Metagenomic Analysis of Microbial and Viral Communities in Low Oxygen Marine Environments

Gutierrez, Frank

UNIVERSITY OF CALIFORNIA, MERCED

Metagenomic Analysis of Microbial and Viral Communities in Low Oxygen Marine Environments

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Quantitative and Systems Biology

by

Frank Gutierrez

Committee in Charge:
Professor Carolin Frank, Chair
Professor J. Michael Beman, Graduate Advisor
Professor Rudy M. Ortiz
Simon Roux

2024

The Dissertation of Frank Gutierrez is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

J. Michael Beman, Ph.D. (Graduate Advisor)

Rudy M. Ortiz, Ph.D.

Simon Roux, Ph.D.

Carolin Frank, Ph.D. (Chair)

University of California, Merced

2024

Dedication

I dedicate this dissertation to my mother, father, and my sisters.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my advisor J. Michael Beman who has mentored me for the last five and half years at UC Merced. With his guidance I have grown from a novice undergraduate to a much more well-rounded scientist and researcher. This advancement in creative, analytical, and technical experience that I've gained from my graduate career will stay with me now till my future career endeavors. Mike was very helpful in identifying my research weaknesses and elevating my strengths. I very much appreciate Mike always willing to lend a hand whenever I veered off-track, keeping me on course towards stellar research. Thanks Mike!

I would also like to thank my lab mates, both past and present, who made me feel welcomed at UC Merced and the graduate student community. It amazes me how our first meeting was as formal colleagues but has evolved into long lasting friendships. There were moments both inside and outside the lab that I will never forget and helped me grow as a person.

I would like to thank my committee members for their assistance during the tougher stages of my PhD career. From their expertise in their field, I gained additional mentorship and valuable feedback on my dissertation projects. I would also like to specifically thank Simon Roux, as he was my mentor during my summer internship at the Joint Genome Institute (JGI). Lastly, I would like to thank the QSB program and Graduate Division at UC Merced, giving me the opportunity to pursue a PhD as a first-generation college student. I appreciate the education that I received at UC Merced, and I am glad to have been taught by many great faculty members who contributed to my academic success.

This work was supported by NSF CAREER Grant OCE-1555375 to J.M.B., and UCMEXUS-CONACyT Collaborative Grants Program joint awards to J.M.B. and J.Q.G.M.

Curriculum Vitae

Education

University of California Merced (Merced, CA) Expected Date (Dec 2024)
Doctor of Philosophy, PhD
Quantitative and Systems Biology

University of California Merced (Merced, CA) May 2023
Master of Science, M.S.
Quantitative and Systems Biology

California State University Los Angeles (Los Angeles, CA) May 2019
Bachelor of Science, B.S.
Biology

Research Experience

University of California Merced (UCM); Merced, CA Aug 2019 - present
Principal Investigator: Dr. Michael Beman

- Microbial and Viral Communities in the Oxygen Minimum Zone

  Conducting metagenomic analysis on microbial and viral communities in the Eastern Tropical North Pacific (ENTP) Oxygen Minimum Zone (OMZ). Exploring microbial functionality and viral Auxiliary Metabolic Genes (AMGs) in their role towards biogeochemical cycling under hypoxic and anoxic conditions along with varying environmental gradients throughout the water column.

Joint Genome Institute (JGI); Berkeley, CA June 2020 - Aug 2020
Principal Investigator: Dr. Simon Roux

- Large-scale profiling of public Metagenomes

  Generated a script to identify the number of genomes within a metagenome using single copy marker genes based on Pfam domains. The script was then used by Integrated Microbial Genomes & Microbiomes system (IMG/M) to access metagenomic statistics of the metagenomic data.

California State University Los Angeles
(CSULA); Los Angeles, CA June 2016 – May 2019
Principal Investigator: Dr. Andres Aguilar

- Genetic variation within the Lake Baikal endemic sculpin genus *Limnocottus*

  Understanding the process of adaptive radiation and speciation in freshwater fish. Determining the genetic differences of the Limnocottus species through phylogenetic analysis from mitochondrial and nuclear DNA.

University of Colorado Boulder
(CU Boulder); Boulder, CO June 2018 - Aug 2018
Principal Investigator: Dr. Robin Dowell

- The Role of Metadata Analysis on Nascent Transcription

  Constructed a database revolving around nascent transcription data from various publications on different organisms. Nascent transcription data can be used to identify bidirectional transcription activity, which are signals for promoters and enhancers in the genome.

University of California Los Angeles (UCLA); Los Angeles, CA June 2017 - Aug 2017
Principal Investigator: Dr. Tracy Johnson

- Identifying the regulatory function of histone acetyltransferase Gcn5 in the Saccharomyces cerevisiae pre-mRNA splicing pathway

  Computational analysis on the pre-mRNA splicing pathway in the S. cerevisiae. Identified the role of Gcn5 in co-transcriptional splicing and how it is affected by the chromatin environment, specifically Histone H3 N terminal tail K9-K14.

Publications

Aguilar, A., Truong, B. R., & **Gutierrez, F**. (2018). Complete mitochondrial DNA genomes for two northeast Pacific mesopelagic fishes, the Mexican lampfish (Triphoturus mexicanus) and black-belly dragonfish (Stomias atriventer). Mitochondrial DNA Part B, 3(1), 21-23. https://doi.org/10.1080/23802359.2017.1413293

Manuscripts in Preparation

Gutierrez, F., Vargas, S., Machado-Perez, F., Wilson, J., García-Maldonado, J.Q., & Beman, J.M., Microbial community metagenomics in the Eastern Tropical North Pacific Oxygen Minimum Zone reveals functional differences along spatial and biogeochemical gradients. Manuscript Submitted to Environmental Microbiology

Presentations

Gutierrez, F.; Vargas, S.; Machado-Perez, F.; García-Maldonado, J. Q.; Beman, M.; Microbial Community Metagenomics in the Eastern Tropical North Pacific Oxygen Minimum Zone Reveals Functional Differences with Depth and Geography. Talk presented at Ocean Sciences Meeting. New Orleans, LA. Feb 2024.

Gutierrez, F.; Sandel, M; Kirilchik, S.; Aguilar, A.; *Genetic variation within the Lake Baikal endemic sculpin genus Limnocottus*. Poster presented at Society for Advancement of Chicanos/Hispanics and Native Americans in Science. San Antonio, TX. Oct 2018.

Gutierrez, F.; Allen, M.; Tripodi, I.; Dowell, R.; *The Role of Metadata Analysis on Nascent Transcription*. Poster presented at the Annual Biomedical Research Conference for Minority Students. Indianapolis, IN. Nov 2018.

Fellowships/Research Programs

JGI - UC Merced Genomics Summer Internship Program	June 2020 - Aug 2020
Minority Access to Research Careers-Undergraduate	June 2017 - May 2019

Student Training for Academic Research (MARC-U*STAR) Program	
CU Boulder Summer Multicultural Access program To Research Training (SMART) program	June 2018 - Aug 2018
The Leadership Alliance Summer Research – Early Identification Program (SR -EIP) at CU Boulder	June 2018 - Aug 2018
UCLA Bruins in Genomics (B.I.G.) Summer Research Program	June 2017 - Aug 2017
Minority Biomedical Research Support-Research Initiative for Scientific Enhancement (MBRS-RISE) Program	July 2016 - June 2017

Teaching Assistantship

University of California Merced

ESS 001: Introduction to Earth Systems Science	Spring 2023, Spring 2024
BIO 005: Biology Today	Spring 2021, Spring 2022
BIO 001L: Contemporary Biology Lab	Fall 2020
BIO 002L: Molecular Biology Lab	Spring 2020
BIO 003: Molecular Basis of Health and Disease	Fall 2019

Honors/Awards

Annual Biomedical Research Conference for Minority Students (ABRCMS) Travel Scholarship	Nov 2018
CSULA Dean’s List 2018	May 2018
CSULA Dean’s List 2017	May 2017
CSULA Dean’s List 2016	May 2016
CSULA Freshman Honors at Entrance Award	Mar 2015

Abstract

Metagenomic Analysis of Microbial and Viral Communities in Low Oxygen Marine Environments

Frank Gutierrez

Doctor of Philosophy in Quantitative and System Biology

University of California, Merced 2024

Chair: Carolin Frank
Graduate Advisor: J. Michael Beman

Oxygen Minimum Zones (OMZs) are regions of the ocean with dissolved oxygen concentrations of <20 µM that have formed due to natural processes but are expanding in consequence of climate change. Despite the extreme conditions of low oxygen, OMZs contain unique microbial communities that play key roles in global biogeochemical cycling. In addition, recent research has shown that OMZs also harbor viral communities, which have been discovered with auxiliary metabolic genes (AMGs) that augment host metabolisms and indirectly contribute to these global cycles. Although both microorganisms and viruses have been found in OMZs, it remains unclear how both groups shape and are shaped by physical and environmental gradients in the OMZ water column. In this dissertation, I have used shotgun metagenomic sequencing and bioinformatic tools to investigate both microbial and viral communities in the ocean's largest and intense OMZ, which resides in the Eastern Tropical North Pacific. In chapter 1, shotgun sequencing was used to identify microbial functional and taxonomic composition in key distinct layers of the water column across different geographical stations in the ETNP OMZ. Read-based analysis showed that gene abundances were much higher in the OMZ core and microbial communities that resided in the productive regions of the OMZ were mainly composed of heterotrophic prokaryotes. Chapter 2 focused on viral communities and AMG composition in the OMZ, along with comparative analysis on metagenomic tools for AMG detection. Main findings from this chapter are that most viruses identified as cyanophage and *Pelagibacter* phage, while most AMGs were involved in photosynthesis and purine synthesis. Lastly, Chapter 3 focused on induvial Metagenome Assembled Genomes (MAGs) of ammonia oxidizing archaea (AOA) for comparative genomic analysis. AOA were identified with many different metabolic capabilities and phylogenomic analysis showed AOA formed distinct clades, each sharing and containing unique core functional genes. Overall, the results within this dissertation demonstrate that both microbial and viral community composition and functionality correlate with key biogeochemical gradients in the ETNP OMZ, and contribute to the field of microbial and viral ecology through metagenomic analysis.

Introduction

Microbial Background

Microbial communities consist of prokaryotes such as bacteria and archaea, unicellular eukaryotes, and to some extent viruses. Microbes are present in all environments, diverse with various metabolic processes, and play a critical role in regulating biogeochemical cycling across all biomes (Thompson et al. 2017; Oren 2009; Allen et al. 2023). Due to microbial metabolic diversity, they have been identified with key enzymes that drive these global biogeochemical cycles (Falkowski et al. 2008). Our oceans are pivotal sites of biogeochemical cycling and host diverse ecosystems that contain many microbial communities. In the carbon cycle, our oceans are a natural carbon sink, and aquatic microbial communities disproportionately impact the global carbon cycle through respiration and primary production (DeVries 2022; Hurley et al. 2021; Williams 1984; Robinson & Williams 2005). Microbes in our oceans also contribute to the sulfur cycle via sulfate reduction and sulfur oxidation (van Vliet et al. 2021). Lastly, biogeochemical manipulation can also occur for the nitrogen cycle through processes such as nitrogen fixation, nitrification, and denitrification. Denitrification leads to nitrogen loss, subsequently converting nitrate and nitrite into nitrous oxide and dinitrogen gas (Kuypers et al. 2018). These are several ways that microbes can alter biogeochemistry, and microbial communities have also been shown to correlate with different environmental and ecosystem factors (Graham et al. 2016; Lozupone & Knight 2007; Dinsdale et al. 2008).

Marine Viruses

Although microorganisms drive global biogeochemical cycles and are the most abundant living organisms on the planet, they are highly outnumbered by viruses (Stern & Sorek 2011; Fuhrman 1999; Suttle 2007). This is especially true in marine environments, as viruses are far more abundant than microbial cells (Cassman et al. 2012; Suttle 2007; Wilhelm & Suttle 1999). Viruses can infect members from all domains of life and impact many biological communities. In the ocean, viruses infect organisms ranging from marine bacteria, zooplankton, phytoplankton, algae, fish, to whales (Suttle 2005). Most viruses in the ocean are bacteriophages, which are viruses that infect bacteria (Wilhelm & Suttle 1999; Clokie et al. 2011). Not only are viruses diverse by the organisms that they infect, but also diverse through method of infection and genomic makeup. Many of the bacteriophages that infect prokaryotes typically have double stranded DNA genomes, and infect through a temperate or lytic approach, while also impacting their host physiology (Berg et al. 2021; Warwick-Dugdale et al. 2019; Suttle 2007). Although viruses' main purpose is to infect their hosts and reproduce, like microorganisms they play a very important part in the global engines that run our planet.

