METAGENOMIC CHARACTERIZATION OF ARCHAEAL BIOGEOGRAPHY AND COLD ADAPTATION IN THE WESTERN ARCTIC OCEAN

by

Kari Green

Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Department of Environment and Geography
Clayton H. Riddell Faculty of Environment, Earth, and Resources
University of Manitoba
Winnipeg

Copyright © 2023 by Kari Green

ABSTRACT

Compared to other microbes such as Bacteria or algae, our knowledge of archaeal biology is limited. The most abundant Archaea in the oceans are the Marine Groups I, II, and III. Previous studies have found that the biogeographical distributions of each group are distinct, specific, and consistent within the groups. The differences in spatial distributions are suggested to be influenced by light penetration and temperature. Temperature greatly affects the biochemical processes and physiological states of microorganisms. Many indices have been developed to detect temperature adaptation in thermophilic microbes, however, this becomes much more difficult for microbes living in low temperature environments. Metagenomic sequence data was collected from seawater in the Western Arctic Ocean during the GEOTRACES 2015 expedition. This thesis aims to categorize the distributions patterns of archaeal MAGs in relation to their biogeography and proteomic presentations. Members of all three marine groups were identified and it was determined that their relative abundances were highly connected with Arctic stratification and water mass layers. Additionally, it was observed here for the first time that bins from all three marine groups displayed variable distribution ecotypes. Using a collection of archaeal genomes from the Bacterial and Viral Bioinformatics Resource Center database, amino acid compositional indices were used to detect signatures of cold adaptation in their proteomes. Although signatures of cold adaptation were not detected, it was determined that there were differences based on the phylogenetics and environmental sources. Higher arginine to lysine ratios were observed in the sequences collected from sea sponges, further supported by significant differences observed in the codon usage and amino acid presence, suggesting a strong selective pressure to obtain host-adaptation. Overall, this thesis contributes to our understanding of archaeal ecology, especially in an understudied environment like the Arctic Ocean.

ACKNOWLEDGEMENTS

First, I would like to greatly acknowledge my supervisor Dr. Eric Collins as well as my committee members Drs. CJ Mundy and Jose Luis Rodriguez-Gil for their continued help, guidance, patience, and knowledge throughout my degree. I would like to thank Anais Gentilhomme for her contributions to the analysis of the cold-adaptation amino acid indices. I would like to acknowledge the hard work of Katlin Bowman, Kyle Dilliplaine, Pete Morton, and everyone aboard the CGC Healy during the 2015 GEOTRACES cruise that was involved in the collection of the metagenomic samples used.

I would like to thank all of the friends that I have made through this program. Thank you to Zakhar Kazmurik, Patricia Montalvo, and Kim Rouanet for your support, encouragement, and laughs. I will treasure the experiences we have shared forever. Kallie Strong, it was an absolute pleasure mentoring you and you have taught me so much through this process. I know you have great things ahead of you.

I would like to thank my friends who have kept me sane outside of the office. Thor Henricksson for always keeping me fed. Ali Doucet and Shelby Reimer for loving and supporting me through everything. You both have been instrumental in defining in me what I want to be as a woman, as a woman in science, and as a friend.

I would like to particularly thank my dad, Mike Green, and my partner, Erik Tycholiz for all of your endless love and support. It was with your belief, care, love, and guidance that I am where I am today.

TABLE OF CONTENTS

ABSTRACT……i
ACKNOWLEDGEMENTS……ii
LIST OF TABLES……...v
LIST OF FIGURES……vi
CHAPTER 1. INTRODUCTION……...1
1.1 THE ARCTIC OCEAN……..….1
1.2 CHALLENGES IN STUDYING ARCHAEA……..5
1.3 ARCHAEAL TAXONOMY……7
1.4 OBJECTIVES……..10
CHAPTER 2. IDENTIFICATION OF ARCHAEAL MAGS IN THE WESTERN ARCTIC OCEAN……..12
2.1 INTRODUCTION…….…...12
2.1.1 ARCHAEAL BIOGEOGRAPHY……..……12
2.2 METHODS……..15
2.2.1 DATA COLLECTION…….….15
2.2.2 METAGENOMIC ANALYSIS……17
2.2.3 MAG ANNOTATION AND TAXONOMIC ASSIGNMENT…….19
2.2.4 BIOGEOGRAPHY MAPPING AND VISUALIZATION……...20
2.3 RESULTS……22
2.3.1 WATER MASS DETERMINATION AND PHYSICAL PARAMETERS FROM THE 2015 GEOTRACES EXPEDITION……..22
2.3.2 ARCHAEAL MAG IDENTIFICATION……..32
2.3.3 ARCHAEAL RELATIVE ABUNDANCE AT STATIONS 26 AND 30……....37
2.3.4 ARCHAEAL BIOGEOGRAPHY IN THE WESTERN ARCTIC OCEAN…….45
2.4 DISCUSSION……..51
2.4.1 ARCHAEAL MAG IDENTIFIECATION AND THE USE OF METAGENOMIC ANALYSIS…….51

2.4.2 ARCHAEAL DISTRIBUTIONS INFLUENCED BY WATER MASSES RATHER THAN STRICTLY LIGHT PENETRATION AND TEMPERATURE IN THE WESTERN ARCTIC OCEAN…….53
CHAPTER 3. PROTEOMIC SIGNATURES OF COLD ADAPTATION IN ARCHAEA…….58
3.1 INTRODUCTION……58
3.1.1. MG-I AMMONIA OXIDIZING PROCESSES……...58
3.1.2 ARCHAEAL COLD ADAPTATION……..61
3.1.3 MG-I HOST ADAPTATION IN SEA SPONGES……64
3.2 METHODS……..66
3.2.1 CALCULATION OF AMINO ACID COMPOSITIONAL INDICES FOR THE DETECTION OF COLD ADAPTATION SIGNATURES…….66
3.2.2 CALCULATION OF CODON AND AMINO ACID USAGE…….67
3.3 RESULTS……68
3.3.1 AMINO ACID COMPOSITIONAL INDICES SHOWS LITTLE VARIATION BETWEEN MG-I MAGS……68
3.3.2 EURYARCHAEOTA MAGS BRANCH INTO FIVE CLUSTERS BASED ON AMINO ACID COMPOSITIONAL INDICES……..…….76
3.3.3 DIFFERENCES IN ARGININE TO LYSINE RATIOS, CODON USAGE, AND AMINO ACID USAGE OBSERVED IN MG-I MAGS ASSOCIATED WITH SEA SPONGES……..……80
3.3.4 MG-I PHYLOGENY IS NOT CONSISTENT WITH GENE CONTENT……87
3.4 DISCUSSIONS……93
3.4.1 ASSESSING THE EFFICACY OF THE AMINO ACID COMPOSITIONAL INDICES FOR ASSESSING COLD ADAPTATION IN ARCHAEA……93
3.4.2 HOST-ADAPTATION IN ARCHAEA……96
3.4.3 EVOLUTION OF MG-I……99
CHAPTER 4. CONCLUSION……101
4.1 SUMMARY OF MAJOR FINDINGS……...101
4.2 RECOMMENDATIONS FOR FUTURE RESEARCH…….103

CHAPTER 5. APPENDIX……..105
CHAPTER 6. REFERENCES……113

LIST OF TABLES

Table 1. Classification of archaeal groups mentioned Chapter 1.3…….10
Table 2. Quality scores for the archaeal MAGs determined using CheckM……...34
Table 3. Archaeal MAG characteristics as described by BV-BRC whole-genome analysis tool from the 2015 GEOTRACES cruise……...37
Table 4. Averaged Amino acid compositional indices for the five clusters observed for the Euryarchaeota……..79
Table 5. List of the Sponge-associated MAGS, their host, and their location of origin……..82
Supplementary Table 1. CheckM scores of completeness and contamination for MG-I MAGs…….106

LIST OF FIGURES

Figure 1. Geographical map displaying the stations from the 2015 GEOTRACES expedition…..17
Figure 2. Metagenomic pipeline used in the analysis…….19
Figure 3. Section plot of N* values for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01)…….24
Figure 4. Section plot of silicate measurements for the upper 500 m of the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01)……..25
Figure 5. Section plot of temperature measurements for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01)……26
Figure 6. Section plot of salinity measurements for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01)……27
Figure 7. Water mass determinations…….28
Figure 8. Section plot of dissolved oxygen concentrations for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01)……....30
Figure 9. Section plot of nitrate concentrations for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01)……31
Figure 10. Section plot of nitrite concentrations for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01)……..32
Figure 11. Phylogenetic tree of the archaeal MAGs identified in the Western Arctic Ocean during the 2015 GEOTRACES expedition (GN01)……..36
Figure 12. Depth profiles for salinity, temperatures, nitrate, and nitrite from the 2015 GEOTRACES expedition from stations 26 and 30……...39
Figure 13. Heat map displaying abundance-based clustering of each bin and each sample at Stations 26 and 30 from the GEOTRACES 2015 expedition (GN01)……..41
Figure 14. Relative abundance measurements for Archaea in the Western Arctic Ocean……43
Figure 15. Archaeal relative abundance by depth for MG-I, MG-II, and MG-III at station 26 and 30 from the 2015 GEOTRACES expedition (GN01)……44
Figure 16. Relative abundance distributions MG-I archaeal MAGs identified in the Western Arctic Ocean……..47

Figure 17. Relative abundance distributions of the MG-II archaeal MAGs identified in the Western Arctic Ocean displaying a surface ecotype……48
Figure 18. Relative abundance distributions of the MG-II archaeal MAGs identified in the Western Arctic Ocean displaying a mid-depth ecotype……..49
Figure 19. Relative abundance distributions of the MG-II archaeal MAGs identified in the Western Arctic Ocean displaying a deep ecotype……50
Figure 20. Relative abundance distributions of the MG-II archaeal MAGs identified in the Western Arctic Ocean……51
Figure 21. The process of nitrification and denitrification occurring in the ocean…….59
Figure 22. Geographical map showing the sampling locations of the MG-I genomes obtained from the BV-BRC database……..68
Figure 23. Heatmaps displaying the standardized amino acid compositional indices for the MG-I MAGs for each global protein family……70
Figure 24. Indices distances against the phylogenetic distances for MG-I highlighting the Western Arctic Archaea……..71
Figure 25. Comparison of indices from the Group A and Group B MG-I……..73
Figure 26. Boxplot of the comparison between fresh water and marine water for Group A and Group B MG-I for the amino acid indices……..75
Figure 27. Geographical map showing the sampling locations of the Euryarchaeota genomes obtained from the BV-BRC database……...76
Figure 28. Heatmaps displaying the standardized amino acid compositional indices for the Euryarchaeota genomes/MAGs for each global protein family…….77
Figure 29. Arginine to lysine Z-scores for the MG-I sponge-associated MAGs……81
Figure 30. Heatmap displaying the standardized frequencies of the aminos acids used in the MG-I MAGs……...84
Figure 31. Bar plot displaying the percentage of each amino acid observed in the genomes of the sponge-associated MAGs and non-sponge associated MAGs……85
Figure 32. Heatmap of codon usage in MG-I…….86
Figure 33. GC content of the sponge associated MAGs compared to the other members of MG-I……87
Figure 34. Presence/absence heatmap for the MG-I genomes……89

Figure 35. Heat map of phylogeny (y-axis) and presence/absence of global protein distances for the MG-I……90
Figure 36. Indices distances against the phylogenetic distances for MG-I……92
Supplementary Figure 1. Phylogenetic tree of 16S rRNA gene sequencing data from samples collected in the Chukchi Sea……...105
Supplementary Figure 2. Amino acid indices distribution for the three MG-I genera analyzed……...……109
Supplementary Figure 3. Annotated phylogenetic tree of the MG-I MAGs……...110
Supplementary Figure 4. Annotated phylogenetic tree of the Euryarchaeota……111
Supplementary Figure 5. Pair-wise distances of the presence or absence of archaeal global protein families in the MG-I MAGs……...……112

CHAPTER 1. INTRODUCTION

1.1 THE ARCTIC OCEAN

The Arctic Ocean is an environment characterized by habitats with extremes in light, temperature, and salinity (Collins et al., 2010; Junge et al., 2001). Globally, the Arctic Ocean is small, accounting for ~3% of the world's oceans by area (Chang & Devol, 2009; Jensen et al., 2019), however, it contains ~20% of the world's continental shelf ($5.1x10^6$ km^2) (Chang & Devol, 2009). This ocean is semi-enclosed by land with approximately 50% of its surface area (9.5×10^{12} m^2) overlying continental shelves (Jensen et al., 2019). The Arctic Ocean consists of two major basins, the Eurasian and the Amerasian Basins, separated by the Lomonosov Ridge (Rudels, 2015). These larger basins can be subdivided by bathymetric regions called subbasins. The Eurasian Basin is further branched into the Nansen Basin and Amundsen Basin (the deepest, ~4500 m), by the Gakkel Ridge, while the Western Amerasian Basin is comprised of the Canada and Makarov subbasins separated by the Alpha and Mendeleev Ridges (Jensen et al., 2019; Rudels, 2015).

Water circulation into the Arctic Ocean occurs in both the east, where Atlantic water enters through the deep passage (2600 m) of the Fram Strait (Rudels, 2015), and the west, where northern Pacific water enters through the narrow and shallow (50 m) Bering Strait (Jensen et al., 2019; Rudels, 2015). Atlantic water moves through the Fram Strait and Barents Sea via the Nordic Sea and through the Canadian Archipelago by Baffin Bay and Davis Strait (Rudels, 2015). Fram Strait is the only passage that allows two-way movement of water into and out of the Arctic Ocean. Upon entering the Arctic Ocean, the warm Atlantic water encounters, and melts sea ice. In the uppermost layer, the water is cooled and homogenized as ice forms through brine rejection and thermohaline convection (Rudels, 2015). In the lower layers, the Atlantic water cools and sinks (Whitmore et.

al., 2019). Pacific water entering the Chukchi Sea through Bering Strait undergoes biochemical and physical changes, such as surface heating and cooling, and internal mixing, during its transit over the shelves and into the Arctic Ocean Basins (Gong & Pickart, 2016). The western region is strongly defined by its thermocline and macronutrient characteristics. Surface water circulation in the Amerasian Basin is strongly influenced by the anticyclonic Beaufort Gyre and the Transpolar Drift, a strong current that advects water from the Chukchi, Eastern Siberian, and Laptev Seas across the central Arctic Ocean (Chang & Devol, 2009; Gong & Pickart, 2016; Whitmore et al., 2019, 2022).

Generally, the Arctic Ocean can be separated vertically into three major layers: The polar surface water (PSW), Atlantic water, and deep bottom waters, although variations may be observed within the different basins. In the Amerasian Basin, surface waters, commonly referred to as the Polar Mixed Layer (PML) of the PSW (Rudels, 2015), originate primarily from Pacific Ocean water that is advected through the Bering Strait (Jensen et al., 2019). The PML is fresher and colder and can extend variably 0-50 m (Farmer et al., 2021; Jensen et al., 2019). The stratification observed here is strong enough to reduce the amount of mixing with lower layers which allows the low salinity water in the PML to cool enough for ice to form in the winter. This ice layer persists throughout the year in the central Arctic Ocean (Rudels, 2015). Ice formation allows the PSW to become homogenized in winter through freezing, brine release, and thermohaline convection processes. In summer, this ice will melt resulting in re-stratification, which again is removed during freezing in the fall and winter. Below the PML, salinity sharply increases through the halocline with temperatures remaining close to freezing. Pacific water entering the Bering Strait, less saline than the Atlantic water, but more saline than the PML, forms an intermediate layer. Summer Bering Sea water sits above, and winter Bering Sea water sits below (Rudels, 2015) and

mainly remains within the Amerasian Basin. The water below the halocline and intermediate layer is known as the Atlantic layer coming in from Fram Strait. This water enters the Arctic Ocean and travels in a counterclockwise motion towards Svalbard (Fram Strait Branch; FSB) and further eastward toward the Barents Sea (Barents Sea Branch; BSB) becoming colder and more saline (Jensen et al., 2019). As the FSB travels, it mixes with the colder, less saline water above the halocline and continues to cool as it moves into the Makarov and Canada Basins (Jensen et al., 2019; Woodgate et al., 2001). Within these Basins, the FSB can typically be found between 350 and 800 m. The BSB lies below the FSB and extends to 1150 m. Water below 1200 m originates from the far North Atlantic (Greenland and Norwegian Seas) both entering and exiting through the Fram Strait (Jensen et al., 2019). The residence time in the Makarov and Canada basins differ (300 and 400 years, respectively) due to the separating Alpha-Mendeleev Ridge (Jensen et al., 2019). Although distinct, these basins are still subject to mixing.

The Arctic Ocean is highly climate relevant due to its impact on the formation of North Atlantic Deep water, a driver of the Atlantic Meridional Overturning Circulation, as well as the albedo effect it provides due to high levels of ice coverage (Whitmore et al., 2022). In recent years we have witnessed the loss of sea ice, particularly multi-year sea ice due to increased Arctic warming (Box et al., 2019; Chang & Devol, 2009; Edwards et al., 2020; Garcia-Soto et al., 2021). In the last decade, the rate of Arctic Ocean warming was observed to be 4.5 times greater than the global mean (Garcia-Soto et al., 2021). Circulation models indicate a greater increase in average annual surface air temperatures in the Arctic compared to other areas of the planet (Johannessen et al., 2004). This accelerated warming observed in the Arctic is caused by a positive feedback loop fueled by the melting of snow-covered sea ice. As sea ice melts due to increasing atmospheric temperatures, absorption rates of solar radiation increase resulting in even further sea ice melt.

This process is called Arctic amplification (Rantanen et al., 2022). It has also been observed that there is a general freshening of the Arctic Ocean and surrounding seas in the last decade due to increases in precipitation, river discharge, and sea ice melt (Chang & Devol, 2009; Peterson et al., 2006). This has the potential to affect deep water formation, sea ice extent and thickness, overall ecosystem function, and the livelihood of surrounding communities (Chang & Devol, 2009; Whitmore et al., 2022).

Microbes are often referred to as the first responders in an ecosystem. Short generation times allow microbes to act as markers for environmental changes. They are also likely to be one of the first organisms impacted by very small changes in their local surroundings. Specifically, many microbial niches are situated at the margins between frozen and liquid water, for example, in brine channels in sea ice, permafrost, glacial weathering crusts, or englacial vein boundaries (Edwards et al., 2020). In these micro-environments, liquid water is limited, hindering microbial activity. Thus, even small shifts in temperature can radically alter the niches available to microbial populations (Edwards et al., 2020). Ocean warming may affect these organisms and their productivity in many ways due to changes in sea ice and snow coverage (Box et al., 2019), temperature (Cheng et al., 2017), nutrient availability (Kwiatkowski et al., 2020), competition, pH (Hoegh-Guldberg et al., 2017), and dissolved oxygen concentrations (Nayfach et al., 2021; Schmidtko et al., 2017). As the Arctic continues to change and temperatures rise, marine microbial communities are expected to undergo changes in cell size and structure (Morán et al., 2010), biogeography (Falkowski & Oliver, 2007), and food web dynamics (Baer et al., 2014; O'Connor et al., 2009). One group of microbes, called Archaea, are important contributors to biogeochemical cycling but remain relatively uncharacterized in the Arctic Ocean.

1.2 CHALLENGES IN STUDYING ARCHAEA

Using DNA sequencing technologies, Woese & Fox (1977) provided the first evidence that Archaea constituted the third Domain of life alongside Eukarya and Bacteria (Baker et al., 2020; Offre et al., 2013), but more than 40 years later our knowledge of archaeal biology is still limited compared to the other Domains. Archaea are a unique group of single-celled organisms that appear in many ways like Bacteria but have numerous biochemical and genetic features that are more similar to Eukaryotes (Roy Chowdhury & Dutta, 2012). Although it appears that Archaea can be found in all environments, the chemistry of their ecological context is still of great interest. Offre et al. (2013) describe two principal issues in obtaining a full understanding of Archaea: i) identifying the entirety of biogeochemical processes that can be performed by members of the Domain Archaea; and ii) measuring the total archaeal biomass, their productivity, and the nutrient cycling processes performed in their respective environments. Historically, our understanding of microorganisms has been hindered by our ability to cultivate them, and many studies relied on culture-dependent methods (Baker et al., 2020). However, only six of the 27 proposed archaeal phyla currently have cultured representatives (Adam, Borrel, Brochier-Armanet, & Gribaldo, 2017; Baker et al., 2020), making the majority of these organisms identifiable solely by culture-independent molecular studies (Rinke et al., 2013).

Microbiomes consist of all the organisms, their genomes, and the surrounding habitat from a defined environment (Frioux et al., 2020). The most common method for culture-independent microbiome analyses utilizes high-throughput DNA sequencing (HTS) with either a targeted approach (amplicon sequencing of a barcode sequence) or an untargeted approach (shotgun metagenomic sequencing) (Frioux et al., 2020; Y. X. Liu et al., 2021; Teeling & Glöckner, 2012). Although more expensive and time-consuming, metagenomic sequencing is more informative

compared to amplicon sequencing as it can provide both taxonomic and functional information (Adam et al., 2017; Baker et al., 2020; Y. X. Liu et al., 2021). Briefly, DNA is sequenced as short reads which are assembled into longer contigs. These contigs are then binned to create partial or complete metagenome-assembled genomes (MAGs). Metagenomic approaches can identify a higher diversity of archaeal lineages than 16S ribosomal RNA gene amplicon sequencing, which relies on commonly used universal primer sets, potentially missing entire clades of organisms (Adam et al., 2017; Y. X. Liu et al., 2021; Teeling & Glöckner, 2012). Metagenomic binning is a relatively new development in the world of bioinformatics and is typically the sole source of genetic information for uncultured organisms (Frioux et al., 2020). A considerable hurdle for metagenomic studies is determining the origin of the derived genome bins (Rinke et al., 2013), especially for assembling low-abundance and closely related members of the community (Bickhart et al., 2022). This often depends on the sequencing depth and quality. It becomes very difficult to assemble genomes closely related to a highly abundant species as those bins are often combined. It is possible to obtain high-quality MAGs but typically only after a large amount of manual polishing and curation. If a community has many members, the problem of creating reliable MAGs quickly becomes an overwhelming task. Binning algorithms have undergone many advancements that include better clustering approaches of contigs and their abundance profile, GC content, and k-mer content (Frioux et al., 2020). GC content refers to the proportion of guanine (G) and cytosine (C) in the genomes of the microbes. In bioinformatics, k-mers refer to substrings of length k contained within a biological sequence, for example, a k-mer of length four (4-mer) would represent a DNA sequence that is four bases long (*i.e.,* ATTC). The ability to reconstruct complete or nearly complete MAGs is influenced by numerous factors such as microbial diversity, biomass, sequencing effort, read length, intrapopulation sequence, gene diversity, and the assembly and

binning algorithms used (Meziti et al., 2021). Reconstructing these genomes becomes extremely important when trying to understand the small-scale changes that are occurring in a specific environment (Frioux et al., 2020).

The majority of uncultured genomes originate from sediments and water columns and represent a sizeable proportion of microbial life (Adam et al., 2017; Karner, Delong, & Karl, 2001; Offre et al., 2013; Reji et al., 2019) making the Arctic Ocean a desirable location for researching microorganisms. Developing a unifying concept of life is often hindered by a lack of available data, especially from microbes in cold environments. Understanding the genomic data and metabolic potential of archaeal representatives in the Arctic Ocean is an important step toward understanding their role in the environment and providing a new perspective on their metabolic diversity.

1.3 ARCHAEAL TAXONOMY

Two primary archaeal phyla, the Euryarchaeota and Crenarchaeota, were initially proposed based on phenotypic characteristics (Baker et al., 2020; Offre et al., 2013), but analysis of genomic data and advances in sequencing technologies including HTS and metagenomic assembly and binning, have had a strong impact on archaeal systematics leading to the description of 20 additional phyla and multiple new clades in the last decade (Adam et al., 2017; Baker et al., 2020). Currently, there are three superphyla recognized as the Asgard, DPANN, and TACK (Baker et al., 2020; Guy & Ettema, 2011; Rinke et al., 2013). The DPANN and TACK superphyla were originally named based on the phyla present within them: DPANN referring to the Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanohaloarchaeota, and Nanoarchaeota; and TACK to the Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota.

The TACK Superphylum is one of the most extensively studied clades today (Adam et al., 2017; Baker et al., 2020). Thaumarchaeota is thought to be one of the most numerically abundant clades of Archaea on the planet, nearing 20% of total microbes in some deep ocean and soil environments (Baker et al., 2020; Brochier-Armanet et al., 2008; Karner et al., 2001; Reji & Francis, 2020; Reji et al., 2019; Santoro et al., 2011; Spang et al., 2015). A highly studied representative of this group, *Nitrosopumilus maritimus*, has adapted to life under extreme nutrient limitations and high specific oxidation rates of ammonium at concentrations found in open oceans (Martens-Habbena et al., 2009). Currently, one phylum, the Euryarchaeota, does not fall into a superphylum but contains the greatest diversity of cultured lineages including members of the classes Methanopyri, Methanococci, Methanobacteria, Methanomicrobia, Archaeoglobi, Halobacteria, Thermococci, and Thermoplasmata (Baker et al., 2020; Offre et al., 2013). Euryarchaeota are broadly associated with diverse metabolisms including methanogenesis, methane-oxidation, denitrification, sulfate reduction, iron oxidation, and organotrophy (Offre et al., 2013). Culture-independent studies have shown that these organisms are important contributors to global biogeochemical cycling, such as the production of methane and the remineralization of nitrate (Baker et al., 2020; Offre et al., 2013). However, these classifications are always evolving. Rinke *et al.* (2021) proposed a major reclassification of the Domain consisting of 16 phyla, combining the TACK superphylum into one, and separating the Euryarchaeota into five new phyla.

The first environmental archaeal sequences detected in marine environments were separated into two groups: Marine Group I (MG-I), now known as Thaumarchaeota, and Marine Group II (MG-II), within Euryarchaeota (Brochier-Armanet et al., 2008; Yilmaz et al., 2016). This structure has been further expanded to contain three main marine groups. Within Euryarchaeota, two groups were identified: Poseidoniales or MG-II (Pereira et al., 2019; Rinke et al., 2020), and

Pontarchaea or Marine Group III (MG-III) (Baker et al., 2020; Yilmaz et al., 2016). Currently, MG-II representatives are not well understood as there are no cultured representatives from this group (Pereira et al., 2019). 16S rRNA gene sequences have revealed that MG-II may be divided into two main monophyletic groups, the MG-IIa, now known as *Candidatus Poseidonaceae*, and the MG-IIb, now known as *Candidatus Thalassarchaeaceae* (Rinke et al., 2020). The MG-III are commonly found in oceans worldwide, however, they are typically found in very low abundances. They have been shown to contain genes for recycling high molecular weight organic compounds (Baker et al., 2020), but the majority of their metabolic functions are unknown. For the remainder of the text, the archaeal clades of interest will be referred to by their marine group name (MG-I; MG-II; MG-III).

Table 1. Classification of archaeal groups mentioned Chapter 1.3.

Superphylum/phylum	Clade	Marine Group
TACK	Thaumarchaeota	I
	Aigarchaeota	_
	Crenarachaeota	_
	Korarchaeota	_
DPANN	Diapherotries	_
	Parvarchaeota	_
	Aenigmarchaeota	_
	Naneloarchaeota	_
	Nanoarchaeota	_
Asgard	_	_
Euryarchaeota	Poseidoniales	II
	Pontarchaea	III

1.4 OBJECTIVES

The objectives of this thesis are two-fold: to characterize the diversity and distribution of archaeal communities residing in Arctic seawater, and to identify signatures of cold adaptation in the proteomes of Archaea found in cold Arctic seawater. The first objective is accomplished through two intermediate objectives to: 1) Identifying archaeal MAGs collected from Arctic seawater concerning their biogeography; 2) Applying environmental context to explain relative abundance distributions. The second objective is accomplished by investigating archaeal evolution by analyzing genomic and proteomic signatures within their genomes. Based on previously

collected 16S rRNA gene sequence data (Suppl. Fig. 1), MAGs for commonly occurring taxa including *Nitrosopumilus*, as well as other members of the MG-I are expected to be identified. In addition, other species of Archaea within the Euryarchaeota (MG-II and MG-III) are identified. An examination of associations between these genomes and environmental metadata is performed. Using these genomes, differences in abundance patterns are explored and changes in the proteome associated with living in a colder climate are assessed using six amino acid indices for inferring cold adaptation. The first set of hypotheses for this project are that water masses in the Arctic Ocean will be the strongest factor influencing archaeal community structure and that the phylogenetics of the marine Archaea will reflect their environmental distributions. The second set of hypotheses is that there would be signatures of temperature adaptation in the proteomes of Archaea and that environmental temperature will be a factor influencing the gene content and amino acid usage in Archaea.

This thesis will be subdivided into two main parts. Chapter 2 will discuss the biogeographical characterization of archaeal MAGs identified in the western Arctic. The methods will be discussed including water sampling, DNA extraction, and DNA sequencing, as well as any computational tools used to analyze the physical and biological data collected. Chapter 2 will cover the results from the first portion of the thesis including the water mass characterizations and the biogeographical analysis of the Archaeal MAGs in the Western Arctic. This chapter will conclude with a discussion of the relative abundance distributions and their connection to the difference water mass layers observed in the Arctic. Chapter 3 will discuss the identification of cold adaptation signatures in the proteomes of MG-I and MG-II Archaea using six amino acid compositional indices. This chapter will cover the methods used to calculate the amino acid indices and any relevant statistical tests. Additionally, this chapter will discuss the efficacy of these indices

and the evolutionary status of the MG-I, expanding to include host adaptation. Chapter 4 will conclude with a summary of major results and recommendations for future research.

CHAPTER 2. IDENTIFICATION OF ARCHAEAL MAGS IN THE WESTERN ARCTIC OCEAN

2.1 INTRODUCTION

2.1.1 ARCHAEAL BIOGEOGRAPHY

Certain groups of Archaea have been found to preferentially reside in colder environments, often lower in the water column (Pereira et al., 2019), while others have shown a preference for warmer photic zones (Rinke et al., 2019). MG-I are highly abundant in deeper waters below the thermocline (Ionescu et al., 2009), while MG-II often surpass MG-I abundances in surface waters, especially in summer months (Pereira et al., 2019; Yilmaz et al., 2016). The MG-IIa (Candidatus Poseidoniales) is typically composed of surface microbes, while the MG-IIb (Candidatus Thalassarchaeaceae) are often found at depths below 200 m (Pereira et al., 2019). Most of these studies did not sample below 500 m. These groups have been shown to have different seasonal variations depending on their location of origin (Pereira et al., 2019). For example, MG-IIa are found in higher abundance in the polar oceans in winter but were predominant in summer months in the Mediterranean (Pereira et al., 2019). It has been proposed that MG-IIa have evolved to occupy many diverse niches while remaining surface residing (Rinke et al., 2019). Some rarer phyla have also been identified in deep-sea waters including Bathyarchaeota and Woesearchaeota from the Indian Ocean and North Pacific Ocean (Pereira et al., 2019).

Liu *et al.* (2018) collected samples off the coast of Georgia, USA from April to November 2014 (Q. Liu et al., 2018). Nitrite was found at low concentrations in the surface waters but increased below ~100 m (Q. Liu et al., 2018). Consistent with other studies (Ionescu et al., 2009; Pereira et al., 2019; Reji et al., 2019; Yilmaz et al., 2016), the majority (80%) of MG-I 16S ribosomal RNA gene sequences were recovered at 200 m (Q. Liu et al., 2018). These results contrast with studies from high-latitude regions such as the North Sea (Herfort et al., 2007; Pitcher et al., 2011; Wuchter et al., 2006) and polar oceans (Church et al., 2003) where MG-I abundance was highest at even greater depths (>200 m). Liu *et al*. proposed that these differences in abundances may be attributed to the enhanced vertical mixing that occurs at lower latitudes in summer months (Q. Liu et al., 2018). Although the MG-I abundance increased around 200 m, Euryarchaeota remained the dominant archaeal phyla in all the samples (Q. Liu et al., 2018). This is consistent with results from the Northeastern China Sea (H. Liu et al., 2017).

It is commonly suggested that the distribution of MG-I may be related to photoinhibition, as light may be an inhibitor for growth, however, it is not clear if this inhibition is a result of light directly (phototaxis) or if photochemical reactive oxygen species play a role (Merbt et al., 2012; Qin et al., 2014). This may also explain the decrease of MG-I in the summer months in the polar zones, as the levels of sunlight increase dramatically during that time of year, and their preference is to reside deeper within the water column. Lui *et al*. also suggest that at sites where light penetration was consistently low, the water temperature was a secondary control on the bloom potential (Q. Liu et al., 2018). Seasonal variability in the Arctic Ocean will likely significantly impact the communities of microbes residing there: this ocean is ice-covered during much of the winter and complete darkness can last up to two months (Christman et al., 2011). These conditions

are expected to support the growth of ammonia-oxidizing Archaea (AOA) as shown by other studies (Baer et al., 2014; Christman et al., 2011; Q. Liu et al., 2018).

Arctic stratification due to salinity is a key factor influencing circulation in the Arctic Ocean, ice coverage, and productivity (Farmer et al., 2021). Surface waters tend to be nutrient-poor and colder than their underlying warmer, nutrient-rich water (Farmer et al., 2021). The PML tends to be comprised of colder, fresher water where primary production typically takes place (Oziel et al., 2022). The subsurface is comprised of the saltier, nutrient-replete waters of Pacific and Atlantic origin (Farmer et al., 2021; Oziel et al., 2022). Water inputs into the Arctic Ocean are supplied from distinct Atlantic and Pacific sources. Salty Atlantic waters originate from inflows through the Fram Strait and the Barents Sea. Atlantic waters dominate the eastern and central Arctic Ocean. In the Western Arctic, this Atlantic water is covered by fresher Pacific water coming in from the Bering Strait (Farmer et al., 2021). Stratification is stronger in the Western Arctic as compared to the eastern Arctic (Farmer et al., 2021). The PML stratification is due to excess precipitation, river outflow/discharge, sea ice melt, and the advection of low-salinity ocean waters across the Bering Strait (Farmer et al., 2021). Understanding the various water masses and the distribution of different microbial clades can provide insight into the evolution and ecology of the organisms as they travel through the Arctic Ocean. Environmental variables such as light, nutrient availability, water column mixing, and water mass characteristics may control the presence or absence of different microbial clades (Arandia-gorostidi et al., 2022). Genome-resolved metagenomics offers an opportunity to identify the causative factors for these trends at the genomic level, for example, due to the presence or absence of particular genes and pathways, or proteomic adaptation to low temperature.

To investigate the ecology of Archaea in the Western Arctic Ocean, metagenomic seawater samples were collected during the GEOTRACES 2015 expedition. Samples were collected in late summer from 66 stations in the Western Arctic Ocean (Figure 1). Physical parameters collected included depth, temperature, salinity, nutrient concentrations, and heavy metal concentrations. After sample filtration, DNA extraction, and DNA purification, metagenomic sequencing was performed. The goals were to catalogue and identify the archaeal taxa collected from the Western Arctic and to identify the population structure of archaeal communities in the Western Arctic using metagenome-assembled genomes (MAGs). Previous studies have suggested that temperature and light penetration are the main factors driving archaeal abundance patterns, however, I hypothesize that water masses in the Arctic Ocean will be the strongest factor influencing archaeal community structure. Each water mass layer has difference physical and chemical properties that are likely to influence any abundance variations observed. For functional convenience, however, the layers themselves will be used proxy for the underlying chemical and physical properties that are driving the abundance distributions observed.

2.2 METHODS

2.2.1 DATA COLLECTION

An oceanographic expedition (HLY1502; (Kadko & Landing, 2015)) that departed from Dutch Harbor, Alaska was conducted from August to October in the Western Arctic as part of the GEOTRACES program (GN01) which mainly focused on characterizing the distribution of trace elements and other chemical parameters in the oceans. GN01 transited through the Bering and Chukchi Seas to the North Pole in two transects (Figure 1): 1) through the Makarov Basin; and 2) through the Canada Basin. Microbial community seawater samples were collected in the Western

Arctic Ocean during this expedition using a CTD rosette. The collected water was filtered onto 0.22 µm Sterivex filters using a peristaltic pump, allowing sample water to run through the tubing equipment before filter attachment. As much water as possible was removed from the filters using a syringe and then stored at -80°C until sequencing could be performed. At each sampling site, depth, salinity, and temperature were measured using a CTD cast. Other variables including particulate organic nitrogen (PON), particulate organic carbon (POC), nitrate (NO_3^-), nitrite (NO_2^-), ammonium (NH_4^+), and dissolved oxygen (DO) were also measured according to the GEOTRACES protocols (Cutter et al., 2017). Briefly, water was collected via the ship's rosette using approximately 250 mL per depth (50 mL in triplicate, plus bottle rinses). Samples were filtered via a pressure filtration using 0.22 um Sterivex filter and stored in 60 mL square wide-mouth HDPE bottles. And stored at -20°C. Dissolved Oxygen was measured using the automated Winkler method with a detection limit of 1 µmol L^{-1}. Ammonia, nitrite, and nitrate were measured using an automated colorimetric method with detection levels of 0.1 µmol L^{-1}. Silicate was measured in the same way with a detection level of 0.4 µmol L^{-1}.

