole community member with a nitrogen fixation pathway and was likely a keystone species in this ecosystem. Another landmark study was the metagenomic analysis of the more complex microbial communities of the Sargasso Sea, resulting in the identification of over 1 million previously unknown genes and 148 previously unknown microbial phylotypes (Venter *et al.*, 2004).

The success of such studies inspired efforts to collect large-scale metagenomic data from marine environments worldwide, including 41 samples from the global ocean as part of the

Global Ocean Survey (Rusch *et al.*, 2007) and, most notably, the TARA Oceans expedition which collected more than 200 metagenomic samples from marine ecosystems throughout the globe (Tully *et al.*, 2017). The analysis of the TARA Ocean's dataset resulted in the generation of 2,631 metagenome assembled genomes including genomes from enigmatic microbes such as: Group II and Group III Euryarchaeota, SAR324 and Marinimicrobia (Tully *et al.*, 2017). More recently, metagenomic data was generated from 610 metagenomic samples collected in the Atlantic and Pacific Oceans as part of the GEOTRACES program (Biller *et al.*, 2018).

Targeted metagenomic studies can also be performed to discover novel marine microbial functions. Examples of this include the discovery of genes for phototrophy in marine bacteria that were previously thought to rely exclusively on the oxidation of organic matter for energy (Beja *et al.*, 2000) and the discovery of novel chemolithotrophic energy-generating pathways in marine bacteria and archaea (Hallam *et al.*, 2006; DA Walsh *et al.*, 2009; Walker *et al.*, 2010; Anantharaman *et al.*, 2013). Other metagenomic analyses have been used to reconstruct the genomes of enigmatic microbes, including the SAR324 Delta-proteobacteria (Sheik *et al.*, 2013), to gain insight into their distribution and metabolic potential. Moreover, comparative metagenomics can be used to elucidate differences in the metabolic potential of different communities. For instance metagenomic analysis of Antarctic surface water communities at two contrasting times of the annual cycle (Summer and Winter) revealed shifts in community composition and functional capacities between the two seasons with a significantly higher phylogenetic and functional diversity observed in the winter (Grzymski *et al.*, 2012). Metagenomic sequences from this study were then used as part of a searchable database for a metaproteomic analysis of the same region.

Although powerful, one of the main limitations of metagenomic analysis is that it is only capable of providing information on community structure and the functional potential of a community without providing information about the *in situ* or realized metabolism of the community. In order to determine which metabolic genes are actually being expressed in a given environment, post-genomic analysis is required.

1.2.3 Post-genomic approaches

Genomic analyses like rRNA gene sequencing and metagenomics allow for better understanding of microbial community structure and metabolic potential of a community, but one of the fundamental questions of microbial ecology is whether an organism or a gene allows us to conclude anything about particular biogeochemical functions in an ecosystem. Due to the plasticity of gene expression, post-genomic analyses such as metatranscriptomics or metaproteomics can provide a more complete understanding of microbial function and help answer this question (Schneider and Riedel, 2010). Metatranscriptomics, or the analysis of *in situ* gene expression, is useful because RNA transcripts provide sequence data, that can be used to describe community structure, as well as metabolic gene expression data that can be used to describe communiy metabolism (Ottesen *et al.*, 2011; Hewson *et al.*, 2010; Poretsky *et al.*, 2009; McLean, 2013; Moran *et al.*, 2012; Gilbert *et al.*, 2010). However, mRNA levels do not necessarily correspond to their protein levels due to translational regulation, RNA half-life, and the fact that multiple protein copies can be generated for every mRNA (Vogel and Marcotte, 2012). Thus, the metatranscriptome will not necessarily represent the final gene expression profiles of a community.

For this reason, metaproteomics is an important tool used in environmental microbiology. Metaproteomic analysis uses a shotgun proteomic approach where the near-complete complement of proteins from a complex environmental sample is purified and analyzed. Proteins are usually enzymatically digested into peptides and subsequently sepparated by liquid chromatography and subjected to tandem mass spectrometry (MS/MS)-based analysis. The resulting mass spectra are searched against a protein sequence database to identify the amino acid sequences of the peptides and their possible proteins of origin (Hettich *et al.*, 2013). Thanks to increases in genomic data for database searching, available from single cell amplified genomes, metagenomic studies and cultivation efforts, and the increasing sensitivity and resolving power of mass spectrometers, high-throughput protein identification and quantification can now be performed (von Bergen *et al.*, 2013; Hettich *et al.*, 2013).

Early MS/MS-based metaproteomic studies in the ocean were used to identify specific proteins in targeted microbial lineages, with the first study focusing on the light driven proton pump proteorhodopsin in SAR11 marine Alpha-proteobacteria (Giovannoni, Bibbs, *et al.*, 2005). The complete genome of SAR11 isolates were previously sequenced and the gene for proteorhodopsin was identified (Giovannoni, Tripp, *et al.*, 2005; Rappé *et al.*, 2002). These genes were also identified from metagenomes from the Sargasso Sea (Venter *et al.*, 2004). Isolates were then cultured and proteomic analysis was performed showing the expression of proteorhodopsin in cultured SAR11. Because protein expression patterns of cultured bacteria do not necessarily represent the expression pattern of bacteria in nature, a metaproteomic analysis of environmental samples was performed, identifying expressed SAR11 proteorhodopsin proteins in natural communities. Studies such as this demonstrated that metaproteomics can be used as a powerful tool to confirm the expression of predicted genes or metabolic pathways in the environment.

One area of interest is the study of expressed membrane transporters because investigating nutrient transport proteins such as components of the ATP-binding cassette (ABC) transporters, TonB-dependent transporters (TBDTs) and tripartite ATP-independent periplasmic (TRAP) transporters can provide insight into the nutrient availability of the system as well as differences in microbial metabolic strategies and nutrient preferences (Williams and Cavicchioli, 2014). Examples of this include studies of both coastal and open ocean systems. In a nutrient-rich coastal system off the coast of Oregon, transporters for amino acids, taurine, polyamines, as well as glutamine synthetase, an enzyme involved in the assimilation of ammonium into amino acids known to be expressed when dissolved organic or inorganic nitrogen is limiting in the environment, were among the most frequently detected proteins. In contrast, transporters for phosphorus were rare, supporting predictions that carbon and nitrogen are more limiting than phosphorus in this environment (Sowell *et al.*, 2011). A similar metaproteomic study performed in an oligotrophic open ocean system in the Sargasso Sea identified transporters for phosphate, amino acids, phosphonate, sugars and spermidine, supporting the idea that competition for multiple nutrients in oligotrophic systems is extreme (Sowell *et al.*, 2009). These two studies show differences in the nutrient availability between the two systems, with the more nutrient-rich coastal system metaproteome indicating an environment less limited in phosphorus than the oligotrophic open ocean system. More recently, a metaproteomic analysis of Atlantic Ocean waters showed that the relative abundance of specific transport proteins changes with depth and that bathypelagic communities were more geared towards the utilization of solubilized particulate organic matter (POM) and aromatic compounds (Bergauer *et al.*, 2017).

Metaproteomics can also be used to explore the metabolic versatility of microbial lineages. For instance, the uncultured lineage Arctic96BD-19 of Gamma-proteobacteria and the SAR324 clade of Delta-proteobacteria are widespread in the surface (Marshall and Morris, 2013; DeLorenzo *et al.*, 2012) and deep ocean waters (Swan *et al.*, 2011; Sheik *et al.*, 2013). Single-cell genome sequencing of SAR324 and Arctic96BD-19 revealed the capacity for sulfur oxidation and CO_2 fixation (Swan *et al.*, 2011) while subsequent genomic and transcriptomic work on SAR324 revealed the potential for a mixotrophic lifestyle, including sulfur oxidation, autotrophic CO_2 fixation, C1 metabolism, as well as heterotrophy (Sheik *et al.*, 2013). A metaproteomic analysis of the coastal North Atlantic identified transport proteins originating from Arctic96BD-19 and SAR324 corresponding to a heterotrophic lifestyle, with no proteins corresponding corresponding to sulfur oxidation, autotrophic CO_2 fixation or C1 metabolism identified (Georges *et al.*, 2014). Therefore, studying the transport proteins of a metaproteome, allows insight into the metabolic activities of bacterial clades with less well, or uncharacterized metabolic capabilities, as well as their metabolic strategies in a given environment.

Comparative metaproteomic analyses have also been used to elucidate differing protein expression patterns between complex communities. These studies include both temporal variations such as those between seasons, as well as spatial differences (e.g. between geographic locations or depths). Examples of studies exploring temporal shifts in metabolism include a study performed on the coastal Northwest Atlantic Ocean demonstrating the importance of chemolithoautotrophic metabolism in winter months, with a shift in the spring to a more heterotrophic community (Georges *et al.*, 2014). Metaproteomic studies exploring shifts in community structure and metabolism between winter and summer seasons in the coastal waters of the Antarctic Peninsula have also been performed (Williams *et al.*, 2012), demonstrating different metabolic strategies between seasons. Another temporal metaproteomic study investigated the bacterioplankton response to a coastal spring phytoplankton bloom (Teeling *et al.*, 2012) where three phases of succession represented by three distinct major phylogenetic groups as well as their relative nutrient transport systems were described.

Examples of variations in protein expression patterns across spatial scales include a study along a geographical transect from a low-nutrient ocean gyre to a highly productive coastal upwelling system (Morris *et al.*, 2010), revealing shifts in nutrient utilization and energy transduction along the environmental nutrient gradient. Another example is a comparative metaproteomic analysis of different depths from an oxygen minimum zone (OMZ), the expanding and widespread regions of deep ocean characterized by O_2 concentrations < 20 μM. Exploring the water column of an OMZ revealed differential protein expression of OMZ microbes across the redoxcline (Hawley *et al.*, 2014). In this case, nitrification and inorganic carbon fixation pathways associated with *Thaumarchaeota* dominated dysoxic waters while the suboxic and anoxic waters were dominated by denitrification, sulfur oxidation, and inorganic carbon fixation pathways associated with the uncultivated SUP05 clade of bacteria. Comparative analysis of protein expression patterns of marine microbial populations can therefore lead to a better understanding of microbial community responses to environmental conditions.

Although powerful, metaproteomic analysis does have limitations (reviewed in (Heyer *et al.*, 2017)). One limitation is that many peptide spectra are not assigned to proteins because representative proteins are absent in the protein reference database. Moreover, it is possible that some peptides may originate from different phylogenetic lineages not represented in the protein database. This is especially problematic if the proteins are highly conserved or if their genes are prone to lateral gene transfer between marine microbes. These limitations can potentially be minimized by including the metagenome from the same sample as the analyzed metaproteomic dataset. Another solution is the inclusion of the increasing number of MAGs and SAGs that are becoming available for poorly characterized bacteria and archaea into the database (Rinke *et al.*, 2013), which will result in a more comprehensive database for peptide identification. The third major limitation of metaproteomic analysis is that it is only semi-quantitative. In general, the number of peptide spectra that match a protein relative to the total number of spectral matches for a sample is used as a measure. This is a semi-quantitative measurement that can help elucidate the relative expression of certain proteins compared to other, but does not actually give absolute quantitative values of protein expression.

1.3 Dissolved organic matter

1.3.1 Dissolved organic matter in marine environments

One fundamental role of microbial ecologists is to uncover the microbe–molecule interactions that govern global biogeochemical cycles. In marine environments a substantial component of these interactions are between microbes and dissolved organic matter (DOM). DOM consists of a heterogeneous mix of compounds made up of dissolved organic carbon, dissolved organic nitrogen and dissolved organic phosphorus, and generally occurs in functional groups common to biopolymers found in marine (mDOM) (Benner *et al.*, 1992; McCarthy *et al.*, 1997; Clark *et al.*, 1998; Ogawa and Tanoue, 2003) and terrestrial organisms (tDOM). Marine DOM is the largest ocean reservoir of reduced carbon, at 662 PgC, making it over 200 times the carbon inventory of marine biomass (Hansell, 2013; Hansell *et al.*, 2009) and approximately equivalent to the reservoir of atmospheric CO_2 (Ogawa et al, 2001). The dissolved organic carbon (DOC) component of DOM can be further differentiated into several fractions based on how biologically labile it is. The biologically labile fraction, produced primarily from photosynthetic production in ocean surface waters, constitutes a small fraction of the ocean DOC inventory (< 0.2 PgC) due to rapid turnover times of hours to days (Hansell, 2013). Semi-labile DOC is produced and transformed from both autotrophic and heterotrophic processing of carbon and is an important DOC component in the euphotic, mesopelagic zones of the oceans. Semi-labile DOC has a higher global inventory at 6 +/- 2 PgC and a turnover rate of months to years depending on its composition, the availability of macronutrients and the microbial community structure of the ecosystem. Semi-labile DOC tends to be made up of carbohydrates including monosaccharides and acyl-oligosaccharides (Aluwihare *et al.*, 1997), amino acids, neutral sugars and amino sugars (Hansell, 2013). Recalcitrant DOC has turnover rates of thousands of years, a global inventory of ~630 PgC and is composed of relatively small molecules, most notably, carboxyl-rich alicyclic molecules (CRAM) and molecules at concentrations too low for biological uptake. Labile and semi-labile DOC can also be transformed into more recalcitrant forms of DOC as heterotrophic organisms metabolize and use the most reactive moieties of a compound (Hansell, 2013). Although the processes involved in DOM transformation into more recalcitrant forms are not yet well understood, isotopic and amino acid enantiomer tracers indicate that both direct autotrophic production and heterotrophic processing lead to the accumulation of recalcitrant DOM in the oceans. Biotic transformation of labile and semi-labile DOM is also complemented by photolytic transformation of labile DOM and semi-labile DOM into recalcitrant DOM. Conversely, recalcitrant DOM can also be removed from a system abiotically via photooxidation, making it more labile to biological metabolism (Hansell, 2013).

1.3.2 Terrestrial-derived DOM

Another major component of the DOM pool in marine ecosystems is terrestrial-derived DOM (tDOM). This allochthonous DOM enters the marine system from terrestrial runoff or the erosion of organic-rich coastal sediments. Because the majority of tDOM originates from terrestrial plants, and lignin is a phenolic polymer synthesized exclusively by vascular plants (Hedges and Mann, 1979), dissolved lignin and lignin oxidation products are well established biomarkers of tDOM in marine environments (Hernes and Benner, 2003; Opsahl and Benner, 1997). Concentrations of lignin phenols in the Atlantic and Pacific oceans suggest that ocean margins might be important hotspots for tDOM transformation because residence times in these regions are decades to centuries and only a small fraction of old, refractory DOM found in the deep oceans is made up of tDOM (Benner, 1997; Opsahl and Benner, 1997). Although the degradation of lignin phenols can be accomplished through photochemical processes (Opsahl and Benner, 1998), microbial degradation is the dominant removal process in coastal and shelf waters (Fichot and Benner, 2014; Lu *et al.*, 2016). However, the molecular processes that govern this removal are still not fully understood. Additionally, the molecular composition and transformation of low molecular weight terrestrial DOM, in marine systems is an area that remains unexplored (Seidel *et al.*, 2017).

1.3.3 Chromophoric DOM

Another sub-category of the DOM pool in marine systems is the light absorbing chromophoric DOM (CDOM), which also contains the fluorescent DOM (FDOM) component. CDOM is ubiquitous in marine environments (Nelson and Siegel, 2013) but the fluorescence intensity of CDOM is low in the surface waters and increases with depth, most likely due to decreasing rates of photooxidation with depth (Yamashita and Tanoue, 2008). Its identification throughout marine environments, from the surface to the ocean floor has led to the suggestion that a substantial portion of CDOM is composed of long-lived and biologically refractory DOM. Contributors to the CDOM pool include aromatic amino acids (Yamashita and Tanoue, 2008), lignin phenols, and uncharacterized humic substances that are operationally characterized by their absorption and fluorescence properties (Coble, 1996). The optical properties of CDOM, particularly its fluorescence, can be used to distinguish compositional characteristics and discriminate between terrestrial and marine DOM sources (Coble, 1996; Blough and Del Vecchio, 2002; Guéguen *et al.*, 2012). Components of CDOM include terrestrial lignin phenols (fluorophores from higher-plant decomposition) (Hernes and Benner, 2003), terrestrial humic and fulvic acids in the coastal transition zone (Benner, 1997; Carder and Steward, 1989), as well as aromatic amino acids whose proportion of the CDOM pool declines with depth (Yamashita and Tanoue, 2008; Murphy *et al.*, 2008; Jørgensen *et al.*, 2011). Most of the remaining CDOM is loosely classified as marine humic material, but their structures and the processes by which they are formed are not understood.

CDOM abundance is globally variable, with surface waters showing low values in the subtropical gyres and higher values in subpolar oceans and regions of persistent upwelling (Kitidis *et al.*, 2006; Swan *et al.*, 2009; Yamashita and Tanoue, 2009; Nelson *et al.*, 2007, 2010; Ortega-Retuerta *et al.*, 2010). CDOM represents a small portion of the open ocean's total DOC pool (Nelson and Siegel, 2013). However, in coastal regions, where the terrestrial-marine interface supply high-CDOM/DOC waters to the oceans, higher proportions of CDOM are often observed (Nelson and Siegel, 2013). Although it was initially thought that CDOM in the oceans was a direct result of primary production, it was later suggested that microbial degradation of organic matter is responsible for the majority of autochthonous CDOM production (Nelson *et al.*, 1998). Culture-based studies of microbial production and degradation of CDOM (Nelson *et al.*, 2004) have shown that the microbially-produced labile CDOM fraction is rapidly degraded and would not necessarily be detectable in the water column. Therefore, it is hypothesized that the CDOM that can be detected tends to be made of recalcitrant DOM.

1.3.4 Microbe-DOM interactions

Understanding how the microbial community interacts with DOM is essential for understanding marine biochemical pathways and the interdependence of microbial cells that participate in different parts of these reactions. Marine microbes can interact with DOM in several ways. First and foremost, marine microbes can be a source of DOM. Autotrophic organisms can fix carbon into biomass, releasing it as DOM after cell death. DOM found in the water column can also be entirely or partially metabolized by microbial populations as a source of carbon or energy. Heterotrophic organisms can also be a source of DOM, releasing it as metabolites, extracellular enzymes, exported molecules, or after cell death. This partial metabolism or modification of DOM by heterotrophic organisms may lead to DOM with altered biochemical properties and biological lability. Alternatively, DOM can be left unmetabolized by microbial cells if it is found at concentrations that are too low or require metabolic pathways that the cells do not possess. Although it is well known that DOM in marine systems plays an important role in marine biogeochemical cycles, the composition of bacterially produced DOM is not fully resolved (Carlson, 2002). In general, a small proportion of marine DOM has been identified as specific known biomolecules (Opsahl and Benner, 1997; Benner, 1998; Hedges *et al.*, 2000), indicating that the molecular structure of marine DOM is often modified.

Just as microbial communities can alter DOM composition, so can the composition of DOM alter microbial community structure due to taxon specific preferences for certain concentrations and types of DOM. Enrichment experiments with high molecular weight (HMW) and low molecular weight (LMW) estuarine DOM showed that communities enriched with LMW DOM are dominated by Gamma-proteobacteria, Epsilon-proteobacteria and, to a lesser degree, Alpha-proteobacteria, while those enriched with HMW DOM are more diverse, with representation from the Alpha-proteobacteria, Beta-proteobacteria, Gamma-proteobacteria, and Bacteroidetes (Covert and Moran, 2001). Moreover, a shift in bacterial community structure from more oligotrophic taxa to more copiotrophic taxa was observed in coastal Arctic Ocean waters amended with Arctic river tDOM (dominated by lignin-like compounds) (Sipler *et al.*, 2017). Other studies have identified Alpha-proteobacteria as an important proponent for DOM metabolism throughout the marine environment (Alonso-sáez *et al.*, 2007), specifically SAR11, which were found to dominate amino acid and glucose uptake in the oligotrophic open surface ocean (Malmstrom *et al.*, 2005). However, Roseobacter, Gamma-proteobacteria and Bacteroidetes seem to play important roles in DOM uptake in coastal regions where DOM is found at higher concentrations (Alonso and Pernthaler, 2006; Alonso-Sáez and Gasol, 2007; Alonso-sáez *et al.*, 2007). Single-cell genomics as well as meta-omic analyses including metatranscriptomics and metaproteomics have also revealed taxon specific preferences for certain DOM compounds. Incubation experiments that followed transport protein transcripts of a coastal marine community amended with differing DOM sources (phytoplankton or plant-derived) revealed that SAR11 transporters were more important during incubations with phytoplankton-derived DOM, while Flavobacterial transport proteins became more important during incubations with plant-derived DOM (Poretsky *et al.*, 2010).

However, the ill-defined nature of marine DOM makes linking DOM to microbial metabolic processes difficult, especially when investigating microbe-recalcitrant DOM interactions. For instance, meta-omic studies attempting to establish links between DOM metabolism and microbial communities often use the identification of transport proteins annotated based on databases of reference organisms to determine the types of DOM compounds important to the community. However, given the fact that up to 75% of marine DOM is unknown, using functional annotations based on reference organisms makes inferring actual microbial-DOM interactions difficult and begs the question of whether the identified transport proteins actually transport the inferred substrate or something else? For this reason, the populations of microbes responsible for the degradation of these compounds, and how, is still unknown, making the biological degradation of recalcitrant DOM in marine systems an active area of research.

1.4 Study systems

1.4.1 Lower Saint Lawrence Estuary

The Saint Lawrence Estuary (SLE) is one of the largest estuaries in North America, located downstream of the Great Lakes and emptying into the Gulf of St. Lawrence in Eastern Canada via the St. Lawrence River. The SLE is a highly productive body of water characterized by strong vertical and geographic gradients that include temperature, salinity, nutrient and oxygen concentrations, as well as organic matter composition, making it an ideal environment to study how varying environmental conditions can affect microbial community structure and metabolism.

The Upper St. Lawrence Estuary (USLE) is vertically homogeneous and extends from Ile d'Orleans to the Saguenay River/Ile aux Coudres and transitions to the stratified three-layered Lower SLE (LSLE) which starts at the Saguenay River and ends at Pointe–des–Monts. This three-layer structure is comprised of a warm and brackish surface layer sitting atop a cold intermediate layer composed of sunken winter waters, and a warmer, hypoxic bottom water layer of Atlantic origin. The near bottom waters of the LSLE are also characterized by a particle–rich layer known as the nepheloid layer which is caused by the friction of tides above the seabed (Mulder and Alexander, 2001). Since the deep waters of the SLE are isolated from the atmosphere by a permanent pycnocline at 100-150 m of depth, the chemical properties of the deep SLE waters are changing. As the bottom water flows landward from the Atlantic, it gradually loses oxygen through respiration and microbial degradation of organic matter. Since oxygen cannot be replenished at depths below 150 m, the concentration of oxygen in the bottom water is sensitive to the fluxes of organic material and rates of oxygen supply with the landward flow. Dissolved oxygen concentrations in the bottom water have decreased by more than 50% over the last 70 years and an area of 1,300 km^2 of the Laurentian trough is now overlaid by hypoxic waters (Gobeil, 2006).

1.4.2 North Water

Moving North along the coast from the SLE is the Eastern Canadian Arctic Ocean, specifically the Labrador Sea, Baffin Bay and the North Water between Ellesmere Island and Greenland. The North Water region is a coastal ecosystem situated between Canada and Greenland that supports one of the largest, most productive and diverse marine ecosystems in the Arctic (Bâcle *et al.*, 2002; Ardyna *et al.*, 2011). It owes this productivity to its unique water column characteristics, which are produced by interactions from various water sources. Cold, phosphate-rich Arctic waters partially derived from the Pacific Ocean flow southward through Nares Straight and converge with locally produced Baffin Bay waters via the Baffin Current. The eastern North Water waters, flowing along the Greenland coast, originate from the Atlantic Ocean and travel North via the West Greenland Current (WGC) (Bâcle *et al.*, 2002). These northward flowing waters transport warmer, saltier, nitrate-rich Atlantic Ocean waters into the North Water and are eventually forced South by cyclonic circulation. These unique current dynamics also lead to a longer open water season compared to surrounding areas, exemplified by the formation of a polyna (open water surrounded by sea ice) in the winter, which enables early access to resources by migrating seabirds and marine mammals. The North Water is characterized by high productivity due to enhanced vertical mixing and is a major net outflow of Arctic waters to the North Atlantic, acting as a gateway connecting these two ocean regions.

The North Water has a fall phytoplankton bloom (Ardyna *et al.*, 2011) and the majority of primary production occurs in the eastern region. Roughly 45% of this production occurs during a 43-day build-up between May and July. Primary production in the West is lower than in the East (Tremblay *et al.*, 2002; Klein *et al.*, 2002), and a substantial part of this production occurs in May and early June. However, chlorophyll a accumulation does occur in July (Tremblay and Smith Jr., 2007) and is likely to contribute to the organic matter pool used by marine bacterioplankton on the western side of the North Water. The abundance of actively respiring bacteria (from depths of 8 to 90 m) increases ~10-fold from May to July, while the proteolytic activities of actively respiring bacteria associated with sinking particles (collected at depths of 50–136 m) triple in the same time period (Tremblay *et al.*, 2002).

1.4.3 Beaufort Sea

The Arctic Ocean comprises about 1% of the global ocean in terms of volume, but receives over 11% of the world's river discharge (Carmack *et al.*, 2016), collecting the greatest loads of freshwater and terrestrial derived organic matter of any ocean on a per volume basis (Hansell *et al.*, 2004). The Arctic Ocean is an enclosed ocean with four restricted openings: the Bering Strait which connects it to the Pacific Ocean and the Barents Sea, the channels in the Canadian Arctic Archipelago, as well as the Fram Strait which connects it to the North Atlantic Ocean. It is comprised of 4 major oceanic basins: the Canada Basin, the Markarov Basin, the Amundsen Basin and the Nansen Basin.

In stark contrast to the North Water, the Beaufort Sea is a highly stratified, oligotrophic marine system. The Beaufort Sea is made up of waters from various sources with inputs of Atlantic-origin through the Fram Strait and the Barents Sea, inputs of Pacific-origin through the Bering Strait to the Chukchi Sea shelf, and freshwater inputs from the surrounding continents of Eurasia and North America (Carmack, 2000; Carmack *et al.*, 2016). Nutrient-rich Pacific and Atlantic-origin waters flowing into the Arctic are more saline and therefore more dense than surface waters that have been diluted from freshwater input and ice melt. These saltier in-flowing waters sink below the surface waters making primary production in this region dependant on upwelling or mixing of the nutrient-rich waters up into the freshwater-influenced euphotic zone (Carmack *et al.*, 2016). The Beaufort Sea is fed primarily by the inflowing Mackenzie River and the brackish Alaska Coastal Current that contains outflow from the Yukon River and numerous smaller rivers (Okkonen *et al.*, 2009; Weingartner *et al.*, 1992; Williams and Carmack, 2015).

The surface waters of the Beaufort Sea known as the polar mixed layer (~ first 50 m) are composed of low salinity waters that are influenced by runoff, sea-ice formation and melt, and surface exchange processes. This is followed by a complex halocline structure (from 50 m - 250 m) made up of two Pacific water-derived haloclines (Bering Sea Summer Waters at 32.4 PSU and Bering Sea Winter Waters at 33.1 PSU) and an Atlantic water-derived halocline (34.4 PSU). Below this lies a more saline (34.4-34.9 PSU), 400 m - 600 m thick Atlantic water layer followed by saltier and warmer Deep Arctic Waters (Guéguen *et al.*, 2012; Rudels *et al.*, 1991).

The surface of the Beaufort Sea bacterioplankton community is composed primarily of Alpha-proteobacteria, specifically SAR11 and Rhodobacterales, Gamma-proteobacteria and, to a lesser extent, Bacteroidetes, particularly Flavobacteriales (Kirchman *et al.*, 2010; Han *et al.*, 2014; Lovejoy *et al.*, 2011). Although not as abundant as bacteria in the surface waters, the archaeal surface water community is composed primarily of MGII Euryarchaeota followed by MGI Thuamarchaeota (Galand *et al.*, 2006). However, the relative abundance of bacterioplankton decreases with depth, while the relative abundance of MGI Thaumarchaeota increases with depth and can represent up to 40% of the microbial community in the deep Beaufort Sea (Kirchman *et al.*, 2007). The majority of the deep Beaufort Sea bacterioplankton community is composed of Alpha-proteobacteria, primarily SAR11, Delta-proteobacteria composed primarily of SAR324, a diversity of Gamma-proteobacteria including Arctic96BD-19, and Chloroflexi, composed primarily of the SAR202 cluster (Galand *et al.*, 2010). In terms of photosynthetic organisms, the Beaufort Sea, and the Arctic Ocean at large, lack representation from Prochlorococcus and Synechococcus cyanobacteria whose roles are instead performed by picoeukaryotic algae (Li *et al.*, 2009; Pedrós-Alió *et al.*, 2015) including Micromonas.

Time series experiments have revealed changes in microbial community composition in the Beaufort Sea in correlation with changing physicochemical properties of the water column. For instance, a 9-year 16S rRNA gene time series analysis showed a marked increase in surface water SAR11 and a decrease in Bacteroidetes, which are typically associated with more productive ecosystems as a result of record-breaking sea ice melt in 2007 (Comeau *et al.*, 2011). Similar increases in the ratios of picoeukaryotes to nanoeukaryotes in response to fresher Beaufort Sea surface waters were observed in a 9-year time series from 2004 to 2012 (Li *et al.*, 2013).

Although several explorations of the Beaufort Sea microbial communities composition have been performed, there has been little work on the metabolic potential and expressed metabolism of these communities. Previous studies have focused on identifying types of biomolecules that can by used by the communities (Kirchman *et al.*, 2007; Sala *et al.*, 2008) or specific metabolisms like nitrogen assimilation, nitrification and dark carbon fixation (Alonso-Sáez *et al.*, 2012, 2010; Connelly *et al.*, 2014; Kirchman *et al.*, 2007).

1.4.3.1 DOM in the Beaufort Sea

The Arctic Ocean is of particular interest in terms of carbon cycling due to its naturally high loads of terrestrial organic matter input, and the relatively high rates of surface water warming and freshening, impacting its ecosystem in terms of dissolved organic matter concentrations, nutrient availability, primary production, and microbial community structure and function.

Rivers draining into the Arctic Ocean make up ~10% of global river flux and have high DOC loads (Anderson, 2002). Estimates of the annual flux of DOC to the Arctic Ocean from 5 major Arctic rivers (Yenisey, Lena, Ob', Mackenzie, and Yukon rivers) report fluxes ~2.5 times greater than temperate rivers with similar watershed sizes and water discharge (Raymond *et al.*, 2007). High concentrations of lignin oxidation products and a depletion of ^{13}C in ultrafiltered dissolved organic matter (UDOM) throughout the surface Arctic Ocean also indicate that terrigenous UDOM accounts for a much greater fraction of the UDOM in the surface Arctic (5–33 %) waters than in the Pacific and Atlantic oceans (0.7–2.4 %) (Opsahl *et al.*, 1999). The dominant source of tDOM to the coastal western Arctic is the Mackenzie River (Gordeev, 2006). The Mackenzie River supplies the Beaufort Sea with approximately $1.4x10^9$ kgs of DOC annually (Raymond *et al.*, 2007). This tDOM is long-lived in the ocean margins, but less than 50 % of tDOC entering the Arctic Ocean survives export to the North Atlantic (Amon, 2003; Hansell *et al.*, 2004). The major source of tDOM in the Arctic is gymnosperm vegetation (Opsahl *et al.*, 1999) and concentrations of tDOM made of lignin-derived phenols decrease from the Mackenzie River plume to the Canadian Archipelago (Walker *et al.*, 2009). In spring, a pulse of freshwater that occurs due to melting of winter snow and ice accumulations contributes to >50 % of annual freshwater river discharge within a 2-week period on the Alaskan Arctic coast, (McClelland *et al.*, 2014). This discharge is accompanied by a large input of tDOM and inorganic nutrients (McClelland *et al.*, 2014), thereby exposing coastal Arctic microbial communities to concentrated pulses of tDOM annually. The tDOM supplied by these Arctic rivers were once thought to be highly refractory (Opsahl *et al.*, 1999; Dittmar and Kattner, 2003; Amon and Meon, 2004). However, several studies have since found that a portion (7 – 40 %) of this riverine tDOM is bioavailable on time scales of weeks to months (Hansell *et al.*, 2004; Holmes *et al.*, 2008; Mann *et al.*, 2012; Vonk *et al.*, 2012; Sipler *et al.*, 2017).

1.5 How the chapters are linked

This thesis work focuses on the microbial community composition, metabolic processes and dynamics of northern marine ecosystems. The work aims to relate microbial community composition to metabolic processes and to further link both of those to the availability and composition of organic and inorganic nutrients in the environment. The main unifying objectives of the following chapters are to resolve the community composition of relatively underexplored marine systems and to gain insight into the metabolic processes important to those systems. These data can then be used to generate and test hypotheses that would otherwise be impossible in these underexplored marine systems. In addition, these data represent some of the first describing the community metabolism of the Arctic Ocean and can be used as refrence when studying changes over time.

In order to first explore how stratification and terrestrial-derived water can influence the microbial community structure and function of marine ecosystems, as well as to develop a meta-omic pipeline for future studies, the microbial community structure and metabolism of the water column at one geographic location in the highly stratified LSLE was investigated using a joint metagenomic-metaproteomic analysis (Chapter 2). Just like the LSLE, the Beaufort Sea is a highly stratified system that is influenced by an increasing amount of terrestrial-derived freshwater making it more and more brackish. The fresher surface waters overlay deeper, more saline waters that originate from the Atlantic and Pacific Ocean. Given the similarities, I utilized my investigation into the microbial community structure and metabolism of the LSLE to develop both molecular and bioinformatic techniques that could then be used in the Arctic Ocean.

The metaproteomic techniques developed in the LSLE were then applied to perform a comparative analysis of the Eastern (Greenland) and Western (Canadian) sides of the North Water (Chapter 4). The North Water is a highly productive marine region influenced by two main sources of water; the Atlantic and Arctic Oceans. The Eastern side of the North Water is influenced by Atlantic Ocean waters while the Western side is influenced by Arctic Ocean waters. As the first metaproteomic study performed in an Arctic ecosystem, my goal was to explore the microbial metabolic processes identified in these regions, as well as the taxa responsible for those processes, and to determine whether some were more relevant to a specific side of the North Water. One of the main limitations of metaproteomic analysis is its reliance on a representative searchable protein database for protein identification. At the time of this study no Arctic Ocean metagenomic sequence data was available to build a more accurate searchable database. In 2015, as part of the Joint Ocean Ice Study, samples from a latitudinal transect from coastal waters to more open ocean waters at three depths within the Beaufort Sea were collected and used to generate the first metagenomic dataset of the Beaufort Sea (Chapter 5). The data was used to construct 360 metagenome assembled genomes. An abundance of Chloroflexi MAGs (which are typically associated with the deep ocean) were identified in the FDOM maximum of the Beaufort Sea and at least one MAG which was specialized in the degradation of aromatic terrestrial-derived dissolved organic matter appeared to have acquired this capacity, at least in part, from lateral gene transfer events from terrestrial organisms. These findings highlight the potential effects increasing terrestrial runoff could have on the Arctic Ocean.

Chapter 2: Metaproteomics of aquatic microbial communities in a deep and stratified estuary

Colatriano D, Ramachandran A, Yergeau E, Maranger R, Gélinas Y, Walsh DA. (2015). Metaproteomics of aquatic microbial communities in a deep and stratified estuary. *Proteomics* **15**:1-14.

2.1 Abstract

Here we harnessed the power of metaproteomics to assess the metabolic diversity and function of stratified aquatic microbial communities in the deep and expansive Lower St. Lawrence Estuary, located in eastern Canada. Vertical profiling of the microbial communities through the stratified water column revealed differences in metabolic lifestyles and in carbon and nitrogen processing pathways. In productive surface waters, we identified heterotrophic populations involved in the processing of high and low molecular weight organic matter from both terrestrial (e.g. cellulose and xylose) and marine (e.g. organic compatible osmolytes) sources. In the less productive deep waters, chemosynthetic production coupled to nitrification by MG-I Thaumarchaeota and Nitrospina appeared to be a dominant metabolic strategy. Similar to other studies of the coastal ocean, we identified methanol oxidation proteins originating from the common OM43 marine clade. However, we also identified a novel lineage of methanol-oxidizers specifically in the particle rich bottom (i.e. nepheloid) layer. Membrane transport proteins assigned to the uncultivated MG-II Euryarchaeota were also specifically detected in the nepheloid layer. In total, these results revealed strong vertical structure of microbial taxa and metabolic activities, as well as the presence of specific "nepheloid" taxa that may contribute significantly to coastal ocean nutrient cycling.

2.2 Statement of significance of the study

Aquatic microbial communities are diverse and critical to ecosystem function. While genomics and metagenomics has revolutionized our understanding of the metabolic diversity and versatility of microbial communities, these technologies only provide insight into the functional potential of microbes rather than *in situ* functional activities. Metaproteomics, that is the identification of expressed proteins and metabolic pathways operating in a microbial community, is one promising approach that moves beyond functional potential and towards functional activity. In this study, we used metaproteomics to assess the metabolic diversity of microbial communities inhabiting the stratified waters of a large and deep estuary. The results provide insights into metabolic pathway expression patterns, and hence function, of these poorly described microbial communities with respect to carbon and nitrogen cycling in aquatic ecosystems.

2.3 Introduction

Marine microbial communities are diverse and their metabolism is central to global biogeochemical cycles and food web structure (Falkowski *et al.*, 2008; Acinas *et al.*, 2004; Thompson *et al.*, 2005). Recently, genomic and metagenomic studies have provided considerable insight into the metabolic diversity of marine bacteria and archaea and the discovery of sometimes novel and ubiquitous metabolic pathways. For example, Beja et al. (Beja *et al.*, 2000) reported the presence of genes for phototrophy in marine bacteria that were previously presumed to rely exclusively on the oxidation of organic matter for energy. More recently, novel chemolithotrophic energy-generating pathways were reported for marine microbes, including hydrogen sulfide (H_2S) and hydrogen (H_2) oxidation pathways in marine bacteria (DA Walsh *et al.*, 2009; Anantharaman *et al.*, 2013), as well as ammonia oxidation (NH_3) in marine archaea (Walker *et al.*, 2010; Hallam *et al.*, 2006) . These studies and others have not only revealed novel metabolic capabilities, but have also shown that many microbial taxa are metabolically versatile, possessing the genetic information to subsist on numerous organic carbon compounds, perform autotrophic carbon fixation, and/or use inorganic compounds as energy source (Newton *et al.*, 2010). Such discoveries have challenged our understanding of the role these important microbes play in marine ecosystems. Meeting this challenge will require a thorough investigation of how the metabolic activities of microbes vary in response to environmental and biotic conditions in the ocean.

One useful approach for assessing the metabolic activities of marine microbes in relation to their environment is through metaproteomic analysis, which has the potential to provide information on *in situ* protein expression for whole microbial communities (VerBerkmoes and Denef, 2009). In metaproteomics, samples are obtained directly from the environment, proteins are extracted from the biomass and subjected to tandem mass spectrometry (MS) and resulting spectra are searched against a protein sequence database to deduce the amino acid sequences of peptides, providing the identification of proteins. Since its first application in the ocean (Kan *et al.*, 2005), metaproteomics has been successfully applied to a range of marine environments. Seasonal studies performed in the coastal waters of the Northwest Atlantic Ocean and the Antarctic Peninsula have shown temporal structure in microbial metabolism, including the importance of chemoautotrophic metabolism in the winter months, and a shift to a heterotrophic community as waters warm (Williams *et al.*, 2012; Georges *et al.*, 2014). Similarly, Teeling et al. used metaproteomics to investigate the metabolic diversity of bacteria over the course of a phytoplankton bloom in the North Sea (Teeling *et al.*, 2012). Other studies have reported on the variation in protein expression patterns across spatial scales, including along a geographical transect from a highly productive coastal system to a low nutrient ocean gyre (Morris *et al.*, 2010).

In the study presented here, we integrated metagenomic and metaproteomic analyses to elucidate protein expression patterns of microbial communities residing in the stratified waters of the Lower St. Lawrence Estuary (LSLE) located in eastern Canada. The waters of the LSLE contain strong geographic and vertical gradients in salinity, nutrient concentration, organic matter composition, pH and oxygen (Gobeil, 2006; Gilbert *et al.*, 2005; Tremblay and Gagné, 2009; Lehmann *et al.*, 2009; Mucci *et al.*, 2011). In this study, we focus on the structure and functional potential of microbial communities inhabiting distinctive strata with contrasting environmental characteristics.

2.4 Materials and Methods

2.4.1 Sampling and metadata collection.

Seawater samples were collected from Station 21 (49°05.60'N/67°17.00'W) in the LSLE using a CTD rosette during an oceanographic cruise in May 2011. For this study, water samples (10 L) for metagenomics and metaproteomics were collected from the surface layer (3 m), the deep layer (180 m) and the bottom nepheloid layer (314 m). Water was processed by prefiltering through a Whatman GF/D filter (2.7-µm cutoff) via peristaltic pump and cells were collected in-line on a 0.22-µm Sterivex-GV filter. After filtration, 1.8 mL of sucrose lysis buffer (40 mM EDTA pH 8.0, 50 mM Tris pH 8.3, 0.75 M sucrose) was added and filters were stored at -80°C until DNA or protein extraction in June 2012.

Bacterial production was measured using the ^{3}H-leucine incorporation method (Smith and Azam, 1992). Water samples (1.2 mL) were dispensed, in triplicate, into clean 2 mL microcentrifuge tubes preloaded with 50 µL ^{3}H-leucine (115.4 Ci mmol-1, Amersham) to produce a final leucine concentration of 40 nM. Samples were incubated in the dark in an ice-filled isothermic container at *in situ* temperature for approximately 1 h. Leucine incorporated into cell protein was collected after precipitation by trichloroacetic acid (TCA) and centrifugation. Tubes were filled with 1.25 mL liquid scintillation cocktail (ScintiVerse, Fisher Scientific), and radioactivity was measured using Perkin Elmer Liquid Scintillation Analyzer. Rates of leucine incorporation were corrected for radioactivity adsorption using TCA-killed controls and converted to bacterial C production (BP) using 3.1 kg C per mol-1 ^{3}H-leucine (Simon and Azam, 1989). Bacterial abundance was measured by flow cytometry. Cells were stained with SYBR Gold and counted using a FACScan flow cytometer (Becton Dickinson, Mountain View, Calif.), equipped with a 15-mW, 488-nm, air-cooled argon-ion laser, and a 70-um nozzle.

2.4.2 Elemental and isotopic analyses

A known volume of water from the rosette was vacuum filtered over combusted (500 ºC, 4 hrs) pre-weighed GF/F filters (nominal pore-size 0.7 µm). Filters capturing suspended particulate matter (SPM) were stored at -20 ºC until time of analysis. Particulate-bearing filters were lyophilized and the dry filters were weighed to determine total SPM mass, sub-sampled with a hole-puncher and placed in Ag capsules to obtain 20-30 mg of SPM. SPM were de-carbonated by exposure to HCl fumes in a sealed container for 10-12 hrs, placed in an oven at 55 ºC for ~1 hour to remove traces of acid, and then folded and wrapped in a tin capsule. Elemental (C, N) and stable isotopic composition ($\delta^{13}C$) were determined on a GV Instruments (now Elementar) EuroVector IsoPrime EA-IRMS. Isotope data was calibrated with and corrected for daily instrumental drift with certified sucrose (IAEA-CH-6, δ13C = -10.47 ‰) and in-house β-alanine (SigmaAldrich, δ13C = -26.0 ‰) standards. Elemental composition is reported in terms of as mg L^{-1} for SPM as well as particulate organic carbon (POC) and particulate total nitrogen (PTN). Isotopes are reported using the V-PDB convention against reference gas of known isotope composition.

2.4.3 Metagenomics

A single metagenome was generated for each of the three sampling depths. DNA was extracted from Sterivex filters as described in (Zaikova *et al.*, 2010). Total DNA was sheared, size selected and subjected to library construction using an Ion Plus Fragment Library Kit. Samples were pooled and a total of 3.5×10^7 molecules were used in an emulsion PCR using an Ion OneTouch 200 template kit (Life Technologies) and OneTouch and OneTouch ES instruments (Life Technologies, Carlsbad, CA) according to the manufacturer's protocol. The sequencing of the pooled samples was performed using an Ion Torrent personal genome machine (PGM) system and a 316 chip with the Ion Sequencing 200 kit according to the manufacturer's protocol. All DNA sequence reads that were less than 100 bp or had an average quality < Q17 were removed prior to further analysis. Identification of putative protein-encoding genes was performed using FragGeneScan (Rho *et al.*, 2010) and allowing 1% sequencing error rate. The 16S rRNA gene sequences in the metagenomics were identified by blastn searching against the Silva rRNA database. The identified 16S rRNA sequences were assigned to taxonomic groups using Mothur and the Wang method with a bootstrap value cutoff of 60% (Wang *et al.*, 2007). Metagenomic data was deposited at the European Nucleotide Archive under study accession number PRJEB9523.

2.4.4 Metaproteomics

Protein extraction from Sterivex filters and in-gel trypsin digestion was performed as described in Georges et al. 2014. LC-MS was performed using an LTQ Orbitrap Velos (Thermo Fisher Scientific) mass spectrometer and digested peptides were previously separated using a EASY-nLC (Proxeon) nanoHPLC. Samples were directly injected onto a nano column (75 µm inner diameter by 15 cm) containing C18 media (Jupiter 5μ m, 300A, Phenomenex). Separation was achieved by applying a 1 to 27% acetonitrile gradient in 0.1% formic acid over 80 minutes at 400 nl / min then followed by a linear gradient of 27% to 52% over 20 minutes at 400 nl / min. Electrospray ionization was enabled through applying a voltage of 2.0 kV through a PEEK junction at the inlet of the nano column. operated in positive mode with data-dependent acquisition mode. A survey scan was obtained with the Orbitrap mass detector at m/z of 360 to 2000 with a resolution of 60 000 at m/z 400. The automatic gain control settings were 3 x 10^6 ions and 2.5 x 10^5 ions for survey scan. Ions were selected for fragmentation on the LTQ linear trap when their intensity was greater than 500 counts. The 10 most intense ions were isolated for ion trap CID with charge states of 2 to 4 and sequentially isolated for fragmentation within the linear ion trap using collision induced dissociation energy set at 35% with q activation set at 0.25 and an activation time of 10ms at a target value of 30,000 ions. Ions selected for MS/MS were dynamically excluded for 30 s. Raw MS/MS data files are available from the authors upon request.

To identify peptide sequences, MS/MS spectra were compared with a protein database consisting of translated predicted protein-coding sequences from 205 marine and archaeal genomes (Georges *et al.*, 2014), the three LSLE metagenomes, as well as metagenomic and metatranscriptomic data previously generated from stratified waters off the coast of Chile (Canfield *et al.*, 2010; Ganesh *et al.*, 2014). All MS/MS spectra were searched with the SEQUEST (Eng *et al.*, 1994) algorithm implemented in Proteome Discoverer (Thermo Scientific) with the following settings: enzyme type, trypsin; maximum missed cleavage sites, 2; maximum modification for peptide, 4; static modifications, +57 Da (iodoacetamide modification of cysteine) and +16 Da (oxidation of methionine). Percolator, a supervised machine learning tool to discriminate between correct and decoy spectrum (Käll *et al.*, 2007) was employed to increase the rate and confidence of peptide identification. Percolator settings were as follows: maximum delta Cn, 0.05; Target False Discover Rate (strict), 0.01; and validation was based on q-values. Using these stringent settings we include peptides identified only once in a sample. In Proteome Discoverer, we used the “Enable Protein Group” setting to group proteins that share sets of identified peptides (i.e. redundant proteins) and select the highest scoring protein in the group to serve as the representative protein (i.e. master protein).

For taxonomic assignment of proteins, each master protein was searched against the RefSeq (v52) protein database using BLASTP and the top 10 hits were reported in the output. The BLASTP search results were loaded into MEGAN and taxonomic assignment was performed using the lowest common ancestor (LCA) algorithm using a minimum bit score equal to 80 (Huson *et al.*, 2011). The master proteins were also queried against the Clusters of Orthologous Genes (COG) database to identify a probable function. Only proteins with high sequence similarity (E-value cut-off < $1x10^{-10}$) to a known COG were annotated with the corresponding function. The COG annotation and the RefSeq analysis were both used to assign function to the proteins. Protein expression levels were estimated using the spectral counts method by summing all the spectra for the collection of peptides matching each protein (VerBerkmoes and Denef, 2009).

2.5 Results and discussion

2.5.1 Study area and environmental conditions

During an oceanographic cruise in May 2011, we assessed the vertical structure of the water column and collected samples for metagenomic and metaproteomic analyses from a station located at the seaward (Station 21, S21) end of the LSLE (**Figure 2.1a**). The typical three-layer structure that defines the LSLE was apparent, with a warm and brackish surface layer sitting on top of a cold intermediate layer and warmer, hypoxic bottom water (**Figure 2.1b**). To investigate vertical structure and activity of microbial communities we collected three samples from the S21 water column, including the surface layer (S21-3 m), the deep layer (S21-180 m), and the hypoxic particle rich bottom nepheloid layer (S21-314 m). Differences in the composition of the organic matter (OM) were apparent along the depth profile. For example, the concentration of solid particulate matter (SPM) was highest at the bottom, an indication that we did indeed sample the particle-rich nepheloid layer (**Table 2.1**). However, particular organic carbon and nitrogen (POC and PON, respectively) were twice as high at the surface than at the bottom, while the bottom water POC was additionally more depleted in ^{13}C than surface POC. These observations indicate that the organic matter at the surface was more labile and accessible to heterotrophic bacteria than that in the bottom waters. As such, bacterial abundance decreased nearly two-fold (from 3.6 x 10^5 cells/ml to 2 x 10^5 cells/mL) (**Table 2.1**) with depth. Bacterial productivity measured using ^{3}H-leucine incorporation decreased by nearly two orders of magnitude (from 26 µg C/L/day to 0.7 x µg C/L/day) (**Table 2.1**), revealing a strong vertical gradient in the heterotrophic activity of the microbial community.