In our oceans, viruses contribute to the microbial loop via the viral shunt, in which carbon and nutrients are transferred from microorganisms to dissolved organic matter (DOM) pools via viral infection (Breitbart et al. 2018; Suttle 2005; Weitz & Wilhelm 2012). When viruses infect prokaryotes and unicellular eukaryotes, they die due to lysis, and these unicellular organisms burst. These busted cells leave behind organic material for heterotrophic bacteria to feed on, benefiting these microbial groups. These heterotrophs (and autotrophs) are then consumed by grazers, which eventually get consumed by larger predators (Weitz & Wilhelm 2012). This is how viruses are key players in moving carbon and nutrients in the environment, while new studies now depict another way how viruses contribute to biogeochemical cycling. This alternative path is by using auxiliary metabolic genes (AMGs), which are viral genes that can control host metabolic processes and augment key pathways during infection to enhance viral replication (Breitbart et al. 2007; Mara et al. 2020). These viral AMGs have been identified in biogeochemical important pathways involved in carbon, sulfur, and nitrogen cycling (Anantharaman et al. 2014; Gazitúa et al. 2021; Sullivan et al. 2006). Although we know that microbes and viruses can alter biogeochemical cycling, we do not fully understand their interactions and implications for biogeochemical cycling under extreme conditions.

Oxygen Minimum Zones

This is particularly relevant in environments that experience low levels of dissolved oxygen (DO), as oxygen is a key element for many of these biogeochemical cycles (Falkowski & Godfrey 2008; Banerjee et al. 2019). In the hydrosphere, oxygen that is present in the water column is called dissolved oxygen (DO), which supports many aerobic aquatic organisms to survive and is essential for marine ecosystems (Hull et al. 2008; Stramma et al. 2012). Similar to the carbon and nitrogen cycles, oxygen can be recycled in our oceans as DO can be produced via microbial primary production but is then depleted by heterotrophic organisms through respiration. DO is also being lost in our oceans due to warmer temperatures, which can lead to widespread anoxia and hypoxia in aquatic environments (Paulmier & Ruiz-Pino 2009; Robinson 2019; Keeling et al. 2010). DO is a strong regulating force for microbes because oxygen is used as an electron acceptor for aerobic forms of metabolism, and current research indicates that microbial community diversity and distribution correlate with high or low levels of DO (Canfield & Kraft 2022; Beman & Carolan 2013).

There are regions of the ocean that are devoid of oxygen and mimic the early conditions of our ancient earth. These regions of our oceans with DO levels < 20 µM are known as oxygen minimum zones (OMZs), which account for about 8% of the ocean and are one of the most important low oxygen environments (Paulmier & Ruiz-Pino 2009). There are three large OMZs, which are the Eastern Tropical North Pacific (ETNP), Eastern Tropical South Pacific (ETSP), and the Arabian Sea (Gilly et al. 2013). These OMZs have been formed through warming of the earth, thermal stratification, slow subsurface circulation, low ocean ventilation, natural upwelling, eutrophication, anthropogenic activity and microbial heterotrophic activity (Long et al. 2021). Heterotrophic organisms that perform respiration and anthropogenic activity are expanding hypoxic conditions, thus providing anaerobic microbes an environment to thrive in (Wright et al. 2012). Not all OMZ are similar as they can be diverse both laterally and vertically. There are OMZs that have regions in the water column where dissolved oxygen is essentially undetectable, this is termed anoxic marine zones (AMZs). In the AMZ core, there is a substantial accumulation of nitrite generated by anaerobic activity, called the Secondary Nitrite Maxima (SNM). In contrast, the Primary Nitrite

Maxima (PNM) are found throughout the ocean at the base of the euphotic zone (Ulloa et al. 2012). The AMZ core also contains a secondary chlorophyll maximum (SCM) inhabited by low light-adapted *Prochlorococcus* (Ulloa et al. 2012, Beman et al. 2021). Overall, not all OMZs are homogeneous, as these AMZs show variation within OMZs by being fully anoxic areas in our oceans (Ulloa et al. 2012).

Since OMZs are environments devoid of oxygen, there are rarely large aerobic multicellular organisms that inhabit these regions, and they are mainly occupied by microbial life such as bacteria and archaea (Stramma et al. 2012). In OMZs, nitrogen and carbon cycling are influenced by the variation of oxygen content throughout the water column, where there is more oxygen towards the surface (above the oxycline), and less oxygen in deeper depths (below the oxycline) (Tiano et al. 2014). In the upper oxycline, aerobic processes such as primary production, respiration, ammonia oxidation, and nitrite oxidation occur (Long et al. 2021). Below the oxycline, in the OMZ/AMZ core it is dominated by anaerobic microbes that perform denitrification and anaerobic ammonia oxidation (anammox) (Ulloa et al. 2012; Ward et al. 2009). Denitrification and anammox ultimately release nitrogen gas back into the atmosphere, but during this entire process, nitrous oxide, a greenhouse gas is also produced leading to more entrapment of heat, making OMZs sites of nitrogen loss and production of greenhouse gasses (Long et al. 2021; Ward et al. 2009; Fuchsman et al. 2019; Gilly et al. 2013; Gazitúa et al. 2021). OMZs also contribute to sulfur cycling through sulfur oxidation and methane cycling through methane oxidation in deeper and anoxic depths (Long et al. 2021; Anantharaman et al. 2014; Carolan et al. 2015; Thamdrup et al. 2019). These biogeochemical cycles that occur within the OMZs are ultimately driven by microbial communities and are distributed in many layers of the water column. Although microbes are typically associated with OMZs, they're not the only organisms that inhabit these extreme conditions, as OMZs contain viruses as well (Cassman et al. 2012).

Ammonia Oxidizing Archaea Background

Nitrogen is cycled in OMZs, and key microorganisms that contribute to this process are Ammonia-oxidizing archaea (AOA). AOA populations have been detected in OMZs, along with the *amoA* gene abundance that correlated with oxygen levels (Beman et al. 2008). AOA belong to the Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota (TACK) superphylum of archaea, specifically the *Thaumarchaeota* clade (Baker et al. 2020). *Thaumarchaeota*, specifically AOA, have gone through different stages of evolution. The main contributor was oxygen, as oxygen led to loss of anerobic processes in favor of aerobic processes and the transition from terrestrial to ocean environments (Ren et al. 2019). AOA currently live in the open ocean at the surface and in moderate temperatures, but also have been found in deeper depths ranging from the epipelagic to the hadopelagic zone (Karner et al. 2001; Könneke et al. 2005; Francis et al. 2005). Genetic studies on AOA have shown that they have gone through adaptive radiation events which allowed AOA to reside in these different environments and contain different metabolic capabilities (Qin et al. 2020; Swan et al. 2014).

Utilizing Metagenomics for Microbial Analysis

Since both prokaryotes and viruses are plentiful in our environment and contribute to global biogeochemical cycling, it is important to determine which microbes and viruses are present in OMZs and what functionality they contain. The most common gene used to identify and compare microbes is the 16S ribosomal RNA (rRNA) gene as it is a universally conserved marker in all bacteria and archaea, but still contains variable regions to differentiate microbial groups (Hugerth et al. 2017; Tringe & Hugenholtz 2008; Woese 1987). Unlike microbes, viruses cannot be identified with the 16S rRNA gene, as they do not contain any universal gene that is conserved across all viruses. Instead, viruses can be identified with metagenomic sequencing (Harris & Hill 2021). With metagenomic sequences, we can utilize a variety of bioinformatic tools to determine both microbial and viral taxonomy and functionality on raw shotgun metagenomic reads or assembled contigs (Nayfach & Pollard 2016; Ruscheweyh et al. 2022; Kim et al. 2016). In addition to this, there are bioinformatic tools that can assemble these shotgun metagenomic reads into full genomes, known as metagenomic assembled genomes (MAGs) (Chivian et al. 2023; Setubal 2021). With metagenomics we can classify and determine the functionality of both microbes and viruses in the harsh conditions of the OMZ/AMZ (Dávila-Ramos et al. 2019). **In this dissertation I will focus on the biogeochemical cycling that occurs in the OMZs along with the microbes and viruses that inhabit them. My dissertation will specifically be examining microbes and viruses in the ETNP OMZ—as this OMZ is the largest, shallowest, and most intense OMZ—through metagenomic analysis (Fig. 1)**

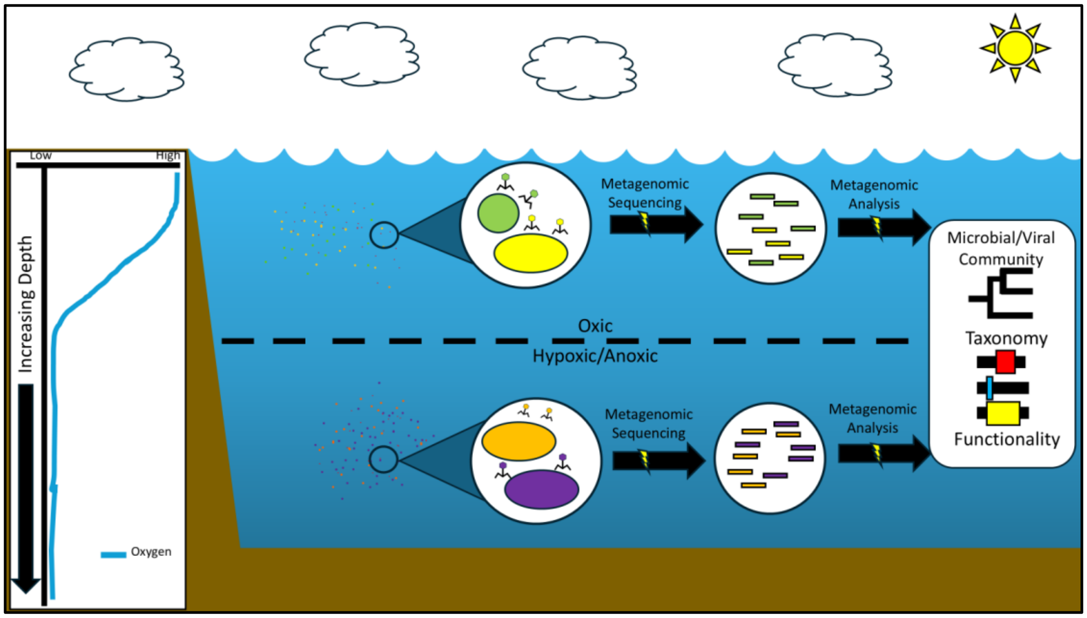


Figure 1: Generalized model of sampling, metagenomic sequencing, and metagenomic analysis of microbial and viral communities in the OMZ.

Dissertation Aims

In the first chapter of my dissertation, I will primarily be focusing on prokaryotic communities that inhabit the ETNP OMZ and determine if biogeochemical variations shape microbial communities and their functionality. I will look at key regions in the ETNP OMZ throughout the water column such as the euphotic zone, oxycline, and OMZ/AMZ core. I expect different microbial communities and specific genes to reside in their respective regions of the water column based on their oxygen gradient. I am also comparing microbial communities and their genes in OMZs that are nearshore and offshore. In the first chapter, I will be implementing metagenomic techniques to build upon previous research in the OMZ that utilized amplicon sequencing. This study is important as I am investigating microbial players in a continuing expanding OMZ, and the first chapter will explore to what extent microbes are contributing or affected by the expansion.

The second chapter will focus on viral communities that inhabit the ETNP OMZs through metagenomics. Similar to the microbial communities, this chapter will explore the role of viruses and their contribution to global biogeochemical cycling under hypoxic and anoxic marine conditions. The chapter will also give insight into viral communities that are unique to the OMZ and add on to previous studies that looked at viral groups from other major OMZ regions. This chapter will also tackle if viruses and viral AMGs are shaped by environmental factors or if they are shaped by host distribution patterns. I will also highlight metagenomic tools and techniques that help identify and classify viruses that inhabit the OMZ. Lastly, with metagenomic sequencing I aim to provide a guide on how to detect viral genes in extreme conditions and determine which AMGs are much more dominant throughout the water column. The focus is to show viruses being an integral part of the OMZ marine community and an active contributor to these global biogeochemical cycling environments devoid of oxygen.

Chapter 3 will focus specifically on AOA in the ETNP OMZ. Since AOA are one of the most abundant prokaryotic groups in our oceans, they have been found in different environmental conditions with unique metabolic capabilities. In the third chapter of my dissertation, I will conduct genomic comparison of AOA that reside in the biogeochemically active areas of the ETNP OMZ. This will be done by assembling AOA MAGs and searching metabolic pathways, then determining which genes are unique or similar across all MAGs. With the MAGs I will also conduct phylogenomic analysis, to determine phylogenetic relatedness and similarity. These MAGs will also allow us to taxonomically identify AOA at the genomic level and provide potential draft genomes for future analysis. Lastly, this chapter will aim to compare metagenomic analysis with traditional 16S rRNA analysis that have been conducted on AOA previously in the ETNP OMZ.