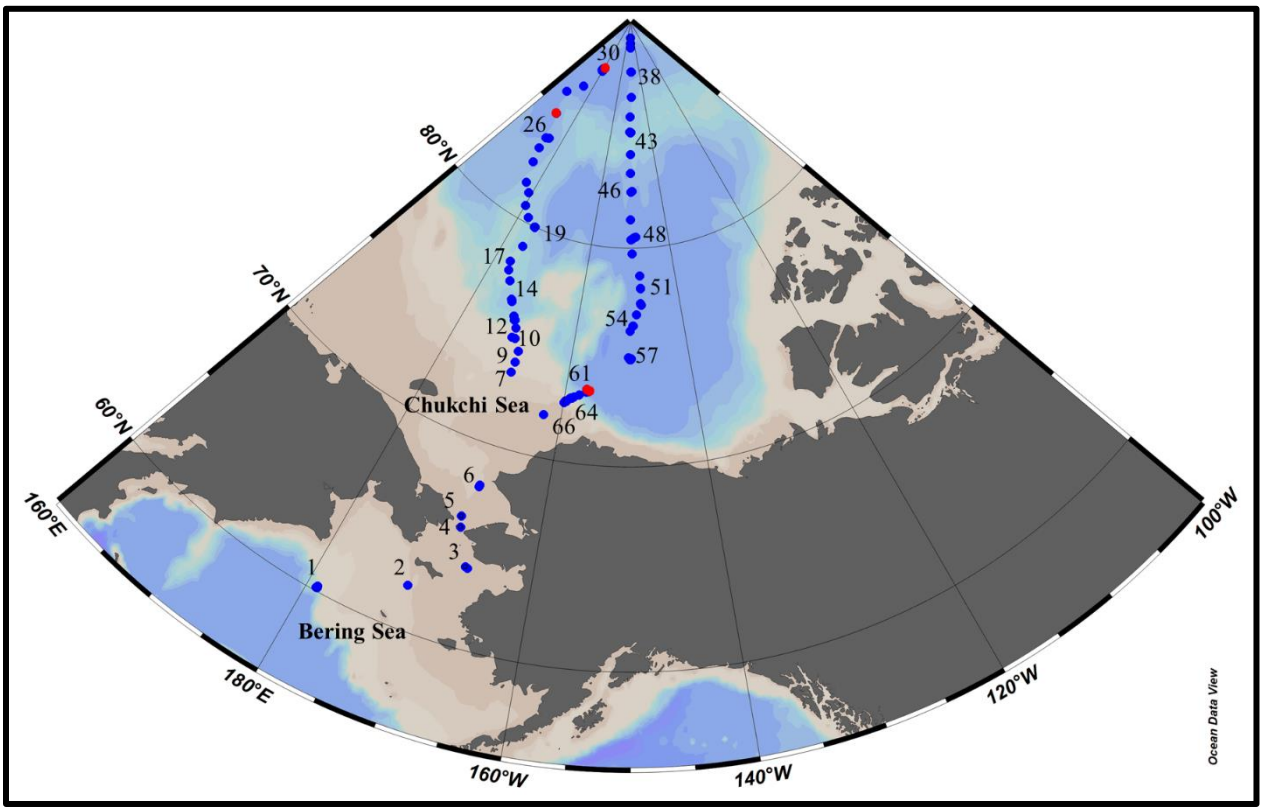


Figure 1. Geographical map displaying the stations from the 2015 GEOTRACES expedition. Red indicates the stations (26, 30, 60, 61) where metagenomic data was collected.

2.2.2 METAGENOMIC ANALYSIS

DNA was extracted and purified from filters using the Qiagen PowerWater Sterivex kit. DNA quality was measured using a Qubit and sequenced on an Illumina MiSeq at 2x250 base pairs (bp). The following metagenomic pipeline is shown in Figure 2. BBtools (Bushnell et al., 2017) was used to remove unwanted DNA sequences (*i.e.,* barcodes, adaptors, sequencing artifacts, and low-complexity sequences) and perform quality controls using the following commands: optical duplication removal (clumpify.sh); low-quality region removal (filterbytile.sh); adapter trimming, and sequencing artifact and low-entropy sequencing removal (bbduk.sh); error correction (bbmerge.sh, clumpify.sh, tadpole.sh); and normalization (bbnorm.sh). Reads were additionally normalized (bbnorm.sh) before co-assembly. Clustering

with minhash was performed for the co-assembly. A co-assembly with 128 samples was performed using metaSPades (Nurk et al., 2017) to elucidate the lower abundance members within each sample. Error-corrected reads were mapped back to the co-assembly (bbmap.sh; percent identity = 100%) to obtain the relative abundances of each contig in each sample. Metagenomic binning was performed by placing contigs into putative microbial genomes (metagenomic assembled genomes or MAGS) using MetaBAT2 (Kang et al., 2019), Maxbin2 (Wu et al., 2016), and MetaBinner (Wang et al., 2023). The binning outputs were consolidated using DASTools (Sieber et al., 2018), which is an automated method the integrates the results of numerous binning algorithms to calculate an optimized, non-redundant set of bins from a single assembly. Completeness and contamination measures were determined using CheckM (Parks et al., 2015). To do this, CheckM uses prodigal (Hyatt et al., 2010) to identify genes in your sequence by first placing your genome on a tree and using that phylogenetic placement to identify sets of genes that should be in your genome. Then it looks for those genes within your genome using HMMER, which uses profile hidden Markov models to identify sequence homologs. The completeness is an estimate of the fraction of genes that are expected to be there which are actually found. The contamination of based on identifying the number of single copy genes that should only be there once.

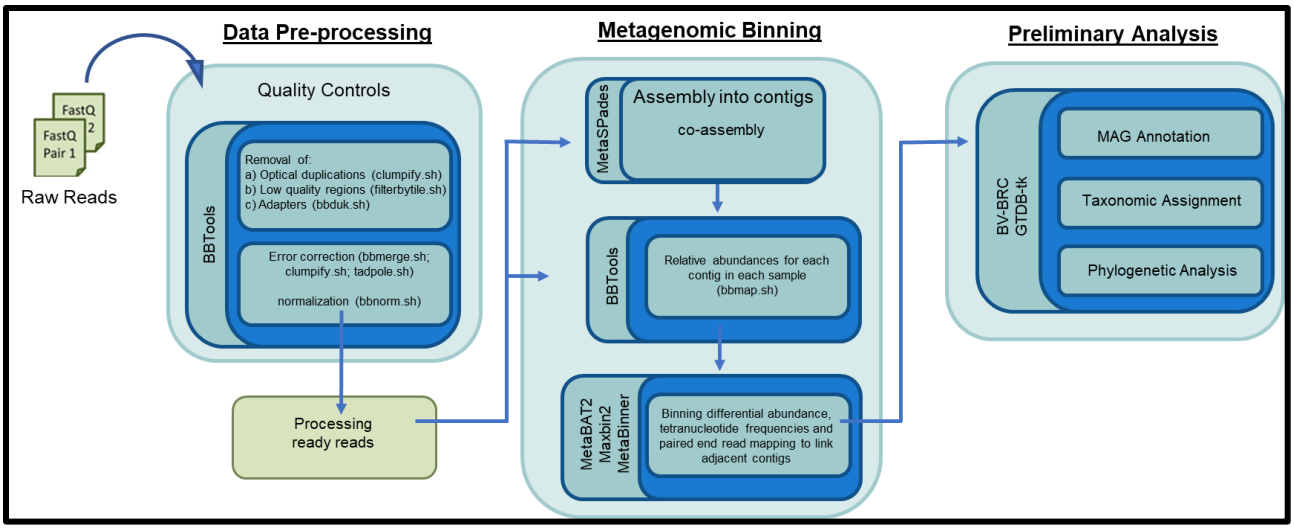


Figure 2. Metagenomic pipeline used in the analysis.

2.2.3 MAG ANNOTATION AND TAXONOMIC ASSIGNMENT

MAGs were annotated using the whole genome analysis tool available from Bacterial and Viral Bioinformatics Resource Center (BV-BRC), formally called PATRIC, which is an online portal for genomes and metagenomics (Davis et al., 2020). BV-BRV uses the RAST toolkit (RASTtk) (Brettin et al., 2015) to identify open reading frames, translate open reading frames into amino acids sequences, compare amino acid sequences to databases of known proteins, and assign functional categories to each gene in the genomes. BV-BRC was also used to calculated other bin quality measurements like the N50 and L50. Contig N50 is the sequence length of the shortest contig at 50% of the total assembly length and captures how much of the genome is covered by large contigs. Contig L50 is the number of contigs that are equal to or longer than the contig N50 value. Therefore, bins with a high N50 and low L50 are considered to be of higher quality. Bins were tentatively classified using Minhash Sketching (sendsketch.sh) and the similar genome finder tool in BV-BRC. Taxonomic assignment was completed using the Genome Taxonomy Database (GTDB-tk; (Chaumeil et al., 2022)). GTDB-tk is a computation toolkit that provides taxonomic classification of bacterial and archaeal genomes by placing them into a domain-specific,

concatenated protein reference tree. Briefly, genomes are assigned to a domain (Bacteria, Archaea, or Eukarya) with the highest proportion of shared marker genes. The marker genes are then aligned using HMMER and concatenated into a single multiple sequence alignment (MSA) which is then placed into a domain-specific reference tree using pplacer (Matsen et al., 2010). The taxonomic classification of a query genome is based on its placement in the reference tree, as well as its Average Nucleotide Identity (ANI) to reference genomes. Phylogenetic analysis was performed using the phylogenetic tree tool in BV-BRC (50 conserved single-copy core genes, allowing for up to 10 duplicates and 10 deletions). The phylogenetic tree tool uses the codon tree method which analyzes aligned protein and coding DNA sequences of single copy genes using protein global families as homology groups using the RaxML program (Stamatakis, 2006). Trees were rooted and manually curated using iTOL (Letunic & Bork, 2007), which is an online tool to display, annotate, and manage phylogenetic trees.

2.2.4 BIOGEOGRAPHY MAPPING AND VISUALIZATION

Ocean Data View (Schlitzer, Reiner, Ocean Data View, odv.awi.de, 2023) was used in the creation of section plots for various parameters (temperature, salinity, dissolved oxygen concentrations, etc.) using the default DIVA gridding settings. Contours were added and colour mapping was auto adjusted. Pacific Water identification was performed using measured silicate concentrations and a linear combination of nitrate and phosphate concentrations ((N* = 0.87(N-16P+2.9); Gruber & Sarmiento, 1997). The latter equation is used to measure the distribution of nitrification and denitrification, where N denotes nitrate concentrations (μmol kg^{-1}) and P denotes phosphate concentrations (μmol kg^{-1}). In this equation, the constant (2.9 μmol kg^{-1}) was determined by forcing the global mean of N* based on the Geochemical Ocean Sections Study

(GEOSECS) data (1972-1978) to be zero. This equals the value that would be calculated using the GEOSECS-based global mean nitrate of 30.38 μmol kg^{-1} and global mean phosphate of 2.08 μmol kg^{-1}. This equation uses values of stoichiometric N:P ratios for the mineralization of organic matter (the Redfield ratio of 16:1; Redfield et al., 1963). This ratio is derived from the global average of biomass in the oceans due to primary and secondary productivity. Lower values indicate nitrate depletion relative to Redfield ratios due to microbial denitrification processes that reduce nitrate back to nitrite. Nitrate can then be utilized by microbes as an energy source to produce nitric oxide (NO) and nitrous oxide (NO_2), both gases that are poorly soluble in water and are consequently released into the atmosphere. Therefore, a lower N* value represents a state where the nitrate depletion is greater relative to what would be expected if nitrate was only being utilized for the production of biomass. The depth of the PML was identified by locating the temperature minimums of each cast. Temperature and salinity measurements were used in the determination of all other water masses following Pisareva et al. 2015).

R Studio (R Core Team, 2021) were used to map the distribution of taxa in relation to environmental parameters such as depth, temperature, salinity, light, and nutrients. Relative abundance was used in this analysis, which is a measure of the proportion of total microbial abundance and does not capture the absolute cell counts or abundances. Therefore, an increase in relative abundance could be caused by a decrease in the abundance of other organisms while the absolute counts may remain the same throughout the water column. Statistical analyses were done in R using the VEGAN package Oksanen, et al., 2022).

2.3 RESULTS

2.3.1 WATER MASS DETERMINATION AND PHYSICAL PARAMETERS FROM THE 2015 GEOTRACES EXPEDITION

The GEOTRACES 2015 Expedition (GN01) took place in late summer departing from Alaska. In total 66 stations were measured for temperature, salinity, depth, and nutrient concentrations. Of the 66 stations, the deepest were stations 30 (3951 m), 32 (4239 m), 34 (4221 m), and 56 (4394 m), which were closest to the North Pole. The shallowest sampling depths were from stations 3 (33 m), 2 (49 m), and 66 (34 m) along with other stations from the southern coastal regions and the Chukchi Sea. Stratification due to temperature and salinity was observed in the northern regions, however, it was more difficult to detect stratification in the southern coastal regions of the Bering Strait and Chukchi Sea using the same metrics.

A linear combination of nitrate and phosphate concentrations (Gruber & Sarmiento, 1997) was used to measure the distribution of nitrification and denitrification, as described in Chapter 2.2.4. N* measurements were used to identify the Pacific Ocean water coming through the Bering Strait and over the shelf (Figure 3). Negative N* values are indicative of a sink of N* due to benthic denitrification and indicate Pacific Ocean water coming into the Western Arctic Ocean through the Bering Strait and Chukchi Sea (Deutsch et al., 2001; Gruber & Sarmiento, 1997) and can be seen between approximately 25 m and 150 m up to 84 ^{o}N. In the southern coastal areas (>75 ^{o}N), there was an increase in the N* values observed in the uppermost 25 m. This tool for determining water masses is not effective in surface waters due to the added influence of other microbial processes that interfere with nutrient content. Therefore, in the surface waters, temperature was used as the main determining factor of water mass differentiation. Another common tracer of Pacific Ocean water entering the Arctic Ocean is elevated silicate concentrations (Jensen et al.,

2019; Jones & Anderson, 1986). Silicate concentrations (Figure 4) were observed to be elevated in the same region as N* values were lower. It was observed that the low silicate surface layer corresponded with higher N* values, increased temperature, and decreased salinity in the upper 25–50 m. The temperatures in this region were much warmer, indicating the summer surface layer (SSL). Temperatures, which ranged from -1 to 14°C, were warmest in the southern regions (Figure 5) and displayed a thermocline in the northern regions (>75 ^{o}N). The temperature minimums for each cast were used to determine the bottom of the Polar Mixed Layer (PML). The PML near the North Pole was colder than its underlying water masses (~ -1°C).

Since the PML is often derived from Pacific water moving into the Arctic in the study region, they will be classified together as the PML in this analysis. Salinity ranged from 25.01 PSU – 35.96 PSU throughout the region (Figure 6). This designation is appropriate as the PML salinities observed here were similar to Jensen et al. 2019. The water below the PML consisted of the more saline Atlantic water and was observed approximately between 200 m and 1000 m. For this analysis, both the Fram Strait Branch (FSB) and the Barents Sea Branch (BSB) have been grouped together. The water below the BSB was characterized as the Arctic Ocean Basin water. Using temperature, salinity, silicate, and N* measurements the water masses observed were classified as SSL, PML, Atlantic water (AW), and Arctic Ocean basin water (ABW) (Figure 7).

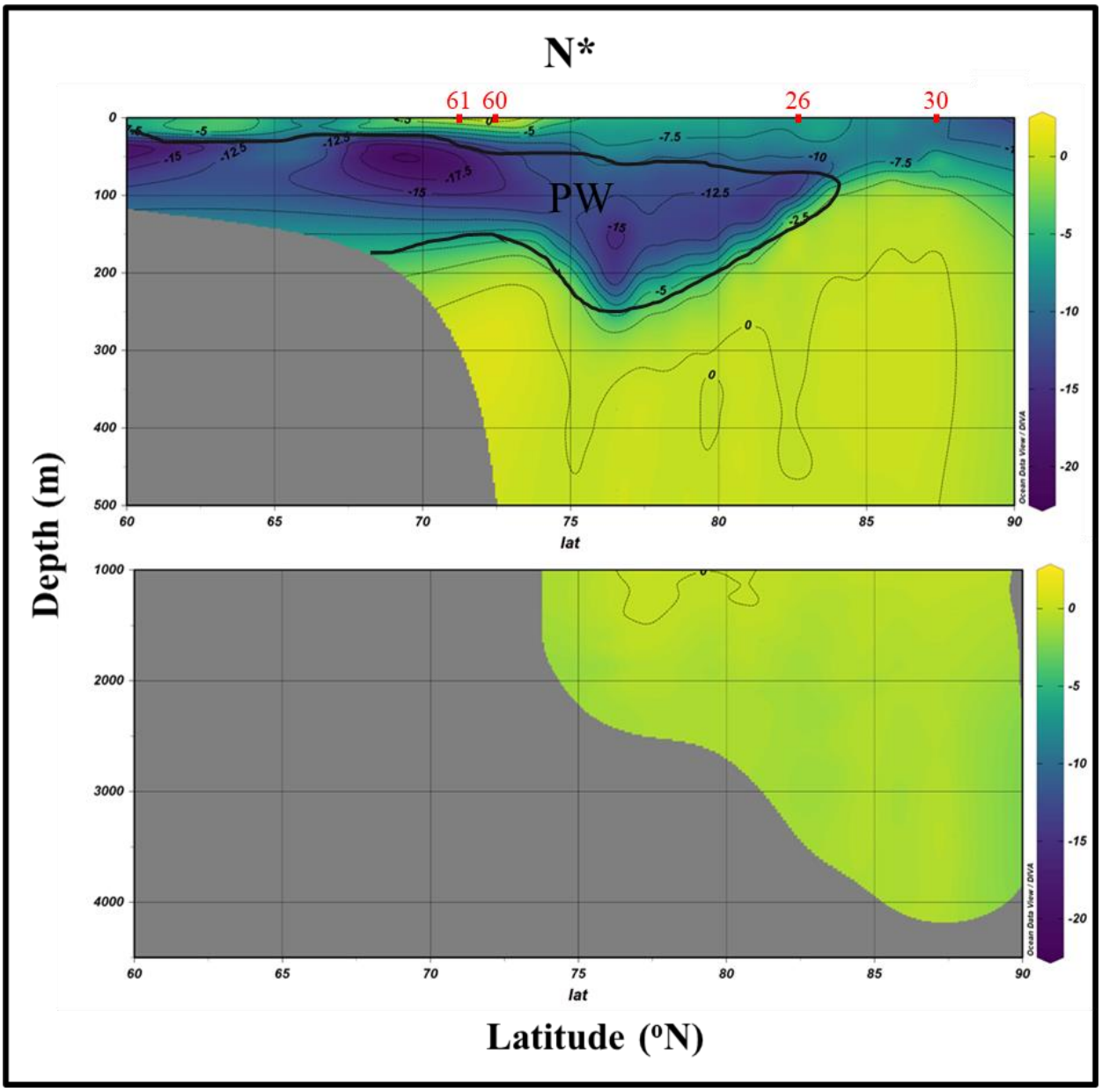


Figure 3. Section plot of N* values for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01). Pacific Ocean water is indicated in dark blue as N* values are expected to be lower. The inferred Pacific Ocean water layer is highlighted in black. Values were gridded using weighted-average gridding in ODV. Metagenomic sampling data was collected at stations listed at the top (red).

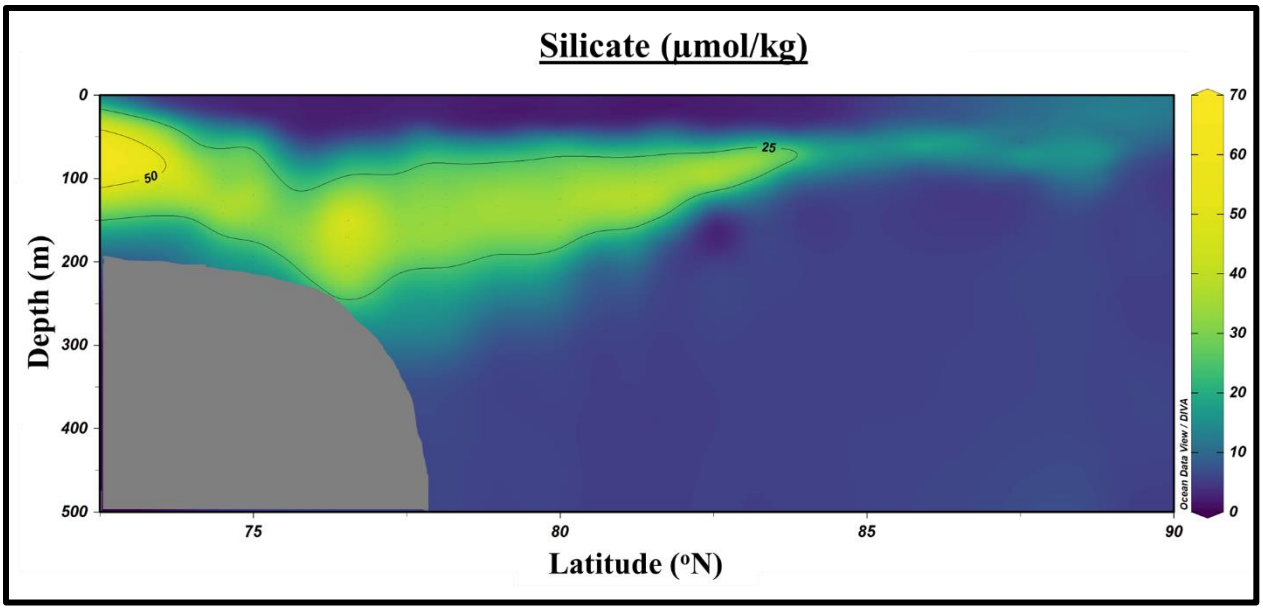


Figure 4. Section plot of silicate measurements for the upper 500 m of the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01). Values were gridded using weighted-average gridding in ODV.

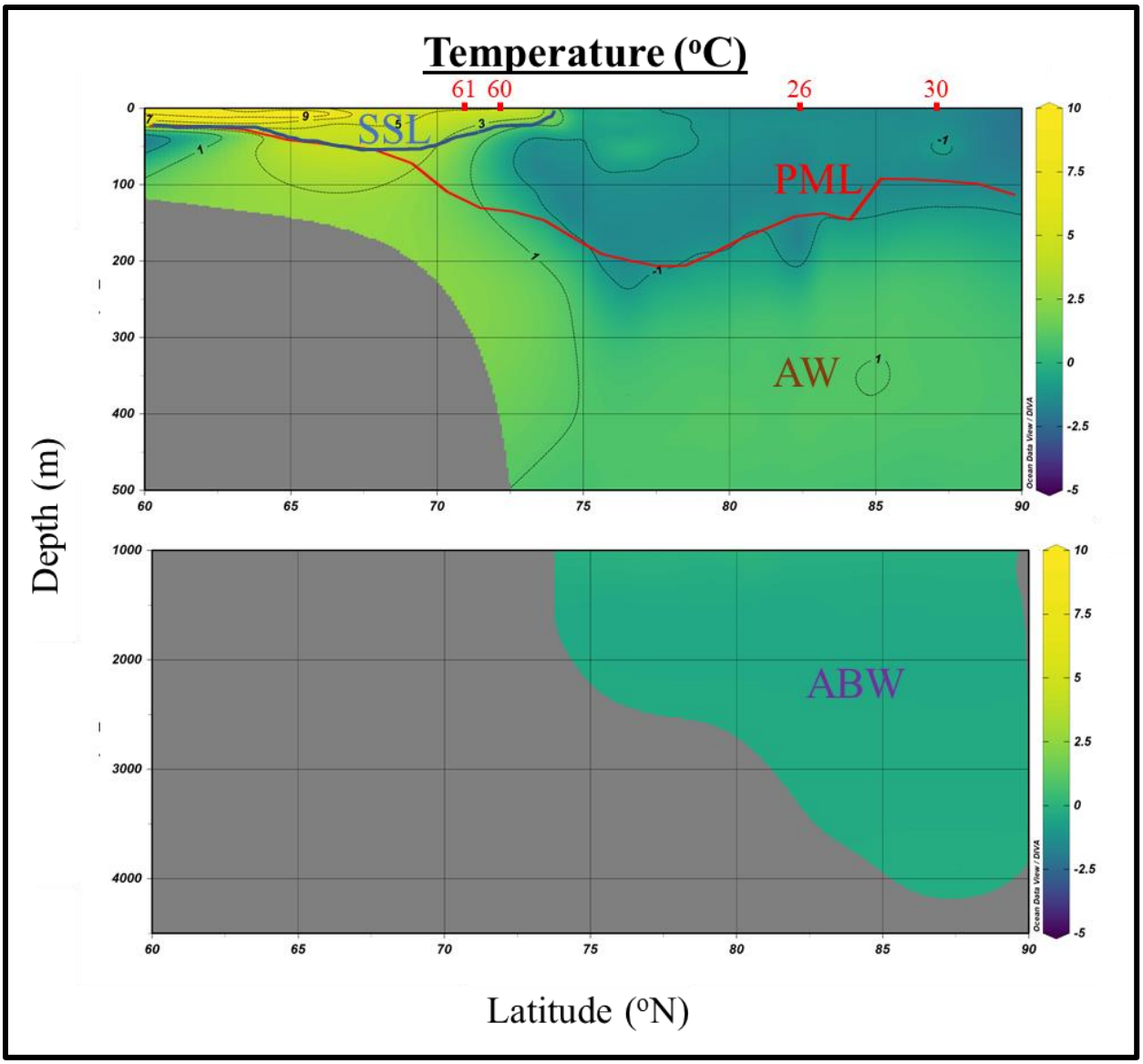


Figure 5. Section plot of temperature measurements for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01). Temperature minimums for each cast were used to determine the bottom of the Polar Mixed Layer (PML; red). The summer surface layer (SSL; light blue) is derived from warmer temperatures in the south. The Atlantic water (AW; brown) is warmer than the PML above and warmer than the Arctic Ocean Basin water (ABW; purple) layer below. Values were gridded using weighted-average gridding in ODV. Metagenomic sampling data was collected at stations listed at the top (red).

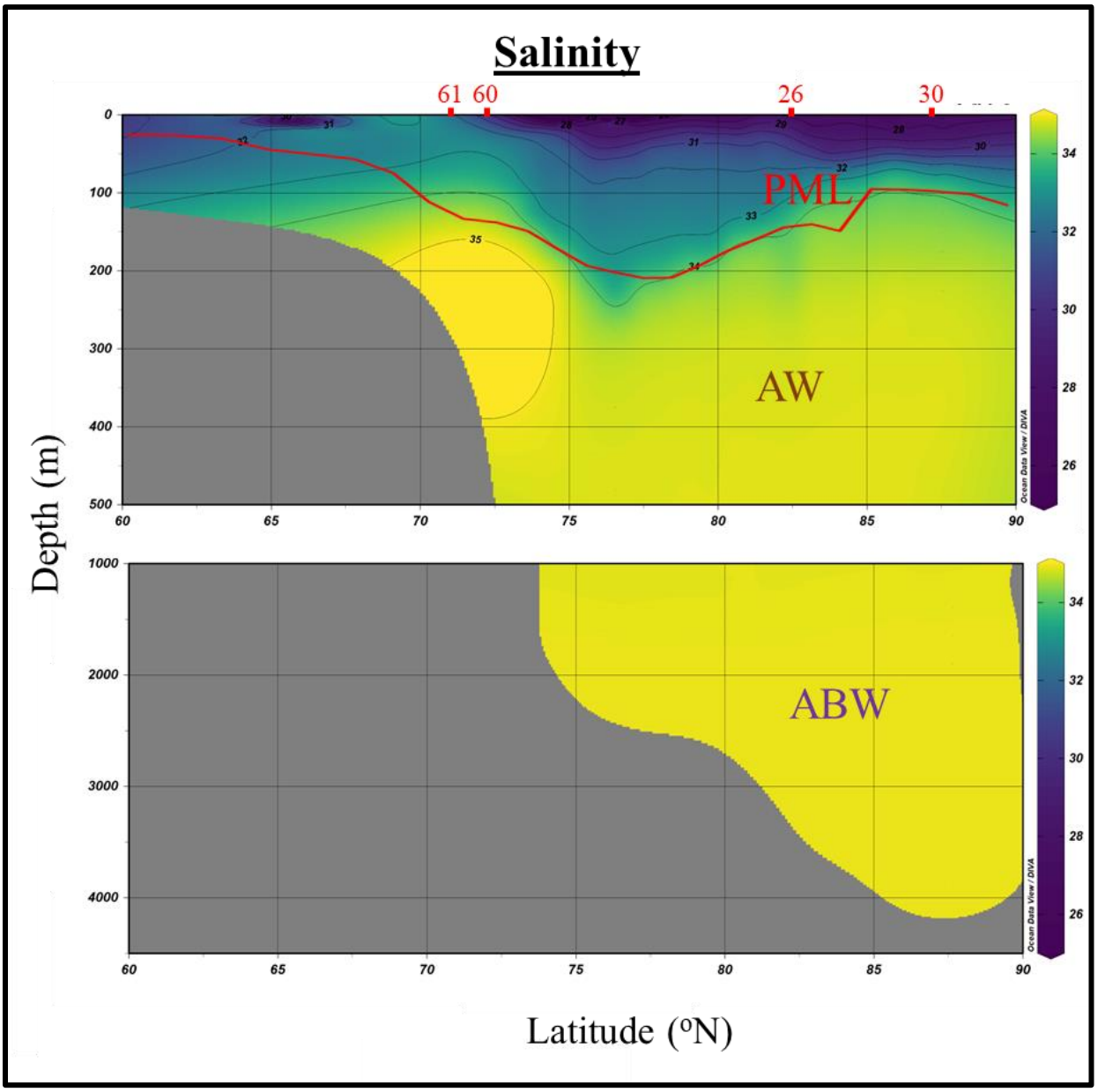


Figure 6. Section plot of salinity measurements for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01). Temperature minimums for each cast were used to determine the bottom of the Polar Mixed Layer (PML; red). The Atlantic water (AW; brown) and Arctic Ocean Basin water (ABW; purple) are much more saline than the PML above. Values were gridded using weighted-average gridding in ODV. Metagenomic sampling data was collected at stations listed at the top (red).

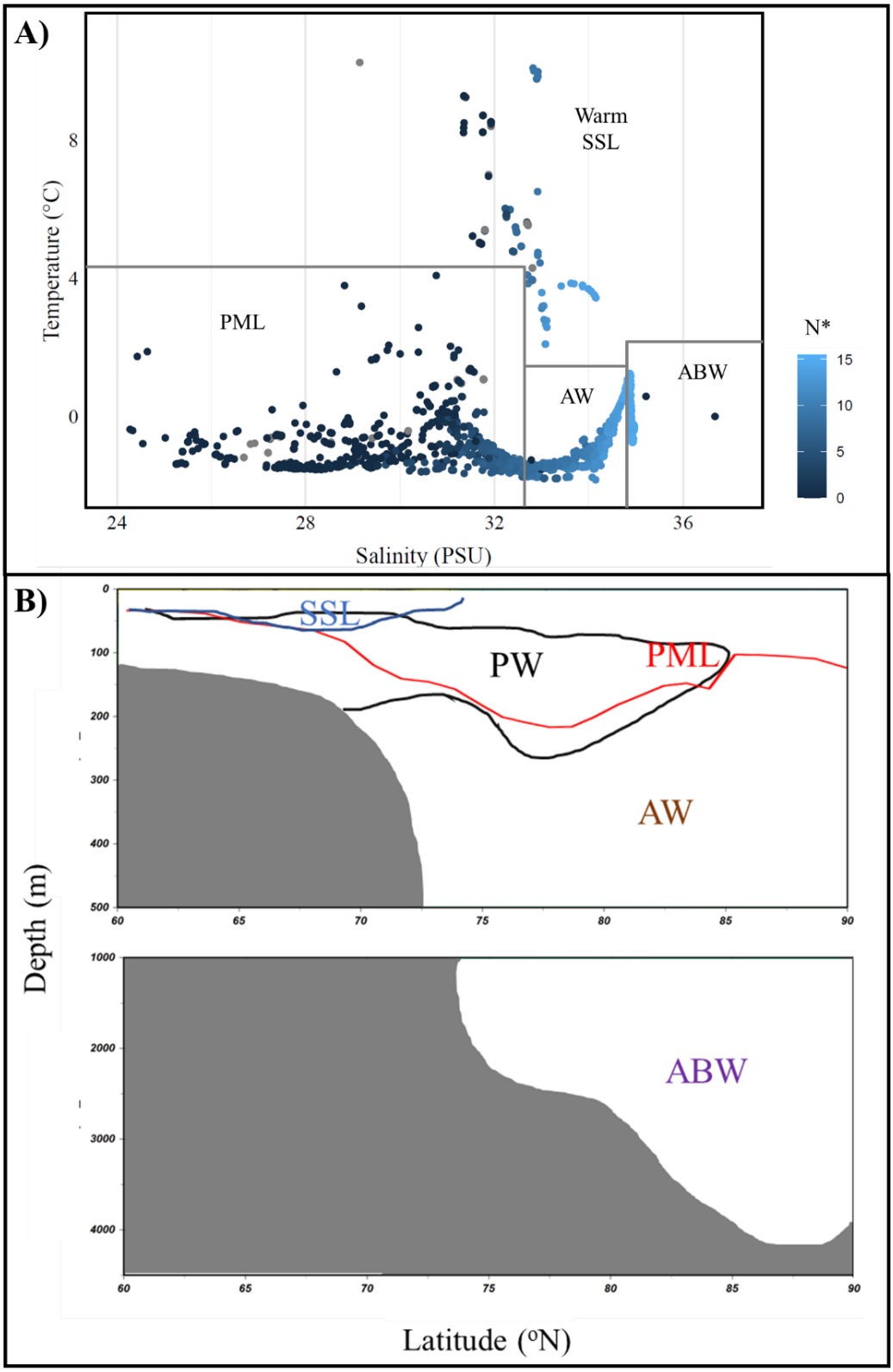


Figure 7. Water mass determinations: A) Temperature-Salinity plots of water sampled; B) Section plot of estimated water layers during the GEOTRACES 2015 expedition (GN01) designating the water masses observed.

Dissolved oxygen concentrations ranged from 167.11 to 572.40 µmol kg^{-1} and were lower in the southern coastal regions, likely caused by increased heterotrophic microbial activity, as well as in the ABW (Figure 8). The upper 50 m of the higher latitude stations had higher concentrations of dissolved oxygen, suggestive of oxygen production by photosynthesis and uptake from the atmosphere. Nitrate ranged from 0 (below detection) to 17.6 µmol L^{-1} and nitrite ranged from 0 (below detection) to 1.03 µmol kg^{-1}. Nitrate and nitrite concentrations were lower in the surface waters for stations below 75°N and increased below 20 m (Figures 9 and 10). Nitrite concentrations were lower than nitrate at all stations and were essentially zero in the ABW. Ammonia concentrations ranged from 0.002 to 8.700 µmol kg^{-1}. Phosphate ranged from 0.168 to 2.816 µmol kg^{-1}. Silicate ranged from 0.6 to 61.2 µmol kg^{-1}.

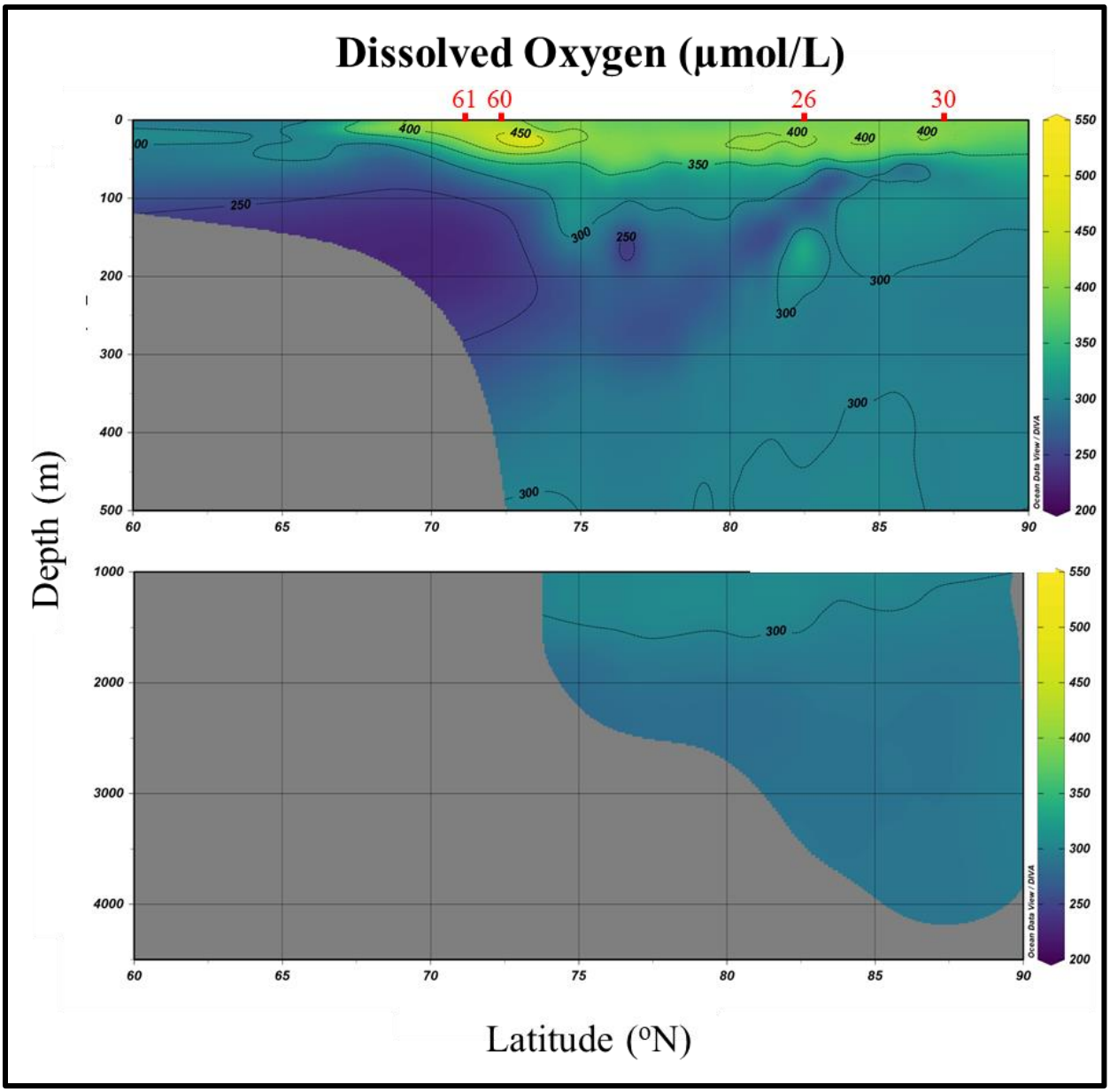


Figure 8. Section plot of dissolved oxygen concentrations for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01). Values were gridded using weighted-average gridding in ODV. Metagenomic sampling data was collected at stations listed at the top (red).