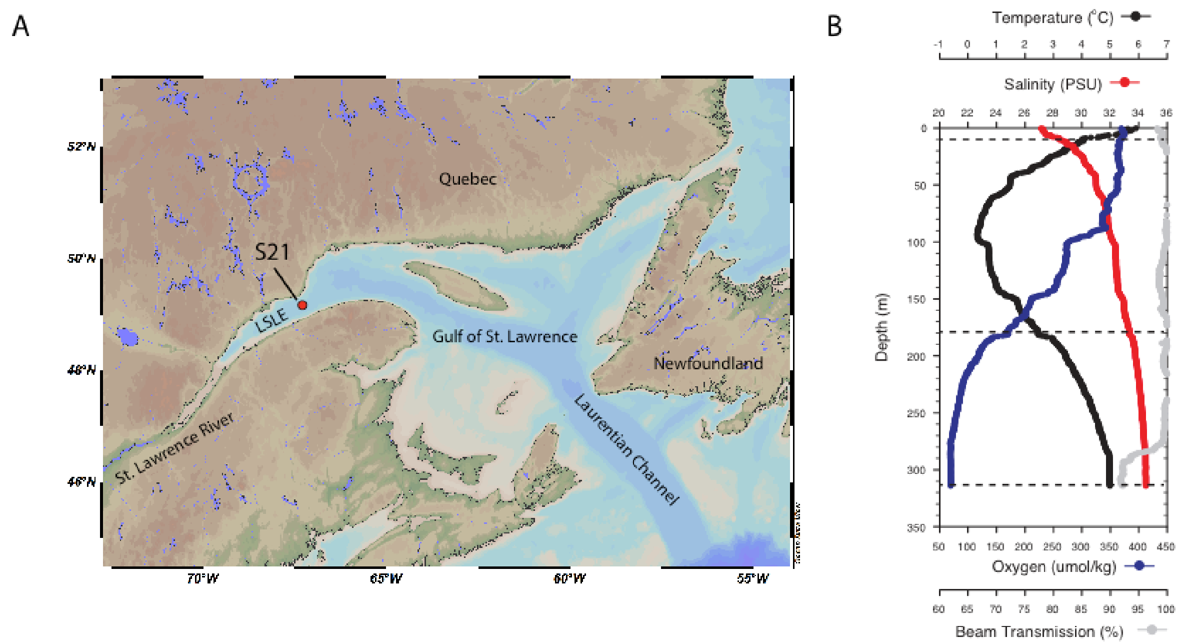


Figure 2.1. (A) Sampling station for this study (S21) in the Lower St. Lawrence Estuary and (**B**) the physiochemical conditions of S21 plotted against depth.

Table 2.1. Metadata from the LSLE

Sample	Cell abundance (x10^5 ml^{-1})	BP (ug C/l/day)	SPM (mg/L)	POC (mg/L)	13C-POC (‰ vs. VPBD)	PON (mg/L)
S21-003m	3.6	26	2.69	0.215	-21.55	0.0238
S21-180m	2.6	1.68	n.d.	n.d	n.d	n.d.
S21-314m	2.0	0.67	6.27	0.110	-23.48	0.0106

2.5.2 Coupled metagenomic-metaproteomic profiling

We combined metagenomics with MS/MS-based proteomics to examine the genomic diversity and protein expression patterns of microbial communities at S21 in the LSLE. In total, we generated 441 Mb of metagenomic DNA sequence data, which contained over 1.2 million protein-encoding genes (**Table 2.2**). These genes, as well as protein-encoding genes from microbial reference genomes and additional Chilean coastal ocean metagenomes and metatranscriptomes, formed the protein sequence database used for identification of peptides and proteins from the MS/MS data. Searching of peptide spectra against this database resulted in the identification of 8,404 distinct proteins. The inclusion of metagenomic data from the LSLE and Chile resulted in a significant increase in the number of spectra assigned to peptides, compared to the use of reference genomes alone. This was particularly true for the bottom waters; 75% of spectra from 3 m were identified from reference genomes, while only 26% of spectra from 314 m were assigned to reference genome proteins, highlighting the bias in reference genome databases towards surface lineages (**Table 2.2**). In contrast, over half of the 314 m spectra were assigned to proteins from coastal Chile, while an additional 10% of spectra from each depth was identified using proteins from the LSLE metagenomes. The higher percent of spectra assigned to Chilean proteins compared to LSLE is likely a result of the LSLE metagenomes being comprised of fewer sequences that the Chilean data. This combined metagenomic and metaproteomic dataset provided the basis for taxonomic and functional analysis of the microbial communities inhabiting the LSLE water column (**Supplementary Tables 2.1 and 2.2**).

Table 2.2. Summary of meta-omic data

	S21-003m	S21-180m	S21-314m
Metagenomes			
No. of metagenomic reads	275,842	246,721	815,722
No. of protein-encoding ORFs	239,542	224,923	735,891
Metaproteomes			
Total no. of spectra	198,837	103,459	111,594
Total no. of spectra assigned to proteins (% assigned)	30,809 (16)	22,758 (22)	26,407 (24)
% of identified spectra assigned to reference genomes	74.9	35.5	25.6
% of identified spectra assigned to SAGs	0.7	7.5	10.5
% of identified spectra assigned to Chilean metagenome	10.3	32.8	38
% of identified spectra assigned to Chilean metatranscriptome	4.5	13.2	15.7
% of identified spectra assigned to SLE metagenome	9.5	11	10.2
No. of master proteins	2282	2421	2522

2.5.3 Taxonomic stratification of microbial lineages and expressed proteins

Vertical structure in the composition of LSLE community was apparent from both the metagenomic and metaproteomic profiles. Based on the relative abundance of 16S rRNA gene sequences extracted from the metagenomes, it was clear that the microbial taxa inhabiting the productive and brackish surface layer (3 m) were distinct compared to microbial taxa located in deeper marine waters (180 m and 314 m) (**Figure 2.2**). Most striking was the dominance of 16S rRNA genes from the Flavobacteriales and the Rhodobacterales orders of the Alpha-proteobacteria. These lineages are often associated with productive coastal systems where they are involved in decomposition of phytoplankton-derived organic matter (El-Swais *et al.*, 2015; Gilbert *et al.*, 2012; Fernández-Gómez *et al.*, 2013; Gómez-Pereira *et al.*, 2010; Kirchman, 2002). Moreover, Flavobacteriales and Rhodobacterales proteins also appeared to make up the majority of peptide spectra in the surface layer, indicating a relatively high contribution to metabolic processes occurring at the surface. Proteins assigned to Flavobacteriales and Rhodobacterales were present in both the deep water (180 m) and the nepheloid layer (314 m). Although this may reflect extension of their ecological niches into the deep water, it may also be a result of sinking microbial cells from the productive surface layer.

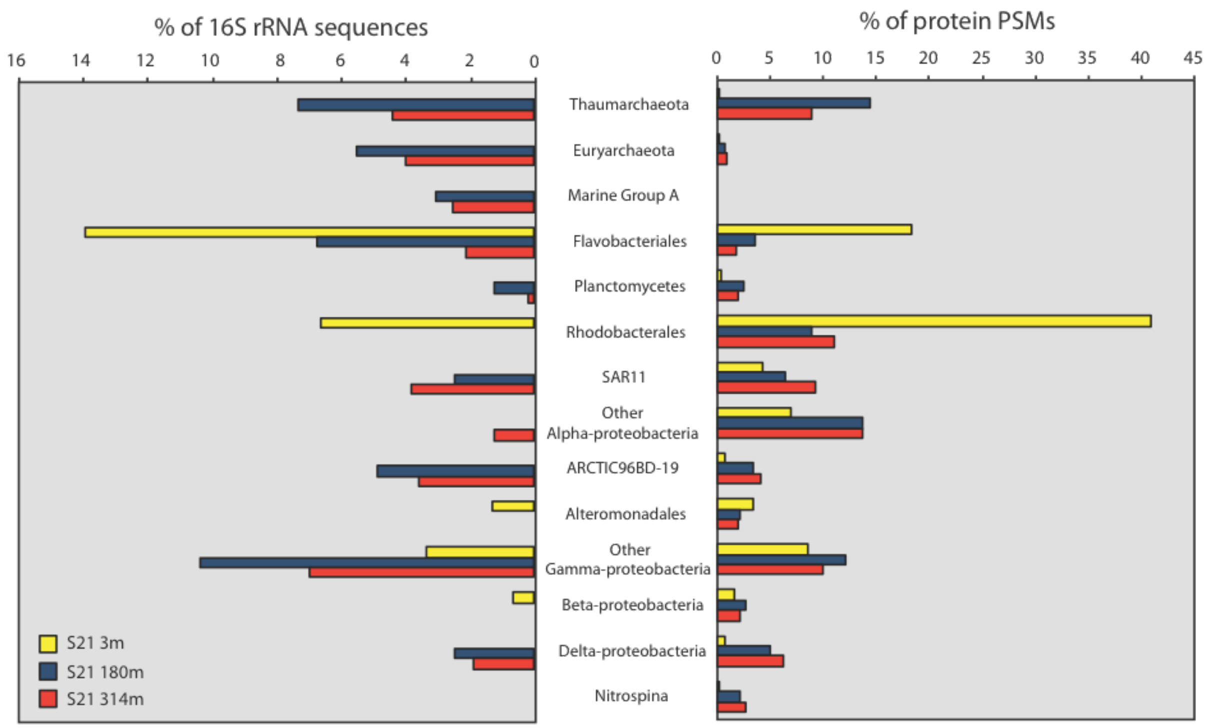


Figure 2.2. Taxonomic composition of the LSLE microbial community based on 16S rRNA genesequences (**left)** and expressed proteins (**right**).

Lineages within the Beta- and Gamma-proteobacteria were also present in the metagenomic and metaproteomic data from the surface layer. These included the

Alteromonadales which share a similar copiotrophic lifestyle to the Flavobacteriales and Rhodobacterales and are also common heterotrophs associated with productive waters (Ivars-Martinez *et al.*, 2008). In addition, the OM43 clade of Beta-proteobacteria was detected at relatively low abundance. The OM43 clade is comprised of known methylotrophs that are common to the coastal ecosystems and capable of subsisting on one-carbon (C1) compounds such as methanol (Giovannoni *et al.*, 2008).

The taxonomic composition of the metagenomes and metaproteomes from the deep water was unique compared to the surface layer and numerous lineages were only detected in metagenomes and metaproteomes from below the surface. These include 16S rRNA sequences and expressed proteins assigned to known chemolithoautotrophs such as the ammonia-oxidizing Marine Group I (MG-I) Thaumarchaeota (Hallam *et al.*, 2006; Karner *et al.*, 2001) and the nitrite-oxidizing Nitrospina group (Lücker *et al.*, 2013). Additional lineages that were enriched in the deep waters included the SAR11 clade of Alpha-proteobacteria, the ARCTIC96BD-19 clade of Gamma-proteobacteria and the SAR324 clade of Delta-proteobacteria. Recently, genome analyses of the ARCTIC96BD-19 and SAR324 bacteria demonstrated that these lineages are metabolically versatile and likely possess a mixotrophic lifestyle (DA Walsh *et al.*, 2009; Swan *et al.*, 2011; Sheik *et al.*, 2013; Mattes *et al.*, 2013). Both lineages possess genes for the membrane transport of sugars and amino acids, autotrophic CO_2 fixation, and the oxidation of reduced sulfur compounds (DA Walsh *et al.*, 2009; Swan *et al.*, 2011; Sheik *et al.*, 2013; Mattes *et al.*, 2013). Mixotrophy may be a widespread adaptation to survival in the bottom waters of the LSLE where the availability of labile organic carbon compounds may be limiting. However, we did not detect any sulfur oxidation proteins or autotrophic CO_2 fixation pathways that could be assigned to the ARCTIC96BD-19 or SAR324 clades.

Although the taxonomic composition of the metagenomic and metaproteomic profiles agreed for the most part, there were some marked differences. For example, the Marine Group A (MGA) bacteria and the Marine Group II (MG-II) Euryarchaeota were significant components of the deep-water metagenomes, yet were either barely detected (i.e. MG-II) or not detected at all (i.e. MGA) in the metaproteomes. This observation highlights the current limitations of metaproteomics studies. Until very recently, no MGA genomic data was publically available for use in protein identification. Moreover, although we would expect MGA representation in the LSLE metagenomes, without MGA reference genomes it is difficult to assign metagenomic sequences to this lineage. These limitations are set to decrease as novel cultivation (Santoro *et al.*, 2015) and single cell genomics technologies (Rinke *et al.*, 2013; Hedlund *et al.*, 2014; Stepanauskas, 2012; Swan *et al.*, 2013) provide more and more genome sequence data for poorly represented marine lineages.

2.5.4 Functional stratification of expressed proteins

To investigate the metabolic diversity of stratified microbial communities in the LSLE, we assigned the 8,404 proteins identified in the metaproteomes to metabolic functions using the Clusters of Orthologous Genes (COG) database (Tatusov *et al.*, 2000). Using a stringent protein sequence similarity cutoff (expect value < $1x10^{-10}$), 4,515 proteins were assigned to a total of 614 COG functions. We then mapped the presence/absence patterns of COG functions across the three sample depths (**Figure 2.3**). Approximately 27% (185) of the COG functions were ubiquitous throughout the water column, and proteins assigned to these common COGs also made up the vast majority of peptide spectra. Proteins in this "core" set of COGs included many essential housekeeping proteins such as DNA-directed RNA polymerase subunits, ribosomal proteins, translation elongation factors (e.g. EF-Tu), chaperonins (e.g. HSP60), and ATP synthase subunits, as well as a wide variety of membrane transporters. However, metabolic partitioning of the surface assemblage from the deep assemblages was apparent from the distribution of COG functions. Some 161 functions were only detected at the surface. In contrast to the surface, only 53 and 65 functions were unique to 180 m and 314 m, respectively. Moreover the metabolic similarity between the 180 m and 314 m communities was shown by the detection of 71 common functions, compared to only 33-46 functions in common with the 3 m proteins. Many of the proteins that distinguished the surface from the deep water were from the archaea, owing to the abundant MG-I Thaumarchaeota at depth.

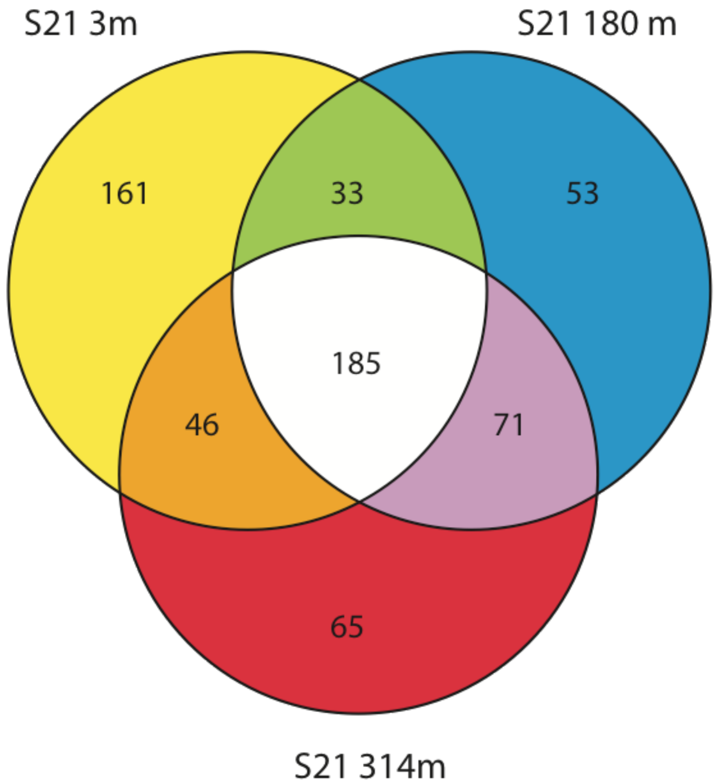


Figure 2.3. Venn diagram depicting the presence/absence pattern across the three sample depths.

Among the COGs uniquely identified at the surface was the ribulose bisphosphate carboxylase (RuBisCo) small subunit (RbcS), an essential component of the Calvin-Benson-Bassham cycle of CO_2 fixation. As expected, one RbcS subunit was most similar to a homolog from Ostreococcus, a photosynthetic eukaryotic member of the picoplankton. A second RbcS subunit was most similar to a Rhodopseudomonas palustris (Alpha-proteobacteria) homolog, suggesting that CO_2 fixation by photoautotrophic bacteria may also contribute to primary production in LSLE surface waters. The RuBisCo large subunit (RbcL) was identified at the surface, but also surprisingly in the bottom nepheloid layer. These RbcL proteins were most similar to Ostreococcus, Rhodopseudomonas and Micromonas, which is another photosynthetic eukaryotic picoplankton. Since the RbcL proteins identified in the nepheloid layer were the same as detected at the surface, their presence in the bottom of the LSLE is most likely a result of sinking cells, rather than an active CO_2 fixing population. Interestingly, it has previously been suggested that RuBisCo proteins may serve as a unique tracer for the origin, accumulation, and transport of organic matter in the deep ocean (Orellana and Hansell, 2012).

2.5.5 Microbial processing of organic matter

The composition of organic matter in estuaries is complex and can originate from autochthonous production by phytoplankton and bacterioplankton or from allochthonous input of terrestrial organic matter from freshwater runoff (Gobeil, 2006; Goñi *et al.*, 2003). To provide insight into the processing of organic matter by microbial communities in the LSLE, we analyzed the abundance, distribution, and phylogenetic identity of membrane transporters specific to the uptake of organic substrates. As in other metaproteomic studies, some of the most prevalent proteins we identified were the solute-binding components of high affinity ATP-binding cassette (ABC) transporters and tripartite ATP-independent periplasmic (TRAP) transporters (Williams *et al.*, 2012; Georges *et al.*, 2014; Morris *et al.*, 2010; Sowell *et al.*, 2009, 2011). These proteins are implicated in the transport of a wide range of low molecular weight (LMW) organic compounds including sugars, amino acids, peptides, dicarboxylates and compatible osmolytes such as glycine betaine and taurine (**Figure 2.4**). Interestingly, all transporters were detected at all depths. However, the relative expression levels and the phylogenetic identity of the transport proteins varied, indicating a depth-dependent preference for certain compounds, as well as metabolic partitioning between phylogenetic groups. For example, the xylose ABC-type transporter was present at all depths, yet the number of peptide spectra was >10-fold higher in the surface layer compared to the deep layer. These transporters were almost all assigned to the Rhodobacterales. In addition, xylose isomerase, a key enzyme in xylose degradation, most similar to homologs from Rhodobacterales, was also only detected in the surface layer. Xylose is an abundant component of plant-derived hemicellulose, and taken together these results indicate an important role for the Rhodobacterales in the processing of terrestrial-derived organic matter entering the LSLE.

In addition to xylose utilization, Rhodobacterales in the surface layer also expressed a diversity of transporters for sugars, amino acids, peptides, and dicarboxylates, which is in agreement with this lineage's generalist role in the transformation of labile carbon compounds in the ocean (Gilbert *et al.*, 2012). In contrast to the Rhodobacterales, we did not detect any ABC-type or TRAP transporters that could be assigned to the Flavobacteriales, even though this lineage was one of the most abundant in the surface waters. This observation is in agreement with the known specialization of Flavobacteriales for high molecular weight (HMW) compounds. Comparative genome analyses have shown that marine Flavobacteriales have a reduced number of transporters for LMW compounds, which is compensated for by transporters for HMW compounds (Fernández-Gómez *et al.*, 2013). Indeed, we did identify TonB dependent outer membrane receptors assigned to Flavobacteriales (**Figure 2.4**). TonB transporters are implicated in the import of bulky organic compounds such as oligosaccharides into the periplasmic space (Eisenbeis *et al.*, 2008). Additional TonB transporters were assigned to the Alteromonadales, which are also implicated in degradation of HMW organic matter (McCarren *et al.*, 2010). We also identified Cellulase M proteins assigned to Flavobacteriales, further supporting their role in the degradation and utilization of HMW compounds in the LSLE surface layer.

Compared to the surface, transporters identified in the deep water were assigned to a more phylogenetically diverse group of bacteria and archaea. In addition to Rhodobacterales, we identified transporters most similar to the SAR11, SAR116, and BAL199 lineages of Alpha-proteobacteria, ARCTIC96BD-19 and other Gamma-proteobacteria, the SAR324 clade, and the MG-II Euryarchaeota. In addition to higher phylogenetic diversity of transporters, differences in the level of substrate specialization were apparent between phylogenetic lineages. For example, SAR324 and ARCTIC96BD-19 appeared to have a generalist lifestyle in the LSLE, capable of transporting numerous compounds, including sugars, peptides, amino acids, and dicarboxylates. In contrast, SAR116 appears to be much more specialized for the transport of amino acids, while BAL199 seems to exhibit a preference for compounds imported through the mannitol/chloroaromatic transporter. Proteins assigned to the MG-II amino acid transporters were specifically identified within the nepheloid layer, suggesting a preference for this particle rich environment. In support of this idea, Orsi et al. (Orsi *et al.*, 2015) showed that the MG-II indeed interact with particles in the ocean. Further investigation of the metabolic diversity of the particle-associated fraction of the nepheloid layer is warranted in order to better understand the ecophysiology of these deep water MG-II populations.

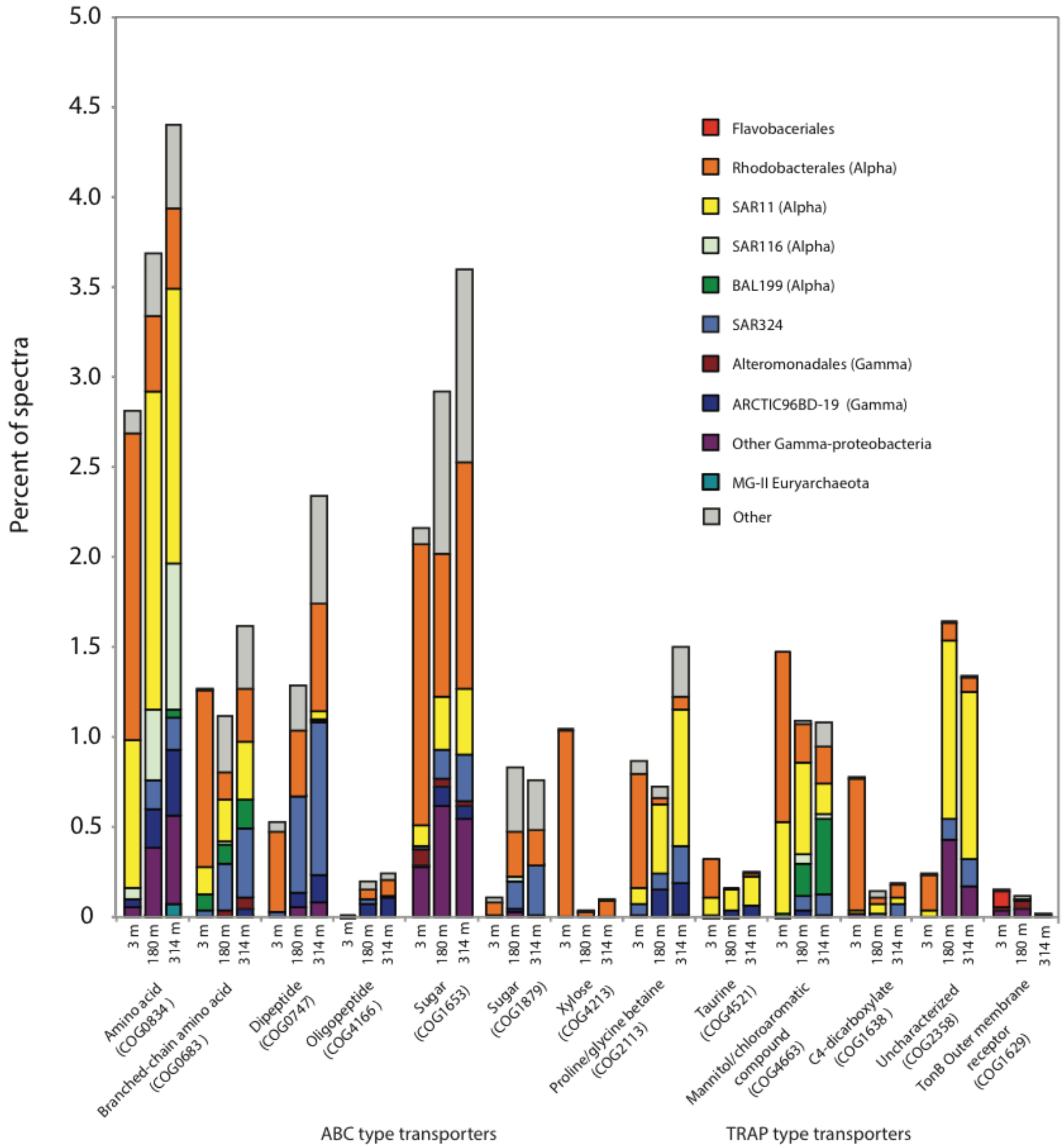


Figure 2.4. The relative abundance of solute transport proteins based on spectra counts identified at each of the three sample depths separated by taxa.

2.5.6 Complete nitrification linked to autotrophic CO_2 fixation

Nitrification is a critical step in the global nitrogen cycle and consists of the oxidation of ammonia to nitrite, followed by nitrite oxidation to nitrate. In the ocean, the MG-I Thaumarchaeota are known to couple ammonia oxidation to energy conservation and therefore play an important role in marine nitrification (Hallam *et al.*, 2006). In the deep water of the LSLE, proteins assigned to the MG-I Thaumarchaeota were among the most abundant we detected. In support of their previously described role in ammonia oxidation, expression of the beta subunit of the archaeal ammonia monooxygenase (AmoB) protein was identified in the bottom nepheloid layer. Archaeal ammonia oxidation can be coupled to autotrophic CO_2 fixation via the 3-hydroxypropionate/4 hydroxybutyrate pathway (Berg *et al.*, 2010). In the LSLE, two key enzymes of this pathway, acetyl-CoA carboxylase and 4-hydroxybutyryl CoA dehydratase were identified in the metaproteomes. Archaeal acetyl-CoA carboxylase was found exclusively at 314 m and the archaeal 4-hydroxybutyryl CoA dehydrates was identified at both 314 m and 180 m.

Much less studied in the ocean is the second step of nitrification, the transformation of nitrite to nitrate by nitrite-oxidizing bacteria (NOB). The key enzyme for nitrite oxidation is nitrite oxidoreductase (NxrABC), which is a member of the large DMSO reductase enzyme family (Jormakka *et al.*, 2004). In the LSLE, we detected NxrA- and NxrB-like proteins at all depths. The NXR-like proteins were all derived from metagenomic data, so we performed phylogenetic analysis to validate that they were indeed NXR subunits and to assess their likely phylogenetic origin. In both the NxrA and NxrB trees, the expressed proteins identified in the LSLE were affiliated with homologs from Nitrospina (**Figure 2.5a and 2.5b**). Previous molecular methods have demonstrated that the genus Nitrospina is common in the ocean (Lücker *et al.*, 2013; Fuchs *et al.*, 2005; Labrenz *et al.*, 2007; DeLong *et al.*, 2006) and we did identify a Nitrospina 16S rRNA sequence at 180 m. The detection of these Nitrospina-like NXR proteins supports the role of Nitrospina as an important NOB in the ocean. The distribution of peptide spectra from NXR proteins paralleled the distribution of ammonia-oxidizing archaea, as the abundance was much higher in the deep water compared to the surface (**Figure 2.5c**).

Two main types of NXR proteins are known which can be distinguished based on their subcellular orientation into the cytoplasm or periplasm (Lücker *et al.*, 2013; Sorokin *et al.*, 2012; Lücker *et al.*, 2010; Spieck and Bock, 2005). The NXR of Nitrospina has been posited to be periplasmically oriented (Lücker *et al.*, 2013). NOB with the periplasmic NXR can grow at lower nitrite concentrations than NOB with the cytoplasmic enzyme. This periplasmic orientation should be energetically advantageous because H^+ released by NO_2^- oxidation in the periplasm and proton consumption by O_2 reduction in the cytoplasm contribute to the membrane potential. Also, there is no requirement to transport NO_3^-/NO_2^- across the membrane. Hence the periplasmic orientation seems to be better adapted to low NO_2^- concentrations. Since NO_2^- rarely accumulates in the environment, the highly efficient use of this substrate is likely a prominent reason for the competitive success and wide distribution of Nitrospina in the ocean.

The type strain of the genus Nitrospina (*Nitrospina gracilis*) couples nitrite oxidation to autotrophic growth with CO_2 as a sole source of carbon (Lücker *et al.*, 2013; Watson and Waterbury, 1971). Genomic data for *N. gracilis* showed that this organism uses the reductive tricarboxylic acid (rTCA) cycle for carbon fixation. In the LSLE, we identified two key enzymes of the rTCA cycle: 2-oxogluturate:ferredoxin oxidoreductase (OGOR) and pyruvate:ferredoxin oxidoreductase (POR). Nitrospina-like proteins encoding multiple subunits (alpha, beta, and gamma) of both the OGOR and POR enzymes were identified in the deep water (**Figure 2.5d**). These results, in combination with the detection of MG-I CO_2 fixation pathways suggest that, in tandem, nitrifying archaea and bacteria play an important role in chemosynthetic production in the deep, dark LSLE.

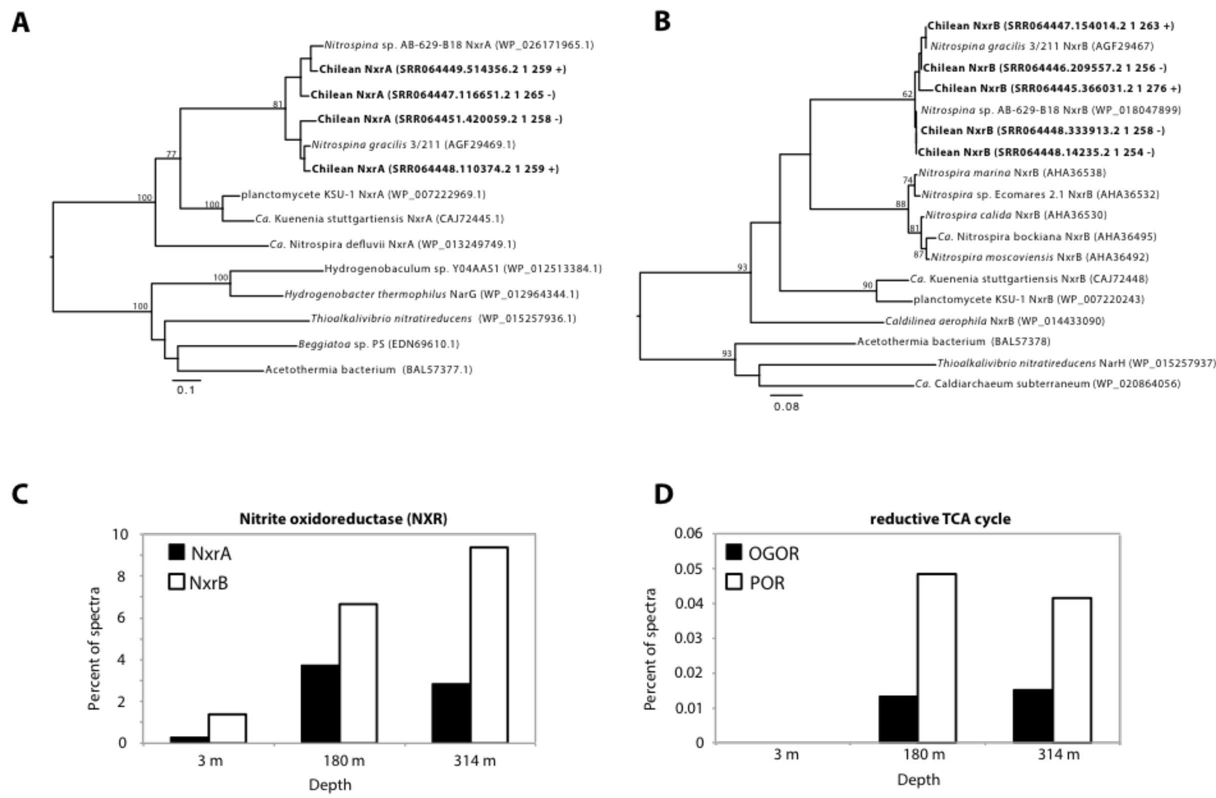


Figure 2.5. Phylogenetic analysis of nitrite oxidoreductase (NXR) proteins. (**A**) NxrA homologs and (**B**) NxrB homologs matching peptide spectra identified in the LSLE (**in bolface**). Phylogenies were inferred using maximum likelihood implemented in MEGA. Names of validated enzymes and bootstrap values are indicated. (**C**) The relative abundance of nitrite oxidoreductase subunits (NxrA and NxrB) and (**D**) key rTCA cycle enzymes (2-oxogluturate:ferredoxin oxidoreductase, OGOR; pyruvate:ferredoxin oxidoreductase, POR) based on spectra counts.

2.5.7 Methanol metabolism

Methylotrophic bacteria in seawater were reported decades ago (Murrell *et al.*, 1992). However, there is a renewed interest in this metabolic group, owing to the recent discovery of short methanol turnover times (J. L. Dixon *et al.*, 2011) and high rates of methanol carbon assimilation in the ocean (Dixon *et al.*, 2013). All known methanol-oxidizing bacteria employ a pyrroloquinoline quinone (PQQ)-containing methanol dehydrogenase (Keltjens *et al.*, 2014). In the LSLE metaproteomes, we identified eight PQQ-dependent dehydrogenase proteins, two of which were XoxF4 proteins from OM43 reference genomes (strain HTCC2181 and KB13).

Recent cultivation, genomic, and metaproteomic work has shown that the OM43 marine clade is comprised of methanol-oxidizing bacteria that are probably ecologically important in the ocean (Georges *et al.*, 2014; Giovannoni *et al.*, 2008; Sowell *et al.*, 2011). Phylogenetic analysis of the metagenome-derived proteins demonstrated that an additional two also belonged to the XoxF4 clade, while four were affiliated with a novel XoxF5 clade of putative methanol dehydrogenase proteins **(Figure 2.6)**. These XoxF4 proteins were identified in all samples, but peaked in the surface layer. However, the XoxF5 proteins were only detected in the deep water, and two were specifically associated with the nepheloid layer **(Figure 2.6)**. These observations suggest that methanol may serve as an important carbon and/or energy source for microbes in the particle-rich low productivity nepheloid layer and that these deep-water populations are distinct from the more common OM43 clade commonly associated with surface waters.

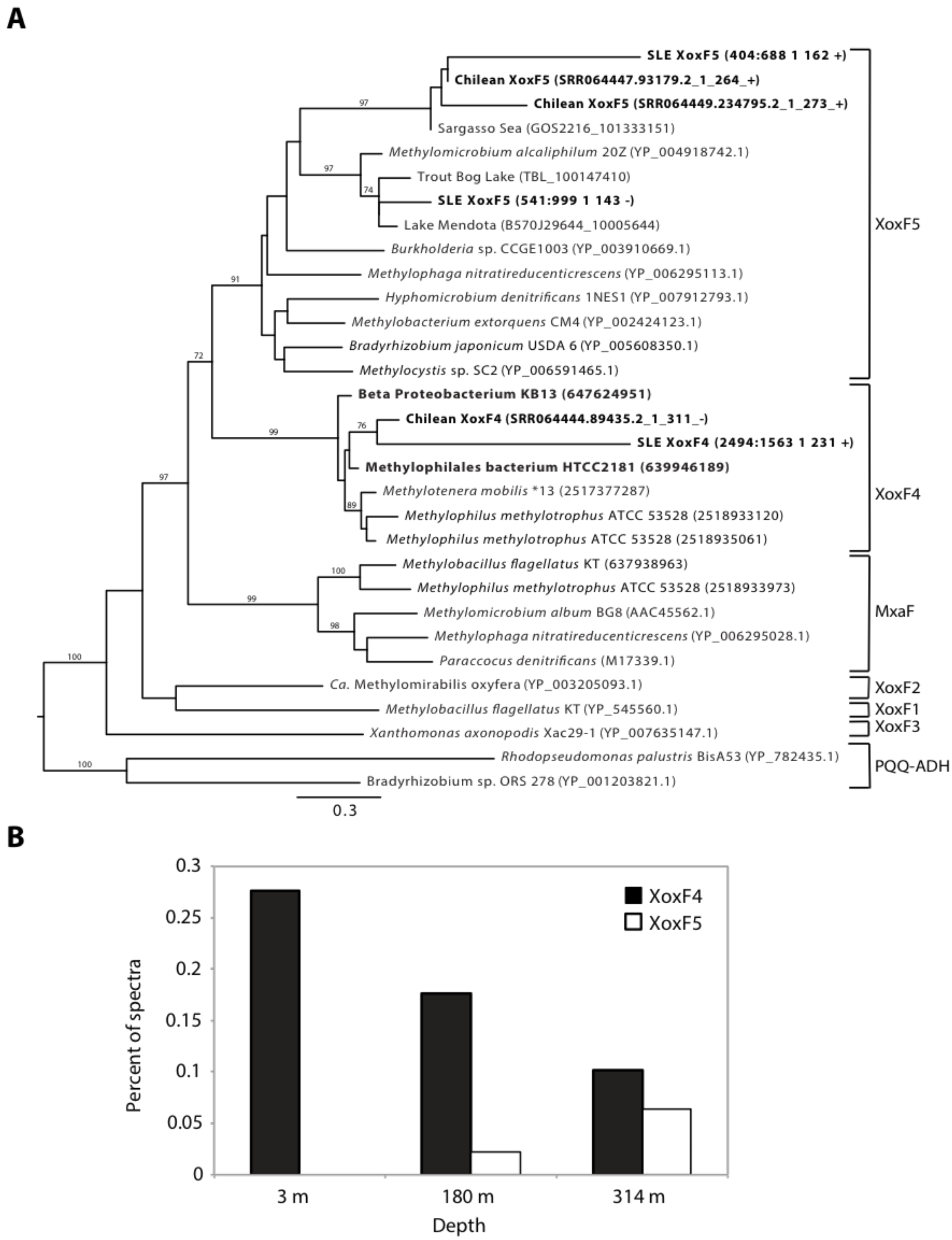


Figure 2.6. Phylogenetic analysis of PQQ-dependent dehydrogenase matching peptides (**A**) (**boldface**). The tree was inferred using maximum likelihood implemented in MEGA. Names of

validated enzymes and bootstrap values are indicated. (**B**) The relative abundance of PQQ-dependent dehydrogenase (XoxF4 and XoxF5) based on spectra counts.

2.6 Concluding remarks

In this study, we used metaproteomics to provide the first assessment of microbial metabolic pathway diversity in the stratified waters of the LSLE. Profiling of the microbial communities revealed differences in metabolic lifestyles and in carbon and nitrogen processing pathways throughout the water column. In the productive brackish surface waters, we identified heterotrophic Flavobacteriales and Rhodobacterales populations involved in the processing of HWM and LMW terrestrial OM (e.g. cellulose and xylose), respectively. Rhodobacterales were also implicated as the dominant processors of the labile organic carbon pool at the surface, while a much higher diversity of lineages appear to be involved in the uptake of labile compounds in the deep water. Chemosynthetic production coupled to nitrification by MG-I Thaumarchaeota and Nitrospina appeared to be an important metabolic strategy in the deep-waters. Although putatively mixotrophic lineages such as ARCTIC96BD-19 and SAR324 were common in the deep, the detection of proteins involved in transport of organic compounds and the lack of identified proteins for lithoautotrophic pathways suggests these groups are preferentially scavenging carbon compounds, rather than fixing CO_2. Common to other metaproteomic studies of the coastal ocean, we identified methanol oxidation proteins (e.g. XoxF4) originating from the OM43 clade. However, we also identified a novel and previously undescribed lineage of XoxF5-containing methanol-oxidizing population specifically in the bottom nepheloid layer. Membrane transport proteins assigned to the uncultivated MG-II Euryarchaeota were also specifically detected in the nepheloid layer, suggesting the presence of microbial taxa that are specialized for the particle rich nepheloid layer.

Although metaproteomics is a powerful approach for assessing metabolic diversity of microbial communities, the approach is not without limitations. One limitation is that many peptide spectra were simply not assigned to proteins because representative proteins were not present in the protein reference database. A second limitation is that although we validated the specificity of peptides originating from various clades (Nitrospina), it is possible that some may have originated from different phylogenetic lineages not represented in the protein database. This is particularly problematic if these proteins are highly conserved (and share common peptides) or their genes are prone to lateral gene transfer between marine microbes. Approaches to addressing these limitations are, as performed in this study, to analyze metagenomic and metaproteomic datasets from the same samples. Another solution is the increasing number of single-cell amplified genomes that are becoming available for a wide variety of poorly characterized bacteria and archaea (Rinke *et al.*, 2013), which will continually provide more comprehensive databases for peptide matching. Finally, the incorporation of *de novo* peptide prediction tools into metaproteomic studies may also increase the resolution of metaproteomic datasets by identifying novel peptide sequences that may not be represented in the protein database.

2.7 Acknowledgments

We wish to thank the captains and crews of the R/V Coriolis II for providing essential assistance during field operations. We thank Paul del Giorgio for providing bacterial abundance data, Sylvie Sanschagrin from the NRC-BRI for help in generating the metagenomics data, as well as Jean-Pierre Falgueyret, Marcos DiFalco and Anna Georges for their help with generating the metaproteomics data. This work was supported by NSERC Discovery (402214-2011), NSERC Shiptime, and CRC (950-221184) research grants. D.C. and A.R. were supported by Concordia Institute for Water, Energy, and Sustainable Systems. D.C. was further supported by FQRNT. We also thank two anonymous reviewers for helpful comments on the manuscript.

Bridging text

In Chapter 2 we identified differences in metabolic lifestyles and carbon and nitrogen processing between the surface and deep waters of the LSLE using a combined metagenomic-metaproteomic analysis. Heterotrophic populations made up of Flavobacteriales and Rhodobacterales were identified in the surface waters and were involved in the processing of HMW and LMW terrestrial-derived DOM (e.g. cellulose and xylose). A higher diversity of organisms were involved in DOM processing in the deep waters, where chemosynthetic production coupled to nitrification by MGI Thaumarchaeota and Nitrospina were also identified. Marine microbial community samples from the North Water were collected in the summer of 2013 with the intention of performing a metaproteomic analysis of these samples. Before the metaproteomic analysis was conducted, I developed a novel method for the isolation of both community DNA and protein from marine microbial samples preserved in RNAlater in order to maximize the amount of data that could be retrieved from each sample ((Colatriano and Walsh, 2015); Chapter 3).

Chapter 3: An aquatic microbial metaproteomics workflow: from cells to tryptic peptides suitable for tandem mass spectrometry-based analysis

Colatriano D, Walsh DA. (2015). An Aquatic Microbial Metaproteomics Workflow: From Cells to Tryptic Peptides Suitable for Tandem Mass Spectrometry-based Analysis. *J Vis Exp* 1–8.

3.1 Short Abstract

This protocol is for the extraction and concentration of protein and DNA from microbial biomass collected from seawater, followed by the generation of tryptic peptides suitable for tandem mass spectrometry-based proteomic analysis.

3.2 Long Abstract

Meta-omic technologies such as metagenomics, metatranscriptomics and metaproteomics can aid in the understanding of microbial community structure and metabolism. Although powerful, metagenomics alone can only elucidate functional potential. On the other hand, metaproteomics enables the description of the expressed *in situ* metabolism and function of a community. Here we describe a protocol for cell lysis, protein and DNA isolation, as well as peptide digestion and extraction from marine microbial cells collected on a cartridge filter unit (such as the Sterivex filter unit) and preserved in an RNA stabilization solution (like RNAlater). In mass spectrometry-based proteomics studies, the identification of peptides and proteins is performed by comparing peptide tandem mass spectra to a database of translated nucleotide sequences. Including the metagenome of a sample in the search database increases the number of peptides and proteins that can be identified from the mass spectra. Hence, in this protocol DNA is isolated from the same filter, which can be used subsequently for metagenomic analysis.

3.3 Introduction

Microorganisms are ubiquitous and play essential roles in Earth's biogeochemical cycles (Madsen, 2011). Currently, there are numerous molecular approaches available for characterizing microbial community structure and function. Most common is the analysis of 16S rRNA gene sequences PCR-amplified from environmental DNA (El-Swais *et al.*, 2015; Galand *et al.*, 2010; Lane *et al.*, 1986). A disadvantage of 16S rRNA gene analysis is that it only provides information on phylogenetic identity and community structure, with little information on metabolic function. In contrast, approaches such as metagenomics, metatranscriptomics and metaproteomics provide information on community structure and metabolism. Metagenomics, or the analysis of the gene content of an assemblage of organisms, provides information about the structure and functional potential of the community (Williams *et al.*, 2012; Sheik *et al.*, 2013; Venter *et al.*, 2004; Tyson *et al.*, 2004). Although powerful, this functional potential may not correspond to the metabolic activities of the organisms. An organism's genotype is represented by its genes, each of which can be transcribed to RNA and further translated to protein, resulting in a phenotype. Thus, to aid in the understanding of microbial functional activity in an environment, post-genomic analysis should be performed (Schneider and Riedel, 2010). Metatranscriptomics, or the analysis of RNA transcripts is useful because it reveals which genes are transcribed in any given environment. However, mRNA levels do not always match their corresponding protein levels due to translational regulation, RNA half-life, and the fact that multiple protein copies can be generated for every mRNA (Vogel and Marcotte, 2012).

For these reasons metaproteomics is now recognized as an important tool for environmental microbiology. Common metaproteomic analyses use a shotgun proteomic approach where the near full complement of proteins in a complex sample are purified and analyzed simultaneously, usually through enzymatic digestion into peptides and analysis on a mass spectrometer. Subsequent tandem mass spectrometry (MS/MS) "peptide fingerprinting" is used to determine the peptide sequence and potential protein of origin by protein database searching (for a review see (Hettich *et al.*, 2013)). Proteomic work has come a long way in the past 25 years thanks to the increase in genomic data availability and the increase in the sensitivity and accuracy of mass spectrometers allowing for high-throughput protein identification and quantification (von Bergen *et al.*, 2013; Hettich *et al.*, 2013). Since proteins are the final product of gene expression, metaproteomic data can help determine which organisms are active in any given environment and what proteins they are expressing. This is advantageous when trying to determine how a particular set of environmental variables will affect the phenotype of an organism or community. Early on, MS/MS-based metaproteomic studies in the ocean were used to identify specific proteins in targeted microbial lineages, with the first study focusing on the light driven proton pump proteorhodopsin in SAR11 marine bacteria (Giovannoni, Bibbs, *et al.*, 2005). More recently, comparative metaproteomic analyses have elucidated differential protein expression patterns between complex communities. Examples include the identification of temporal shifts in metabolism in the coastal Northwest Atlantic Ocean (Georges *et al.*, 2014) or the Antarctic Peninsula (Williams *et al.*, 2012). Other studies have described variations in protein expression patterns across spatial scales, for instance, along a geographical transect from a low-nutrient ocean gyre to a highly productive coastal upwelling system (Morris *et al.*, 2010). For further reviews of metaproteomics we recommend Schneider et al. (2010) (Schneider and Riedel, 2010) and Williams et al. (2014) (Williams and Cavicchioli, 2014). Targeted proteomics has also been employed in recent years to quantify expression of specific metabolic pathways in the environment (Saito *et al.*, 2014; Bertrand *et al.*, 2013).

There are three main phases in metaproteomic analysis. The first phase is sample preparation, which includes sample collection, cell lysis and concentration of protein. Sample collection in marine microbiology often entails the filtration of seawater through a pre-filter to remove larger eukaryotic cells, particles and particle-associated bacteria, followed by filtration for the capture of free living microbial cells, commonly with the use of a 0.22 μm cartridge filter unit (Hawley *et al.*, 2013; David A Walsh *et al.*, 2009). These filters are incased in a plastic cylinder and a cell lysis and protein extraction protocol that can be performed within the filter unit would be a valuable tool. Once biomass is obtained, the cells must be lysed to allow for protein extraction. Several methods can be employed, including guanidine-HCl lysis (Thompson and Chourey, 2008) and sodium dodecyl sulfate (SDS)-based lysis methods. Although detergents like SDS are very efficient at disrupting membranes and solubilizing many protein types, concentrations as low as 0.1% can interfere with downstream protein digestion and MS analysis (Lu and Zhu, 2005). Of major concern is the negative effects of SDS on trypsin digestion efficiency, resolving power of reversed phase liquid chromatography and ion suppression or accumulation inside the ion source (Sharma *et al.*, 2012).

The second phase is fractionation and analysis, where proteins are subjected to enzymatic digestion followed by LC MS/MS analysis, resulting in a m/z fragmentation pattern that can be used to ascertain the primary amino acid sequence of the initial tryptic peptide. Various digestion methods can be performed depending on the types of detergents used, as well as the downstream mass spectrometry workflow. In our protocol, 1-D PAGE electrophoresis followed by removal of SDS from the gel is utilized in order to remove any detergent contamination. The analysis of proteins that are difficult to solubilize, such as membrane proteins, requires the use of high concentrations of SDS or other detergents. This leads to compatibility issues with SDS-gel electrophoresis. If the objective of a study requires the solubilization of these hard to solubilize proteins, the tube-gel system can be used (Santoro *et al.*, 2015; Lu and Zhu, 2005). The tube-gel method incorporates proteins within the gel matrix without the use of electrophoresis. Subsequently any detergents used for solubilization are removed before protein digestion.