References:

Allen, B., Gonzalez-Cabaleiro, R., Ofiteru, I. D., Øvreås, L., Sloan, W. T., Swan, D., & Curtis, T. (2023). Diversity and metabolic energy in bacteria. FEMS Microbiology Letters, 370, fnad043.

Anantharaman, K., Duhaime, M. B., Breier, J. A., Wendt, K. A., Toner, B. M., & Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. Science, 344(6185), 757-760.

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.

Banerjee, A., Chakrabarty, M., Rakshit, N., Bhowmick, A. R., & Ray, S. (2019). Environmental factors as indicators of dissolved oxygen concentration and zooplankton abundance: Deep learning versus traditional regression approach. Ecological indicators, 100, 99-117.

Beman, J. M., & Carolan, M. T. (2013). Deoxygenation alters bacterial diversity and community composition in the ocean’s largest oxygen minimum zone. Nature Communications, 4(1), 2705.

Beman, J. M., Popp, B. N., & Francis, C. A. (2008). Molecular and biogeochemical evidence for ammonia oxidation by marine Crenarchaeota in the Gulf of California. The ISME Journal, 2(4), 429-441.

Beman, J. M., Vargas, S. M., Vazquez, S., Wilson, J. M., Yu, A., Cairo, A., & Perez-Coronel, E. (2021). Biogeochemistry and hydrography shape microbial community assembly and activity in the eastern tropical North Pacific Ocean oxygen minimum zone. Environmental Microbiology, 23(6), 2765-2781.

Berg, M., Goudeau, D., Olmsted, C., McMahon, K.D., Yitbarek, S., Thweatt, J.L., Bryant, D.A., Eloe-Fadrosh, E.A., Malmstrom, R.R., & Roux, S. (2021). Host population diversity as a driver of viral infection cycle in wild populations of green sulfur bacteria with long standing virus-host interactions. The ISME Journal, 15, 1569 - 1584.

Breitbart, M. Y. A., Thompson, L. R., Suttle, C. A., & Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. Oceanography, 20(2), 135-139.

Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. Nature microbiology, 3(7), 754-766.

Canfield, D. E., & Kraft, B. (2022). The ‘oxygen’in oxygen minimum zones. Environmental microbiology, 24(11), 5332-5344.

Carolan, M. T., Smith, J. M., & Beman, J. M. (2015). Transcriptomic evidence for microbial sulfur cycling in the eastern tropical North Pacific oxygen minimum zone. Frontiers in Microbiology, 6, 334.

Cassman, N., Prieto-Davó, A., Walsh, K., Silva, G. G., Angly, F., Akhter, S., ... & Dinsdale, E. A. (2012). Oxygen minimum zones harbour novel viral communities with low diversity. Environmental microbiology, 14(11), 3043-3065.

Chivian, D., Jungbluth, S. P., Dehal, P. S., Wood-Charlson, E. M., Canon, R. S., Allen, B. H., ... & Arkin, A. P. (2023). Metagenome-assembled genome extraction and analysis from microbiomes using KBase. Nature Protocols, 18(1), 208-238.

Clokie, M. R., Millard, A. D., Letarov, A. V., & Heaphy, S. (2011). Phages in nature. Bacteriophage, 1(1), 31-45.

Dávila-Ramos, S., Castelán-Sánchez, H. G., Martínez-Ávila, L., Sánchez-Carbente, M. D. R., Peralta, R., Hernández-Mendoza, A., ... & Batista-García, R. A. (2019). A review on viral metagenomics in extreme environments. Frontiers in microbiology, 10, 2403.

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., ... & Rohwer, F. (2008). Functional metagenomic profiling of nine biomes. Nature, 452(7187), 629-632.

DeVries, T. (2022). The ocean carbon cycle. Annual Review of Environment and Resources, 47(1), 317-341.

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive Earth's biogeochemical cycles. science, 320(5879), 1034-1039.

Falkowski, P. G., & Godfrey, L. V. (2008). Electrons, life and the evolution of Earth's oxygen cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1504), 2705-2716.

Francis, C. A., Roberts, K. J., Beman, J. M., Santoro, A. E., & Oakley, B. B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proceedings of the National Academy of Sciences, 102(41), 14683-14688.

Fuchsman, C. A., Palevsky, H. I., Widner, B., Duffy, M., Carlson, M. C., Neibauer, J. A., ... & Rocap, G. (2019). Cyanobacteria and cyanophage contributions to carbon and nitrogen cycling in an oligotrophic oxygen-deficient zone. The ISME Journal, 13(11), 2714-2726.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. Nature, 399(6736), 541-548.

Gazitúa, M. C., Vik, D. R., Roux, S., Gregory, A. C., Bolduc, B., Widner, B., ... & Sullivan, M. B. (2021). Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. The ISME Journal, 15(4), 981-998.

Gilly, W. F., Beman, J. M., Litvin, S. Y., & Robison, B. H. (2013). Oceanographic and biological effects of shoaling of the oxygen minimum zone. Annual review of marine science, 5, 393-420.

Graham, E. B., Knelman, J. E., Schindlbacher, A., Siciliano, S., Breulmann, M., Yannarell, A., ... & Nemergut, D. R. (2016). Microbes as engines of ecosystem function: when does community structure enhance predictions of ecosystem processes?. Frontiers in microbiology, 7, 214.

Harris, H. M., & Hill, C. (2021). A place for viruses on the tree of life. Frontiers in Microbiology, 11, 604048.

Hugerth, L. W., & Andersson, A. F. (2017). Analysing microbial community composition through amplicon sequencing: from sampling to hypothesis testing. Frontiers in microbiology, 8, 1561.

Hull, V., Parrella, L., & Falcucci, M. (2008). Modelling dissolved oxygen dynamics in coastal lagoons. Ecological Modelling, 211(3-4), 468-480.

Hurley, S. J., Wing, B. A., Jasper, C. E., Hill, N. C., & Cameron, J. C. (2021). Carbon isotope evidence for the global physiology of Proterozoic cyanobacteria. Science Advances, 7(2), eabc8998.

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.

Keeling, R. F., Körtzinger, A., & Gruber, N. (2010). Ocean deoxygenation in a warming world. Annual review of marine science, 2(1), 199-229.

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., & Zhan, X. (2016). FMAP: functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. BMC bioinformatics, 17, 1-8.

Könneke, M., Bernhard, A. E., de La Torre, J. R., Walker, C. B., Waterbury, J. B., & Stahl, D. A. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature, 437(7058), 543-546.

Kuypers, M. M., Marchant, H. K., & Kartal, B. (2018). The microbial nitrogen-cycling network. Nature Reviews Microbiology, 16(5), 263-276.

Long, A. M., Jurgensen, S. K., Petchel, A. R., Savoie, E. R., & Brum, J. R. (2021). Microbial ecology of oxygen minimum zones amidst ocean deoxygenation. Frontiers in Microbiology, 12, 748961.

Lozupone, C. A., & Knight, R. (2007). Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences, 104(27), 11436-11440.

Mara, P., Vik, D., Pachiadaki, M. G., Suter, E. A., Poulos, B., Taylor, G. T., ... & Edgcomb, V. P. (2020). Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline. *The ISME Journal*, *14*(12), 3079-3092.

Nayfach, S., & Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. Cell, 166(5), 1103-1116.

Oren, A. (2009). Metabolic diversity in prokaryotes and eukaryotes. Biological Science Fundamentals and Systematics, 2, 40-76.

Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. Progress in Oceanography, 80(3-4), 113-128.

Qin, W., Zheng, Y., Zhao, F., Wang, Y., Urakawa, H., Martens-Habbena, W., ... & Ingalls, A. E. (2020). Alternative strategies of nutrient acquisition and energy conservation map to the biogeography of marine ammonia-oxidizing archaea. The ISME Journal, 14(10), 2595-2609.

Ren, M., Feng, X., Huang, Y., Wang, H., Hu, Z., Clingenpeel, S., ... & Luo, H. (2019). Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. The ISME journal, 13(9), 2150-2161.

Robinson, C. (2019). Microbial respiration, the engine of ocean deoxygenation. Frontiers in Marine Science, 5, 533.

Robinson, C., & Williams, P. L. B. (2005). Respiration and its measurement in surface marine waters. Respiration in aquatic ecosystems, 2005, 147-180.

Ruscheweyh, H. J., Milanese, A., Paoli, L., Karcher, N., Clayssen, Q., Keller, M. I., ... & Sunagawa, S. (2022). Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. Microbiome, 10(1), 212

Setubal, J. C. (2021). Metagenome-assembled genomes: concepts, analogies, and challenges. Biophysical reviews, 13(6), 905-909.

Stern, A., & Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. Bioessays, 33(1), 43-51

Stramma, L., Prince, E. D., Schmidtko, S., Luo, J., Hoolihan, J. P., Visbeck, M., ... & Körtzinger, A. (2012). Expansion of oxygen minimum zones may reduce available habitat for tropical pelagic fishes. *Nature Climate Change*, *2*(1), 33-37.

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., & Chisholm, S. W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS biology, 4(8), e234.

Suttle, C. A. (2005). Viruses in the sea. Nature, 437(7057), 356-361.

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. Nature reviews microbiology, 5(10), 801-812.

Swan, B. K., Chaffin, M. D., Martinez-Garcia, M., Morrison, H. G., Field, E. K., Poulton, N. J., ... & Stepanauskas, R. (2014). Genomic and metabolic diversity of Marine Group I Thaumarchaeota in the mesopelagic of two subtropical gyres. PloS one, 9(4), e95380.

Thamdrup, B., Steinsdóttir, H. G., Bertagnolli, A. D., Padilla, C. C., Patin, N. V., Garcia-Robledo, E., ... & Stewart, F. J. (2019). Anaerobic methane oxidation is an important sink for methane in the ocean's largest oxygen minimum zone. Limnology and Oceanography, 64(6), 2569-2585.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., ... & Knight, R. (2017). A communal catalogue reveals Earth’s multiscale microbial diversity. Nature, 551(7681), 457-463.

Tiano, L., Garcia-Robledo, E., Dalsgaard, T., Devol, A. H., Ward, B. B., Ulloa, O., ... & Revsbech, N. P. (2014). Oxygen distribution and aerobic respiration in the north and south eastern tropical Pacific oxygen minimum zones. Deep Sea Research Part I: Oceanographic Research Papers, 94, 173-183.

Tringe, S. G., & Hugenholtz, P. (2008). A renaissance for the pioneering 16S rRNA gene. Current opinion in microbiology, 11(5), 442-446.

van Vliet, D. M., von Meijenfeldt, F. B., Dutilh, B. E., Villanueva, L., Sinninghe Damsté, J. S., Stams, A. J., & Sánchez-Andrea, I. (2021). The bacterial sulfur cycle in expanding dysoxic and euxinic marine waters. Environmental Microbiology, 23(6), 2834-2857.

Ward, B. B., Devol, A. H., Rich, J. J., Chang, B. X., Bulow, S. E., Naik, H., Pratihary, A., & Jayakumar, A. (2009). Denitrification as the dominant nitrogen loss process in the Arabian Sea. Nature, 461(7260), 78-81.

Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J., & Temperton, B. (2019). Host-hijacking and planktonic piracy: how phages command the microbial high seas. Virology journal, 16, 1-13.

Weitz, J. S., & Wilhelm, S. W. (2012). Ocean viruses and their effects on microbial communities and biogeochemical cycles. F1000 biology reports, 4.

Wilhelm, S. W., & Suttle, C. A. (1999). Viruses and nutrient cycles in the sea: viruses play critical roles in the structure and function of aquatic food webs. Bioscience, 49(10), 781-788.

Williams, P. L. (1984). A review of measurements of respiration rates of marine plankton populations. Heterotrophic activity in the sea, 357-389.

Woese, C. R. (1987). Bacterial evolution. Microbiological reviews, 51(2), 221-271.

Wright, J. J., Konwar, K. M., & Hallam, S. J. (2012). Microbial ecology of expanding oxygen minimum zones. Nature Reviews Microbiology, 10(6), 381-394.

Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M., & Stewart, F. J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences, 109(40), 15996-16003.