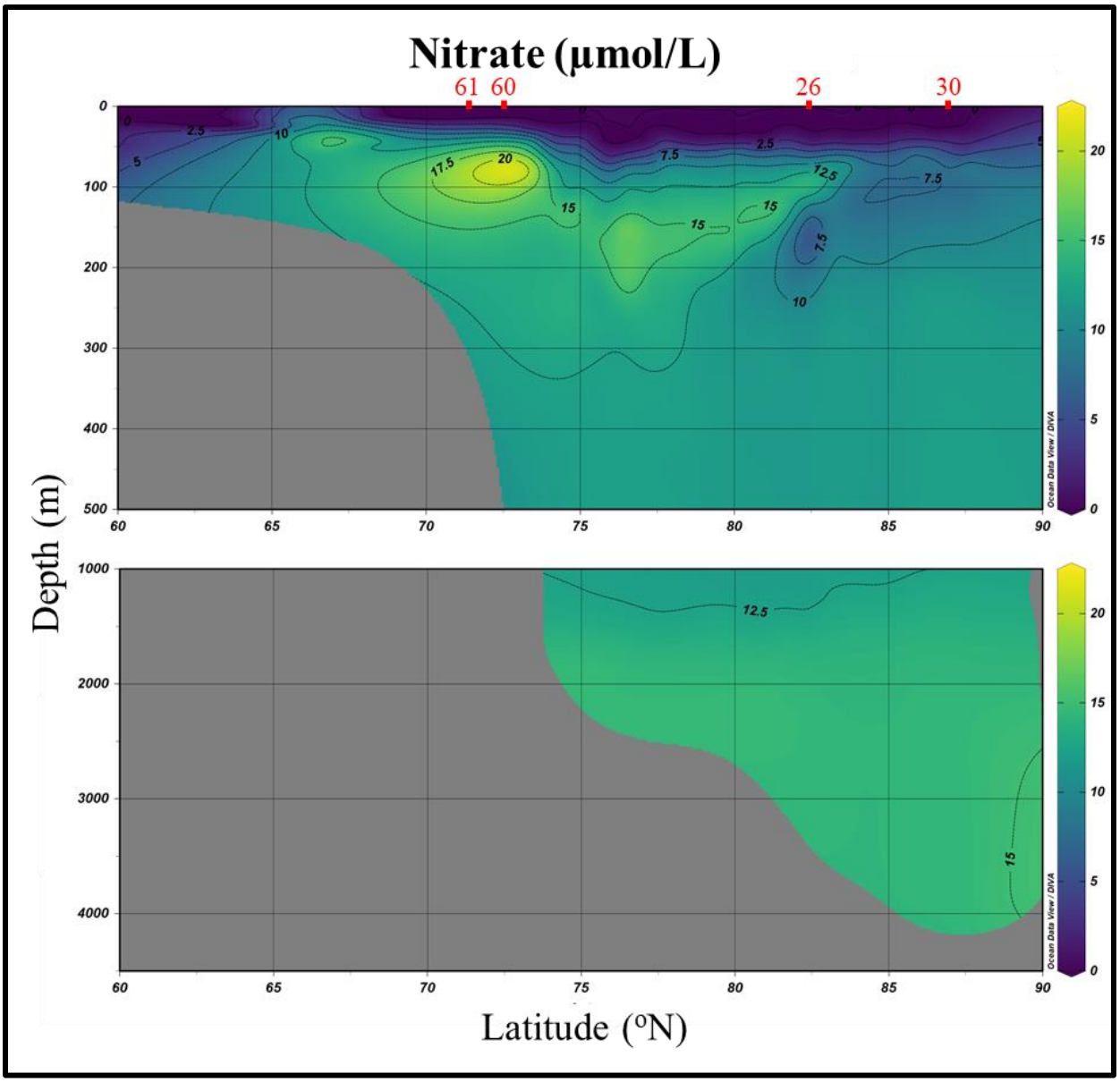


Figure 9. Section plot of nitrate concentrations for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01). Values were gridded using weighted-average gridding in ODV. Metagenomic sampling data was collected at stations listed at the top (red).

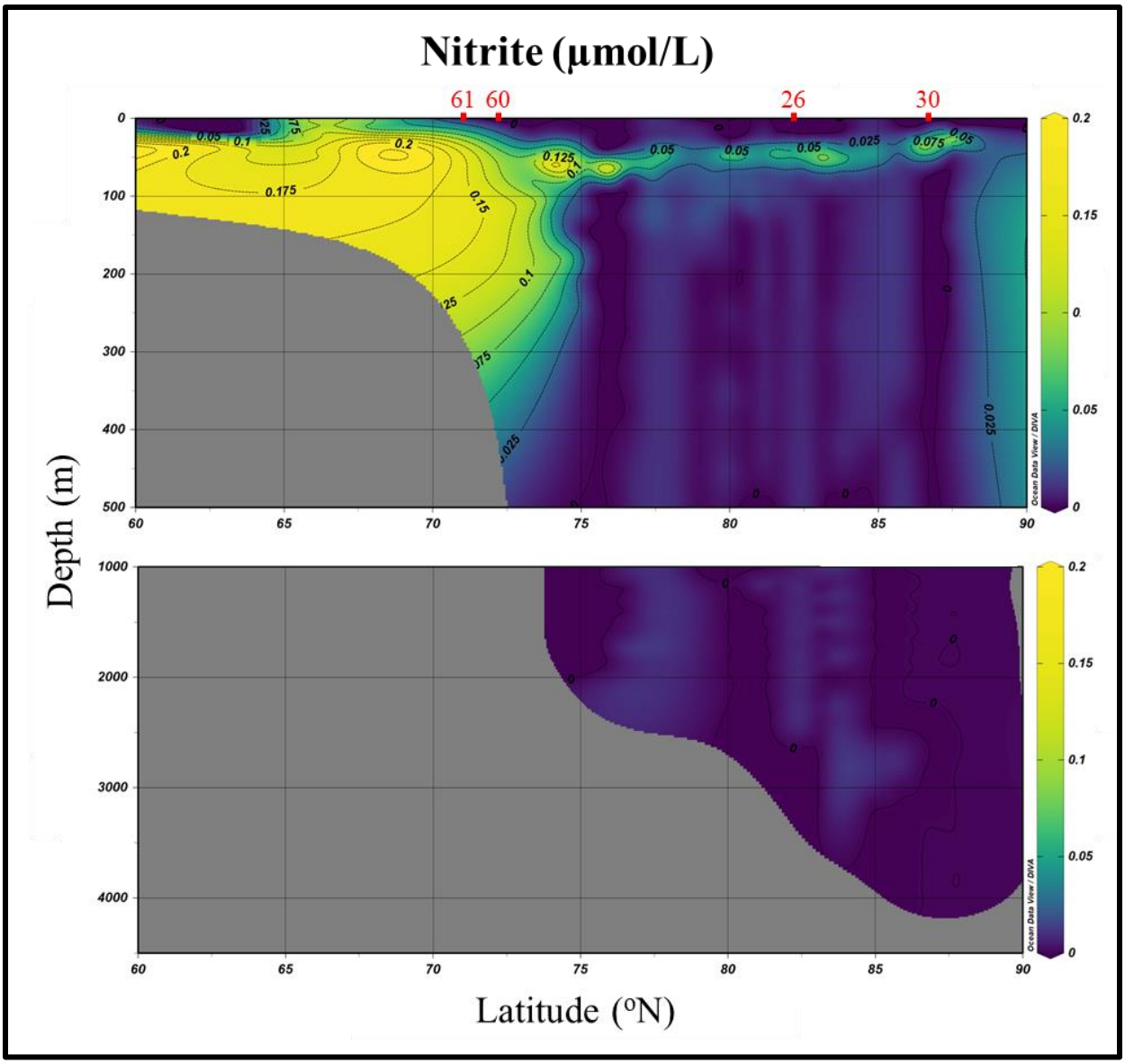


Figure 10. Section plot of nitrite concentrations for the Western Arctic Ocean during the GEOTRACES 2015 expedition (GN01). Values were gridded using weighted-average gridding in ODV. Metagenomic sampling data was collected at stations listed at the top (red).

2.3.2 ARCHAEAL MAG IDENTIFICATION

Metagenomic data from stations 26, 30, 60, and 61 were used for the identification of archaeal taxa as described in Chapters 2.2.1 to 2.2.3. MetaSpades was used to produce 128 co-assemblies resulting in 2,735,798 contigs. Contigs smaller than 1500 bp were removed resulting in 224,116 contigs (867,388,435 bp) that were used for downstream processing. Three

metagenomic binning algorithms (Metabat2; METAbinner; Maxbin2) were used to identify bins that were consolidated using DASTools. A total of 155 bins with an average size of 2,312,350 bp (range: 211,901 to 14,059,051 bp) were recovered. Using GTDB-tk for taxonomic assignment, 17 of the 155 bins were determined to be of archaeal origin (Table 2). Bins were identified in each of the three marine groups, although MG-II (Poseidoniales) contained the greatest amount of richness (11 bins). Bin Quality was measured using CheckM (Table 2). CheckM completeness and contamination scores of > 90% and < 5% ("high-quality draft" as defined by the MIMAG standard, https://pubmed.ncbi.nlm.nih.gov/28787424/), were only met by bin 35_sub and bin 29, the remainder were classified as "medium-quality drafts" (> 50% completion, < 10% contamination), "low-quality drafts" (< 50% complete, < 10% contamination), or simply drafts (all others).

Table 2. Quality scores for the archaeal MAGs determined using CheckM.

Marine Group	Bin ID	Completeness	Contamination
MG-III	35_sub	93.56	1.07
	105_sub	74.81	9.6
MG-II	maxbin.078_sub	84.36	4.4
	55_sub	82.49	0
	54	82.36	0.8
	maxbin.001_sub	82.23	2.46
	63	81.24	0
	58	79.33	0
	82_sub	78.95	1.93
	118	73.92	5.65
	108	63.88	1.92
	145_sub	59.05	3.6
	161	50.92	1.85
MG-I	29	93.69	0.32
	maxbin.015_sub	76.38	20.49
	maxbin.074_sub	71.77	13.79
	166_sub	57.58	6.3

The BV-BRC Comprehensive Genome Analysis tool was used to annotate the bins (Table 3). Bin 54 (MG-II) contained the fewest number of contigs (73), while bin 105_sub (MG-III) and bin 108 (MG-II), both contained the greatest number of contigs (530). The average genome size was 1,470,528 bp, which was within the normal size range for archaeal genomes. The smallest genomes were observed in the MG-I (maxbin.074_sub (908,512 bp); 166_sub (898,496 bp); 29 (956,250 bp)), while the largest was observed in MG-II (Maxbin.078_sub (2,434,869 bp). Contig N50 and L50 values were also calculated as a measure of bin quality. Bin 35_sub, bin 54, and bin 63 displayed the most desirable values (N50 > 20,000; L50 < 20) which were fairly consistent with the quality scores from CheckM. Lower quality bins determined by CheckM (145_sub, 161, and 166_sub) consequently displayed poorer N50 and L50 scores (N50 < 4000; L50 > 100 contigs). The number of coding sequences (CDSs) ranged from 1090 (bin 29) to 2395 (bin maxbin.078_sub), consistent with the genome sizes mentioned above. The lowest genomic GC percentages were observed in the MG-I MAGs (maxbin.074_sub (30.93%); 166_sub (33.31%); maxbin.015_sub (34.70); 29 (34.73%)) but are well within the expected bounds of microbial GC content, which is typically between 30-50% GC. Both MG-III bins (35_sub (64.23%); 105_sub (60.59%)) had very high GC content, alongside members of MG-II (55_sub (59.05%); 54 (59.80%); maxbin.001_sub 60.59%); 63 (55.33%); 118 (59.81%); 108 (60.59%)). Phylogenetic analysis was performed on the archaeal bins (Figure 11). The phylogenetic analysis was consistent with the taxonomic assignments for all MAGs.

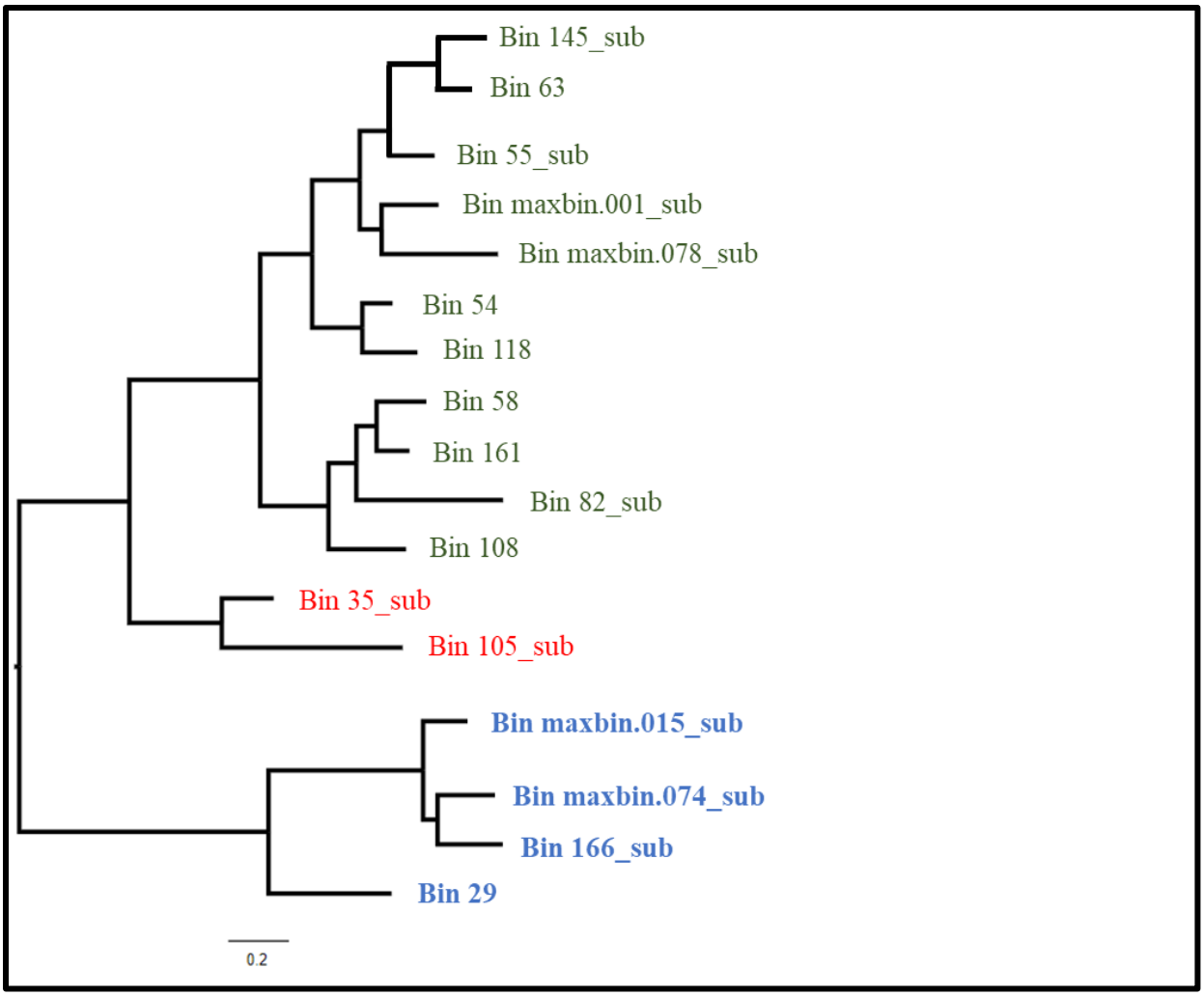


Figure 11. Phylogenetic tree of the archaeal MAGs identified in the Western Arctic Ocean during the 2015 GEOTRACES expedition (GN01). Clades are identified as MG-I (blue), MG-II (green), and MG-III (red).

Table 3. Archaeal MAG characteristics as described by BV-BRC whole-genome analysis tool from the 2015 GEOTRACES cruise.

Bin	Genome Length (bp)	GC content (%)	Contigs	Contig N50	Contig L50	CDS
35_sub	1292743	64.23	117	28161	13	1309
105_sub	1196390	60.59	530	2914	173	1623
Maxbin.078_sub	2434869	44.39	405	18006	23	2395
55_sub	1364045	59.05	110	18854	23	1256
54	1857474	59.80	73	36416	17	1561
maxbin.001_sub	1541469	55.96	152	21754	22	1465
63	1400658	55.33	133	22859	17	1277
58	1775200	52.89	113	22387	25	1535
82_sub	1956988	44.54	217	13314	43	1757
118	1886358	59.81	501	4367	140	1961
108	1486156	60.59	530	2914	173	1623
145_sub	1362659	56.59	422	3500	123	1412
161	1508422	58.12	522	3005	163	1734
29	956250	34.73	104	13853	21	1090
maxbin.015_sub	1172288	34.70	442	2698	147	1625
maxbin.074_sub	908512	30.93	326	2912	103	1360
166_sub	898496	33.31	366	2566	125	1263

2.3.3 ARCHAEAL RELATIVE ABUNDANCE AT STATIONS 26 AND 30

To better understand abundance trends as a function of depth, abundance data for all bins was analyzed at stations 26 and 30. These stations were chosen for further analysis because archaeal relative abundance was high at these two stations, the water columns were deep, and the clearest distinction of water masses were observed here. The temperature minimums were observed at ~100 m for both stations (station 26: -1.37 ^{o}C; station 30: -1.44 ^{o}C) which marked the bottom of the PML (Figure 12). The temperature increased throughout the AW layer to a maximum at 450 m (~1 ^{o}C), which began to decrease again moving into the BSB and ABW. Salinity was low at the surface (29 PSU) and increased until it reached a maximum at 300 m (~35 PSU). Nitrite (NO_2^-) concentrations were much lower than nitrate (NO_3^-) concentrations throughout the depth

profile (Figure 13). The relationship between the two was inversely proportional, where nitrite was higher at the surface and decreased with depth, and nitrate was lower at the surface and increased with depth. Nitrate concentrations reached a maximum at approximately 500 m for both stations (11.91 μmol L^{-1}, Station 26; 12.37 μmol L^{-1}, Station 30). Nitrite concentrations became undetectable below ~500 m for both stations.

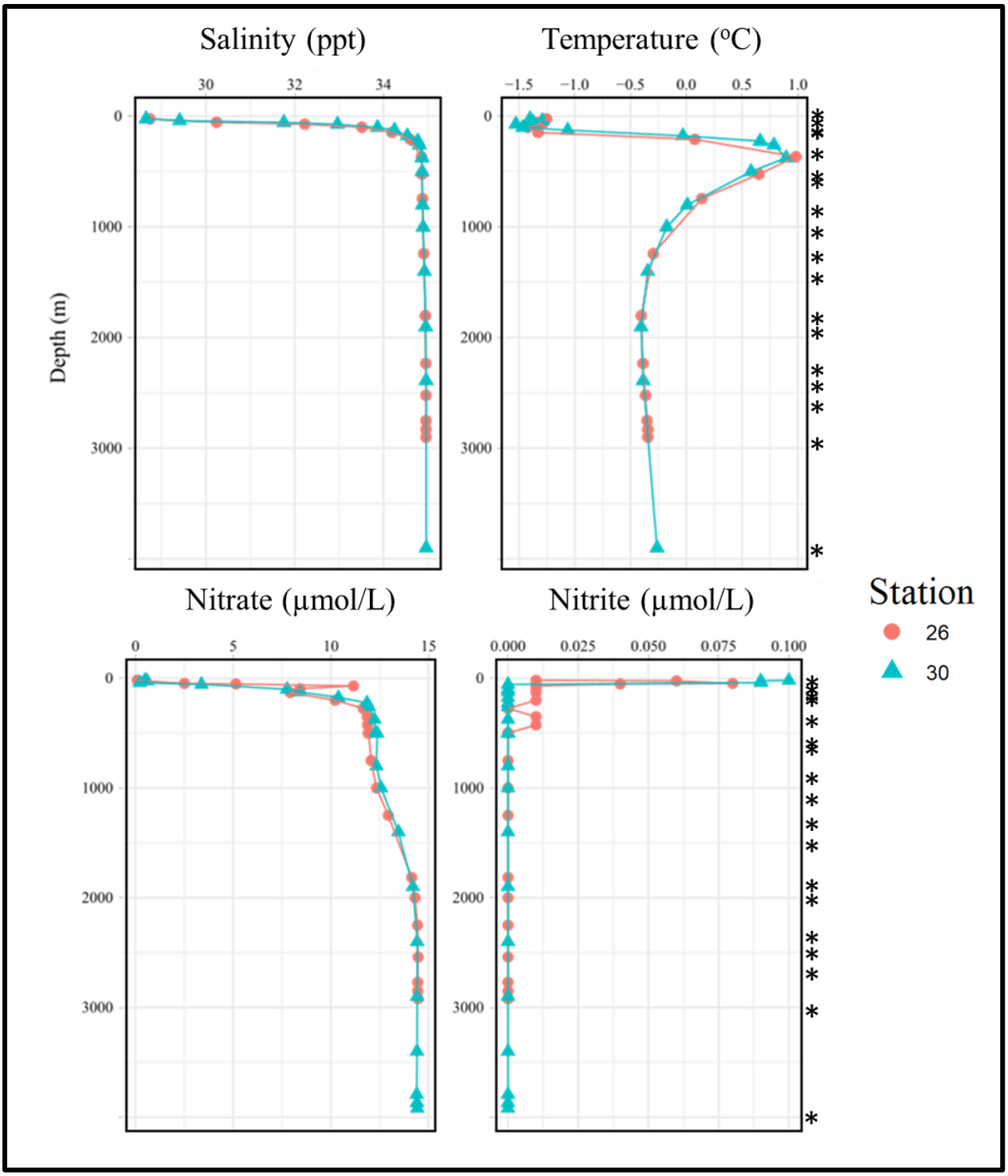


Figure 12. Depth profiles for salinity, temperatures, nitrate, and nitrite from the 2015 GEOTRACES expedition from stations 26 (red circles) and 30 (blue triangle). The depths of metagenomic sampling are indicated on the right (*).

Clustering based on relative abundance patterns of all MAGs (155) uncovered four distinct communities based on depth (Figure 13). Using the water mass determinations from Chapter 2.3.1. the sample clusters from different depths could be associated with a water mass: PML, AW, and ABW. The ABW cluster consisted of samples below ~1000 m. The exceptions were samples ARODd.30.8 and RjKKY.30.8 both collected from a depth of 226 m. The AW cluster contained samples from depths of ~200 m to ~1100 m. The FSB and BSB can be seen with the AW cluster which were separated by depths between ~200 to ~400 m and ~500 to ~1000 m, respectively. The final cluster, the PML, contained samples from ~25 m to ~200 m. It is important to note that these water mass distinctions are not rigid. Relative abundance was generally higher in the PML than the other layers for most bins, as was expected with the presence of phytoplankton at the surface.

The bins were hierarchically clustered and divided into four clusters for analysis (Figure 14). Most bins (35_sub, 105_sub, maxbin.078_sub, 63, 54, 58, 82_sub, 29, maxbin.074_sub, and 166) clustered into cluster A, with bins that displayed a higher relative abundance in the PML and AW layers. Bin 55_sub was found in cluster B which contained bins that displayed a mixed pattern of relative abundance across the samples. The final clusters (C and D) contained bins (maxbin.00.1_sub, 108, 118, 145_sub, 161, and maxbin.015_sub) with higher abundance in the ABW. Bin 118 was the only bin to cluster into cluster C which showed relatively high abundances at the very deep samples as well as some of the shallower samples collected. Bins from within the same phylogenetic lineage did not display similar abundance patterns at these two stations despite previously established distribution trends for the archaeal marine groups (Ionescu et al., 2009; Pereira et al., 2019; Rinke et al., 2019; Yilmaz et al., 2016).

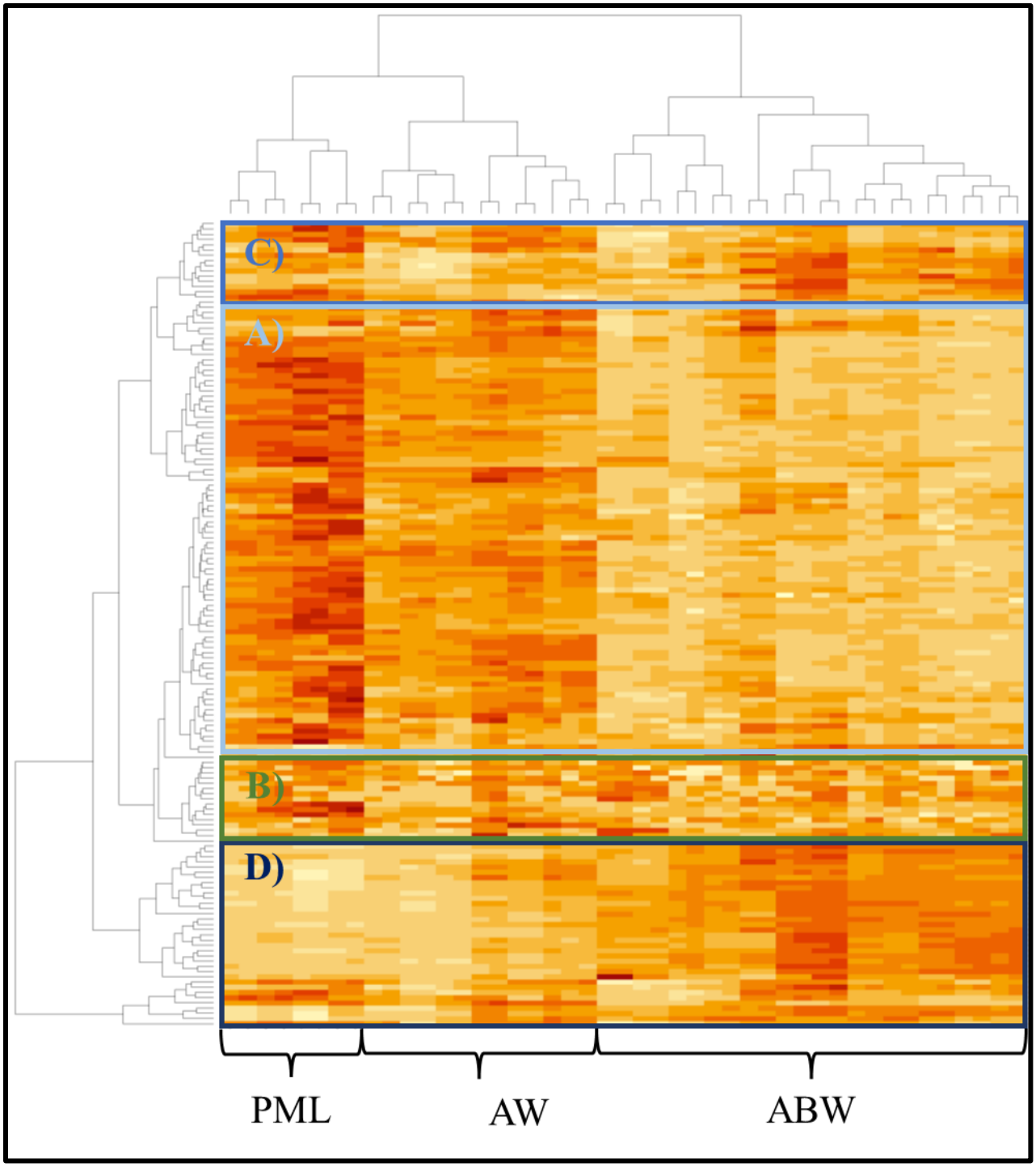


Figure 13. Heat map displaying abundance-based clustering of each bin (y-axis) and each sample (x-axis) at Stations 26 and 30 from the GEOTRACES 2015 expedition (GN01). Samples formed four main clusters based on depth: Arctic Ocean Basin water (ABW); Atlantic water (AW); Pacific water (PW); Polar Mixed Layer (PML).

Overall, archaeal relative abundance at these stations appeared to increase with depth (Figure 14A). The MAGs were also analyzed by the proportion of total archaeal relative abundance (Figure 14B). The MG-I MAGs made up the largest proportion, up to 80% of the community, in the upper 400 m. The MG-II MAGs made up the largest proportion at most depths, particularly below 500 m. The MG-III MAG abundance proportion was fairly consistent throughout the water column (approximately 10% of the archaeal community). Relative abundances of each MAG with respect to the over all microbial community was also investigated. It was observed that the MG-I bins were found in greater abundances in the upper 500 m (Figure 15A). The MG-II varied greatly in their overall abundance but did increase with depth (Figure 15B). Bin 82_sub was the most abundant from this group. MG-III were much more abundant below 800 m dominated by bin 35_sub (Figure 15C), however their overall relative abundances were much lower than the other two marine groups.

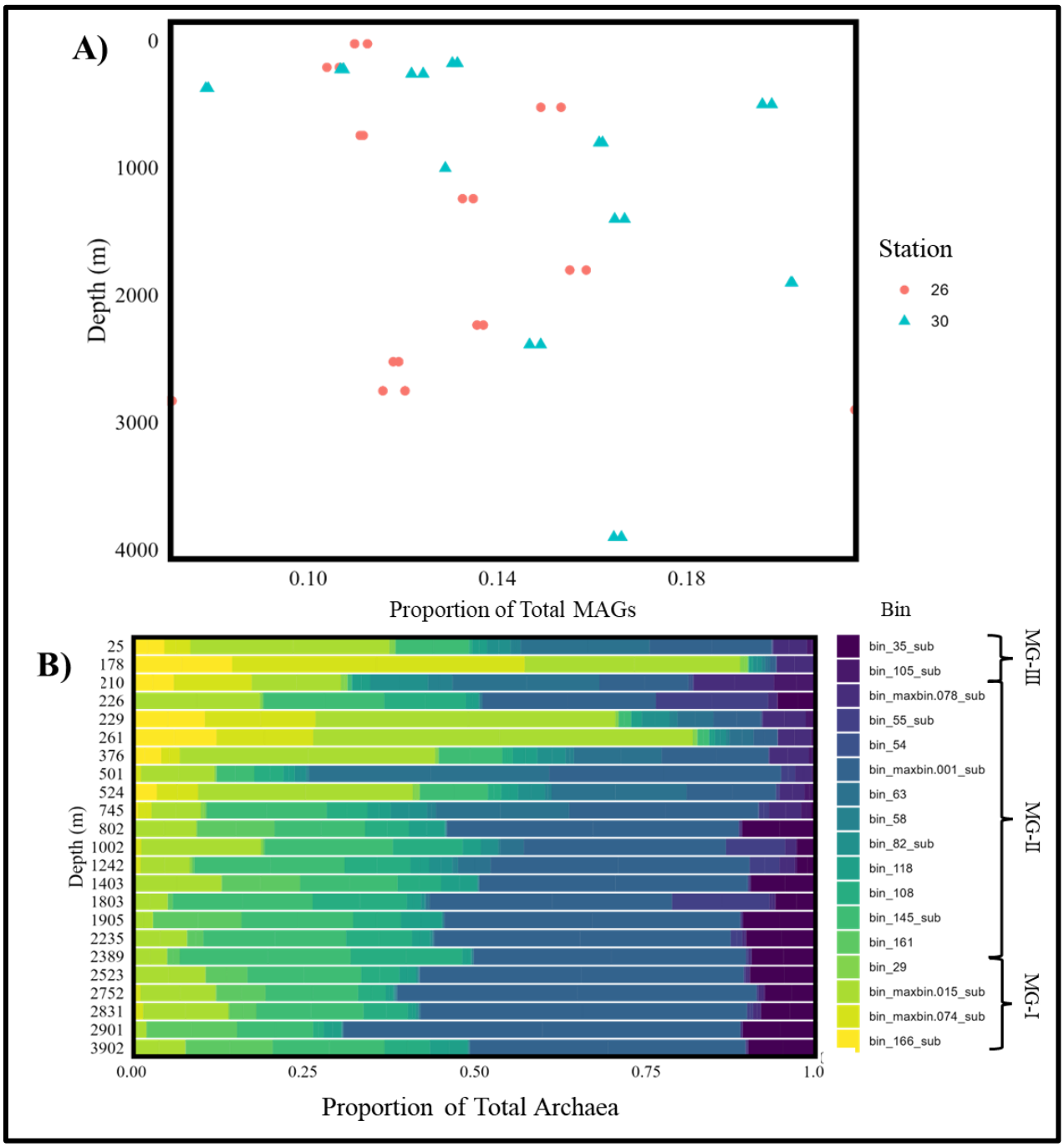


Figure 14. Relative abundance measurements for Archaea in the Western Arctic Ocean: A) total archaeal relative abundance as a proportion of total MAG abundance by depth; B) proportion of total archaeal abundance for all bins by depth.

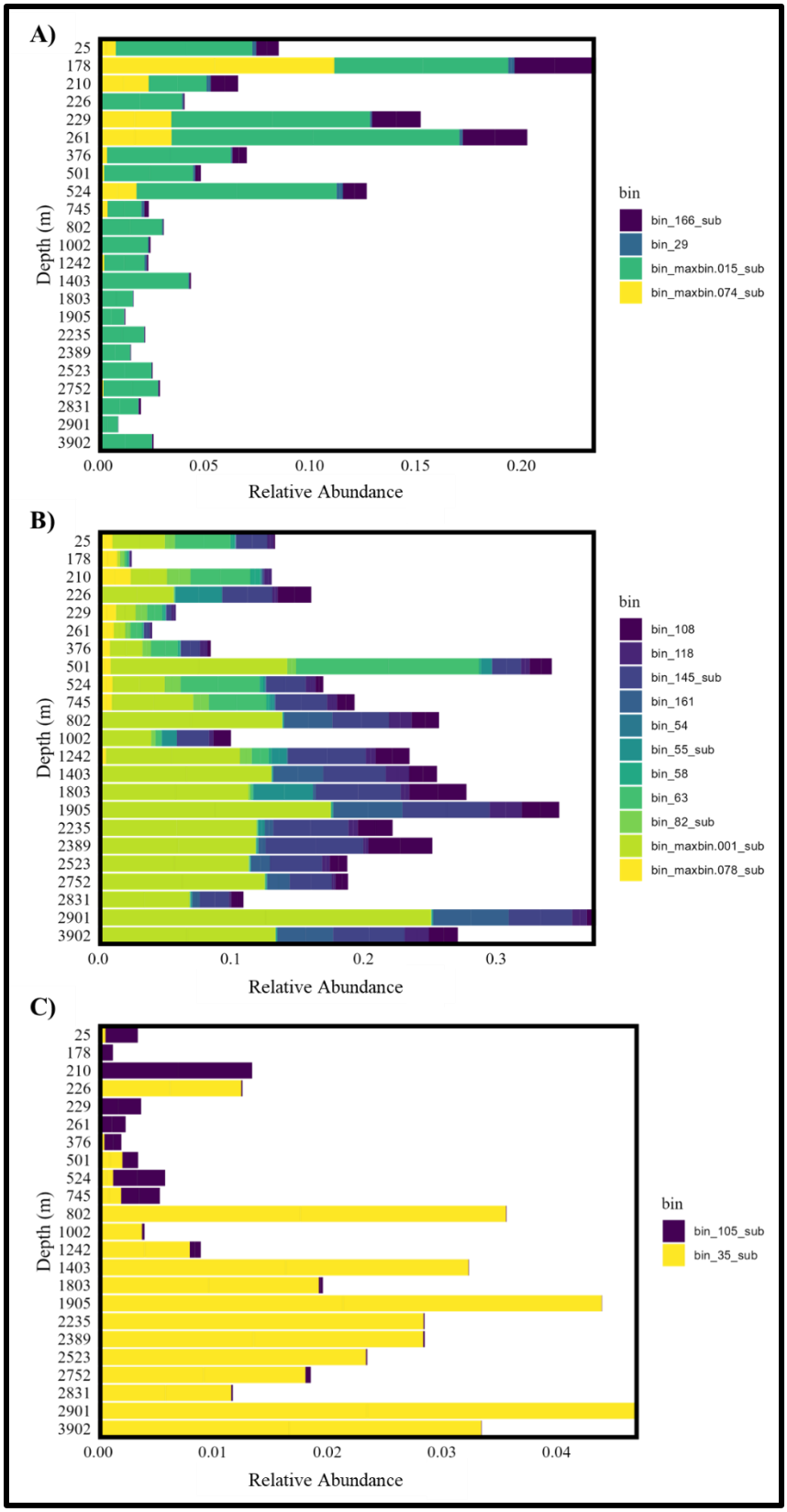


Figure 15. Archaeal relative abundance by depth for MG-I (A), MG-II (B), and MG-III (C) at station 26 and 30 from the 2015 GEOTRACES expedition (GN01).

2.3.4 ARCHAEAL BIOGEOGRAPHY IN THE WESTERN ARCTIC OCEAN

To further explore the archaeal distribution in the Western Arctic Ocean, the relative abundance of each bin was plotted by latitude and depth across a dataset of several hundred samples. Three main abundance distribution patterns were observed, which were variable within each marine group (Figures 16-20). These three abundance distribution ecotypes corresponded with the water masses determined in Chapter 2.2.1: PML, AW, and ABW. Again, these descriptions of water mass layers are not definitive, and thus the ecotypes were simplified to surface, mid-depth, and deep ecotypes.

The MG-I Archaea demonstrated surface and deep distributions (Figure 16). In fact, maxbin.074_sub had the highest relative abundance of the four bins and was found in greater amounts in the surface layers just north of the coast (Figure 15A). The abundances observed did correspond with regions with elevated nitrite concentrations (Figure 10). It was observed that the MG-II Archaea displayed all three distribution ecotypes: Bins 54 and maxbin.078_sub displayed a surface ecotype (Figure 17); Bins 58, 118, 82_sub, and 63 displayed a mid-depth ecotype (Figure 18); and Bins 55_sub, maxbin.001_sub, 161, 108, and 145_sub displayed a deep ecotype (Figure 19). Bin maxbin.078_sub was very abundant at the surface and was found at lower latitudes where water temperatures were much higher, as was expected. The mid-depth MG-II were lower in their abundances overall compared to maxbin.078_sub. In contrast, maxbin.001_sub exhibited comparable abundance measurements to the surface residing maxbin.078_sub while presenting a deep ecotype. The MG-III Archaea displayed both surface (bin 105_sub; Figure 21A) and deep (35_sub; Figure 20B) ecotypes. Interestingly, bin 105_sub, although surface residing, did display a different pattern of distribution where the abundance was greatest closer to the pole rather than the shelf.

The relative abundance distributions were compared to the phylogenetic tree structure (Figure 11) to see if the surface, mid-depth, and deep ecotypes were more closely related to one another. There was no consistency between the phylogenetic structure and the water profile ecotypes displayed by the MAGs. For example, maxbin.074_sub which was identified as a surface residing MG-I was not less related to the three other deep residing MG-I MAGs.