The third phase is the bioinformatic analysis. In this phase the MS/MS peptide data are searched against a database of translated nucleotide sequences to determine which peptides and proteins are present in the sample. The identification of peptides is dependent on the database it is searched against. Marine metaproteomic data are commonly searched against databases comprised of reference genomes, metagenomic data such as the Global Ocean Sampling dataset (Rusch *et al.*, 2007), as well as single cell amplified genomes from uncultivated lineages (Swan *et al.*, 2011; Rinke *et al.*, 2013). Protein identification can also be increased by the inclusion of metagenomic sequences from the same sample as the metaproteomic data was derived (Williams *et al.*, 2012).

Here we provide a protocol for the generation of peptides suitable for MS/MS-based analysis from microbial biomass collected by filtration and stored in an RNA stabilization solution. The protocol described here allows for DNA and protein to be isolated from the same sample so that all steps leading up to the protein and DNA precipitations are identical. From a practical perspective, less filtration is required since only one filter is required for both protein and DNA extraction. We would also like to acknowledge that this protocol was created through the combination, adaptation and modification of two previously published protocols. The cell lysis steps are adapted from Saito et al. (2011) (Saito *et al.*, 2011) and the in-gel trypsin digest component is adapted from Shevchenko et al. (2007) (Shevchenko *et al.*, 2007).

3.4 Protocol

Refer to Appendix A.

3.5 Representative results

As a demonstration, we performed the protocol on two seawater samples collected from the surface and the chlorophyll maximum of the coastal ocean in Northern Canada. While at sea, 6-7 L of seawater was passed through a 3 μm GF/D prefilter, then microbial cells were collected onto a 0.22 μm cartridge filter unit following the protocol of Walsh et al. (David A Walsh *et al.*, 2009). Cells were immediately stored in an RNA stabilization solution until further processing. Upon returning to the lab, we performed the protocol as it is presented here. The concentrated cell lysate was divided; protein was precipitated from 90% of the volume, while DNA was precipitated from the remaining 10% of the volume. We recovered 24-26 μg of protein and 250-308 ng of high quality DNA from these samples (**Figure. 3.1**). After the in-gel trypsin digestion and peptide extraction, we subjected the peptides to MS/MS analysis using a nano-LC coupled to the Orbitrap Elite mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA). From the peptides, we generated over 23,000 MS/MS spectra per sample. Peptides and proteins were then identified by searching these spectra against a custom in-house sequence database using the PEAKS bioinformatics tool (BSI, Waterloo, ON, Canada). The database was comprised of predicted proteins from marine reference genomes and metagenomes. The search resulted in the identification of around 1000 peptides and 700-800 proteins for each sample. Naturally, these results are dependent on microbial cell abundance, MS instrumentation, and protein search database and algorithms. Nonetheless, these results demonstrate that this protocol has the potential to produce adequate tryptic peptides suitable for identify hundreds of proteins in the environment. Moreover, since metagenomic libraries can be constructed from as little as 100 ng of DNA (Thomas *et al.*, 2012), this protocol also has potential to provide adequate quantities of DNA to generate matched metagenomic-metaproteomic datasets.

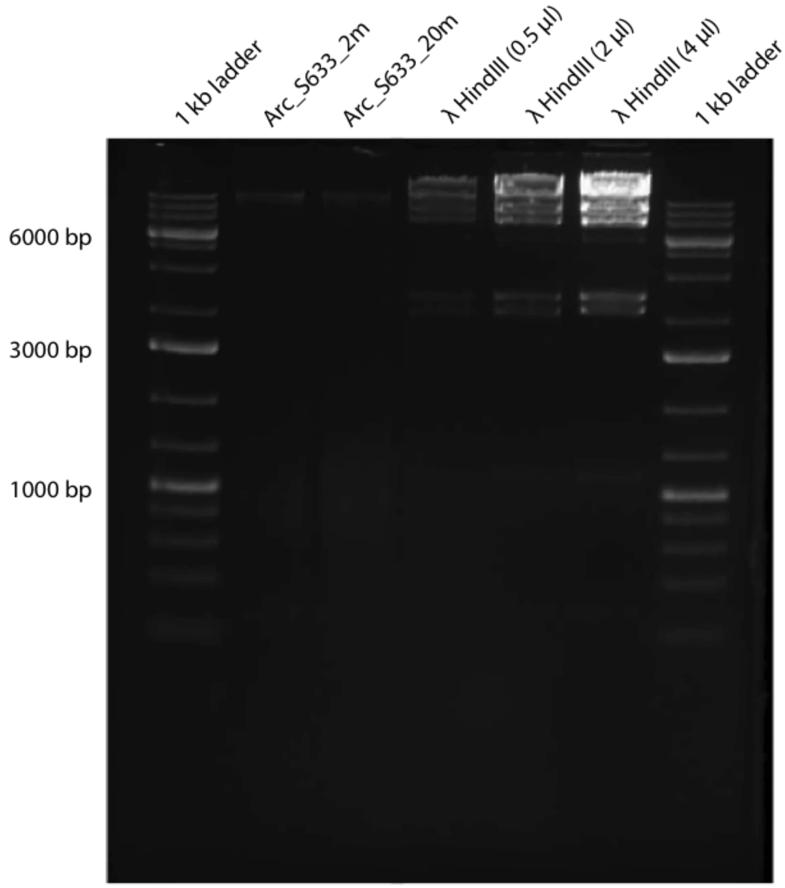


Figure 3.1. Genomic DNA from 2 depths at Arctic station S633. The first lane contains 4 µl of a 1 kb DNA ladder, lane 2 contains 3 µl of genomic DNA extraction from S633_2 m, lane 3 contains 3 µl of genomic DNA extraction from S633_20 m and lanes 4-6 contain 0.5 µl (85 ng), 2 µl (333 ng) and 4 µl (667 ng) of HindIII digested lambda DNA.

Taxonomic and functional composition of the metaproteomes was analyzed using a combination of BLASTp and the MEGAN (Metagenome analyzer) software package (Huson *et al.*, 2007, 2011) **(Figure. 3.2)**. Proteins assigned to Alpha-proteobacteria were the most highly represented in the dataset, the vast majority of which were assigned to the SAR11 clade. The Rhododobacterales clade of Alpha-proteobacteria was also highly represented and identified most often in surface waters. Proteins assigned to *Bacteroidetes* were evenly distributed between the surface and chlorophyll maximum, but Flavobacteria proteins were identified to a greater degree at the chlorophyll maximum. Gamma-proteobacterial proteins were evenly distributed throughout the water column while Beta-proteobacterial proteins were found predominately in the surface. From a functional perspective, a wide range of metabolic pathways were identified. Vertical structuring of these metabolic pathways was apparent. For example, proteins associated with amino acid metabolism, carbohydrate metabolism and prokaryotic carbon fixation pathways were identified primarily at the surface, and nitrogen metabolism was found exclusively at the surface. Photosynthetic carbon fixation proteins were observed primarily at the chlorophyll maximum while proteins involved in photosynthesis were identified evenly between the surface and chlorophyll maximum. These results demonstrate that a wide variety of proteins from a diversity of microbial taxa can be detected using the protocol presented here.

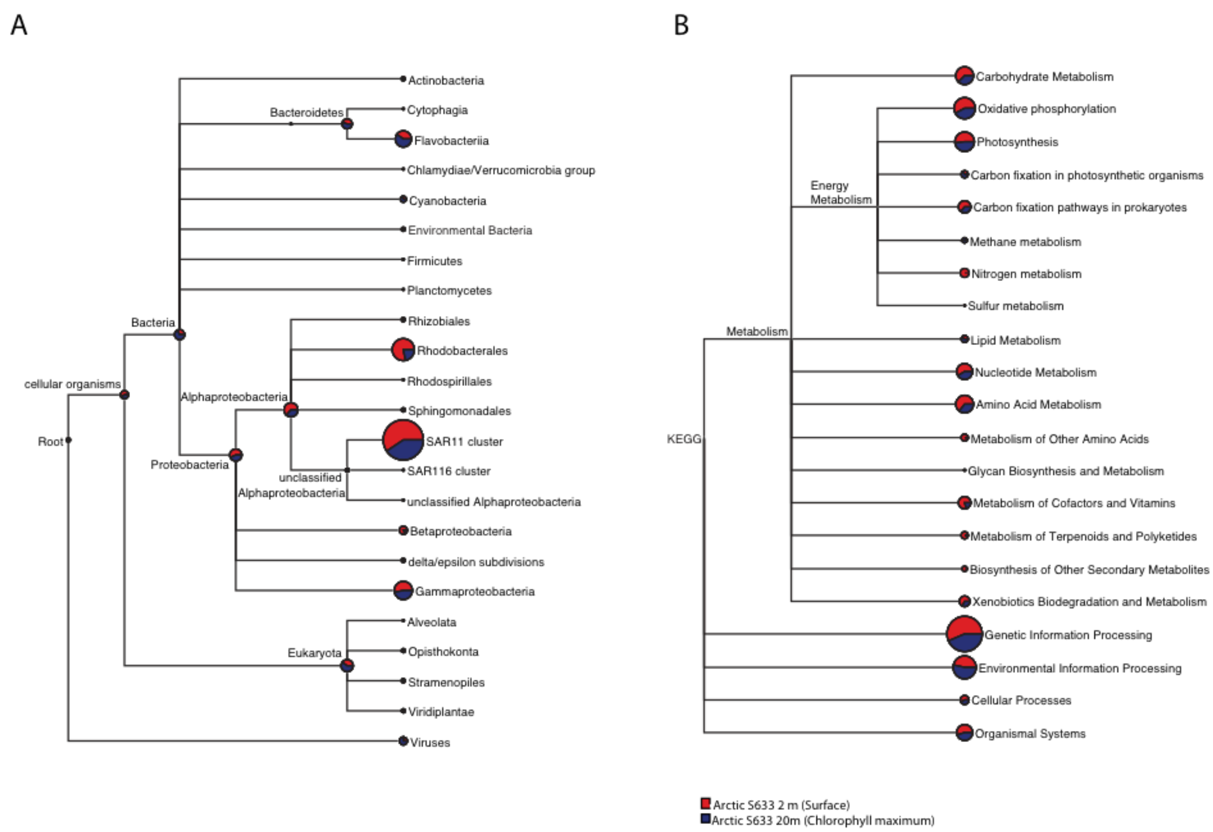


Figure 3.2. Taxonomic and functional analysis of 2 depths at Arctic station S633. Taxonomic diversity comparison of the Arctic station 633 surface and chlorophyll maximum waters (**A**) created using MEGAN. Functional diversity comparison of the Arctic station 633 surface and chlorophyll maximum waters (**B**) created using MEGAN to query against the KEGG database.

3.6 Discussion

Sample preservation is key to metaproteomic studies and previous work demonstrated that an RNA stabilization solution is a useful storage buffer for storing cells prior to protein extraction (Saito *et al.*, 2011). Ideally, samples would be preserved *in situ* to negate shifts in protein expression during handling (Ottesen *et al.*, 2011; Feike *et al.*, 2012). In fact, *in situ* sampling and fixation technologies have been developed, which allow for the autonomous collection and preservation of samples by ship-deployed instruments. However, access to these technologies is not always feasible. In the common case that it is not, samples should be preserved as soon as possible after collection.

Here we present a protocol for extracting protein from RNA stabilization solution stored cells collected on a cartridge filter unit, which is commonly used in aquatic microbiology. The protocol includes cell lysis using an SDS-lysis solution and heating, followed by a protein concentration step using ultracentrifugal filter units that doubled as a necessary desalting step. It must be noted that the concentrating and desalting steps cannot be overlooked. We found that a minimum of three buffer exchange steps was required to desalt our concentrate. Due to the high salt concentration of the RNA stabilization solution, if proper desalting does not occur, too much salt will be precipitated during the overnight protein precipitation step and the desalting and precipitation step will have to be repeated. Additionally, if desalination is not properly performed the 1D-PAGE will not work and the samples will be lost.

Next, the concentrated lysate was divided so that both protein precipitation and DNA precipitation could be performed. This is useful as it is often desirable that metagenomic and metaproteomic data be generated from the same samples. If a protein is not represented in the protein sequence database then the peptide will not be identified. Including the genomic data from the same sample as the proteomic data reduces the risk of not being able to identify a protein due to its absence from the database.

Although this protocol was optimized for use with cartridge filter units and validated to work on coastal ocean microbial communities, it can be adapted for use with other types of environmental samples and filters. However, it should be clearly stated that the success of this protocol is dependent on an adequate amount of starting biomass. Therefore in aquatic ecosystems where biomass may be very low, we recommend increasing the volume of water filtered accordingly.

3.7. Acknowledgements:

The authors would like to acknowledge Marcos Di Falco for his expertise and advice with the preparation of the samples for nano-LC MS/MS as well as Dr. Zoran Minic from the University of Regina for the LC MS/MS analysis. This work was supported by NSERC (DG402214-2011) and CRC (950-221184) funding. D.C. was supported by Concordia Institute for Water, Energy, and Sustainable Systems and FQRNT.

Bridging text

Using the computational techniques developed in the analysis of Chapter 2 and the new DNA/protein extraction method developed in Chapter 3, I performed the first metaproteomic analysis of the North Water.

Chapter 4: Metaproteomics reveals a strong association between community phenotype and taxonomic composition of main bacterioplankton in the North Water

4.1 Abstract

The North Water, situated between Canada and Greenland, supports one of the most productive marine ecosystems in the Arctic and is a major gateway for Arctic waters entering the North Atlantic Ocean. Although the North Water plays an integral role in supporting local animal populations and influences the biogeochemical cycles of the North Atlantic there is a lack of information about the microbial diversity and metabolism in this region. Here, we investigated microbial community structure and function across the North Water using a combination of 16S rRNA gene sequencing and metaproteomic analyses. Three distinct microbial community types with differing metabolic functions were identified: one in the polar mixed layer on the Canadian side (W-PML), one in the polar mixed layer on the Greenland side (E-PML) side and one in the bottom waters (BTM). Analysis of taxonomically resolved COG functions revealed differences in resource allocation and microbial food web structure between the different microbial community types. The W-PML community was more typical of those associated with dynamic ecosystems that sustain pulses of productivity and relied more heavily on the scavenging of phytoplankton-derived high molecular weight dissolved organic matter. The W-PML also harboured a microbial community with potentially higher growth rates than the E-PML. The E-PML was defined by its participation in methanol oxidation, a community composition more typical of a stable phytoplankton community, and constituted a community that utilized a larger proportion of high affinity membrane transport proteins, perhaps as a means of cycling and competing for limited low molecular weight DOM. The bottom waters were characterized by a more diverse and metabolically versatile array of microbes with an abundance of SAR324 and Arctic96BD-19 bacteria. Additionally, the identification of an abundance of proteins involved in nitrification assigned to Thaumarchaeota and Nitrospinae in the bottom waters could suggest an important role of these microbes in providing nitrate to help drive the high productivity of this region.

4.2 Introduction

The North Water (locally referred to as Pikialasorsuaq), situated between Canada and Greenland, supports one of the most productive marine ecosystems in the Arctic (Bâcle *et al.*, 2002; Ardyna *et al.*, 2011). This productivity is due, in part, to a longer open water season compared to surrounding areas, exemplified by the formation of a polyna (open water surrounded by sea ice) in the winter, which enables early access to resources by migrating seabirds and marine mammals. The North Water is contiguous with Nares Strait, which, along with Jones Sound and Lancaster Sound comprises a major net outflow of Arctic waters to the North Atlantic. Arctic Ocean water flows directly southward on the Canadian side of the North Water, transporting phosphate rich waters into the North Atlantic which influences the nutrient balance and species composition in the North Atlantic Ocean. A northward flow along the west coast of Greenland (West Greenland Current) transports warmer, saltier, nitrate-rich Atlantic Waters into the North Water. The northward flow is then forced South, driven by cyclonic circulation, and this mixing of water masses results in the high primary productivity characteristic of the North Water region.

The Arctic is highly sensitive to climate change, with impacts on primary production and carbon fluxes (Frey and McClelland, 2009; Vonk *et al.*, 2012; Tremblay *et al.*, 2011; Arrigo *et al.*, 2008). In the North Water, the location of the spring-summer chlorophyll signal is variable but is observed to be occurring further North over time (Marchese *et al.*, 2017). Recent increasing northern penetration of Atlantic waters along the west coast of Greenland and melting of the Greenland Ice Sheet could alter the environment of Nares Strait and the North Water region as a whole. Given the rapidly changing Arctic climate, the importance of the North Water to regional Arctic productivity, and the role of this region as a gateway linking the Arctic-Atlantic system, an understanding of microbial distributions and metabolic activities, which form the base of the marine food web, within this region is warranted.

A paucity of information exists for bacterioplankton diversity and metabolism in the North Water, although numerous studies in other regions of the Arctic exist. What little data there is focuses on specific taxa. For example, Fu et al. (Fu *et al.*, 2013) determined that the Rhodobacterales communities in the different water masses of the North Water had distinct phylogenetic compositions and that community structure changed with depth. Similarly, a study focusing on archaeal community composition between three distinct water masses within the North Water suggested that community structure and metabolic potential is not necessarily dependent on depth stratification, but rather on parent water masses (Galand *et al.*, 2009). Ammonia monooxygenase genes were also more highly represented in one of the three water masses sampled, indicating that distinct water masses could harbour distinct biogeochemical functions. One meta-analysis exploring the differences in community composition of various Arctic Ocean regions, including Baffin Bay/North Water, Chukchi Sea, Lancaster sound, Cambridge Bay, Hudson Bay and Laptev Sea resolved the distribution of major taxonomic groups (Pedrós-Alió *et al.*, 2015). Work exploring the temporal and spatial trends in phytoplankton and protist community structure over a four month period at several regions within the North Water (Lovejoy *et al.*, 2002), as well as a comparison of the diurnal changes in phytoplankton and protist composition between the Western and Eastern side of the North Water (Joli *et al.*, 2018) has also been performed, and showed that the western North Water seemed to be a more dynamic system in terms of phytoplankton community composition through time than the eastern region. Several of these studies, as well as others concentrating on other Arctic or marine systems, imply that water masses with distinct physico-chemical properties, nutrient composition and concentrations harbour different microbial community structures despite their close geographic proximity (Fu *et al.*, 2013; Lovejoy *et al.*, 2002; Hamilton *et al.*, 2008; Varela, Van Aken, Sintes, *et al.*, 2008; Galand *et al.*, 2009, 2010). However, little is known about the difference in metabolic potential or protein expression of the microbial communities in these different water masses.

One useful approach for assessing the metabolic activity of marine microbes in relation to environmental conditions is through metaproteomic analysis. Metaproteomic analysis has the potential to provide information on *in situ* protein expression for whole microbial communities (VerBerkmoes and Denef, 2009). In metaproteomics, samples are obtained directly from the environment, proteins are extracted from the biomass and subjected to MS/MS analysis. The resulting mass spectra are searched against a protein sequence database to determine the amino acid sequences of the peptides, and their protein of origin. Metaproteomics has been applied to a range of marine environments to determine temporal variability in microbial metabolism through seasons (Williams *et al.*, 2012; Georges *et al.*, 2014) and phytoplankton blooms (Teeling *et al.*, 2012). It has also been used to investigate the variation in protein expression patterns across spatial scales, including along a geographical transect from a highly productive coastal system to a low nutrient ocean gyre (Morris *et al.*, 2010) or depth in a stratified estuary (chapter 2; (Colatriano *et al.*, 2015)). Proteomic analysis of marine microbial isolates have also been used to elucidate the protein expression changes that occur when marine microbes change physiological states (*i.e.* from exponential to stationary phase) (Sowell *et al.*, 2008; Muthusamy *et al.*, 2017). These experiments demonstrate the different ways marine bacteria cope with unfavourable growth conditions. For instance, SAR11 increases the abundance of several proteins that contribute to homeostasis rather than undergoing a major remodeling of its proteome (Sowell *et al.*, 2008) while a change from a dominant translation-related protein profile to a nutrient-scavenging-related profile was observed for typical copiotrophic marine bacteria (Muthusamy *et al.*, 2017).

Given the fact that atmospheric warming is rapidly changing Arctic ecosystems and will likely impact marine microbial community structure and function, understanding the degree to which metabolic functions are redundant in marine systems is necessary for properly assessing how changes in microbial community structure might affect the metabolic processes of the system. The high diversity of bacterioplankton identified in marine systems with limited sets of resources have led to the hypothesis that marine microbial communities are highly functionally redundant, or that different microbial communities are capable of performing the same metabolic processes (Allison and Martiny, 2008) and can therefore readily replace each other (Louca *et al.*, 2018). However, this hypothesis is not universally accepted, with others positing that the high diversity of bacterioplankton instead reflects a large diversity of microbial metabolisms that is not fully captured by annotations of genes and pathways from the cultured representatives found in databases (Galand *et al.*, 2018). Furthermore, the idea that changes in microbial community composition should not affect microbial-mediated processes does not take into account organisms that share some functions but differ in others, or have other ecological requirements. The differences in water mass origin, and phytoplankton dynamics between the two sides of the North Water allows for an interesting exploration into the functional redundancy of communities and how differences in ecosystem properties might change both community composition and/or the metabolic processes taking place.

In this study, we investigated microbial community structure and function across the Canadian and Greenland sides of the North Water using a combination of 16S rRNA gene sequencing and metaproteomic analyses. We posit that, although the Canadian and Greenland sides of the North Water region are found within the same eco-region (Lawrence *et al.*, 2015), differences in average primary production, hydrological processes and water origins will result in distinct bacterioplankton community compositions with distinct metabolic functions and nutrient acquisition strategies. Based on previous work investigating primary production in the North Water (Joli *et al.*, 2018), we hypothesized that microbial communities on the Greenland side would be comprised of organisms typically associated with ecosystems of stable nutrient concentration and composition, while the Canadian side would be comprised of communities typically associated with dynamic ecosystems subjected to nutrient pulses.

4.3 Methods

4.3.1 Sample collection and DNA/protein extraction

Arctic marine microbial samples were collected during an expedition from July 26th 2013 to September 5th 2013 aboard the CCGS Amundsen. Water samples were collected using a CTD rosette and subsequently pumped through a 20 µm pore size mesh, a 3µm pore size GF/D prefilter, followed by a 0.22 µm pore size GP Sterivex filter with the use of a peristaltic pump. Once filtration was completed, the 3 µm pore size filters were stored in 2 ml of RNAlater. Approximately 1.6 ml of RNAlater was added to the Sterivex filter for storage. Both filters were then stored at -80 °C. Protein and DNA extraction was performed as in chapter 3, (Colatriano and Walsh, 2015).

Oceanographic data was collected as in Joli et al. 2018 (Joli *et al.*, 2018) aboard the CCGS Amundsen using a rosette system equipped with a conductivity, temperature, depth (CTD) profiler (Sea-Bird SBE-911 CTD), relative nitrate (In-Situ Ultraviolet Spectrometer, ISUS, Satlantic), oxygen (Seabird SBE-43), chlorophyll fluorescence (Seapoint), fluorescent colored dissolved organic matter (fCDOM; Wetlabs ECO) and photosynthetically available radiation (PAR, 400–700 nm; Biospherical Instruments QDP2300) sensors. The oxygen sensor was calibrated onboard against Winkler titrations (Martin *et al.*, 2010).

4.3.2 Bacterial 16S rRNA gene analysis

Isolated DNA was used as a template for 16S rRNA gene amplification and analysis. Community DNA was diluted to a concentration of 1 ng/µl and used as a template for nested PCR amplification (two-step) modified from Tran et al. 2019 (Tran *et al.*, 2019). The primers used in the first amplification reaction were 515R and 341F. The reaction mix consisted of 0.5 µM of each primer, 1 X Phire Reaction Buffer containing 1.5 mM $MgCl_2$, 0.2mM deoxynucleotide triphosphates and 1 U of *Phire* Hot Start II DNA polymerase (Finnzymes Thermofischer Scientific). Cycling conditions involved an initial denaturing step at 98 °C followed by 30 cycles of 10 s at 98 °C, 20 s at 55 °C and 15 s at 72 °C, and a final elongation step of 5 minutes at 72 °C. The template was amplified using non-barcoded PCR primers for 20 cycles, then 1 µl of the PCR product was amplified for an additional 10 cycles with barcoded reverse PCR primers with specific IonXpress sequences to identify samples. PCR products were purified using QIAquick Gel Extraction Kit (Qiagen), quantified using Quantifluor dsDNA System (Promega), pooled at equimolar concentration and sequenced using an Ion Torrent PGM system on a 316 chip with the Ion Sequencing 400 kit.

16S rRNA sequences were analyzed using the open-source MOTHUR pipeline (Schloss *et al.*, 2009). Sequences with an average quality of < 17, length < 100 bp or that did not match the IonXpress barcode and both the PCR forward and reverse primer sequences were discarded. Sequences were clustered into OTUs at 97% identity using the furthest neighbour algorithm. Sequences were subsampled to a depth of 24119 sequences per sample. Sequences and OTUs were assigned to taxonomic groups using the Silva database (Quast *et al.*, 2013), the Wang method and a bootstrap value cut-off of > 60% (Wang *et al.*, 2007).

4.3.3 Metaproteomics

Once protein extraction was completed, trypsin digestion of extracted proteins into peptides followed by LC MS/MS was performed. Chromatographic separation of peptides was performed on a Proxeon EASY nLC 1000 (Proxeon, Mississauga, ON, Canada) nano high-performance liquid chromatograph. Samples were directly injected into a nano column (C18 column, 10 cm x 75 μm ID, 3 μm, 100 Å) employing a water/acetonitrile/0.1% formic acid gradient over 100 min at a flow rate of 0.30 μl/min. Peptides were then separated using 1% acetonitrile, increasing to 3 % acetonitrile in the first 2 min and then a linear gradient from 3 % to 24 % acetonitrile for 74 min, followed by a linear gradient from 24 % to 100 % acetonitrile for 14 minutes, followed by a 10 minute wash of 100 % acetonitrile. Eluted peptides were directly sprayed into an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific) using positive electrospray ionization (ESI) at an ion source temperature of 250 °C and an ion spray voltage of 2.1 kV. Full-scan MS spectra (m/z 350–2000) were acquired in the Orbitrap at a resolution of 60 000 (m/z 400). The automatic gain control setting was $1e^6$ for full FTMS scans and $5e^4$ for MS/MS scans. Fragmentation was performed with collision-induced dissociation (CID) in the linear ion trap when an ion's intensity was >1500 counts. The 15 most intense ions were isolated for ion trap CID with charge states ≥2 and sequentially isolated for fragmentation using the normalized collision energy set at 35%, activation Q at 0.250 and an activation time of 10 ms. Ions selected for MS/MS were dynamically excluded for 30 s.

4.3.4 Protein identification

To identify peptide sequences, spectra were searched against a custom in-house made protein database comprised of 676 reference genomes, 77 metagenomes, 107 single cell amplified genomes from uncultivated lineages and the common Repository of Adventitious Proteins (cRAP) database. The PEAKS (Bioinformatics Solutions, Waterloo, ON, Canada) database search tool was employed to search MS/MS spectra against the constructed database with settings: enzyme type, trypsin, error tolerance parent ion, 10.0 ppm using monoisotopic mass, fragment ion 0.8 Da; maximum missed cleavage sites, 3; static post translational modification, Carbamidomethylaiton (+57 Da, iodoacetamide modification of cysteine); variable post translational modification, Oxidation (+16 Da, oxidation of methionine); maximum allowed PTM per peptide, 3; estimate false discovery rate (FDR) with decoy-fusion. An FDR cutoff of 0.1% was used. The search input consisted of 209,159 spectra.

4.3.5 Taxonomic and functional annotation of proteins

All identified proteins assigned to a member of the common repository of adventitious proteins (cRAP) database were excluded from further analysis. All other identified proteins were searched against the RefSeq (10_2015) protein database using DIAMOND (Buchfink *et al.*, 2015) and the top 10 hits with an e-value less than e^{-5} were reported. The DIAMOND search results were then loaded into MEGAN and taxonomic assignment was performed using the lowest common ancestor (LCA) algorithm (Huson *et al.*, 2011).

Proteins were first assigned to function based on their COG annotation from JGI. Proteins that did not have a JGI annotation were queried against the Cluster of Orthologous Genes (COG) database to identify a probable function. Only proteins with a sequence similarity of $1xe^{-5}$ to a COG function were annotated with the corresponding function. Additionally, proteins assigned to MGI Thaumarchaeota and identified as having no COG functional category were then queried against Refseq nr (using BLASTp) to identify ammonia monooxygenase proteins. Methanol dehydrogenase proteins were identified by first aligning all identified proteins assigned to COG4993 (Glucose dehydrogenase), along with 32 PQQ-dependent dehydrogenase reference sequences using MUSCLE (implemented in MEGA6). Phylogenetic reconstructions were conducted by maximum likelihood using MEGA6-v.0.6 and the following settings: JTT substitution model, gamma distribution with invariant sites model for the rate variation with four discrete gamma categories, and the nearest-neighbor interchange (NNI) heuristic search method (Tamura *et al.*, 2013) with a bootstrap analysis using 100 replicates (**Supplementary Figure 4.1)**. Taxonomic identification of the 2 proteins assigned to nitrate reductase beta was performed by aligning them against 13 reference NxrB reference genomes from various phyla using MUSCLE (implemented in MEGA6). Phylogenetic reconstruction was performed as in the methanol dehydrogenase phylogeny (**Supplementary Figure 4.2**).

4.3.6 Statistical analysis

Principal coordinate analyses of samples based on OTU abundance were plotted by first constructing a Bray-Curtis distance matrix based on the relative abundance of OTUs per sample. A PCoA was then constructed using vegdist and plotted using ordiplot which are both part of the vegan package in R (Oksanen *et al.*, 2019). Environmental data was then fitted to the PCoA using the envfit function as implemented in the vegan package (Oksanen *et al.*, 2019).

Hierarchical clustering of samples based on COG functions was performed using the hclust module implemented in R (Müllner, 2013) and an average distance measure on a matrix of the relative abundance (based on peptide spectra matches) of proteins assigned to COG functions per sample (with those assigned to No COG removed). Principal coordinate analyses of samples based on COG functions were plotted by first constructing a Bray-Curtis distance matrix based on the relative abundance of peptide spectra matches (PSMs) assigned to COG functions per sample (with those assigned to No COG removed). A PCoA was then constructed using vegdist and plotted using ordiplot which are both part of the vegan package in R (Oksanen *et al.*, 2019). Environmental data was then fitted to the PCoA using the envfit function as implemented in the vegan package (Oksanen *et al.*, 2019).

Principal coordinate analyses of samples based on taxonomically-resolved COG (tr-COG) functions were plotted by first constructing a Bray-Curtis distance matrix based on the relative abundance (based on PSMs) of proteins assigned to COG functions linked to their taxonomic assignment per sample (with those that could only be taxonomically resolved below phylum or that were assigned to No COG removed). A PCoA was then constructed using vegdist and plotted using ordiplot which are both part of the vegan package in R (Oksanen *et al.*, 2019). Environmental data was then fitted to the PCoA using the envfit function as implemented in the vegan package (Oksanen *et al.*, 2019).

COG and Tr-COG function log-base 2 differences between the W-PML and E-PML were calculated by first performing a Laplace correction on the PSMs of identified proteins in each region. Corrected PSM values were then used to calculate percent spectra for each region before determining the absolute $\log_2$ ratios between the two regions.

4.4 Results

4.4.1 Oceanographic setting

Sampling was conducted in mid to late August 2013, coinciding with the period that follows the highly productive spring/early summer bloom. Samples were collected from four locations, consisting of a northern and southern station on each of the Canadian (stations SC-108 and NC-117) and Greenland (stations SG-115 and NG-126) sides of the North Water (**Figure 4.1**). At the time of sampling, the polar mixed layer (PML) reached a depth of ~50 m at all stations, although a more complex thermocline was observed at station NC-117. Nutrient concentrations increased with depth at all stations and a subsurface chlorophyll maximum (SCM) was consistently observed between 20-33 m (**Figure 4.1**). The vertical structure of the water column at SG-115 was unique compared to the others. Specifically, the SG-115 surface water was fresher and warmer, and the maximum chlorophyll concentration (5.4 mg/m^3) observed during the study was within the SG-115 SCM, which was associated with more saline waters. The fresher surface water observed at SG-115 is a result of freshwater input from melting Greenland glaciers, while the elevated salinity of the underlying water is a result of northward flow of saltier Atlantic water with the West Greenland Current.

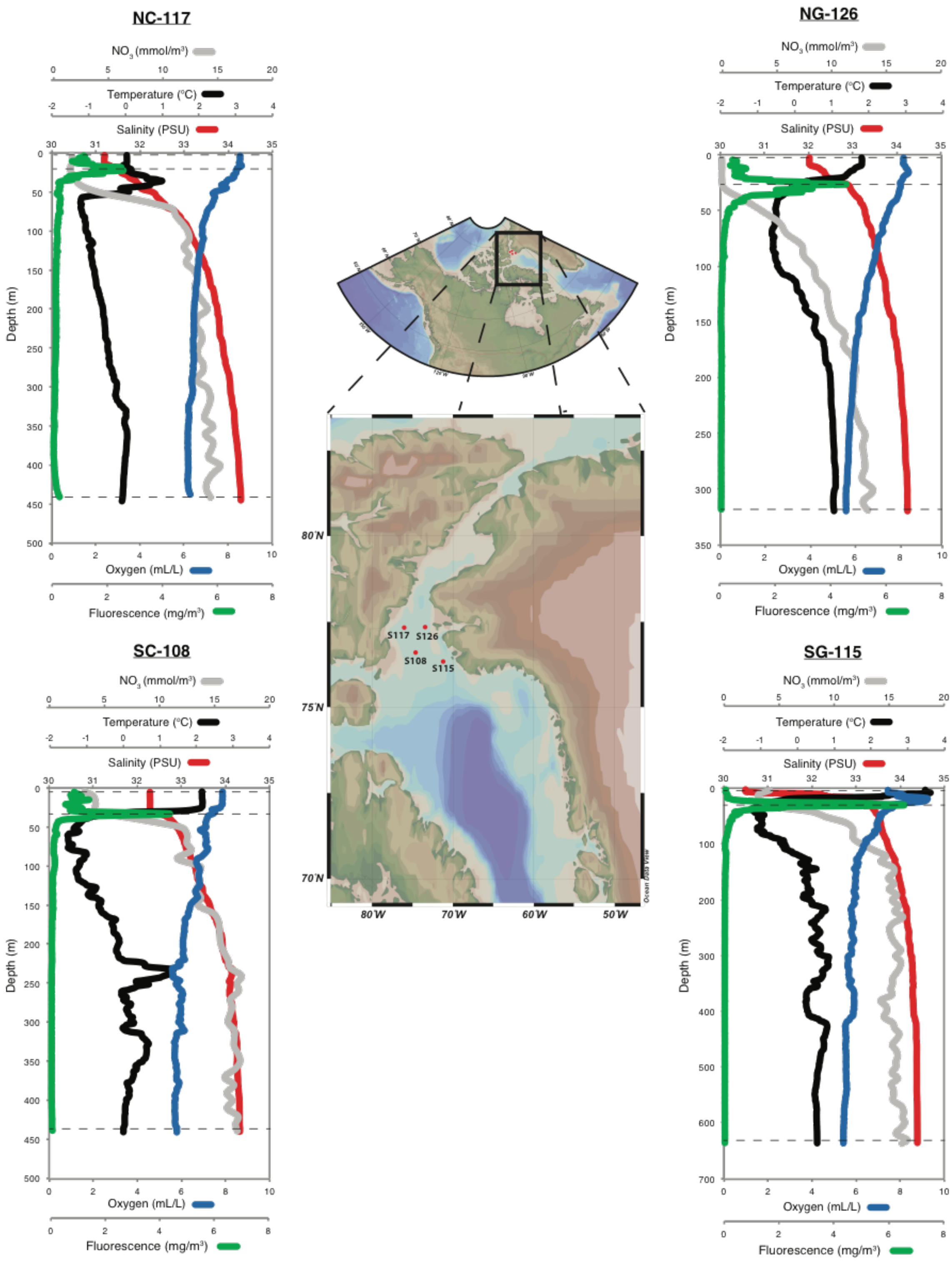


Figure 4.1. Sampling stations for this study in the Canadian North Water (NC-117, SC-108, NG-126, SC-115) and the physicochemical conditions each station plotted against depth.

4.4.2 Bacterial community structure

Bacterial community composition was assessed in North Water samples originating from the surface (1 m), subsurface chlorophyll mximum (SCM) (20-33 m), and bottom waters (319-636 m) using 16S rRNA gene analysis. Principal coordinate (PCo) analysis of communities based on the distribution of operational taxonomic units **(Figure 4.2a)** separated samples originating from deeper (> 300 m) and saltier water (including the SG-115 SCM) from the rest of the PML samples along axis 1. PML samples from the Canadian and Greenland sides were further separated from each other in a longitudinal fashion along axis 2 of the PCo ordination plot. Based on the PCo analysis, we assigned North Water samples to three distinct community types, consisting of the Western PML (W-PML), Eastern PML (E-PML) and bottom water (BTM) samples, although we acknowledge that the SG-115 SCM sample is unique in bacterial community composition.

In agreement with the PCo analysis, the taxonomic composition differed between W-PML, E-PML, and BTM communities **(Figure 4.2b)**. Among the differences, E-PML communities exhibited a higher relative abundance of oligotrophic SAR11 and SAR86 taxa compared to the W-PML. In contrast, W-PML communities harboured a higher average abundance of Rhodobacterales, ZA2333c and other Gamma-proteobacteria, Flavobacteriales, and Verrucomicrobia. In addition to 16S rRNA sequences of bacterial origin, a significant portion of chloroplast-related 16S rRNA sequences were recovered. Sequences related to diatoms (i.e. Haslea) and dinoflagellates (i.e. Dynophysis) chloroplasts, as well as sequences related to the unidentified eukaryote OM81 were common, but on average more abundant in the W-PML (**Figure 4.2b**). A striking observation was that more than 30% of the 16S rRNA sequences from the SCM at the northern side of Greenland (NG-126) were related to chloroplasts of coccolithophores (i.e. Emiliania). The BTM communities were differentiated from the PML communities by a higher average abundance of typical deep-water lineages such as SAR324, Gamma sulfur-oxidizer (GSO) clade, Nitrospina, Chloroflexi, Planctomycetes, and Marinimicrobia.

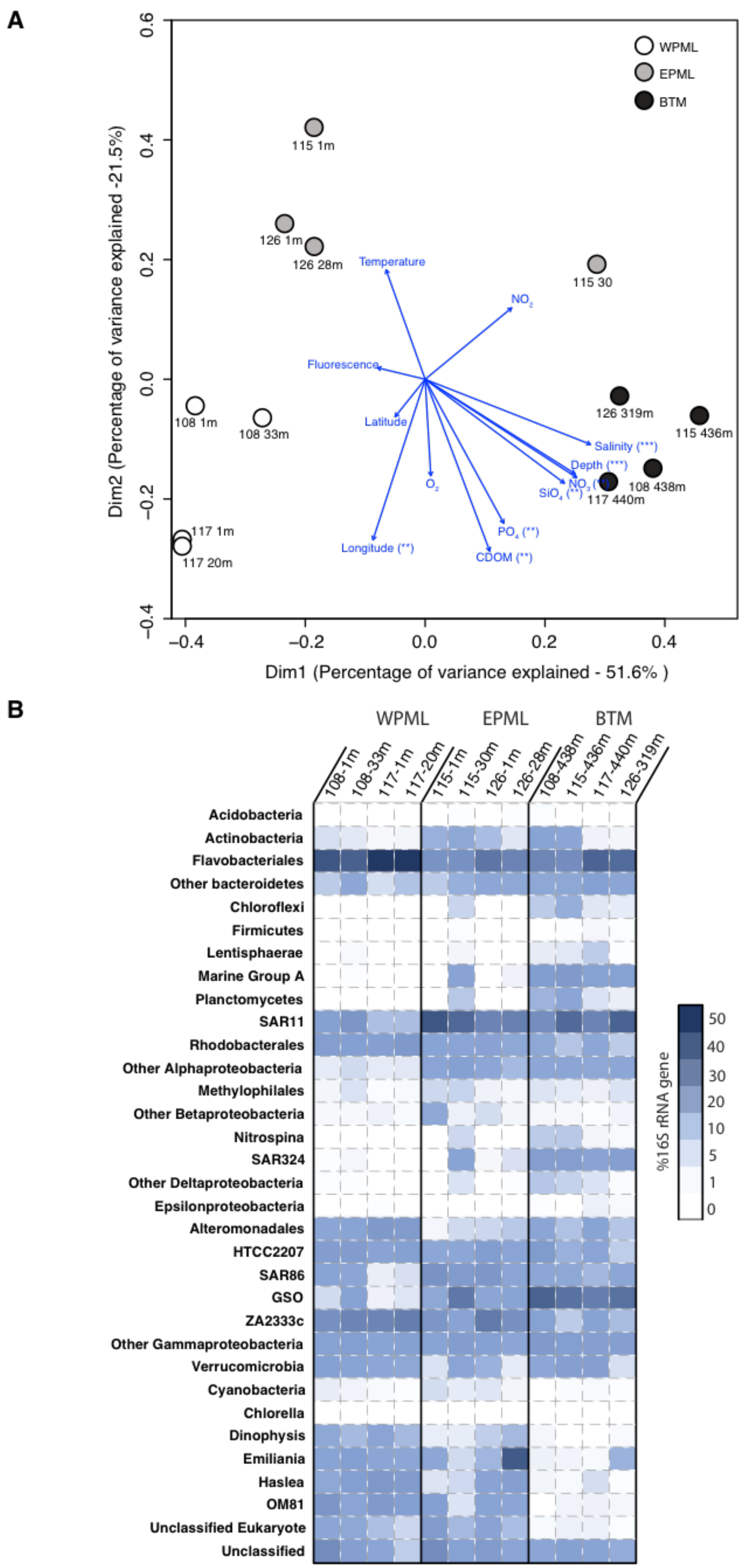

A
WPML
EPML
BTM
115 1m
126 1m
126 28m
115 30
Temperature
NO2
Fluorescence
Latitude
O2
Longitude (**)
CDOM (**)
PO4 (**)
SiO4 (**)
NO3 (**)
Depth (***)
Salinity (***)
126 319m
115 436m
108 438m
117 440m
108 1m
108 33m
117 1m
117 20m
Dim2 (Percentage of variance explained -21.5%)
Dim1 (Percentage of variance explained - 51.6%)
B
WPML
EPML
BTM
108-1m
108-33m
117-1m
117-20m
115-1m
115-30m
126-1m
126-28m
108-438m
115-436m
117-440m
126-319m
Acidobacteria
Actinobacteria
Flavobacteriales
Other bacteroidetes
Chloroflexi
Firmicutes
Lentisphaerae
Marine Group A
Planctomycetes
SAR11
Rhodobacterales
Other Alphaproteobacteria
Methylophilales
Other Betaproteobacteria
Nitrospina
SAR324
Other Deltaproteobacteria
Epsilonproteobacteria
Alteromonadales
HTCC2207
SAR86
GSO
ZA2333c
Other Gammaproteobacteria
Verrucomicrobia
Cyanobacteria
Chlorella
Dinophysis
Emiliania
Haslea
OM81
Unclassified Eukaryote
Unclassified
%16S rRNA gene
50
40
30
20
10
5
1
0

Figure 4.2. Principal coordinate analysis of communities based on the distribution of operational taxonomic units with correlated environmental variables (**A**). Taxonomic composition of the 12 samples, grouped by community (W-PML, E-PML, BTM) based on % 16S rRNA gene abundances (**B**).

<u>4.4.3 Comparative metaproteomic profiling</u>

To investigate the metabolic diversity of microbial communities inhabiting the North Water, we profiled relative protein abundance patterns using a MS/MS-based metaproteomic approach. Searching of 209,159 peptide spectra against a protein database comprised of marine reference genomes, single cell amplified genomes, and cold-water metagenomes resulted in the identification of 8,033 unique peptide sequences mapping to 4,765 proteins (**Supplementary Table 4.1 and Supplementary Table 4.2**). Relative protein abundance patterns were first compared between samples from a functional perspective using the COG database (Tatusov *et al.*, 2000) (**Supplementary Table 4.3**). Hierarchical cluster analysis of the relative abundances of peptide spectra assigned to 379 COG functions revealed a separation of samples into two clusters, consisting of either PML or BTM samples **(Figure 3.3a)**. However, one BTM sample, NC-117 440 m, fell within the PML cluster. Similar separation of BTM and PML samples was observed along axis 1 in a PCo analysis ordination **(Supplementary Figure 4.3)**. These results indicate that the metaproteomic profiles revealed significant differences in functional content associated with the unique microbial communities inhabiting surface and deep Arctic waters.

We then investigated whether or not metaproteomic profiles captured metabolic differences in microbial communities inhabiting the PML of different areas of the North Water. In a PCo analysis of COG functions including PML metaproteomes only, W-PML and E-PML samples were separated from one another along axis 1, with longitude as a significant environmental variable (p-value = 0.003) that correlate strongly with axis 1 (**Figure 4.3b)**. PML samples were also separated along axis 2, which was correlated with latitude, another significant environmental variable (p-value = 0.05).

Although the metaproteomic profiles showed significant functional variation between W-PML, E-PML, and BTM communities, we hypothesized that there were additional COG functions that were common to two or more communities but expressed by taxonomically distinct members. To investigate this we assigned the identified proteins to taxonomic groups, which resulted in a taxonomic structure that was consistent with that obtained through 16S rRNA gene analysis. For instance, relative abundances of Flavobacteriales and Rhodobacterales were higher in the W-PML, relative abundances of SAR11 were higher in the E-PML and the relative abundance of SAR324, Nitrospina, and GSO were higher in the BTM (**Figure 4.3c**). However, there were differences between the datasets that reflect limitations associated with the taxonomic analysis of metaproteomic data; taxa that are not well represented in the protein search database or the reference genome database used for taxonomic assignment will be underrepresented in the metaproteomic profiles compared to the 16S rRNA datasets. This is the case for the Marinimicrobia or ZA2333c Gamma-proteobacteria for example. Nevertheless, the metaproteomic profiles capture some major taxonomic differences between the W-PML, E-PML and BTM communities.

A similar PCo analysis on a larger matrix consisting of taxonomically-resolved COG (tr-COG) functions (**Supplementary Table 4.4**) was then performed allowing for the determination of functional redundancy between communities. In the resulting ordination of all samples, the separation of PML and BTM samples previously observed by COG function alone was amplified (**Supplementary Figure 4.4**). A strong separation between the W-PML and E-PML samples based on the tr-COG functions was also observed and longitude became a marginally more significant explanatory variable (p-value = 0.001) (**Figure 4.3d**), compared to when COG functions were analyzed alone (**Figure 4.3b**). In contrast, the significance of latitude decreased (p-value = 0.056). Overall, these results demonstrate that the metaproteomic profiles captured both the core functional differences between W-PML and E-PML communities as well as the functions that were common across the North Water PML but performed by taxonomically distinct community members.