Chapter 1: Microbial community metagenomics in the eastern tropical North Pacific oxygen minimum zone reveals predictable functional differences along biogeochemical gradients

Manuscript in Preparation

Gutierrez, F.1*; Vargas, S.1; Machado-Perez, F.1; Wilson, J.1; García-Maldonado, J.Q.2; Beman, J.M.1*

1Life and Environmental Science and Sierra Nevada Research Institute, University of California Merced, California, United States of America
2Departamento de Recursos del Mar, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Unidad Mérida, Yucatán, México

KEYWORDS: Metagenomics; Oxygen Minimum Zone; Marine Microbiology

1.1 Abstract

Oxygen Minimum Zones (OMZs) are pivotal regions of the ocean defined by low dissolved oxygen concentrations. However, biogeochemical variations occurring within OMZs—both laterally and with depth—may select for fundamentally different microbial metabolisms important for ocean biogeochemistry. We used metagenome sequencing to investigate these potential differences along biogeochemical gradients with depth and between stations in the eastern tropical North Pacific OMZ. Read-based analysis identified predictable variations in 5,389 functional genes, with high similarity among metagenomes recovered from the secondary chlorophyll maximum and secondary nitrite maximum at multiple stations. 690 genes showed statistically significant differences between different biogeochemical features in the water column, and included key functional genes with relative abundances >2,000 reads per kilobase per million (RPKM) involved in photosynthesis, carbon fixation, anaerobic nitrogen cycling, and sulfur cycling. We also found distinct functional and taxonomic composition on the edge of the OMZ at our most productive station. Metagenome assembled genomes (MAGs) from this sample included multiple *Flavobacteriaceae* and *Rhodobacteraceae* MAGs, with annotated functions contributing to metabolism of carbohydrates and amino acids, as well as aerobic anoxygenic photosynthesis (in *Rhodobacteraceae*). Our results provide new insight into microbial contributions to multiple biogeochemical processes—and their variations—within and across the ocean's largest OMZ.

1.2 Introduction

Oceanic oxygen minimum zones (OMZs) play a central role in ocean biogeochemistry as sites of intense microbially-mediated nitrogen (N), carbon (C), oxygen, and sulfur (S) cycling (Lam and Kuypers 2011; Gilly et al. 2013; Bertagnolli and Stewart 2018; Callbeck et al. 2021). These biogeochemically important regions of the ocean experience low dissolved oxygen (DO) concentrations due to a combination of high surface productivity and slow DO resupply at depth (Wyrtki 1962) and are typically defined by DO concentrations of <20 µM (Paulmier and Ruiz-Pino 2009). However, some tropical OMZ regions are functionally anoxic, with DO concentrations that fall below the limits of detection for even the most sensitive sensors (Thamdrup et al. 2012). These subregions within the broader OMZs are known anoxic marine zones (AMZs; Ulloa et al. 2012) or oxygen-deficient zones (ODZs), and are distinct from OMZs in that there is a pronounced secondary nitrite maximum (SNM) produced through anaerobic N cycling (Ulloa et al. 2012, Bristow et al. 2017). As a result, AMZs host microbial communities possessing a range of metabolisms that are integral to ocean biogeochemistry (Ulloa et al. 2012; Beman and Carolan 2013; Bertagnolli and Stewart 2018).

Despite their biogeochemical importance, multiple lines of recent research indicate that microbial communities present within OMZs/AMZs are more complex and dynamic than previously assumed. In particular, (i) the range of DO levels over which different types of metabolism (both aerobic and anaerobic) are present and active is broader than previously thought; (ii) oxygen supply to OMZs/AMZs is more significant than initially assumed; and (iii) a wider range of microbial processes are now known to be present and active in OMZs/AMZs. For example, above subsurface waters depleted in DO, the depth and severity of the oxycline result in sharp peaks of aerobic activity (ammonia and nitrite oxidation, but also respiration) at the base of the euphotic zone. However, rates of these processes display a tail of activity that can extend deeper into the water column in spite of the low DO concentrations found within OMZs/AMZs (Tiano et al. 2014; Kalvelage et al. 2015; Beman et al. 2021). Consistent with this, many aerobic organisms have high affinity cytochromes that allow them to respire DO even at nM concentrations (Morris and Schmidt 2013, Stolper et al. 2010). This wider oxygen tolerance for respiration blurs the line between aerobic and anaerobic metabolism. This concept importantly also applies to different N cycle processes, which have varying oxygen tolerances that allow aerobic and anaerobic processes to overlap in the water column (Dalsgaard et al. 2014; Zakem and Follows 2017; Zakem et al. 2020; Beman et al. 2021; Sun et al. 2023). Related to this, SAR11 are the dominant aerobic heterotrophic bacteria throughout the ocean (Morris et al. 2002), but some present within OMZs/AMZs are capable of nitrate reduction (Tsementzi et al. 2016), and the level at which different groups of organisms ‘switch’ from aerobic respiration to nitrate reduction and/or complete denitrification is more complex than previously thought (Zakem et al. 2020).

With regard to oxygen supply, an additional feature of AMZs is a secondary chlorophyll maximum (SCM) located below the primary chlorophyll maximum (PCM) and composed of low-light adapted *Prochlorococcus* (Lavin et al. 2010). Oxygenic photosynthesis by *Prochlorococcus* present in the SCM produces oxygen that can be metabolized by aerobic heterotrophs and nitrite oxidizing bacteria, resulting in coupled ‘cryptic’ cycling of carbon, oxygen, and nitrogen (Garcia-Robledo et al. 2017; Beman et al. 2021). Deeper in the AMZ water column, the bottom of the SCM overlaps with the top of the SNM generated by anaerobic N metabolism (Ulloa et al. 2012; Dalsgaard et al. 2014). Here genes involved in sulfur metabolism are present and expressed (Canfield et al. 2010; Carolan et al. 2015), and S and N metabolism may be coupled via chemoautotrophic denitrification present in genomes of several groups of organisms (Walsh et al. 2009). Other AMZ organisms are capable of anaerobically oxidizing methane (CH_4) using nitrite (Ettwig et al. 2010), while recent work indicates that this process, as well as ammonia oxidizing archaea (Kraft et al. 2022), may be able to produce oxygen in the absence of light. Arsenic metabolism is also present in OMZs (Saunders et al. 2019). Finally, all of these forms of metabolism are coupled to diverse carbon fixation pathways present within different groups of organisms present in OMZs/AMZs (Ruiz-Fernández et al. 2020).

Multiple microbial groups with a range of metabolisms may therefore form complex connections within OMZs/AMZs, and potentially respond to, track, or even create the biogeochemical variations found within and across these regions. However, many studies focus on single organisms or processes present within the broader microbial community—despite the fact that overall biogeochemical patterns are sustained by the genomic potential to conduct different biogeochemical transformations distributed across microbial communities in their entirety. As a result, where different metabolisms are located in the water column, and how they may interact, are not well understood. Multiple pathways and processes also map to different microbial groups that vary within and across OMZs (Wright et al. 2012). The presence and prevalence of different pathways, processes, and organisms in OMZs—and how these may vary in space and time—remain poorly known. Finally, significant biogeochemical variations exist within each of the major OMZs, particularly from nearshore regions versus those further offshore. In the ETSP and in the Arabian Sea, for instance, intense, shallow, but seasonably-variable OMZs are associated with nearshore upwelling, while a more stable OMZ is found offshore (Naqvi et al. 2000; Fuenzalida et al. 2009; Kalvelage et al. 2013; Loescher et al. 2016). Although similar variations may occur in the ETNP (e.g., Pennington et al. 2006; Beman and Carolan 2013; Domínguez-Hernández et al. 2020; Kwiecinski and Babbin 2021), the effects of these variations on microbial communities and metabolism are not well studied. Given the biogeochemical importance of OMZs, and strong variations within and across them, there is still relatively little work examining whole microbial community metagenomics across the gradients found within these regions of the ocean.

We used shotgun metagenomics to resolve microbial groups and their functionality across four stations in the ETNP OMZ that capture a range of biogeochemical conditions. We targeted four key features at these stations: (i) the primary nitrite maximum (PNM) at the base of the euphotic zone, where intense biogeochemical cycling occurs (Beman et al. 2012); (ii) the edge of OMZ waters at 20 µM DO, where microbial communities are diverse (Beman and Carolan 2013; Bertagnolli and Stewart 2018); (iii) the SCM, where cryptic biogeochemical cycling occurs between aerobic and anaerobic organisms and processes (Garcia-Robledo et al. 2017; Zakem et al. 2020); and (iv) the SNM, where anaerobic organisms and metabolisms are present and active (Thamdrup et al. 2012; Ulloa et al. 2012). All of these features were present at three AMZ stations (1, 2, and 3) that extend off the coast of Mexico (Figure 1.1), but which differ in the depth of the features (Figure 1.2). For comparison, we also sampled the PNM, OMZ edge, and the single PCM present at an OMZ—but not AMZ—Station 4 (Figure 1.2, Table 1.1). We examined overall patterns based on metagenomic and oceanographic data, tested for significant variations in genes with depth, and assembled metagenome-assembled genomes (MAGs). We expected and found significant changes in genes and organisms involved in key biogeochemical processes with depth and from station to station.

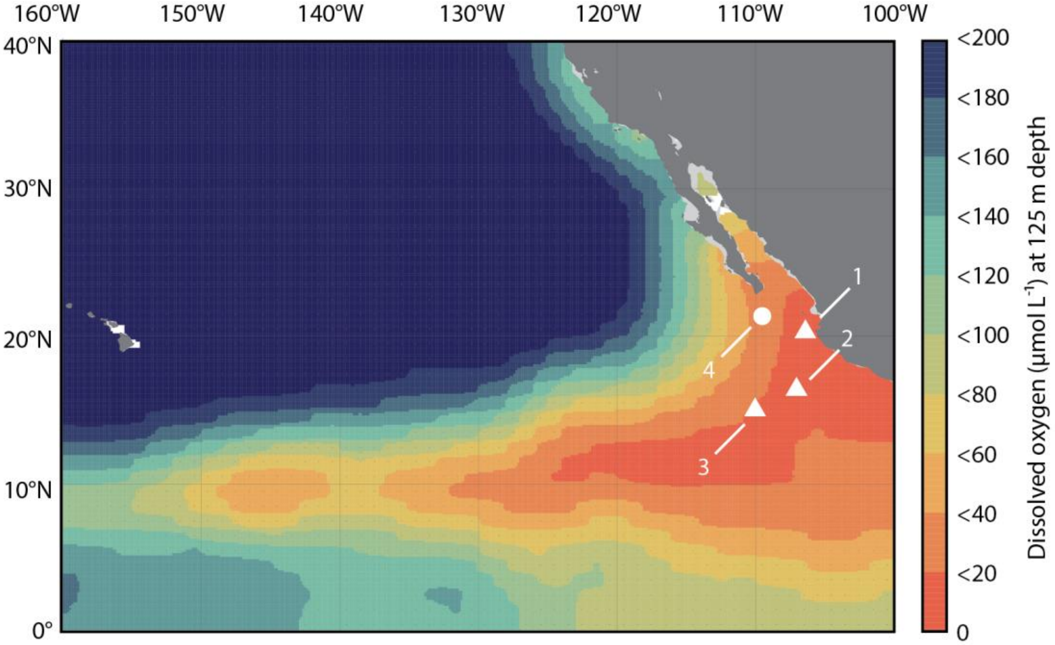


Figure 1.1: Locations of sampling stations in the ETNP (numbered symbols) plotted on dissolved oxygen concentrations (in μM) at 125 m depth from the World Ocean Atlas. Triangles denote anoxic marine zone (AMZ) Stations 1–3 and the circle denotes oxygen minimum zone (OMZ) Station 4. The color scale shows dissolved oxygen concentrations in 20 μM intervals.