The majority of the associated ecotypes were consistent with the relative abundance clustering that was seen at stations 26 and 30 (Figure 13). There were three exceptions. Bin maxbin.015_sub (MG-I) clustered in Group D at stations 26 and 30 which were associated with higher relative abundances from the deeper water samples. Although this bin was determined to have a mid-depth ecotype, the abundances were quite high at 800 – 1000 m which roughly marks the bottom of the AW layer. This group was also found at moderate abundances in the ABW. The other two bins that had discrepancies were MG-II MAGs. Bin 118 was determined to have a mid-depth ecotype, however it clustered with Group C, which was associated with higher abundances at greater depths as well as higher abundances in some of the shallower samples. Bin 55_sub was identified as belonging to Group B, which displayed a mixed abundance pattern at the two stations. Although it was characterized to have a mid-depth ecotype, this mixed abundance behaviour was also seen in the biogeography as abundance was high in the 200 m range and 2000 m range. The addition of more southern stations (60 and 61) has elucidated the biogeography of these organisms that would have been missed with fewer stations.

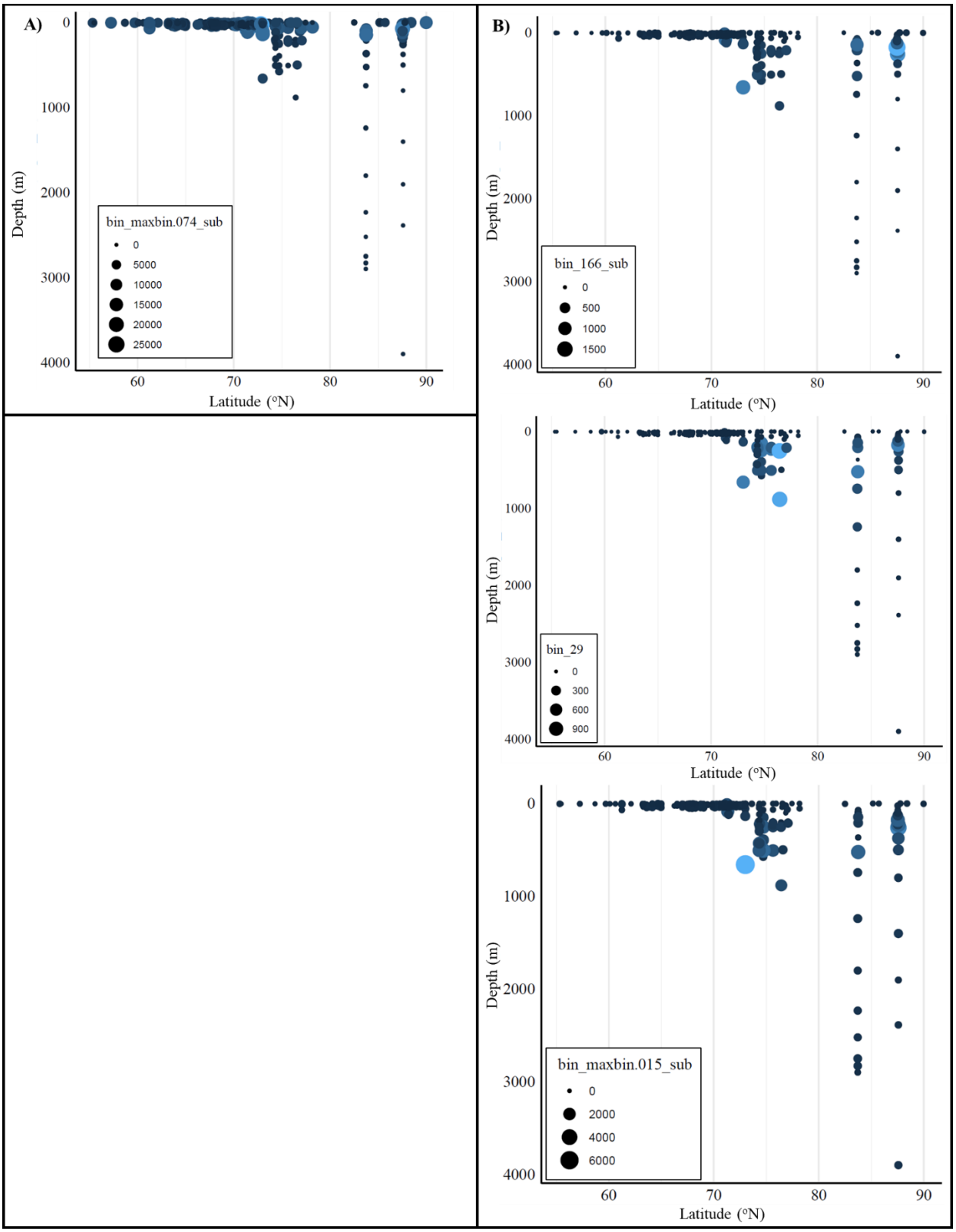


Figure 16. Relative abundance distributions (x1000000) MG-I archaeal MAGs identified in the Western Arctic Ocean displaying a surface ecotype (A) and a mid-depth ecotype (B). Colours correspond with the abundances (light blue higher relative abundance and dark blue is lower relative abundance.

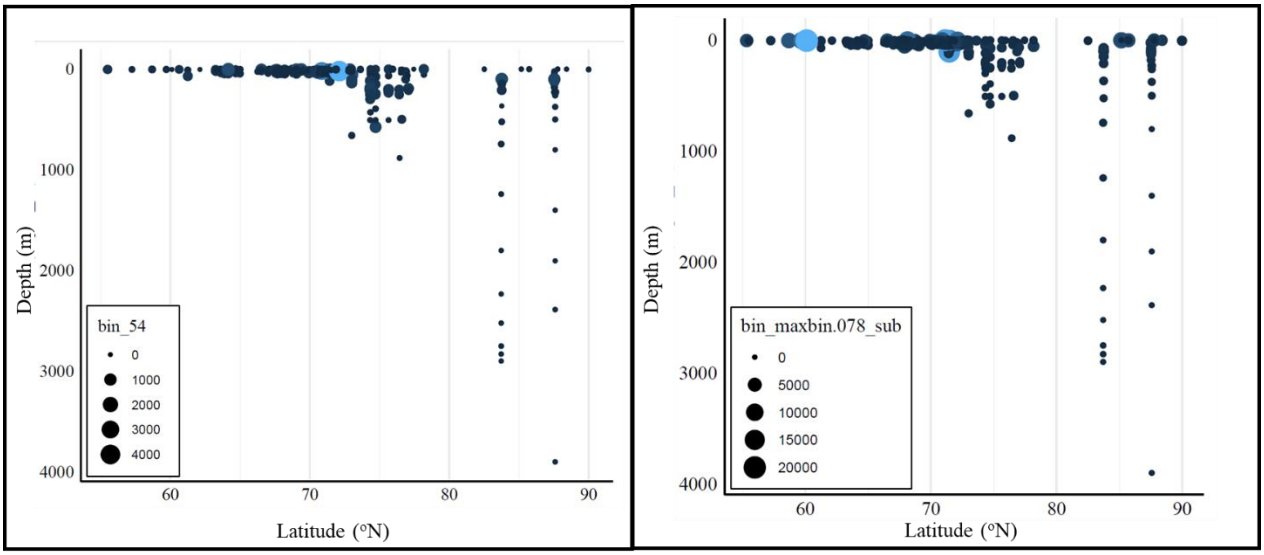


Figure 17. Relative abundance distributions (x1000000) of the MG-II archaeal MAGs identified in the Western Arctic Ocean displaying a surface ecotype. Colours correspond with the abundances (light blue higher relative abundance and dark blue is lower relative abundance.

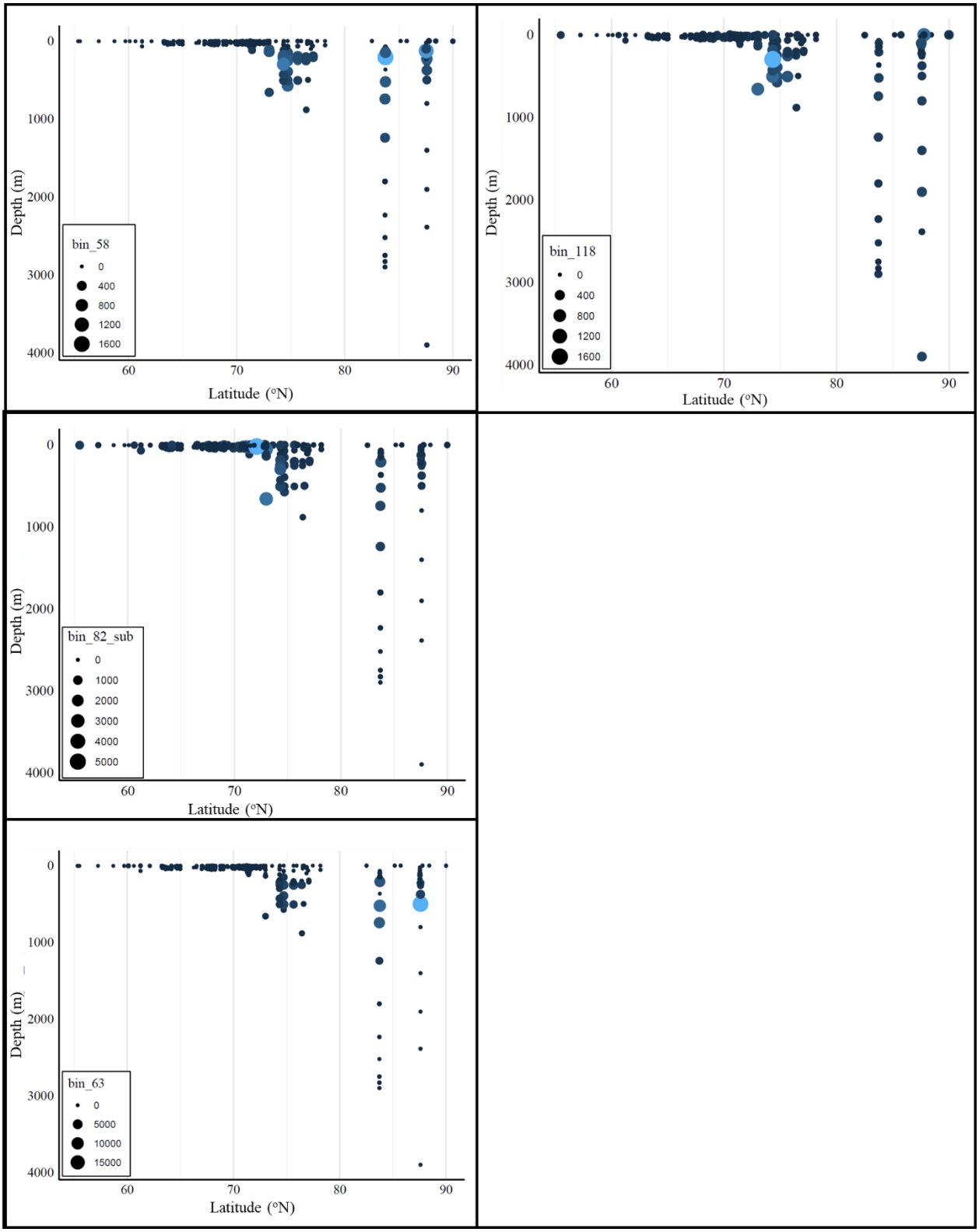


Figure 18. Relative abundance distributions (x1000000) of the MG-II archaeal MAGs identified in the Western Arctic Ocean displaying a mid-depth ecotype. Colours correspond with the abundances (light blue higher relative abundance and dark blue is lower relative abundance.

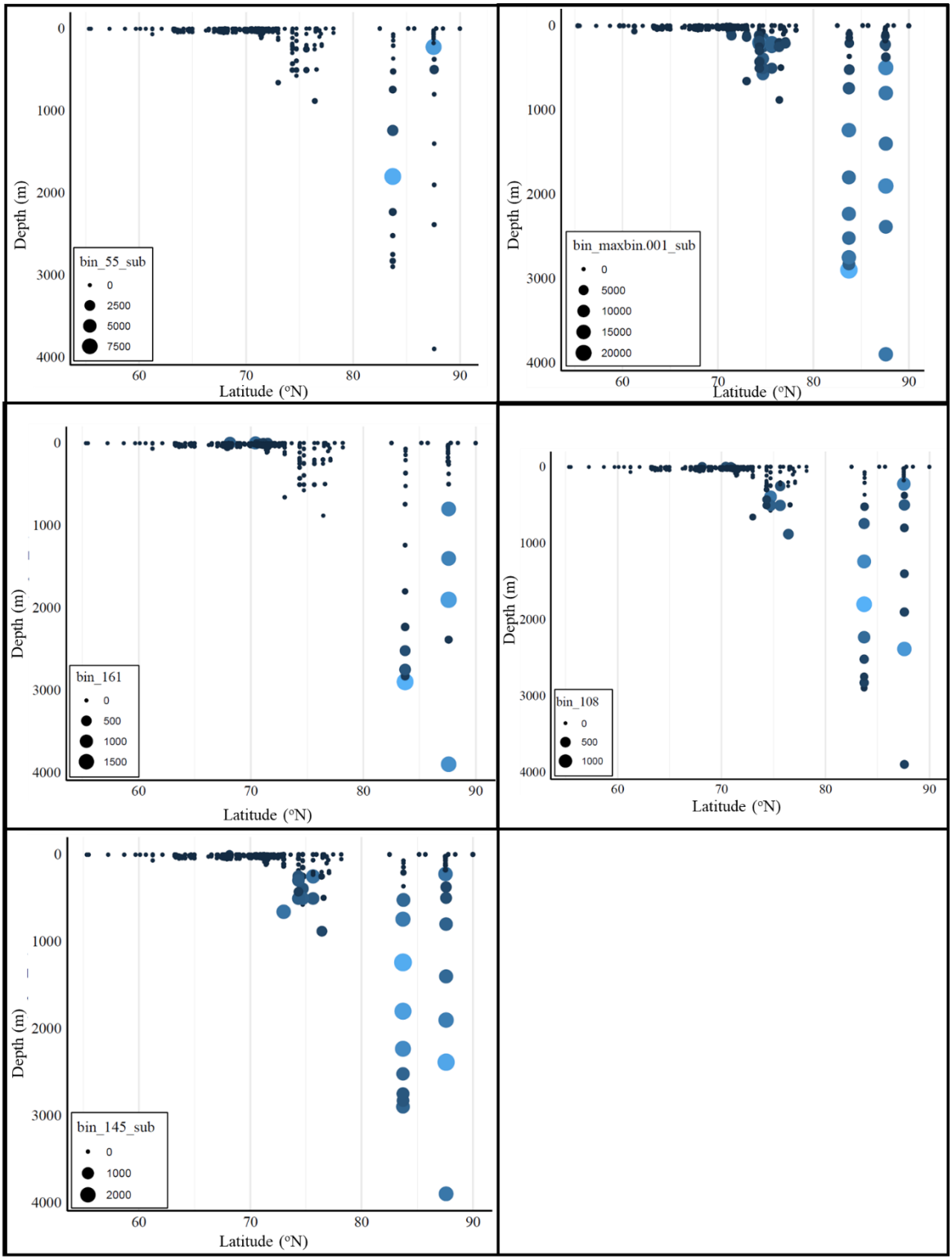


Figure 19. Relative abundance distributions (x1000000) of the MG-II archaeal MAGs identified in the Western Arctic Ocean displaying a deep ecotype. Colours correspond with the abundances (light blue higher relative abundance and dark blue is lower relative abundance.

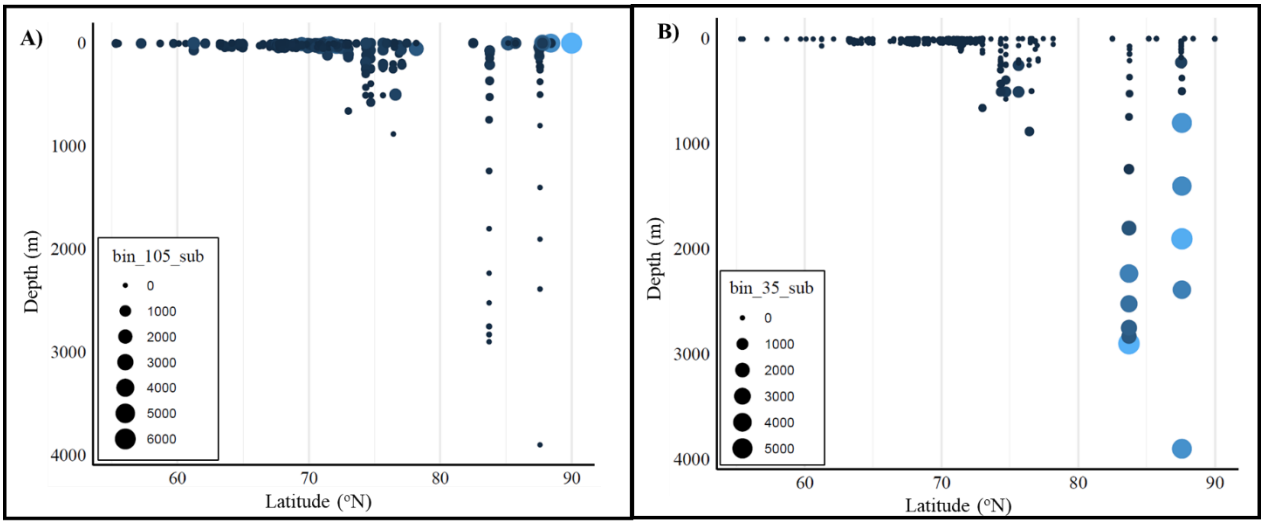


Figure 20. Relative abundance distributions (x1000000) of the MG-II archaeal MAGs identified in the Western Arctic Ocean displaying a surface ecotype (A) and a deep ecotype (B).

2.4 DISCUSSION

2.4.1 ARCHAEAL MAG IDENTIFIECATION AND THE USE OF METAGENOMIC ANALYSIS

Metagenomics is a relatively new technique for the study of environmental microbial samples. Historically, organisms were classified using 16s rRNA sequencing, also referred to as tag sequencing. The general process of tag sequencing involves sequencing a small region of the genome, typically the 16S rRNA region, selected with the use of primers for taxonomic assignment. There are three major drawbacks to this type of analysis. The first is that the primers chosen for the sequencing process can greatly affect the downstream results. There are many aspects to consider was choosing primer sets: the primer length, the specificity, and other physical and chemical properties that can affect the efficacy of the analysis. The chosen primers therefore have potential to completely miss entire clades of organisms, especially those that are rare and relatively unique. The second drawback is that the 16S rRNA region is fairly short, relative to the

entire genome, and may not be effective in elucidating very closely related organisms, resulting in incorrect taxonomic assignments. The final drawback of this technique is that no functional information is provided. Therefore, the biogeochemical processes being preformed in that respective environment are left unknown. Metagenomic analysis attempts to resolve these issues by removing the need to use primers and by sequencing all regions of the genome to obtain a higher level of phylogenetic resolution and provide functional information.

In this study, 17 archaeal MAGs were identified in the Western Arctic Ocean. Although relative abundances and diversity were low, members of all three marine groups were identified. They displayed ranges in completeness and contamination, although most were categorized as medium- or high-quality drafts. Both MG-III bins (35_sub (64.23%); 105_sub (60.59%)) had very high GC content, alongside members of MG-II (55_sub (59.05%); 54 (59.80%); maxbin.001_sub 60.59%); 63 (55.33%); 118 (59.81%); 108 (60.59%)) (FIGURE NUMBER). GC content, the proportion of guanine and cytosine in the DNA, is often used to infer the level of DNA stability within a genome as guanines and cytosines can form three hydrogen bonds. High GC content is often associated with microorganisms living in harsh environments such as those with high temperatures, high salt concentrations, or high levels of predation. These observations are not typically associated with organisms observed in colder habitats.

There were discrepancies observed between previous studies of Archaea in marine environments and this study when exploring the spatial distributions of the archaeal marine groups. It is likely that these differences were a result of using different sequencing technologies. The most probable cause for these differences is that 16S rRNA sequencing causes closely related organisms to be identified as the same lineage. Meaning that the higher abundance organisms would display a stronger genomic signal, therefore biasing the results of where these organisms were found. The

use of metagenomic sequencing has allowed for the separation of these closely related genomes into individual bins so that the diversity of distribution patterns could be observed.

2.4.2 ARCHAEAL DISTRIBUTIONS INFLUENCED BY WATER MASSES RATHER THAN STRICTLY LIGHT PENETRATION AND TEMPERATURE IN THE WESTERN ARCTIC OCEAN

The Arctic Ocean is a small, enclosed basin, that provides a harsh environment for microbial life. The temperatures are low (< 20°C), the salinities are high (> 20 ppt), and light is variable with season and depth. This ocean offers a unique look at the microbial communities residing there, their biogeography, as well as their biogeochemical processes. The water masses of the Western Arctic Ocean were defined by their thermocline and nutrient properties. These water mass layers were simplified to contain the Summer Surface Layer, Polar Mixed Layer, Atlantic Water layer, and Arctic Ocean Basin Water layer. The overall microbial communities were shown to cluster into three major groups (Figure 13). These groups were differentiated based on the water mass layer of collection. It was shown here that the different water mass layers harbour distinct and unique communities.

Previous studies (Ionescu et al., 2009; Liu et al., 2018; Pereira et al., 2019; Rinke et al., 2019) have observed that the three archaeal marine groups have specific and consistent distributions with a marine water column. For example, the MG-I are commonly found to be the most abundant archaeal clade in marine environments overall but are generally found at higher abundances at or below 200 m (Baer et al., 2014; Christman et al., 2011). The MG-II are often found at higher abundances in the upper 200 m, however are present at lower abundances at greater depths (Liu et al., 2017). Finally, the MG-III are often found at very lower abundances fairly deep

within the water column. It has been suggested that these variations in distributions are controlled microbial responses to light penetration into the water, and secondly by temperature.

When analyzing the individual bins from each marine group from all sampling locations and depths it was shown that these results differed from previous studies. Here it was observed that MAGs from the different archaeal marine groups did not show a collective preference for any one water mass (Figures16-20). Bins from within the same phylogenetic lineage did not display similar abundance patterns despite previously established distribution trends for the archaeal marine groups (Ionescu et al., 2009; Pereira et al., 2019; Rinke et al., 2019; Yilmaz et al., 2016). There was a considerable amount of variation within the groups themselves, suggesting that there was niche overlap due to shared environmental preferences and metabolic interactions involving the exchange of nutrients. Many of the previous studies that have looked at archaeal biogeography in the oceans have claimed that the phylogenetic organization of the groups influences the types of niches they reside in and that these characteristics are consistent and specific to each clade of Archaea. However, it was seen here that the type of biogeographical distributions observed were not specific to any one clade. All three marine groups identified in the Western Arctic had multiple biogeographical distributions. Although the temperature and light penetration are connected to the different water mass layers, these parameters do not affect the distributions of the entire clade in the same way.

The distributions of the three marine groups were investigated at Stations 26 and 30. Both of these stations were located above 80°N, closer to the pole. Nitrate concentrations at these stations followed a classic nutrient profile where concentrations were low at the surface and increased with depth (Figure 12). Low levels of nitrate are indicative of nutrient utilization by phyto- and picoplankton, consistent with the expected surface blooms. At these stations, it was

observed that nitrite concentrations were elevated at the surface (Figure 12) which is often associated with a higher prevalence of MG-I activity (Q. Liu et al., 2018) because their ability to oxidize ammonia to nitrite. We did see a moderately elevated MG-I relative abundance at 25 m (Figure 15A); however, it is difficult to correlate these two findings because there are other ammonia-oxidizing microbes that can be influencing this observation, especially in late summer months when phytoplankton blooms are expected to be present. Other studies have observed MG-I at higher abundances below the thermocline (Ionescu et al., 2009), which has been used to suggest that MG-I members prefer an environment that is cooler. It was observed that the MG-I bins were found in greater abundances in the upper 500 m (Figure 15A). There were two distribution ecotypes observed for the MG-I (Figure 16). This remains consistent with other studies that found that MG-I abundances were greater below 200 m (Church et al., 2003; Herfort et al., 2007; Q. Liu et al., 2018; Pitcher et al., 2011; Reji et al., 2019; Wuchter et al., 2006).

MG-II Archaea are generally thought to be surface residing, with many studies observing them at high abundances in the top 200 m of the water column (Pereira et al., 2019). The MG-IIb are commonly seen at depths around 200 m, but most studies did not sample below 500 m (Church et al., 2003; Offre et al., 2013; Pereira et al., 2019; Rinke et al., 2019). MG-II are usually the most abundant Archaea in surface waters, so it is unusual to see the MG-II MAGs dominating the community in the epipelagic zone (Figure 15B). This observation has two explanations. One, that the abundances truly are higher at depths greater then 500 m. Mentioned previously, many studies did not sample below 500 m, and therefore did not observe this finding. The second explanation involves the use of relative abundance as a measurement. The calculation of relative abundance involves mapping the reads from the sequencing to the assembled contigs. This provides a proportion that represents how many times that contig was observed in the sample. This can then

be used to calculate how many times each bin was observed in the sample. Relative abundance is therefore a measure of the total proportion of microbial bins identified. Consequently, when an increase in relative abundance is detected, it could actually be that there were a fewer number of organisms observed, rather than an increase in absolute counts. The MG-II displayed all three distribution ecotypes (Figures 17-19). The group of deep residing MG-II could be a new subgroup that has not yet been defined. This discrepancy between previous studies is likely due to a lack of sampling depth in previous studies. Although previously thought of as surface residing organisms, it is clear now that as a phylogenetic group they are not all surface residing and that their biogeography is much broader than previously stated.

MG-III are among the least understood members of marine microbial communities. They are typically considered low-abundance members of deep-sea environments (Galand et al., 2010). The MG-III MAG relative abundance was fairly consistent throughout the water column (approximately 10% of the archaeal community) (Figure 14B), which is consistent with literature (Haro-Moreno, et al., 2017) as these organisms are often identified at low abundances. Their relative abundances did increase with depth (Figure 15C), as was expected. It is important to note again, that this observation could be due to the use of relative abundance measurement as well. The two MG-III MAGs displayed a surface and deep ecotype. A recent study found members of MG-III in both the photic zones and deep-sea samples (Haro-Moreno et al., 2017). Haro-Moreno et al. (2017) suggest that the epipelagic MG-III evolved from a pelagic origin as they contained many light-protecting proteins and pathways derived from bacterial origin, likely obtained through horizontal gene transfer.

When analyzing the individual bins from each marine group from all sampling locations and depths it was shown that these results differed from previous studies. Here it was observed

that MAGs from the different archaeal marine groups did not show a collective preference for any one water mass (Figures16-20). Bins from within the same phylogenetic lineage did not display similar abundance patterns despite previously established distribution trends for the archaeal marine groups (Ionescu et al., 2009; Pereira et al., 2019; Rinke et al., 2019; Yilmaz et al., 2016). There was a considerable amount of variation within the groups themselves, suggesting that there was niche overlap due to shared environmental preferences and metabolic interactions involving the exchange of nutrients. Many of the previous studies that have looked at archaeal biogeography in the oceans have claimed that the phylogenetic organization of the groups influences the types of niches they reside in and that these characteristics are consistent and specific to each clade of Archaea. However, it was seen here that the type of biogeographical distributions observed were not specific to any one clade. All three marine groups identified in the Western Arctic had multiple biogeographical distributions. Although the temperature and light penetration are connected to the different water mass layers, these parameters do not affect the distributions of the entire clade in the same way.

Beyond the environmental structuring observed here, there may also be biological structuring occurring where the taxa present can be predicted based on other taxa there. There are factors influencing communities to work together rather than as individuals. It can be helpful to think of the microbial community structure like a body of microorganisms that are working within a defined geographical space. The geographical space that holds these microbial bodies are connected to the water mass layers. The water masses are the current assumption influencing the distributions but there could be a deeper link within the community. Investigating the co-occurrence of archaeal taxa with other members of the community is likely to elucidate the biological factors that are influencing this community structure.

In this Chapter it was hypothesized that the water body masses would be the strongest factor influencing the archaeal relative abundance patterns in the Western Arctic Ocean. It was shown that the overall communities within the water mass layers were distinct. It was also shown that the archaeal marine groups did not show a consistent distribution pattern, like what had been previously observed.

CHAPTER 3. PROTEOMIC SIGNATURES OF COLD ADAPTATION IN ARCHAEA

3.1 INTRODUCTION

3.1.1 MG-I AMMONIA OXIDIZING PROCESSES

Nitrogen is a necessary element for life and nitrogenous compounds are major limiting nutrients for marine primary production (Chang & Devol, 2009). Microbial nitrogen cycling is highly diverse but can be conceptualized as two main processes: nitrification (oxidation) and denitrification (reduction). Nitrification is a two-step process involving the oxidation of ammonia to nitrite, and nitrite to nitrate in aerobic environments (Figure 21) (Martens-Habbena et al., 2009). Aerobic ammonia oxidation is the first rate-limiting step in the nitrification process catalyzed by ammonia-oxidizing and nitrite-oxidizing microbes such as MG-I Archaea (Martens-Habbena et al., 2009). Aerobic ammonia oxidation is the only oxidative biological process that links reduced and oxidized pools of inorganic nitrogen in nature (Martens-Habbena et al., 2009). The transformation of nitrogen directly influences primary producers who may differ in their ability to utilize different nitrogen compounds (Ward, 2005).

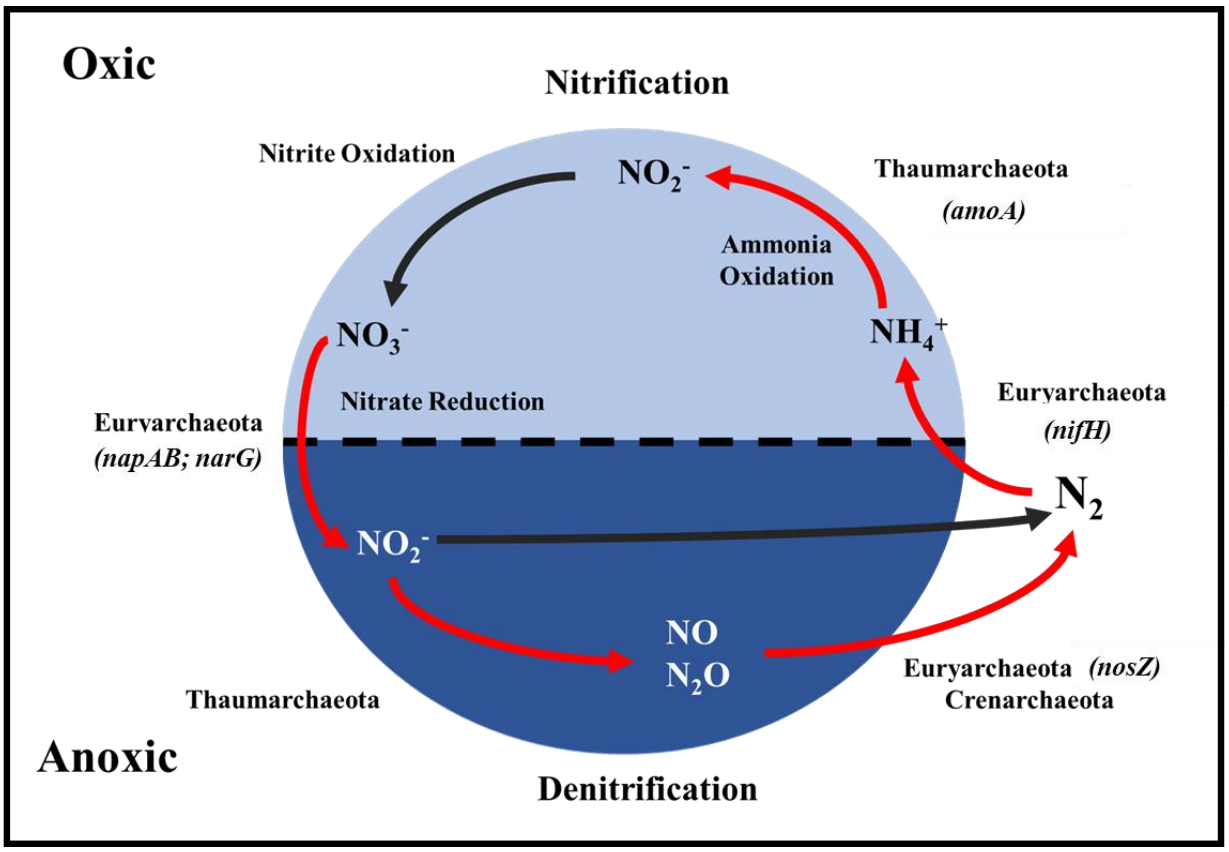


Figure 21. The process of nitrification and denitrification occurring in the ocean. Portions of the cycle in light blue occur in an oxic environment, while dark blue indicates an anoxic environment. Red arrows signify steps within the process that Archaea are believed to play a role in and if known, the main archaeal gene involved is listed. The genes are as follows: *amoA*: ammonia monooxygenase (PGF_02515073); *nifH*: predicted nitrogenase and assimilatory nitrate reductase (PGF_00025954); *napA*: periplasmic nitrate reductase (PGF_00031036); *napB*: nitrate reductase electron transport subunit (PGF_00025698); *narG*: nitrate reductase (PGF_09372748); *nosZ*: nitrous oxide reductase (PGF_00026089) (Yilmaz et al., 2016).

It has been shown that MG-I possess ammonia monooxygenase genes (*amoA*) that oxidize ammonium to nitrite (Baker et al., 2020; Q. Liu et al., 2018). These organisms have been shown to oxidize ammonia at low concentrations (Martens-Habbena et al., 2009) and obtain ammonia from urea and cyanate found in the open ocean (Baker et al., 2020; Christman, Cottrell, Popp, Gier, & Kirchman, 2011; Q. Liu et al., 2018; Palatinszky et al., 2015). Ammonia oxidation rates are often maximized near the bottom of the euphotic zone and are likely influenced by light levels in shallower waters where AOA activity is minimized (Ward, 2005) suggesting a light inhibition. There have also been suggestions that their distribution may be controlled by the availability of

ammonia as their summer blooms coincide with high nitrite summer peaks (Horak et al., 2013; Q. Liu et al., 2018; Martens-Habbena et al., 2009), as observed off the coast of Georgia, USA where MG-I 16S ribosomal RNA genes and archaeal *amoA* genes coincided with peak nitrite concentrations (Q. Liu et al., 2018).

In every environment, there are two possibilities for ammonia consumption: nitrification or assimilation. Ammonia assimilation has been shown to favour increases in temperature compared to nitrification suggesting that as the Arctic continues to warm, assimilation may become more dominant (Baer et al., 2014; Ward, 2005). Nitrification rates and total *amoA* gene abundance have been found to be much higher in winter months than in summer which was attributed to the cold and dark conditions which increase the chemoautotrophic potential of the community (Christman et al., 2011). Baer et al. (2014) demonstrated that nitrification rates were up to 2-fold greater than ammonia assimilation in the ice-covered winter and spring seasons and appear to be driven by a high concentration of ammonia in the environment. Although ammonia oxidation and assimilation were not found to be significantly correlated, there has been an inverse relationship noted where the highest nitrification rates observed were at depths where the lowest assimilation rates were occurring (Ward, 2005). Phytoplankton have been shown to outcompete bacterial and archaeal nitrifiers for ammonia when waters are warmer and light penetration is higher (Baer et al., 2014; Martens-Habbena et al., 2009; Ward, 2005). During the winter, this competition is reduced as primary production is limited by light and bioavailable nutrients under ice coverage (Baer et al., 2014; Oziel, Schourup-Kristensen, Wekerle, & Hauck, 2022). Nutrient limitation is speculated to become a more prominent issue with decreased sea-ice coverage allowing more light to penetrate the water column (Oziel et al., 2022). This is likely to be one of the main controls on ammonia oxidation productivity. However, among bacteria, true

psychrophiles (organisms that grow exclusively at low temperatures) have been shown to be less sensitive to substrate concentrations and have the potential to remain competitive against other phytoplankton for ammonia (Baer et al., 2014; Yager & Deming, 1999).

Environmental factors such as light, competition, substrate, and community composition all support the process of nitrification and the use of ammonia as an energy source in the winter and spring months (Baer et al., 2014). As the Arctic continues to warm, it is expected that light penetration will increase, primary production rates will increase, begin earlier, and end later, and thus it is likely that ammonia competition will become greater. However, if the nitrifiers residing in these environments are "true-psychrophiles", they may be able to maintain the same levels of energy production even when concentrations are reduced.

3.1.2 ARCHAEAL COLD ADAPTATION

Temperature greatly impacts cellular biochemistry because many chemical reactions are limited by diffusion of the reagents to the active site of the enzyme that catalyzes the reaction. In warmer environments chemical processes are sped up, but if the temperature is too high, proteins denature, and cells die. In colder environments, processes slow down, and protein flexibility is greatly hindered. All microorganisms are at a temperature equilibrium with their environment, therefore organisms living in extreme temperatures, like those seen in the Arctic Ocean, typically have genetic and proteomic adaptations to withstand the conditions. Psychrophilic microorganisms can grow at -10 to 15 ^{o}C, metabolize at -20 ^{o}C, and survive at -45 ^{o}C (Cavicchioli, 2006). Psychrophiles typically have decreased enzymatic reaction rates and increased osmotic and oxidative stress (Cimen, Jensen, & Buckler, 2021; Goordial et al., 2016).

These organisms tend to have protein adaptations that enable them to live in sub-zero environments which allow DNA replication, transcription, translation, and membrane fluidity to

be maintained (Goordial et al., 2016). Some adaptations in protein structure can include fewer salt bridges, proline residues, and hydrogen bonds at low temperatures (Goordial et al., 2016). In general, cold-active enzymes will have fewer strong amino acid residue bonds (Cimen et al., 2021). They also tend to increase the copy number of core genes to combat the low reaction rates brought on by colder temperatures (Goordial et al., 2016). The model cold adapted Archaea *Methanococcoides burtonii* has many molecular adaptations associated with psychrophilia including differences in its Elongation Factor 2 (EF2) (Siddiqui et al., 2002; Thomas & Cavicchioli, 2000), changes in its intracellular levels of potassium glutamate and potassium aspartate (Siddiqui et al., 2002), and a higher rate of membrane-lipid unsaturation (Nichols et al., 2004). Generally, proteins from cold adapted Archaea have more non-charged amino acids (Gln and Thr), and fewer hydrophobic residues (Leu) (Cavicchioli, 2006; Saunders et al., 2003). These changes are compensations that increase flexibility, reduce the activation energy of the protein-substrate transition state, and increase catalytic efficiency at low temperatures (Cavicchioli, 2006; Saunders et al., 2003). Phylogenetically, organisms from cold environments are typically more closely related to each other than to their thermophilic counterparts. For example, *N. maritimus*, a representative member of MG-I, identified in low temperature environments were more closely related to each other than thermophilic members in the same clade (Cavicchioli, 2006). Overall, the responses to cold environments in Archaea and Bacteria are quite similar (Goordial et al., 2016).