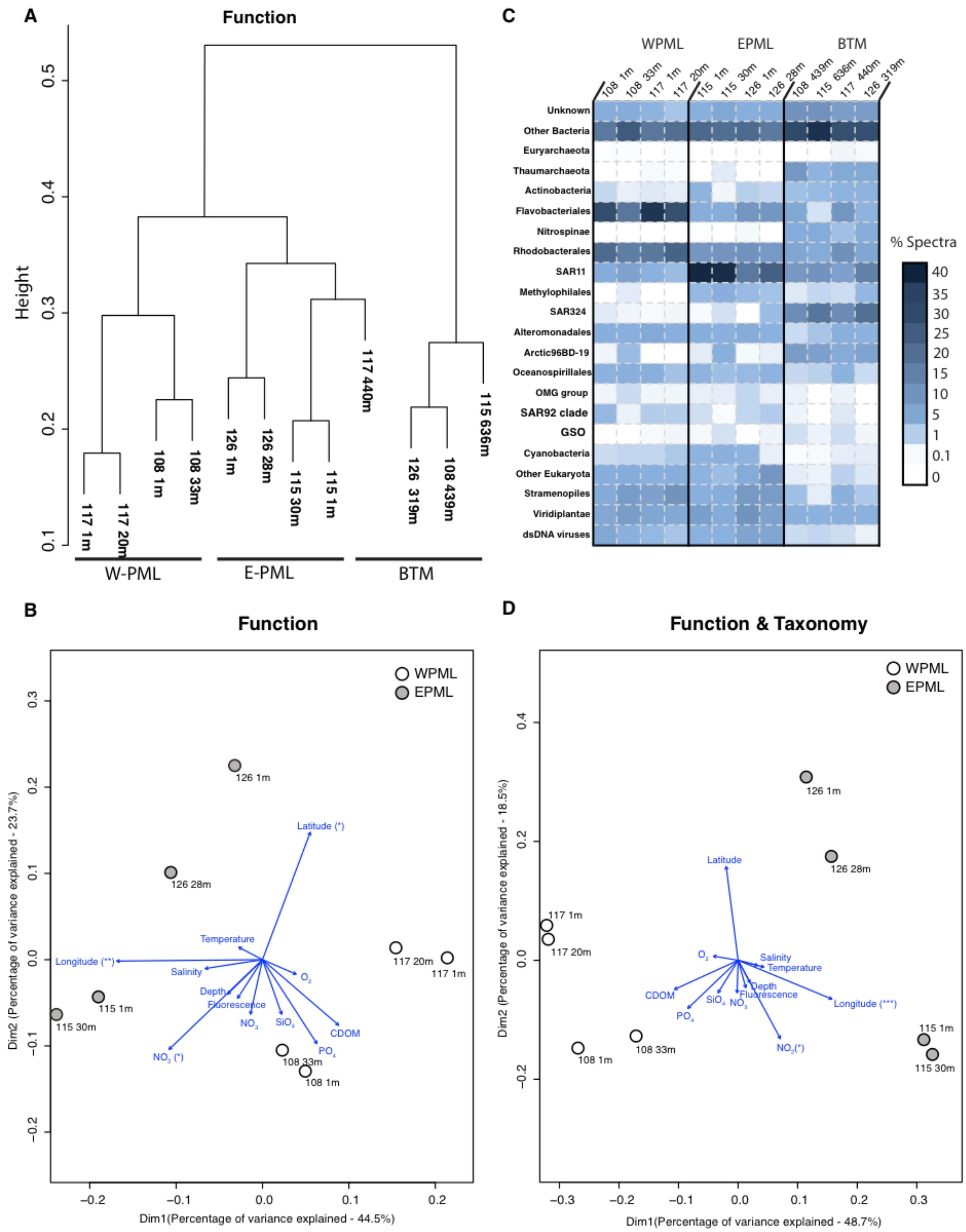


Figure 4.3. Hierarchical cluster analysis of the relative abundances of peptide spectra assigned to 379 COG functions for all 12 samples (**A**). PCo analysis of the relative abundances of peptide

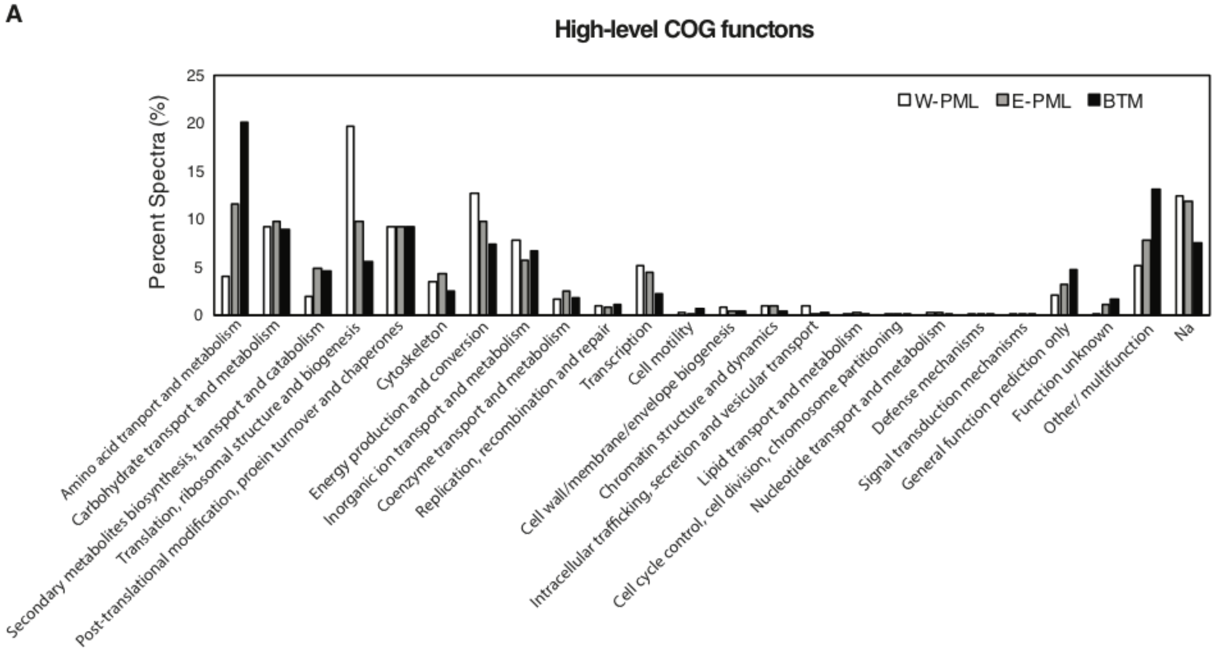
spectra assigned to COG functions including PML metaproteomes only (**B**). Taxonomic composition of the 12 samples, grouped by community (W-PML, E-PML, BTM) based on the relative abundances of peptide spectra assigned to each taxa (**C**). PCo analysis of the relative abundances of peptide spectra assigned to taxonomically resolved COG functions including PML metaproteomes only (**D**).

4.4.4 Differentiation between W-PML and E-PML metaproteomes

A broad range of COG functional categories were represented in the North Water metaproteomes (**Figure 4.4a**), but the relative abundance of peptide spectra assigned to specific protein functions was dissimilar between W-PML, E-PML and BTM samples. To elucidate these differences in more detail, we compared the proportional contribution of COG functions to the communities using $\log_2$ ratios, starting with a comparison between the W-PML and E-PML. Sixty seven of the 379 COG functions exhibited strong differences in their relative contributions (defined as $\log_2$ ratios of > |2|) to the W-PML (41 functions) and E-PML (26 functions) metaproteomes (**Figure 4.4b, Supplementary Table 4.5**). Ribosomal proteins were among the functions most prevalent in the W-PML metaproteomes. In contrast, functions more common in the E-PML metaproteomes were related to amino acid transport and metabolism.

Comparison of tr-COG functions provided further insight into diversity and metabolic activity of microbial taxa in the W-PML and E-PML. Of the 1,105 tr-COG functions, 258 (43%) had a $\log_2$-fold difference > |2| between the two communities (**Supplementary Table 4.6**). Overall, 139 of the total tr-COG functions identified in the PML metaproteomes were represented by four housekeeping proteins (EF-Tu, GroEL, and ATPase alpha and beta subunits). In combination, these four tr-COG functions comprised 18.4% and 16.2% of PSMs in metaproteomes from the W-PML and E-PML, respectively. Of the 139, 16 were found at a $\log_2$-fold difference > |2| between the W-PML and E-PML. The high taxonomic diversity assigned to the four housekeeping proteins (as indicated by the large number of tr-COG functions assigned to the four housekeeping proteins), as well as their high relative abundance in the North Water communities, supports the notion that the taxonomic composition of the different microbial communities in the North Water play a large role in the separation of the three communities observed in the PCoA.

A



B

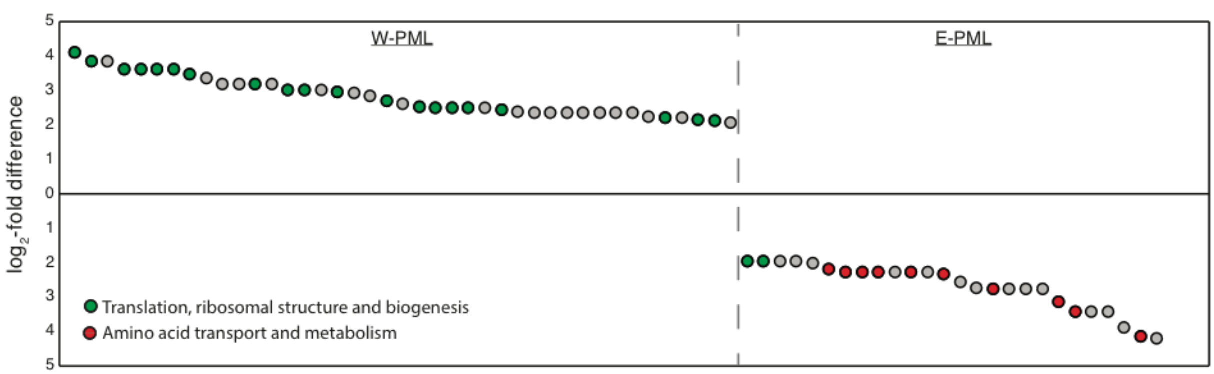


Figure 4.4. Relative abundance of peptide spectra assigned to high-level COG functions for each community (W-PML, E-PML and BTM) (**A**). Log_2 ratios of the Sixty-seven COG functions identified with a Log_2 ratio of > |2|. Green circles represent COGs implicated in translation, ribosome structure and biogenesis and red circles represent COGs implicated in amino acid transport and metabolism (**B**).

4.4.5 Membrane transport

Substrate-binding components of ATP binding cassette (ABC)-type transport systems, as well as tripartite ATP-independent periplasmic (TRAP)-type transport systems and TonB-dependent transport (TBDT) proteins represented a significant portion of the PML metaproteomes with 27.8 % of the W-PML and 39.4 % of the E-PML metaproteomes. Substrate specificities and taxonomic assignment of these transport proteins differed between the W-PML and E-PML. In general, high affinity transport proteins identified in the W-PML were primarily assigned to Rhodobacterales while those identified in the E-PML were primarily assigned to SAR11. Overall, the majority of high affinity transport proteins identified in the North Water PML were identified at a higher relative abundance in the E-PML and assigned primarily to SAR11 (**Figure 4.5a**). This included transport proteins for amino acids (2.39 log_2-fold difference), including proline/glycine betaine (4.37 log_2-fold difference), branched-chain amino acids (6.80 log_2-fold difference), the polyamines spermidine/putrescine (3.58 log_2-fold difference), tricarboxylate transport proteins (5.7 log_2-fold difference) and general sugar transport proteins **(Figure 4.5a, Figure 4.5b)**. Additionally, TRAP-type mannitol/chloroaromatic compound transporters were identified ~2.5-fold more in the E-PML than the W-PML and represented a large portion of the E-PML metaproteome (4.83 %). The majority of the mannitol/chloroaromatic compound transport proteins were assigned to SAR11 (3.13 log_2-fold difference). In contrast, only two transport proteins, the ABC-type transport proteins for xylose and Fe^{3+}, were identified at a higher relative abundance in the W-PML and were assigned primarily to Rhodobacterales.

In addition to SAR11 and Rhodobacterales, transport proteins assigned to other taxa were also identified primarily in one of the two PML communities. General sugar transport proteins assigned to unknown Beta-proteobacteria and Arctic96BD-19, as well as oligopeptide, amino acids and proline/glycine betaine transport proteins assigned to unknown Gamma-proteobacteria, and branched-chain amino acid transport proteins assigned to unknown Beta-proteobacteria were only identified in the E-PML. Spermidine/putrescine ABC-type transport proteins assigned to unknown Alpha-proteobacteria, unknown Gamma-proteobacteria and SAR324 were also only identified in the E-PML. Although dipeptide transport proteins were identified at a higher relative abundance in the E-PML, they were primarily assigned to Rhodobacterales in both regions. However, additional dipeptide transport proteins assigned to SAR11, Arctic96BD-19 and SAR324 were exclusively identified in the E-PML. General sugar transport proteins assigned to Rhizobiales, Oceanospirillales and unknown Gamma-proteobacteria, were exclusively identified in the W-PML while xylose transport proteins assigned to unknown Alpha-proteobacteria were identified at a higher relative abundance in the W-PML (log_2-fold differences of 3.27). Additionally, branched-chain amino acid transport proteins assigned to Alteromonadales were identified to a greater degree in the W-PML (log_2-fold differences of 1.53). Taurine transport proteins were common to the PML but identified to a higher degree in the E-PML with proteins assigned to SAR324 and Arctic96BD-19 only identified in the E-PML. In addition to SAR11, Mannitol/chloroaromatic transport proteins assigned to unknown Alpha-proteobacteria were also identified at higher relative abundances in the E-PML (2.07. log_2-fold difference).

Glycerol-3-phosphate transport proteins were identified at a similar relative abundance between the two PML communities. The main difference between the two regions was a higher relative abundance of Rhodobacterales assigned proteins in the W-PML (1.81 log_2-fold difference) and the high relative abundance of SAR11 assigned proteins identified exclusively in the E-PML (1.16 %). However, differences also included a higher relative abundance of unknown Alpha-proteobacteria, Alteromonadales and unknown Gamma-proteobacteria in the W-PML and a higher abundance of SAR324, Oceanospirillales, Arctic96BD-19 and OMG group in the E-PML. Phosphate/phosphonate ABC-type transport proteins were also identified at similar relative abundances on either side of the PML. Those assigned to Rhodobacterales were common to the PML, while those assigned to Thaumarchaeota were only identified in the E-PML. Similarly, C4-dicarboxylate transport proteins were identified at a similar relative abundance between the two PML communities but those transport proteins assigned to Rhodobacterales were more abundant in the W-PML (2.44 log_2-fold difference) and those assigned to SAR11 were more prevalent in the E-PML (5.73 log_2-fold difference). Transport proteins for long-chain fatty acids were common in the PML with those assigned to Chromatiales identified only in the E-PML and those assigned to Flavobacteriales only identified in the W-PML.

A striking number of TBDT proteins were identified in the North Water PML. In fact, 9.36 % and 6.43 % of the W-PML and E-PML metaproteomes were identified as TBDT proteins respectively (**Figure 4.5a**). TBDT proteins are known to transport high molecular weight compounds (>600 daltons) like Ni-, Cu-, Fe-chelates, proteins, vitamins B_1, cobalamin (vitamin B_{12}), and polysaccharides across the outer membrane (Schauer *et al.*, 2008). TBDT proteins annotated as transporting iron-complexes, biopolymers and cobalamin were identified in the PML and a difference in the relative contribution of certain taxa to the expression of these proteins between regions was observed. The majority of TBDT proteins identified in the W-PML were assigned to Flavobacteriales while those in the E-PML were assigned to a broader array of taxa, including more pronounced contributions from Alteromonadales, the OMG group of Gamma-proteobacteria, unknown Gamma-proteobacteria and unknown Proteobacteria.

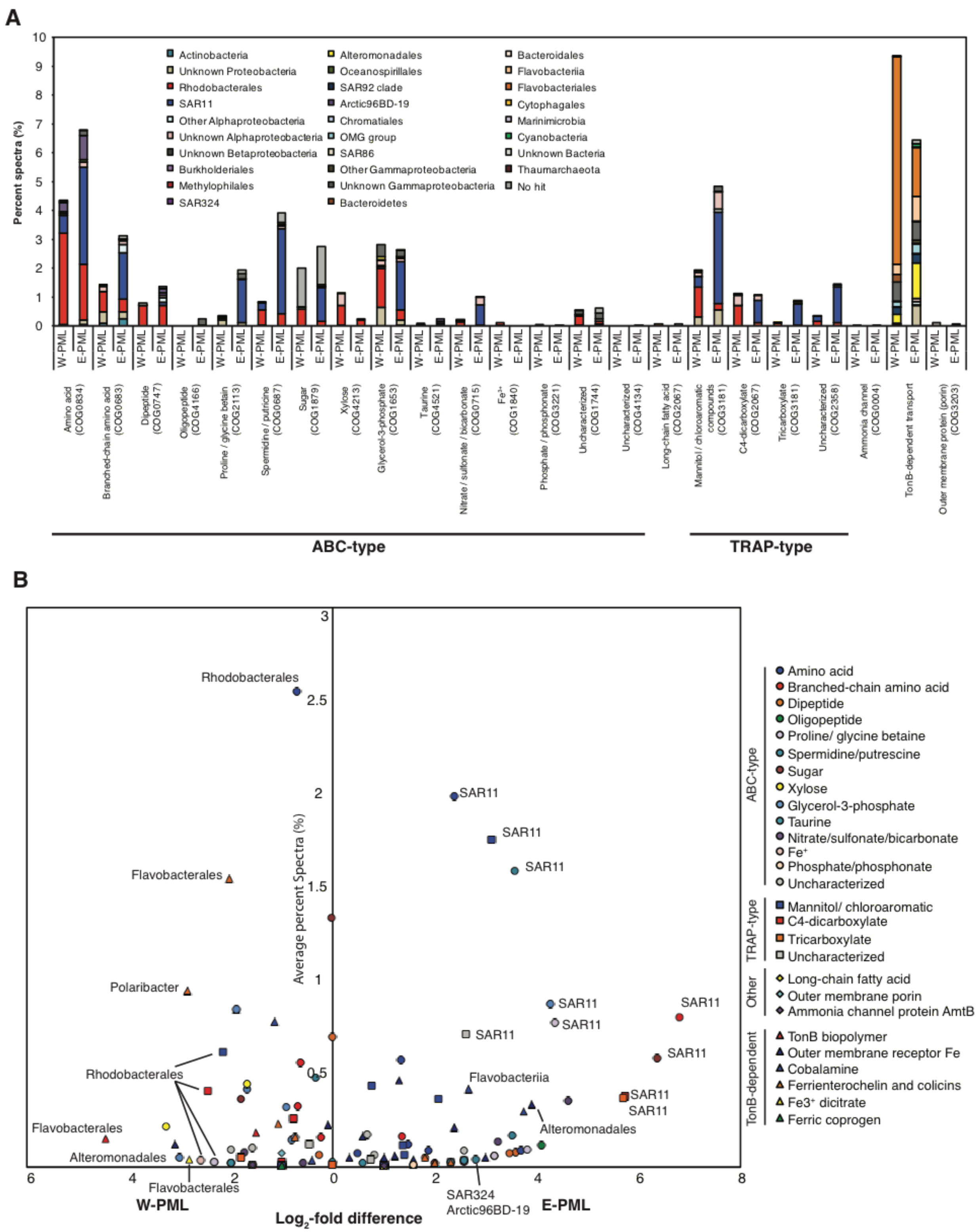

A
Percent spectra (%)
Actinobacteria
Unknown Proteobacteria
Rhodobacterales
SAR11
Other Alphaproteobacteria
Unknown Alphaproteobacteria
Unknown Betaproteobacteria
Burkholderiales
Methylophilales
SAR324
Alteromonadales
Oceanospirillales
SAR92 clade
Arctic96BD-19
Chromatiales
OMG group
SAR86
Other Gammaproteobacteria
Unknown Gammaproteobacteria
Bacteroidetes
Bacteroidales
Flavobacteriia
Flavobacteriales
Cytophagales
Marinimicrobia
Cyanobacteria
Unknown Bacteria
Thaumarchaeota
No hit
W-PML
E-PML
Amino acid (COG0834)
Branched-chain amino acid (COG0683)
Dipeptide (COG0747)
Oligopeptide (COG4166)
Proline / glycine betain (COG2113)
Spermidine / putricine (COG0687)
Sugar (COG1879)
Xylose (COG4213)
Glycerol-3-phosphate (COG1653)
Taurine (COG4521)
Nitrate / sulfonate / bicarbonate (COG0715)
Fe^{3+} (COG1840)
Phosphate / phosphonate (COG3221)
Uncharacterized (COG1744)
Uncharacterized (COG4134)
Long-chain fatty acid (COG2067)
Mannitol / chloroaromatic compounds (COG3181)
C4-dicarboxylate (COG2067)
Tricarboxylate (COG3181)
Uncharacterized (COG2358)
Ammonia channel (COG0004)
TonB-dependent transport
Outer membrane protein (porin) (COG3203)
ABC-type
TRAP-type
B
Average percent Spectra (%)
Log$_2$-fold difference
W-PML
E-PML
Rhodobacterales
SAR11
Flavobacterales
Polaribacter
Flavobacteriia
Alteromonadales
SAR324
Arctic96BD-19
Amino acid
Branched-chain amino acid
Dipeptide
Oligopeptide
Proline/ glycine betaine
Spermidine/putrescine
Sugar
Xylose
Glycerol-3-phosphate
Taurine
Nitrate/sulfonate/bicarbonate
Fe^{+}
Phosphate/phosphonate
Uncharacterized
Mannitol/ chloroaromatic
C4-dicarboxylate
Tricarboxylate
Uncharacterized
Long-chain fatty acid
Outer membrane porin
Ammonia channel protein AmtB
TonB biopolymer
Outer membrane receptor Fe
Cobalamine
Ferrienterochelin and colicins
Fe^{3+} dicitrate
Ferric coprogen
ABC-type
TRAP-type
Other
TonB-dependent

Figure 4.5. The relative abundance of solute transport proteins based on spectra counts identified in each of the two PML communities (W-PML and E-PML) separated by taxa (**A**). Log_2 ratios versus average % abundance based on spectra counts of all identified transport tr-COGs (**B**).

4.4.6 Inorganic nitrogen transport and assimilation

Differences in transport proteins for inorganic nitrogen compounds were observed between the W-PML and E-PML. Of the 39 nitrate/sulfonate/bicarbonate ABC-transporters, 15 were confirmed to be nitrate transport proteins by querying the 39 transport proteins against Refseq nr (using BLASTp). Nitrate transport proteins were common to the PML, but found at a relative abundance of ~2-fold higher in the E-PML. Those assigned to Rhizobiales and Flavobacteriales were only identified in the W-PML, and those assigned to Rhodobacterales were identified most often in the W-PML. The majority of nitrate transport proteins identified in the E-PML were assigned to unknown Alpha-proteobacteria. Unlike nitrate transport proteins, ammonia channel proteins (AmtB) were common to the PML and expressed at a similar relative abundance in the W-PML and E-PML. Ammonia transport proteins were only assigned to SAR11 and unknown Alpha-proteobacteria. Those assigned to SAR11 were only identified in the E-PML, and those assigned to unknown Alpha-proteobacteria only identified in the W-PML. Proteins involved in ammonia assimilation, including aspartate aminotransferase and glutamate synthase proteins were exclusively identified in the E-PML and assigned to a variety of taxa. Aspartate aminotransferase proteins were assigned to unknown Proteobacteria while glutamate synthase proteins were assigned to unknown Proteobacteria, SAR11, Actinobacteria and Rhodobacterales.

4.4.7 Metabolism of one carbon (C1) compounds

Proteins assigned to COG functions involved in C1 metabolism were identified at a higher relative abundance in the E-PML than the W-PML. Seventeen proteins in the dataset were annotated as PQQ-dependent dehydrogenases, a protein family that includes glucose dehydrogenases, methanol dehydrogenases and other alcohol dehydrogenases. Phylogenetic analysis revealed that 12 of the 17 were most closely related to methanol dehydrogenase proteins

(**Supplementary figure 4.1**). Two of the 12 were identified in all three North Water communities, 7 were exclusively identified in the E-PML, 2 were exclusively identified in the BTM and 1 was identified in the E-PML and BTM. In the PML, methanol dehydrogenase proteins were exclusively assigned to Methylophilales at a high log_2-fold difference of 4.21. All 10 of the methanol dehydrogenase proteins identified in the PML fell within the XoxF4 clade of methanol dehydrogenase proteins. However, the two methanol dehydrogenase proteins identified exclusively in the BTM were assigned to Rhodobacterales and Thiotrichales and fell within the XoxF5 clade. Formate dehydrogenase (anaerobic selenocysteine-containing dehydrogenase) proteins assigned to Rhodobacterales and Alteromonadales were also identified exclusively in the E-PML. Additionally, CO or xanthine dehydrogenase subunit proteins were also identified in the North Water PML and assigned to unknown Alpha-proteobacteria and Rhodobacterales. However, no great difference in relative abundance between the W-PML and E-PML was observed.

4.4.8 BTM community metaproteomes

In contrast to the W-PML, but similar to the E-PML, the most abundant proteins identified in the BTM were those involved in amino acid transport (12.9 % spectra) followed by the chaperonin GroEL (6.65 % spectra). Proteins involved in substrate transport, translation elongation (EF-Tu), protein folding (GroEL) and dissimilatory nitrogen metabolism (dissimilatory nitrate reductase) were highly expressed in the bottom waters, representing the top 15 proteins in terms of relative abundance.

COG functions from the BTM were further divided by taxonomy, resulting in 433 tr-COG functions. Proteins assigned to SAR324, SAR11, Rhodobacterales, Arctic96BD-19, MGI Thaumarchaeota, Flavobacteriales and Nitrospinae were all expressed at high relative abundances in the BTM. Additionally, high relative abundance of proteins were assigned to unknown Gamma-proteobacteria and Alpha-proteobacteria. Among the top 10 tr-COG functions were those involved in amino acid transport (Arctic96BD-19, unknown Alpha-proteobacteria, SAR11 and Rhodobacterales), dipeptide transport (SAR324), proline/glycine betaine transport (unknown proteobacteria, unknown Gamma-proteobacteria and SAR11), nitrate reductase (Nitrospinae) and GroEL (SAR324).

4.4.8.1 SAR324 and Arctic96BD-19

Tr-COG functions assigned to the metabolically versatile bacteria SAR324 and Arctic96BD-19 were among the most highly identified in the BTM. Tr-COG functions assigned to SAR324 were the most abundantly identified with 14.5 % of spectra while those assigned to Arctic96BD-19 were among the top 10 with 5.52 % of spectra. More than half of the relative abundance of proteins assigned to SAR324 (~64 %) were devoted the membrane transport of various substrates including oligopeptides, amino acids, branched-chain amino acids, proline/glycine betaine, spermidine/putricine, maltose, glycerol-3-phosphate, mannitol/chloroaromatics, taurine, tricaboxylates, nitrate/sulfonate/bicarbonate, phosphate/phosphonate, Fe^{3+} and biopolymers. An even larger percentage of the Arctic96BD-19 proteome in the BTM was mapped to transport proteins with ~98 % of all spectra assigned to Arctic96BD-19 identified as involved in amino acid, glycerol-3-phosphate, taurine, tricarboxylates, sugars or mannitol/chloroaromatics transport.

4.4.8.2 Nitrification

Ammonia channel proteins (AmtB), and an ammonia monooxygenase protein assigned to the ammonia oxidizing MGI Thaumarchaeota were also identified in the BTM. Initial COG annotation of the identified proteins revealed no MGI Thaumarchaeota-assigned ammonia monooxygenase proteins. This was surprising considering the known role MGI Thaumarchaeota play in nitrification in deep marine waters (Wuchter *et al.*, 2006; Swan *et al.*, 2014; Hallam *et al.*, 2006). The proteins assigned to MGI Thaumarchaeota and identified as having no COG functional category were then queried against Refseq nr (using BLASTp) and one ammonia monooxygenase/methane monooxygenase protein was identified in the BTM. Additionally, 4 multi-copper oxidases assigned to MGI Thaumarchaeota were identified in the BTM and could be implicated in ammonia oxidation (Kozlowski *et al.*, 2016).

Dissimilatory nitrate reductase proteins assigned to Nitrospinae, were among the most highly identified proteins in the BTM in terms of % spectra, and were common throughout the BTM. Proteins assigned to Nitrospinae and annotated as nitrite oxidoreductase alpha and beta subunits both represented 1.8 % of each of the BTM metaproteome. Nitrite oxidation in Nitrospinae has also been linked to autotrophic CO_2 fixation via the rTCA cycle, but the only enzyme associated to the rTCA cycle identified in the BTM was pyruvate:ferredoxin oxidoreductase and it could only be assigned to unknown Bacteria. In addition to nitrite oxidoreductase proteins, ammonia channel proteins were also identified and assigned to Nitrospinae.

4.5 Discussion

Given the sensitivity of the Arctic to climate change and the key role the North Water plays in linking the Arctic-Atlantic systems, thereby affecting the nutrient budget of the Atlantic Ocean, an understanding of its microbial community composition as well as the metabolisms they mediate is needed. Microbial community structure and function across the North Water was analyzed using a combination of 16S rRNA gene sequencing and metaproteomics. Clear distinctions between the western and eastern polar mixed layers in terms of community structure and community metabolism were observed. In general, the W-PML was comprised of bacterioplankton typically associated with dynamic ecosystems with pulses of primary production and the metaproteomes were enriched in translational machinery. In contrast, the E-PML had a higher prevalence of typical oligotrophic bacteria found in more stable environments like SAR11 and the metaproteomes were enriched in proteins involved in amino acid metabolism.

4.5.1 Distinct W-PML and E-PML community composition and resource allocation

Although all sampling locations in this study are from the same ecozone, in close geographic proximity to one another, and essentially consist of productive coastal waters, the western and eastern side of the North Water are comprised of water of differing origins. Generally, polar surface waters on the Western side of the North Water are richer in silicate, phosphate and fluorescent dissolved organic matter FDOM (Joli *et al.*, 2018) and flow southward, while the eastern waters of the North Water are richer in nitrate and flow northward from the Atlantic Ocean. Like most polynyas, the North Water has a Fall phytoplankton bloom due to a long open-water period allowing phytoplankton to exploit nutrients supplied by the late season increase in convective mixing and upwelling (Ardyna *et al.*, 2011). The majority of primary production in the North Water occurs in the Eastern region (Klein *et al.*, 2002; Tremblay *et al.*, 2002) between May and July. Following nitrate exhaustion and the collapse of the bloom in late June, a secondary bloom can be observed in September. We observed a more prominent subsurface chlorophyll maximum (as inferred by fluorescence) and a corresponding oxygen maximum peak in the E-PML (primarily at SG-115) in association with a well-stratified water column and a sharper halocline. A higher proportion of proteins assigned to Cyanobacteria, as well as a higher proportion of 16S rRNA genes assigned to Cyanobacteria and chloroplasts was observed in the E-PML. Additionally, a dynamic phytoplankton community structure over a short time-frame of 24 hours in the W-PML in conjunction with a stable phytoplankton community structure in the E-PML was observed at the time of sampling (Joli *et al.*, 2018). These different dynamics are likely due to the more highly stratified water column in the E-PML and a higher degree of mixing of different water masses in the W-PML (Joli *et al.*, 2018). The dynamic nature of the W-PML was also exemplified by the observation of a phytoplankton bloom just north of the W-PML sampling sites (personal correspondence with members of the Jean-Eric Tremblay lab).

The bloom-like dynamics of the W-PML likely results in pulses of phytoplankton-derived dissolved organic matter (DOM) to the W-PML and could account for the observed higher proportion of proteins assigned to bacterioplankton that are typically associated with more dynamic ecosystems like Rhodobacterales, Alteromonadales, Flavobacteriales, Chromatiales, Oceanospirillales and SAR92. These organisms generally have more versatile metabolisms and are capable of transporting higher molecular weight compounds, and are therefore better adapted to respond to nutrient pulses. In contrast, the E-PML community was characterized by organisms typically associated with more stable environments like the open ocean, such as SAR11, SAR86, as well as methylotrophic Beta-proteobacteria (**Figure 4.2a**). These findings highlight the importance of continuous time-series sampling, as well as broad geographical sampling in order to identify and understand the dynamic relationships between microbial community structure, DOM production and hydrological dynamics of the water column.

In general, the COG functional categories most represented in the North Water metaproteomes were translation, ribosomal, structure and biogenesis, post-translational modification, protein turnover, and chaperones, energy production and conversion, amino acid transport and carbohydrate transport and metabolism (**Figure 3.4a**). However, as expected based on the multivariate statistical analyses presented above, the relative abundance of peptide spectra assigned to specific protein functions was dissimilar between W-PML, E-PML and BTM samples, suggesting differences in metabolic strategy and cellular physiology between these microbial communities.

A distinguishing characteristic between the W-PML and E-PML was the relative abundance and diversity of ribosomal proteins identified. The W-PML communities seemed to invest relatively more cellular resources in ribosome biogenesis and cell division while the E-PML communities invested relatively more cellular resources in energy conservation and nutrient acquisition. The difference in the number of identified ribosomal proteins could be due to several factors, including 1) the taxonomic differences in community composition between the W-PML and E-PML and 2) the cellular characteristics of the two communities, including cell size, physiological state and growth rate. Not all bacteria have the same complement of ribosomal proteins, however, of the 19 bacterial ribosomal proteins that were identified at a $\log_2$-fold change of 2 or greater, 18 were found in all 995 bacterial genomes analyzed by Yutin et al. (Yutin et al. 2012) and one was identified in 994 of the bacterial genomes (not found in *Mycoplasma penetrans*). The ribosomal proteins were also assigned to a broad taxonomic diversity. It is therefore unlikely that this difference in ribosomal protein abundance was solely due to taxonomic differentiation between the W-PML and E-PML.

However, differences in cellular characteristics like cell size, physiological state and growth rate between the different communities could still be a contributing factor. Growth rate and the physiological state of an organism are linked. Doubling times in exponential phase are shorter than in stationary phase, and faster-growing cells require higher rates of protein synthesis which necessitates higher numbers of ribosomes and thus more ribosomal proteins (Schaechter *et al.*, 1958; Kemp *et al.*, 1993; Molenaar *et al.*, 2009). Because of this, it is conceivable to use the relative number of ribosomal proteins found between environments to determine the growth rate of the organisms, similarly to studies that correlate RNA content with growth rate (Kemp *et al.*, 1993). Therefore, the higher number and greater diversity of ribosomal proteins identified in the W-PML could indicate a faster growing community than in the E-PML. Furthermore, a shift from the expression of proteins related to cell growth, cell division and protein biosynthesis to proteins involved in nutrient scavenging is observed in model organisms transitioning from exponential growth to stationary phase (Folio *et al.*, 2004; Houser *et al.*, 2015). In environmental isolates, a decrease in the expression of genes involved in protein and RNA metabolism with a corresponding increase in the expression of genes involved in membrane transport, carbohydrate and amino acid metabolism, among others (Muthusamy *et al.*, 2017) are also observed. However, the W-PML community is also primarily comprised of organisms that are generally larger and faster-growing compared to those organisms that dominate the E-PML, and cell size can also play a role in determining the number of ribosomes present in a cell. Although difficult to assess because of differences in ribosome abundance caused by differences in growth rate, in general, smaller cells contain less ribosomes than larger cells (Zhao *et al.*, 2017; Kemp *et al.*, 1993).

In summary, the higher relative abundance of ribosomal proteins identified in the W-PML, a community made up of taxa typically associated with productive waters supports the idea that the W-PML contains more available nutrients, allowing for a community of larger, more rapidly dividing cells than in the more stable E-PML, where nutrient pulses are more rare. Due to this the W-PML community can invest more cellular resources in ribosome biogenesis and cell division while E-PML communities invest more cellular resources in energy conservation and nutrient acquisition instead of costly ribosomal protein and rRNA biosynthesis.

4.5.2 Divergent strategies for nutrient acquisition

Although both the W-PML and E-PML are within productive coastal waters of the same eco-region, an enrichment of TonB-dependent transport proteins and xylose transport proteins were identified in the W-PML. In contrast, the E-PML was characterized by an enrichment in transport proteins for nitrogen-containing compounds like oligopeptides, branched-chain amino acids, amino acids, proline/glycine betaine, spermidine/putrescine and nitrate. The difference in transport protein profiles suggests divergent strategies for nutrient acquisition between the two communities and correspond to those observed during phytoplankton blooms and the metabolic succession that follows. Specifically, the initial increase in the expression of proteins involved in the degradation and transport of high molecular weight DOM (TBDT) followed by the expression of high affinity transporters to compete for smaller molecules once the initial nutrient pulse has worn off (Teeling *et al.*, 2012; Williams *et al.*, 2012). In general, porins, permeases and major facilitator superfamily proteins, especially those for nitrogenous compounds like nitrate, nitrite, urea and ammonium were more common in less productive samples and transport proteins for dissolved organic carbon compounds like taurine, and carboxylates, as well as phosphonate transporters and TonB dependent transporters are more abundant in productive coastal water samples (Morris *et al.*, 2010).

Based on these findings, the W-PML community seems to rely more heavily on the scavenging of phytoplankton-derived HMW DOM associated with phytoplankton bloom and decay, while the E-PML community utilizes a higher proportion of high-afinity membrane transport proteins, perhaps as a means of cycling LMW DOM excreted by phytoplankton and other bacterioplankton.

Additionally, transport proteins in the E-PML metaproteome were predominately assigned to SAR11 and involved in the transport of amino acids, spermidine/putrescine, mannitol/chloroaromatic compounds and the compatible osmolytes proline/glycine betaine. Similar patterns of high-affinity membrane transport protein abundances were identified in SAR11 proteomes from the oligotrophic Sargasso Sea (Sowell *et al.*, 2009). However, unlike in the oligotrophic Sargasso Sea, where phosphate transport proteins were the most abundantly detected proteins, very few phosphate/phosphonate transport proteins from SAR11 were identified in the North Water. SAR11 in the North Water instead invested more resources in the acquisition of organic nutrients, including nitrogen containing compounds. Even though phosphate/phosphonate transport proteins represented a small fraction of the North Water metaproteome, they were still identified to a higher degree in the E-PML than the W-PML, so although the North Water might not be phosphorus-limited, there seems to be greater competition for phosphate in the E-PML than in the W-PML.

4.5.3 Methanol oxidation differentiates the W-PML from the E-PML

In addition to their divergent strategies for nutrient acquisition, methanol oxidation also differentiated the W-PML and E-PML communities. The key enzyme responsible for the initial oxidation of methanol, methanol dehydrogenase, was identified primarily in the E-PML and assigned to Methylophilales. Methanol is an abundant organic volatile compound in marine ecosystems and can serve as a carbon and energy source for methylotrophic microorganisms. Major sources of atmospheric methanol include both plant growth (Fall and Benson, 1996; Millet *et al.*, 2008), and phytoplankton growth (Mincer and Aicher, 2016). Atmospheric methanol as well as phytoplankton-derived methanol are the major contributors to marine methanol (Millet *et al.*, 2008). We hypothesize that the stable nature of the E-PML allows for relatively constant phytoplankton production which would lead to stable methanol production, allowing for the establishment of methylotrophs within this community. On the other hand, the higher degree of mixing and dynamic nature of the W-PML, as well as the high biological turnover rates of methanol already observed in marine systems (Joanna L Dixon *et al.*, 2011; Dixon *et al.*, 2013) may limit the establishment of obligate methylotrophs in that community.

Evidence for methanol oxidation was also observed in the BTM. Methanol dehydrogenase proteins assigned to Methylophilales, Rhodobacterales and Thiotrichales were identified in the bottom North Waters, although to a lesser degree than in the PML. The source of methanol in these deep waters has not yet been resolved. One possibility is that methanol production could persist in sinking phytoplankton detritus (Mincer and Aicher, 2016). In the absence of phytoplankton blooms and close proximity to terrestrial runoff, methanol may also form as a product of fermentation in anoxic microenvironments or from the oxidation of methane (Krause *et al.*, 2016). Another possible source of methanol in the deep North Water could be the demethylation of organic compounds in the deep waters. Lastly, the identification of methanol dehydrogenase proteins in the BTM could also be due to the sinking of methylotrophic microbial cells from the productive surface layer.

4.5.4 TonB-dependent transport is important in the North Water PML

A striking number of TBDT proteins were identified in the North Water PML. Although TBDT proteins were traditionally thought to be involved in the transport of iron complexes and cobalamine, recent genomic and proteomic studies have determined that the breadth of substrates that TBDT proteins can transport is actually much greater and includes the transport of maltose and maltodextrins (Neugebauer *et al.*, 2005) and algal-derived polysaccharides like laminarin and alginate (Kabisch *et al.*, 2014; Unfried *et al.*, 2018). Other studies have suggested a role of TBDT in transporting dissolved proteins or oligopeptides (Orsi *et al.*, 2016). Therefore, although the TBDT proteins identified in the North Water are primarily annotated as being involved in the transport of iron complexes and cobalamin, it is conceivable, and likely, that they represent transporters of higher substrate diversity.

Some of these more novel functions of TBDT proteins were discovered through experimentation (Blanvillain *et al.*, 2007), but many of the substrate specificities identified have been inferred from the genomic context of the TBDT gene (i.e. found within a cluster of genes involved in the metabolism of a particular compound) (Schauer *et al.*, 2008; Kabisch *et al.*, 2014). Because this study was a metaproteomic study, it is difficult to determine actual TBDT protein substrate specificity by genomic context. However, by using a combined metagenomic-metaproteomic approach where metaproteomes are searched against a database that includes the metagenome of the same sample, it would be possible to gain insight into the expression level of TBDT proteins as well as the genomic context in which they are found, to better determine their substrate specificity.

4.5.5 The metaproteome of the BTM is distinct from the PML

Proteins responsible for membrane transport represented a large fraction of the metaproteome for the BTM community, with 58.9 % of spectra. This is considerably higher than in the PML communities (27.79 % in the W-PML and 36.68 % in the E-PML). A similar increase in membrane transport protein abundance with depth was reported in the Atlantic Ocean (Bergauer *et al.*, 2017), with transport proteins representing ~23 %, 32 % and 39 % abundances in the euphotic, mesopelagic and bathypelagic waters respectively. Similar to the Atlantic Ocean communities, a depth-dependent stratification of substrate-responsive phyla was observed in the North Water. Most notably, a higher relative abundance of transport proteins assigned to Alpha-proteobacteria and Bacteroidetes was identified in the PML, while those assigned to Gamma-proteobacteria (predominantly Arctic96BD-19) were identified in the BTM. However, although the relative abundance of transport proteins assigned to SAR11 in the BTM was lower than the E-PML, it was higher than in the W-PML. This could be due to sinking of SAR11 cells from the more SAR11-rich E-PML community, or because the E-PML and BTM environments are more similar than the BTM and W-PML environments in terms of available organic matter, and therefore support the growth of similar organisms.

A higher relative proportion of Arctic96BD-19 and SAR324 assigned proteins were identified in the BTM than the PML. Genomic analysis of the Arctic96BD-19 clade of Gamma-proteobacteria and the SAR324 clade of Delta-proteobacteria have demonstrated that they are both metabolically versatile lineages, possessing genes for the membrane transport of sugars and amino acids, autotrophic CO_2 fixation, and the oxidation of reduced sulfur compounds (DA Walsh *et al.*, 2009; Swan *et al.*, 2011; Sheik *et al.*, 2013; Mattes *et al.*, 2013). Therefore, it is hypothesized that these lineages possess a mixotrophic lifestyle. The high abundance of membrane transport proteins for organic substrates assigned to these lineages and the lack of proteins for sulfur oxidation or CO_2 fixation pathways suggests that SAR324 and Arctic96BD-19 preferentially scavenge carbon compounds in the deep North Water. Additionally, the high relative abundance of proteins assigned to these metabolically versatile clades indicated that they are likely important contributors to the BTM community in the North Water.

One of the main metabolic processes that differentiates the PML and the BTM is nitrification. Tr-COG functions involved in nitrification were only identified in the BTM and were among those proteins identified with the highest relative abundance in the BTM. The first step in nitrification is the oxidation of ammonia to nitrite by either ammonia oxidizing bacteria (AOB) or archaea (AOA), followed by the oxidation of nitrite to nitrate by nitrite oxidizing bacteria. Work investigating ammonia monooxygenase gene abundance in the North Water observed the highest ammonia monooxygenase abundances in water masses corresponding to the Arctic Basin halocline (ABH) intrusions, made up of mixing upper Pacific-derived (Arctic) and lower Atlantic waters (Galand *et al.*, 2009). Unfortunately ammonia oxidation abundance in the bottom waters were not explored in the 2009 study and the mesopelagic waters of the ABH were not analyzed in the current study making comparisons between the two difficult. However, the relatively high number of ammonia monooxygenase genes identified in the mesopelagic waters making up the ABH and the low relative abundance of ammonia monooxygenase proteins identified in the BTM waters in this study could indicate a greater importance of ammonia oxidation in the mesopelagic waters of the North Water than in the bottom waters. Ammonia oxidation in archaea can be coupled to autotrophic CO_2 fixation via the 3-hydroxypropionate/4 hydroxybutyrate pathway (Berg *et al.*, 2010), but none of the enzymes involved in the 3-hydroxypropionate/4 hydroxybutyrate pathway were identified the BTM. However, the lack of identified transporters for organic compounds assigned to Thaumarchaetoa, would suggest a reliance on autotrophic carbon assimilation for Thaumarchaeota communities in the BTM.

The second step in nitrification is the oxidation of nitrite to nitrate by nitrite oxidizing bacteria. A striking amount of nitrite oxidoreductase protein subunits A and B assigned to Nitrospinae, a nitrifying bacteria common to the oceans (Lücker *et al.*, 2013; Fuchs *et al.*, 2005; Labrenz *et al.*, 2007; DeLong *et al.*, 2006) that utilizes the NxrABC protein complex (a member of the large DMSO reductase enzyme family) (Jormakka *et al.*, 2004), were identified exclusively in the BTM. In fact, nitrite oxidoreductase proteins assigned to Nitrospinae represented one of the tr-COG functions with the highest relative abundance identified in the BTM. Genomic data from *Nitrospina gracilis* has also revealed the potential for autotrophic CO_2 fixation coupled to nitrite oxidation via the reductive tricarboxylic acid (rTCA) cycle (Lücker *et al.*, 2013; Watson and Waterbury, 1971). Pyruvate:ferredoxin oxidoreductase (POR), one of the key enzymes of the rTCA cycle, was identified in the bottom waters of the North Water suggesting that, Nitrospinae could play a role in chemosynthetic production in the deep waters of the Canadian North Water. The high relative abundance of nitrite oxidoreductase proteins identified in the BTM are indicative of high nitrite oxidation to nitrate in the bottom waters. Given that the North Water is the most productive region in the Arctic and dissimilatory ammonia oxidation to nitrate is an important source of nitrate for primary production it is possible that nitrification by MGI Thaumarchaeota and Nitrospinae in the deeper waters are important drivers of primary production in the North Water PML.

4.5.6 Functional redundancy within different North Water communities

In this study we assigned identified proteins to taxonomically resolved COGs in order to better understand which functions were characteristic of each region, which functions were redundant between regions, and which populations of organisms were contributing to the expression of proteins related to any given function. The three North Water community types were more dissimilar from one another when tr-COG functions, rather than when COG functions alone, were used to construct PCo ordinations. This increased dissimilarity when taxonomy is included as a variable indicates some degree of functional redundancy between the communities. However, the significance of latitude diminished when tr-COG functions were used instead of COG functions (p-value of 0.094 to 0.792), while the significance of longitude and depth increased (p-value of 0.292 to 0.010 and 0.002 to 0.001). The slight increase in the significance of depth between the COG and tr-COG PCo analyses for all three communities compared to the large increase in significance of longitude indicates lower functional redundancy between the PML and the BTM compared to the W-PML and E-PML.

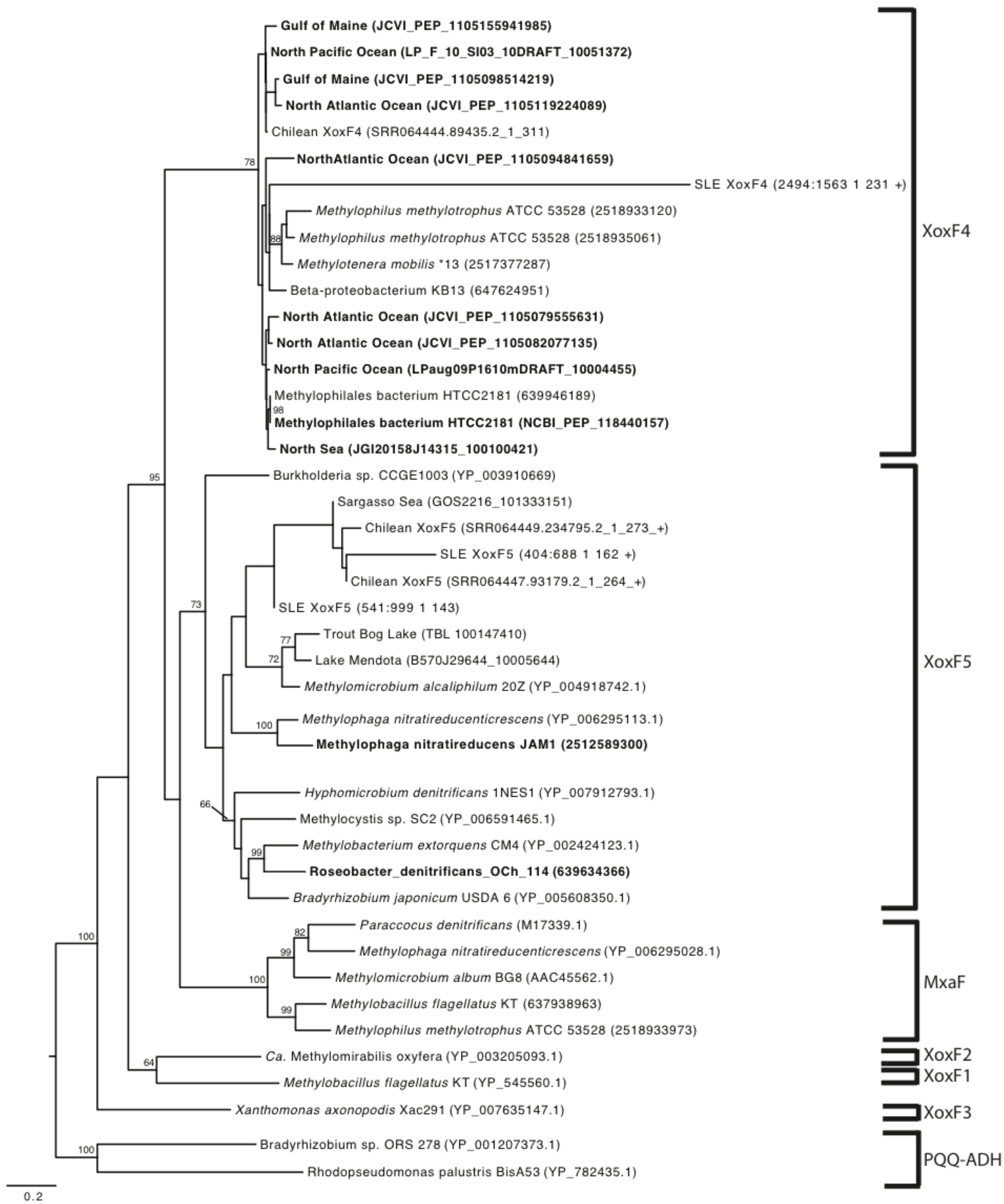
When just the PML samples were analyzed, an increase in the amount of variance explained by axis 1, as well as increased separation between the W-PML and E-PML samples was observed when tr-COG functions were used to build the PCoA rather than COG functions alone. This suggests some degree of functional redundancy, a large portion of which likely has to do with housekeeping functions, as can be seen by the significant percent spectra associated with just four housekeeping proteins (EF-Tu, GroEL, and ATPase alpha and beta subunits) in both sets of metaproteomes (i.e. W-PML and E-PML). The separation between Northern and Southern E-PML samples also increased when tr-COG functions were used instead of COG functions alone, indicating a higher degree of functional redundancy between the Northern and Southern E-PML communities compared to the Northern and Southern W-PML communities. However, the marginal increase in separation between the W-PML and E-PML, as well as the slight increase in percent variation explained by axis one suggests that although the W-PML and E-PML communities do have some degree of functional redundancy, they also contain functional differences. In essence, the W-PML and E-PML communities have higher functional redundancy with each other than do the PML communities and the BTM. The idea that deeper marine samples are functionally distinct from surface samples has already been explored, and clear vertical partitioning of metabolic functions haa already been demonstrated in coastal marine and estuary samples (Colatriano *et al.*, 2015; Georges *et al.*, 2014).