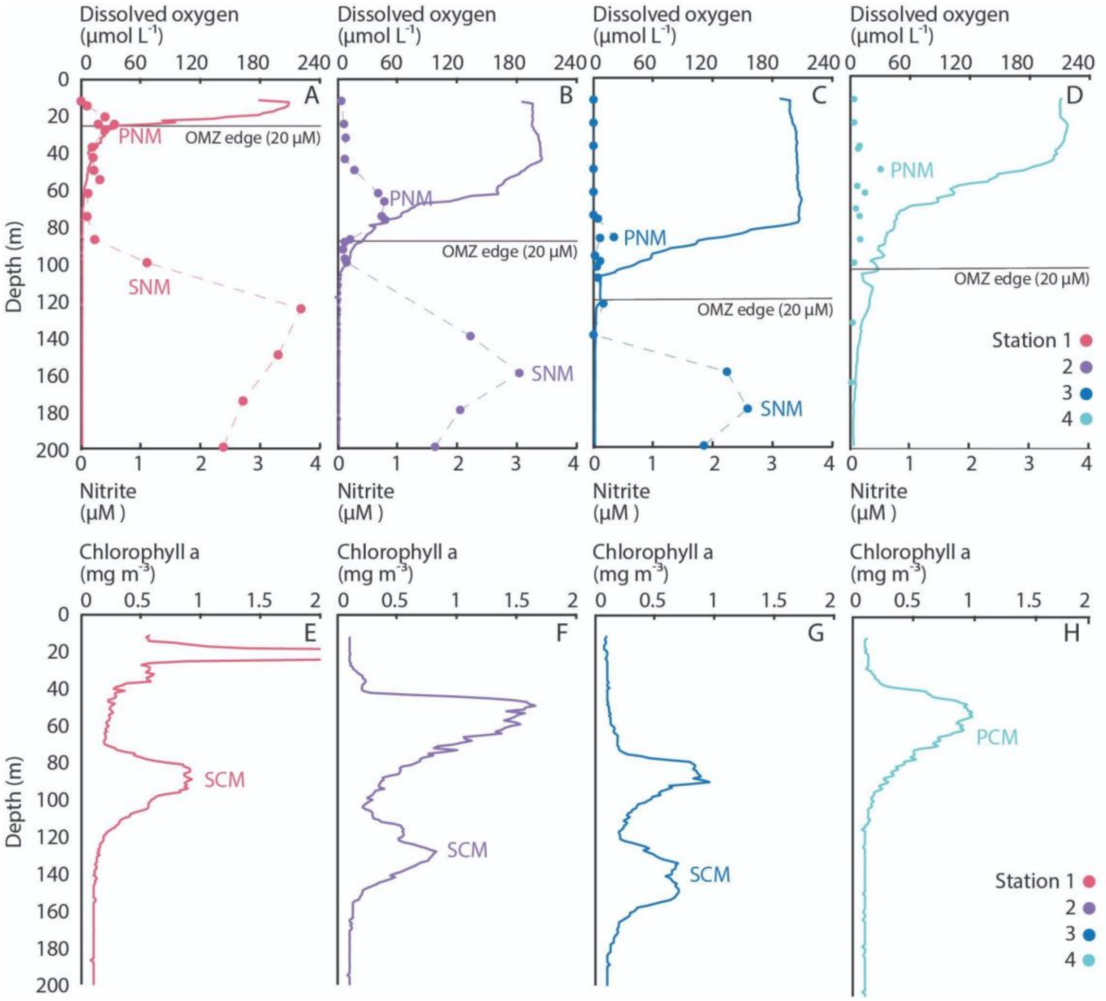


Figure 1.2: Oceanographic profiles illustrate biogeochemical differences among sampling stations (1-4 from left to right and color coded) and with depth (to 200 m). (A-D) Stations 1-3 all exhibit sharp declines in oxygen at different depths (with the OMZ edge marked by the horizontal black line), as well as the accumulation of nitrite at a primary nitrite maximum (PNM) and a higher concentration secondary nitrite maximum (SNM) at depth. Oxygen profiles are shown as a solid line with values on the top axis, and nitrite concentrations by data points with values on the bottom axis. (E-H) Stations 1-3 also show a secondary chlorophyll maximum (SCM) below the primary chlorophyll maximum. Station 4 exhibits a less severe decline in oxygen and a single PNM and PCM.

Table 1.1: Details of sequenced metagenomes

Station	Depth (m)	Sample type	Number of reads	Base pairs
1	25	PNM/OMZ edge	35,454,892	5,282,293,481
1	87.5	SCM	37,300,966	5,523,816,421
1	100	SNM	40,292,372	5,983,420,646
2	67	PNM	42,660,318	6,324,738,264
2	89	OMZ edge	36,166,538	5,380,769,020
2	130	SCM	44,805,440	6,656,012,730
2	160	SNM	38,539,128	5,728,222,387
3	87	PNM	38,499,244	5,720,610,903
3	123	OMZ edge	39,160,822	5,829,653,969
3	140	SCM	41,698,180	6,188,407,419
3	180	SNM	39,881,790	5,935,792,428
4	50	PNM	38,644,992	5,738,423,810
4	75	PCM	25,954,334	3,861,979,553
4	100	OMZ edge	34,336,278	5,100,306,914

1.3 Experimental Procedures

Sampling, DNA Extraction, and Metagenome Sequencing

Samples were collected in April 2017 aboard the *R/V Oceanus*, with samples collected in Mexican territorial waters under Instituto Nacional de Estadística y Geografía (INEGI) permit EG0062017 and Permiso de Pesca de Fomento permit PPFE/DGOPA-016/17. At each station, conductivity/salinity, temperature, depth, pressure, chlorophyll fluorescence, and photosynthetically active radiation (PAR) were measured by a SeaBird SBE-9plus CTD, SBE-3F temperature sensor, SBE-43 DO sensor, WetLabs ECO-FLR Fluorometer, and Biospherical QCP2200 PAR sensor. Nutrient samples were analyzed and community respiration rates measured as described in Beman et al. (2021).

Water samples were collected for DNA extraction and sequencing using sampling bottles deployed on the CTD rosette. At each depth, 2L samples were filtered through 0.22 µm filters (Millipore, Darmstadt, Germany) using a peristaltic pump, then filters were submerged in Sucrose-Tris-EDTA (STE) buffer in pre-prepped Lysis Matrix E tubes and frozen at -80°C until extraction. DNA was extracted from filters following Beman et al. (2012) and DNA samples were sent for metagenome sequencing in the Vincent J. Coates Genome Sequencing Laboratory (GSL) at the University of California, Berkeley (https://genomics.qb3.berkeley.edu/), which is supported by NIH S10

OD018174 Instrumentation Grant. For each sample, 250 ng of genomic DNA was sheared and libraries were prepared using the KAPA HyperPrep Kit (Kapa Biosystems, Wilmington, MA, USA). Samples were pooled into a single lane and sequenced via 150-cycle paired-end sequencing on the Illumina HiSeq 4000 platform (Illumina, Inc., San Diego, CA, USA). Data were demultiplexed by the GSL and reads were filtered and trimmed using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) with the following parameters: maq=8, maxns=1, minlen=40, minlenfraction=0.6, k=23, hdist=1, trimq=12, qtrim=rl. Forward and reverse reads were then merged using PANDASeq (https://github.com/neufeld/pandaseq; Masella et al. 2012) with default parameters.

Functional Mapping and Analysis of Metagenomes

To assess functional diversity, metagenomic sequences from the ETNP OMZ were input into the Functional Mapping and Analysis Pipeline (FMAP; Kim et al. 2016). FMAP aligns sequences to the KEGG database reference genome and uses statistical analysis to identify the abundance of particular features, such as enriched operons and pathways from differentially abundant genes (Kanehisa et al. 2017). In brief, metagenomes were individually mapped to reference proteins on the KEGG database by using programs USEARCH and DIAMOND (Edgar 2010; Buchfink et al. 2015). Once each metagenome was mapped to the database, a mapping file was created for each sample that was then filtered by setting a cutoff on the e-value at <1 e^{-30}. FMAP was then used to determine the abundance of KEGG Orthologs groups (KO) from the mapping files by calculating the number of reads mapped to the KO and by calculating reads per kilobase per million reads (RPKM). This created an abundance file for each sample containing the identification and relative abundance of each gene in RPKM.

These data were used to detect differentially abundant (DA) genes and perform ‘enrichment analysis’ of pathways and operons. To accomplish this, we conducted a total of five comparisons of samples based on where they were collected within the water column. These comparisons were: (1) PNM vs. OMZ edge, (2) PNM vs. SCM, (3) OMZ edge vs. SCM, (4) OMZ edge vs. SNM, and (5) SCM vs. SNM. Each comparison group underwent Kruskal-Wallis testing for analysis of DA genes, producing a comparison file containing the test statistics of the genes, which were then analyzed through a Fisher's exact test. FMAP then creates a pathway text file containing all significantly different pathways detected during the Fisher’s Exact test (Kim et al. 2016). After pathway detection, we identified the genes that were present in pathways that were statistically significantly different (P <0.05). From these data, we generated histograms based on the FMAP gene abundance table, which contains all the genes identified in the ETNP OMZ. Each histogram was created using the histplot function of the seaborn library in python (Waskom 2021) and the subplots command in matplotlib (Hunter 2007).

To visualize differentially abundant genes, a heatmap was created based on the genes that were only found in statistically significant pathways. Before constructing the heatmap we removed genes that had an abundance less than 2000 RPKM from the gene pathway abundance table (the full heatmap is available in the supplementary information). The modified gene pathway abundance table was then converted into a dataframe by using the pandas library in python (McKinney 2011) and into a hierarchical heatmap using the seaborn library by using the clustermap function (Waskom 2021). For hierarchical clustering, the clustering method was set to ‘average’ and the clustering metric to ‘braycurtis.’ The heatmap displays all the genes with an abundance over 2000 RPKM, the statistically significant pathways each gene appears in, the gene id number, and gene definition. Further customization of the heatmap was carried out using the matplotlib library (Hunter 2007).

Principal component analysis and redundancy analysis of metagenomes

Principal component analysis (PCA) was also conducted on the FMAP abundance table that was generated for all ETNP metagenomic samples using the factoextra, factominer, and the ggforce packages in R. With the factominer package, we used the PCA command on the abundance table (with the data scaled to unit variance before the analysis) to create a PCA matrix (Lê et al. 2008). Once the PCA data matrix was created, the PCA was plotted using the factoextra package, where the points on the PCA were the ETNP samples or the water column they were collected (Kassambara 2016). Modifications made on the PCA such as adding the ellipses were conducted using the ggforce package (Pedersen 2020).

We also conducted redundancy analysis (RDA) on the abundance table generated by FMAP using the Vegan package in R (Oksanen et al. 2013). The RDA function was applied to the abundance table along with environmental variables (depth, density, chlorophyll fluorescence, photosynthetically-active radiation, temperature, dissolved oxygen, NH_4^+, NO_2^-, nitrate, phosphate, salinity, and community respiration). After the appliance of the RDA function, the abundance table and the environmental variables were converted into an RDA Matrix and plotted using the ggord package in R. Modifications made to the ggord plot are done with ggplot2 (Wickham 2011). The RDA displays all samples and environmental conditions that were found in the ETNP OMZ.

Microbial Abundance and Profiling with mOTUs

We used the program mOTUs tool (Ruscheweyh et al. 2022) for microbial abundance and profiling of the ETNP OMZ. Similar to FMAP, mOTUs tool accepts raw shotgun metagenomic sequences; each individual shotgun metagenome is inputted into the mOTUs workflow, which creates a text file containing all the identified microbial species and their estimated relative abundance for each sample. We profiled, identified, and estimated microbial abundances at the family level in the ETNP OMZ and all taxonomic files were merged together into one master file summarizing the microbial families and relative abundance across all samples.

We then generated a heatmap from the mOTUs merged master file that displays the identified microbial families and their relative abundance across all samples in the ETNP OMZ, and following the approach above used for statistically DA genes. Families with a relative abundance less than 0.001 were removed.

Metagenome Assembled-Genomes (MAGs)

Metagenome Assembled Genome (MAG) extraction was performed using the Department of Energy (DOE) systems Biology Knowledgebase (Kbase) following the workflow published by Chivian et al. 2023. In brief, metagenomes were assembled into contigs using metaSPADES (Nurk et al. 2017) and then binned into draft MAGs using the three binning applications MaxBin2, MetaBAT2, and CONCOCT (Wu et al. 2016; Kang et al. 2019; Alneberg et al. 2014). After the binning process, binned genomes were input into the DAS-Tool app to obtain optimized consensus bins of each individual program (Sieber et al. 2018). Optimized bins were checked for quality using the CheckM application to determine the completion and contamination of the bins. We then filtered bins by their quality using the CheckM app (Parks et al. 2015). Based on previous research, we elected to filter and keep bins with a genome completion of >70 percent and a contamination of <5 percent.

Filtered, optimized bins were extracted for assemblies and then used for annotation using RAST (Brettin et al. 2015). RAST accomplishes annotation using the curated collection of gene families functional annotations known as SEED (Overbeek et al. 2014). Generates MAGs containing information such as genome length, taxonomy, gene identification, number of identified genes, and genes involved in SEED functions. ETNP MAGs were then taxonomically classified using the app GTDB-Tk (Chaumeil et al 2022).

Data Availability

CTD data generated in this study have been deposited in the Rolling Deck to Repository, with 2017 data available under accession number OC1704A. Nutrient data are available through the Biological and Chemical Oceanography Data Management Office under project number 863208 (https://www.bco-dmo.org/project/863208). Sequence data are available in the Sequence Read Archive, with metagenome data available under BioProject PRJNA634212 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA634212).