It is not currently feasible to experimentally determine the optimal growth temperature (OGT) of every species in Arctic seawater, but there are several metrics that can be used to infer OGT by identifying differences in genomic, transcriptomic, and proteomic composition using -omics techniques. Several indices have been suggested as a measure of cold adaptation in proteins

including arginine to lysine ratio, proline residue abundance, grand average of hydropathicity (GRAVY), aromaticity, aliphatic index, and acidic residue abundance (Goordial et al., 2016). Each index is assigned a score that can be used to determine whether the protein is cold adapted. It is expected that these scores will be lower in cold adapted organisms. A lower arginine to lysine score indicates a preference for using lysine, which is smaller and requires fewer nitrogen atoms. GRAVY is the measure of hydropathicity of each amino acid. It is expected that amino acids that are less hydrophobic and amino acids with smaller aliphatic side chains would allow for increased flexibility.

The environmental conditions preferred by individuals or populations of any microbial taxa are encoded in its genome, yet predicting those preferences from genomic data is an unsolved problem for many environmentally relevant parameters, including temperature and salinity. For example, predictions of optimal growth temperature in Bacteria and Archaea using only genomic information are robust for temperatures above 37°C, but poor at lower temperatures (Li et al., 2019). The benefits of being able to make such predictions would be widespread, including incorporating rates of microbial growth and metabolism into numerical models of ocean biogeochemistry based on easily accessible metagenomic sequencing. A number of genomic targets have been considered for such predictions in the literature, including nucleotide base composition, amino acid usage, and gene content. To assess cold adaptation in Archaea, a collection of archaeal metagenome-assembled genomes (MAGs) and genomes from cultured isolates were obtained from the BV-BRC database. Additional MAGs assembled from Arctic Ocean metagenomes (see Chapter 3.3-3.5) as part of this study were included in the collection. Metadata and sequences including nucleotide, amino acid, and coding regions were downloaded. The genomes obtained from the BV-BRC portal were of diverse origin encompassing fresh water

(lakes, rivers, ponds, and river sediment), marine water (oceans, seas, marine sediment, and saltwater aquariums), or other (wastewater, soil, feces, and brackish water). These samples were also collected from various years, seasons, and locations. Using the amino acid sequences of the annotated genes, six amino acid indices were calculated: arginine to lysine ratio, aliphatic index, aromaticity, acidic residue proportion, GRAVY index, and proline residue proportion.

The objective of this chapter is to identify signatures of cold adaptation in the amino acid usage and gene content of Archaea, to work towards future genomic predictions of optimal growth temperature in these important marine microbes. Based on previous work in cold adapted Bacteria (Gentilhomme, 2020; Goordial et al., 2016; Raymond-Bouchard et al., 2018), I hypothesized that cold adaptation would be apparent in the genomes of cultured and uncultured Archaea at the proteomic level when measured with six amino acid usage indices, and that environmental temperature will be a factor influencing the amino acid usage in Archaea, and finally that phylogenetics reflect the environmental niches of Archaea.

3.1.3 MG-I HOST ADAPTATION IN SEA SPONGES

Sponges are one of the first groups of organisms that evolved to become multicellular (Kenny et al., 2020). Most have a highly conserved body structure which consists of channels that filter water, up to thousands of liters daily, using pumping cells (Kenny et al., 2020; Slaby et al., 2017). Sponges have very large genomes, approximately double that of humans (Kenny et al., 2020). The sponge microbiome is known to include as many as 52 different microbial phyla, however the type and amounts may vary between species (Slaby et al., 2017; Webster & Thomas, 2016). There can be thousands of symbiont lineages associated with a single sponge, typically dominated by Proteobacteria (mostly Alpha- and Gammaproteobacteria), Acidobacteria,

Actinobacteria, Chloroflexi, Cyanobacteria, Crenarchaeota, as well as symbionts of several candidate phyla (Webster & Thomas, 2016). The sponge-associated microbes can perform carbon, nitrogen, and possibly sulfur and vitamin metabolism (Bayer et al., 2008; Hentschel et al., 2012; Horn et al., 2016). The microbiome of sponges has been shown to preferentially express genes related to cellular defense (restriction-modification systems, toxin-antitoxin systems, host colonization, and extracellular matrix utilization (Slaby et al., 2017). This observation is likely due the extensive water filtering performed by sponges that increase the contact with potential pathogens, toxins, and other stress factors (Degoeij et al., 2009; Slaby et al., 2017).

Generally, it has been observed that sponge symbiont genomes are larger than free living organisms and they contain a higher GC-content (50-60%) (Horn et al., 2016; Slaby et al., 2017). Enrichment of certain systems, such as amino acid transport and metabolism, replication, recombination, and repair were seen in the symbionts, while transcription, signal transduction mechanisms, cell wall and membrane biosynthesis, and cell mobility were depleted (Slaby et al., 2017). These findings were all found to occur in Bacteria, while no archaeal analysis was performed. A large issue with identifying these traits in Archaea is the lack of high-quality annotations available for archaeal proteins. The genetic changes seen in sponge symbionts are likely due to increased exposure of free DNA and other pathogens the microbes encounter during sponge filtration and reflects an evolutionary pressure to host adaptation. High levels of HGT are likely occurring and high GC-content and defense mechanisms appear to be a requirement for sponge adaptation (Horn et al., 2016).

3.2 METHODS

3.2.1 CALCULATION OF AMINO ACID COMPOSITIONAL INDICES FOR THE DETECTION OF COLD ADAPTATION SIGNATURES

Archaeal genomes and metadata were collected and downloaded from the BV-BRC portal (https://www.bv-brc.org). The genomes were all classified as "good" in the database. Cold adaptation indices were calculated using existing scripts developed at the University of Alaska Fairbanks (https://github.com/agentilhomme/Genomic-Signatures-of-OGT-2020-). Briefly, global and local protein family assignments for each gene in each genome were downloaded. Using the amino acid sequences of each protein, statistical analysis of cold adaptation as a function of phylogeny was done in R Studio (R Core Team, 2021). For each protein in each sequence, the amino acid indices were calculated. The proteins were averaged based on their global protein family assignments. The indices were then standardized using Z-scores, which are numerical measurements that describe the relationship of the value to the mean and are measured in terms of standard deviations from the mean. The analysis was based on the phyla of Archaea and the ward distances of the Global protein families. Ward's distance method uses a minimum variance criterion where the total within cluster variances is minimized. To implement this method, it will find a pair of clusters that result in the smallest increase in total within-cluster variance when merged. Statistical analysis was performed comparing various groups (phylogenetic, environmental, etc.) for each indices using T-Tests as all indices followed a normal distribution and had unknown variances. The null hypothesis assumed that the two groups in comparison have equal variances. The null hypothesis was rejected when the P-value<0.05 indicating that the data readings were strong and probably not due to chance. The type of T-Test used was an equal

variance independent T-Test because the number of samples (genomes) in each group being compared was different. Degrees of freedom were calculated as $n_1 + n_2 - 2$.

3.2.2 CALCULATION OF CODON AND AMINO ACID USAGE

Sequences of all the genomic features (protein sequences and open reading frame DNA sequences) for each genome used in amino acid compositional indices analysis were downloaded from the BV-BRC database. Calculations of total amino acid usage by each genome were totaled in R. Codon usage analysis was performed in R using the seqinr package (Charif & Lobry, 2007). Briefly, codons within each open reading frame DNA sequence were counted, where a codon represents three DNA bases that correspond to an amino acid during protein translation. Total codon usage was presented as either a tally of each codon used in that genome and as a proportion of all the codons identified in that genome. Statistical analysis was performed comparing the sponge-associated genomes against the general population for the overall amino acid usage using T-Tests. The null hypothesis assumed that the two groups in comparison have equal variances. The null hypothesis was rejected when the P-value<0.05 indicating that the data readings were strong and probably not due to chance. The type of T-Test used was an equal variance independent T-Test because the number of samples (genomes) in each group being compared was different. Degrees of freedom were calculated as $n_1 + n_2 - 2$.

3.3 RESULTS

3.3.1 AMINO ACID COMPOSITIONAL INDICES SHOWS LITTLE VARIATION BETWEEN MG-I MAGS

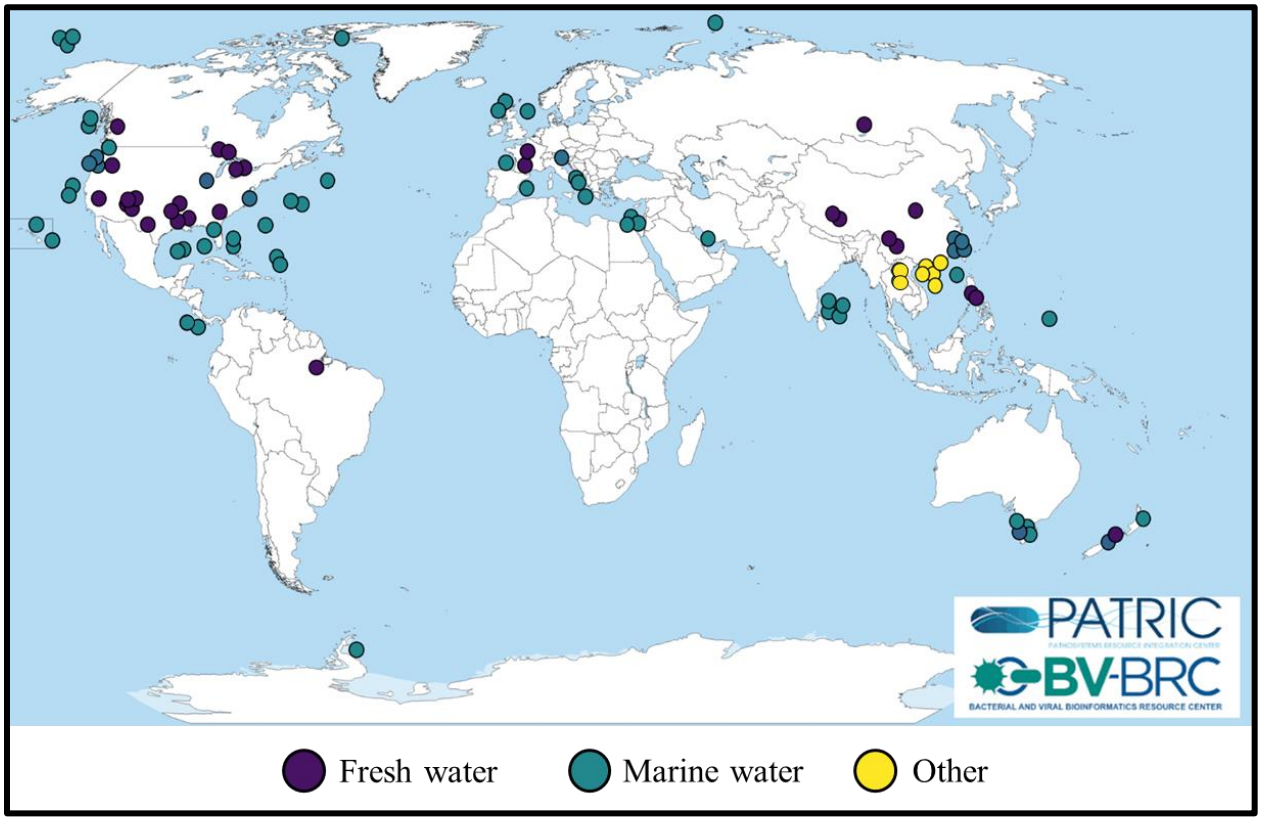


Figure 22. Geographical map showing the sampling locations of the MG-I genomes obtained from the BV-BRC database. Colours of each circle indicate the sampling source: fresh water (dark blue), marine water (turquoise), and other (yellow).

A collection of 142 MG-I MAGs were collected from the BV-BRC database to assess the six established amino acid indices. The source and location of the MAGs can be seen in Figure 22. The genomes were of diverse sources including season, year, environment, and location. The majority of these samples were obtained from mid-latitude regions, with few from the poles. This is due to the lack of available archaeal data in polar regions, exemplifying the importance of the previous analysis on archaeal MAGs in the Western Arctic Ocean. The individual indices for each

protein were averaged across global protein families within each MAG (Figure 23). The outgroups (*Thermococcus alcaliphilus* AEDII12; *Halorientalis regularis* TNN28; *Candidatus Bathyarchaeota* strain NP77; *Candidatus Bathyarchaeota* strain B24-2; *Nanoarchaeum equitans* Kin4-M; *Nanoarchaeota* strain JGI OUT-1) showed distinct differences in indices, specifically in the arginine to lysine ratio and acidic residue proportion compared to MG-I. Little variation was observed across the MG-I for each index when standardized (Figure 23). The index distributions for the three MG-I genera used in this analysis can be seen in Suppl. Fig. 2. In addition, the environmental niches which they occupied were not consistent with the phylogeny.

Archaeal MG-I MAGs collected from the Western Arctic Ocean (Chapter 2) were also added to this analysis and included bins 29, maxbin.015_sub, and maxbin.074_sub. These MAGs were chosen as their completeness was above 70%, although lower than the MIMAG standard of 90%, they were the highest quality bins from the identified MG-I MAGs and were still considered to be medium quality drafts. To investigate any differences between the Arctic MG-I MAGs and the general dataset, the index distances and the phylogenetic distances of each MAG compared to another MAG is the dataset was plotted in Figure 24. The indices for these MAGs were all within the normal distributions of the indices for the population tested. The bins did have slightly elevated distances in both the indices and phylogenetics, but these was not significant.

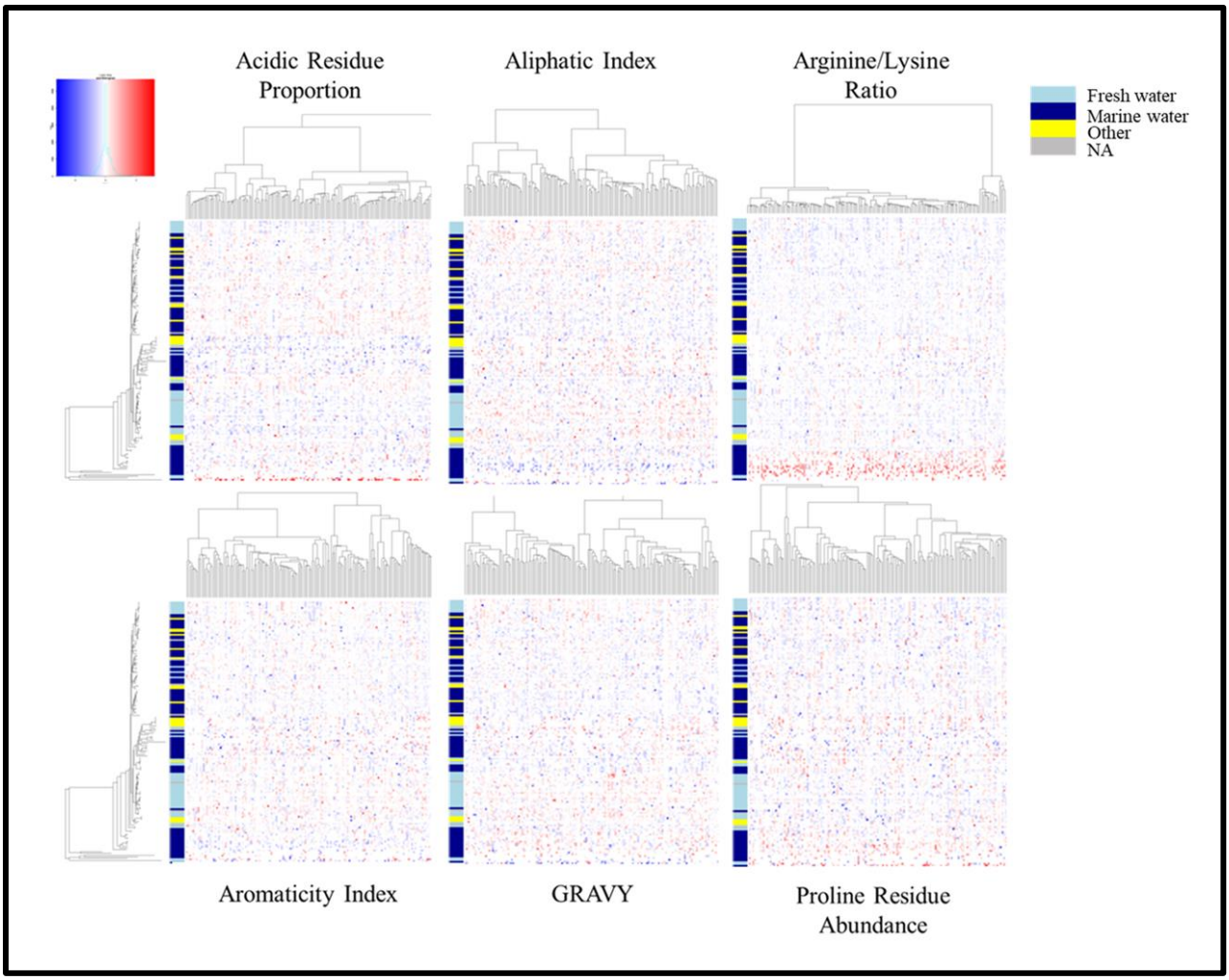


Figure 23. **Heatmaps displaying the standardized amino acid compositional indices for the MG-I MAGs (y-axis; phylogenetic distance of each sequence) for each global protein family (x-axis; ward distances).** Red indicates a high Z-score, while blue indicates a low Z-score. White indicates a score of zero or an NA. Coloured bars on the lefthand side of each plot indicate the source of each genome (fresh water (turquoise); marine water (dark blue); other (yellow)). The outgroup taxa are at the bottom.

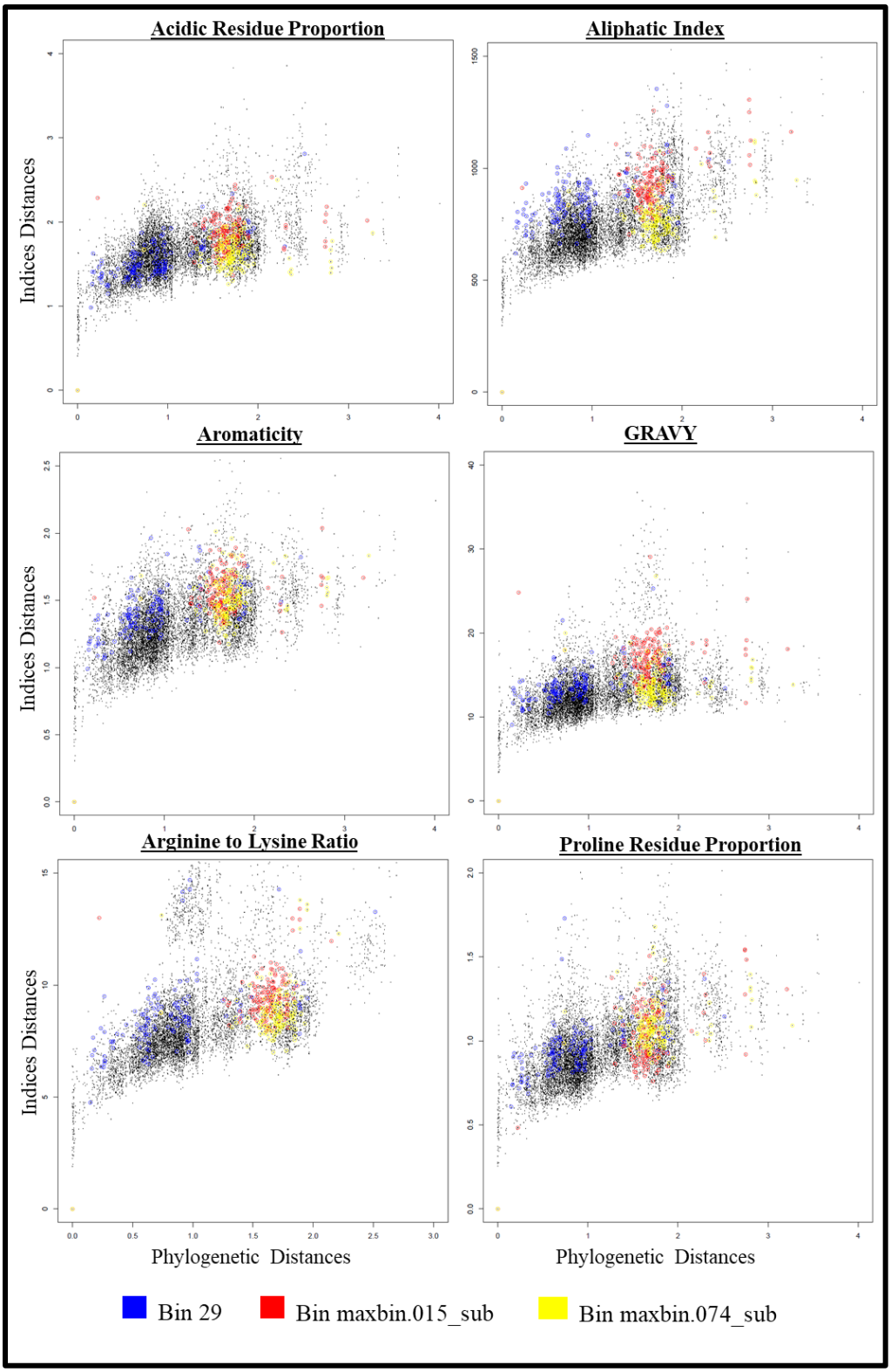


Figure 24. Indices distances against the phylogenetic distances for MG-I highlighting the Western Arctic Archaea. Three of the MG-I MAGs (Chapter 3.3 - 3.5) from the Western Arctic were included in this analysis.

Although there were not any detectable signatures of cold adaptation when comparing the Arctic MAGs to the general archaeal population, there were differences observed in the indices tested based on two other factors: phylogeny and environment. There were two main clades of MG-I (Suppl. Fig. 3) encompassing the genera *Nitrosopumilus* and *Nitrosomarinus* (Group A) and *Nitrosoarchaeum* (Group B) which were identified based on phylogenetic tree structure. It is important to note that there was a collection of *Nitrosopumilus* MAGs associated with sea sponges that clustered phylogenetically with Group B. The phylogenetic tree building tool used single copy genes to measure the relatedness between the different MAGs. Many of the MAGs included in this analysis did not have completeness or contamination measurements included in their metadata. CheckM was used to assess bin quality and can be seen in Suppl. Table 1. This will be discussed further in Chapter 3.3.3. Between these two phylogenetic groups ($n_1 = 65$ (Group A); $n_2 = 76$ (Group B); Degrees of Freedom = 139), it was observed that there was a significant difference in the indices for the arginine to lysine ratio ($P<0.0001$), acidic residue ($P<0.0001$), GRAVY ($P<0.0001$), and proline residue proportion ($P<0.001$) indices (Figure 25). Group B was found to have higher scores for all of the significantly different indices, excluding the acidic residue proportion. These differences were determined to be due to the outlier Nitrosopumilaceae MAGs in Group B identified above. In addition, there was also a collection of MAGs (2202732.58, 2202732.18, 2202732.35, 2202732.59, 2202732.61, and 2202732.62) that were not associated with sponges that were also members of this outlier Nitrosopumilaceae group, although they were not identified at the genus level. Their completeness and contamination scores ranged from 77.02-92.23% and 0.07-0.97%, respectively, suggesting that this observation was not caused by poor bin quality. With the removal of the outlier Nitrosopumilaceae MAGs, there was a lack of phylogenetic signals

influencing the indices and could speak to the evolutionary conservation of this group, which will be further explored in Chapter 3.4.2.

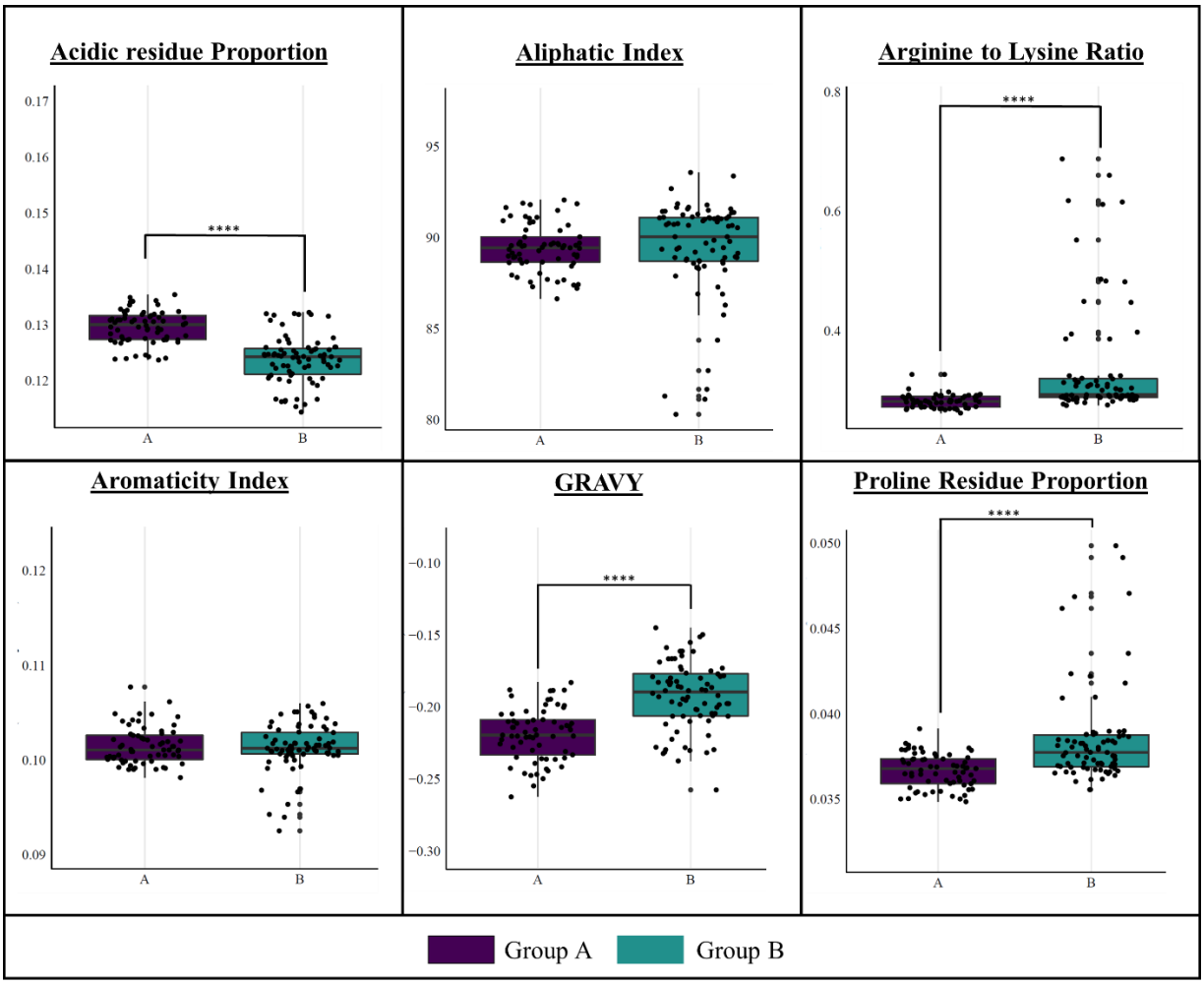


Figure 25. Comparison of indices from the Group A (Dark blue) and Group B (Turquoise) MG-I.

The second factor influencing variance within the indices was the collection environment. The majority of the MAGs (53%) were collected from marine environments, and it was observed that there were significant differences between the fresh and marine water MAGs in all of the amino acid indices tested in at least one of the phylogenetic groups (Group A: $n_1 = 9$ (fresh water), $n_2 = 42$ (marine water), Degrees of Freedom = 49; Group B: $n_1 = 28$ (fresh water), $n_2 = 33$ (marine

water), Degrees of Freedom = 59). This observation was consistent within the two phylogenetic groups as the aliphatic ($P<0.0001$), acidic residue proportion ($P<0.0001$), GRAVY ($P<0.0001$), and proline residue proportion ($P<0.001$) indices were significantly different in Group A, and the arginine to lysine ratio ($P<0.0001$), aliphatic ($P<0.0001$), aromaticity ($P<0.001$), and proline residue proportion ($P<0.0001$) indices were significantly different within Group B when comparing the fresh water and marine water MAGs (Figure 26). The acidic residue proportion, the arginine to lysine ratios, and the proline residues were lower in the MAGs from freshwater environments in both groups. The aliphatic and GRAVY indices were lower in the MAGs from marine environments. The aromaticity index was also lower in the marine MAGs in Group B.

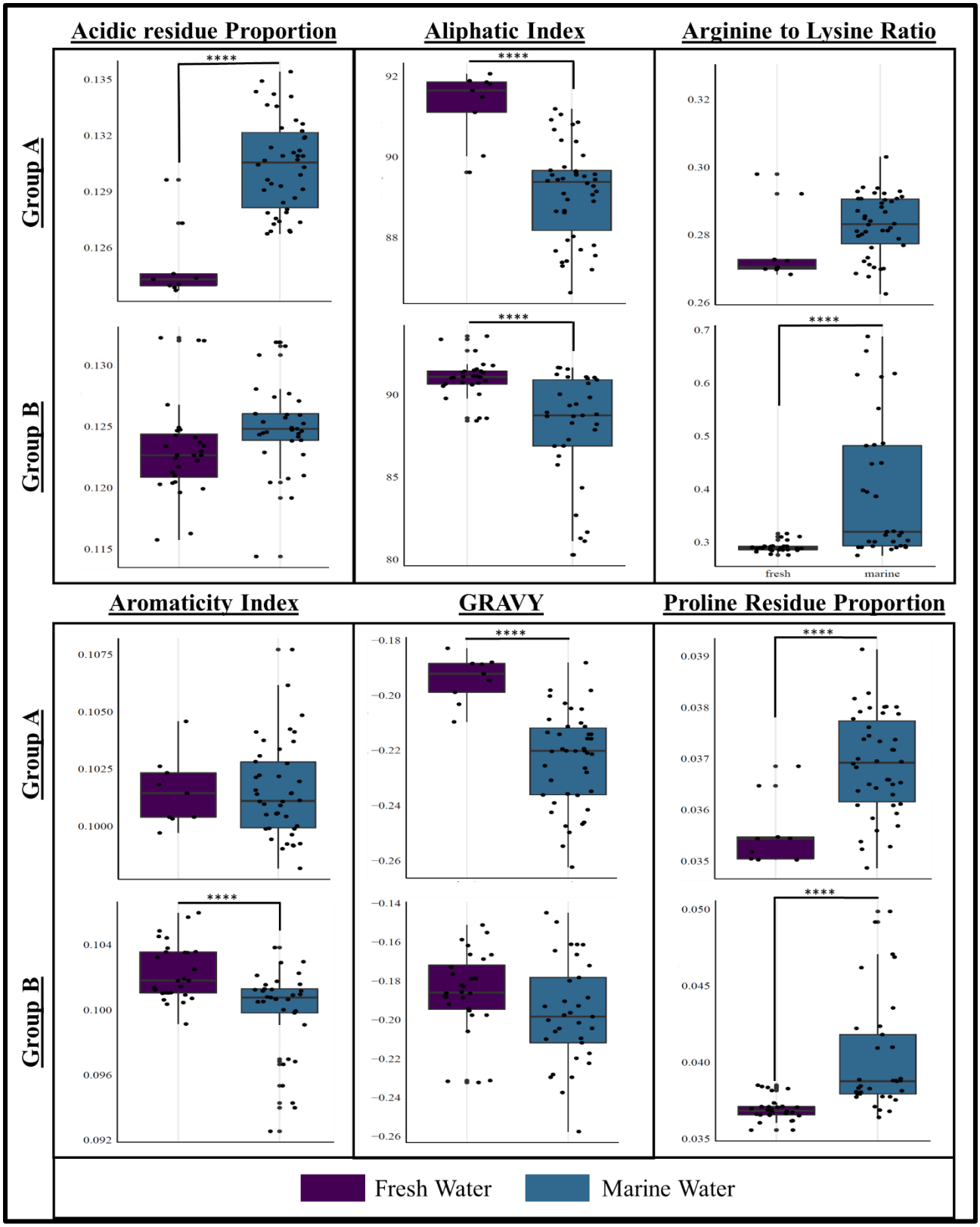


Figure 26. Boxplot of the comparison between fresh water (dark blue) and marine water (medium blue) for Group A and Group B MG-I for the amino acid indices.

3.3.2 EURYARCHAEOTA MAGS BRANCH INTO FIVE CLUSTERS BASED ON AMINO ACID COMPOSITIONAL INDICES

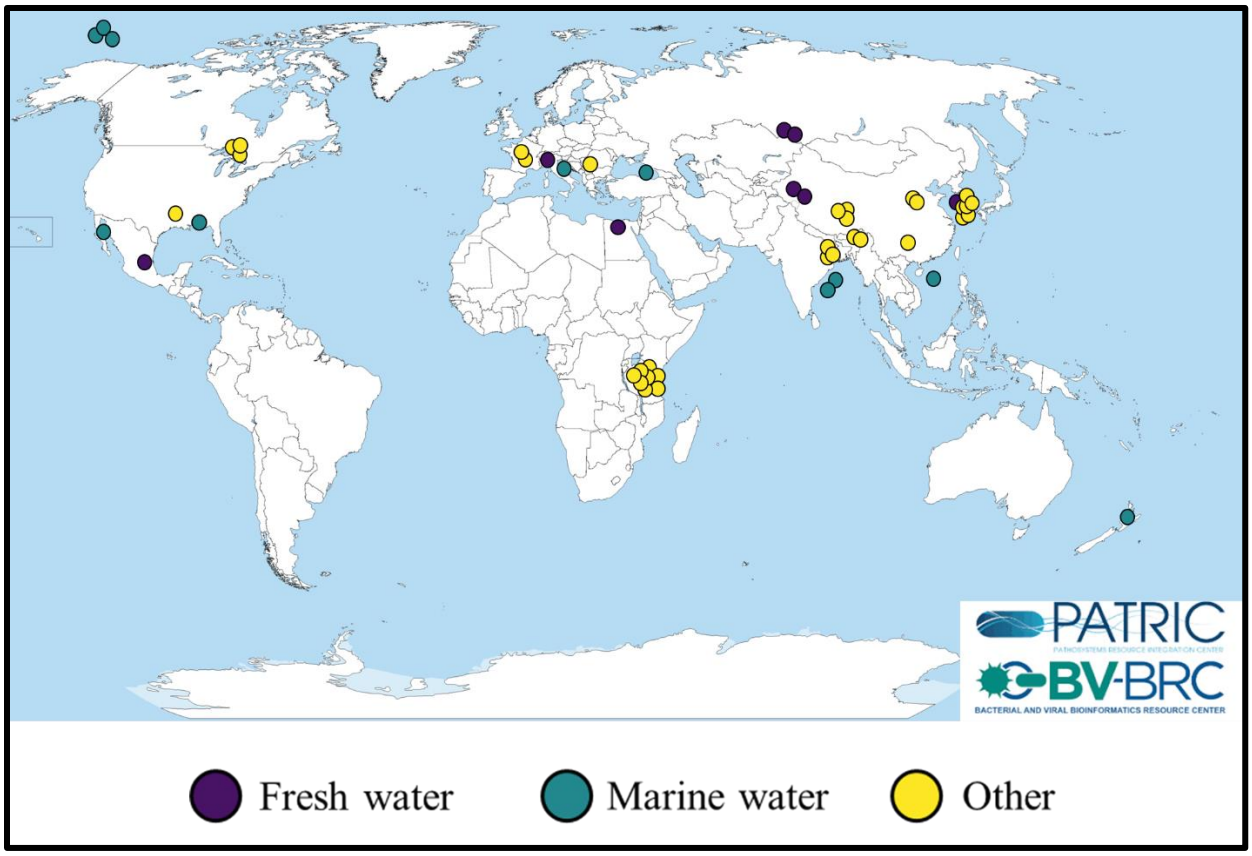


Figure 27. Geographical map showing the sampling locations of the Euryarchaeota genomes obtained from the BV-BRC database. Colours of each circle indicate the sampling source: fresh water (dark blue), marine water (turquoise), and other (yellow).

Using a collection of 146 Euryarchaeota genomes collected from the BV-BRC portal (Figure 27), the same analysis was performed to determine if the lack of variance in the amino acid usage was consistent across a larger set of Archaea or was specific to MG-I. The outgroups used in this analysis were *Nitrosopumilus* sp. UBA526, *Nitrosopumilus* sp. UBA241, *Candidatus Woesearchaeota* CG07, and *Candidatus Woesarchaeota* CG08. Contrary to the results for MG-I, Euryarchaeota appeared to have a consistent clustering pattern of five clades across all indices

measured (Figure 28). This clustering trend appeared to be consistent with the phylogenetic grouping of the organisms (Suppl. Fig. 4), rather than the source of their isolation.