The contrast between stable/similar functional compositions and variable taxonomic compositions of microbial communities observed in several studies, compiled in a recent synthesis (Louca *et al.*, 2018) suggests that a high degree of global functional redundancy exists between microbial populations. Recent work also suggests that because taxonomic composition within functional groups seem to be shaped by different factors than those shaping the functional structure of communities, taxonomic composition and functional composition (genetic potential) appear to be decoupled (Louca *et al.*, 2016, 2018) However, the marginally greater separation between the W-PML and E-PML based on tr-COG functions observed here implies that there are taxon-specific differences in the relative expression of common proteins, as well as taxon-specific functions found primarily on one side, for example, methanol oxidation by Methylophilales in the E-PML. These findings support the ideas put forth by Galand et al. that a shift in community composition can alter the overall functional attributes of communities across temporal and spatial scales and that partial redundancy between populations, where organisms share some specific functions but differ in others or might have other ecological requirements is more likely the case (Galand *et al.*, 2018). This has important implications for community metabolism in a changing ecosystem where the current community structure and thus, community metabolism could shift. It also implies that functional structure and taxonomic composition might not be completely decoupled. The implication that a function in a community is redundant simply because several taxonomic groups in the community possess the potential to perform it does not take into account the fact that distinct taxa can have different reaction kinetics resulting in different biochemical flux rates, and can depend on different environmental conditions and syntrophic interactions. Additionally, because organisms share some functions but differ in others, the dynamics of many metabolic processes can change due to a change in microbial community composition.

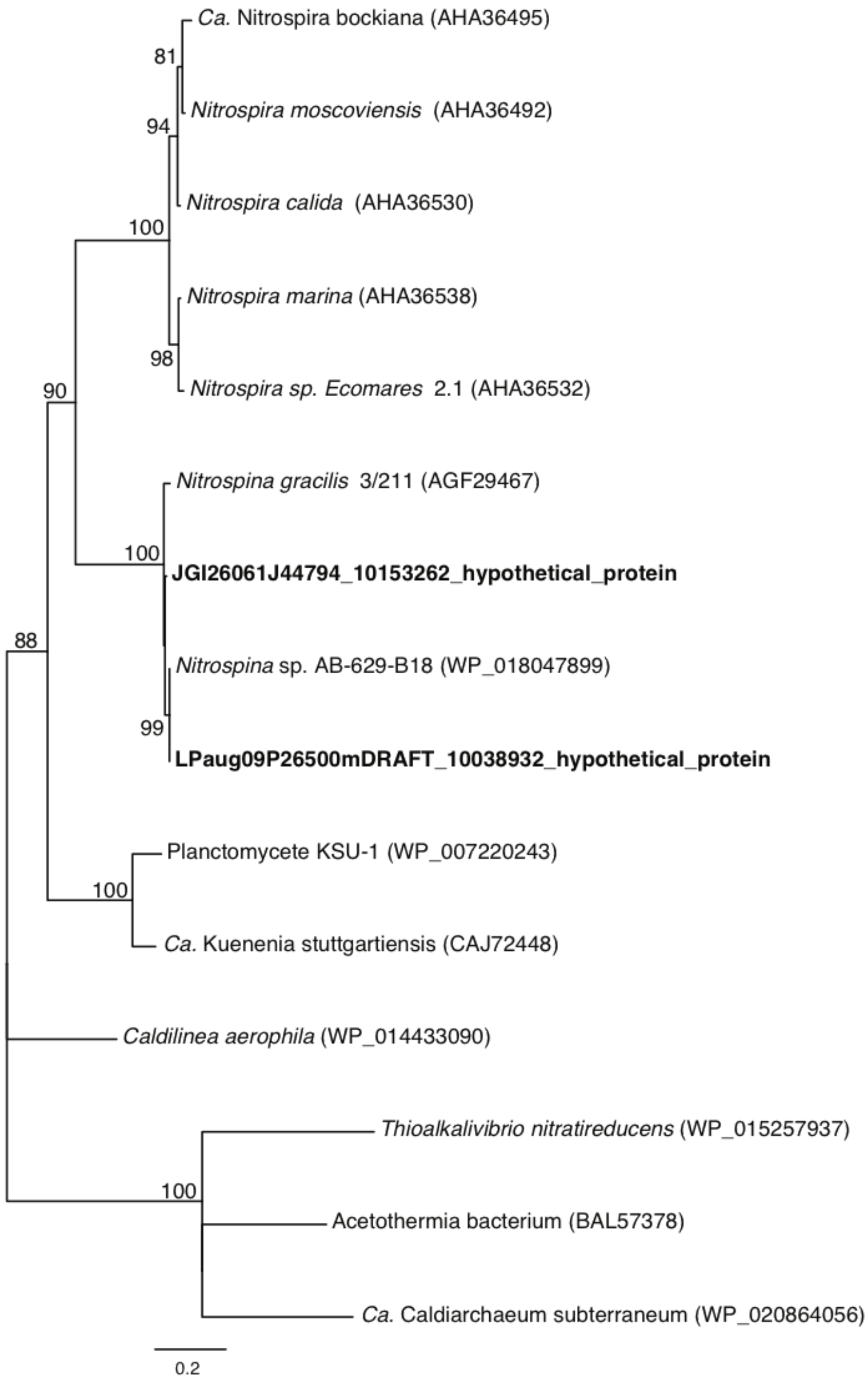
4.6 Conclusions

To our knowledge, this is the first work to compare microbial taxonomy and protein expression in the North Water at several geographic locations and depths. Metaproteomes from distinct water masses in the North Water differed in microbial community structure and metabolic functions of expressed proteins. Based on the differences in observed community composition and tr-COG functions, specifically those associated with membrane transport, we determined that the community in the W-PML was more typical of one associated with dynamic waters that undergo nutrient pulses compared to the E-PML community. Based on the difference in the number of ribosomal proteins identified on each side of the PML, we hypothesize that the W-PML might harbour a microbial community with higher growth rates than the E-PML at the time of sampling. These findings also show that 16S rRNA gene analysis, in conjunction with metaproteomics could be used to indicate past production that cannot be resolved by traditional methods, but is instead captured by the metaproteomic signature. The deep waters were characterized by metaproteomes originating from a more diverse and metabolically versatile array of microbes compared to the surface. Notably, proteins assigned to the uncultivated marine SAR324 and Arctic96BD-19 clades were among the most abundantly detected. Additionally, proteins involved in nitrification assigned to MGI Thaumarchaeota and Nitrospinae were identified in the bottom waters, suggesting a role for nitrifying archaea and bacteria in chemosynthetic production in the deep North Water and providing nitrate to drive the high productivity of this region. The potential importance of methanol as a carbon and energy source for two distinct populations of organisms, both in the surface and deep waters was also identified. This work illustrates how dynamic and different marine microbial communities can be, even when sampled in close geographic proximity and underscores the importance of higher resolution time series and geographic sampling to better understand the dynamics of microbial communities and their relationships to changes in their environment.

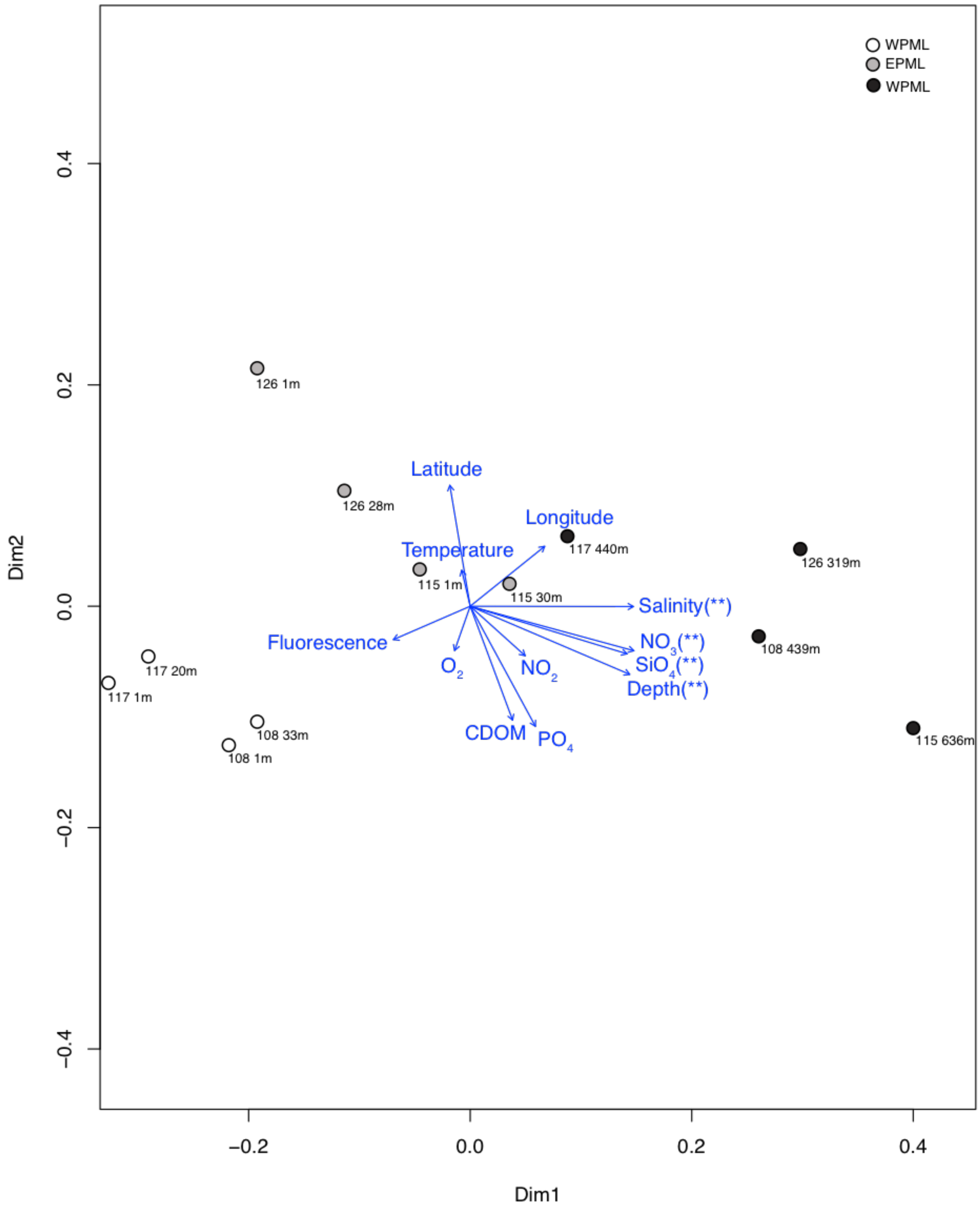
4.7 Supplementary figures



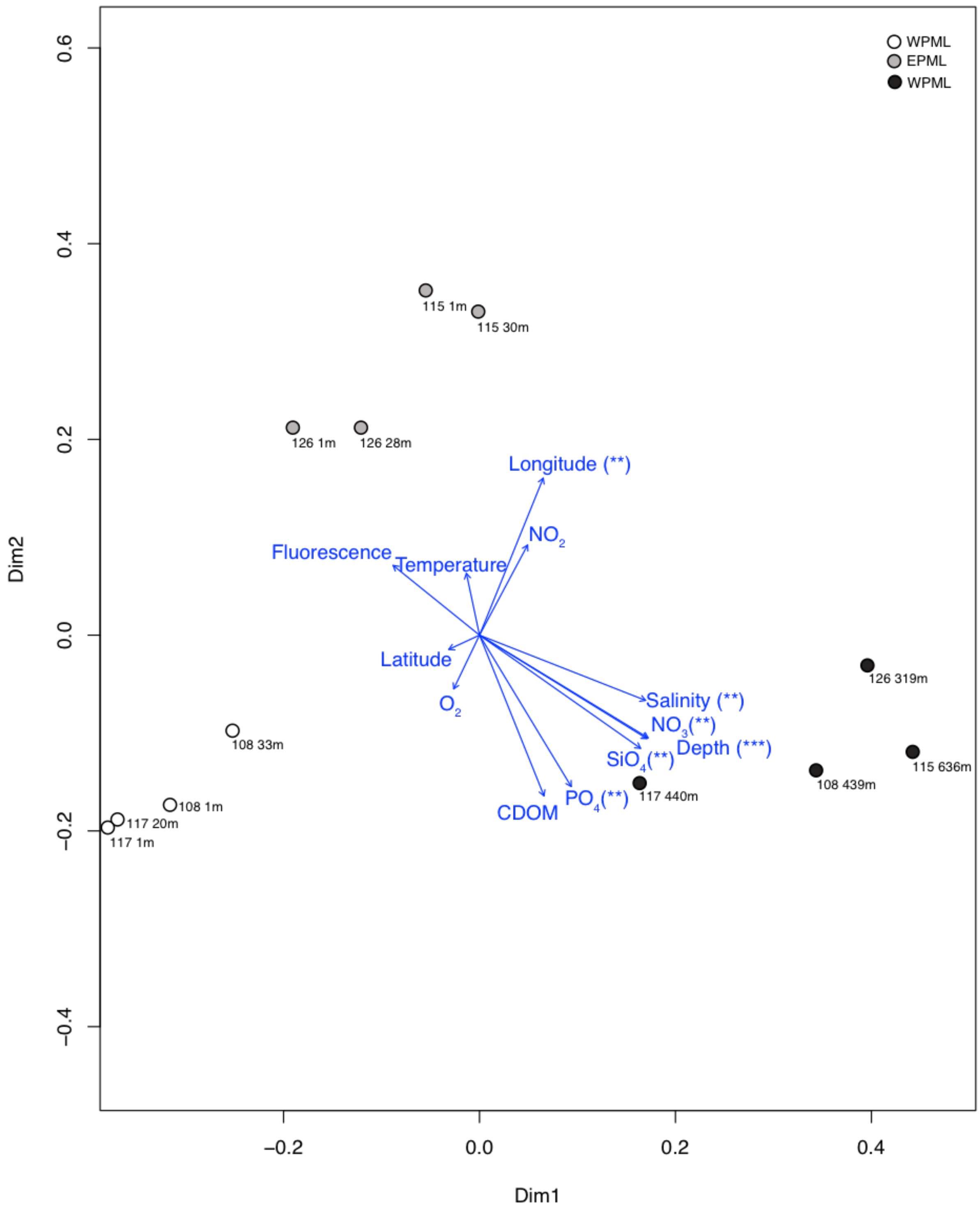
Supplementary figure 4.1. Phylogenetic analysis of identified methanol dehydrogenase homologs matching peptide spectra identified in the North Water (boldface). Phylogeny was inferred using maximum likelihood implemented in MEGA. Bootstrap values are indicated.



Supplementary figure 4.2. Phylogenetic analysis of nitrite oxidoreductase beta (NxrB) homologs matching peptide spectra identified in the North Water (**boldface**). Phylogeny was inferred using maximum likelihood implemented in MEGA. Bootstrap values are indicated.



Supplementary figure 4.3. PCo analysis of the relative abundances of peptide spectra assigned to COG functions including all 12 samples.



Supplementary figure 4.4. PCo analysis of the relative abundances of peptide spectra assigned to tr-COG functions including all 12 samples.

Bridging text

In Chapter 4, 16S rRNA gene analysis and metaproteomics were used to compare microbial community composition and protein expression in the North Water. Three distinct microbial communities (the Western PML, the Eastern PML and the bottom waters) were identified based on differences in observed community composition and taxonomically resolved COG functions. The PML communities differed in their community composition, and strategies for nutrient acquisition and resource allocation. The bottom waters were characterized by a more diverse and metabolically versatile array of microbes including the SAR324 and Arctic96BD-19 clades. The importance of methanol as a carbon and energy source for two distinct populations of organisms, both in the surface and deep waters was also identified. This work was performed using a custom searchable protein database made up of various reference genomes, metagenomes and single cell amplified genomes. At the time of this analysis, no Arctic Ocean metagenomes were publicly available, which, if included in the searchable database might have improved the database search, increasing the number of peptide spectral matches. To remedy the lack of metagenomic representation from the Arctic Ocean we performed the first metagenomic analysis of the Arctic Ocean on 12 samples from 4 research stations in the Beaufort sea, resulting in the construction of 360 metagenome assembled genomes.

Of these 360 metagenome assembled genomes, 6 Chloroflexi MAGs were of particular interest. Marine Chloroflexi are typically comprised of a monophyletic clade called the SAR202 cluster. The SAR202 cluster of Chloroflexi, first recovered as an environmental gene clone from marine waters of the Bermuda Atlantic time series (Giovannoni *et al.*, 1996), represents a ubiquitously distributed clade, having then been isolated from the deep subsurface sediments, soils, marine sponges, freshwater and marine water sources (Giovannoni *et al.*, 1996; Chandler *et al.*, 1998; Urbach *et al.*, 2001; Dunbar *et al.*, 2002; Morris *et al.*, 2004; Varela, Van Aken and Herndl, 2008; Schmitt *et al.*, 2011; Okazaki *et al.*, 2013; Landry *et al.*, 2017; Thrash *et al.*, 2017). SAR202 are an important and abundant component of the free-living microbial community in the mesopelagic (Milici *et al.*, 2016) and bathypelagic oceanic zone worldwide (Salazar *et al.*, 2015; Varela, Van Aken and Herndl, 2008) and represent ~10.2 % of all DNA containing bacterioplankton between 500 m and 4000 m (Morris *et al.*, 2004), and 5 % of 16S rRNA gene sequences from the deep Arctic Ocean waters (Galand *et al.*, 2010). However, recent evidence suggests that SAR202 might represent a higher contribution to marine microbial communities than previously thought (Guerrero-Feijóo *et al.*, 2016). SAR202 is hypothesized to be well adapted to the environmental conditions of the dark ocean and likely contribute a major fraction of the deep-water bacterial activity (Varela, Van Aken and Herndl, 2008; Landry *et al.*, 2017). Although marine Chloroflexi are found to be an important contributor to deep marine microbial communities, relatively little is known about their metabolic properties.

Chapter 5: Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria

Colatriano D, Tran P, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. (2018). Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. *Commun Biol* **1**:1-9

5.1 Abstract

The Arctic Ocean currently receives a large supply of global river discharge and terrestrial dissolved organic matter. Moreover, an increase in freshwater runoff and riverine transport of organic matter to the Arctic Ocean is a predicted consequence of thawing permafrost and increased precipitation. The fate of the terrestrial humic-rich organic material and its impact on the marine carbon cycle are largely unknown. Here, the first metagenomic survey of the Canada Basin in the Western Arctic Ocean showed that pelagic Chloroflexi from the Arctic Ocean are replete with aromatic compound degradation genes, acquired in part by lateral transfer from terrestrial bacteria. Our results imply marine Chloroflexi have the capacity to use terrestrial organic matter and that their role in the carbon cycle may increase with the changing hydrological cycle.

5.2 Introduction

The Arctic Ocean accounts for 1.4% of global ocean volume but receives 11 % of global river discharge (Carmack *et al.*, 2016). Up to 33 % of the dissolved organic matter in the Arctic Ocean is of terrestrial origin and a major fraction of this terrestrial dissolved organic matter (tDOM) originates from carbon-rich soils and peatlands (Benner *et al.*, 2004; Opsahl *et al.*, 1999). With thawing permafrost and increased precipitation occurring across the Arctic (Bintanja and Andry, 2017), increases in freshwater runoff and riverine transport of organic matter to the Arctic Ocean are predicted, which will increase tDOM fluxes and loadings (Frey and McClelland, 2009; Vonk *et al.*, 2012). The additional tDOM may represent new carbon and energy sources for the Arctic Ocean microbial community and contribute to increased respiration, which would result in the Arctic being a source of dissolved inorganic carbon to the ocean. Alternatively, as it moves from its source of origin to the Arctic Ocean tDOM could become more recalcitrant to bacterial metabolism and represent a long term sequestration of the newly released carbon making the Arctic more carbon neutral (Jiao *et al.*, 2010). However, an estimated 50% of Arctic Ocean tDOM is removed before being released to the Atlantic, at least in part by microbial processes (Kaiser *et al.*, 2017). As input of tDOM increases, knowledge on its microbial transformation will be critical for understanding changes in Arctic carbon cycling.

The marine SAR202 is a diverse and uncultivated clade of Chloroflexi bacteria that comprise roughly 10% of planktonic cells in the dark ocean (Giovannoni *et al.*, 1996; DeLong *et al.*, 2006; Morris *et al.*, 2004; Varela, Van Aken and Herndl, 2008; Schattenhofer *et al.*, 2009). SAR202 is also common in marine sediments and deep lakes (Urbach *et al.*, 2001, 2007; Yamada and Sekiguchi, 2009). It has long been speculated that SAR202 may play a role in the degradation of recalcitrant organic matter (DeLong *et al.*, 2006; Varela, Van Aken and Herndl, 2008), and the recent analysis of SAR202 single-cell amplified genomes (SAGs) lends support to this notion (Landry *et al.*, 2017). More generally, Chloroflexi, including those in the SAR202 clade, are also present in the upper layers of the Arctic Ocean (Bano and Hollibaugh, 2002), leading to the hypothesis that recalcitrant organic compounds present in high Arctic tDOM could be utilized by this group.

5.3 Results

In this study, we analyzed Chloroflexi metagenome assembled genomes (MAGs) generated from samples collected from the vertically stratified waters of the Canada Basin in the Western Arctic Ocean (**Figure 5.1a**). A metagenomic co-assembly was generated from samples originating from the surface layer (5 - 7 m), the subsurface chlorophyll maximum (25 - 79 m) and a layer corresponding to the terrestrially-derived DOM fluorescence (FDOM) maximum previously described within the cold CB halocline comprised of Pacific-origin waters (177 - 213 m) (Guéguen *et al.*, 2012). The Pacific-origin FDOM maximum is due to sea ice formation and interactions with bottom sediments on the Beaufort and Chukchi shelves, which themselves are influenced by coastal erosion and river runoff (Guéguen *et al.*, 2012). Binning based on tetranucleotide frequency and coverage resulted in 360 MAGs from a diversity of marine microbes (**Figure 5.1b**).

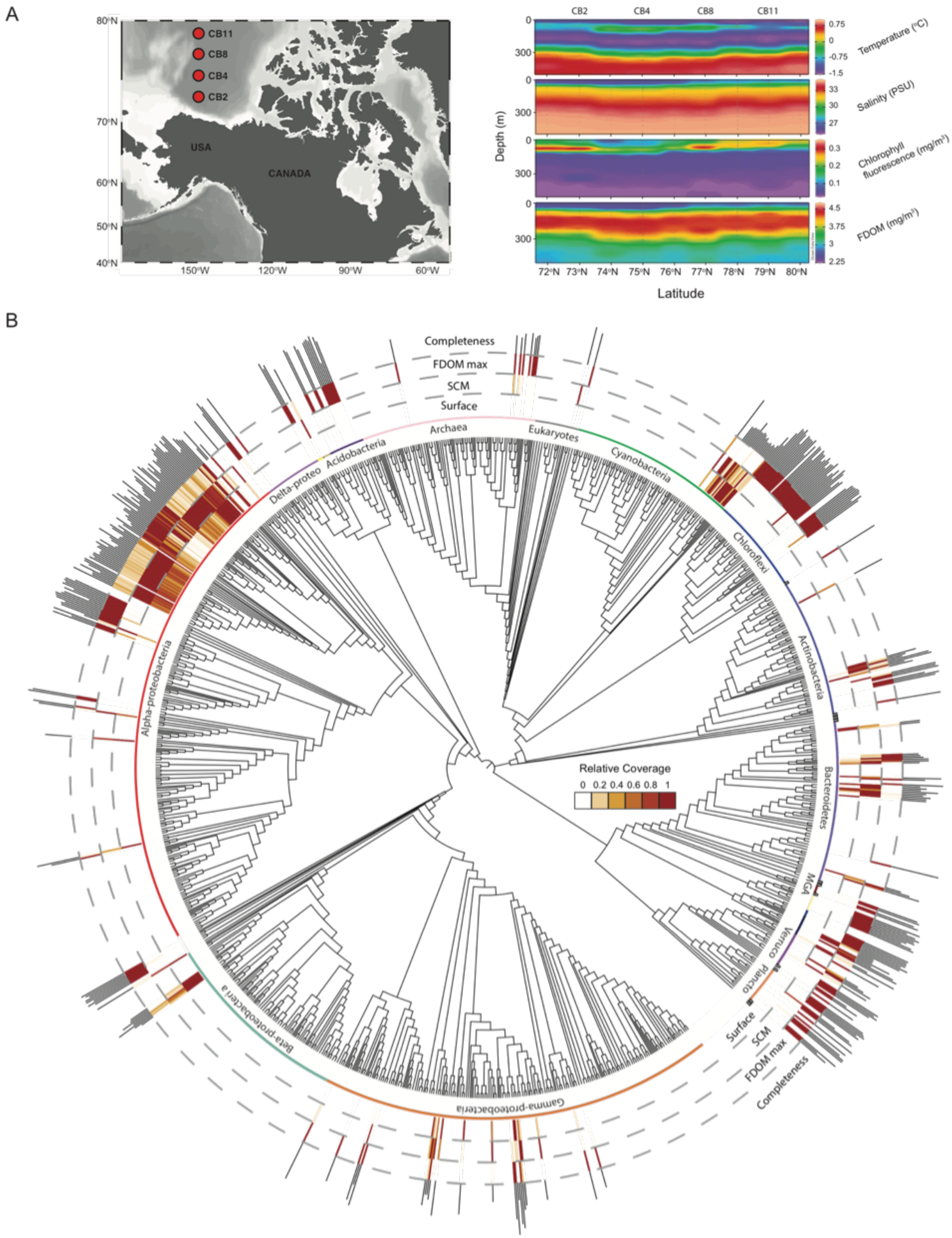


Figure 5.1. Metagenomic survey of microbial diversity in the Canada Basin. **A** Sampling locations and environmental profiles of the Canada Basin. Sample locations and the associated

environmental datasets were plotted using Ocean Data View version 4.7.7 (Schlitzer, 2016). **B** Concatenated protein phylogeny of 360 Arctic Ocean MAGs inferred by MetaWatt and visualized in iTOL. The three inner tracks present relative coverage of MAGs averaged across samples collected from surface waters, subsurface chlorophyll maximum (SCM) and the fluorescent dissolved organic matter maximum (FDOM max). The outer track presents estimated MAG completeness as inferred by MetaWatt. MAG completeness ranged from 25 to 94%.

Six near complete Chloroflexi MAGs were identified. Based on 16S rRNA gene phylogeny, these MAGs represented three distinct SAR202 subclades (SAR202-II, -VI, -VII), the AncK29 clade and the TK10 clade (**Figure 5.2a**). Estimated MAG completeness ranged from 77 to 99%, while contamination ranged from 0 to 2.3% (**Table 5.1**). All MAGs exhibited highest coverage just below the subsurface chlorophyll maximum (**Figure 5.2b**) which is consistent with earlier findings on SAR202 distribution in the North Pacific Ocean (Morris *et al.*, 2004). However, the concentration and composition of the FDOM maximum in the Canada Basin is significantly different compared to the North Pacific Ocean (Dainard and Guéguen, 2013) and the North Atlantic Subtropical Gyre (Dainard *et al.*, 2015). A concatenated protein phylogeny demonstrated that the SAR202 MAGs were distinct from previously published MAGs from the deep ocean (Landry *et al.*, 2017) and oxygen minimum zones (Thrash *et al.*, 2017) (**Supplementary figure 5.1**). Fragment recruitment of 21 TARA Ocean metagenomic datasets spanning epipelagic to mesopelagic waters at 7 locations and 4 separate bathypelagic metagenomes indicated that the Canada Basin Chloroflexi MAGs were not widely distributed in the oceans (**Figure 5.2c, Supplementary table 5.1**). These findings are evidence that the Chloroflexi MAGs represent genotypes that are rare outside Arctic marine waters.

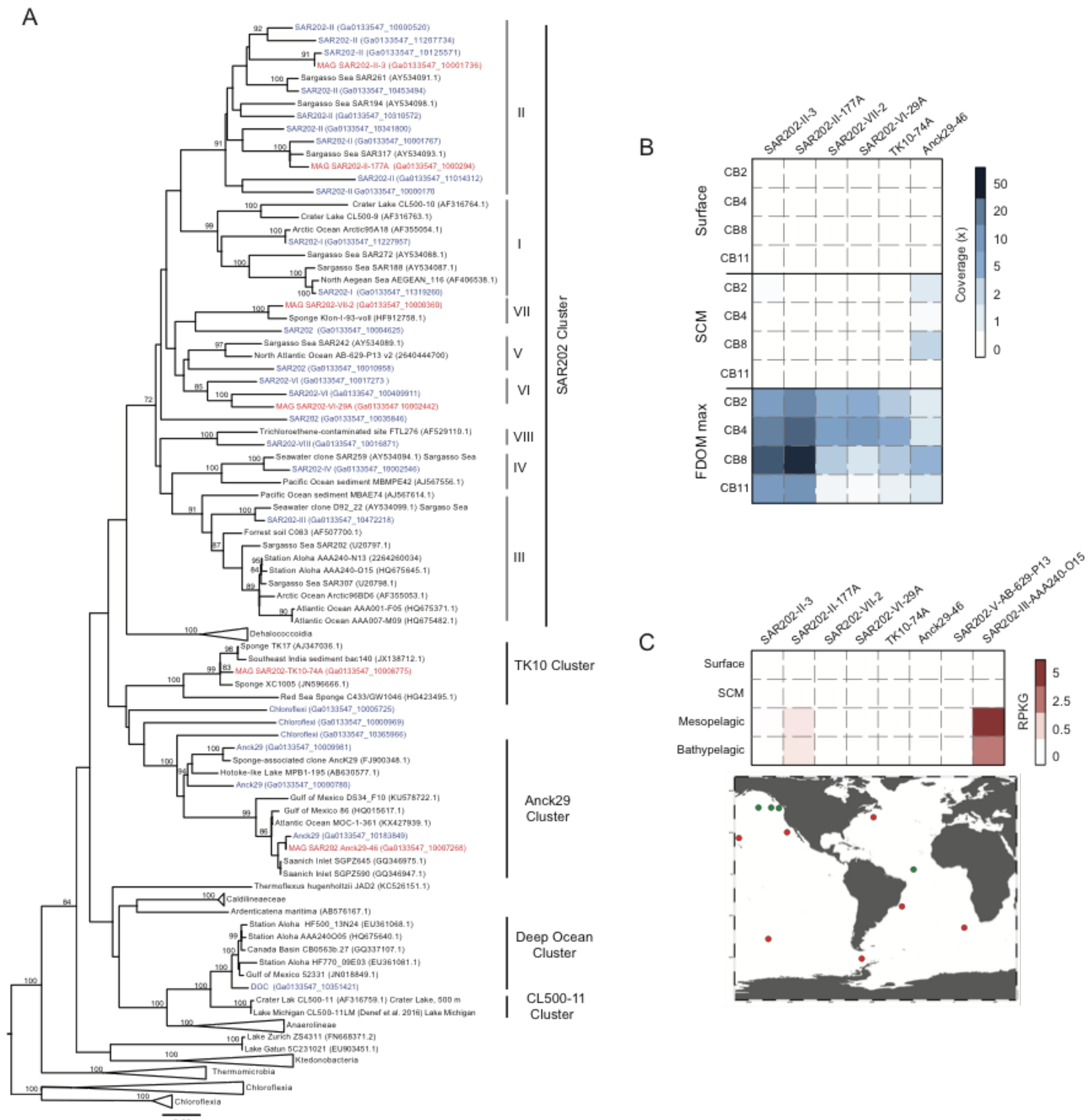


Figure 5.2. Diversity and distribution of Arctic Ocean Chloroflexi MAGs. **A** Maximum likelihood tree of Chloroflexi based on partial 16S rRNA gene sequences. Blue taxa are from Canada Basin metagenomes and red taxa are from the six Chloroflexi MAGs. Bootstrap values of >70% are shown (100 replicates). **B** Distribution of MAGs in the Canada Basin based on metagenome coverage at the surface, subsurface chlorophyll maximum (SCM) and fluorescent dissolved organic matter maximum (FDOM max). c Distribution of MAGs in global ocean metagenomes based on fragment recruitment. Two deep ocean SAR202 SAGs from Landry et al. (Landry *et al.*, 2017) were included for comparison. Location of TARA ocean metagenomes

(red) and bathypelagic metagenomes (green) are shown on the map generated using Ocean Data View version 4.7.7 (Schlitzer, 2016).

Table 5.1. Genomic characteristics of MAGs

	Size (Mb)	Cov (x)	GC (%)	Completeness (%)	Contamination (%)	N50 (kb)	# of Contigs
SAR202-II-3	1.36	68	39	80	0	35	46
SAR202-II-177A	1.62	101	42	82	1	36	52
SAR202-VII-2	2.78	16	59	99	0	246	8
SAR202-VI-29A	1.52	16	46	97	2	59	2
TK10-74A	2.45	12	69	81	2.3	38	76
Anck29-46	1.11	11	32	77	0	75	22

The Chloroflexi MAGs contained many genes implicated in the degradation of aromatic compounds typically associated with humic-rich tDOM (**Supplementary table 5.2**). A single MAG (SAR202-VII-2) from a previously undescribed clade (SAR202-VII) exhibited a striking enrichment in these genes (**Figure 5.3a**). Partial pathways for the catabolism of aromatic compounds were recently reported from deep ocean SAR202 SAGs (Landry *et al.*, 2017). To assess whether the abundance and diversity of SAR202-VII-2 genes involved in aromatic compound catabolism is unique to Arctic Ocean MAGs or is a more broad characteristic of marine Chloroflexi, we compared gene content between SAR202-VII-2 and two SAGs (SAR202-V-AB-629-P13 and SAR202-III-AAA240-O15) reported in Landry et al. (Landry *et al.*, 2017). Of the 117 SAR202-VII-2 orthologs implicated in aromatic compound degradation, 12 were identified in SAR202-III-AAA240-O15 and only one was identified in SAR202-V-AB-629-P13, implying distinct and less diverse pathways in deep ocean SAR202 compared to the Arctic Ocean populations (**Supplementary table 5.2**).

Proteins for the modification and degradation of monoaryl and biaryl compounds were predicted, including a diversity of aromatic ring-cleaving dioxygenases (Barry and Taylor, 2014; Kasai *et al.*, 2005; Werwath *et al.*, 1998). A total of 42 ring-cleaving dioxygenases targeting compounds related to catechol, protocatechuate and gentisate were present in the six MAGs, with 25 dioxygenases predicted in SAR202-VII-2 alone (**Figure 5.3a-b**). Ring demethylation, hydroxylation and decarboxylation are important prerequisite steps to prime diverse aromatic compounds for downstream oxidative cleavage (Fetzner, 2012; Fuchs *et al.*, 2011). Thirty ring-demethylating monooxygenases, ten ring-hydroxylating dioxygenases, and eleven ring-decarboxylases were annotated in the SAR202-VII-2 MAG (**Figure 5.3c, Supplementary table 5.2**). Proteins involved in the conversion of ring-cleavage products to central intermediates of the citric acid cycle were also present in the SAR202-VII-2 MAG, including dehydrogenases (*i.e.* 2,3-dihydroxy-2,3-dihydrophenylpropionate dehydrogenase), decarboxylases (*i.e.* oxaloacetate B-decarboxylase), aldolases (*i.e.* HMG aldolase and 4-carboxymuconolactone decarboxylase), hydratases (*i.e.* 4-oxalmescanoate hydratase and 2-oxopent-4-enoate hydratase), isomerases (*i.e.* mycothiol maleulpyruvate isomerase and muconolactone isomerase) and hydrolases (*i.e.* 3-oxoadipate enol-lactonase and β-ketoadipate enol-lactone hydrolase) (**Supplementary table 5.2, Supplementary figure 5.2**). We note that we were unable to identify a single complete reference pathway for humic-like aromatic compound degradation. Since estimated genome completeness for SAR202-VII-2 was 99%, it is unlikely the genes were missed due to an incomplete genome. Another explanation is that marine Chloroflexi genomes encode novel pathway variants. Indeed, numerous metal-dependent hydrolases, hydrolases of the HAD family and NAD(P)-dependent dehydrogenase were clustered in genomic regions with the ring-modifying oxygenases, decarboxylases, and demethylases described above. In addition to the array of aromatic compound degradation genes, the SAR202-VII-2 MAG also contained 34 copies of the flavin mononucleotide (FM)/F420-dependent monooxygenase catalytic subunit (FMNO) proteins previously implicated in activation of recalcitrant organic compounds in the deep ocean (Landry *et al.*, 2017). These results are consistent with Chloroflexi in the Arctic Ocean having the metabolic potential to access carbon and energy available in aromatic compounds typically associated with tDOM.

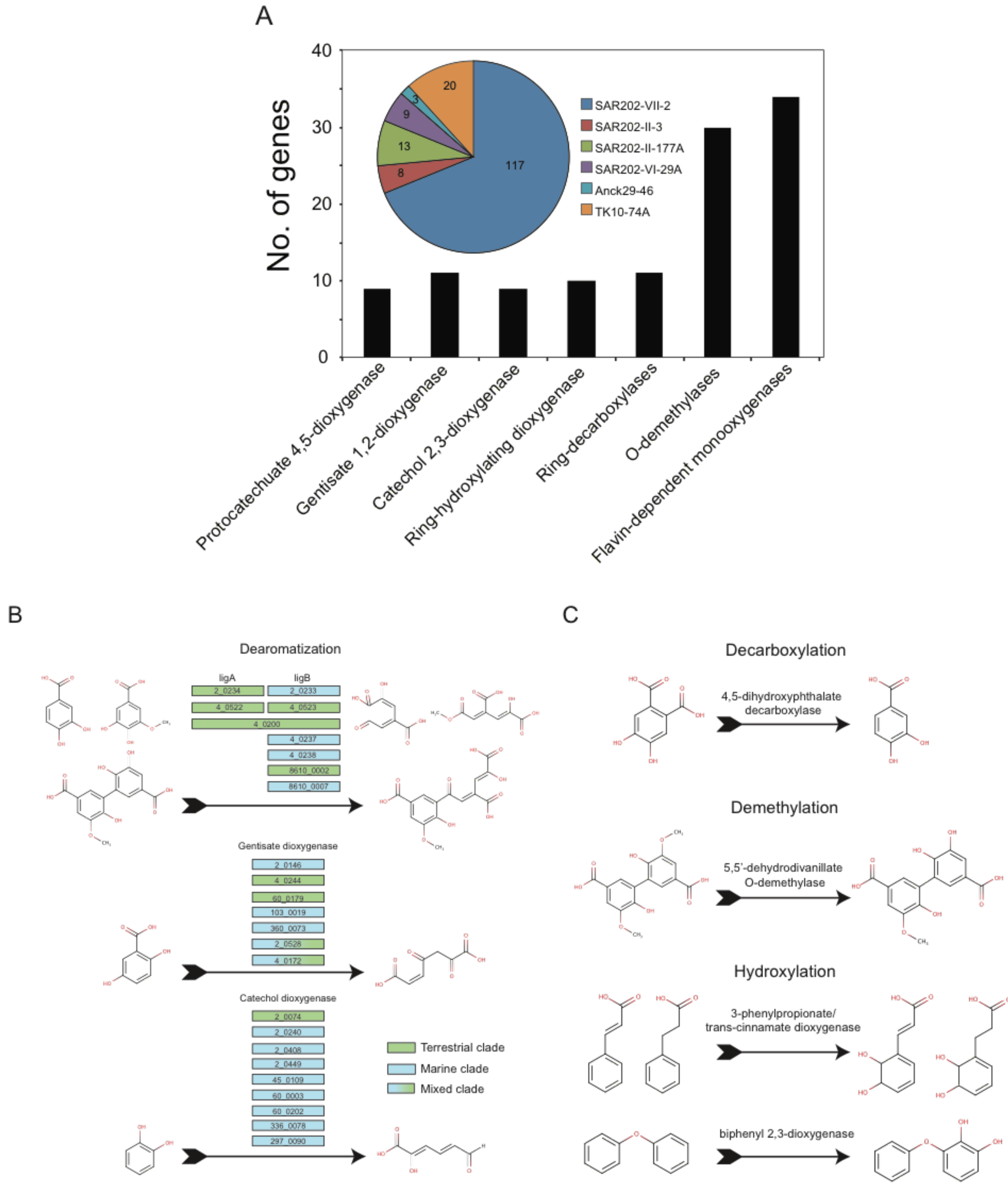


Figure 5.3. Aromatic compound degradation genes and pathways in Arctic Ocean Chloroflexi MAGs. **A** Abundance of aromatic compound degradation genes in Chloroflexi MAGs from the Canada Basin (pie chart) and a breakdown of those found specifically in SAR202-VII-2 (column plot). **B** Predicted aromatic ring-opening enzymatic reactions identified in SAR202-VII-2 with

gene loci displayed in the colored boxes. Genes in green and blue boxes were most closely related to homologs from terrestrial or marine bacteria, respectively. Genes in the blue/green boxes were in clades containing diverse environmental bacteria. **c** Examples of predicted aromatic ring-modifying enzymatic reactions identified in SAR202-VII-2.

The diversity of SAR202-VII-2 genes implicated in aromatic compound degradation lead us to hypothesize that they may have originated by lateral gene transfer from terrestrial bacteria. To test this, we targeted aromatic compound degradation genes (the ring-cleaving dioxygenases, specifically) in the Chloroflexi MAGs for in-depth phylogenetic analyses. The genomic diversity of marine Chloroflexi was expanded in our analysis by including 130 Chloroflexi MAGs recently assembled and binned from the TARA Oceans project (Tully *et al.*, 2017). A number of the SAR202-VII-2 ring-cleaving dioxygenase homologs were most closely related to proteins from the TARA Ocean Chloroflexi MAGs and other marine originating genomes, particularly the catechol dioxygenases (**Supplementary figure 5.3**), 3 gentisate 1,2 dioxygenases (**Figure 5.4**) and methylgallate dioxygenases (**Supplementary figure 5.4**) indicating that aromatic compound degradation in Chloroflexi is not restricted to the Arctic Ocean. However, lateral gene transfer from terrestrial bacteria was also evident. For example, an annotated gentisate dioxygenase gene was positioned within a clade of terrestrial Actinomycetes (**Figure 5.4**). Additional genes involved in the degradation of structures related to catechol, protocatechuate and gentisate were phylogenetically associated with homologs from terrestrial Acidobacteria (**Supplementary figure 5.5**), Actinobacteria (**Supplementary figure 5.6**), Armatimonadetes (**Figure 5.4**), Delta-proteobacteria (**Supplementary figures 5.3 and 5.6**), Beta-proteobacteria (Supplementary figure 5.7) and a clade of diverse terrestrial phyla (**Supplementary figure 5.8**). Additionally, 2 gentisate 1,2-dioxygenase genes and 1 protocatechuate dioxygenase *ligB* gene were phylogenetically associated to a clade of genes from both terrestrial Delta-proteobacteria and marine microbes (**Fig. 5.4 and Supplementary figure 5.8**). These putative gene acquisitions were unlikely due to contaminating scaffolds because the genes were located on long scaffolds that were assigned to Chloroflexi with high confidence based on tetranucleotide frequencies and the phylogenetic identity of house-keeping genes. Such a phylogenetic pattern supports the hypothesis that marine Chloroflexi acquired the capacity for aromatic compound degradation, at least in part, by lateral gene transfer from terrestrial bacteria.

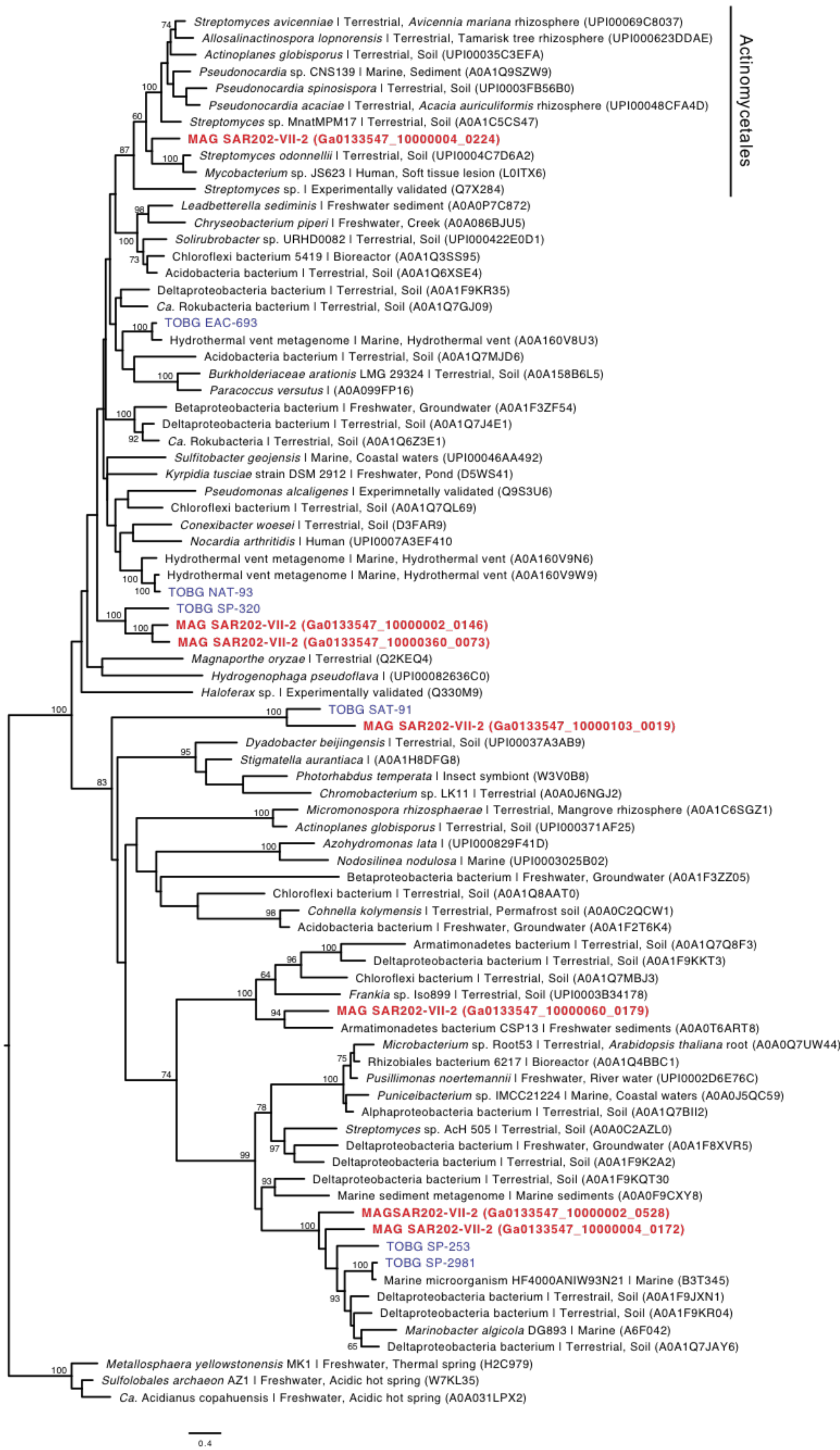

Actinomycetales
Streptomyces avicenniae | Terrestrial, Avicennia mariana rhizosphere (UPI00069C8037)
Allosalinactinospora lopnorensis | Terrestrial, Tamarisk tree rhizosphere (UPI000623DDAE)
Actinoplanes globisporus | Terrestrial, Soil (UPI00035C3EFA)
Pseudonocardia sp. CNS139 | Marine, Sediment (A0A1Q9SZW9)
Pseudonocardia spinosispora | Terrestrial, Soil (UPI0003FB56B0)
Pseudonocardia acaciae | Terrestrial, Acacia auriculiformis rhizosphere (UPI00048CFA4D)
Streptomyces sp. MnatMPM17 | Terrestrial, Soil (A0A1C5CS47)
MAG SAR202-VII-2 (Ga0133547_10000004_0224)
Streptomyces odonnellii | Terrestrial, Soil (UPI0004C7D6A2)
Mycobacterium sp. JS623 | Human, Soft tissue lesion (L0ITX6)
Streptomyces sp. | Experimentally validated (Q7X284)
Leadbetterella sediminis | Freshwater sediment (A0A0P7C872)
Chryseobacterium piperi | Freshwater, Creek (A0A086BJU5)
Solirubrobacter sp. URHD0082 | Terrestrial, Soil (UPI000422E0D1)
Chloroflexi bacterium 5419 | Bioreactor (A0A1Q3SS95)
Acidobacteria bacterium | Terrestrial, Soil (A0A1Q6XSE4)
Deltaproteobacteria bacterium | Terrestrial, Soil (A0A1F9KR35)
Ca. Rokubacteria bacterium | Terrestrial, Soil (A0A1Q7GJ09)
TOBG EAC-693
Hydrothermal vent metagenome | Marine, Hydrothermal vent (A0A160V8U3)
Acidobacteria bacterium | Terrestrial, Soil (A0A1Q7MJD6)
Burkholderiaceae arationis LMG 29324 | Terrestrial, Soil (A0A158B6L5)
Paracoccus versutus | (A0A099FP16)
Betaproteobacteria bacterium | Freshwater, Groundwater (A0A1F3ZF54)
Deltaproteobacteria bacterium | Terrestrial, Soil (A0A1Q7J4E1)
Ca. Rokubacteria | Terrestrial, Soil (A0A1Q6Z3E1)
Sulfitobacter geojensis | Marine, Coastal waters (UPI00046AA492)
Kyrpidia tusciae strain DSM 2912 | Freshwater, Pond (D5WS41)
Pseudomonas alcaligenes | Experimnetally validated (Q9S3U6)
Chloroflexi bacterium | Terrestrial, Soil (A0A1Q7QL69)
Conexibacter woesei | Terrestrial, Soil (D3FAR9)
Nocardia arthritidis | Human (UPI0007A3EF410)
Hydrothermal vent metagenome | Marine, Hydrothermal vent (A0A160V9N6)
Hydrothermal vent metagenome | Marine, Hydrothermal vent (A0A160V9W9)
TOBG NAT-93
TOBG SP-320
MAG SAR202-VII-2 (Ga0133547_10000002_0146)
MAG SAR202-VII-2 (Ga0133547_10000360_0073)
Magnaporthe oryzae | Terrestrial (Q2KEQ4)
Hydrogenophaga pseudoflava | (UPI00082636C0)
Haloferax sp. | Experimentally validated (Q330M9)
TOBG SAT-91
MAG SAR202-VII-2 (Ga0133547_10000103_0019)
Dyadobacter beijingensis | Terrestrial, Soil (UPI00037A3AB9)
Stigmatella aurantiaca | (A0A1H8DFG8)
Photorhabdus temperata | Insect symbiont (W3V0B8)
Chromobacterium sp. LK11 | Terrestrial (A0A0J6NGJ2)
Micromonospora rhizosphaerae | Terrestrial, Mangrove rhizosphere (A0A1C6SGZ1)
Actinoplanes globisporus | Terrestrial, Soil (UPI000371AF25)
Azohydromonas lata | (UPI000829F41D)
Nodosilinea nodulosa | Marine (UPI0003025B02)
Betaproteobacteria bacterium | Freshwater, Groundwater (A0A1F3ZZ05)
Chloroflexi bacterium | Terrestrial, Soil (A0A1Q8AAT0)
Cohnella kolymensis | Terrestrial, Permafrost soil (A0A0C2QCW1)
Acidobacteria bacterium | Freshwater, Groundwater (A0A1F2T6K4)
Armatimonadetes bacterium | Terrestrial, Soil (A0A1Q7Q8F3)
Deltaproteobacteria bacterium | Terrestrial, Soil (A0A1F9KKT3)
Chloroflexi bacterium | Terrestrial, Soil (A0A1Q7MBJ3)
Frankia sp. Iso899 | Terrestrial, Soil (UPI0003B34178)
MAG SAR202-VII-2 (Ga0133547_10000060_0179)
Armatimonadetes bacterium CSP13 | Freshwater sediments (A0A0T6ART8)
Microbacterium sp. Root53 | Terrestrial, Arabidopsis thaliana root (A0A0Q7UW44)
Rhizobiales bacterium 6217 | Bioreactor (A0A1Q4BBC1)
Pusillimonas noertemannii | Freshwater, River water (UPI0002D6E76C)
Puniceibacterium sp. IMCC21224 | Marine, Coastal waters (A0A0J5QC59)
Alphaproteobacteria bacterium | Terrestrial, Soil (A0A1Q7BII2)
Streptomyces sp. AcH 505 | Terrestrial, Soil (A0A0C2AZL0)
Deltaproteobacteria bacterium | Freshwater, Groundwater (A0A1F8XVR5)
Deltaproteobacteria bacterium | Terrestrial, Soil (A0A1F9K2A2)
Deltaproteobacteria bacterium | Terrestrial, Soil (A0A1F9KQT30
Marine sediment metagenome | Marine sediments (A0A0F9CXY8)
MAGSAR202-VII-2 (Ga0133547_10000002_0528)
MAG SAR202-VII-2 (Ga0133547_10000004_0172)
TOBG SP-253
TOBG SP-2981
Marine microorganism HF4000ANIW93N21 | Marine (B3T345)
Deltaproteobacteria bacterium | Terrestrail, Soil (A0A1F9JXN1)
Deltaproteobacteria bacterium | Terrestrial, Soil (A0A1F9KR04)
Marinobacter algicola DG893 | Marine (A6F042)
Deltaproteobacteria bacterium | Terrestrial, Soil (A0A1Q7JAY6)
Metallosphaera yellowstonensis MK1 | Freshwater, Thermal spring (H2C979)
Sulfolobales archaeon AZ1 | Freshwater, Acidic hot spring (W7KL35)
Ca. Acidianus copahuensis | Freshwater, Acidic hot spring (A0A031LPX2)
0.4

Figure 5.4. Maximum likelihood tree of predicted gentisate1,2-dioxygenases. Bootstrap values of >60% are shown (100 replicates). Homologs from SAR202-VII-2 are highlighted in red and homologs from TARA Ocean MAGs are highlighted in blue.