1.4 Results

ETNP oceanographic sampling context and metagenome sequencing

We sampled multiple stations across the ocean's largest OMZ in the ETNP, with three sampling stations (1, 2, and 3) extending from near the coast of Mexico to well offshore (Figure 1.1), and an additional sampling station located further north (Station 4; Figure 1.2). Oxygen declined rapidly with depth below the euphotic zone at all of these stations (Fig. 1.2B and F). However, as expected, a notably shallow OMZ at station 1 (25 m) was followed by a progressively deeper OMZ at station 2 (89 m), and station 3 (106 m). These three stations also displayed a pronounced SCM (Fig. 1.2C) and SNM (indicative of anaerobic N cycling) (Fig 1.2D). The SCM and SNM are diagnostic features of AMZs that were, in contrast, absent at station 4. Similar to the progression in OMZ depth that occurred moving from nearshore to offshore, the SCM and SNM were also shallower at station 1 (75-125 m and >100 m, respectively), and deeper at station 2 (SCM 105-155 m and SNM >140 m) followed by station 3 (120-180 m and >160 m). All four stations showed primary nitrite maxima (PNM), ammonium (NH_4^+) maxima, and primary chlorophyll maxima (PCM) near the base of the euphotic zone (ranging from 20-90 m depth; Fig. 1.2). However, Station 1 located nearest the coast had notably high chlorophyll values (5 mg m^{-3}) and NH_4^+ concentrations (433 nM), and the PNM was also co-located with the OMZ at just 25 m depth at this station.

We sequenced metagenomes at several of these diagnostic and biogeochemically-relevant features to examine variations in the functional capabilities of resident microbial communities. At all stations, this included the PNM located at the base of the euphotic zone, as well as the edge of the OMZ (20 µM DO concentration). (These two features were co-located at 25 m depth at Station 1.) At AMZ stations 1-3, we also sampled the SCM and the SNM, while at OMZ station 4 (which lacks an SCM and SNM), we sampled the primary chlorophyll maximum (PCM). For each of these metagenomes, we annotated reads, assigned functions, and tested for significant differences in metagenomic content across samples using the Functional Mapping and Analysis Pipeline (FMAP; Kim et al. 2016).

We found that all samples showed the same overall power law relationship between gene abundance values (reads per kilobase per million mapped reads; RPKM) and frequency, with many genes that were uncommon and a few genes that were very common. For example, across all samples, a total of only 112 genes had RPKM values >2000 (Figure 1.3). Within this broader pattern, there were differences between sample types, with greater skew for PNM samples, while SNM samples were typically less uneven. These patterns suggest metagenomic differences across regions of the water column that we analyzed in multiple ways.

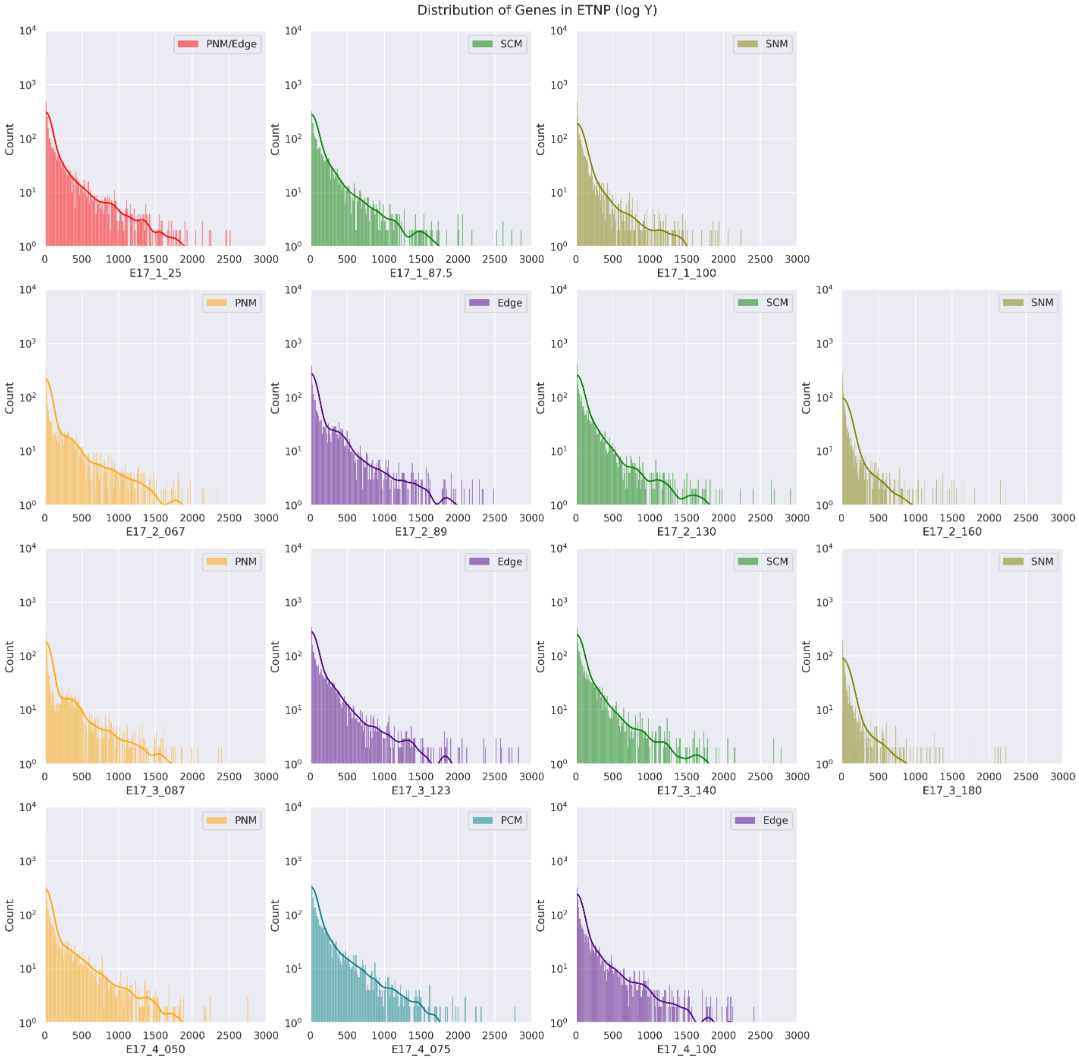


Figure 1.3: Histograms display the trade off between relative abundance and number of genes across different metagenomes. Relative abundance (RPKM) values are shown across the horizontal axis, while the vertical axis (log scale) indicates the total number of genes reaching each RPKM value. Metagenomes from different stations (ordered from top to bottom) and key regions of the water column (color coded and ordered by depth from left to right) are shown with identical axes ranges and the sample name at bottom.

Microbial metagenomics across stations and depth regions

To provide an initial overview of microbial metagenomics in the ETNP, we used Principal Components Analysis (PCA) to examine variations in FMAP-based gene abundances across all stations and depths (Figure 1.4). This analysis showed that samples typically clustered together based on the region of the water column in which they were collected, with several important exceptions. Consistent with distinct biogeochemical conditions expected in the SNM, samples from the SNM (station 1 at 100 m, station 2 at 160 m, and station 3 at 180 m) clustered to the left of the PCA biplot, exhibiting negative values on the PC1 axis and positive values on the PC2 axis. A second group of samples from station 1 at 87.5 m, station 2 at 130 m, station 3 at 140 m and station 4 at 100 m also clustered together on the negative portion of PC1 axis but near 0 on the PC2 axis. Three of these samples were collected from the SCM found at AMZ stations 1-3; however, the station 4 sample was collected at 100m on the OMZ edge, and no SCM was present at this station. A third group of samples exhibited more variation along the PC1 axis, and included OMZ edge samples at station 2 (89 m) and 3 (123 m), as well as a sample collected above the OMZ edge but below the PNM at Station 4 (at 75 m). Finally, three of four PNM samples clustered together on the right side of the diagram (station 2 at 67 m, station 3 at 87 m, and station 4 at 50 m). In contrast, the station 1 PNM sample, which was also located on the edge of the OMZ (collected at 25 m), was distinct from all other samples, falling in the positive regions of both axes. Gene abundance profiles for each sample therefore corresponded with the region of the water column in which they were collected, with the two exceptions of the Station 1 25 m sample and the Station 4 OMZ edge sample (100 m), examined in greater detail below.

In addition to PCA, we used multivariate redundancy analysis (RDA) to examine variation in gene abundances in ETNP metagenomes as a function of environmental variation with depth and between stations (Figure 1.4). Measured variables included a range of oceanographic parameters and properties from CTD sensors, nutrient measurements, and community respiration (CR) rates. We used RDA to examine variations between samples, and also determined the individual gene abundances that drive these variations between samples. Overall, we found that 99% of the variations in FMAP-based gene abundances across samples was explained by environmental variation. 58% of the variation was explained by the RDA1 axis, 15% of variation by the RDA2 axis, and 12% by RDA3, with the remaining axes each contributing a few percent.

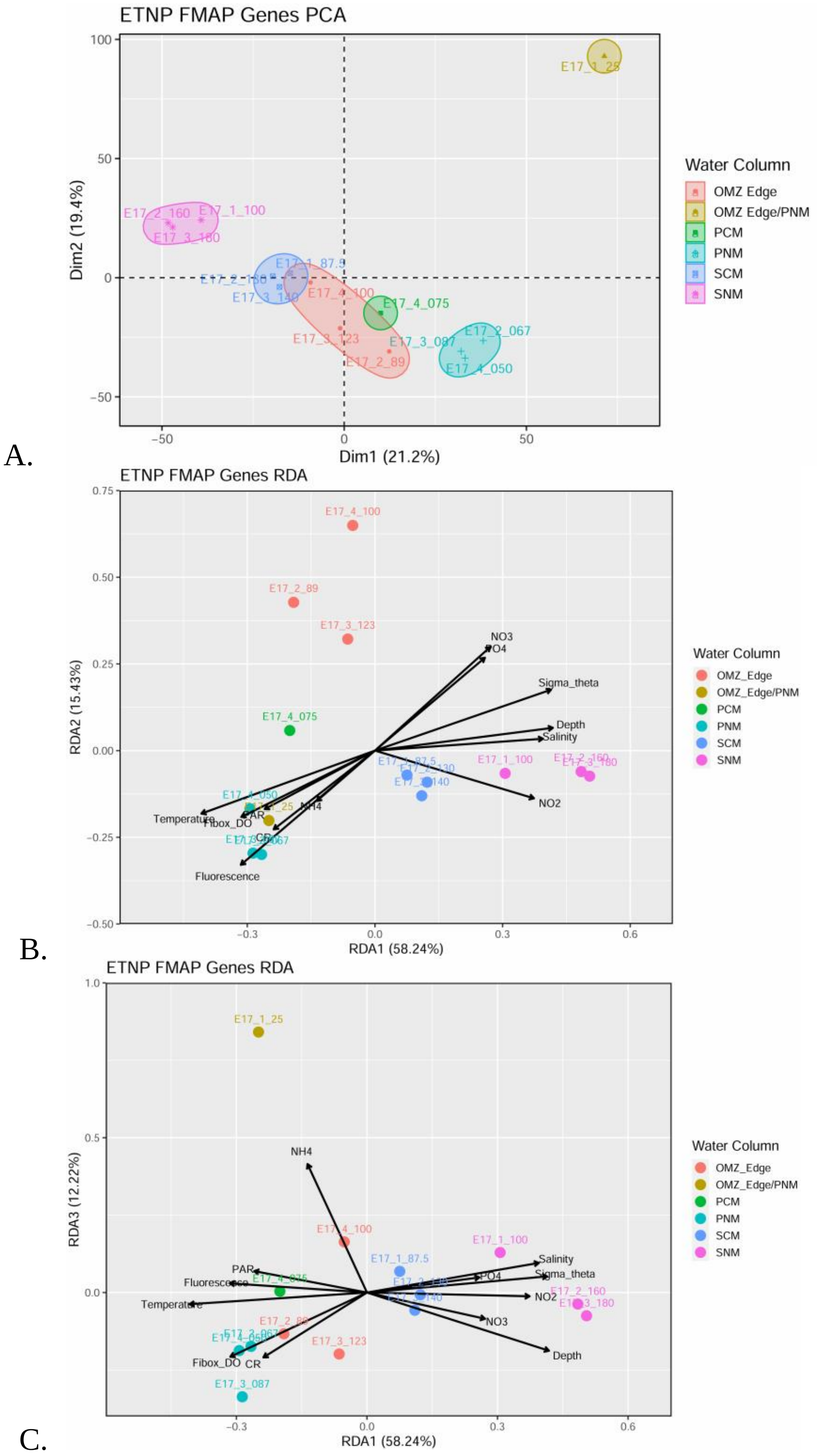


Figure 1.4: (A) Principal Component Analysis (PCA) and (B and C) Redundancy Analysis (RDA) of metagenomes show distinct patterns in metagenome composition based on biogeochemistry of the water column. (A) PCA based on FMAP relative abundances shows close clustering of SNM, SCM, and PNM samples collected across different stations. (B) RDA axes 1 and 2 and (C) RDA axes 1 and 3 show that variations in metagenomes are related to variations in multiple

environmental parameters. Samples are color coded by sample type, and figure axes show percent variation explained by a particular PCA component or RDA axis.