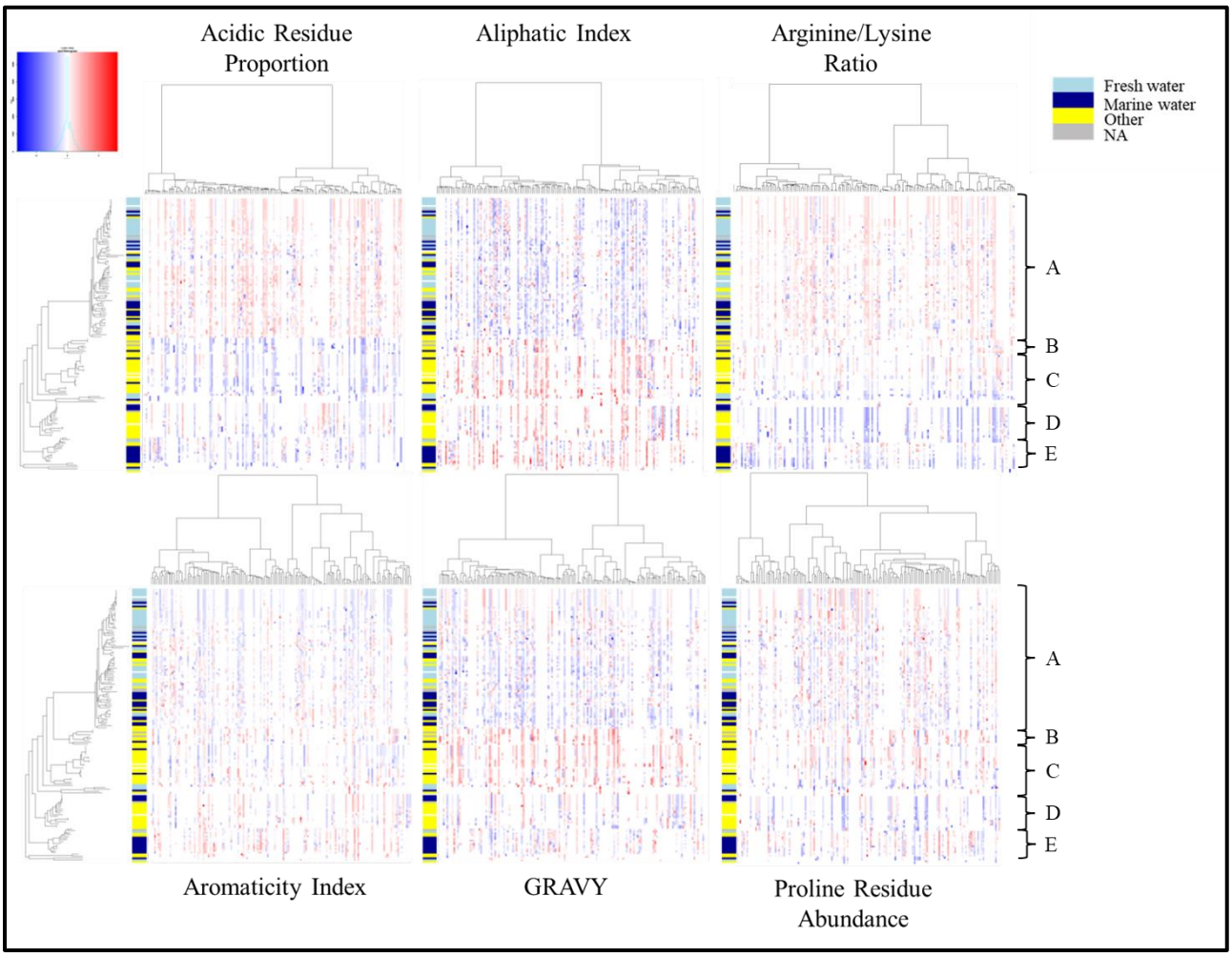


Figure 28. Heatmaps displaying the standardized amino acid compositional indices for the Euryarchaeota genomes/MAG (y-axis; phylogenetic distance of each sequence) for each global protein family (x-axis; ward distance). Red indicates a high Z-score, while blue indicates a low Z-score. White indicates a score of zero or an NA. Coloured bars on the lefthand side of each plot indicate the source of each genome (fresh water (turquoise); marine water (dark blue); other (yellow)). The letters on the right-hand side of the plots indicate the clustering. The outgroup taxa are at the bottom.

The first and largest cluster (Group A) was monophyletic and contained members of the class Halobacteria. This group showed higher arginine to lysine ratios (0.642 +/- 0.182) than the other groups (Table 4). The second cluster (Group B) was monophyletic and contained members

from the order Methanomicrobiales. This group displayed higher aliphatic indices (92.106 +/- 3.328), and lower arginine to lysine ratios (0.364 +/- 0.172) and proline residue proportions (0.036 +/- 0.005) indices. The third cluster (Group C) was monophyletic and contained members from the orders Methanosarcinales and Methanotrichales. This group showed higher aliphatic (88.55 +/- 4.23) and GRAVY indices (-0.138 +/- 0.075). The fourth cluster (Group D) was polyphyletic and contained members of the orders Archaeoglobales and Methanobacteriales. This group showed lower arginine to lysine ratios (0.595 +/- 0.191), as well as higher GRAVY (-0.137 +/- 0.081) and aliphatic indices (89.006 +/- 5.130). The fifth cluster (Group E) was monophyletic and contained members from the order Thermococcales. This group showed higher average arginine to lysine ratios (0.700 +/- 0.147) and lower average GRAVY index (-0.207 +/- 0.067).

Table 4. Averaged Amino acid compositional indices for the five clusters observed for the Euryarchaeota.

Group	Family	Arginine to Lysine	Acidic Residue Proportion	GRAVY Index	Proline Residue Proportion	Aromaticity	Aliphatic Index
A	Haloarculaceae Halobacteriaceae Halococcaceae Haloferacaceae Halorubraceae Natrialbaceae	0.645 +/- 0.182	0.154 +/- 0.018	-0.181 +/- 0.056	0.045 +/- 0.004	0.093 +/- 0.006	86.570 +/- 5.060
B	Methanocalculaceae Methanocorpusculaceae Methanomicrobiaceae Methanoregulaceae	0.364 +/- 0.172	0.145 +/- 0.017	-0.167 +/- 0.046	0.036 +/- 0.005	0.098 +/- 0.005	92.106 +/- 3.628
C	Candidatus Methanoperedenaceae Methanosarcinaceae Methanotrichaceae	0.614 +/- 0.133	0.146 +/- 0.019	-0.138 +/- 0.075	0.045 +/- 0.004	0.094 +/- 0.005	88.55 +/- 4.230
D	Archaeoglobaceae	0.595 +/- 0.191	0.146 +/- 0.021	-0.137 +/- 0.081	0.043 +/- 0.006	0.097 +/- 0.006	89.006 +/- 5.130
E	Thermococcaceae	0.700 +/- 0.147	0.164 +/- 0.017	-0.207 +/- 0.067	0.045 +/- 0.002	0.095 +/- 0.008	85.375 +/- 5.760
O	NA	0.675 +/- 0.171	0.157 +/- 0.016	-0.157 +/- 0.060	0.044 +/- 0.002	0.096 +/- 0.008	87.467 +/- 6.965

3.3.3 DIFFERENCES IN ARGININE TO LYSINE RATIOS, CODON USAGE, AND AMINO ACID USAGE OBSERVED IN MG-I MAGS ASSOCIATED WITH SEA SPONGES

In investigating the amino acid compositional indices for cold adaptation, one specific group of MG-I were found to have unusually high arginine to lysine ratios (Figure 29A). This group of MAGs were all found to be isolated from within or on sea sponges (Table 5). These genomes were all members of the Nitrosopumilaceae family, and all originated from marine environments. Additionally, this set of MAGs, which will be referred to as sponge associated MAGs (SAMs) for the remainder of the text, were also previously identified as members of the outlier *Nitrosopumilus* MAGs that were incorrectly phylogenetically clustering with the Group B MG-I in chapter 4.2. Not all of these genomes were SAMs, and not all the SAMs were grouped here. The SAMs included in this list were Nitrosopumilaceae archaeon strain Plut_88866, Nitrosopumilaceae archaeon strain Plut 88877, Nitrosopumilaceae archaeon strain PNGco C binSS26, *Nitrosopumilus* sp. strain PNGco C binVS21, Nitrosopumilus sp. strain PNGco S binVS9, *Nitrosopumilus* sp. H8, *Nitrosopumilus* sp. D6, *Nitrosopumilus* sp. H13, *Nitrosopumilus* sp. B06, Nitrosopumilaceae archaeon p16x60bins.31, Nitrosopumilaceae archaeon p16x44bins.32, Nitrosopumilaceae archaeon 277c 17, and Nitrosopumilaceae archaeon 277c 18.

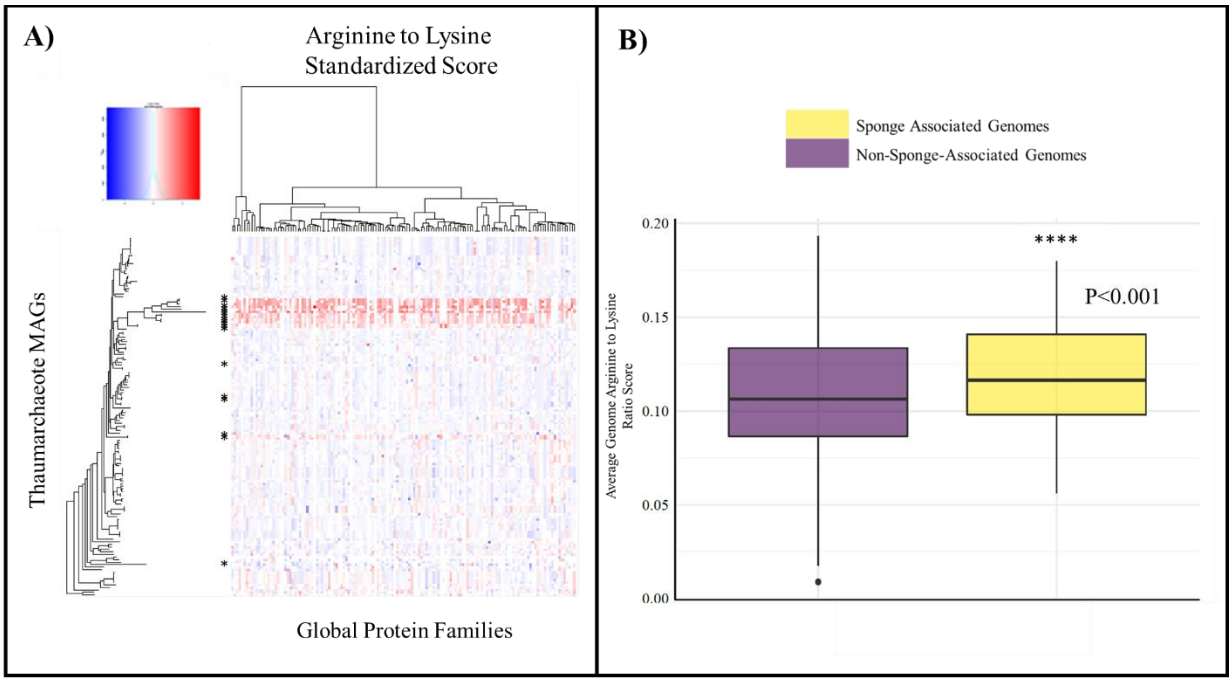


Figure 29. Arginine to lysine Z-scores for the MG-I sponge-associated MAGs. A) Heatmap displaying the standardized arginine to lysine Z-scores for the Euryarchaeote genomes/MAG (y-axis; phylogenetic distance of each sequence) for each global protein family (x-axis; ward distance). Red indicates a high score, while blue indicates a low Z-score. White indicates a Z-score of zero or an NA. an asterisk (*) indicated a sponge-associated genome. B) Boxplot of the average arginine to lysine ratios of the sponge-associated MAGs compared to the other MG-I genomes.

Table 5. **List of the Sponge-associated MAGS, their host, and their location of origin.**

Genome Name	Host Name	Location
Nitrosopumilus sp. Nsub	*Suberites*	NA
Nitrosopumilus sp. strain HM N	*Hymedesmia methanophila*	Mexico
Nitrosopumilaceae archaeon strain Plut 88866	*Porites lutea*	Australia
Nitrosopumilaceae archaeon strain Plut 88877	*Porites lutea*	Australia
Nitrosopumilaceae archaeon strain PNGco C binSS26	*Coelocarteria singaporensis*	Papua New Guinea
Nitrosopumilus sp. strain PNGco C binVS21	*Coelocarteria singaporensis*	Papua New Guinea
Nitrosopumilus sp. strain PNGco S binVS9	*Coelocarteria singaporensis*	Papua New Guinea
Nitrosopumilus sp. H8	*Hexadella dedritifera*	Atlantic Ocean
Nitrosopumilus sp. D6	*Hexadella dedritifera*	Atlantic Ocean
Nitrosopumilus sp. H13	*Hexadella dedritifera*	Atlantic Ocean
Nitrosopumilus sp. B06	*Hexadella dedritifera*	United Kingdom
Nitrosopumilaceae archaeon p16x60bins.31	*Ircinia* sp.	Panama
Nitrosopumilaceae archaeon p16x44bins.32	*Ircinia* sp.	Panama
Nitrosopumilaceae archaeon 277c 17	*Petrosia ficiformis*	Israel
Nitrosopumilaceae archaeon 277c 18	*Petrosia ficiformis*	Israel
Nitrosopumilaceae archaeon 277c 19	*Petrosia ficiformis*	Israel
Nitrosopumilus sp. strain PfA	*Petrosia ficiformis*	Israel
Nitrosopumilus sp. strain MH-Pat-all metabat2 16	*Mycale hentscheli*	New Zealand
Nitrosopumilus sp. LS AOA	*Lophophysema* sp. RT-2015	NA

Figure 29B shows that the arginine to lysine ratio indices for the SAMs were significantly higher ($p<0.001$) than MG-I found in other environments ($n_1 = 19$ (SAMs; $n_2 = 128$ (fresh, marine, and other), Degrees of freedom = 145); however, the variance was greater in the non-SAMs group.

Significant differences were also observed in the aliphatic ($P<0.00001$), aromaticity ($P<0.05$), and proline residue proportion ($P<0.000001$) indices. In Figure 30, it is seen that the patterns of amino acid usage by MG-I are consistent across the majority of the genomes measured. The SAMs, however, showed a differential pattern of amino acid utilization. The only amino acids that did not show a significant difference in the SAMs were alanine and lysine (Figure 31). Amino acids used less frequently in the SAMs were tyrosine, serine, arginine, proline, asparagine, methionine, leucine, phenylalanine, glutamic acid, and aspartic acid, while the amino acids used more frequently were tryptophan, valine, threonine, glutamine, isoleucine, histidine, glycine, cysteine.

Acidic amino acids, like glutamic acid and aspartic acid were used less frequently in these organisms which was consistent with the averaged indices for the acidic residue proportion (0.125 in SAMs and 0.127 in MG-I), although this difference was not significant ($P>0.05$). The contradiction between the acidic residue proportions and the individual glutamic acid and aspartic acid utilization frequencies were due to the nature of the calculations. The index calculates a total count of acidic amino acids in the proteins of each genome, while the frequency calculation takes into consideration the length of the protein sequences, therefore, removing the bias of different genome sizes. Bulky aromatic amino acids like tyrosine and phenylalanine were utilized less frequently, however within the same class of amino acid, tryptophan was used more frequently. The aromaticity index was consistent with these results as the indices for the SAMs (0.010) were lower than the rest of MG-I (0.102) ($P=0.02$). Histidine was used more frequently than arginine, both basic amino acids. This is supported by the aliphatic index which was significantly different ($P<0.0001$) in the SAMs. Valine and isoleucine were also more frequently used, having shorter aliphatic side chains than methionine and leucine which were not used as frequently. Glycine and cysteine were observed more frequently. Alternatively, proline was observed less frequently.

Glutamine and asparagine are both polar uncharged amino acids which show differential utilizations. Glutamine being used more frequently, however larger, is coded by the codons CAA and CAG, which have more C/G than the codons encoding asparagine (AAU and AAC).

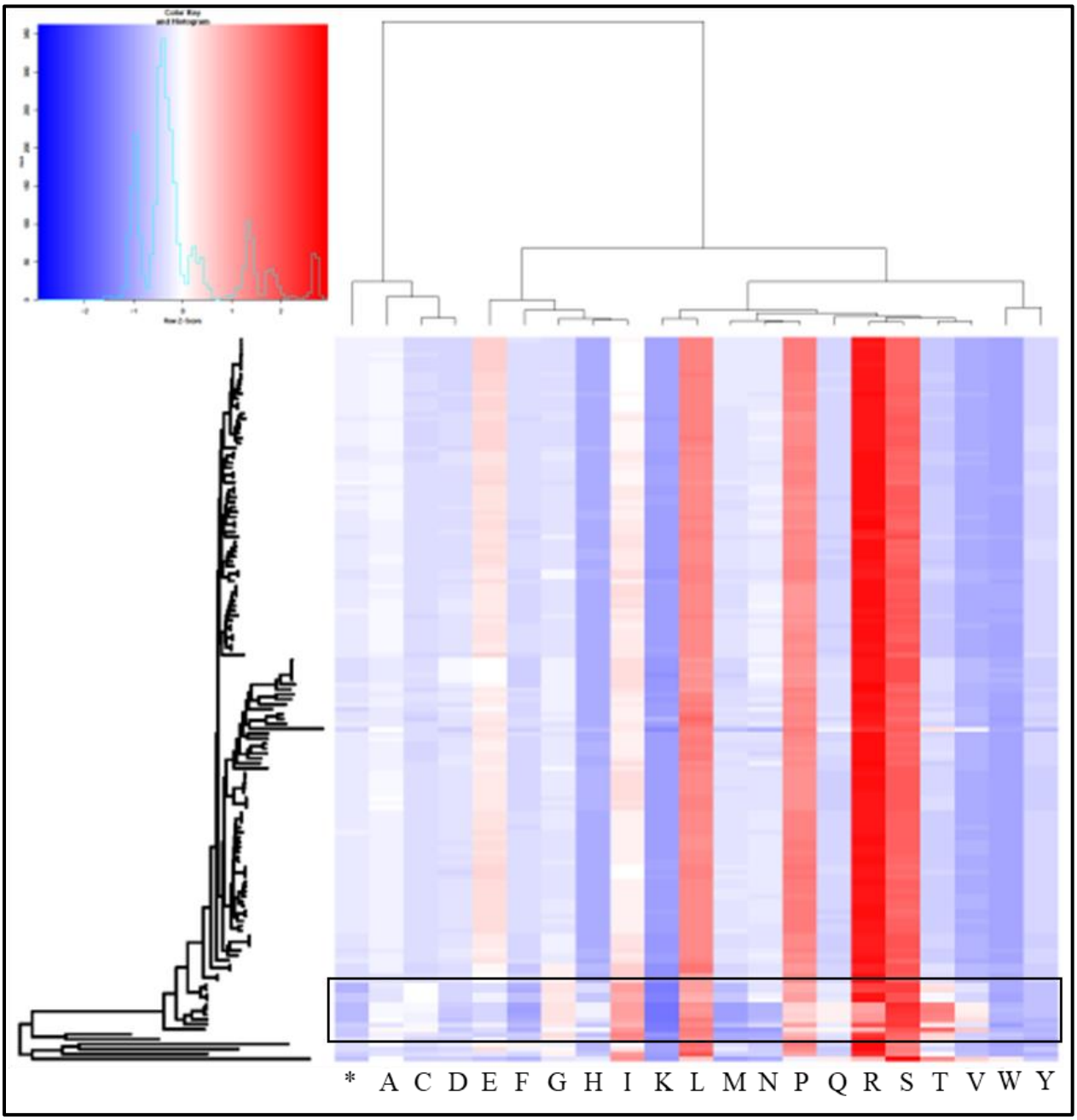


Figure 30. **Heatmap displaying the standardized frequencies of the aminos acids (x-axis) used in the MG-I MAGs** (y-axis) Blue indicates lower usage and red indicates higher usage.

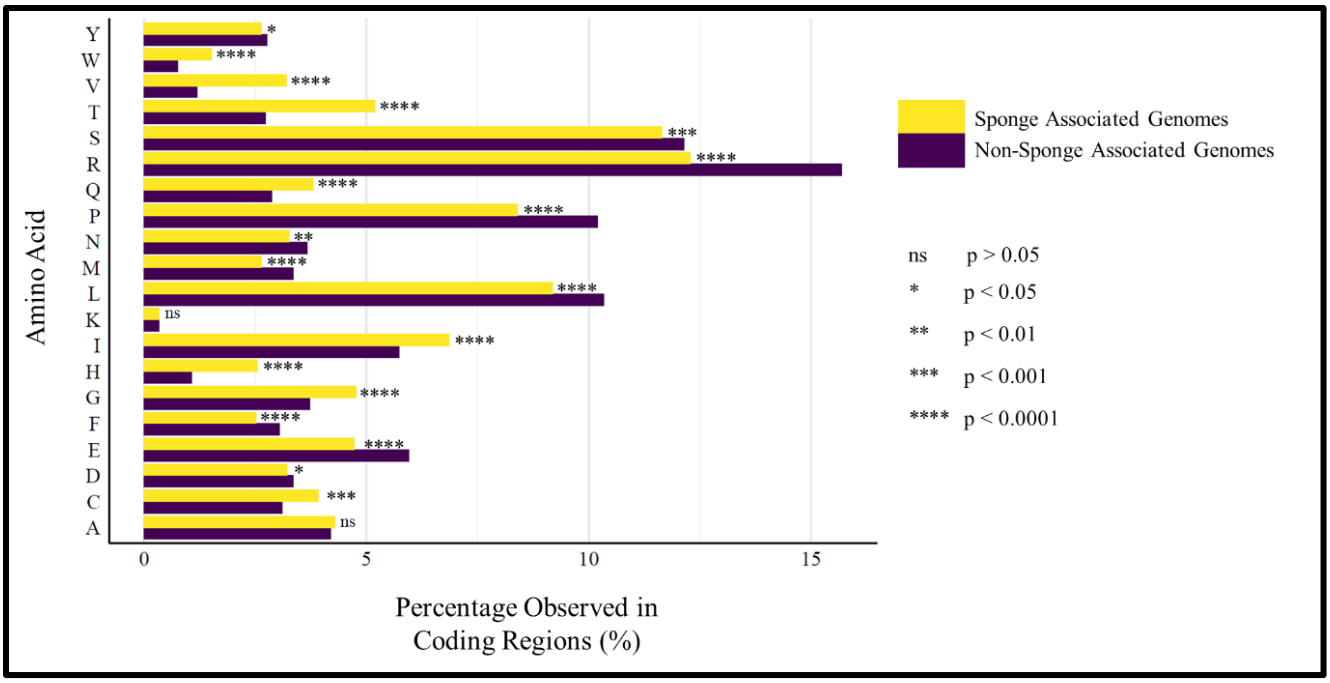


Figure 31. Bar plot displaying the percentage of each amino acid observed in the genomes of the sponge-associated MAGs and non-sponge associated MAGs.

To further explore these differences, codon usage by each group was analyzed. It is known now that these organisms show preferential expression of different amino acids, however, the codons used to encode these amino acids were also different, specifically in the wobble position of the codon. Figure 32 shows that the SAMs differed in their codon usage compared to the MG-I. The SAMs also had higher overall genome GC content (Figure 33A) than the others. Additionally, within the coding regions of the genomes, the SAMs showed a preference towards codons with a higher GC content (Figure 33B).

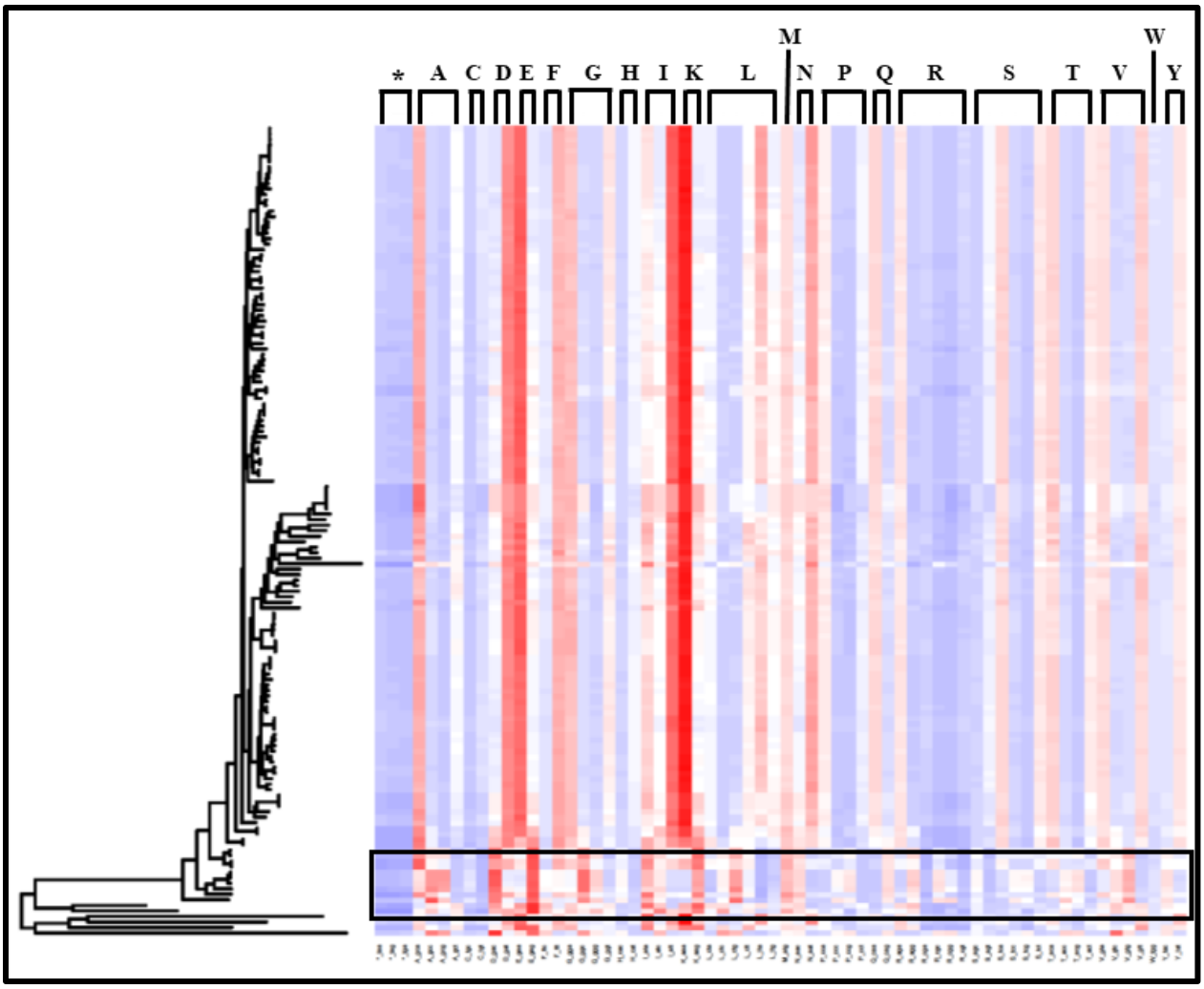


Figure 32. Heatmap of codon usage in MG-I. Y-axis is the phylogenetic tree of the MG-I genomes. X-axis are the codons differentiated by the amino acids these encode. The amino acids are listed at the top as their single letter code. The black box highlights the sponge-associated genomes.

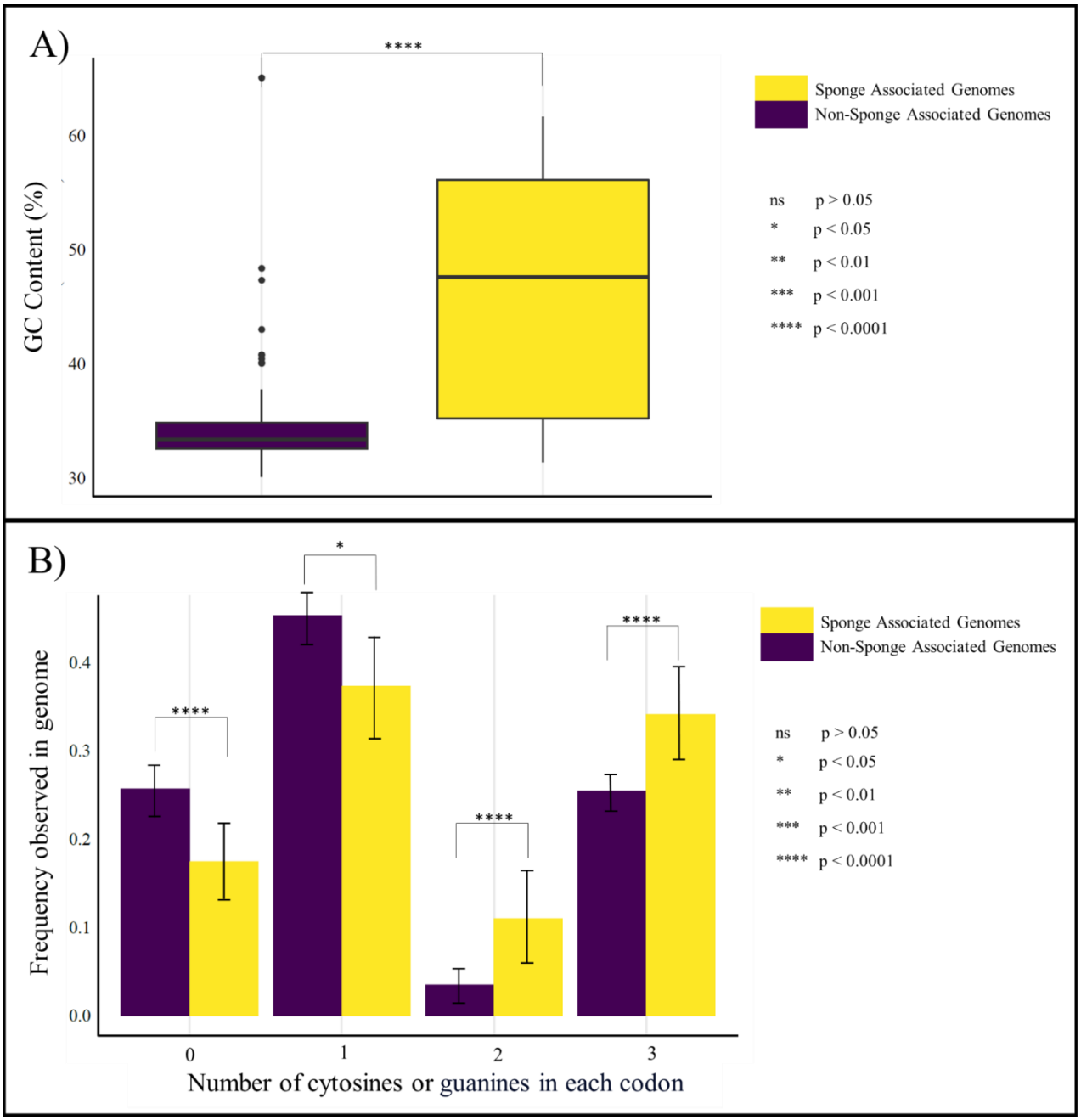


Figure 33. **GC content of the sponge associated MAGs compared to the other members of MG-I**: A) overall genomic GC content. B) the frequency of codons observed in the MAGs which have either zero, one, two or three guanines or cytosines.

3.3.4 MG-I PHYLOGENY IS NOT CONSISTENT WITH GENE CONTENT

MG-I evolution has been a continuing topic throughout this chapter. It was observed that there were very little differences in the six amino acid indices measures and that expected trends

for phylogeny were not followed. The expected trends for the phylogenetic and environmental niches displayed varied patterns for the indices, although not associated with cold adaptation. In light of these variations, the presence/absence of archaeal global protein families (GPFs) was analyzed to determine changes in proteomic composition that may affect cold adaptation (Figure 34). The MAGs grouped into two large clusters, with the latter dominated by lower quality MAGs. This latter group also had a greater proportion of absent gene families. Associated genomes with similar gene content could be clustered together because they either share a common ancestor or because the genomes are not complete. The relationship between the phylogenetic grouping of the MAGs and distance measurements of the global protein family presence or absence data was also analyzed (Figure 35), showing that the phylogenetic grouping is not consistent with the distance measurements. Within the cluster of good-quality MAGs, the phylogeny is determining the dendrogram structure.

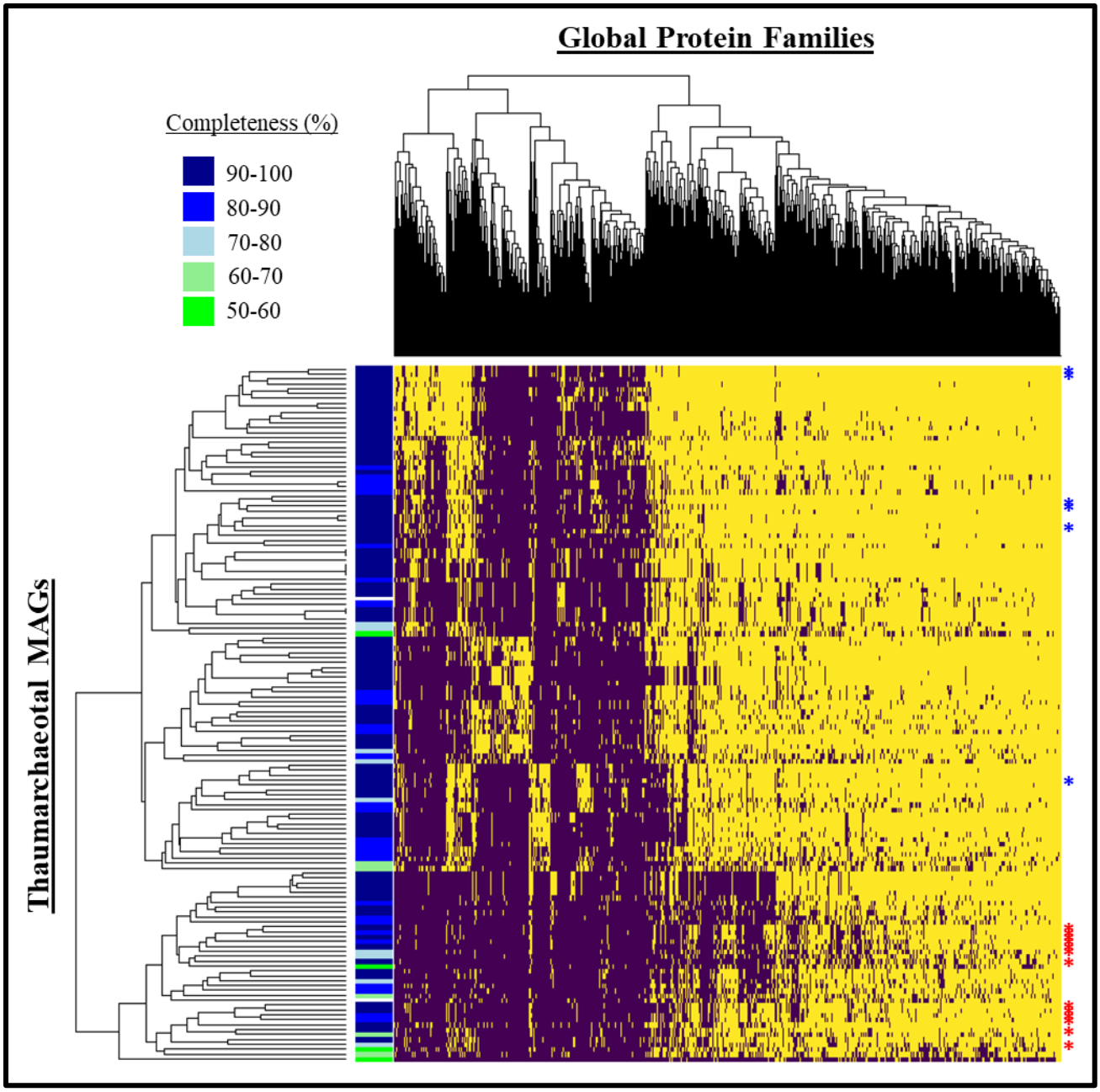


Figure 34. Presence/absence heatmap for the MG-I genomes. Yellow indicates presence and purple eindicates absence. Coloured bar on the left indicates completeness. Sponge-associated MAGs (SAMs) belonging to MG-I phylogenetic group A (blue) and B (red) are identified by asterisks (*) in the right-hand side.

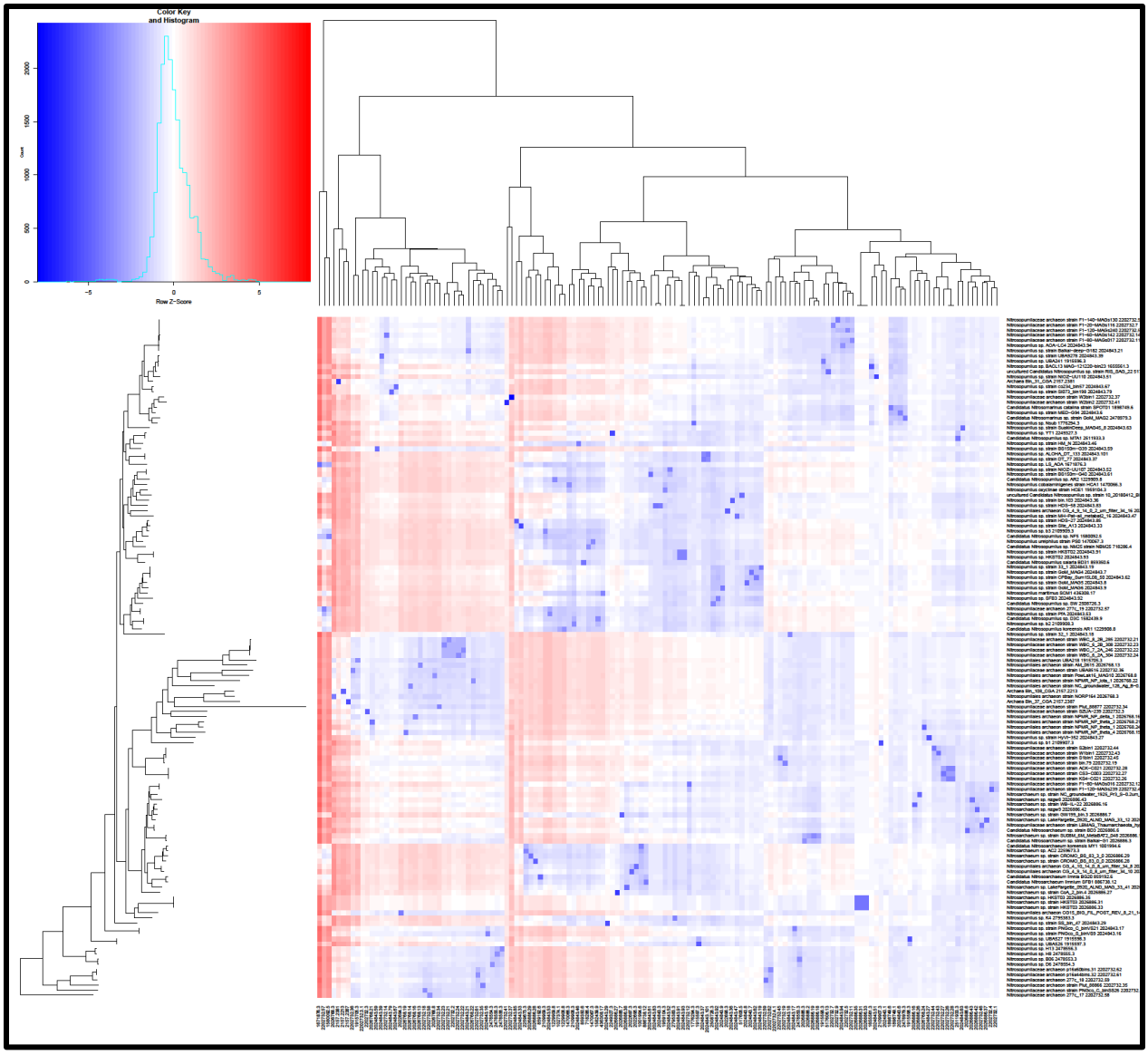


Figure 35. Heat map of phylogeny (y-axis) and presence/absence of global protein distances (x-axis) for the MG-I. Blue indicates a lower distance and red indicates a larger distance between MAGs. Dark blue indicates a direct comparison between the same MAG.