5.4 Conclusion

In total, these results are consistent with Chloroflexi playing a role in tDOM transformation in waters of the Arctic Ocean. This is the first study to our knowledge to associate a specific microbial group with tDOM metabolism in the Arctic Ocean and it expands on recent studies contributing to our understanding of the metabolic diversity of the abundant yet uncultivated marine Chloroflexi (Landry *et al.*, 2017; Thrash *et al.*, 2017). Moreover, lateral gene transfer from terrestrial bacteria appears to have contributed to the evolution of aromatic compound degradation capabilities within marine Chloroflexi, particularly in regions of the Arctic Ocean impacted by terrestrial input.

The majority of MAGs were restricted to the humic-rich Pacific-origin halocline of the CB, however it is the surface waters that will be most immediately affected by increased freshwater input (Carmack *et al.*, 2016). Hence, our initial observations suggest a need for further research on the distribution of tDOM-utilizing microbes in other Arctic water masses with an aim to establish how common and phylogenetically widespread tDOM metabolism is in the Arctic Ocean. These water masses could include coastal surface waters at the mouth of the Mackenzie River, as well as regions of differing DOM composition such as the East Siberian Sea (Guéguen *et al.*, 2012). Moreover, metagenomic studies such as this are, in essence, hypothesis-generating and future work that includes targeted cultivation, *in situ* gene expression analysis, and rate measurement-based approaches are required to validate and quantify microbial metabolic contributions to nutrient cycling. Overall it is likely that marine Chloroflexi have the capacity to degrade tDOM, and their role in the Arctic carbon cycle may increase as Arctic warming leads to greater inputs of terrestrial organic matter.

5.5 Methods

5.5.1 Sampling and DNA extraction

Twelve samples for metagenomics were collected in September 2015 during the Joint Ocean Ice Study cruise to the Canada Basin . For each sample, 4-8 L of seawater was sequentially filtered through a 50 µm pore mesh, followed by a 3 µm pore size polycarbonate filter and a 0.22 µm pore size Sterivex filter (Durapore; Millipore, Billerica, MA, USA). Filters were preserved in RNAlater and stored at -80 °C until processed in the laboratory. DNA was extracted from the Sterivex filter using the following method: filters were thawed on ice and RNAlater was removed. The Sterivex was then rinsed twice with a sucrose-based lysis buffer, and filled with 1.8 mL of the lysis buffer. Filters were treated with 100 µL of 125 mg mL^{-1} lysozyme and 20 µL of 10 µg mL^{-1} RNAse A and left to rotate at 37 °C for 1 hour. After incubation, 100 µL of 10 mg mL^{-1} proteinase K and 100 µL of 20% SDS was added. Filters were left to rotate for 2 hours at 55 °C. Lysate was removed from the filters. Protein was precipitated and removed with 0.583 volumes of MPC Protein Precipitation Reagent (Epicentre, Madison, WI, USA) and centrifugation at 10,000 g at 4 °C for 10 minutes. The supernatant was transferred to a clean tube. DNA was precipitated with cold isopropanol, and resuspended in low TE buffer.

5.5.2 Metagenomic sequencing, assembly, annotation, and binning

DNA sequencing of 12 samples was performed at the Department of Energy Joint Genome Institute (Walnut Creek, CA, USA) on the HiSeq 2500-1TB (Illumina) platform. Paired-end sequences of 2 × 150bp were generated for all libraries. A metagenome coassembly of all raw reads was generated using MEGAHIT (Li *et al.*, 2015) with kmer sizes of 23,43,63,83,103,123. Gene prediction and annotation was performed using the DOE Joint Genome Institute's Integrated Microbial Genomes (IMG) database tool (version 4.11.0) (Huntemann *et al.*, 2016). Metagenomic binning was performed on scaffolds greater than or equal to 10 Kb in length using MetaWatt (Strous *et al.*, 2012). Relative weight of coverage binning was set to 0.75 and the optimize bins and polish bins options were set to on. The taxonomic identity of MAGs was assessed using a concatenated phylogenetic tree based on 138 single copy conserved genes as implemented in MetaWatt (Strous *et al.*, 2012). Visualization of the tree and mapping of data on to taxa was performed with iTOL (Letunic and Bork, 2016). Estimation of MAG completeness and contamination was performed using CheckM (Parks *et al.*, 2015). Six Chloroflexi MAGs were selected for further analysis based on the presence of a 16S rRNA gene, high completeness and low contamination. Manual curation of the six Chloroflexi MAGs was performed and suspected contaminating scaffolds (single copy genes most similar to non-Chloroflexi taxa) were removed prior to further analysis of MAGs.

5.5.3 16S rRNA phylogenetic analysis

Chloroflexi diversity in the metagenomic assembly was assessed by 16S rRNA gene analysis. All 16S rRNA genes in the co-assembly were assigned to taxonomic groups using mothur (Schloss *et al.*, 2009) and the Wang method with a bootstrap value cutoff of 60 % (Wang *et al.*, 2007). Chloroflexi 16S rRNA genes greater than 360 bp were included in a phylogenetic analysis with Chloroflexi reference sequences. A multiple sequence alignment was generated using MUSCLE (implemented in MEGA6) (Edgar, 2004). Phylogenetic reconstructions were conducted by maximum likelihood using MEGA6-v.0.6 and the following settings: general time reversible model, gamma distribution model for the rate variation with four discrete gamma categories, and the nearest-neighbour interchange (NNI) heuristic search method (Tamura *et al.*, 2013) with a bootstrap analysis using 100 replicates.

5.5.4 Single protein and concatenated protein phylogenies

A concatenated protein phylogeny was constructed using 30 Chloroflexi reference genomes and the 6 Canada Basin Chloroflexi MAGs. Orthologous genes in the 36 Chloroflexi genomes were identified using ProteinOrtho (Lechner *et al.*, 2011). Fifty orthologs present in at least 34 of the 36 genomes were selected for concatenated phylogenetic analysis (**Supplemetary table 4.3**). Each orthologous protein family was aligned using MUSCLE (implemented in MEGA6) and alignment positions were masked using the probabilistic masker ZORRO (Wu *et al.*, 2012), masking columns with weights less than 0.5. The concatenated alignment consisted of 14,815 amino acid positions. Phylogenetic reconstructions were conducted by maximum likelihood using MEGA6-v.0.6 and the following settings: JTT substitution model, gamma distribution with invariant sites model for the rate variation with four discrete gamma categories, and the nearest-neighbour interchange (NNI) heuristic search method (Tamura *et al.*, 2013) with a bootstrap analysis using 100 replicates.

For phylogenetic analysis of ring-cleaving dioxygenase sequences identified in SAR202-VII-2, query sequences were searched against UniRef90 and 130 Chloroflexi MAGs constructed from the TARA Oceans dataset (Tully *et al.*, 2017). The TARA Ocean Chloroflexi MAGs were used as is with no manual curation. UniRef90 sequences and the top TARA Ocean MAG hits for each dioxygenase were aligned with their respective SAR202-VII-2 homologs with MUSCLE (implemented in MEGA6) and alignment positions were masked using the probabilistic masker ZORRO (Wu *et al.*, 2012), masking columns with weights less than 0.5. Phylogenetic reconstruction was conducted using the same settings as the concatenated phylogeny.

5.5.5 Comparative genomics and metabolic reconstruction

The distribution of orthologs across Arctic Ocean genomes, as well as the identification of orthologs shared with the deep ocean SAGs, was determined using proteinortho (Lechner *et al.*, 2011). Inference of protein function and metabolic reconstruction was based on the IMG annotations provided by the JGI, including KEGG, Pfam, EC numbers, and Metacyc annotations. Metabolic reconstruction was also facilitated by generated pathway genome databases for each MAG using the pathologic software available through Pathway Tools (Karp *et al.*, 2011).

5.5.6 Metagenomic fragment recruitment

The distribution of the Canada Basin MAGs in the global ocean was determined using best-hit reciprocal blast analysis similar to Landry et. al 2017 (Landry *et al.*, 2017). Unassembled metagenomic data from 25 samples (**Supplementary figure 5.1**) was first recruited to the six Canada Basin Chloroflexi MAGS as well as two SAR202 SAGs originating from the deep North Pacific Ocean from the Hawaiian ocean time series (HOTS) and the deep North Atlantic Ocean (Landry *et al.*, 2017). Metagenomes from the TARA Ocean project used here were representative of the surface, chlorophyll maximum and mesopelagic waters from the North Atlantic, South

Atlantic, North Pacific, Coastal North Pacific, South Pacific, Coast of Brazil and the Antarctic peninsula. To reduce computational demand, only part 1 (1 Gbp of a random subset of reads) of each metagenomic dataset available at EBI was used (Supplementary table 5.1). Additional bathypelagic metagenomes from the North Pacific and South Atlantic Oceans (LineP P04, P12 and P26, and Knorr S15 2500 m) were also included. All hits from the initial blast were then reciprocally queried against the Canada Basin Chloroflexi MAGs, bathypelagic SAR202 SAGs, and 130 Chloroflexi MAGs constructed from the TARA oceans data. The best hit was reported. Only hits with an alignment length greater than or equal to 100 bp and a percent identity of 95% or more were counted (lower % identity cut-offs did not alter the number of reads recruited in any significant manner). To compare the results among the different data sets, the number of recruited reads was normalized to total number of reads in each sample. The final coverage results were expressed as the number of reads per kilobase of the MAG per gigabase of metagenome (rpkg).

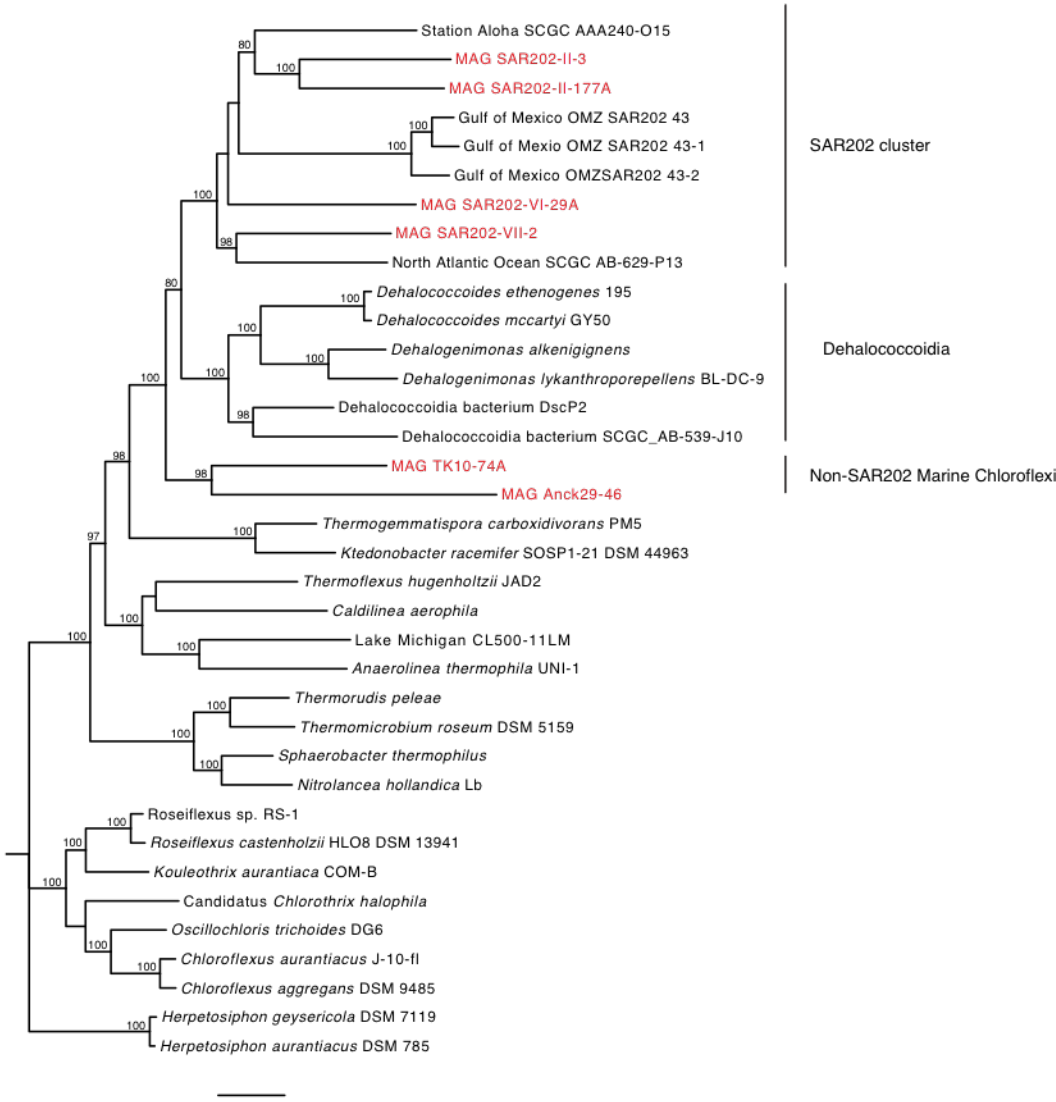
5.6 Data Availability

The metagenomic data generated in this study are available in the Integrated Microbial Genomes database at the Joint Genome Institute at https://img.jgi.doe.gov, GOLD Project ID: Ga0133547. Metagenome assembled genome projects have been deposited at DDBJ/ENA/GenBank under the Bioproject PRJNA471535 and accession numbers QGNM00000000 (for SAR202-II-3), QGNN00000000 (for SAR202-II-177A), QGNO00000000 (for SAR202-VI-29A), QEVV00000000 (for SAR202-VII-2), QGNP00000000 (for Anck29-46) and QGNQ00000000 (for TK10-74A). The versions described in this paper are versions QGNM01000000 (for SAR202-II-3), QGNN01000000 (for SAR202-II-177A), QGNO01000000 (for SAR202-VI-29A), QEVV01000000 (for SAR202-VII-2), QGNP01000000 (for Anck29-46) and QGNQ01000000 (for TK10-74A).

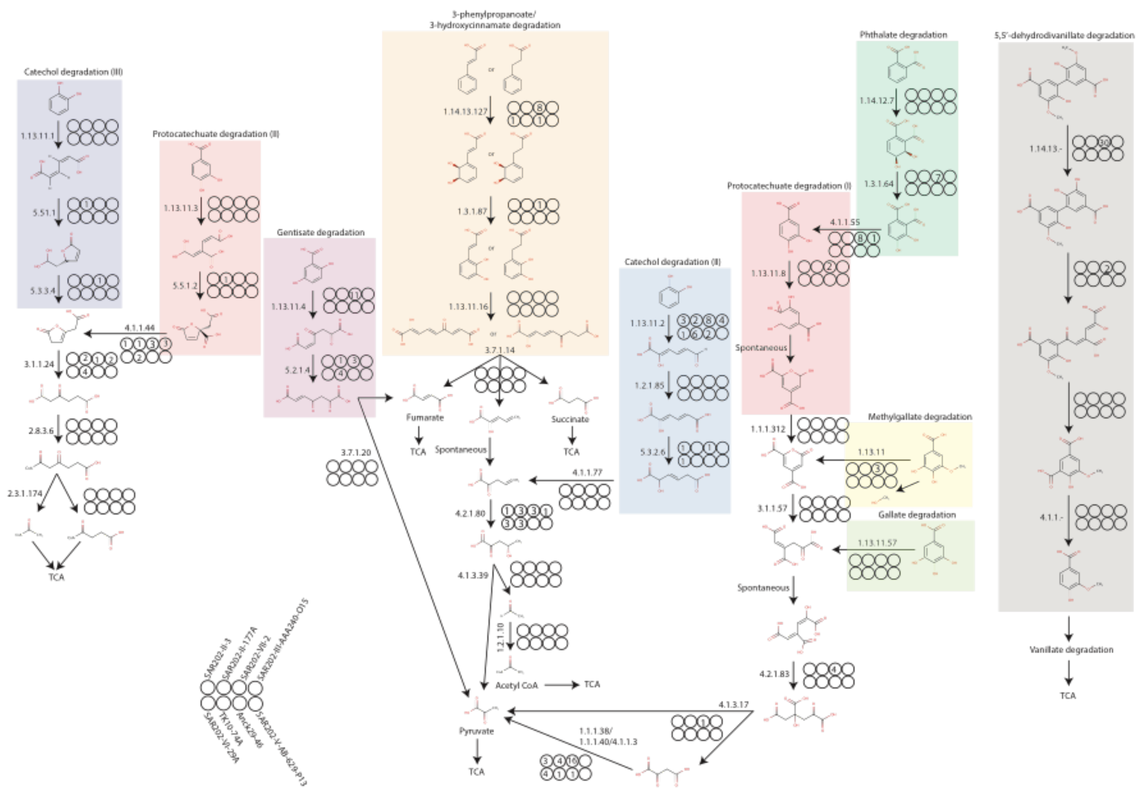
5.7 Acknowledgments

The data were collected aboard the CCGS Louis S. St-Laurent in collaboration with researchers from Fisheries and Oceans Canada at the Institute of Ocean Sciences and Woods Hole Oceanographic Institution's Beaufort Gyre Exploration Program and are available at http://www.whoi.edu/beaufortgyre. We would like to thank both the Captain and crew of the CCGS Louis S. St-Laurent and the scientific teams aboard. The work was conducted in collaboration with the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and was supported under Contract No. DE-AC02-05CH11231. Funding from the Canadian Natural Science and Engineering Research Council (NSERC) Discovery grants (D.W., C.G. and C.L.) and the Canada Research Chair Program (D.W., C.G.) are acknowledged. D.C. was supported by FRQNT and Concordia's Institute for Water, Energy and Sustainable Systems.

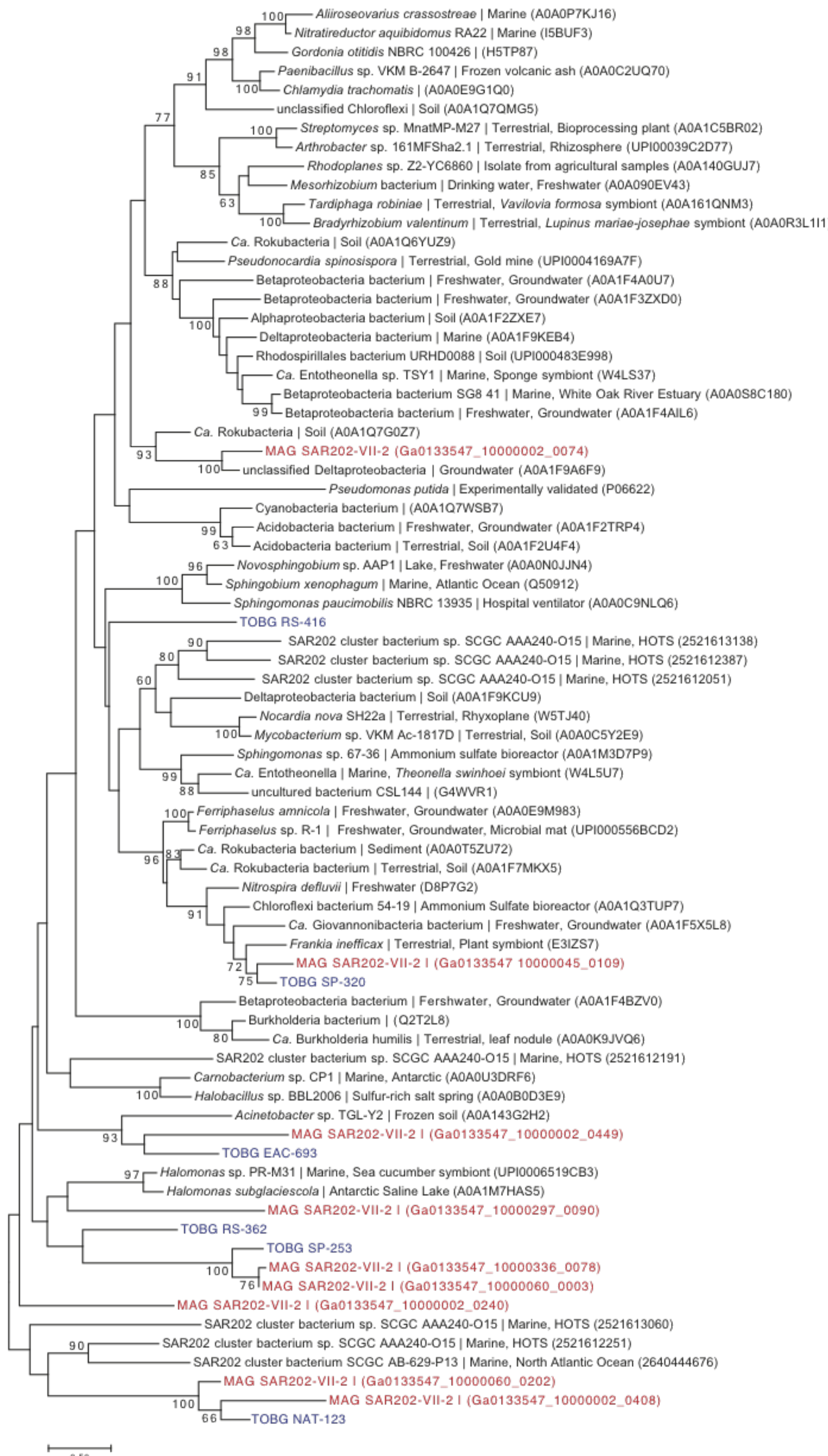
5.8 Supplementary figures



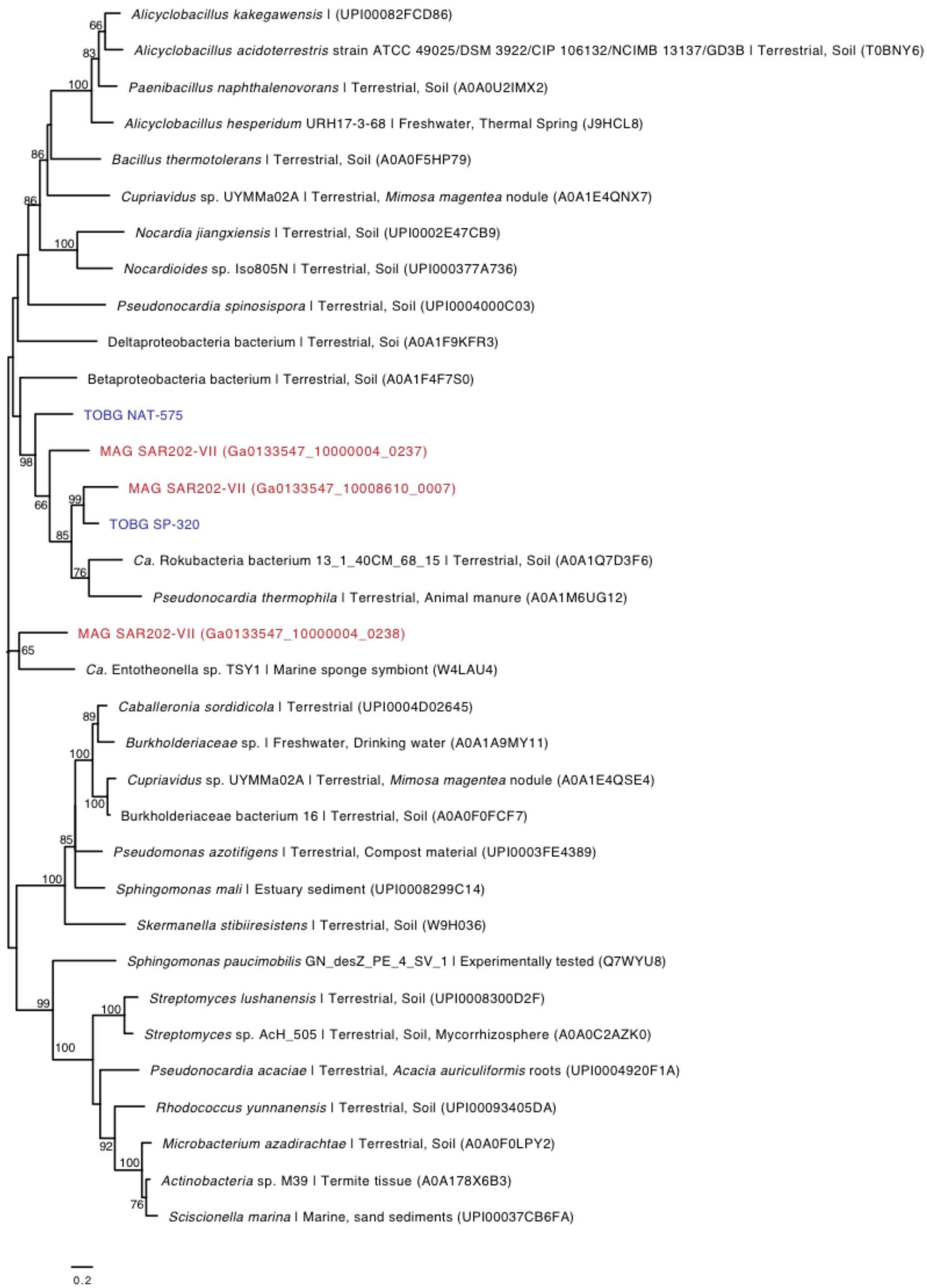
Supplementary figure 5.1. A concatenated protein phylogeny of Chloroflexi genomes including 30 Chloroflexi reference genomes and the 6 Canada Basin Chloroflexi MAGs. This is a maximum likelihood tree based on concatenated sequence alignment of fifty amino acid sequences (representing 14,815 amino acid positions), present in at least 24 of the 36 Chloroflexi genomes. Bootstrap values are based on 100 replicates.



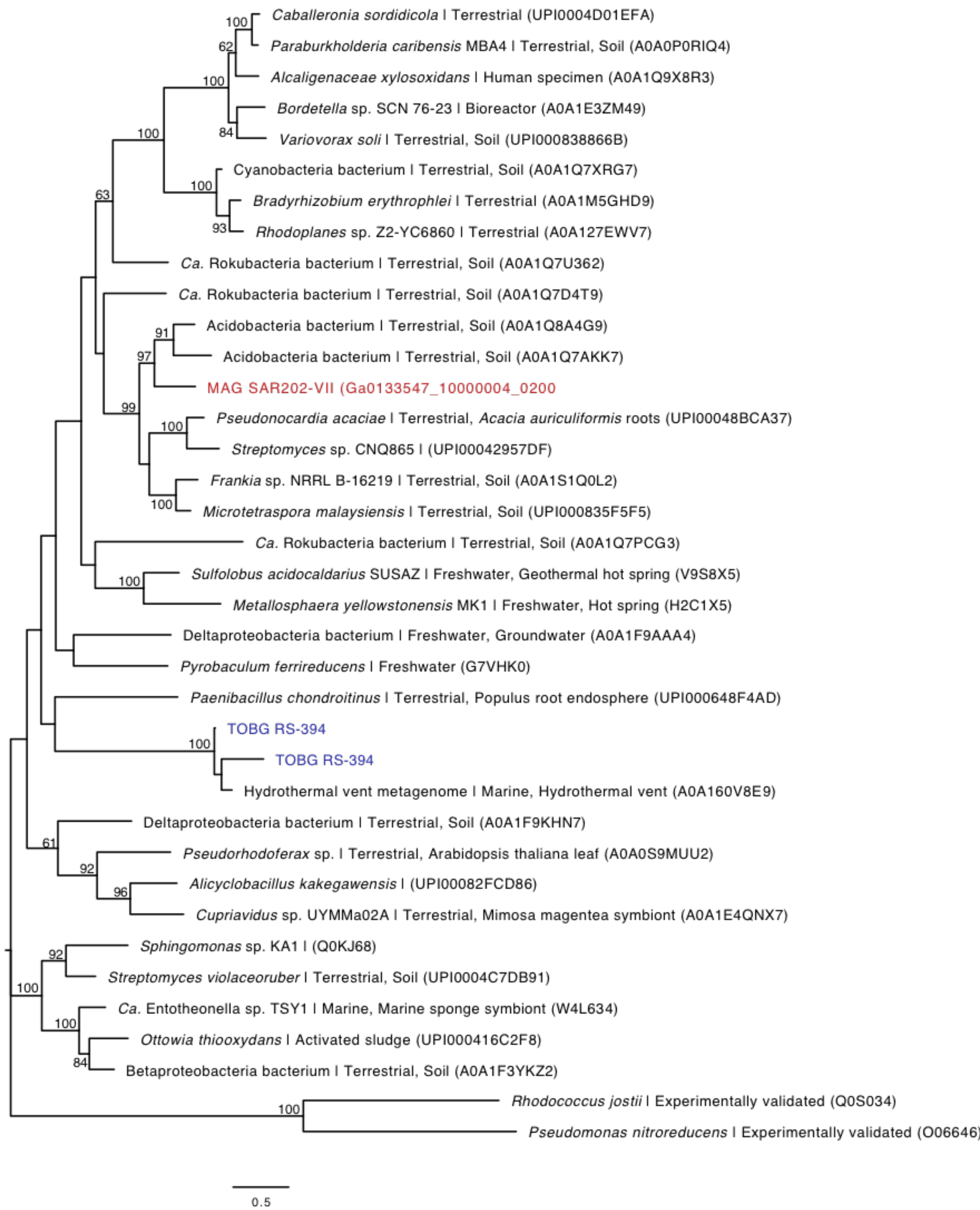
Supplementary figure 5.2. Aromatic compound degradation pathways identified in the Arctic Ocean and deep ocean Chloroflexi genomes. Circles represent the six Arctic Ocean MAGs and two deep ocean SAGs. The values in the circles represent the number of homologs identified in each respective genome. These reference metabolic pathways from MetaCyc were populated with proteins based on functional annotations available at the Integrated Microbial Genomes database.


Aliiroseovarius crassostreae | Marine (A0A0P7KJ16)
Nitratireductor aquibidomus RA22 | Marine (I5BUF3)
Gordonia otitidis NBRC 100426 | (H5TP87)
Paenibacillus sp. VKM B-2647 | Frozen volcanic ash (A0A0C2UQ70)
Chlamydia trachomatis | (A0A0E9G1Q0)
unclassified Chloroflexi | Soil (A0A1Q7QMG5)
Streptomyces sp. MnatMP-M27 | Terrestrial, Bioprocessing plant (A0A1C5BR02)
Arthrobacter sp. 161MFSha2.1 | Terrestrial, Rhizosphere (UPI00039C2D77)
Rhodoplanes sp. Z2-YC6860 | Isolate from agricultural samples (A0A140GUJ7)
Mesorhizobium bacterium | Drinking water, Freshwater (A0A090EV43)
Tardiphaga robiniae | Terrestrial, Vavilovia formosa symbiont (A0A161QNM3)
Bradyrhizobium valentinum | Terrestrial, Lupinus mariae-josephae symbiont (A0A0R3L1I1)
Ca. Rokubacteria | Soil (A0A1Q6YUZ9)
Pseudonocardia spinosispora | Terrestrial, Gold mine (UPI0004169A7F)
Betaproteobacteria bacterium | Freshwater, Groundwater (A0A1F4A0U7)
Betaproteobacteria bacterium | Freshwater, Groundwater (A0A1F3ZXD0)
Alphaproteobacteria bacterium | Soil (A0A1F2ZXE7)
Deltaproteobacteria bacterium | Marine (A0A1F9KEB4)
Rhodospirillales bacterium URHD0088 | Soil (UPI000483E998)
Ca. Entotheonella sp. TSY1 | Marine, Sponge symbiont (W4LS37)
Betaproteobacteria bacterium SG8 41 | Marine, White Oak River Estuary (A0A0S8C180)
Betaproteobacteria bacterium | Freshwater, Groundwater (A0A1F4AIL6)
Ca. Rokubacteria | Soil (A0A1Q7G0Z7)
MAG SAR202-VII-2 (Ga0133547_10000002_0074)
unclassified Deltaproteobacteria | Groundwater (A0A1F9A6F9)
Pseudomonas putida | Experimentally validated (P06622)
Cyanobacteria bacterium | (A0A1Q7WSB7)
Acidobacteria bacterium | Freshwater, Groundwater (A0A1F2TRP4)
Acidobacteria bacterium | Terrestrial, Soil (A0A1F2U4F4)
Novosphingobium sp. AAP1 | Lake, Freshwater (A0A0N0JJN4)
Sphingobium xenophagum | Marine, Atlantic Ocean (Q50912)
Sphingomonas paucimobilis NBRC 13935 | Hospital ventilator (A0A0C9NLQ6)
TOBG RS-416
SAR202 cluster bacterium sp. SCGC AAA240-O15 | Marine, HOTS (2521613138)
SAR202 cluster bacterium sp. SCGC AAA240-O15 | Marine, HOTS (2521612387)
SAR202 cluster bacterium sp. SCGC AAA240-O15 | Marine, HOTS (2521612051)
Deltaproteobacteria bacterium | Soil (A0A1F9KCU9)
Nocardia nova SH22a | Terrestrial, Rhyxoplane (W5TJ40)
Mycobacterium sp. VKM Ac-1817D | Terrestrial, Soil (A0A0C5Y2E9)
Sphingomonas sp. 67-36 | Ammonium sulfate bioreactor (A0A1M3D7P9)
Ca. Entotheonella | Marine, Theonella swinhoei symbiont (W4L5U7)
uncultured bacterium CSL144 | (G4WVR1)
Ferriphaselus amnicola | Freshwater, Groundwater (A0A0E9M983)
Ferriphaselus sp. R-1 | Freshwater, Groundwater, Microbial mat (UPI000556BCD2)
Ca. Rokubacteria bacterium | Sediment (A0A0T5ZU72)
Ca. Rokubacteria bacterium | Terrestrial, Soil (A0A1F7MKX5)
Nitrospira defluvii | Freshwater (D8P7G2)
Chloroflexi bacterium 54-19 | Ammonium Sulfate bioreactor (A0A1Q3TUP7)
Ca. Giovannonibacteria bacterium | Freshwater, Groundwater (A0A1F5X5L8)
Frankia inefficax | Terrestrial, Plant symbiont (E3IZS7)
MAG SAR202-VII-2 I (Ga0133547 10000045_0109)
TOBG SP-320
Betaproteobacteria bacterium | Fershwater, Groundwater (A0A1F4BZV0)
Burkholderia bacterium | (Q2T2L8)
Ca. Burkholderia humilis | Terrestrial, leaf nodule (A0A0K9JVQ6)
SAR202 cluster bacterium sp. SCGC AAA240-O15 | Marine, HOTS (2521612191)
Carnobacterium sp. CP1 | Marine, Antarctic (A0A0U3DRF6)
Halobacillus sp. BBL2006 | Sulfur-rich salt spring (A0A0B0D3E9)
Acinetobacter sp. TGL-Y2 | Frozen soil (A0A143G2H2)
MAG SAR202-VII-2 I (Ga0133547_10000002_0449)
TOBG EAC-693
Halomonas sp. PR-M31 | Marine, Sea cucumber symbiont (UPI0006519CB3)
Halomonas subglaciescola | Antarctic Saline Lake (A0A1M7HAS5)
MAG SAR202-VII-2 I (Ga0133547_10000297_0090)
TOBG RS-362
TOBG SP-253
MAG SAR202-VII-2 I (Ga0133547_10000336_0078)
MAG SAR202-VII-2 I (Ga0133547_10000060_0003)
MAG SAR202-VII-2 I (Ga0133547_10000002_0240)
SAR202 cluster bacterium sp. SCGC AAA240-O15 | Marine, HOTS (2521613060)
SAR202 cluster bacterium sp. SCGC AAA240-O15 | Marine, HOTS (2521612251)
SAR202 cluster bacterium SCGC AB-629-P13 | Marine, North Atlantic Ocean (2640444676)
MAG SAR202-VII-2 I (Ga0133547_10000060_0202)
MAG SAR202-VII-2 I (Ga0133547_10000002_0408)
TOBG NAT-123
0.50

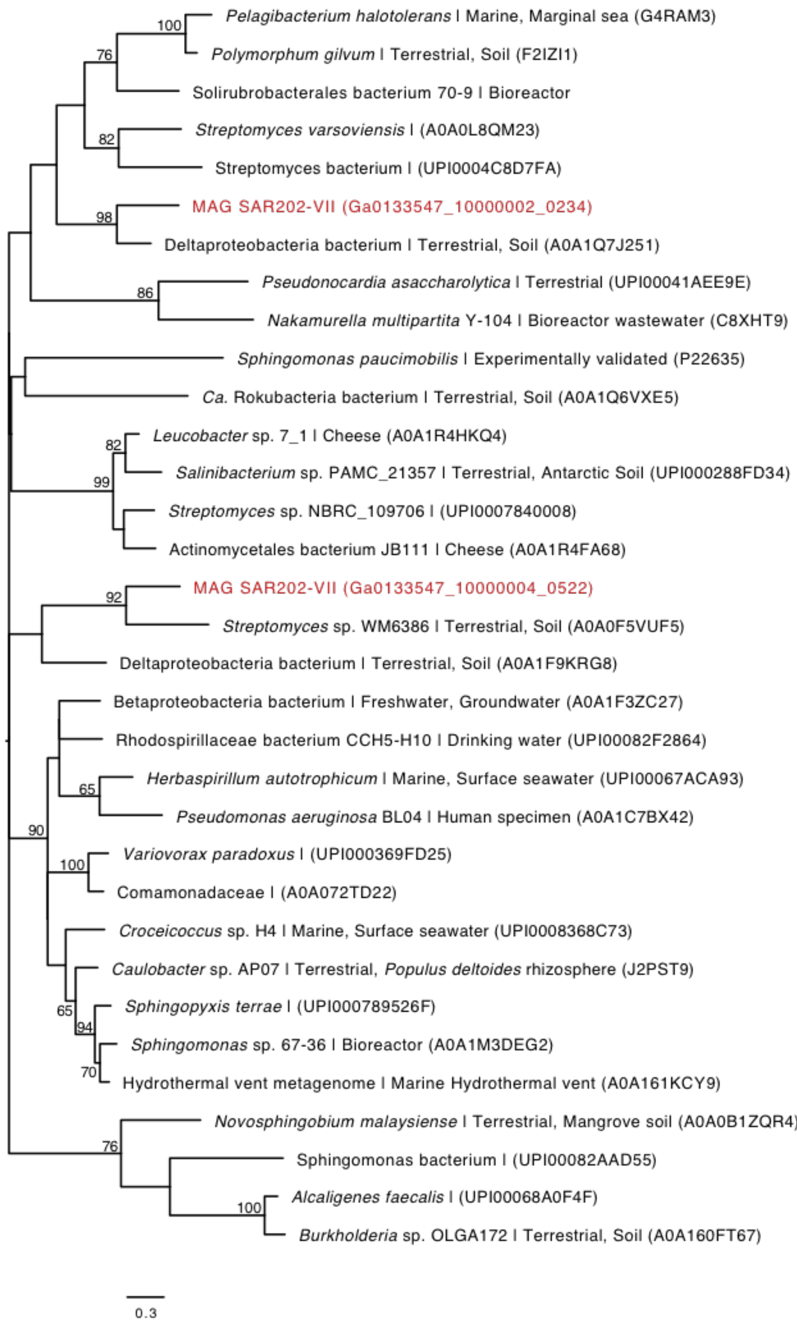
Supplementary figure 5.3. Phylogenetic analysis of predicted catechol 2,3-dioxygenase protein sequences identified in SAR202-VII-2. This is a maximum likelihood tree based on the alignmnet of catechol 2,3-dioxygenaase sequences identified in SAR202-VII-2, UniRef90 sequences and Chloroflexi MAGs constructed from the TARA Oceans dataset. Bootstrap values of >60% are shown (100 replicates). SAR202-VII-2 sequences are shown in red while sequences originating from the TARA Oceans MAGs are shown in blue.



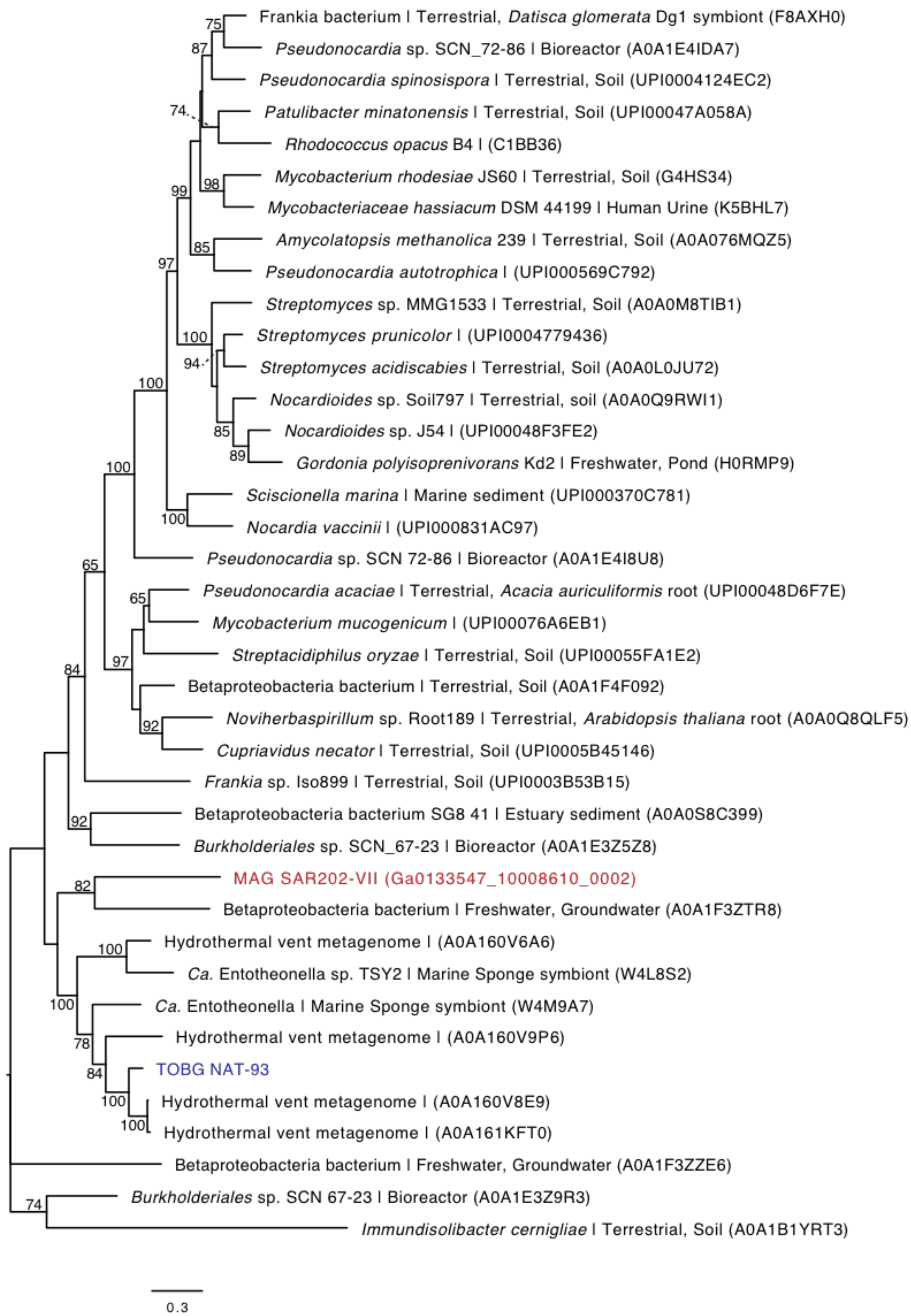
Supplementary figure 5.4. Phylogenetic analysis of predicted 3-O-methylgallate dioxygenase protein sequences identified in SAR202-VII-2. Method and description as described in Supplementary figure 5.3.