In general, the first two RDA axes were composites of variables that either increased (depth, salinity, density, nitrate, phosphate) or decreased (temperature, light, DO) through the water column, with the former falling in the negative regions of both axes and the latter falling in positive regions of both axes. Variables that exhibited mid-depth peaks included chlorophyll fluorescence, NH_4^+, and nitrite (NO_2^-). Of these, the NO_2^- vector was orthogonal to most of the other variables. All SNM and SCM samples from stations 1-3 fell in this lower right region of the biplot, with the SNM samples obviously strongly related to NO_2^-. SCM samples were influenced by additional variables, falling to the left of SNM samples along the RDA1 axis. PNM samples plotted in the lower left portion of the biplot, reflecting the prevailing conditions found in this depth region (higher temperature, DO, light, etc.). OMZ edge samples, and the 75 m sample from Station 4, were located in the upper portion of the diagram at similar values for RDA1, but were separated along RDA2. Given that RDA3 explained a similar amount of variance as RDA2, we also plotted the RDA3 axis versus RDA1. Many samples showed similar patterns in this biplot, with the clear exception of Station 1 25 m: NH_4^+ was the major contributor to the RDA3 axis, and the Station 1 25 m sample was separated along this axis and strongly related to NH_4^+ concentrations.

We examined overall patterns between the samples and specific variables by regressing their RDA values against each other, and found that the strongest overall correspondence was between the Station 1 25m metagenome and NH_4^+, with an r value of 0.99. This notably high r value explains why this sample is distinct on the PCA biplot, but not in the RDA (specifically RDA3 axis). As one would expect, SNM samples (station 1 at 100 m, station 2 at 160 m, and station 3 at 180 m) were positively related to NO_2^-, with r values of 0.65-0.96. Likewise, SCM samples (station 1 at 87.5 m, station 2 at 130 m, and station 3 at 140 m) were significantly related to chlorophyll fluorescence (r values of 0.15-0.40). PNM samples (station 1 at 25 m, station 2 at 67 m, station 3 at 87 m, station 4 at 50 m) were also related to chlorophyll fluorescence (r values of 0.34-0.52), as well as other variables that are typically elevated in the upper water column (PAR, temperature, DO). Finally, some OMZ edge samples correlated with dissolved nitrate and phosphate—consistent with their spread along the RDA1 axis.

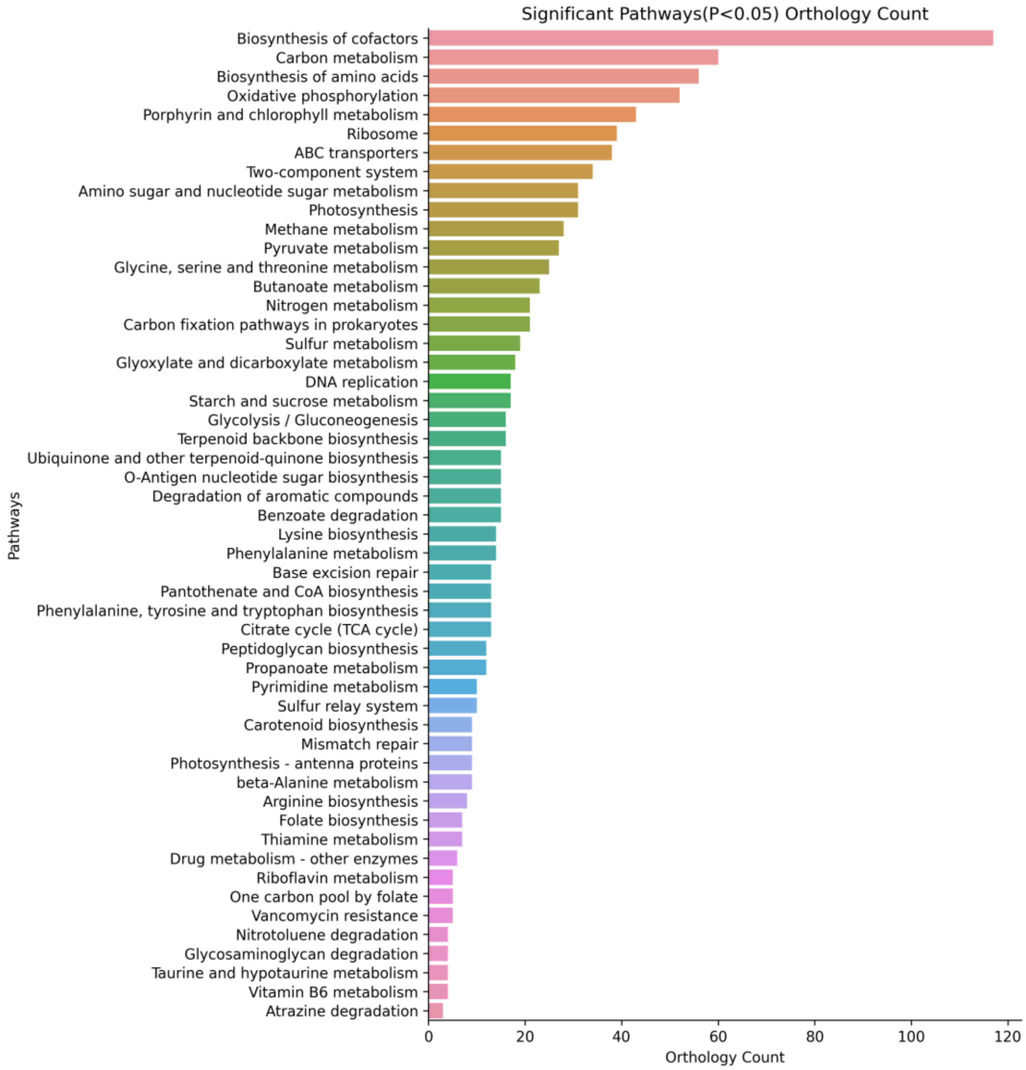


Figure 1.5: Pathways with statistically significant differences between different regions of the water column. Pathways are ordered from highest to lowest orthology count (the number of unrepeated genes identified within each pathway).

Significant differences in metagenomes across different regions of the water column

We used FMAP to test for significant differences (see Experimental Procedures) in the relative abundance of specific genes that drive metagenomic variations between different depth regions (PNM, OMZ edge, SCM, SNM). Overall, 690 genes falling within 52 different pathways showed significant differences in the ETNP water column. The pathway with the highest orthology count between the different habitats sampled was Biosynthesis of cofactors, with 117 genes that were significantly different (Figure 1.5). Biosynthesis of cofactors had nearly double the pathway with the second largest orthology count, which was carbon metabolism with 60 significantly different genes. Many other forms of metabolism and biosynthesis showed significant differences in multiple genes, including in particular: Biosynthesis of amino acids, Porphyrin and chlorophyll metabolism, Amino acid and sugar metabolism, Photosynthesis, Methane metabolism, Nitrogen metabolism, and Sulfur metabolism.

We examined the 690 genes that varied significantly across different regions of the water column by generating a heatmap of gene abundances within samples (Figure 1.6). We also clustered samples based on the genes that were significantly different. In line with the patterns observed for all genes (Figure 1.3), the full heatmap showed that most genes exhibiting significant differences were relatively low in abundance (Supplementary Figure 1). Clustering samples based on these genes also showed that the SNM and SCM samples formed a separate grouping from the OMZ edge and PNM samples. Within these two broad groupings, SNM samples formed a distinct subcluster that was separate from the SCM samples. OMZ edge samples from stations 2, 3, and 4 also clustered together, as did PNM samples from these stations, along with the PCM sample from station 4. However, the PNM/OMZ edge sample from station 1 at 25 m was distinct from these clusters.

These differences were driven by a combination of many lower abundance genes, and several higher abundance genes, that varied significantly between samples collected in different regions of the ETNP water column. The most substantial variations occurred in 16 genes with maximum RPKM values greater >2000 (Fig. 1.6). Of these, nearly all were present in higher abundances in samples collected from the SCM and SNM, with the key exception of the *psbA* gene, which encodes for the D1 protein in photosystem II and is directly involved in oxygenic photosynthesis. As expected based on its role in photosynthesis, *psbA* showed strong variation across different depth regions, with an abundance over 6000 RPKM in the PNM sample station 3 at 87m and over 3000 RPKM in the PNM samples from stations 2 and 4—consistent with the fact that these samples were collected at the base of the euphotic zone where active photosynthesis takes place. *psbA* was also present in OMZ edge samples and SCM samples, but declined to lower abundance in the SNM. Several other genes involved in photosynthesis (*psbB, psbD, psaA*, *psaB*) showed similar patterns at slightly lower overall abundances.

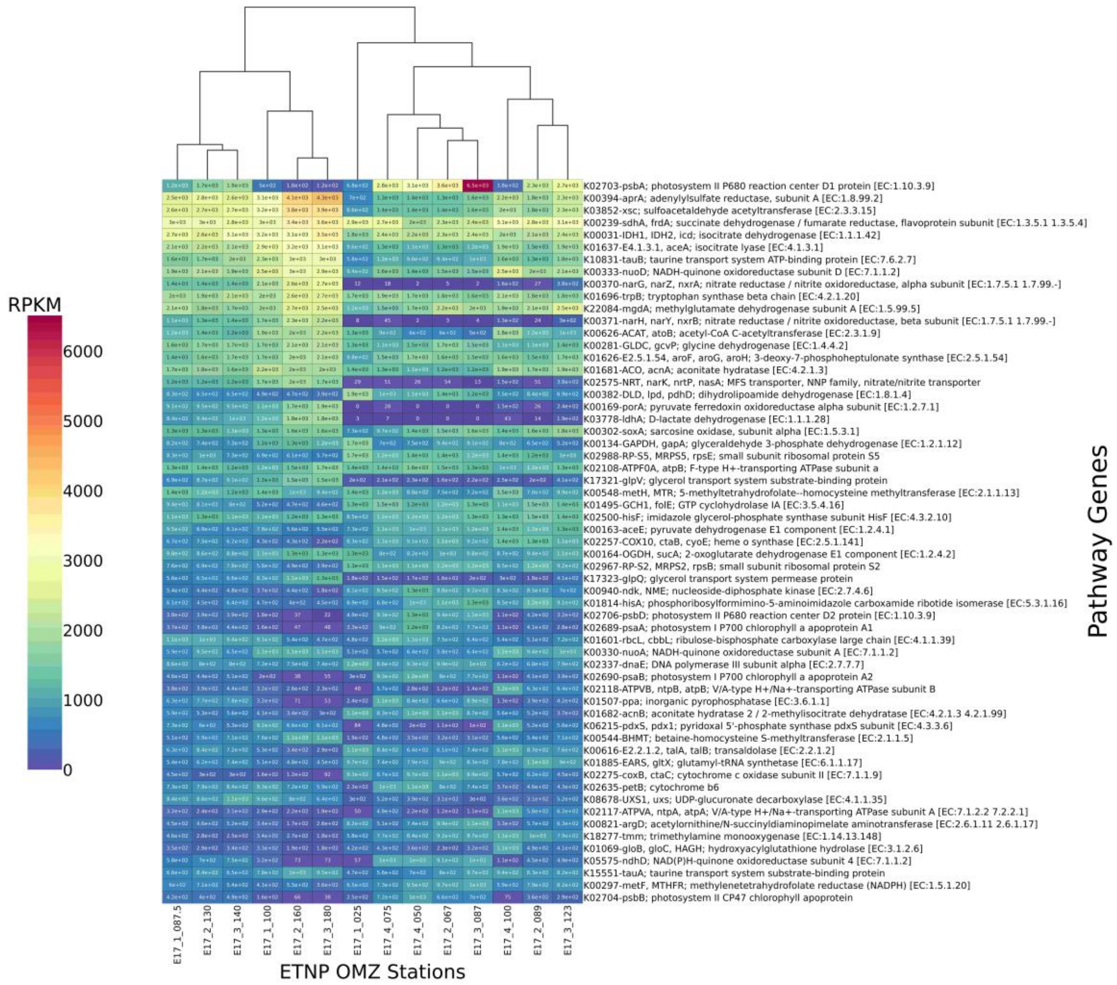


Figure 1.6: Heatmap showing the relative abundance of genes with maximum relative abundances >2000 RPKM (the complete heat map is available in the supplementary information). Genes are shown by color scale at left, ordered by their maximum relative abundance, and listed at right. Sampling clustering is shown at top, and samples are listed along the bottom of the figure. SNM and SCM samples cluster on the left side of the heatmap, while the PNM and OMZ edge samples are located to the right, and the station 1 25 m PNM/OMZ edge sample forms its own cluster in the middle of the heatmap.