The index distances and phylogenetic distances between all of the MG-I MAGs was analyzed (Figure 36). The outgroups (one representative is highlighted in grey) are the most phylogenetically distant, as expected. One of the SAMs was chosen (2202732.34) as a representative to display that they were more phylogenetically related to the MG-I than the outgroups, but they generally were the most phylogenetically distant from the main population. The arginine to lysine ratio distances for

the SAMs were distinct, as many of them clustered in the top left signifying a greater distance in the index compared to other MG-I at similar phylogenetic distances. In Chapter 3.3.1. (Figure 23) (a zoomed in portion of the same figure) it was observed that Western Arctic archaeal representatives were closer to the general MG-I population in both index distances and phylogenetic distances.

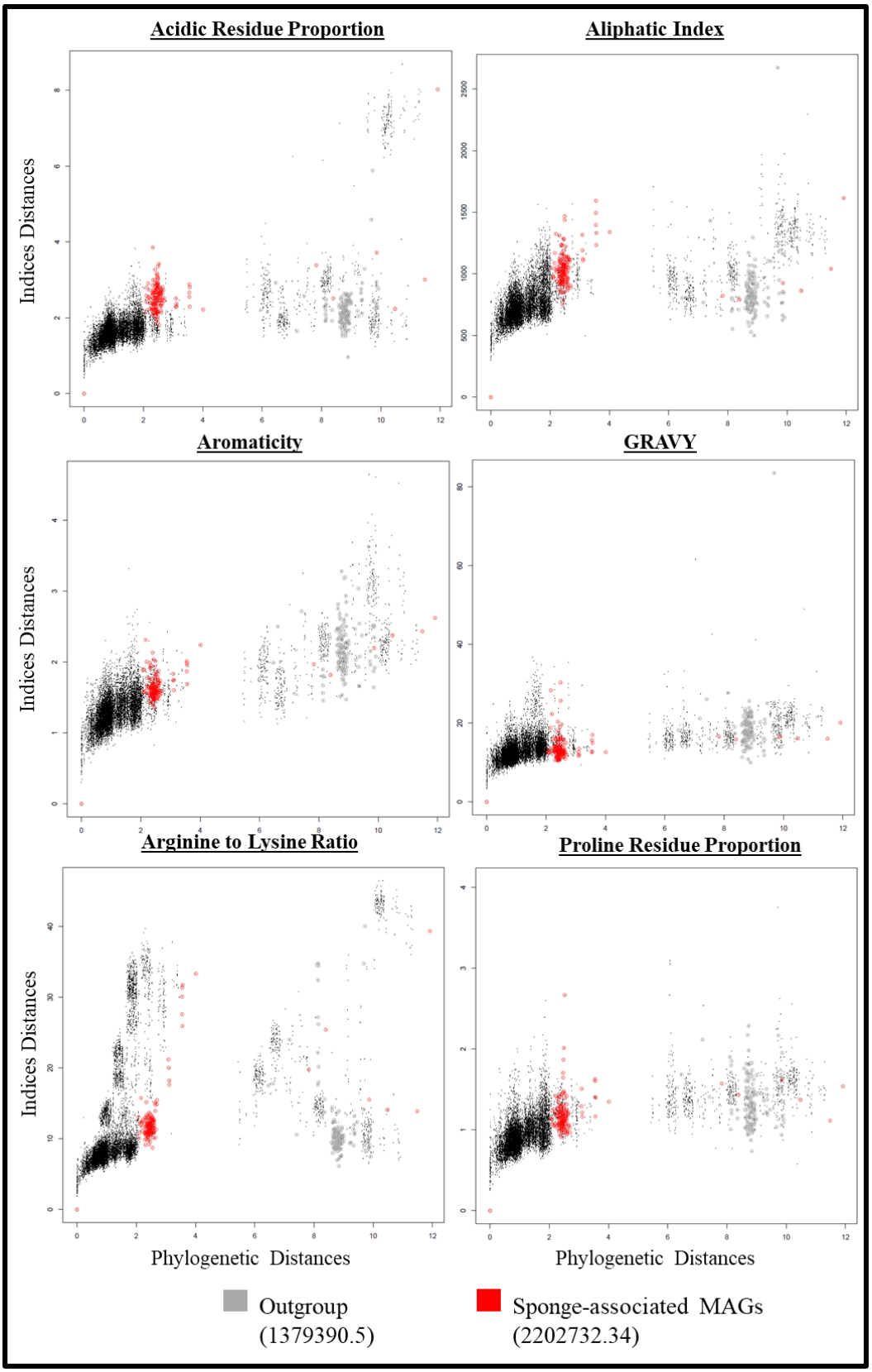


Figure 36. **Indices distances against the phylogenetic distances for MG-I.** One SAM (genome ID: 2202732.34) is highlighted here alongside one of the outgroup MAGs (genome ID: 1379390.5).

3.4 DISCUSSION

3.4.1 ASSESSING THE EFFICACY OF THE AMINO ACID COMPOSITIONAL INDICES FOR ASSESSING COLD ADAPTATION IN ARCHAEA

Inferred cold adaptation in Archaea was measured using six established amino acid compositional indices: arginine to lysine ratio, acidic residue proportion, GRAVY index, aliphatic index, proline residue proportion, and aromaticity index. . The hypotheses were that there would be signatures of temperature-adaptation in the proteomes of marine Archaea and that environmental temperature would be a factor influencing the gene content and amino acid usage in Archaea. Based on analyses perform with Bacteria (Gentilhomme, 2020), it was expected that the amino acid indices would be lower in MAGs from low temperature environments and that the phylogeny would reflect the environmental niches of marine Archaea. Since a large collection of archaeal MAGs was being analyzed, with diverse sampling locations, sources, seasons, and years, it was expected that there would be a high level of variation amongst the MAGs that would be consistent with their phylogenetics and environment niches. The phylogenetic and amino acid compositional index distances were calculated (Figure 23 & 24) and showed that the Arctic MG-I MAGs (Chapter 2) were all well within the normal distribution of distances in both the index distances and phylogenetic distances. Since these MAGs were collected from an environment with lower temperatures, it was expected that the distances for the indices would be greater for them compared to the rest of the population. Therefore, it was concluded that these indices did not show apparent signatures of cold adaptation in Archaea. Potentially this speaks to the efficacy of the indices themselves. A review study (Li et al., 2019) that looked at a collection of different genomic and proteomic indices and metrics for inferring optimal growth temperature (OGT) found that although most of these tools are quite good for high temperatures, they do not behave the same

way for psychrophiles. For example, GC content is often used as a metric to measure heat-adaptation, where higher GC contents are seen in thermophilic (heat loving) microorganisms. However, in psychrophilic microbes GC content does not correlate with optimal growth temperature (Li et al. 2019). The explanation as to why is unclear.

Although the Arctic MG-I MAGs did not show differences based on the amino acid indices tested, there were other parameters that did seem to influence the amino acid usage in this group. The first factor was the phylogenetics. The MG-I organisms were organized phylogenetically into two groups. It was observed that four of the indices tested were significantly different between the groups (Figure 25). It is important to note that the three Arctic MG-I MAGs included in this analysis were found in both of the phylogenetic groups. Additionally, it was determined that these differences were likely a results of the SAMs incorrected clustering with phylogenetic group B due to long-branch attraction. This will be discussed in the following section. The second factor was the environment or source of collection. The collection sources were categorized as marine water, fresh water, or other. It was shown that there were significant differences in all indices between the marine water and the freshwater MG-I MAGs in at least one of the phylogenetic groups (Figure 26). This observation was consistent within the two phylogenetic groups as the aliphatic ($P<0.0001$), acidic residue proportion ($P<0.0001$), GRAVY ($P<0.0001$), and proline residue proportion ($P<0.001$) indices were significantly different in Group A, and the arginine to lysine ratio ($P<0.0001$), aliphatic ($P<0.0001$), aromaticity ($P<0.001$), and proline residue proportion ($P<0.0001$) indices were significantly different within Group B. This suggests that the environmental conditions may be correlated with the amino acid composition seen in MG-I. This observation also implies that there in possible convergent evolution occurring where organisms from different lineages are obtaining similar evolutionary traits based on their environments.

It was observed that the Euryarchaeota and the MG-I displayed very different patterns of amino acid usage. While there was little variation between the MG-I clades after the removal of the SAMs, the Euryarchaeota had very distinct patterns based on five phylogenetic groups. As an example, the first and largest cluster (Group A) which contained members of the class Halobacteria displayed higher arginine to lysine ratios (0.642 +/- 0.182) than the other groups (Table 3). All members of this class are halophilic (salt-loving) and are known to have generally higher arginine to lysine ratios (Fukuchi et al., 2003) likely due to the protein stability required in a highly saline environment. While there were differences between the different clades of Euryarchaeota, there were very little differences within the MG-I. Recently, it has been proposed that the Euryarchaeota should actually be reclassified from one phylum into five different phyla (Rinke et al., 2020). This new classification was consistent with the phylogenetic grouping seen here. The analysis performed on the Euryarchaeota was consequently performed at the superphylum level, rather than at the phylum level, following this new classification. Rinke et al. (2020) also suggested that the Thaumarchaeota (MG-I), which are members of the TACK superphylum, should be reclassified. They further suggested that the entire TACK superphylum should be combined to create one phylum. This also changes the context in which these results are understood. The resolution at which these indices remain useful may be at a higher taxonomic level in Archaea.

These indices have also been tested in Bacteria, such as Cowelliaceae, and were found to show distinctive differences within the family (Gentilhomme, 2020). In that analysis, DNA from isolated strains of Colwelliaceae was sequenced and analyzed in comparison to their OGT. It was found that the amino acid compositional indices tested were different within the different clades of Colwelliaceae. This is inconsistent with the archaeal results even with a much higher phylogenetic resolution in the Colwelliaceae study. In that study, the indices tested were all

correlated with OGT, however, not all were positively correlated, as was expected (Gentilhomme, 2020). Aliphatic and GRAVY indices were negatively correlated with OGT in Colwelliaceae. These indices are treated as independent measures in this analysis although the underpinning molecular processes are interconnected. The negative correlations indicate the presence of more hydrophobic amino acids in these genomes, driven by a greater proportion of short-chained hydrophobic amino acids. Additionally, only three of the six indices were moderately or strongly associated with OGT (GRAVY, proline residue proportion, and aliphatic index).

Although these indices were useful in determining the proteomic characteristics within the different clades of Colwelliaceae, they continued to lack efficacy in determining genomic cold adaptation characteristics in Archaea. They may not be the most useful indices for determining cold adaptation because they did not take into consideration the interconnections between the indices themselves as well as other aspects of the genomes that are influencing cold adaptation. There are other components within the cells that provide cold tolerance or cold adaptation that were not yet discussed here. These include alterations in signal transduction, nucleotide-, co-enzyme-, and amino acid metabolism, defense mechanisms, energy production and conversion, DNA replication and processing, RNA and protein synthesis, cellular division, tRNA modifications, motility, and secretion (Cavicchioli, 2006). As an individual tool, these indices do not capture the entirety of cold adaptation mechanisms in microbes.

3.4.2 HOST-ADAPTATION IN ARCHAEA

During the phylogenetic analysis of the MG-I organisms, it was observed that there was one group of Nitrosopumiliceae that were incorrectly grouping with the Group B MG-I. In phylogenetics, this observation is called long-branch attraction where distantly related lineages are

often incorrectly associated with lineages that they are not related to (Susko & Roger, 2021). This occurs when there are enough changes or differences in the genome that they appear similar to another long-branch lineage. Rather than actually being closely related to one another, they are grouped together solely because they have both undergone a high degree of molecular changes. This bias is partially understood and is likely caused by assuming an overly simplistic model of evolution. This explains why some of the SAMs clustered together and separately from their *Nitrosopumilus* family members. The state changes likely influencing this phylogenetic outcome will be discussed in this section.

This same group was also shown to have unusually high arginine to lysine ratios (Figure 29), and it was determined that they had all been isolated from within or on sea sponges. Further analysis showed that there were significant differences in all but two amino acids in the proteomes of the SAMs when compared to the general population (Figure 31). There were increased in tryptophan which is a very large amino acid with an indole ring which can stabilize protein structures and domains. While there was a decrease in phenylalanine which produces rigidity within a protein and reduces flexibility. Histidine was used more frequently than arginine, both basic amino acids. These results suggest a preference for smaller aliphatic side chains and aromatic properties that can increase the conformational stability of the proteins. Valine and isoleucine were also more frequently used, having shorter aliphatic side chains than methionine and leucine which were not used as frequently. Glycine and proline both influence the kinetics of protein folding involving loops (the patternless regions that connect secondary structures to one another, such as beta sheets and alpha helices). Glycine, observed more frequently, is often found to accelerate this process as its small size destabilizes the secondary structure to form a loop. Also due to its smaller size, it has the ability to create bends within the protein allowing for flexibility and movement.

Alternatively, proline, which was observed less frequently, disrupts these secondary structures by lacking the ability to form hydrogen bonds creating a rigid break in the secondary structure. An increase in the use of cysteine suggests a push towards more stability with its ability to form disulfide bonds. Overall, there appears to be a selection for increase protein flexibility, while maintain some level of moderate stability.

Additionally, it was seen that the SAMs also had very high GC content in their DNA compared to the general population (Figure 33). Archaeal amino acid composition is generally dominated by a GC bias but can be further influenced by environmental factors (Roy Chowdhury & Dutta, 2012). GC content is often associated with a higher level of DNA stability to prevent denaturation, but also as a possible defense mechanism against predation (Slaby et al., 2017). Microbes often have defense mechanisms to detect foreign DNA from viruses. These mechanisms often work by using the GC content of incoming foreign DNA as a determinant for whether or not to degrade that DNA. Having a higher microbial GC content can allow the cell to differentiate foreign DNA from self DNA.

Sponges consist of channels that filter thousands of litres of water daily introducing foreign DNA, viruses, nutrients, and contaminants (Kenny et al., 2020; Slaby et al., 2017). It had been shown in Bacteria that that sponge symbionts have greater expression levels of genes associated with defense, amino acid transport and metabolism, recombination, and repair. (Slaby et al. 2017). Although archaeal representatives from the Crenarchaeota have been identified with sea sponges, the gene content and expression of these organisms have not been explored. It appears that the differences in amino acid usage and the higher GC content observed in the SAMs are consistent with bacterial symbionts from sea sponges. It is plausible to assume that the other state changes occurring in Bacteria may also be occurring in Archaea in response to sea sponge host adaptation.

For example, higher expression of genes involved in recombination are often seen in bacterial symbionts as a mechanism for horizontal gene transfer. Horizontal gene transfer (HGT) is the uptake of DNA via direct cell to cell contact, free DNA in the environment, or viral infection. Microbes can quickly pick up genes from other members in their community that can provide the necessary tools to withstand the conditions. Following this logic, this could explain the long-branch attraction that was observed in the SAMs as a result of horizontally transfers genes contributing to their diversity and rates of evolution. The differences observed in the SAM's amino acid usage, codon usage, GC content, and index and phylogenetic distances suggests that the environmental conditions of the host-adaptation SAMs have influenced the evolution of these organisms.

3.4.3 EVOLUTION OF MG-I

As previously discussed, the phylogenetics (evolutionary history) of this portion of the project were not consistent with the taxonomy (classification or naming system). The amino acid indices tested in the proteomes of the MG-I did not have any distinct trends with regards to the phylogenetic grouping. The relationship between the phylogenetic grouping of the MAGs and distance measurements of the global protein family presence or absence data was also analyzed (Figure 36), showing that the phylogenetic grouping was not consistent with the distance measurements. One explanation for why this may be occurring is that the number of genes shared between highly distant MAGs is low, but they are similar because those genes require genes (*i.e.* DNA machinery; protein synthesis; etc). Another explanation is that clusters of blue in Figure 34 represent two genomic states: one cluster of MAGs sharing a last common ancestor and one that

contains MAGs with low completeness. Within the cluster of good-quality MAGs, the phylogeny is determining the dendrogram structure.

Although it was determined that these indices were not effective in identifying cold adaptation, they did provide insight into the evolutionary status and evolutionary rates of the MG-I. There are two explanations for these observations: that the MG-I are evolving very slowly or are evolving very quickly. The rates of evolution in Archaea in general may be different than Bacteria simply because they harbor different DNA replication machinery. Many of the DNA replication components found in Archaea are more similar to those in eukaryotic cells than those in bacterial cells (Dey et al., 2016; Klinger et al., 2016). If the MG-I are evolving much slower than other microbes in their environment it may mean that they maintain highly conserved genomes because they grow very slowly. R strategists are organisms that live in unstable environments and will typically undergo rapid reproduction in order to stabilize themselves, while K strategists typically live in stable environments, typically found in large populations, and do not need to undergo the same rapid rates of replication. These differences in lifestyle will affect how they structure their proteins. It could also mean that they grow solitarily to the point where they are not competing with other organisms. This second explanation is less likely because there have been strong interactions observed between the MG-I and picoplankton for ammonia competition (Martens-Habbena et al., 2009). If the MG-I growth rates are very low, then it could mean that they don't have time to "catch-up" once in a colder Arctic environment. It could also mean there is not enough pressure to change and that they already harbour the necessary machinery to survive, but not thrive.

If the alternative explanation is true, that the MG-I are actually evolving at very high rates, it means that they are obtaining adaptations through other means. Archaea and Bacteria share similar genome sizes that are both gene dense, however a recent study (Kellner et al., 2018) found

that bacterial genome size changed in a larger range than in Archaea and that the overall diversity in Archaea was much lower. Ultimately, there may be other factors that are more important for adaptation depending on the microbial community and the environment. Although we did not see evidence of cold adaptation, there may be other environmental or biological factors that are stronger evolutionary influences. Overall, it appears that at a proteomic level these organisms are fairly conserved. One exception to this claim that was observed was with the MAGs associated with sea sponges.

CHAPTER 4. CONCLUSION

4.1 SUMMARY OF MAJOR FINDINGS

In this thesis the biogeography of Western Arctic Ocean archaeal MAGs was explored. In addition, the MAGs were used in a larger analysis of cold adaptation signatures in Archaea. The objectives of this thesis were to characterize the diversity and distribution of archaeal communities residing in Arctic seawater, and to identify signatures of cold adaptation in the proteomes of Archaea found in cold Arctic seawater. The first objective was accomplished by identifying archaeal MAGs collected from Arctic seawater concerning their biogeography and applying environmental context to explain relative abundance distributions. The second objective was accomplished by investigating archaeal evolution by analyzing genomic and proteomic signatures within their genomes. The hypotheses for this thesis were that water mass layers in the Arctic Ocean would be the strong factor influencing the biogeography of these organisms, that there would be signatures of cold adaptation in the proteomes of these organisms, and that the environmental niches would influence the phylogenetic organization of these organisms.

Through the analysis of the relative abundance distributions of the individual archaeal MAGs, this thesis expanded on our current understanding of archaeal marine biogeography. While it was currently suggested that the three marine had specific and consistent distributions, it has been shown here, with increased phylogenetic resolution, expanded sampling depths, expanded water mass characterizations, and the use of metagenomic tools, that bins within each marine group displayed variability in their abundance mapping, suggesting that the previous claim that these traits are specific and consistent are not appropriate statements for this natural model.

Although the use of the established amino acid compositional indices for the detection of cold adaptation did not prove to be affective in Archaea, these measures did shed light into the evolutionary history of these organisms in terms of environmental adaptation. It did not appear that the Arctic MAGs had significant changes in their amino acid composition and were very similar to other MG-I organisms found in other, more temperate locations. This could suggest that these organisms are very conserved and do not evolve quickly. It could also suggest that these organisms are simply migrating through the Arctic Ocean. The rates of evolution are slowed here, due to lower temperatures that slow reactions rates and replications times. It was observed that the majority of MG-I MAGs were fairly similar to one another, with the exceptions of those associated with sea sponges, suggesting that host-adaptation provides an evolutionary pressure.

This thesis contributes to the understanding of archaeal ecology, especially in the context of an Arctic environment. Although previously stated in many studies, it appears that phylogenetics are a weaker factor influencing the biogeography of marine Archaea in the Arctic, as all three marine groups had variations in their abundance distributions. It was determined that the established amino acid compositional indices commonly used to assess cold adaptation in Bacteria do not appear to be as useful with Archaea. This analysis also provided more information

regarding the evolutionary states of MG-I organisms. Overall, this thesis contributes to natural model building in terms of Archaea in marine environments, specifically in the Arctic Ocean, and the assessment of cold adaptation which can lead to the development of techniques to infer optimal growth temperature. These findings in conjunction can be used to create numerical models of the biological system. Although, it appears that these indices were not useful in Archaea, this information remains useful in the larger picture of inferring optimal growth temperature, cold adaption, and host adaptation.

4.2 RECOMMENDATIONS FOR FUTURE RESEARCH

To further understand the ecological context of Arctic Archaea, determining absolute abundance measurements is the first step. Additionally, further exploration of the environmental parameters associated with the different water masses that may be influencing the abundance distributions will be helpful. As was discussed above, there are also biological factors that influence microbial communities. Investigating the co-occurrence of different taxa in the environment would be the first step in understanding the biological interconnectedness.

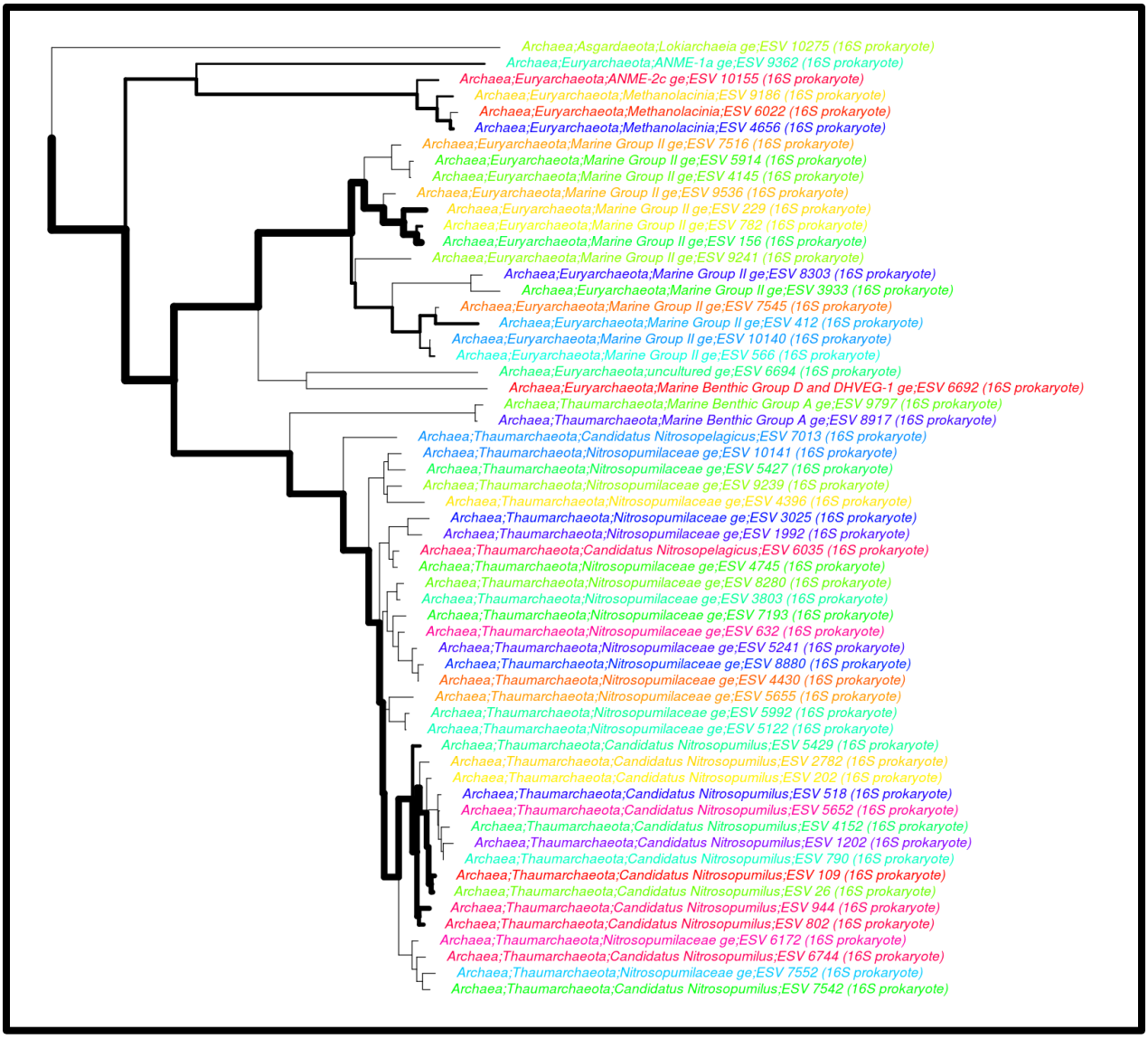
Being able to infer preferred growth temperatures of microorganisms is important, especially for uncultured taxa. Although the amino acid compositional indices did not prove to be a useful tool in this manner, there may be other metrics or indices that can be explored. Wang et al. (2019) recently looked at the differences in prevalence of certain protein domains in a large collection of bacterial proteomes to understand their niche adapted lifestyles. This work was done with bacterial halophiles and thermophiles but was not yet done in psychrophilic Bacteria or Archaea. Since there were differences observed in the former environments, it may be worthwhile

to look for domain differences in psychrophiles as well. This analysis could be done in Archaea, however there is little data on archaeal domains.

Potentially taking a broader approach that considers more of the connected molecular systems rather than individual indices may be helpful. It has also been discussed that there are many other aspects of cellular biology that are affected by lower temperatures. These include the presence of cold shock proteins, modifications in the lipid membrane, and DNA replication machinery. In this analysis, I looked at cold adaptation from a larger taxa point of view, but it could be useful to further investigate the archaeal taxa from the Western Arctic. One way to accomplish this is to attempt to culture archaeal representatives from this environment and experimentally determine each of their OGTs, which can be used to correlate various indices to OGT.

To understand the evolution of the MG-I to a fuller extent, the degree to which horizontal gene transfer influences the genetic makeup of these organisms can be explored. For the sponge associated MAGs, it would be helpful to analyze the genomic composition and understand what it is about the sponge environment that is influencing the evolution of this group. It would also be a good idea to try to sequence and identify viral DNA to determine how much viral infection and predation affects these cells.

CHAPTER 5. APPENDIX



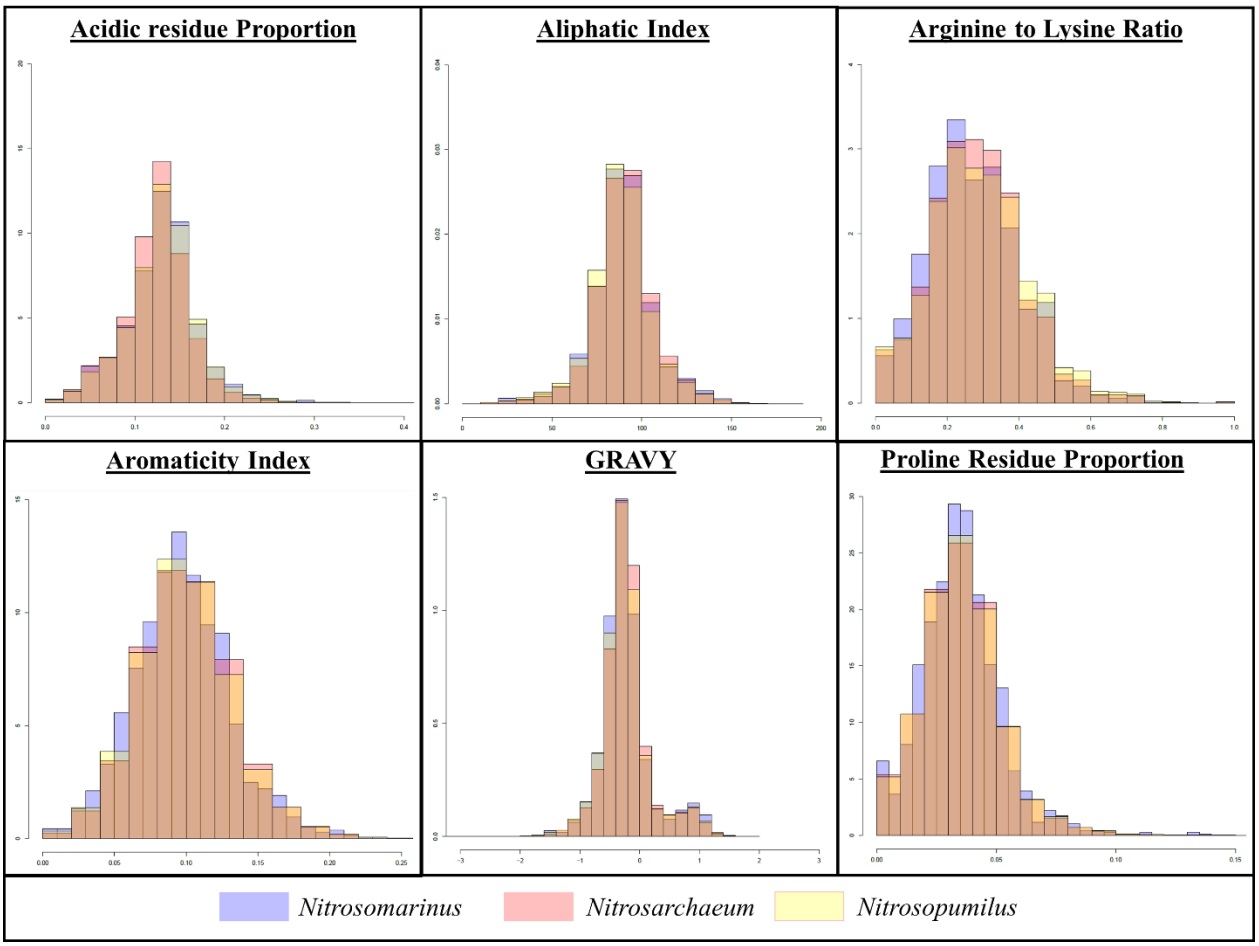
Supplementary Figure 1. **Phylogenetic tree of 16S rRNA gene sequencing data from samples collected in the Chukchi Sea by the Collins lab** (plotted using Microscape (http://cryomics.org/microscape). Metagenomic data from a subset of this sample set will be analyzed for this project. Each leaf on the tree is an exact sequence variant (ESV), corresponding approximately to the species or ecotype level. The line weights are proportional to the mean relative abundance of each ESV in the dataset. Branch lengths represent the relative evolutionary time between each node.

Supplementary Table 1. CheckM scores of completeness and contamination for MG-I MAGs used in the analysis in Chapter 4.

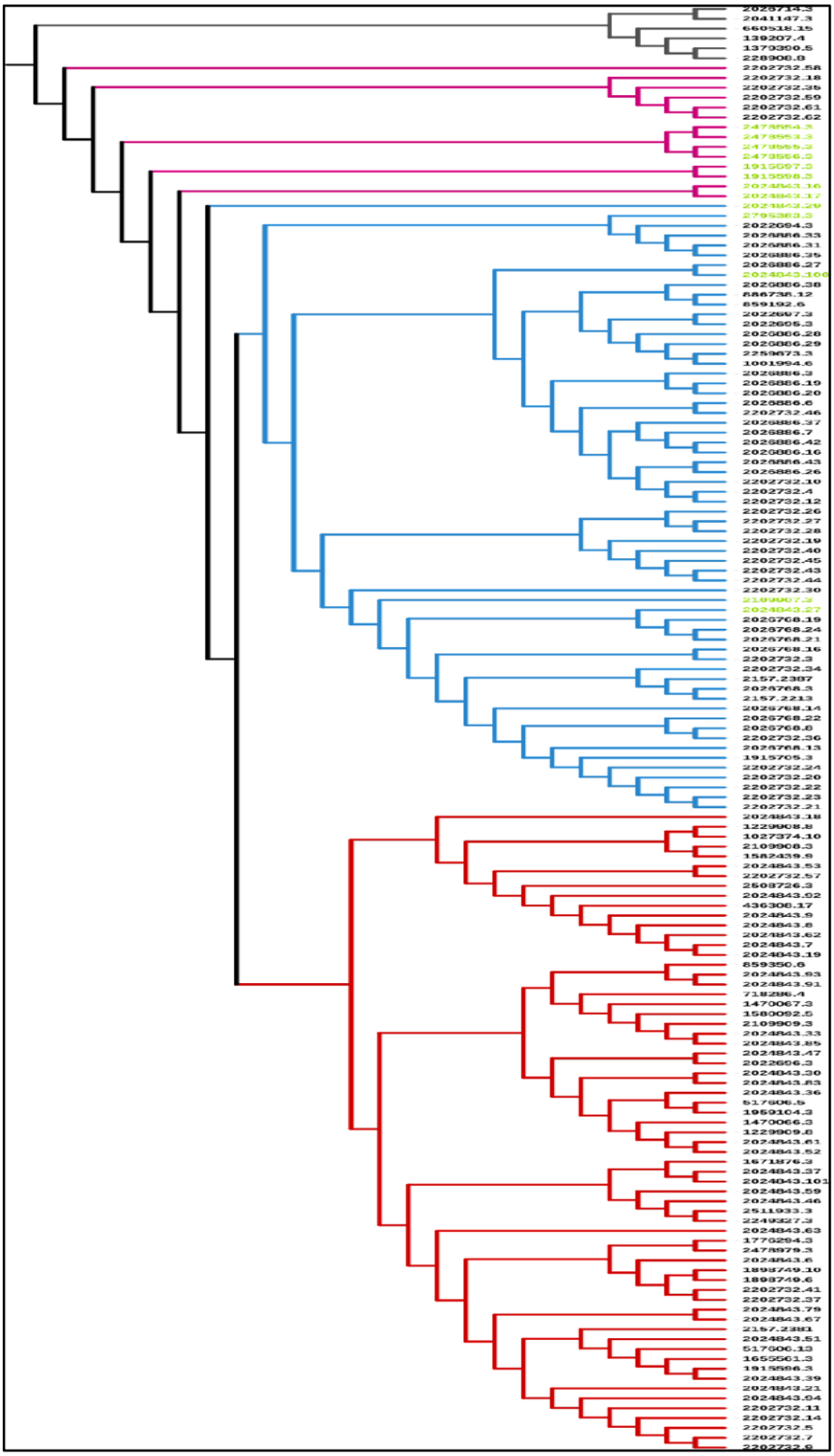
Bin Id	Completeness	Contamination
139207.4	99.5	0.5
228908.8	73.13	0
436308.17	100	0.97
517606.13	80.74	0.97
517606.5	95.63	7.61
660518.15	99.57	1.7
718286.4	100	0
859192.6	99.03	5.83
859350.6	92.39	1.94
886738.12	98.06	0
1001994.6	100	0
1229908.8	94.66	0
1229909.8	97.09	0
1379390.5	73.83	2.8
1470066.3	98.54	0
1470067.3	100	0
1580092.5	100	0
1582439.9	100	0.97
1655561.3	99.51	0.97
1671876.3	100	0
1776294.3	100	0
1898749.6	100	0
1915596.3	95.15	0
1915597.3	93.2	0
1915598.3	79.21	0
1915705.3	81.91	0
1959104.3	99.51	0
2022694.3	62.81	0.97
2022695.3	84.63	0.56
2022696.3	80.1	3.4
2022697.3	93.95	1.94
2024843.101	99.51	0.49
2024843.16	68.61	0.97
2024843.17	91.35	0.97
2024843.18	89.32	0
2024843.19	97.09	1.94
2024843.21	94.25	0
2024843.27	83.1	5.83
2024843.29	77.47	4.85
2024843.33	58.86	10.36
2024843.36	89	5.74

2024843.37	99.51	0.49
2024843.39	78.4	0
2024843.46	97.57	0
2024843.47	98.54	0.97
2024843.51	81.18	3.88
2024843.52	93.69	0.97
2024843.53	98.06	0
2024843.59	52.43	0
2024843.6	100	0
2024843.61	87.38	0.97
2024843.62	95.15	1.46
2024843.63	95.15	0.97
2024843.67	85.41	6.8
2024843.7	95	1.94
2024843.79	82.78	6.37
2024843.8	92.72	2.91
2024843.83	87.34	2.91
2024843.85	77.62	5.83
2024843.9	91.26	0.97
2024843.91	95.15	0
2024843.92	100	0.97
2024843.93	95.15	0
2024843.94	99.03	0
2026714.3	92.99	3.31
2026768.13	99.51	0.97
2026768.14	85.44	2.17
2026768.16	86.23	5.42
2026768.19	76.24	6.31
2026768.21	93.69	4.05
2026768.22	98.54	0.97
2026768.24	99.51	2.91
2026768.3	58.82	31.96
2026768.8	94.17	0
2026886.16	84.63	3.64
2026886.19	97.73	0.97
2026886.26	80.42	1.05
2026886.27	74.92	0.47
2026886.28	76.54	5.9
2026886.29	93.2	7.77
2026886.3	99.03	1.94
2026886.31	97.09	0
2026886.33	97.09	0
2026886.35	97.09	0
2026886.37	91.42	1.05
2026886.38	99.03	1.46
2026886.42	96.12	1.94
2026886.43	99.03	1.46
2026886.6	98.54	1.94
2026886.7	93.13	6.63