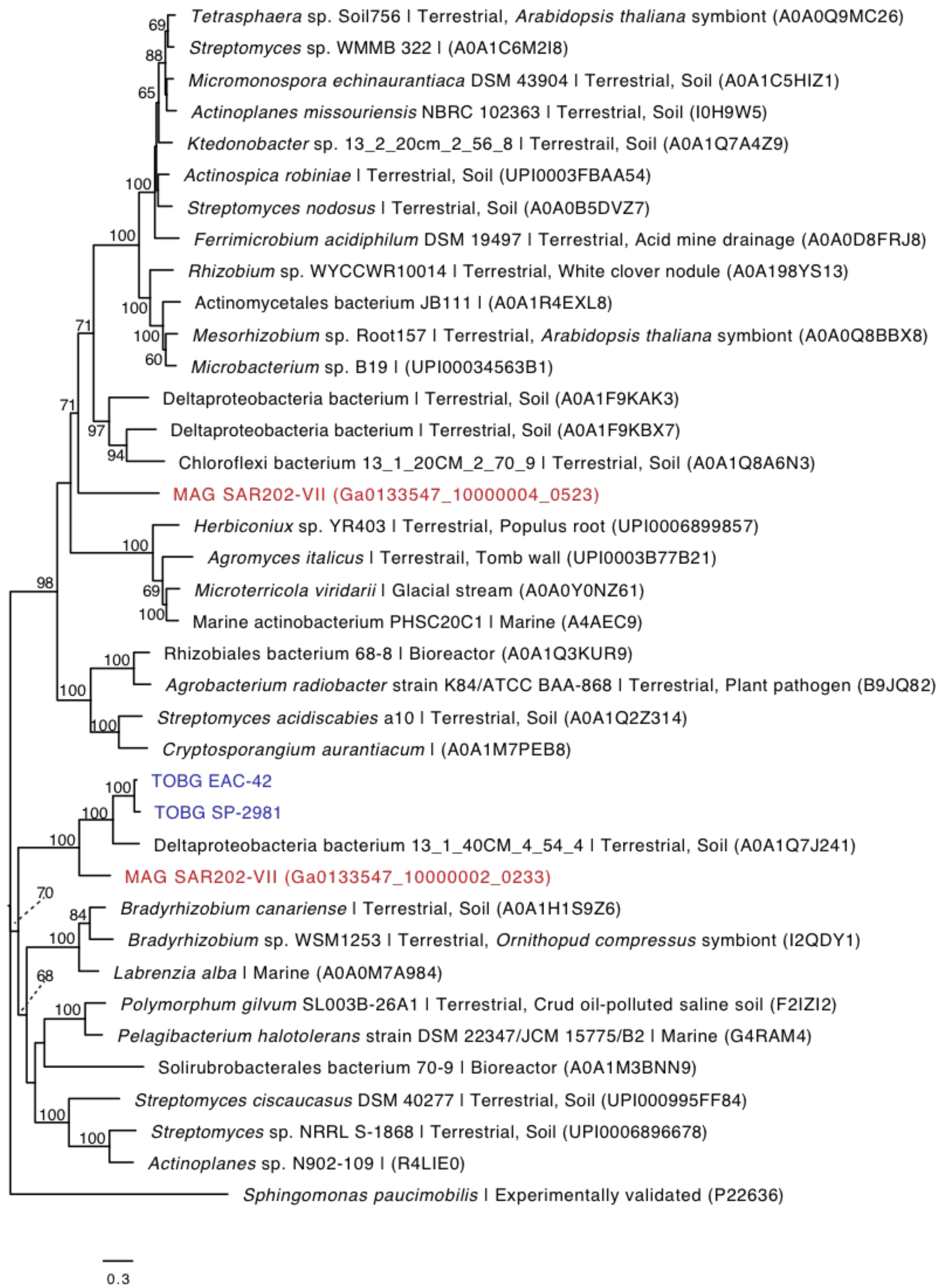
Supplementary figure 5.5. Phylogenetic analysis of a predicted merged LigA/LigB fusion protein sequence identified in SAR202-VII-2. Method and description as described in Supplementary figure 5.3.



Supplementary figure 5.6. Phylogenetic analysis of predicted LigA dioxygenase protein sequences identified in SAR202-VII-2. Method and description as described in Supplementary figure 5.3.



Supplementary figure 5.7. Phylogenetic analysis of a predicted LigB dioxygenase protein sequence identified in SAR202-VII-2. Method and description as described in Supplementary figure 5.3.



Supplementary figure 5.8. Phylogenetic analysis of predicted beta subunits of the protocatechuate 4,5-dioxygenase protein sequences identified in SAR202-VII-2. Method and description as described in Supplementary figure 5.3.

Chapter 6: Conclusions, discussion and future directions

The Arctic Ocean is comprised of a wide diversity of ecosystems, ranging from the highly stratified and terrestrially-influenced waters of the Beaufort Sea, to the North Water, the most productive region north of the Arctic Circle. However, the Arctic Ocean is one of the most understudied marine systems, and as such, our understanding of the microbial metabolic pathways and the community members that govern them in these regions is lacking. Moreover, in a time of rapid climate change, knowledge of microbial community structure and metabolism in the region most affected by warming will be invaluable for predicting how the ecology of the Arctic Ocean ecosystem will change.

In this thesis I defined the microbial community composition of the Lower Saint Lawrence Estuary, Canada Basin and the North Water. In Chapter 2, differences in community composition, nutrient acquisition and metabolic strategies between the surface and deep LSLE were identified. In Chapter 3, I used samples recovered from the North Water to develop a novel method to isolate microbial community DNA and proteins from the same environmental sample preserved in RNAlater. In Chapter 4, a combined 16S rRNA gene sequence and metaproteomic analysis of the North Water revealed two distinct microbial community compositions and nutrient acquisition and resource allocation strategies in the PML on either side of the North Water, as well as a third distinct community in the bottom waters. Both the functional and taxonomic analysis of the North Water suggest a microbial community more typical of those associated with waters that undergo dynamic pulses of primary production on the Canadian side, and those associated with a more steady level of primary production on the Greenland side. In Chapter 5, the first Arctic Ocean metagenomic dataset was generated and used to construct three hundred and sixty metagenome assembled genomes from the polar mixed layer of the Beaufort Sea. These MAGs can now be used as a resource for future studies investigating the Arctic Ocean marine microbial community composition and metabolism. While constructing these MAGs, an abundance of genes involved in aromatic compound degradation assigned to Chloroflexi were revealed, prompting a more in-depth analysis of the aromatic compound degrading potential of Chloroflexi in the Arctic Ocean. Evidence that some of the genes involved in aromatic compound degradation were acquired through lateral gene transfer events from terrestrial organisms was identified. Given these results, and the fact that the Arctic Ocean is highly influenced by terrestrial-derived water, we hypothesize that this ecosystem might be a hot-spot for terrestrial DOM degradation. Moreover, that terrestrial runoff might transport genes as well as tDOM and could contribute to local adaptations to increased tDOM in the Beaufort Sea.

6.1 Beaufort Sea Chloroflexi might be more abundant at depth

In Chapter 5, Chloroflexi were found to be relatively abundant below the chlorophyll maximum, at the FDOM maximum, in the cold halocline waters made up of the Pacific winter waters of the Beaufort Sea. Chloroflexi are typically associated with deep marine communities and can represent up 5% of the bacterioplankton community in the deep Beaufort Sea (Galand *et al.*, 2010). However, no metagenomic data was collected from depths below the FDOM maximum in this work. Because of this we do not know whether or not the abundance of Chloroflexi and genes involved in aromatic compound degradation observed in chapter 5 are actually the proverbial “tip of the iceberg” and would be identified at greater abundance with increasing depth. To determine whether or not this is true, a metagenomic analyses of samples collected from the deep and bottom waters of the Beaufort Sea could be performed.

However, the deeper Canada Basin waters are made up of Atlantic-derived waters. Because the North Water is the gateway between the Arctic and Atlantic Oceans, linking the two oceans we performed a new protein database search on the North Water MS/MS data generated in Chapter 4, and used the Beaufort Sea metagenomic data generated in Chapter 5 as a searchable database. We hypothesized that we would identify an abundance of Chloroflexi peptides in the North Water if they were abundant in the Atlantic Ocean, and therefore abundant in the deep Beaufort Sea waters. We then combed the resulting metaproteomic data looking for Chloroflexi-derived proteins. Only two proteins were assigned to the six Chloroflexi MAGs constructed in Chapter 5, none of which were assigned to SAR202-VII-2 or involved in aromatic compound degradation. These preliminary results suggest that they might not be as abundant in the deep Atlantic-derived waters, or that the Chloroflexi identified in the FDOM maximum of Beaufort Sea are not common to other Arctic regions. However, comparing proteomic data to genomic data is flawed because we do not know whether the aromatic compound degradation genes identified in the Beaufort Sea Chloroflexi are in fact being expressed. In order to truly compare the two datasets, one would have to either perform metaproteomic or metagenomic analyses on both regions.

With this in mind, a preliminary metaproteomic analysis of three samples from the Beaufort Sea FDOM maximum was conducted using the Beaufort Sea metagenomic data as a search database. Only seventy-two identified proteins were assigned to Chloroflexi, one of which was assigned to SAR202-VII-2 and none of which were involved in aromatic compound degradation. This could indicate a lack of expression of these aromatic compound degradation pathways, or, the expression of Chloroflexi assigned proteins might be low, in which case their signal could be overwhelmed by proteins expressed at higher abundance. The Walsh lab is currently conducting a more in-depth analysis of the microbial communities of the Arctic Ocean, utilizing a tandem metagenomic, metatranscriptomic and metaproteomic approach of Beaufort Sea microbial community samples from various depths including those below the FDOM maximum. However, to explore how common and phylogenetically widespread tDOM metabolism is in the Arctic Ocean it would be interesting to perform these meta-omic experiments in other regions of the Arctic Ocean, such as the East Siberian Sea, Baffin Bay and Canadian Archipelago. It would also be interesting to perform a meta-omic analysis of other marine regions that are highly influenced by terrestrial waters, like the mouth of the Amazon River, to establish if this enrichment in tDOM-degrading genes is localized to the Arctic, or a more widespread phenomena.

6.2 Incomplete aromatic compound degradation pathways

An abundance of genes involved in aromatic compound degradation were identified in Arctic Ocean Chloroflexi bacteria in chapter 5. However, we were unable to identify any complete metabolic pathways that went from aromatic compounds to TCA intermediates. The four main hypotheses we can draw from this are that either: 1) Genes are missing from the pathway due to the incompleteness of the MAGs, 2) Arctic Ocean Chloroflexi utilize novel pathways not used by reference organisms, and/or 3) Arctic Ocean Chloroflexi do not possess these metabolic steps and instead participate in a distributed metabolism utilizing metabolic handoffs, or 4) Arctic Ocean Chloroflexi do in fact have incomplete pathways and do not metabolize terrestrial-derived aromatic compounds.

The estimated completeness of the Cholorflexi MAG harbouring the majority of the aromatic compound degrading genes (SAR202-VII-2) was 99%. It is therefore unlikely that the incomplete metabolic pathways are due to incompleteness of the genome. However, it is conceivable that aromatic compound degradation might follow a different metabolic path in Arctic Ocean Chloroflexi than in the reference genomes currently used for metabolic pathway reconstructions. In fact, different pathways for aromatic compound degradation are known even within reference genomes from soil microbes. In *Pseudomonas putida,* protocatechuate is degraded via the ortho-cleavage pathway (Frazee *et al.*, 1993), while in *Sphingomonas paucimobilis,* protocatechuate is degraded via the meta-cleavage pathway (Noda *et al.*, 1990). Moreover, a major challenge in environmental metagenomic analysis is the annotations of abundant hypothetical proteins found in MAGs. Although full pathways similar to those found in reference genomes could not be identified, it is not to say that many of the hypothetical protein genes identified could not act as part of an alternate, previously undescribed pathway.

The third hypothesis is that Chloroflexi might be involved in some form of metabolic cross-feeding whereby they perform parts of the biochemical cycles and "handoff" the resulting metabolites to other participating organism. Typically, metabolic byproducts are "handed off" to other microbial phyla and used as nutrient sources (reviewed in: (D'Souza *et al.*, 2018)) including: carbon (Steven J. Biller *et al.*, 2014; Belenguer *et al.*, 2006), nitrogen (Dekas *et al.*, 2009), amino acids (Mee *et al.*, 2014; Rosenthal *et al.*, 2011; Payne *et al.*, 1957) and vitamins (Garcia *et al.*, 2015; Rosenthal *et al.*, 2011), but they can also be used in electron exchange (Boone *et al.*, 1989). An example of cross-feeding relationships is that of auxotrophic organisms acquiring necessary metabolites that are produced from other members of the community (Glen *et al.*, 2014; Morris *et al.*, 2012; Wintermute and Silver, 2010; Pande *et al.*, 2014). Moreover, auxotrophic interactions have also been shown to create additional syntrophic interactions that control community carbon and energy flux (Embree *et al.*, 2015). Another example would be between *Escherichia coli* mutants capable of surviving on the acetate produced by the wild-type community as a byproduct of glucose metabolism (Helling *et al.*, 1987; Treves *et al.*, 1998).

More recently, work in subsurface aquifers revealed interconnected networks of activities between organisms that produce metabolic products required by others in the community (Hug *et al.*, 2016). Additionally, this type of metabolic handoff was hypothesized to take place in another subterranean aquifer microbial community where a striking number of incomplete pathways within members of the community were identified (Anantharaman *et al.*, 2016). As in the Arctic Ocean Chloroflexi, members of the aquifer community were predicted to only perform parts of the biochemical pathways.

By performing meta-omic analyses on the Beaufort Sea samples in conjunction with new computational tools like TreeSAPP (Morgan-Lang *et al.*) or NetCooperate (Levy *et al.*, 2015), we can also gain insight into whether Chloroflexi are the sole phyla with the capabilities to degrade these aromatic compounds or if metabolic handoffs are occuring. TreeSAPP allows for a visual representation of the abundance of expressed proteins of a given metabolic pathway plotted on a phylogenetic tree of all organisms identified in metagenomic dataset, while NetCooperate helps to define metabolic complementarity between co-occurring microbes. In this way, community constituents with complimentary portions of metabolic pathways can be further investigated.

Along with Chloroflexi, another microbial phylum predominantly identified in the FDOM maximum was Acidobacteria. Very little is known about the metabolism of marine Acidobacteria as they are predominantly thought of as terrestrial organisms (Kielak *et al.*, 2016). Because the majority of the aromatic compounds identified in the FDOM maximum of the Beaufort Sea waters are thought to be of terrestrial origin (Guéguen *et al.*, 2012), and Acidobacteria are typically thought of as terrestrial organisms, we hypothesize that if metabolic handoffs are indeed taking place between Chloroflexi and other organisms in the Arctic Ocean, Acidobacteria would be a good candidate for further investigation. Preliminary data analysis suggests that Acidobacteria from a variety of subgroups were identified in the Beaufort Sea FDOM maximum. Like the Arctic Ocean Chloroflexi, these Acidobacteria seem to be most prevalent in the Arctic Ocean as indicated through reciprocal BLAST analysis (results not included). Initial analysis of the metabolic reconstruction of several Arctic Ocean Acidobacteria MAGs compared to the Chloroflexi MAGs did not suggest any form of metabolic handoffs between the two in terms of aromatic compound degradation. However, different groups of Acidobacteria MAGs do seem to complement one another.

Since the results in Chapter 5 are based on metagenomic data, whether or not these genes are actually expressed in the Arctic Ocean remains unknown. The final possibility for the incomplete metabolic pathways could be that these aromatic compound degradation pathways are no longer advantageous, and are in the process of being lost by genomic streamlining. One method to address this would be to search for pseudogenes in the Chloroflexi genomes. An abundance of pseudogenes have already been identified in insect symbionts in the process of streamlining (Burke and Moran, 2011; Toh *et al.*, 2006) as well as coral symbionts (Kwan *et al.*, 2012). In order to confirm the expression of these genes, post-genomic analysis of these samples should be performed. Performing a metaproteomic analysis using the metagenomes as a searchable protein database would provide this information. In light of this, a metaproteomic analysis of Beaufort Sea microbial communities is currently being conducted in the Walsh lab.

6.3 DOM characterization and degradation

One of the conclusions derived from Chapter 5 is that the Beaufort Sea might be a hotspot for terrestrial-derived dissolved organic matter transformation in the Arctic Ocean. This conclusion was based on the abundance of genes implicated in the degradation of aromatic compounds including lignin breakdown products identified in the metagenomic dataset. However, one of the major limitations of studying marine microbial organic matter processing, and in particular microbe-DOM dynamics, is the ill-defined nature of marine DOM owing to the low proportion of DOM identified as specific known biomolecules (Williams and Druffel, 1988; Benner, 1998; Hedges *et al.*, 2000). The majority of current research involved in the uptake and utilization of DOM in marine systems is based on the functional annotation of genes and proteins in metagenomic and metaproteomic datasets inferred by searching databases of canonical metabolic pathways found in reference organisms. Therefore, it is conceivable that in enigmatic bacteria collected from environments replete with a diversity of unknown DOM compounds, the reference metabolic pathways may not apply, and that the specific gene products in question may act on substrates different than their reference database orthologs. Although bacterial transformation of DOM can profoundly modify the biogeochemical behaviour of organic matter (Azam, 1998) and microbial DOM transformation is integral to the biogeochemical cycles of our planet, a lack of knowledge on the chemical composition of marine DOM, and the undefined nature of the annotated genes used for marine meta-omic studies results in a limited understanding of the molecular mechanisms of DOM transformation in the oceans.

To increase our current understanding of microbe-DOM interactions, analyses to elucidate the chemical composition of DOM found in meta-omic samples, as well as the metabolic pathways responsible for their transformation should be performed. One way to accomplish this would be to collect paired sets of samples alongside those meant for meta-omic analysis that can be used for ultra-high resolution mass spectrometry like fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) analysis. Using FT-ICR-MS, the elemental composition of molecules making up the complex DOM pool can be resolved (Koch *et al.*, 2005). Although chemical structures cannot be resolved, Van Krevelen plots of the elemental compositions can then be used to determine the different compound classes found in the complex samples. Similar methods have already been used in fresh water systems to elucidate the succession of DOM degradation along a soil-stream-river continuum (Hutchins *et al.*, 2017). In this study, biopolymer and low molecular weight (LMW) compounds were the first to be degraded at the soil-stream interface, followed by humic-like aromatic DOM degradation (Hutchins *et al.*, 2017). In marine systems, this method was adopted to explore the molecular structure of marine DOM for thermogenic signatures in different water masses of the Southern Ocean (Dittmar and Koch, 2006) and to identify the elemental composition of ~300 DOM molecules in the deep Atlantic Ocean (Reemtsma *et al.*, 2008). By performing this type of DOM analysis on multiple samples over a geographic or temporal range and in tandem with meta-omic analysis, correlations between DOM degradation gene (or gene product) abundances and changes in the abundance of certain compound classes or chemical formulas can be made. Therefore, it would be very interesting to perform a joint meta-omic DOM analysis along a geographic transect from the mouth of the Mackenzie River to more open Arctic ocean waters, or from the surface waters to the deeper waters of the Arctic Ocean.

Additionally, meso- or microcosm experiments that utilize joint FT-ICR-MS and post-genomic analyses, like metatranscriptomics or metaproteomics, can be employed to determine metabolic pathway expression levels that correlate with the production or disappearance of compound classes. Terrestrial-derived DOM amended mesocosm experiments in Northern ecosystems have already been successfully implemented to showcase changes in community composition after allochthonous DOM additions (Lindh *et al.*, 2015; Müller *et al.*, 2018). These studies utilized 16S rRNA gene sequencing to identify changes in community structure, but by using post-genomic approaches, changes in metabolic gene expression could be established to determine which genes are most important for tDOM metabolism. Joint transcriptomic-FT-ICR-MS work has already been used to identify the DOM compounds that are most important for microbial communities off the coast of Georgia, as well as the probable genes that drive DOM metabolism (Vorobev *et al.*, 2018).

By performing short time series experiments combining meta-omic analyses, FT-ICR-MS and meso- or microcosms enriched with tDOM, we would be able to track DOM compound class profiles over time as well as the gene expression levels that correlate to them. The genes or metabolic pathways that correlate with DOM compound class changes could then be targeted for further exploration. Not only would this allow for the potential identification of genes important for tDOM degradation/transformation, but mesocosm experiments could, in theory, also create enrichment cultures of community constituents important for tDOM degradation. These enrichment cultures with simplified communities could then be further explored to determine if metabolic handoffs are being performed by their members.

6.4 Improved metaproteomic analysis

Since protein sequences contain taxonomic information and proteins represent a major metabolic investment, proteomic analysis can be used to infer the functional community composition and resource allocation strategies of the community, as seen in Chapters 2 and 4. However, our ability to do so is restricted by the limitations of metaproteomics. Those limitations include incorrect phylogenetic assignment of spectra, a semi-quantitative assessment of the data rather than absolute quantification, and an abundance of unassigned spectra, with generally only ~10–30% of MS/MS spectra matched to peptide sequences with high confidence. Several factors contribute to this low assignment. Parent peptides might not have proper representation in the search database, have post-translational modifications, missed cleavage sites, or truncations (Picotti *et al.*, 2007). Additionally, peptides with similar m/z ratios may co-elute, resulting in a contaminated parent ion that will be co-fragmented. As a result, the chimeric spectra will have reduced searched scores (Houel *et al.*, 2010).

Issues of co-eluting parent peptides can be minimized by reducing the complexity of the sample by fractionation prior to trypsin digestion, or by increasing elution times. Additionally, programs such as reSpect, aimed at disentangling chimeric spectra, can be used to increase peptide identification (Shteynberg *et al.*, 2015). Performing a joint metagenomic-metaproteomic analysis on the same samples could also help mitigate some limitations associated with a non-representative protein search database. By using a metagenomic dataset from the same sample as the metaproteomic dataset as the searchable protein database, the results would be less prone to errors in phylogenetic assignment. We would also be able to use ratios of PSMs to number of reads to further our understanding of which organisms are most present and which are the most active for a particular metabolism in a given ecosystem. Another way to address the problem of conserved peptide sequences between taxa is to employ a peptide-centric method combined with lowest common ancestor approaches. One such method is UniPept (Mesuere *et al.*, 2012). Programs like UniPept can be used to identify peptide biomarkers for target taxa (taxon-specific peptides) which may then allow absolute quantification by targeted proteomics (Mesuere *et al.*, 2016). Additionally, Unipept, and more recent programs like MetaGOmics, can also be used as a peptide-centric approach for not only taxonomic, but functional analysis of metaproteomic data (Gurdeep Singh *et al.*, 2019; Riffle *et al.*, 2017).

Once a taxon or metabolism of interest is identified, specific metaproteomic biomarkers can be identified, and targeted proteomics can be employed to quantify their expression. Targeted metaproteomics uses isotopically-labelled peptide reference standards and selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) mass spectrometry to identify and quantify peptides representing the proteins of interest from a given taxon in a specific environment. This method has already been successfully employed in the Central Pacific

Ocean to identify differences in nitrogen regulatory protein utilization preferences between different cyanobacteria (Saito *et al.*, 2015), to confirm the expression of vitamin B_{12}-synthesizing genes from novel bacteria in the Ross Sea (Bertrand *et al.*, 2011), and vitamin B_{12} metabolism genes in diatoms from McMurdo Sound (Bertrand *et al.*, 2013). Given the observed enrichment in aromatic compound degrading genes identified in the Canada Basin in Chapter 5, and the hypothesis that these genes were enriched to degrade terrestrial dissolved organic matter, targeted metaproteomics could be of particular interest for determining the expression levels of these proteins and how expression levels change from the coast to the open ocean. Using a combination of the peptide-centric approach for the taxonomic and functional characterization of the communities, and the identification of peptide biomarkers for proteins involved in interesting metabolisms (like terrestrial organic matter degradation proteins), followed by targeted metaproteomics to quantify these peptide biomarkers between samples collected at different depths and distances from shore could help to answer questions like: 1) does terrestrial dissolved organic matter metabolism become more or less important with depth and 2) does terrestrial dissolved organic matter metabolism become more or less important with distance from shore? The opposing reasonings behind these questions are that regions closer to the source of tDOM, e.g. closer to the shore or higher up in the water column, will have higher tDOM concentrations and will enrich for tDOM degrading microbial populations. The opposing view is that regions closer to the shore or higher up in the water column have higher concentration of labile organic matter, and favour the enrichment of communities that can readily use the labile DOM. As labile DOM becomes more scarce (moving away from shore or deeper in the water column) tDOM metabolism might become more important.

6.6 Long-term monitoring

Given the rapid environmental changes being experienced by the Arctic and our relative lack of understanding about the Arctic Ocean microbial community, a long-term goal performing yearly metagenomic analyses would be interesting. One way to do this would be to perform a similar study to the one described in Chapter 5, but taking a function-centric approach rather than a taxonomic-centric approach to track changes in microbial community function over time. A long-term 16S rRNA gene sequence analysis has already been used to reveal changes in microbial community composition over time and in response to a loss of sea ice extent leading to fresher surface waters (Comeau *et al.*, 2011). Using a metagenomic approach rather than 16S rRNA gene sequence analysis would allow for the discovery of metabolic pathways that are more susceptible to the changing ecosystem. Tracking the relative abundance of genes or metabolic pathways through time, and correlating them with observed changes could provide meaningful data to help model the types of functional changes we can expect from communities that will be affected by climate change.

6.7 Implications for a warming Arctic Ocean

Finally, although the findings discussed in this thesis help broaden our understanding of marine microbial diversity, community composition and metabolism in northern marine ecosystems, they also have implications for our understanding of how climate change will affect the microbial communities and metabolic processes occurring in these same ecosystems. Increased atmospheric warming is leading to increased glacier melt, as well as thawing permafrost, which in turn will lead to increased freshwater runoff from terrestrial sources. Given that an estimated 60% of soil organic matter resides in taiga and tundra soils found in Arctic drainage basins (Dixon *et al.*, 1994), thawing permafrost due to global warming could elevate the amount of terrigenous DOM discharged into the Arctic Ocean. Based on the findings in Chapter 5 we hypothesize that an increase in terrestrial runoff will not only result in greater transport of allochthonous carbon to the Arctic Ocean, but also a potentially greater abundance of terrestrial genes. The increase in tDOM might increase the importance of Arctic marine Chloroflexi and other members of the microbial community capable of tDOM degradation. Additionally, as it seems that Arctic Ocean Chloroflexi obtained at least some of its aromatic degradation genes from lateral gene transfer (LGT) events from terrestrial organisms, there is the potential for an increase in LGT events from terrestrial to marine organisms as the Arctic warms.

Atmospheric warming is perhaps already affecting the North Water. The North Water is a relatively small area that is known to be the most productive region north of the Arctic Circle. However, as described in Chapter 4, both sides of the North Water have distinct community composition and metabolic profiles. The Eastern Greenland side is composed primarily of bacterioplankton typical of steady primary production, typically seen in the open ocean, while the Canadian side has a bacterioplankton composition typical of those associated with more blooming productive ecosystems. One reason for this discrepancy could be the increased freshwater runoff from Greenland glacier melt on the eastern side. The freshwater entering on the eastern side decreases the salinity of the marine surface waters which in turn results in greater density stratification of the water column, altering nutrient cycling in the water column, minimizing phytoplankton access to nutrients found in the deeper waters. As freshwater continues to flow into the Arctic Ocean, overall Arctic Ocean microbial communities might start to shift to resemble those identified in the E-PML in terms of community composition and metabolism.

The Arctic is the region of the Earth that is most quickly being affected by climate change. Given that physical and chemical changes to the Arctic Ocean as a result of atmospheric warming are already being observed, and the fact that these changes are unlikely to slow down or stop any time in the near future, it is imperative that further work be performed to understand how these changes will affect this unique marine ecosystem, and the other ocean systems that these waters feed.

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, *et al.* (2004). Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**:551–554.

Allison SD, Martiny JBH. (2008). Resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci* **105**:11512–11519.

Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, *et al.* (2014). Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**:1144–1146.

Alonso-sáez L, Arístegui J, Pinhassi J, Gómez-consarnau L, González JM, Vaqué D, *et al.* (2007). Bacterial assemblage structure and carbon metabolism along a productivity gradient in the NE Atlantic Ocean. *Aquat Microb Ecol* **46**:43–53.

Alonso-Sáez L, Galand PE, Casamayor EO, Pedrós-Alió C, Bertilsson S. (2010). High bicarbonate assimilation in the dark by Arctic bacteria. *ISME J* **4**:1581–1590.

Alonso-Sáez L, Gasol JM. (2007). Seasonal variations in the contributions of different bacterial groups to the uptake of low-molecular-weight compounds in Northwestern Mediterranean coastal waters. *Appl Environ Microbiol* **73**:3528–3535.

Alonso-Sáez L, Waller AS, Mende DR, Bakker K, Farnelid H, Yager PL, *et al.* (2012). Role for urea in nitrification by polar marine Archaea. *Proc Natl Acad Sci U S A* **109**:17989–94.

Alonso C, Pernthaler J. (2006). Roseobacter and SAR11 dominate microbial glucose uptake in coastal North Sea waters. *Environ Microbiol* **8**:2022–2030.

Aluwihare LI, Repeta DJ, Chen RF. (1997). A major biopolymeric component to dissolved organic carbon in surface sea water. *Nature* **387**:166–169.

Amon RMW. (2003). The Role of Dissolved Organic Matter for the Organic Carbon Cycle in the Arctic Ocean The Arctic Ocean Organic Carbon Cycle : doi:10.1007/978-3-642-18912-8.

Amon RMW, Meon B. (2004). The biogeochemistry of dissolved organic matter and nutrients in two large Arctic estuaries and potential implications for our understanding of the Arctic Ocean system. *Mar Chem* **92**:311–330.

Anantharaman K, Brier JA, Sheik CS, Dick GJ. (2013). Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proc Natl Acad Sci* **110**:330–335.

Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, *et al.* (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an

aquifer system. *Nat Commun* **7**:1–11.
Anderson LG. (2002). DOC in the Arctic Ocean. In:*Biogeochemistry of Marine Dissolved Organic Matter*, Hansell, DA & Carlson, CA (eds), Elsevier Inc., pp. 665–683.
Ardyna M, Gosselin M, Michel C, Poulin M, Tremblay JÉ. (2011). Environmental forcing of phytoplankton community structure and function in the Canadian High arctic: Contrasting oligotrophic and eutrophic regions. *Mar Ecol Prog Ser* **442**:37–57.
Arrigo KR. (2005). Marine microorganisms and global nutrient cycles. *Nature* **437**:349–355.
Arrigo KR, van Dijken G, Pabi S. (2008). Impact of a shrinking Arctic ice cover on marine primary production. *Geophys Res Lett* **35**:1–6.
Azam F. (1998). Microbial Control of Oceanic Carbon Flux : The Plot Thickens. *Science* **280**:694–696.
Bâcle J, Carmack EC, Ingram RG. (2002). Water column structure and circulation under the North Water during spring transition: April-July 1998. *Deep Res Part II Top Stud Oceanogr* **49**:4907–4925.
Bano N, Hollibaugh J. (2002). Phylogenetic composition of bacterioplankton assemblages from the Arctic Ocean. *Appl Environ Microbiol* **68**:505–518.
Barry KP, Taylor EA. (2014). Characterizing the Promiscuity of LigAB, a Lignin Catabolite Degrading Extradiol Dioxygenase from Sphingomonas paucimobilis SYK-6. *Biochemistry* **6**:1–16.
Beja O, Aravind L, Koonin E V., Suzuki MT, Hadd A, Nguyen LP, *et al.* (2000). Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science* **289**:1902–1906.
Belenguer A, Duncan SH, Calder a G, Holtrop G, Louis P, Lobley GE, *et al.* (2006). Two Routes of Metabolic Cross-Feeding between Bifidobacterium adolescentis and Butyrate-Producing Anaerobes from the Human Gut Alvaro. *Appl Environ Microbiol* **72**:3593–3599.
Benner R. (1998). Cycling of dissolved organic matter in the ocean. In:*Aquatic Humic Substances: Ecology and Biogeochemistry*, Hessen, DO & L.J., T (eds), Springer: Berlin, Heidelberg, pp. 317–331.
Benner R. (1997). What happens to terrestrial organic matter in the ocean? *Org Geochem* **27**:195–212.
Benner R, Benitez-Nelson B, Kaiser K, Amon RMW. (2004). Export of young terrigenous dissolved organic carbon from rivers to the Arctic Ocean. *Geophys Res Lett* **31**:1–4.

Benner R, Pakulski JD, McCarthy M, Hedges JI, Hatcher PG. (1992). Bulk Chemical Characteristics od Dissolved Organic Matter in the Oceans. *Science* **225**:1561–1564.
Berg IA, Kockelkorn D, Ramos-Vera WH, Say RF, Zarzycki J, Hügler M, *et al.* (2010). Autotrophic carbon fixation in archaea. *Nat Rev Microbiol* **8**:447–460.
Bergauer K, Fernandez-Guerra A, Garcia JAL, Sprenger RR, Stepanauskas R, Pachiadaki MG, *et al.* (2017). Organic matter processing by microbial communities throughout the Atlantic water column as revealed by metaproteomics. *Proc Natl Acad Sci* E400–E408.
von Bergen M, Jehmlich N, Taubert M, Vogt C, Bastida F, Herbst F-A, *et al.* (2013). Insights from quantitative metaproteomics and protein-stable isotope probing into microbial ecology. *ISME J* **7**:1877–85.
Bertrand E, Moran D, McIlvin M, Hoffman J, Allen A, Saito M. (2013). Methionine synthase interreplacement in diatom cultures and communities: Implications for the persistence of B12 use by eukaryotic phytoplankton. *Limnol Oceanogr* **58**:1431–1450.
Bertrand EM, Saito MA, Jeon YJ, Neilan BA. (2011). Vitamin B12 biosynthesis gene diversity in the Ross Sea: the identification of a new group of putative polar B12 biosynthesizers. *Environ Microbiol* **13**:1285–1298.
Biller SJ, Berube PM, Dooley K, Williams M, Satinsky BM, Hackl T, *et al.* (2018). Data Descriptor : Marine microbial metagenomes sampled across space and time. *Sci Data* 1–7.
Bintanja R, Andry O. (2017). Towards a rain-dominated Arctic. *Nat Clim Chang* **7**:263–267.
Blanvillain S, Meyer D, Boulanger A, Lautier M, Guynet C, Denancé N, *et al.* (2007). Plant carbohydrate scavenging through TonB-dependent receptors: A feature shared by phytopathogenic and aquatic bacteria. *PLoS One* **2**. doi:10.1371/journal.pone.0000224.
Blough N V., Del Vecchio R. (2002). Chromophoric DOM in the COastal Environment. In:*Biochemistry of Marine Dissolved Organic Matter*, Hansell, DA & Carlson, CA (eds), Academic Press: London, pp. 509–546.
Boone DR, Johnson RL, Liu Y. (1989). Diffusion of the Interspecies Electron Carriers H2 and Formate in Methanogenic Ecosystems and Its Implications in the Measurement of Km for H2 or Formate Uptake. *Appl Environ Microbiol* **55**:1735–1741.
Buchfink B, Xie C, Huson DH. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**:59–60.
Burke GR, Moran NA. (2011). Massive genomic decay in Serratia symbiotica, a recently

evolved symbiont of aphids. *Genome Biol Evol* **3**:195–208.
Canfield DE, Stewart FJ, Thamdrup B, De Brabandere L, Dalsgaard T, Delong EF, *et al.* (2010). A cryptic sulfur cycle in oxygen-minimum-zone waters off the Chilean coast. *Science* **330**:1375–1378.
Carder KL, Steward RG. (1989). Marine Humic and Fulvic-Acids - Their Effects on Remote-Sensing of Ocean Chlorophyll. *Limnol Oceanogr* **34**:68–81.
Carlson CA. (2002). Production and removal processes. In:*Biogeochemistry of Marine Dissolved Organic Matter*, Hansell, DA & Carlson, CA (eds), Press, Academic, pp. 91–152.
Carmack EC. (2000). The Arctic Ocean's Freshwater Budget: Sources, Storage and Export. In:*The Freshwater Budget of the Arctic Ocean*, Lewis, EL, Jones, EP, Lemke, P, Prowse, TD, & Wadhams, P (eds), Springer Netherlands, pp. 91–126.
Carmack EC, Yamamoto-Kawai M, Haine TWN, Bacon S, Bluhm BA, Lique C, *et al.* (2016). Freshwater and its role in the Arctic Marine System: Sources, disposition, storage, export, and physical and biogeochemical consequences in the Arctic and global oceans. *J Geophys Res G Biogeosciences* **121**:675–717.
Chandler DP, Brockman FJ, Bailey TJ, Fredrickson JK. (1998). Phylogenetic Diversity of Archaea and Bacteria in a Deep Subsurface Paleosol. *Microb Ecol* **36**:37–50.
Clark LL, Ingall ED, Benner R. (1998). Marine phosphorus is selectively remineralized. *Nature* **393**:1998.
Coble PG. (1996). Characterization of marine and terrestrial DOM in seawater using excitation-emission matrix spectroscopy. *Mar Chem* **51**:325–346.
Colatriano D, Ramachandran A, Yergeau E, Maranger R, Gélinas Y, Walsh DA. (2015). Metaproteomics of aquatic microbial communities in a deep and stratified estuary. *Proteomics* **15**:1-14.
Colatriano D, Tran P, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. (2018). Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. *Commun Biol* **1**:1-9
Colatriano D, Walsh DA. (2015). An Aquatic Microbial Metaproteomics Workflow: From Cells to Tryptic Peptides Suitable for Tandem Mass Spectrometry-based Analysis. *J Vis Exp* 1–8.
Comeau AM, Li WKW, Tremblay J-É, Carmack EC, Lovejoy C. (2011). Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. *PLoS One* **6**:e27492.

Connelly TL, Baer SE, Cooper JT, Bronk D a, Wawrik B. (2014). Urea Uptake and Carbon Fixation by Marine Pelagic Bacteria and Archaea During the Arctic Summer and Winter Seasons. *Appl Environ Microbiol* **80**:6013–6022.

Covert JS, Moran MA. (2001). Molecular characterization of estuarine bacterial communities that use high- and low-molecular weight fractions of dissolved organic carbon. *Aquat Microb Ecol* **25**:127–139.

D'Souza G, Shitut S, Preussger D, Yousif G, Waschina S, Kost C. (2018). Ecology and evolution of metabolic cross-feeding interactions in bacteria. *Nat Prod Rep* **35**:455–488.

Dainard PG, Guéguen C. (2013). Distribution of PARAFAC modeled CDOM components in the North Pacific Ocean. *Mar Chem* **157**:216–223.

Dainard PG, Guéguen C, McDonald N, Williams WJ. (2015). Photobleaching of fluorescent dissolved organic matter in Beaufort Sea and North Atlantic Subtropical Gyre. *Mar Chem* **177**:630–637.

Dekas AE, Poretsky RS, Orphan VJ. (2009). Deep-Sea archaea fix and share nitrogen in methane-consuming microbial consortia. *Science* **326**:422–426.

DeLong E, Preston C, Mincer T, Rich V. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**:496–503.

Delong EF. (1992). DeLong, 1992 - Archaea in coastal marine environment. *Proc Natl Acad Sci* **89**:5685–5689.

DeLong EF. (2002). Microbial population genomics and ecology. *Curr Opin Microbiol* **5**:520–524.

DeLorenzo S, Bräuer SL, Edgmont C a, Herfort L, Tebo BM, Zuber P. (2012). Ubiquitous dissolved inorganic carbon assimilation by marine bacteria in the Pacific Northwest coastal ocean as determined by stable isotope probing. *PLoS One* **7**:e46695.

Dittmar T, Kattner G. (2003). The biogeochemistry of the river and shelf ecosystem of the Arctic Ocean: a review. *Mar Chem* **83**:103–120.

Dittmar T, Koch BP. (2006). Thermogenic organic matter dissolved in the abyssal ocean. *Mar Chem* **102**:208–217.

Dixon ARK, Brown S, Houghton RA, Solomon AM, Trexler MC, Dixon RK, *et al.* (1994). Carbon Pools and Flux of Global Forest Ecosystems. *Science* **263**:185–190.

Dixon Joanna L, Beale R, Nightingale PD. (2011). Microbial methanol uptake in northeast

Atlantic waters. *ISME J* **5**:704–16.
Dixon J. L., Beale R, Nightingale PD. (2011). Rapid biological oxidation of methanol in the tropical Atlantic: Significance as a microbial carbon source. *Biogeosciences* **8**:2707–2716.
Dixon JL, Sargeant S, Nightingale PD, Colin Murrell J. (2013). Gradients in microbial methanol uptake: productive coastal upwelling waters to oligotrophic gyres in the Atlantic Ocean. *ISME J* **7**:568–80.
Dunbar J, Barns SM, Ticknor LO, Kuske CR. (2002). Empirical and theoretical bacterial diversity in four Arizona soils. *Appl Environ Microbiol* **68**:3035–45.
Edgar RC. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797.
Eisenbeis S, Lohmiller S, Valdebenito M, Leicht S, Braun V. (2008). NagA-dependent uptake of N-acetyl-glucosamine and N-acetyl-chitin oligosaccharides across the outer membrane of Caulobacter crescentus. *J Bacteriol* **190**:5230–5238.
El-Swais H, Dunn K a., Bielawski JP, Li WKW, Walsh DA. (2015). Seasonal assemblages and short-lived blooms in coastal north-west Atlantic Ocean bacterioplankton. *Environ Microbiol* 1–20.
Embree M, Liu JK, Al-Bassam MM, Zengler K. (2015). Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proc Natl Acad Sci* **112**:15450–15455.
Eng J, McCormack A, Yates J. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* **5**:976–989.
Falkowski PG, Fenchel T, Delong EF. (2008). The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**:1034–9.
Fall R, Benson AA. (1996). Leaf methanol - the simplest natural product from plants. *Trends plant Sci* **1**:296–301.
Feike J, Jürgens K, Hollibaugh JT, Krüger S, Jost G, Labrenz M. (2012). Measuring unbiased metatranscriptomics in suboxic waters of the central Baltic Sea using a new in situ fixation system. *ISME J* **6**:461–70.
Fernández-Gómez B, Richter M, Schüler M, Pinhassi J, Acinas SG, González JM, *et al.* (2013). Ecology of marine Bacteroidetes: a comparative genomics approach. *ISME J* **7**:1026–37.
Fetzner S. (2012). Ring-cleaving dioxygenases with a cupin fold. *Appl Environ Microbiol* **78**:2505–2514.

Fichot CG, Benner R. (2014). The fate of terrigenous dissolved organic carbon in a river-influenced ocean margin. *Global Biogeochem Cycles* 300–318.
Folio P, Chavant P, Chafsey I, Belkorchia A, Chambon C, Hébraud M. (2004). Two-dimensional electrophoresis database of Listeria monocytogenes EGDe proteome and proteomic analysis of mid-log and stationary growth phase cells. *Proteomics* **4**:3187–8201.
Frazee RW, Livingston DM, LaPorte DC, Lipscomb JD. (1993). Cloning, sequencing, and expression of the Pseudomonas putida protocatechuate 3,4-dioxygenase genes. *J Bacteriol* **175**:6194–6202.
Frey KE, McClelland JW. (2009). Impacts of permafrost degradation on arctic river biochemistry. *Hydrol Process* **23**:169–182.
Fu Y, Keats KF, Rivkin RB, Lang AS. (2013). Water mass and depth determine the distribution and diversity of Rhodobacterales in an Arctic marine system. *FEMS Microbiol Ecol* **84**:564–576.
Fuchs BM, Woebken D, Zubkov M, Burkill P, Amann R, Woebken M V, *et al.* (2005). Molecular identification of picoplankton populations in contrasting waters of the Arabian Sea. *Aquat Microb Ecol* **39**:135–157.
Fuchs G, Boll M, Heider J. (2011). Microbial degradation of aromatic compounds — from one strategy to four. *Nat Rev Microbiol* **9**:803–816.
Fuhrman JA, Cram JA, Needham DM. (2015). Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol* **13**:133–146.
Fuhrman JA, McCallum K, Davis AA. (1992). Novel major achaebacterial group from marine plankton. *Nature* **356**:148–149.
Galand P, Lovejoy C, Vincent W. (2006). Remarkably diverse and contrasting archaeal communities in a large arctic river and the coastal Arctic Ocean. *Aquat Microb Ecol* **44**:115–126.
Galand PE, Lovejoy C, Hamilton AK, Ingram RG, Pedneault E, Carmack EC. (2009). Archaeal diversity and a gene for ammonia oxidation are coupled to oceanic circulation. *Environ Microbiol* **11**:971–980.
Galand PE, Pereira O, Hochart C, Auguet JC, Debroas D. (2018). A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *ISME J* **12**:2470–2478.
Galand PE, Potvin M, Casamayor EO, Lovejoy C. (2010). Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *ISME J* **4**:564–76.

Ganesh S, Parris DJ, DeLong EF, Stewart FJ. (2014). Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J* **8**:187–211.

Garcia SL, Buck M, McMahon KD, Grossart HP, Eiler A, Warnecke F. (2015). Auxotrophy and intrapopulation complementary in the interactome of a cultivated freshwater model community. *Mol Ecol* **24**:4449–4459.

Georges AA, El-Swais H, Craig SE, Li WK, Walsh DA. (2014). Metaproteomic analysis of a winter to spring succession in coastal northwest Atlantic Ocean microbial plankton. *ISME J* 1–13.

Gilbert D, Sundby B, Gobeil C, Mucci A, Tremblay G-H. (2005). A seventy-two-year record of diminishing deep-water oxygen in the St. Lawrence estuary: The northwest Atlantic connection. *Limnol Oceanogr* **50**:1654–1666.

Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, *et al.* (2012). Defining seasonal marine microbial community dynamics. *ISME J* **6**:298–308.

Gilbert J a, Dupont CL. (2011). Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci* **3**:347–71.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P, *et al.* (2010). Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *Microbiol Today* **37**:82–85.

Giovannoni S, Stingl U. (2007). The importance of culturing bacterioplankton in the 'omics' age. *Nat Rev Microbiol* **5**:820–826.

Giovannoni SJ, Bibbs L, Cho J-C, Stapels MD, Desiderio R, Vergin KL, *et al.* (2005). Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature* **438**:82–5.

Giovannoni SJ, Britschgi TB, Moyer CL, Fielf KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**:60–63.

Giovannoni SJ, Hayakawa DH, Tripp HJ, Stingl U, Givan S a, Cho J-C, *et al.* (2008). The small genome of an abundant coastal ocean methylotroph. *Environ Microbiol* **10**:1771–82.

Giovannoni SJ, Rappe MS, Vergin KL, Adair NL. (1996). 16S rRNA genes reveal stratified open ocean bacterioplankton populations related to the Green Non-Sulfur bacteria. *Proc Natl Acad Sci* **93**:7979–7984.

Giovannoni SJ, Stingl U. (2005). Molecular diversity and ecology of microbial plankton. *Nature* **437**:343–348.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, *et al.* (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**:1242–1245.

Glen D, Silvio W, Samay P, Katrin B, Christoph K, Christian K. (2014). Less Is More: Selective Advantages Can Explain the Prevalent Loss of Biosynthetic Genes in Bacteria. *Evolution (N Y)* **68**:2559–2570.

Gobeil C. (2006). Biogeochemistry and chemical contamination in the St. Lawrence estuary. *Estuaries* **5**:121–147.

Gómez-Pereira PR, Fuchs BM, Alonso C, Oliver MJ, van Beusekom JEE, Amann R. (2010). Distinct flavobacterial communities in contrasting water masses of the north Atlantic Ocean. *ISME J* **4**:472–87.

Goñi MA, Suga H, Perkeya DW. (2003). Sources and distribution of organic matter in a river-dominated estuary (Winyah Bay, SC, USA). *Estuar Coast Shelf Sci* **57**:1023–1048.

Gordeev V V. (2006). Fluvial sediment flux to the Arctic Ocean. *Geomo* **80**:94–104.

Grzymski J, Riesenfeld C, Williams T. (2012). A metagenomic assessment of winter and summer bacterioplankton from Antarctica Peninsula coastal surface waters. *ISME J* **6**:1901–1915.

Guéguen C, McLaughlin FA, Carmack EC, Itoh M, Narita H, Nishino S. (2012). The nature of colored dissolved organic matter in the southern Canada Basin and East Siberian Sea. *Deep Res Part II Top Stud Oceanogr* **81**–**84**:102–113.

Guerrero-Feijóo E, Nieto-Cid M, Sintes E, Dobal-Amador V, Hernando-Morales V, Álvarez M, *et al.* (2016). Optical properties of dissolved organic matter relate to different depth-specific patterns of archaeal and bacterial community structure in the north Atlantic ocean. *FEMS Microbiol Ecol* 1–14.

Gurdeep Singh R, Tanca A, Palomba A, Van Der Jeugt F, Verschaffelt P, Uzzau S, *et al.* (2019). Unipept 4.0: Functional Analysis of Metaproteome Data. *J Proteome Res* **18**:606–615.

Hallam SJ, Mincer TJ, Schleper C, Preston CM, Roberts K, Richardson PM, *et al.* (2006). Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine Crenarchaeota. *PLoS Biol* **4**:520–536.

Hamilton AK, Lovejoy C, Galand PE, Ingram RG. (2008). Water masses and biogeography of picoeukaryote assemblages in a cold hydrographically complex system. *Limnol Oceanogr* **53**:922–935.

Han D, Kang I, Ha HK, Kim HC, Kim OS, Lee BY, *et al.* (2014). Bacterial communities of surface mixed layer in the pacific sector of the western Arctic Ocean during sea-ice melting. *PLoS One* **9**. doi:10.1371/journal.pone.0086887.

Hansell DA. (2013). Recalcitrant Dissolved Organic Carbon Fractions. *Ann Rev Mar Sci* **5**:421–445.

Hansell DA, Carlson CA, Repeta DJ, Schlitzer R. (2009). Dissolved Organic Matter in the Ocean A Controversy Stimulates New Insights. *Oceanography* **22**:202–211.

Hansell DA, Kadko D, Bates NR. (2004). Degradation of Terrigenous Dissolved Organic Carbon in the Western Arcitc Ocean. *Science* **304**:858–862.

Hawley AK, Brewer HM, Norbeck AD, Pasa-Toli L, Hallam SJ. (2014). Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proc Natl Acad Sci* **111**. doi:10.1073/pnas.1322132111.

Hawley AK, Kheirandish S, Mueller A, Leung HTC, Norbeck AD, Brewer HM, *et al.* (2013). Molecular tools for investigating microbial community structure and function in oxygen-deficient marine waters. 1st ed. Elsevier Inc. doi:10.1016/B978-0-12-407863-5.00016-2.

Hedges JI, Eglinton G, Hatcher PG, Kirchman DL, Arnosti C, Derenne S, *et al.* (2000). The molecularly-uncharacterized component of nonliving organic matter in natural environments. *Org Geochem* **31**:945–958.

Hedges JI, Mann DC. (1979). The lignin geochemistry of marine sediments from the southern Washington coast. *Geochim Cosmochim Acta* **43**:1809–1818.

Hedlund BP, Dodsworth J a, Murugapiran SK, Rinke C, Woyke T. (2014). Impact of single-cell genomics and metagenomics on the emerging view of extremophile ‘microbial dark matter’. *Extremophiles* 865–875.