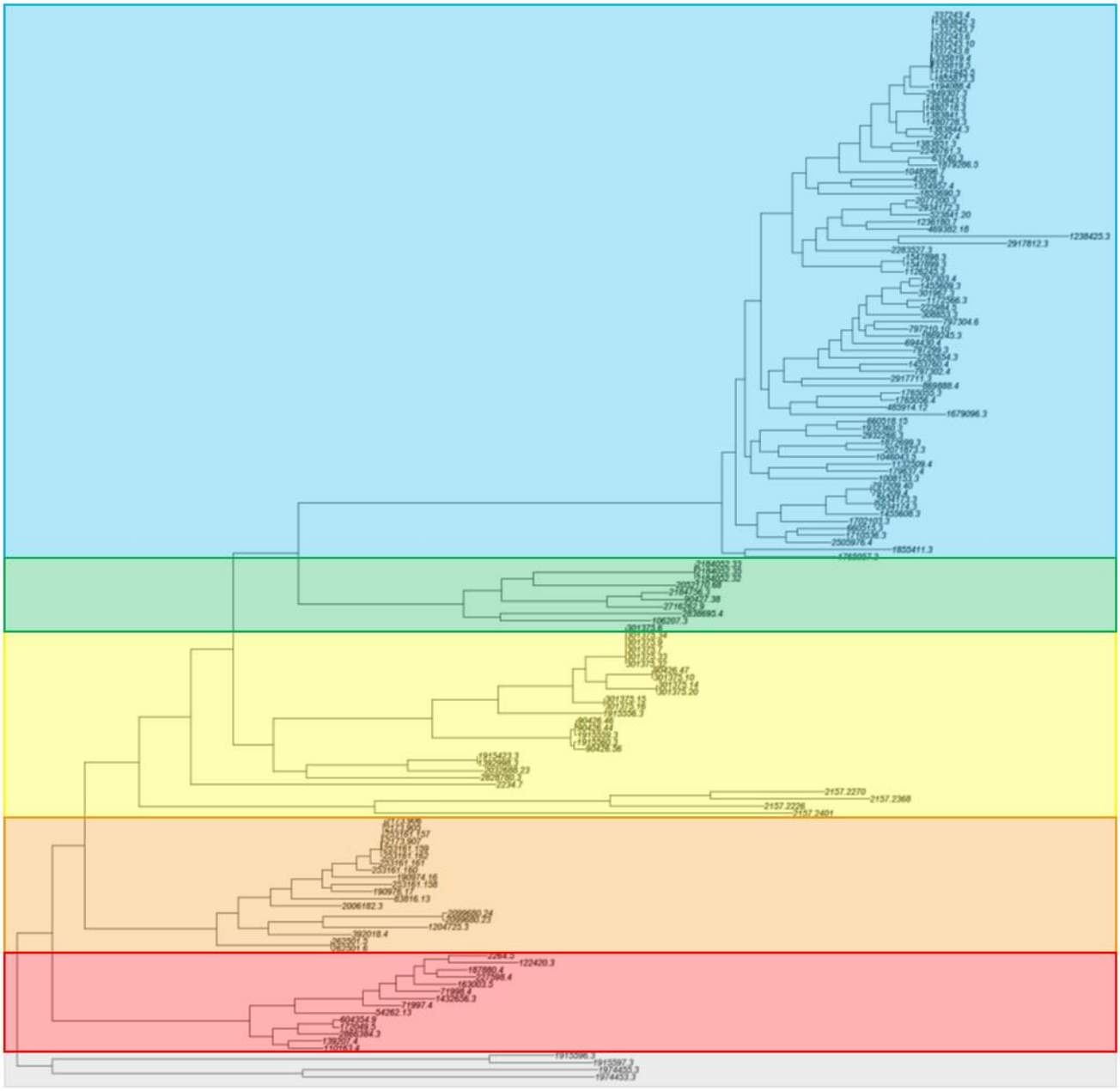
2041147.3	96.88	1.01
2109907.3	98.06	5.96
2109908.3	99.51	0.97
2109909.3	100	0
2202732.11	95.39	0
2202732.12	93.2	0
2202732.14	85.44	1.94
2202732.18	77.02	0.07
2202732.19	87.54	1.94
2202732.21	98.71	1.94
2202732.22	96.6	1.94
2202732.23	94.9	1.94
2202732.24	98.06	1.94
2202732.26	96.12	1.94
2202732.27	97.09	1.94
2202732.28	96.12	1.94
2202732.3	63.11	5.83
2202732.34	99.03	1.61
2202732.35	89.08	0.65
2202732.36	85.6	0
2202732.37	64.89	5.34
2202732.4	89.48	0
2202732.41	89.32	4.85
2202732.43	91.75	4.37
2202732.44	95.63	0.97
2202732.45	91.64	1.94
2202732.46	98.54	0.97
2202732.5	99.03	0.97
2202732.57	100	0.97
2202732.58	90.78	0
2202732.59	92.23	0.97
2202732.61	85.44	0.97
2202732.62	95.15	0.97
2202732.7	83.01	0.81
2202732.9	98.06	3.88
2249327.3	87.7	14.56
2259673.3	100	0
2478553.3	91.75	0.97
2478554.3	92.23	0
2478555.3	89.16	0
2478556.3	89.81	0.07
2478979.3	93.2	0
2508726.3	100	0.97
2511933.3	100	0
2795383.3	99.03	0



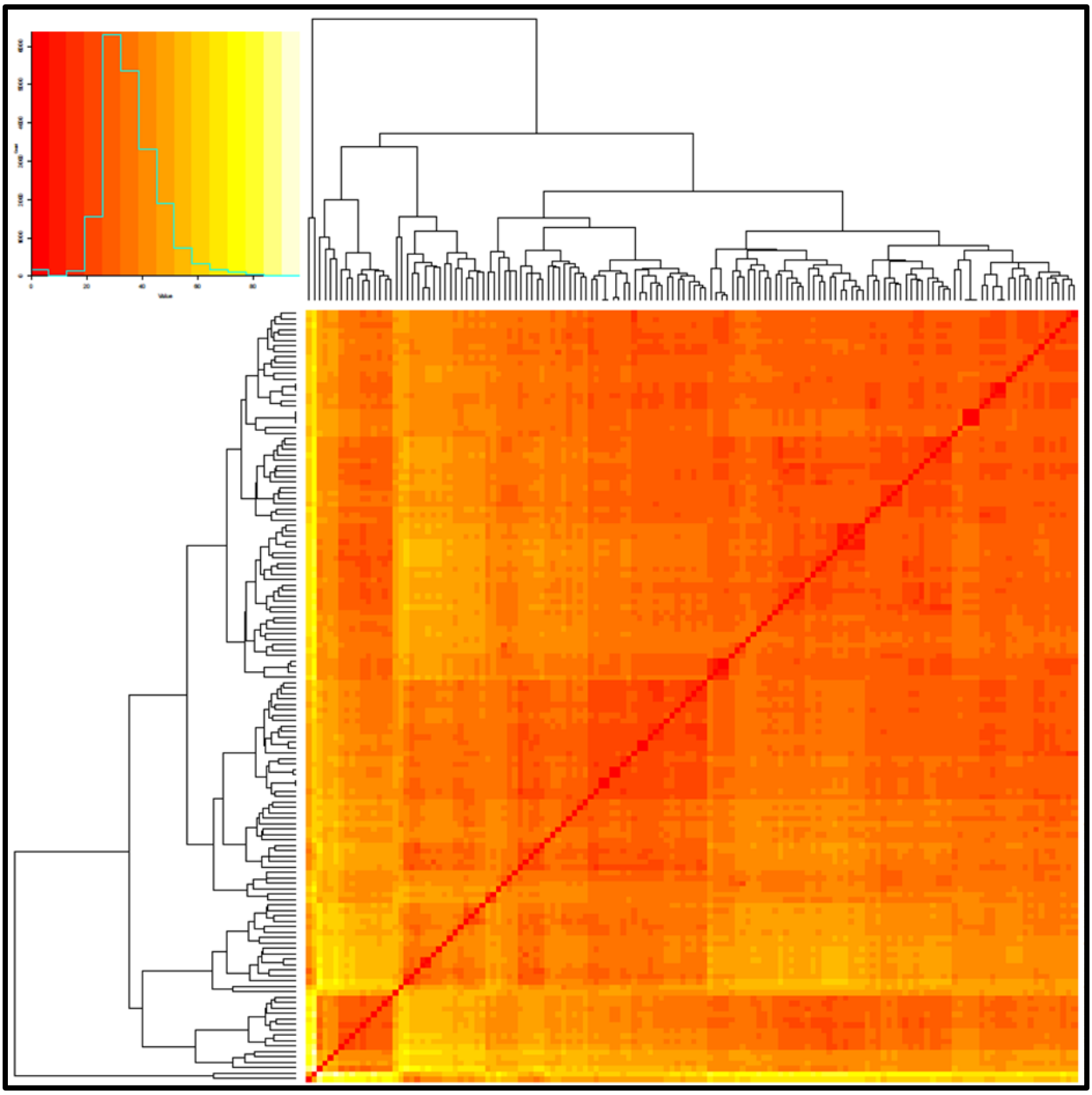
Supplementary Figure 2. Amino acid indices distribution for the three MG-I genera analyzed.



Supplementary Figure 3. Annotated phylogenetic tree of the MG-I MAGs analyzed in Chapter 4.2 and 4.4, including group A branches (blue), group B branches (red), Sponge-associated genomes (green), Nitrosopumilaceae outliers (pink), and the outgroup (grey).



Supplementary Figure 4. Annotated phylogenetic tree of the Euryarchaeota MAGs analyzed in Chapter 4.3 including Groups A (blue), B (green), C (yellow), D (orange), and E (red), and the outgroups (grey).



Supplementary Figure 5. Pair-wise distances of the presence or absence of archaeal global protein families in the MG-I MAGs. Both y- and x-axes are the same. And the dark red line indicates a direct MAG to MAG comparison.

CHAPTER 6. REFERENCES

Adam, P. S., Borrel, G., Brochier-Armanet, C., & Gribaldo, S. (2017). The growing tree of Archaea: New perspectives on their diversity, evolution and ecology. *ISME Journal*, *11*(11), 2407–2425. https://doi.org/10.1038/ismej.2017.122

Arandia-gorostidi, N., Krabberød, A. K., Logares, R., Mora, A. G., & Gonza, F. (2022). *Novel Interactions Between Phytoplankton and Bacteria Shape Microbial Seasonal Dynamics in Coastal Ocean Waters*. *9*(July), 1–14. https://doi.org/10.3389/fmars.2022.901201

Baer, S. E., Connelly, T. L., Sipler, R. E., Yager, P. L., & Bronk, D. A. (2014). Effect of temperature on rates of ammonium uptake and nitrification in the western coastal Arctic during winter, spring, and summer. *Global Biogeochemical Cycles*, (28), 1455–1466. https://doi.org/10.1002/2013GB004765

Baker, B. J., De Anda, V., Seitz, K. W., Dombrowski, N., Santoro, A. E., & Lloyd, K. G. (2020). Diversity, ecology and evolution of Archaea. *Nature Microbiology*, *5*(7), 887–900. https://doi.org/10.1038/s41564-020-0715-z

Bayer, K., Schmitt, S., & Hentschel, U. (2008). Physiology, phylogeny and in situ evidence for bacterial and archaeal nitrifiers in the marine sponge Aplysina aerophoba. *Environmental Microbiology*, *10*(11), 2942–2955. https://doi.org/10.1111/j.1462-2920.2008.01582.x

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., … Smith, T. P. L. (2022). Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nature Biotechnology*. https://doi.org/10.1038/s41587-021-01130-z

Box, J. E., Colgan, W. T., Christensen, T. R., Schmidt, N. M., Lund, M., Parmentier, F. J. W., … Olsen, M. S. (2019). Key indicators of Arctic climate change: 1971-2017. *Environmental Research Letters*, *14*(4). https://doi.org/10.1088/1748-9326/aafc1b

Brettin, T., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Olsen, G. J., … Xia, F. (2015). RASTtk: A modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Scientific Reports*, *5*. https://doi.org/10.1038/srep08365

Brochier-Armanet, C., Boussau, B., Gribaldo, S., & Forterre, P. (2008). Mesophilic crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nature Reviews Microbiology*, *6*(3), 245–252. https://doi.org/10.1038/nrmicro1852

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE*, *12*(10), 1–15. https://doi.org/10.1371/journal.pone.0185056

Cavicchioli, R. (2006). Cold-adapted archaea. *Nature Reviews Microbiology*, *4*(5), 331–343. https://doi.org/10.1038/nrmicro1390

Chang, B. X., & Devol, A. H. (2009). Seasonal and spatial patterns of sedimentary denitrification rates in the Chukchi sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, *56*(17), 1339–1350. https://doi.org/10.1016/j.dsr2.2008.10.024

Charif D., Lobry J., (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Bastolla U, Porto M, Roman H, Vendruscolo M (eds.), *Structural approaches to sequence evolution: Molecules, networks, populations*, series Biological and Medical Physics, Biomedical Engineering, 207-232. Springer Verlag, New York. ISBN : 978-3-540-35305-8.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics (Oxford, England)*, *38*(23), 5315–5316. https://doi.org/10.1093/bioinformatics/btac672

Cheng, L., Trenberth, K. E., Fasullo, J., Boyer, T., Abraham, J., & Zhu, J. (2017). Improved estimates of ocean heat content from 1960 to 2015. *Science Advances*, *3*(3), 1–10. https://doi.org/10.1126/sciadv.1601545

Christman, G. D., Cottrell, M. T., Popp, B. N., Gier, E., & Kirchman, D. L. (2011). Abundance, diversity, and activity of ammonia-oxidizing prokaryotes in the coastal arctic ocean in summer and winter. *Applied and Environmental Microbiology*, *77*(6), 2026–2034. https://doi.org/10.1128/AEM.01907-10

Church, M. J., DeLong, E. F., Ducklow, H. W., Karner, M. B., Preston, C. M., & Karl, D. M. (2003). Abundance and distribution of planktonic Archaea and Bacteria in the waters west of the Antarctic Peninsula. *Limnology and Oceanography*, *48*(5), 1893–1902. https://doi.org/10.4319/lo.2003.48.5.1893

Cimen, E., Jensen, S. E., & Buckler, E. S. (2021). Building a tRNA thermometer to estimate microbial adaptation to temperature. *Nucleic Acids Research*, *48*(21), 12004–12015. https://doi.org/10.1093/nar/gkaa1030

Collins, R. E., Rocap, G., & Deming, J. W. (2010). Persistence of bacterial and archaeal communities in sea ice through an Arctic winter. *Environmental Microbiology*, *12*(7), 1828–1841. https://doi.org/10.1111/j.1462-2920.2010.02179.x

Cutter, G. A., Casciotti, K., Croot, P., Geibert, W., Heimburger, L.-E., Lohan, M., … van de Flierdt, T. (2017). *Sampling and Sample-handling Protocols for GEOTRACES Cruises*.

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., … Stevens, R. (2020). The PATRIC Bioinformatics Resource Center: Expanding data and analysis capabilities. *Nucleic Acids Research*, *48*(D1), D606–D612. https://doi.org/10.1093/nar/gkz943

Degoeij, J. M., De Kluijver, A., Van Duyl, F. C., Vacelet, J., Wijffels, R. H., De Goeij, A. F. R. M., … Schutte, B. (2009). Cell kinetics of the marine sponge Halisarca caerulea reveal rapid cell turnover and shedding. *Journal of Experimental Biology*, *212*(23), 3892–3900. https://doi.org/10.1242/jeb.034561

Deutsch, C., Gruber, N., Key, R. M., Sarmiento, J. L., & Ganachaud, A. (2001). Denitrification and N2 fixation in the Pacific Ocean. *Global Biogeochemical Cycles*, *15*(2), 483–506.

Dey, G., Thattai, M., & Baum, B. (2016). On the Archaeal Origins of Eukaryotes and the Challenges of Inferring Phenotype from Genotype. *Trends in Cell Biology*, *26*(7), 476–485. https://doi.org/10.1016/j.tcb.2016.03.009

Edwards, A., Cameron, K. A., Cook, J. M., Debbonaire, A. R., Furness, E., Hay, M. C., & Rassner, S. M. E. (2020). Microbial genomics amidst the arctic crisis. *Microbial Genomics*, *6*(5), 1–20. https://doi.org/10.1099/mgen.0.000375

Falkowski, P. G., & Oliver, M. J. (2007). Mix and match: How climate selects phytoplankton. *Nature Reviews Microbiology*, *5*(10), 813–819. https://doi.org/10.1038/nrmicro1751

Farmer, J. R., Sigman, D. M., Granger, J., Underwood, O. M., Fripiat, F., Cronin, T. M., … Haug, G. H. (2021). Arctic Ocean stratification set by sea level and freshwater inputs since the last ice age. *Nature Geoscience*, *14*(9), 684–689. https://doi.org/10.1038/s41561-021-00789-y

Frioux, C., Singh, D., Korcsmaros, T., & Hildebrand, F. (2020). From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal*, *18*, 1722–1734. https://doi.org/10.1016/j.csbj.2020.06.028

Fukuchi, S., Yoshimune, K., Wakayama, M., Moriguchi, M., & Nishikawa, K. (2003). Unique amino acid composition of proteins in halophilic bacteria. *Journal of Molecular Biology*, *327*(2), 347–357. https://doi.org/10.1016/S0022-2836(03)00150-5

Galand, P. E., Gutiérrez-Provecho, C., Massana, R., Gasol, J. M., & Casamayor, E. O. (2010). Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). *Limnology and Oceanography*, *55*(5), 2117–2125. https://doi.org/10.4319/lo.2010.55.5.2117

Garcia-Soto, C., Cheng, L., Caesar, L., Schmidtko, S., Jewett, E. B., Cheripka, A., … Abraham, J. P. (2021). An Overview of Ocean Climate Change Indicators: Sea Surface Temperature, Ocean Heat Content, Ocean pH, Dissolved Oxygen Concentration, Arctic Sea Ice Extent, Thickness and Volume, Sea Level and Strength of the AMOC (Atlantic Meridional Overturning Circula. *Frontiers in Marine Science*, *8*(September). https://doi.org/10.3389/fmars.2021.642372

Gentilhomme, A. (2020). *Genomic signatures of optimal growth temperature in the family Colwelliaceae*.
Gong, D., & Pickart, R. S. (2016). Early summer water mass transformation in the eastern Chukchi Sea. *Deep-Sea Research Part II*, *130*, 43–55. https://doi.org/10.1016/j.dsr2.2016.04.015
Goordial, J., Raymond-Bouchard, I., Zolotarov, Y., de Bethencourt, L., Ronholm, J., Shapiro, N., … Whyte, L. (2016). Cold adaptive traits revealed by comparative genomic analysis of the eurypsychrophile Rhodococcus sp. JG3 isolated from high elevation McMurdo Dry Valley permafrost, Antarctica. *FEMS Microbiology Ecology*, *92*(2), 1–11. https://doi.org/10.1093/femsec/fiv154
Gruber, N., & Sarmiento, J. L. (1997). Global patterns of marine nitrogen fixation and denitrofication. *Global Biogeochemical Cycles*, *11*(2), 235–266.
Guy, L., & Ettema, T. J. G. (2011). The archaeal "TACK" superphylum and the origin of eukaryotes. *Trends in Microbiology*, *19*(12), 580–587. https://doi.org/10.1016/j.tim.2011.09.002
Haro-Moreno, J. M., Rodriguez-Valera, F., López-García, P., Moreira, D., & Martin-Cuadrado, A. B. (2017). New insights into marine group III Euryarchaeota, from dark to light. *ISME Journal*, *11*(5), 1102–1117. https://doi.org/10.1038/ismej.2016.188
Hentschel, U., Piel, J., Degnan, S. M., & Taylor, M. W. (2012). Genomic insights into the marine sponge microbiome. *Nature Reviews Microbiology*, *10*(9), 641–654. https://doi.org/10.1038/nrmicro2839
Herfort, L., Schouten, S., Abbas, B., Veldhuis, M. J. W., Coolen, M. J. L., Wuchter, C., … Sinninghe Damsté, J. S. (2007). Variations in spatial and temporal distribution of Archaea in the North Sea in relation to environmental variables. *FEMS Microbiology Ecology*, *62*(3), 242–257. https://doi.org/10.1111/j.1574-6941.2007.00397.x
Hoegh-Guldberg, O., Poloczanska, E. S., Skirving, W., & Dove, S. (2017). Coral reef ecosystems under climate change and ocean acidification. *Frontiers in Marine Science*, *4*(MAY). https://doi.org/10.3389/fmars.2017.00158
Horak, R. E. A., Qin, W., Schauer, A. J., Armbrust, E. V., Ingalls, A. E., Moffett, J. W., … Devol, A. H. (2013). Ammonia oxidation kinetics and temperature sensitivity of a natural marine community dominated by Archaea. *ISME Journal*, *7*(10), 2023–2033. https://doi.org/10.1038/ismej.2013.75
Horn, H., Slaby, B. M., Jahn, M. T., Bayer, K., Moitinho-Silva, L., Förster, F., … Hentschel, U. (2016). An Enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. *Frontiers in Microbiology*, *7*(NOV), 1–15. https://doi.org/10.3389/fmicb.2016.01751
Hyatt, D., Chen, GL., LoCascio, P.F. et al. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* 11, 119. https://doi.org/10.1186/1471-2105-11-119
Ionescu, D., Penno, S., Haimovich, M., Rihtman, B., Goodwin, A., Schwartz, D., … Oren, A. (2009). Archaea in the Gulf of Aqaba. *FEMS Microbiology Ecology*, *69*(3), 425–438. https://doi.org/10.1111/j.1574-6941.2009.00721.x
Jensen, L. T., Wyatt, N. J., Twining, B. S., Rauschenberg, S., Landing, W. M., Sherrell, R. M., & Fitzsimmons, J. N. (2019). Biogeochemical Cycling of Dissolved Zinc in the Western Arctic (Arctic GEOTRACES GN01). *Global Biogeochemical Cycles*, *33*(3), 343–369. https://doi.org/10.1029/2018GB005975
Johannessen, O. M., Bengtsson, L., Miles, M. W., Kuzmina, S. I., Semenov, V. A., Alekseev, G. V., … Cattle, H. P. (2004). Arctic climate change: Observed and modelled temperature and sea-ice variability. *Tellus, Series A: Dynamic Meteorology and Oceanography*, *56A*, 328–341. https://doi.org/10.1111/j.1600-0870.2004.00092.x
Jones, E. P., & Anderson, L. G. (1986). On the origin of the chemical properties of the Arctic Ocean halocline. *Journal of Geophysical Research*, *91*(C9), 10759–10767. https://doi.org/10.1029/jc091ic09p10759
Junge, K., Krembs, C., Deming, J., Stierle, A., & Eicken, H. (2001). A microscopic approach to investigate bacteria under in situ conditions in sea-ice samples. *Annals of Glaciology*, *33*, 304–310. https://doi.org/10.3189/172756401781818275
Kadko, D., & Landing, W. (2015). *U.S. Arctic GEOTRACES: USCGC HEALY (HLY1502)*. Retrieved from https://www.geotraces.org/gn01/

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *2019*(7), 1–13. https://doi.org/10.7717/peerj.7359

Karner, M. B., Delong, E. F., & Karl, D. M. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature*, *409*(6819), 507–510. https://doi.org/10.1038/35054051

Kellner, S., Spang, A., Offre, P., Szöllosi, G. J., Petitjean, C., & Williams, T. A. (2018). Genome size evolution in the Archaea. *Emerging Topics in Life Sciences*, *2*(4), 595–605. https://doi.org/10.1042/ETLS20180021

Kenny, N. J., Francis, W. R., Rivera-Vicéns, R. E., Juravel, K., de Mendoza, A., Díez-Vives, C., … Leys, S. P. (2020). Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge Ephydatia muelleri. *Nature Communications*, *11*(1), 1–11. https://doi.org/10.1038/s41467-020-17397-w

Klinger, C. M., Spang, A., Dacks, J. B., & Ettema, T. J. G. (2016). Tracing the Archaeal Origins of Eukaryotic Membrane-Trafficking System Building Blocks. *Molecular Biology and Evolution*, *33*(6), 1528–1541. https://doi.org/10.1093/molbev/msw034

Kwiatkowski, L., Torres, O., Bopp, L., Aumont, O., Chamberlain, M., R. Christian, J., … Ziehn, T. (2020). Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections. *Biogeosciences*, *17*(13), 3439–3470. https://doi.org/10.5194/bg-17-3439-2020

Letunic, I., & Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics*, *23*(1), 127–128. https://doi.org/10.1093/bioinformatics/btl529

Li, G., Rabe, K. S., Nielsen, J., & Engqvist, M. K. M. (2019). Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima. *ACS Synthetic Biology*, *8*(6), 1411–1420. https://doi.org/10.1021/acssynbio.9b00099

Liu, H., Zhang, C. L., Yang, C., Chen, S., Cao, Z., Zhang, Z., & Tian, J. (2017). Marine Group II dominates planktonic archaea in water column of the Northeastern South China Sea. *Frontiers in Microbiology*, *8*(JUN), 1–11. https://doi.org/10.3389/fmicb.2017.01098

Liu, Q., Tolar, B. B., Ross, M. J., Cheek, J. B., Sweeney, C. M., Wallsgrove, N. J., … Hollibaugh, J. T. (2018). Light and temperature control the seasonal distribution of thaumarchaeota in the South Atlantic bight /704/47 /631/158/855 /38/77 article. *ISME Journal*, *12*(6), 1473–1485. https://doi.org/10.1038/s41396-018-0066-4

Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein and Cell*, *12*(5), 315–330. https://doi.org/10.1007/s13238-020-00724-8

Matsen F. et al. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinform.*, 11, 538.

Martens-Habbena, W., Berube, P. M., Urakawa, H., De La Torre, J. R., & Stahl, D. A. (2009). Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature*, *461*(7266), 976–979. https://doi.org/10.1038/nature08465

Merbt, S. N., Stahl, D. A., Casamayor, E. O., Martí, E., Nicol, G. W., & Prosser, J. I. (2012). Differential photoinhibition of bacterial and archaeal ammonia oxidation. *FEMS Microbiology Letters*, *327*(1), 41–46. https://doi.org/10.1111/j.1574-6968.2011.02457.x

Meziti, A., Rodriguez-R, L. M., Hatt, J. K., Peña-Gonzalez, A., Levy, K., & Konstantinidis, K. T. (2021). The Reliability of Metagenome-Assembled Genomes (MAGs) in Representing Natural Populations: Insights from Comparing MAGs against Isolate Genomes Derived from the Same Fecal Sample. *Applied and Environmental Microbiology*, *87*(6), 1–15. https://doi.org/10.1128/AEM.02593-20

Morán, X. A. G., López-Urrutia, Á., Calvo-Díaz, A., & LI, W. K. W. (2010). Increasing importance of small phytoplankton in a warmer ocean. *Global Change Biology*, *16*(3), 1137–1144. https://doi.org/10.1111/j.1365-2486.2009.01960.x

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., … Eloe-Fadrosh, E. A. (2021).

A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), 499–509. https://doi.org/10.1038/s41587-020-0718-6

Nichols, D. S., Miller, M. R., Davies, N. W., Goodchild, A., Raftery, M., & Cavicchioli, R. (2004). Cold adaptation in the Antarctic archaeon Methanococcoides burtonii involves membrane lipid unsaturation. *Journal of Bacteriology*, *186*(24), 8508–8515. https://doi.org/10.1128/JB.186.24.8508-8515.2004

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). MetaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

O'Connor, M. I., Piehler, M. F., Leech, D. M., Anton, A., & Bruno, J. F. (2009). Warming and resource availability shift food web structure and metabolism. *PLoS Biology*, *7*(8), 3–8. https://doi.org/10.1371/journal.pbio.1000178

Offre, P., Spang, A., & Schleper, C. (2013). Archaea in biogeochemical cycles. *Annual Review of Microbiology*, *67*, 437–457. https://doi.org/10.1146/annurev-micro-092412-155614

Oksanen, K., Simpson, G. L., Guillaume Blanchet, L., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B. Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G. Chirico, M., De Caceres, M., Durand, S., Evangelista, H. B. A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M. O., Lahti, L., McGlinn, D., Ouellette, M. H., Cunha, E. R., Smith, T., Stier, A., Ter Braak C. T. F., and Weedon, F. (2022). vegan: Community Ecology Package. R package version 2.6-2. https://CRAN.R-project.org/package=vegan

Oziel, L., Schourup-Kristensen, V., Wekerle, C., & Hauck, J. (2022). The Pan-Arctic Continental Slope as an Intensifying Conveyer Belt for Nutrients in the Central Arctic Ocean (1985–2015). *Global Biogeochemical Cycles*, *36*(6), 1–19. https://doi.org/10.1029/2021gb007268

Palatinszky, M., Herbold, C., Jehmlich, N., Pogoda, M., Han, P., Von Bergen, M., … Wagner, M. (2015). Cyanate as an energy source for nitrifiers. *Nature*, *524*(7563), 105–108. https://doi.org/10.1038/nature14856

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 1043–1055. https://doi.org/10.1101/gr.186072.114.Freely

Pereira, O., Hochart, C., Auguet, J. C., Debroas, D., & Galand, P. E. (2019). Genomic ecology of Marine Group II, the most common marine planktonic Archaea across the surface ocean. *MicrobiologyOpen*, *8*(9), 1–15. https://doi.org/10.1002/mbo3.852

Peterson, B. J., McClelland, J., Curry, R., Holmes, R. M., Walsh, J. E., & Aagaard, K. (2006). Trajectory shifts in the arctic and subarctic freshwater cycle. *Science*, *313*(5790), 1061–1066. https://doi.org/10.1126/science.1122593

Pisareva, M. N., Pickart R. S., Spall M. N., Nobre, C., Torres, D.J., Moore, G.W.K., Whitledge, T. E. (2015). Flow of Pacific water in the western Chukchi Sea: Results from the 2009 RUSALCA expedition. *Deep Sea Research*. 105, 53-73. https://doi.org/10.1016/j.dsr.2015.08.011.

Pitcher, A., Wuchter, C., Siedenberg, K., Schouten, S., & Damsté, J. S. S. (2011). Crenarchaeol tracks winter blooms of ammonia-oxidizing Thaumarchaeota in the coastal North Sea. *Limnology and Oceanography*, *56*(6), 2308–2318. https://doi.org/10.4319/lo.2011.56.6.2308

Qin, W., Amin, S. A., Martens-Habbena, W., Walker, C. B., Urakawa, H., Devol, A. H., … Stahl, D. A. (2014). Marine ammonia-oxidizing archaeal isolates display obligate mixotrophy and wide ecotypic variation. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(34), 12504–12509. https://doi.org/10.1073/pnas.1324115111

Rantanen, M., Karpechko, A.Y., Lipponen, A. et al. (2022). The Arctic has warmed nearly four times faster than the globe since 1979. *Commun Earth Environ.* 3, 168. https://doi.org/10.1038/s43247-022-00498-3

Raymond-Bouchard, I., Goordial, J., Zolotarov, Y., Ronholm, J., Stromvik, M., Bakermans, C., & Whyte, L. G. (2018). Conserved genomic and amino acid traits of cold adaptation in subzero-growing Arctic permafrost bacteria. *FEMS Microbiology Ecology*, *94*(4), 1–13.

https://doi.org/10.1093/femsec/fiy023
Redfield, A. C., Netchum B. H., and Richards F. A. (1963). The influence of organisms on the composition of sea-water. *Hill, M.N. (Ed.)* The composition of seawater: Comparative and descriptive oceanography. The sea: ideas and observations on progress in the study of the seas, 2: pp. 26-77
Reji, L., & Francis, C. A. (2020). Metagenome-assembled genomes reveal unique metabolic adaptations of a basal marine Thaumarchaeota lineage. *ISME Journal*, *14*(8), 2105–2115. https://doi.org/10.1038/s41396-020-0675-6
Reji, L., Tolar, B. B., Smith, J. M., Chavez, F. P., & Francis, C. A. (2019). Differential co-occurrence relationships shaping ecotype diversification within Thaumarchaeota populations in the coastal ocean water column. *ISME Journal*, *13*(5), 1144–1158. https://doi.org/10.1038/s41396-018-0311-x
Rinke, C., Chuvochina, M., Mussig, A., Chaumeil, P.-A., Davin, A., Waite, D., … Hugenholtz, P. (2020). *Resolving widespread incomplete and uneven archaeal classifications based on a rank-normalized genome-based taxonomy*. https://doi.org/10.1101/2020.03.01.972265
Rinke, C., Rubino, F., Messer, L. F., Youssef, N., Parks, D. H., Chuvochina, M., … Hugenholtz, P. (2019). A phylogenomic and ecological analysis of the globally abundant Marine Group II archaea (Ca. Poseidoniales ord. nov.). *ISME Journal*, *13*(3), 663–675. https://doi.org/10.1038/s41396-018-0282-y
Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352
Roy Chowdhury, A., & Dutta, C. (2012). A pursuit of lineage-specific and niche-specific proteome features in the world of archaea. *BMC Genomics*, *13*(1). https://doi.org/10.1186/1471-2164-13-236
Rudels, B. (2015). Arctic Ocean circulation, processes and water masses: A description of observations and ideas with focus on the period prior to the International Polar Year 2007 – 2009. *Progress in Oceanography*, *132*, 22–67. https://doi.org/10.1016/j.pocean.2013.11.006
Santoro, A. E., Buchwald, C., McIlvin, M. R., & Casciotti, K. L. (2011). Isotopic Signatute of N20 Produced by Marine Ammonia-Oxidizing Archaea. *Science*, *333*(6047), 1282–1285. https://doi.org/10.1126/science.1208239
Saunders, N. F. W., Thomas, T., Curmi, P. M. G., Mattick, J. S., Kuczek, E., Slade, R., … Cavicchioli, R. (2003). Mechanisms of thermal adaptatation revealed from genomes of the anatarctic Archaea Methanogenium frigidum and Methanacoccoides burtonii. *Genome Research*, *13*(7), 1580–1588. https://doi.org/10.1101/gr.1180903
Schmidtko, S., Stramma, L., & Visbeck, M. (2017). Decline in global oceanic oxygen content during the past five decades. *Nature*, *542*(7641), 335–339. https://doi.org/10.1038/nature21399
Siddiqui, K. S., Cavicchioli, R., & Thomas, T. (2002). Thermodynamic activation properties of elongation factor 2 (EF-2) proteins from psychrotolerant and thermophilic Archaea. *Extremophiles*, *6*(2), 143–150. https://doi.org/10.1007/s007920100237
Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843. https://doi.org/10.1038/s41564-018-0171-1
Slaby, B. M., Hackl, T., Horn, H., Bayer, K., & Hentschel, U. (2017). Metagenomic binning of a marine sponge microbiome reveals unity in defense but metabolic specialization. *ISME Journal*, *11*(11), 2465–2478. https://doi.org/10.1038/ismej.2017.101
Spang, A., Saw, J. H., Jørgensen, S. L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A. E., … Zaremba-, K. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*, *521*(7551), 173–179. https://doi.org/10.1038/nature14447.Complex
Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinform.*, 22(21), 2688–2690, https://doi.org/10.1093/bioinformatics/btl446
Susko E., Roger A. J. (2021) Long Branch Attraction Biases in Phylogenetics. *Syst. Bio.*, 70(4), 838–843,

https://doi.org/10.1093/sysbio/syab001
Teeling, H., & Glöckner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis-A bioinformatic perspective. *Briefings in Bioinformatics*, *13*(6), 728–742. https://doi.org/10.1093/bib/bbs039
Thomas, T., & Cavicchioli, R. (2000). Effect of temperature on stability and activity of elongation factor 2 proteins from Antarctic and thermophilic methanogens. *Journal of Bacteriology*, *182*(5), 1328–1332. https://doi.org/10.1128/JB.182.5.1328-1332.2000
Wang, L., Yang, J., Xu, Y., Piao, X., & Lv, J. (2019). Domain-based Comparative Analysis of Bacterial Proteomes: Uniqueness, Interactions, and the Dark Matter. *Current Genomics*, *20*(2), 115–123. https://doi.org/10.2174/1389202920666190320134438
Wang, Z., Huang, P., You, R., Sun, F., & Zhu, S. (2023). MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biology*, *24*(1), 1–18. https://doi.org/10.1186/s13059-022-02832-6
Ward, B. B. (2005). Temporal variability in nitrification rates and related biogeochemical factors in Monterey Bay, California, USA. *Marine Ecology Progress Series*, *292*, 97–109. https://doi.org/10.3354/meps292097
Webster, N. S., & Thomas, T. (2016). The sponge hologenome. *MBio*, *7*(2). https://doi.org/10.1128/mBio.00135-16
Whitmore, L. M., Morton, P. L., Twining, B. S., & Shiller, A. M. (2019). Vanadium cycling in the Western Arctic Ocean is in fl uenced by shelf-basin connectivity. *Marine Chemistry*, *216*(October), 103701. https://doi.org/10.1016/j.marchem.2019.103701
Whitmore, L. M., Shiller, A. M., Horner, T. J., Xiang, Y., Auro, M. E., Bauch, D., … Thomas, H. (2022). Strong Margin Influence on the Arctic Ocean Barium Cycle Revealed by Pan-Arctic Synthesis. *Journal of Geophysical Research: Oceans*, *127*(4), 1–28. https://doi.org/10.1029/2021JC017417
Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090. https://doi.org/10.1073/pnas.74.11.5088
Woodgate, R. A., Aagaard, K., Muench, R. D., Gunn, J., Bjork, G., Rudels, B., … Schauer, U. (2001). The Arctic Ocean Boundary Current along the Eurasian slope and the adjacent Lomonosov Ridge : Water mass properties , transports and transformations from moored instruments. *Deep-Sea I*, *48*, 1757–1792.
Wu, Y. W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638
Wuchter, C., Abbas, B., Coolen, M. J. L., Herfort, L., Van Bleijswijk, J., Timmers, P., … Damsté, J. S. S. (2006). Archaeal nitrification in the ocean. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(33), 12317–12322. https://doi.org/10.1073/pnas.0600756103
Yager, P. L., & Deming, J. W. (1999). Pelagic microbial activity in an arctic polynya: Testing for temperature and substrate interactions using a kinetic approach. *Limnology and Oceanography*, *44*(8), 1882–1893. https://doi.org/10.4319/lo.1999.44.8.1882
Yilmaz, P., Yarza, P., Rapp, J. Z., & Glöckner, F. O. (2016). Expanding the world of marine bacterial and archaeal clades. *Frontiers in Microbiology*, *6*(JAN), 1–29. https://doi.org/10.3389/fmicb.2015.01524