Helling RB, Vargas CN, Adams J. (1987). Evolution of Escherichia coli during growth in a constant environment. *Genetics* **116**:349–358.

Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J* **5**:1571–1579.

Hernes PJ, Benner R. (2003). Photochemical and microbial degradation of dissolved lignin phenols: Implications for the fate of terrigenous dissolved organic matter in marine environments. *J Geophys Res* **108**:1–9.

Hettich RL, Pan C, Chourey K, Giannone RJ. (2013). Metaproteomics: harnessing the power of high performance mass spectrometry to identify the suite of proteins that control metabolic activities in microbial communities. *Anal Chem* **85**:4203–14.
Hewson I, Poretsky RS, Tripp HJ, Montoya JP, Zehr JP. (2010). Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ Microbiol* **12**:1940–56.
Heyer R, Schallert K, Zoun R, Becher B, Saake G, Benndorf D. (2017). Challenges and perspectives of metaproteomic data analysis. *J Biotechnol* **261**:24–36.
Holmes RM, McClelland JW, Raymond PA, Frazer BB, Peterson BJ, Stieglitz M. (2008). Lability of DOC transported by Alaskan rivers to the Arctic Ocean. *Geophys Res Lett* **35**:3–7.
Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. (2010). Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomic studies. *J. Proteome Res* **9**:4152-4160.
Houser JR, Barnhart C, Boutz DR, Carroll SM, Dasgupta A, Michener JK, *et al.* (2015). Controlled Measurement and Comparative Analysis of Cellular Components in E. coli Reveals Broad Regulatory Changes in Response to Glucose Starvation. *PLoS Comput Biol* **11**:1–27.
Hug LA, Thomas BC, Sharon I, Brown CT, Sharma R, Hettich RL, *et al.* (2016). Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ Microbiol* **18**:159–173.
Huntemann M, Ivanova NN, Mavromatis K, Tripp HJ, Paez-Espino D, Tennessen K, *et al.* (2016). The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4). *Stand Genomic Sci* **11**:17.
Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**:377–86.
Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res* **21**:1552–60.
Hutchins RHS, Aukes P, Schiff SL, Dittmar T, Prairie YT, Giorgio PA. (2017). The Optical , Chemical , and Molecular Dissolved Organic Matter Succession Along a Boreal Soil-Stream-River Continuum. *J Geophys Res G Biogeosciences* **122**:2892–2908.
Ivars-Martinez E, Martin-Cuadrado A-B, D'Auria G, Mira A, Ferriera S, Johnson J, *et al.* (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas

macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME J* **2**:1194–1212.
Jiao N, Herndl GJ, Hansell DA, Benner R, Kattner G, Wilhelm SW, *et al.* (2010). Microbial production of recalcitrant dissolved organic matter: Long-term carbon storage in the global ocean. *Nat Rev Microbiol* **8**:593–599.
Joli N, Gosselin M, Ardyna M, Babin M, Onda DF, Tremblay JÉ, *et al.* (2018). Need for focus on microbial species following ice melt and changing freshwater regimes in a Janus Arctic Gateway. *Sci Rep* **8**:1–11.
Jørgensen L, Stedmon CA, Kragh T, Markager S, Middelboe M, Søndergaard M. (2011). Global trends in the fluorescence characteristics and distribution of marine dissolved organic matter. *Mar Chem* **126**:139–148.
Jormakka M, Richardson D, Byrne B, Iwata S. (2004). Architecture of NarGH Reveals a Structural Classification of Mo-bisMGD Enzymes. *Structure* **12**:95–104.
Kabisch A, Otto A, König S, Becher D, Albrecht D, Schüler M, *et al.* (2014). Functional characterization of polysaccharide utilization loci in the marine Bacteroidetes 'Gramella forsetii' KT0803. *ISME J* **8**:1492–1502.
Kaiser K, Benner R, Amon RMW. (2017). The fate of terrigenous dissolved organic carbon on the Eurasian shelves and export to the North Atlantic. *J Geophys Res Ocean* **122**:4–22.
Käll L, Canterbury J, Weston J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**:923–925.
Kan J, Hanson TE, Ginter JM, Wang K, Chen F. (2005). Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems* **1**:7.
Kang DD, Froula J, Egan R, Wang Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**:e1165.
Karner M, DeLong E, Karl D. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* 507–510.
Karp PD, Latendresse M, Caspi R. (2011). The pathway tools pathway prediction algorithm. *Stand Genomic Sci* **5**:424–9.
Kasai D, Masai E, Miyauchi K, Katayama Y, Fukuda M. (2005). Characterization of the gallate dioxygenase gene: Three distinct ring cleavage dioxygenases are involved in syringate degradation by Sphingomonas paucimobilis SYK-6. *J Bacteriol* **187**:5067–5074.

Keltjens JT, Pol A, Reimann J, Op den Camp HJM. (2014). PQQ-dependent methanol dehydrogenases: Rare-earth elements make a difference. *Appl Microbiol Biotechnol* **98**:6163–6183.

Kemp PF, Lee S, Laroche J. (1993). Estimating the growth rate of slowly growing marine bacteria from RNA content. *Appl Environ Microbiol* **59**:2594–601.

Kielak AM, Barreto CC, Kowalchuk GA, van Veen JA, Kuramae EE. (2016). The ecology of Acidobacteria: Moving beyond genes and genomes. *Front Microbiol* **7**:1–16.

Kirchman DL. (2002). The ecology of Cytophaga-Flavobacteria in aquatic environments. *FEMS Microbiol Ecol* **39**:91–100.

Kirchman DL, Cottrell MT, Lovejoy C. (2010). The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol* **12**:1132–1143.

Kirchman DL, Elifantz H, Dittel AI, Malmstrom RR, Cottrell MT. (2007). Standing stocks and activity of Archaea and Bacteria in the western Arctic Ocean. *Limnol Oceanogr* **52**:495–507.

Kitidis V, Stubbins AP, Uher G, Upstill Goddard RC, Law CS, Woodward EMS. (2006). Variability of chromophoric organic matter in surface waters of the Atlantic Ocean. *Deep Res Part II Top Stud Oceanogr* **53**:1666–1684.

Klein B, LeBlanc B, Mei ZP, Beret R, Michaud J, Mundy CJ, *et al.* (2002). Phytoplankton biomass, production and potential export in the North Water. *Deep Res Part II Top Stud Oceanogr* **49**:4983–5002.

Koch BP, Witt M, Engbrodt R, Dittmar T, Kattner G. (2005). Molecular formulae of marine and terrigenous dissolved organic matter detected by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Geochim Cosmochim Acta* **69**:3299–3308.

Kozlowski JA, Stieglmeier M, Schleper C, Klotz MG, Stein LY. (2016). Pathways and key intermediates required for obligate aerobic ammonia-dependent chemolithotrophy in bacteria and Thaumarchaeota. *ISME J* **10**:1–10.

Krause SMB, Johnson T, Samadhi Karunaratne Y, Fu Y, Beck DAC, Chistoserdova L, *et al.* (2016). Lanthanide-dependent cross-feeding of methane-derived carbon is linked by microbial community interactions. *Proc Natl Acad Sci* **114**:358–363.

Kujawinski EB. (2011). The impact of microbial metabolism on marine dissolved organic matter. *Ann Rev Mar Sci* **3**:567–599.

Kwan JC, Donia MS, Han AW, Hirose E, Haygood MG, Schmidt EW. (2012). Genome streamlining and chemical defense in a coral reef symbiosis. *Proc Natl Acad Sci* **109**:20655–20660.

Labrenz M, Jost G, Jürgens K. (2007). Distribution of abundant prokaryotic organisms in the water column of the central Baltic Sea with an oxic-anoxic interface. *Aquat Microb Ecol* **46**:177–190.

Landry ZC, Swan BK, Herndl GJ, Stepanauskas R, Giovannoni SJ. (2017). SAR202 Genomes from the Dark Ocean Predict Pathways for the Oxidation of Recalcitrant Dissolved Organic Matter. *MBio* **8**:1–19.

Lane DJ, Pace B, Olsen GJ, Stahlt DA, Sogint ML, Pace NR. (1986). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A* **83**:4972.

Lawrence J, Popova E, Yool A, Srokosz M. (2015). On the vertical phytoplankton response to an ice-free Arctic Ocean. *J Geophys Res Ocean* **120**:8571–8582.

Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF, Prohaska S. (2011). Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**. doi:10.1186/1471-2105-12-124.

Lehmann MF, Barnett B, Gélinas Y, Gilbert D, Maranger RJ, Mucci A, *et al.* (2009). Aerobic respiration and hypoxia in the Lower St. Lawrence Estuary: Stable isotope ratios of dissolved oxygen constrain oxygen sink partitioning. *Limnol Oceanogr* **54**:2157–2169.

Letunic I, Bork P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**:W242–W245.

Levy R, Carr R, Kreimer A, Freilich S, Borenstein E. (2015). NetCooperate: A network-based tool for inferring host-microbe and microbe-microbe cooperation. *BMC Bioinformatics* **16**:1–6.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**:1674–1676.

Li WKW, Carmack EC, McLaughlin F a., Nelson RJ, Williams WJ. (2013). Space-for-time substitution in predicting the state of picoplankton and nanoplankton in a changing Arctic Ocean. *J Geophys Res Ocean* **118**:5750–5759.

Li WKW, Mclaughlin FA, Lovejoy C, Carmack EC. (2009). Smallest Algae Thrive As the Arctic Ocean Freshens. *Science* **326**:2009.

Lindh M V., Lefébure R, Degerman R, Lundin D, Andersson A, Pinhassi J. (2015). Consequences of increased terrestrial dissolved organic matter and temperature on bacterioplankton community composition during a Baltic Sea mesocosm experiment. *Ambio* **44**:402–412.

Louca S, Parfrey LW, Doebeli M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* **353**:1272–1277.

Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, *et al.* (2018). Function and functional redundancy in microbial systems. *Nat Ecol Evol* **2**:936–943.

Lovejoy C, Galand PE, Kirchman DL. (2011). Picoplankton diversity in the Arctic Ocean and surrounding seas. *Mar Biodivers* **41**:5–12.

Lovejoy C, Legendre L, Martineau MJ, Bacle J, von Quillfeldt CH. (2002). Distribution of phytoplankton and other protists in the North Water. *Deep Res Part II-Topical Stud Oceanogr* **49**:5027–5047.

Lu C-J, Benner R, Fichot CG, Fukuda H, Yamashita Y, Ogawa H. (2016). Sources and Transformations of Dissolved Lignin Phenols and Chromophoric Dissolved Organic Matter in Otsuchi Bay, Japan. *Front Mar Sci* **3**:1–12.

Lu X, Zhu H. (2005). Tube-gel digestion: a novel proteomic approach for high throughput analysis of membrane proteins. *Mol Cell Proteomics* **4**:1948–1958.

Lücker S, Nowka B, Rattei T, Spieck E, Daims H. (2013). The Genome of Nitrospina gracilis Illuminates the Metabolism and Evolution of the Major Marine Nitrite Oxidizer. *Front Microbiol* **4**:27.

Lücker S, Wagner M, Maixner F, Pelletier E, Koch H, Vacherie B, *et al.* (2010). A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci U S A* **107**:13479–84.

Madsen EL. (2011). Microorganisms and their roles in fundamental biogeochemical cycles. *Curr Opin Biotechnol* **22**:456–64.

Malmstrom RR, Cottrell MT, Elifantz H, Kirchman DL, Kiene RP, Yu L, *et al.* (2005). Biomass production and assimilation of dissolved organic matter by SAR11 bacteria in the Northwest Atlantic Ocean Contribution of SAR11 bacteria to dissolved dimethylsulfoniopropionate and amino acid uptake in the North Atlantic ocean Diversity and abunda. *Appl Env Microbiol* **71**:2979–86.

Mann PJ, Davydova A, Zimov N, Spencer RGM, Davydov S, Bulygina E, *et al.* (2012). Controls on the composition and lability of dissolved organic matter in Siberia's Kolyma River basin. *J Geophys Res Biogeosciences* **117**:1–15.
Marchese C, Albouy C, Tremblay JÉ, Dumont D, D'Ortenzio F, Vissault S, *et al.* (2017). Changes in phytoplankton bloom phenology over the North Water (NOW) polynya: a response to changing environmental conditions. *Polar Biol* **40**. doi:10.1007/s00300-017-2095-2.
Marshall KT, Morris RM. (2013). Isolation of an aerobic sulfur oxidizer from the SUP05/Arctic96BD-19 clade. *ISME J* **7**:452–455.
Martin J, Tremblay JÉ, Gagnon J, Tremblay G, Lapoussière A, Jose C, *et al.* (2010). Prevalence, structure and properties of subsurface chlorophyll maxima in Canadian Arctic waters. *Mar Ecol Prog Ser* **412**:69–84.
Martiny AC. (2019). High proportions of bacteria are culturable across major biomes. *ISME J* 3–6.
Massana R, Murray AE, Preston CM. (1997). Vertical Distribution and Phylogenetic Characterization of Marine Planktonic. *Microbiology* **63**:50–56.
Mattes TE, Nunn BL, Marshall KT, Proskurowski G, Kelley DS, Kawka OE, *et al.* (2013). Sulfur oxidizers dominate carbon fixation at a biogeochemical hot spot in the dark ocean. *ISME J* 1–12.
McCarren J, Becker JW, Repeta DJ, Shi Y, Young CR, Malmstrom RR, *et al.* (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc Natl Acad Sci U S A* **107**:16420–16427.
McCarthy M, Pratum T, Hedges J, Benner R. (1997). Chemical composition of dissolved organic nitrogen in the ocean. *Nature* **390**:150–154.
McClelland JW, Townsend-Small A, Holmes RM, Pan F, Stieglitz M, Khosh M, *et al.* (2014). River export of nutrients and organic matter from the North Slope of Alaska to the Beaufort Sea. *Water Resour Res* **50**:1823–1839.
McLean TI. (2013). 'Eco-omics': a review of the application of genomics, transcriptomics, and proteomics for the study of the ecology of harmful algae. *Microb Ecol* **65**:901–15.
Mee MT, Collins JJ, Church GM, Wang HH. (2014). Syntrophic exchange in synthetic microbial communities. *Proc Natl Acad Sci* **111**:E2149–E2156.
Mesuere B, Devreese B, Debyser G, Aerts M, Vandamme P, Dawyndt P. (2012). Unipept:

Tryptic peptide-based biodiversity analysis of metaproteome samples. *J Proteome Res* **11**:5773–5780.
Mesuere B, Willems T, Van Der Jeugt F, Devreese B, Vandamme P, Dawyndt P. (2016). Unipept web services for metaproteomics analysis. *Bioinformatics* **32**:1746–1748.
Milici M, Deng ZL, Tomasch J, Decelle J, Wos-Oxley ML, Wang H, *et al.* (2016). Co-occurrence analysis of microbial taxa in the Atlantic ocean reveals high connectivity in the free-living bacterioplankton. *Front Microbiol* **7**:1–20.
Millet DB, Jacob DJ, Custer TG, de Gouw JA, Goldstein AH, Karl T, *et al.* (2008). New constraints on terrestrial and oceanic sources of atmospheric methanol. *Atmos Chem Phys* **8**:6887–6905.
Mincer TJ, Aicher AC. (2016). Methanol production by a broad phylogenetic array of marine phytoplankton. *PLoS One* **11**:1–17.
Molenaar D, Van Berlo R, De Ridder D, Teusink B. (2009). Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol* **5**:1–10.
Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan L-K, *et al.* (2012). Sizing up metatranscriptomics. *ISME J* **7**:237–243.
Morgan-Lang C, Konwar KM, McLaughlin R, Zhang G, Armstrong Z, Song YC, *et al.* TreeSAPP: Tree-based Sensitive and Accurate Protein Profiler.
Morris JJ, Lenski RE, Zinser ER. (2012). The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *MBio* **3**:1–9.
Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J* **4**:673–85.
Morris RM, Rappé MS, Connon S a, Vergin KL, Siebold W a, Carlson C a, *et al.* (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**:806–810.
Morris RM, Rappé MS, Urbach E, Connon S a, Rappe MS. (2004). Prevalence of the Chloroflexi -Related SAR202 Bacterioplankton Cluster throughout the Mesopelagic Zone and Deep Ocean. *Appl Environ Microbiol* **70**:2836–2842.
Mucci A, Starr M, Gilbert D, Sundby B. (2011). Acidification of Lower St. Lawrence Estuary Bottom Waters. *Atmosphere-Ocean* **49**:206–218.
Mulder T, Alexander J. (2001). The physical character of subaqueous sedimentary density flows

and their deposits. *Photodermatol Photoimmunol Photomed* **48**:269–299.
Müller O, Seuthe L, Bratbak G, Paulsen ML. (2018). Bacterial Response to Permafrost Derived Organic Matter Input in an Arctic Fjord. *Front Mar Sci* **5**:1–12.
Müllner D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *J Stat Softw* **9**:1–18.
Murphy KR, Stedmon CA, Waite TD, Ruiz GM. (2008). Distinguishing between terrestrial and autochthonous organic matter sources in marine environments using fluorescence spectroscopy. *Mar Chem* **108**:40–58.
Murrell JC, McGowan V, Cardy DLN. (1992). Detection of methylotrophic bacteria in natrual samples by molecular probing techniques. *Chemosphere* **26**:1–11.
Muthusamy S, Lundin D, Mamede Branca RM, Baltar F, González JM, Lehtiö J, *et al.* (2017). Comparative proteomics reveals signature metabolisms of exponentially growing and stationary phase marine bacteria. *Environ Microbiol* **19**:2301–2319.
Nelson NB, Carlson CA, Steinberg DK. (2004). Production of chromophoric dissolved organic matter by Sargasso Sea microbes. *Mar Chem* **89**:273–287.
Nelson NB, Siegel DA. (2013). The Global Distribution and Dynamics of Chromophoric Dissolved Organic Matter. *Ann Rev Mar Sci* **5**:447–476.
Nelson NB, Siegel DA, Carlson CA, Swan C, Smethie WM, Khatiwala S. (2007). Hydrography of chromophoric dissolved organic matter in the North Atlantic. *Deep Res Part I Oceanogr Res Pap* **54**:710–731.
Nelson NB, Siegel DA, Carlson CA, Swan CM. (2010). Tracing global biogeochemical cycles and meridional overturning circulation using chromophoric dissolved organic matter. *Geophys Res Lett* **37**:1–5.
Nelson NB, Siegel DA, Michaels AF. (1998). Seasonal dynamics of colored dissolved material in the Sargasso Sea. *Deep Sea Res I* **45**:931–957.
Neugebauer H, Herrmann C, Kammer W, Schwarz G, Nordheim A, Braun V. (2005). ExbBD-Dependent Transport of Maltodextrins through the Novel MalA Protein across the Outer Membrane of Caulobacter crescentus. *J Bacteriol* **187**:8300–8311.
Newton RJ, Griffin LE, Bowles KM, Meile C, Gifford S, Givens CE, *et al.* (2010). Genome characteristics of a generalist marine bacterial lineage. *ISME J* **4**:784–98.
Noda Y, Nishikawa S, Shiozuka KI, Kadokura H, Nakajima H, Yoda K, *et al.* (1990). Molecular

cloning of the protocatechuate 4,5-dioxygenase genes of Pseudomonas paucimobilis. *J Bacteriol* **172**:2704–2709.

Ogawa H, Tanoue E. (2003). Dissolved Organic Matter in Oceanic Waters. *J Oce* **59**:129–147.

Okazaki Y, Hodoki Y, Nakano SI. (2013). Seasonal dominance of CL500-11 bacterioplankton (phylum Chloroflexi) in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiol Ecol* **83**:82–92.

Okkonen SR, Ashjian CJ, Campbell RG, Maslowski W, Clement-Kinney JL, Potter R. (2009). Intrusion of warm Bering/Chukchi waters onto the shelf in the western Beaufort Sea. *J Geophys Res Ocean* **114**:1–23.

Oksanen J, Blanchet GF, Friendly M, Kindt R, Legendre P, McGlinn D, *et al.* (2019). VEGAN: community ecology package.

Opsahl S, Benner R. (1997). Distribution and cycling of terrigenous dissolved organic matter in the ocean. *Nature* **386**:480–482.

Opsahl S, Benner R. (1998). Photochemical reactivity of dissolved lignin in river and ocean waters. *Limnol Oceanogr* **43**:1297–1304.

Opsahl S, Benner R, Amon RMW, Dec N. (1999). Major flux of terrigenous dissolved organic matter through the Arctic Ocean. *Limnol Oceanogr* **44**:2017–2023.

Orellana M V., Hansell D a. (2012). Ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO): A long lived protein in the deep ocean. *Limnol Oceanogr* **57**:826–834.

Orsi WD, Smith JM, Liu S, Liu Z, Sakamoto CM, Wilken S, *et al.* (2016). Diverse, uncultivated bacteria and archaea underlying the cycling of dissolved protein in the ocean. *ISME J* **10**:2158–2173.

Orsi WD, Smith JM, Wilcox HM, Swalwell JE, Carini P, Worden AZ, *et al.* (2015). Ecophysiology of uncultivated marine euryarchaea is linked to particulate organic matter. *ISME J* 1–17.

Ortega-Retuerta E, Siegel DA, Nelson NB, Duarte CM, Reche I. (2010). Observations of chromophoric dissolved and detrital organic matter distribution using remote sensing in the Southern Ocean: Validation, dynamics and regulation. *J Mar Syst* **82**:295–303.

Ottesen E a, Marin R, Preston CM, Young CR, Ryan JP, Scholin C a, *et al.* (2011). Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J* **5**:1881–95.

Pande S, Merker H, Bohl K, Reichelt M, Schuster S, De Figueiredo LF, *et al.* (2014). Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *ISME J* **8**:953–962.
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**:1043–55.
Payne TMB, Rouatt JW, Lochhead AG. (1957). The Relationship Between Soil Bacteria With Simple Nutritional Requirements and Those Requiring Amino Acids. *Can J Microbiol* **3**:73–80.
Pedrós-Alió C, Potvin M, Lovejoy C. (2015). Diversity of planktonic microorganisms in the Arctic Ocean. *Prog Oceanogr* **139**:233–243.
Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, *et al.* (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**:1–16.
Picotti P, Aebersold R, Domon B. (2007). The implications of proteolytic background for shotgun proteomics. *Mol. Cell Proteomics* **6**:1589-1598.
Pomeroy LR, Williams PJ leB., Azam F, Hobbie JE. (1998). The microbial loop in the planktonic communities in lakes with various trophic status. *Oceanog* **20**:28-33.
Pommier T, Canbäck B, Riemann L, Boström KH, Simu K, Lundberg P, *et al.* (2007). Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16**:867–880.
Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP, Moran MA. (2009). Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol* **11**:1358–75.
Poretsky RS, Sun S, Mou X, Moran MA. (2010). Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* **12**:616–627.
Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res* **41**:590–596.
Rappé MS, Connon S a, Vergin KL, Giovannoni SJ. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**:630–633.
Raymond PA, Mcclelland JW, Holmes RM, Zhulidov A V, Mull K, Peterson BJ, *et al.* (2007). Flux and age of dissolved organic carbon exported to the Arctic Ocean : A carbon isotopic study of the five largest arctic rivers. *Global Biochem Cycles* **21**:1–9.

Reemtsma T, These A, Leenheer J, Spitzy A. (2008). Molecular and Structural Characterization of Dissolved Organic Matter from the Deep Ocean by FTICR-MS , Including Hydrophilic Nitrogenous Organic Molecules. *Environ Sci Technol* **42**:1430–1437.

Rho M, Tang H, Ye Y. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Res* **38**:1–12.

Riffle M, May D, Timmins-Schiffman E, Mikan M, Jaschob D, Noble W, *et al.* (2017). MetaGOmics: A Web-Based Tool for Peptide-Centric Functional and Taxonomic Analysis of Metaproteomics Data. *Proteomes* **6**:1–17.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**:431–7.

Rosenthal AZ, Matson EG, Eldar A, Leadbetter JR. (2011). RNA-seq reveals cooperative metabolic interactions between two termite-gut spirochete species in co-culture. *ISME J* **5**:1133–1142.

Rudels B, Larsson A, Sehlstedt P. (1991). Stratification and water mass formation in the Arctic Ocean: some implications for the nutrient distribution. *Polar Res* 19–31.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, *et al.* (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**:e77.

Saito M a., McIlvin MR, Moran DM, Goepfert TJ, DiTullio GR, Post a. F, *et al.* (2014). Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* **345**:1173–1177.

Saito MA, Bulygin V V, Moran DM, Taylor C, Scholin C. (2011). Examination of microbial proteome preservation techniques applicable to autonomous environmental sample collection. *Front Microbiol* **2**:215.

Saito MA, Dorsk A, Post AF, Mcilvin M, Rappé MS, DiTullio G, *et al.* (2015). Needles in the Blue Sea: Sub-Species Specificity in Targeted Protein Biomarker Analyses Within the Vast Oceanic Microbial Metaproteome. *Proteomics* **15**:3521–3531.

Sala MM, Terrado R, Lovejoy C, Unrein F, Pedrós-Alió C. (2008). Metabolic diversity of heterotrophic bacterioplankton over winter and spring in the coastal Arctic Ocean. *Environ Microbiol* **10**:942–9.

Salazar G, Cornejo-Castillo FM, Borrull E, D??ez-Vives C, Lara E, Vaqu?? D, *et al.* (2015).

Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes. *Mol Ecol* **24**:5692–5706.
Santoro AE, Dupont CL, Richter RA, Craig MT, Carini P, McIlvin MR, *et al.* (2015). Genomic and proteomic characterization of " Candidatus Nitrosopelagicus brevis": An ammonia-oxidizing archaeon from the open ocean. *Proc Natl Acad Sci* **112**:1173–1178.
Schaechter M, Maaloe O, Kjeldgaard N. (1958). Dependency on medium and temperature of cell size and chemical composition during balanced growth of Salmonella typhimurium. *Microbiology* **19**:592–606.
Schattenhofer M, Fuchs BM, Amann R, Zubkov M V., Tarran GA, Pernthaler J. (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environ Microbiol* **11**:2078–2093.
Schauer K, Rodionov D a., de Reuse H. (2008). New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci* **33**:330–338.
Schlitzer R. (2016). Ocean data view.
Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, *et al.* (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**:7537–7541.
Schmitt S, Deines P, Behnam F, Wagner M, Taylor MW. (2011). Chloroflexi bacteria are more diverse, abundant, and similar in high than in low microbial abundance sponges. *FEMS Microbiol Ecol* **78**:497–510.
Schneider T, Riedel K. (2010). Environmental proteomics: analysis of structure and function of microbial communities. *Proteomics* **10**:785–98.
Seidel M, Manecki M, Herlemann DPR, Deutsch B, Schulz-Bull D, Jürgens K, *et al.* (2017). Composition and Transformation of Dissolved Organic Matter in the Baltic Sea. *Front Earth Sci* **5**:1–20.
Sharma R, Dill BD, Chourey K, Shah M, VerBerkmoes NC, Hettich RL. (2012). Coupling a detergent lysis/cleanup methodology with intact protein fractionation for enhanced proteome characterization. *J Proteome Res* **11**:6008–18.
Sheik CS, Jain S, Dick GJ. (2013). Metabolic flexibility of enigmatic SAR324 revealed through metagenomics and metatranscriptomics. *Environ Microbiol* **16**:304–17.
Shevchenko A, Tomas JH, Olsen J, Mann M. (2007). In-gel digestion for mass spectrometric

characterization of proteins and proteomes. *Nat Protoc* **1**:2856–2860.
Shteynberg A, Mendoza L, Hoopmann MR, Sun Z, Schmidt F, Deutsch EW. (2015). reSpect: Software for identification of high and low abundance ion species in chimeric tandem mass spectra. *JASMS* **26**:1837-1847**.**
Simon M, Azam F. (1989). Protein content and protein synthesis rates of planktonic marine bacteria . *Mar Ecol Prog Ser* **51**:201–213.
Sipler RE, Kellogg CTE, Connelly TL, Roberts QN, Yager PL, Bronk DA. (2017). Microbial community response to terrestrially derived dissolved organic matter in the coastal Arctic. *Front Microbiol* **8**:1–19.
Smith D, Azam F. (1992). A simple, economical method for measuring bacterial protein synthesis rates in seawater using 3H-leucine. *Mar Microb Food Webs* **6**:107–114.
Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al.* (2011). Microbial Diversity in the Deep Sea and the Underexplored 'Rare Biosphere'. *Handb Mol Microb Ecol II Metagenomics Differ Habitats* 243–252.
Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al.* (2006). Microbial Diversity in the Deep Sea and the Underexplored 'Rare Biosphere'. *Proc Natl Acad Sci* **103**:12115–12120.
Sorokin DY, Lücker S, Vejmelkova D, Kostrikina NA, Kleerebezem R, Rijpstra WIC, *et al.* (2012). Nitrification expanded: discovery, physiology and genomics of a nitrite-oxidizing bacterium from the phylum Chloroflexi. *ISME J* **6**:2245–2256.
Sowell SM, Abraham P, Shah M, Verberkmoes NC, Smith DP, Barofsky DF, *et al.* (2011). Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *ISME J* **5**:856–865.
Sowell SM, Norbeck AD, Lipton MS, Nicora CD, Callister SJ, Smith RD, *et al.* (2008). Proteomic analysis of stationary phase in the marine bacterium 'Candidatus pelagibacter ubique'. *Appl Environ Microbiol* **74**:4091–4100.
Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, *et al.* (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J* **3**:93–105.
Spieck E, Bock E. (2005). The lithoautotrophic nitrite-oxidizing bacteria. In:*Bergey's Manual® of Systematic Bacteriology*, Brenner, D., Krieg, NR, Staley, JT, & Garrity, GM (eds), Springer:

New York, NY, pp. 149–153.
Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fr.pdf. *J Bacteriol* **178**:591–599.
Stepanauskas R. (2012). Single cell genomics: An individual look at microbes. *Curr Opin Microbiol* **15**:613–620.
Steven J. Biller, Florence Schubotz, Sara E. Roggensack, Anne W. Thompson, Roger E. Summons, Sallie W. Chisholm. (2014). Bacterial vesicles in marine ecosystems. *Science* **343**:183–186.
Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol* **3**:1–11.
Swan BK, Chaffin MD, Martinez-Garcia M, Morrison HG, Field EK, Poulton NJ, *et al.* (2014). Genomic and metabolic diversity of marine group i thaumarchaeota in the mesopelagic of two subtropical gyres. *PLoS One* **9**. doi:10.1371/journal.pone.0095380.
Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, *et al.* (2011). Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**:1296–300.
Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, *et al.* (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U S A* **110**:11463–8.
Swan CM, Siegel DA, Nelson NB, Carlson CA, Nasir E. (2009). Biogeochemical and hydrographic controls on chromophoric dissolved organic matter distribution in the Pacific Ocean. *Deep Res Part I Oceanogr Res Pap* **56**:2175–2192.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**:2725–2729.
Tatusov RL, Galperin MY, Natale D a, Koonin E V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**:33–36.
Teeling H, Fuchs B, Becher D, Klockow C, Gardbrecht A, Bennke CM, *et al.* (2012). Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytplankton Bloom. *Science* **336**:608–611.
Temperton B, Giovannoni SJ. (2012). Metagenomics: Microbial diversity through a scratched

lens. *Curr Opin Microbiol* **15**:605–612.
Thomas T, Gilbert J, Meyer F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* **2**:3.
Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, *et al.* (2005). Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**:1311–1313.
Thompson M, Chourey K. (2008). Experimental approach for deep proteome measurements from small-scale microbial biomass samples. *Anal Chem* **80**:9517–9525.
Thrash JC, Seitz KW, Baker BJ, Temperton B, Gillies LE, Rabalais NN, *et al.* (2017). Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico "Dead Zone". *MBio* **8**:1–20.
Toh H, Weiss BL, Perkin SAH, Yamashita A, Oshima K, Hattori M, *et al.* (2006). Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of Sodalis glossinidius in the tsetse host. *Genome Res* **16**:149–156.
Tran P, Ramachandran A, Khawasik O, Beisner BE, Rautio M, Huot Y, *et al.* (2019). Microbial life under ice: Metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered Lakes. *Environ Microbiol* **20**:2568–2584.
Tremblay J-E, Smith Jr. WO. (2007). Chapter 8 Primary Production and Nutrient Dynamics in Polynyas. *Elsevier Oceanogr Ser* **74**:239–269.
Tremblay JÉ, Bélanger S, Barber DG, Asplin M, Martin J, Darnis G, *et al.* (2011). Climate forcing multiplies biological productivity in the coastal Arctic Ocean. *Geophys Res Lett* **38**:1–5.
Tremblay JE, Gratton Y, Fauchot J, Price NM. (2002). Climatic and oceanic forcing of new, net, and diatom production in the North Water. *Deep Res Part II Top Stud Oceanogr* **49**:4927–4946.
Tremblay L, Gagné JP. (2009). Organic matter distribution and reactivity in the waters of a large estuarine system. *Mar Chem* **116**:1–12.
Treves DS, Manning S, Adams J. (1998). Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli. *Mol Biol Evol* **15**:789–797.
Tully BJ, Graham ED, Heidelberg JF. (2017). Data Descriptor : The reconstruction of 2 , 631 draft metagenome-assembled genomes from the global oceans. *Sci Data* **5**;1–8.
Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the

environment. *Nature* **428**:37–43.
Unfried F, Becker S, Robb CS, Hehemann JH, Markert S, Heiden SE, *et al.* (2018). Adaptive mechanisms that provide competitive advantages to marine bacteroidetes during microalgal blooms. *ISME J*. **12**:2894-2906.
Urbach E, Vergin KL, Larson GL, Giovannoni SJ. (2007). Bacterioplankton communities of Crater Lake, OR: Dynamic changes with euphotic zone food web structure and stable deep water populations. *Hydrobiologia* **574**:161–177.
Urbach E, Vergin KL, Young L, Morse A, Larson GL, Giovannoni SJ. (2001). Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnol Oceanogr* **46**:557–572.
Varela MM, Van Aken HM, Herndl GJ. (2008). Abundance and activity of Chloroflexi-type SAR202 bacterioplankton in the meso- and bathypelagic waters of the (sub)tropical Atlantic. *Environ Microbiol* **10**:1903–1911.
Varela MM, Van Aken HM, Sintes E, Herndl GJ. (2008). Latitudinal trends of Crenarchaeota and Bacteria in the meso- and bathypelagic water masses of the Eastern North Atlantic. *Environ Microbiol* **10**:110–124.
Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen J a, *et al.* (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**:66–74.
VerBerkmoes N, Denef V. (2009). Systems biology: functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* **7**:196–205.
Vergin KL, Urbach E, Stein JL, DeLong EF, Lanoil BD, Giovannoni SJ. (1998). Screening of a fosmid library of marine environmental genomic DNA fragments reveals four clones related to members of the order Planctomycetales. *Appl Environ Microbiol* **64**:3075–3078.
Vogel C, Marcotte EM. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**:227–32.
Vonk JE, Sanchez-Garca L, Van Dongen BE, Alling V, Kosmach D, Charkin A, *et al.* (2012). Activation of old carbon by erosion of coastal and subsea permafrost in Arctic Siberia. *Nature* **489**:137-140.
Vorobev A, Sharma S, Yu M, Lee J, Washington BJ, Whitman WB, *et al.* (2018). Identifying labile DOM components in a coastal ocean through depleted bacterial transcripts and chemical signals. *Environ Microbiol* **20**:3012–3030.

Walker CB, de la Torre JR, Klotz MG, Urakawa H, Pinel N, Arp DJ, *et al.* (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci* **107**:8818–23.
Walker SA, Amon RMW, Stedmon C, Duan S, Louchouarn P. (2009). The use of PARAFAC modeling to trace terrestrial dissolved organic matter and fingerprint water masses in coastal Canadian Arctic surface waters. *J Geophys Res* **114**:1–12.
Walsh DA, Zaikova E, Howes C, Song Y. (2009). Metagenome of a Versatile Chemolithoautotroph from Expanding Oceanic Dead Zones. *Science* **326**:578–582.
Walsh David A, Zaikova E, Hallam SJ. (2009). Small volume (1-3L) filtration of coastal seawater samples. *J Vis Exp* 1–2.
Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**:5261–5267.
Watson SW, Waterbury JB. (1971). Characteristics of Two Marine Nitrite Oxidizing Bacteria,. *Microscopy* **77**:203–230.
Weingartner TJ, Danielson S, Sasaki Y, Pavlov V, Kulakov M. (1992). The Siberian Coastal Current: A wind- and buoyancy-forced Arctic coastal current. *J Geophys Res Ocean* **104**:29697–29713.
Werwath J, Arfmann HA, Pieper DH, Timmis KN, Wittich RM. (1998). Biochemical and genetic characterization of a gentisate 1, 2-dioxygenase from Sphingomonas sp. strain RW5. *JBacteriol* **180**:4171–4176.
Whitman WB, Coleman DC, Wiebe WJ. (1998). Prokaryotes: the unseen majority. *Proc Natl Acad Sci* **95**:6578–6583.
Williams PM, Druffel E. (1988). Dissolved Organic Matter in the Ocean: Comments on a Controversy. *Oceanography* **1**:14–17.
Williams TJ, Cavicchioli R. (2014). Marine metaproteomics: deciphering the microbial metabolic food web. *Trends Microbiol* **22**:248–260.
Williams TJ, Long E, Evans F, Demaere MZ, Lauro FM, Raftery MJ, *et al.* (2012). A metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula coastal surface waters. *ISME J* **6**:1883–900.
Williams WJ, Carmack EC. (2015). The 'interior' shelves of the Arctic Ocean: Physical

oceanographic setting, climatology and effects of sea-ice retreat on cross-shelf exchange. *Prog Oceanogr* **139**:24–41.
Wintermute EH, Silver PA. (2010). Emergent cooperation in microbial metabolism. *Mol Syst Biol* **6**:1–7.
Woese CR. (1987). Bacterial Evolution. *Microbiol Rev* **51**:221–271.
Wu M, Chatterji S, Eisen JA. (2012). Accounting for alignment uncertainty in phylogenomics. *PLoS One* **7**:1–10.
Wuchter C, Abbas B, Coolen MJL, Herfort L, van Bleijswijk J, Timmers P, *et al.* (2006). Archaeal nitrification in the ocean. *Proc Natl Acad Sci* **103**:12317–12322.
Yamada T, Sekiguchi Y. (2009). Cultivation of Uncultured Chloroflexi Subphyla: Significance and Ecophysiology of Formerly Uncultured Chloroflexi 'Subphylum I' with Natural and Biotechnological Relevance. *Microbes Environ* **24**:205–216.
Yamashita Y, Tanoue E. (2009). Basin scale distribution of chromophoric dissolved organic matter in the Pacific Ocean. *Limnol Oceanogr* **54**:598–609.
Yamashita Y, Tanoue E. (2008). Production of bio-refractory fluorescent dissolved organic matter in the ocean interior. *Nat Geosci* **1**:579–582.
Zaikova E, Walsh DA, Stilwell CP, Mohn WW, Tortell PD, Hallam SJ. (2010). Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia. *Environ Microbiol* **12**:172–91.
Zhao X, Schwartz CL, Pierson J, Giovannoni SJ, Mcintosh JR, Nicastro D. (2017). Three-Dimensional Structure of the Ultraoligotrophic Marine Bacterium. *Appl Environ Microbiol* **83**:1–14.

Appendix A

1. Prepare reagents

1.1) Prepare SDS-extraction solution: 0.1 M Tris-HCl pH 7.5, 5 % glycerol, 10 mM EDTA and 1 % SDS. Filter-sterilize using a 0.22 μm filter and store at 4°C.

1.2) Prepare stock reagents needed for the polyacrylamide gel.

1.2.1) Prepare 1.5 M Tris-HCL pH 8.8. Filter-sterilize using a 0.22 μm filter and store at room temperature.

1.2.2) 0.5 M Tris-HCL pH 6.8. Filter-sterilize using a 0.22 μm filter and store at room temperature.

1.2.3) Prepare 10 % SDS. Filter sterilize using a 0.22 μm filter and store at room temperature.

1.3) Prepare solutions needed for in-gel trypsin digest and peptide extraction.

1.3.1) Prepare 100 mM ammonium bicarbonate. Filter-sterilize using a 0.22 μm filter and store at 4 °C.

1.3.2) Prepare 1 M DTT stocks. Suspended in 100 mM ammonium bicarbonate and store at -80 °C.

1.3.3) Prepare 550 mM iodoacetamide stocks. Suspended in 100 mM ammonium bicarbonate and store at -80 °C.

1.3.4) Prepare 100 ng/μl trypsin stocks. Aliquot 60 μl into 1.5 ml microcentrifuge tubes and store at -80°C.

1.3.5) Prepare extraction solution: 1 % formic acid, 2 % acetonitrile.

1.3.6) Prepare resuspension solution: 5 % acetonitrile and 0.1 % formic acid.

2. Perform cell lysis in cartridge filter unit with cells preserved in RNA stabilization solution.

2.1) Expel the RNA stabilization solution from the cartridge filter unit (1.5 ml) and into a 2 ml microcentrifuge tube using a 60 ml syringe attached at the luer-lock end of the filter.

2.2) Centrifuge the 2 ml microcentrifuge tube for 10 minutes at 17000 x g to pellet any cellular debris. This is done so that the filter in step 2.3 does not clog.

2.3) Transfer supernatant to a 10 K ultracentrifugal filter unit to capture any proteins originating from cells that have lysed from the freeze/thaw of the sample. Do not discard the pellet. It will be used in step 2.5. Perform centrifugation of the ultracentrifugal filter unit at 3270 x g for 30 minutes or until the volume has reduced to about 600 µl.

2.4) Melt the tip of a p10 tip and block the non luer-lock side of the cartridge filter unit with the open end of the tip. This will ensure that no extraction buffer and biomass escapes during steps 2.5-2.7.

2.5) Suspend the pellet in 1 ml SDS-extraction solution and pipette it into the original cartridge filter unit. The easiest way to pipette into the cartridge filter unit is to stack a p200 tip onto a p1000 tip and use this double tip for pipetting.

2.6) Add 1 ml of SDS extraction solution to the cartridge filter unit so the total volume is about 2 ml and incubate for 10 minutes at room temperature while rotating in a hybridization oven. Place 3 crumpled lab wipes at the bottom of a 50 ml conical tube. Parafilm both ends of the cartridge filter unit closed and put the cartridge filter unit into the 50 ml tube. Close the lid and put the tube into the hybridization oven.

2.7) After 10 minutes place the Sterivex filter on a foam floater and secure it in place with a 5 ml syringe on the luer-lock end. Float and incubate in a 95 °C water-bath for 15 minutes.

2.8) Let the cartridge filter unit cool and rotate at room temperature for 1 hour (as in step 2.6).

2.9) Expel the SDS extraction solution/cell lysate out of the filter with a 60 ml syringe into the same ultracentrifugal filter unit as before (step 2.3). Add 1 ml of fresh SDS extraction solution to the cartridge filter unit and mix for 30 seconds by hand, then expel into the ultracentrifugal filter unit using the 60 ml syringe. This is to rinse the cartridge filter unit to ensure all proteins have been removed.

2.10) Perform centrifugation on the ultracentrifugal filter unit for 45 minutes at 3270 x g or until the volume in the filter unit is less than 600 µl.

2.11) Discard flow-through and top up the ultracentrifugal filter unit with fresh SDS-extraction solution.

2.12) Centrifuge the filter unit for another 45 minutes at 3270 x g.

2.1.3) Repeat steps 2.11 and 2.12 twice more. Ensure the final volume in the ultracentrifugal filter unit is at most 600 µl at the end of the final spin.

2.14) At this point, split the concentrate in two. One fraction will be used for DNA precipitation and the other for protein precipitation.
Note: The fractionation amount depends on the amount of biomass that was filtered and the intended uses of the products. In our case, we split the concentrate with 10 % towards DNA precipitation and 90 % towards protein precipitation.

3. Protein precipitation.
3.1) Add 4 volumes methanol:acetone (50:50) to one volume of concentrate and vortex for 10 seconds. Incubate overnight at -20 °C.

3.2) Spin down at 17,000 x g for 30 minutes. Decant the supernatant and let the pellet (may be invisible) dry in a speedvac for one hour (or until dry). Note: Do not over dry the pellet as this may make it difficult to resuspend.

3.3) Suspend the pellet in 25 µl of SDS-extraction solution. Let sit for one hour then resuspend by pipetting up and down.

3.4) Quantify protein using a protein assay kit and the manufacturers instructions.

4. DNA precipitation.

4.1) Add SDS-extraction solution to the concentrate fraction to be used for DNA precipitation until the 500 µl mark. This step is simply to increase the volume of the solution, making it easier to work with.

4.2) Add 0.583 volumes of a protein precipitation reagent (such as the MPC protein precipitation reagent) and vortex for 10 seconds. You should see a white precipitate form. Note: We have also used phenol:chloroform DNA extraction methods, but it is more difficult due to the low volumes. We obtained better DNA yields using the MPC Protein Precipitation Reagent.

4.3) Centrifuge at 17,000 x g and 4 °C for 10 minutes.

4.4) Transfer supernatant to another 1.5 ml microcentrifuge tube and add 0.95 volumes of isopropanol. Invert 30-40 times.

4.5) Centrifuge for 10 minutes at 4 °C at maximum speed.

4.6) Carefully decant and discard the supernatant.

4.7) Rinse twice with 750 µl of 70 % ethanol.

4.8) Remove as much ethanol as possible by pipetting, then let air-dry. Note: Do not over dry the DNA as this may make it difficult to resuspend.

4.9) Resuspend in 25 µl of low TE buffer (pH 8).

4.10) Quantify the DNA using a dsDNA assay kit and the manufacturers instructions. Perform agarose gel (1%) electrophoresis on 3 µl of the DNA to check its quality.

5. SDS-PAGE gel of proteins.

5.1) Prepare sample buffer (950 µl Laemmli sample buffer and 50 µl β-mercaptoethanol).

5.2) Add the corresponding volume for 15 µg of protein, or a max of 20 µl to equal volume of sample buffer and boil for 4 minutes. Samples can now be stored at -80 °C until SDS-PAGE can be performed.

5.3) Prepare 10 % acrylamide resolving gel with a 5 % acrylamide stacking gel.

5.4) Load the gel with samples and 4 µl of a protein ladder. Run the gel at constant 120 V until the 250 kDa ladder marker has just reached the resolving gel.

5.5) Stain the gel using a coomassie based stain according to the manufacturers instructions.

6. In-gel trypsin digest and peptide extraction

Note: All steps from here until the end are performed in a biological safety cabinet to minimize contamination.

6.1) Cut off excess gel under the 10 kDa ladder mark for each lane. Each lane will be analyzed on the MS/MS separately. Cut each lane into 1mm x 1mm squares to increase surface area and place all squares from the lane into a low-binding micro-centrifuge tube. Repeat this for all lanes. (1 % acetic acid can be used to prevent the gel from drying out and becoming difficult to cut).

Note: Minimizing contamination is critical at this stage, so gel slicing is performed in a biological safety cabinet on the glass SDS-PAGE gel mold used to make the gel.

6.2) De-stain

6.2.1) Dispense 50 µl of 100 mM ammonium bicarbonate into each low-binding tube, close cap and incubate at 37 °C for 10 minutes.

6.2.2) Dispense 50 µl of acetonitrile, close cap and incubate at 37 °C for 5 minutes.

6.2.3) Aspirate and discard 150 µl.

6.2.4) Repeat steps 6.2.1 and 6.2.2.

6.2.5) Aspirate and discard 95 µl.

6.3) Dehydration

6.3.1) Dispense 50 µl acetonitrile, close cap and wait 5 minutes. Aspirate and discard 45 µl and wait 10 minutes.

6.4) Reduction

6.4.1) Dispense 50 µl DTT (10 mM, diluted in 100 mM ammonium bicarbonate), close cap and wait 30 minutes at 37 °C.

6.5) Alkylation

6.5.1) Dispense 50 µl iodoacetamide (55 mM, diluted in 100 mM ammonium bicarbonate), close cap and wait 20 minutes at 37 °C.

6.5.2) Dispense 100 µl of acetonitrile, close cap and incubate for 5 minutes at 37 °C. Aspirate and discard 195 µl.

6.6) Wash

6.6.1) Dispense 50 µl ammonium bicarbonate, close cap and incubate for 10 minutes at 37 °C.

6.6.2) Dispense 50 µl of acetonitrile, close cap and incubate at 37 °C for 5 minutes. Aspirate and discard 120 µl.

6.7) Dehydration

6.7.1) Dispense 50 µl acetonitrile, close cap and wait 5 minutes. Aspirate and discard 45 µl.

6.7.2) Dispense 50 µl of acetonitrile, wait 5 minutes .

6.7.3) Aspirate and discard 75 µl and wait 5 minutes.

6.8) Digestion

6.8.1) Dispense 25-30 µl of 6 ng/µl Trypsin (until all gel fragments are covered). Close cap and wait 30 minutes at room temperature.

6.8.2) Incubate overnight (4.5 hours minimum) at 37 °C.

6.8.3) Let sit at room temperature for 30 minutes.

6.9) Extraction and peptide transfer

6.9.1) Dispense 30 µl of extraction solution and cover with lid for 30 minutes on ice.

6.9.2) Aspirate 30 µl and dispense aspirated volume in a new labeled low-binding tube.

6.9.3) Dispense 12 µl of extraction solution and 12 µl of acetonitrile into the original tube (with gel). Close cap and incubate for 30 minutes on ice.

6.9.4) Aspirate 15 µl and deposit into the new tube.

6.9.5) Repeat steps (6.9.3 and 6.9.4).

6.10) Place new tubes in a SpeedVac until dry.

6.11) Resuspend in 50 µl of resuspension solution. Vortex on medium for 10 minutes to resuspend.

6.12) Pipette peptide solution into either nano/LC vials or 96-well plates suitable for nano/LC MS/MS work.