In contrast with photosynthesis genes, *aprA* gene abundance increased into the OMZ and especially the AMZ: *aprA* abundance near 1000 RPKM in the upper water column increased to over 2000 RPKM in two OMZ edge samples (station 4 at 100 and station 3 at 123m) and all SCM samples, and reached over 3000 RPKM in all SNM samples. *aprA* encodes for the enzyme adenylylsulfate reductase (dissimilatory adenosine-5′-phosposulfate reductase) involved in sulfur oxidation or dissimilarity sulfate reduction (Meyer and Kuever 2007). In addition, the *xsc* gene involved in oxidation and dissimilation of the sulfur-containing amino acid taurine, and the *tauB* gene involved in uptake and transport of taurine and other sulfur compounds, also showed similar and similarly strong variations into the OMZ/AMZ. Like *aprA*, both genes increased from 1000 RPKM in the PNM to >3000 RPKM in the SNM.

A number of genes involved in carbon metabolism and especially the citric acid (TCA) cycle also increased in abundance with depth into the OMZ/AMZ. These included *sdhA/frdA*, *IDH*, *aceA*, *atoB*, *gcvP*, and *acnA*, all of which increased 2-4 fold from the PNM into the SNM. While genes for carbon fixation therefore increased in the OMZ/AMZ, we found that the *nuoD* gene—part of the bacterial complex I, the first enzyme in the respiratory chain, which transfers electrons from NADH to the ubiquinone pool (Efremov et al. 2010)—also increased with depth. This change was less pronounced, as *nuoD* had an abundance >1000 RPKM in all samples except station 1 at 25m, increasing to over 2000 RPKM in the OMZ edge and SNM samples.

Two genes involved in biosynthesis of amino acids, *trpB* and *aroF*, also showed similar increases with depth, as did the *mgdA* gene. This latter gene encodes for methylglutamate dehydrogenase subunit A, which has multiple functions such as the oxidation of methylamine and using N-methylalanine to produce alanine (Hersh et al. 1972; Latypova et al. 2010; Lin and Wagner 1975). *mgdA* had an abundance over 1000 RPKM across all samples and an abundance over 2000 RPKM in all the SNM samples, as well as some SCM (station 1 at 87.5m) and OMZ edge (Stations 2 and 3) samples.

Finally, the most striking changes with depth were observed for several genes involved in nitrogen cycling, specifically the alpha and beta subunits of nitrate reductase/nitrite oxidoreductase. These were present at only low levels outside of the SCM/SNM, but at >1000 RPKM in SCM and SNM samples

Taxonomic differences in relation to functional differences throughout the water column

These genes are present within different groups of organisms that are also expected to vary throughout the ETNP water column; as a result, we used mOTUs (Ruscheweyh et al. 2022) to assign reads to taxonomic groups and profile their variations in relative abundance across different samples. While many reads per sample could not be taxonomically assigned (27-47%), this analysis again identified distinct patterns in the water column based on taxonomic composition.

Clustering showed four main groupings of samples driven by multiple groups of organisms (Figure 1.7). First, SNM and SCM samples grouped together in the middle of the figure and were distinguished from the other samples by three main differences: (1) higher abundances of *Firmicutes*, *Nitrospinaceae*, *Archaea*, and *Chloroflexi*; (2) lower abundances of *Euryarchaeota*, *Prochloraceae*, and *Flavobacteriaceaea*; and (3) the presence of *Desulfobacteraceae* and several groups of *Planctomyces*.

A second sample grouping included OMZ edge samples from stations 3 and 4, as well as the 75 m (PCM) sample from station 4, and was distinguished in particular by high abundances of *Thaumarchaeota* (right side of Fig. 1.7). Multiple other groups showed higher abundances in this cluster of samples but were relatively low abundance overall (typically 0.01-0.1%), including *Acidimicrobia*, *Woeseiaceae, Cytophagales*, *Gemmatimonadaceae*, *Planctomycetia*, *Alteromonadaceae*, and *Deltaproteobacteria*.

This sample grouping also shared some of the same groups as the third sample grouping, which included PNM samples from station 2 at 67 m, station 3 at 87 m, and station 4 at 50 m, as well as the OMZ edge sample at station 2 (89 m). Both clusters included *Flavobacteriaceae* and *Euryarchaeota,* for example; however, the station 2, 3, and 4 PNM samples had high abundances of *Prochloraceae*. Finally, the OMZ edge/PNM sample station 1 at 25 m clustered separately from the remaining samples and was distinguished by high relative abundances of *Flavobacteriaceae*, *Bacteroidetes*, and *Rhodobacteraceae*.

In addition to these differences between depth regions, there were several groups that were ubiquitous and present in high abundance throughout most of the water column (top of Fig. 1.7). This included Bacteria that could not be identified at higher taxonomic resolution, and especially *Pelagibacteracaea* and *Gammaproteobacteria*. This is consistent with their known importance throughout the ocean for *Pelagibacteraceae* (Morris et al. 2002) and in OMZs for *Gammaproteobacteria* (Stevens and Ulloa 2008).

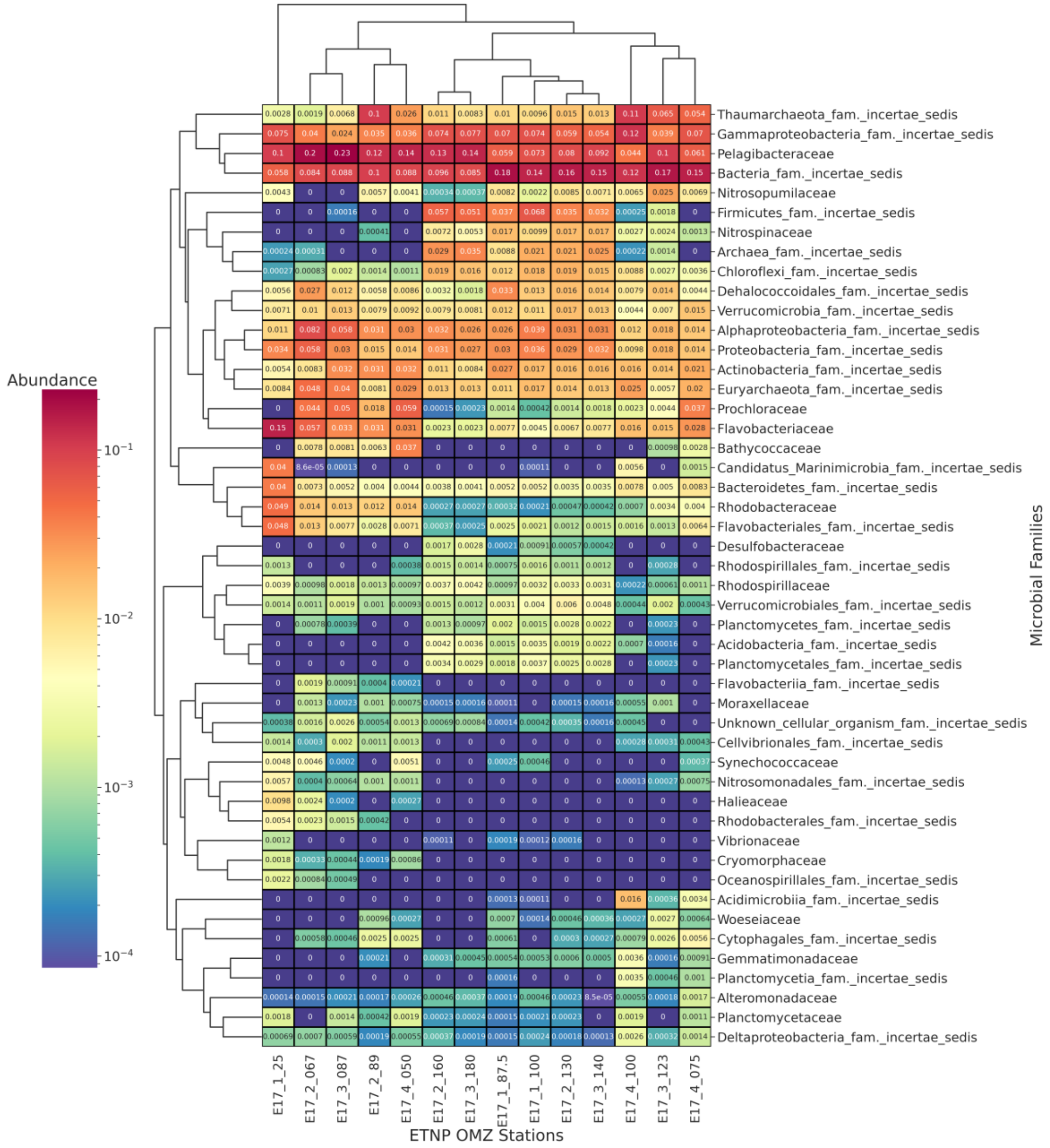


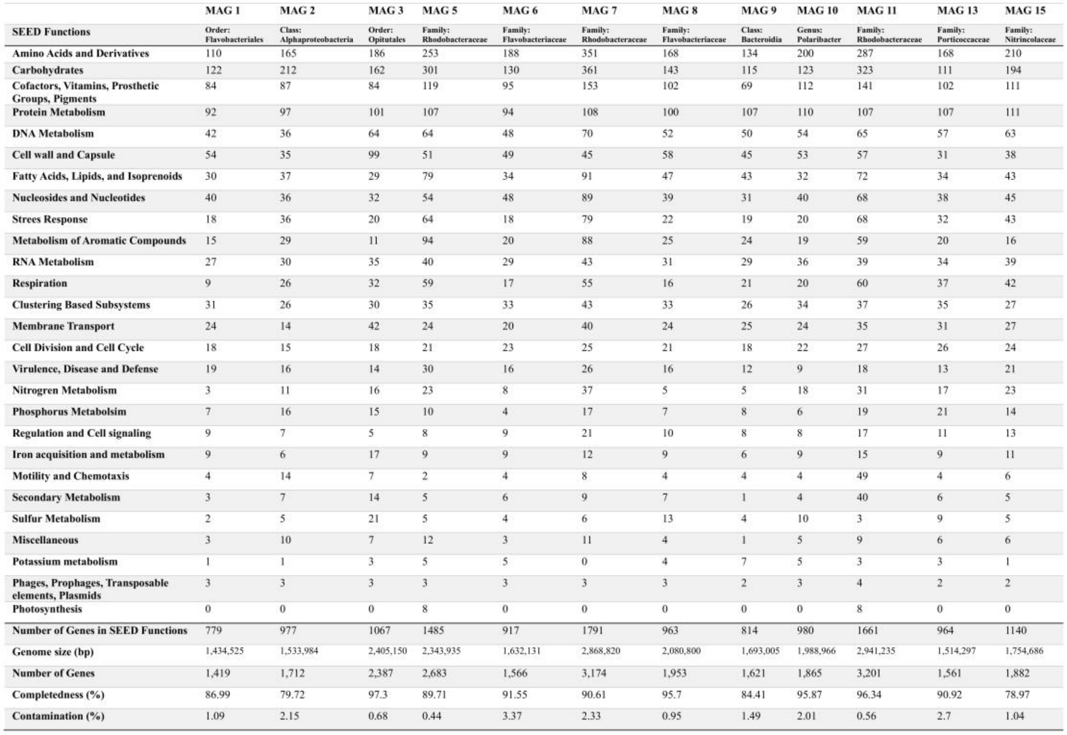
Figure 1.7: Heatmap showing the relative abundance of taxa in metagenomes. Logarithmic color scale at left shows the relative abundance of different taxonomic groups listed on the right side of

the figure. The dendrogram on top of the clustermap shows sample groupings while the dendrogram on the left shows microbial family groupings. All groups with at least 0.001 abundance in one or more samples are shown while unclassified reads are not shown.

Metagenome Assembled Genomes

Due to the unique nature of Station 1 at 25m—where the PNM overlaps with a very shallow OMZ edge, where we observed high chlorophyll (5 mg m^{-3}) and NH_4^+ (433 nM) concentrations, and where metagenomes differed from most other samples (Figs. 1.3, 1.6, 1.7)—we further investigated microbial metagenomics and metabolism through the assembly of metagenomic assembled genomes (MAGs). A total of 12 MAGs with a completion percentage >70% and contamination <5% were assembled at Station 1 at 25m. Of these twelve MAGs, five were identified as *Flavobacteriales* or *Bacteroidia* (with one identified as the genus *Polaribacter* and two within family of *Flavobacteriaceae*) and three others as the family *Rhodobacteraceae* (Table 1.2)—consistent with the prevalence of these groups in our taxonomic analysis (Fig. 1.7). Two other MAGs were identified as *Pseudomonadales*, another MAG as unspecified *Alphaproteobacteria*, and a final MAG as the order *Optutales* under the class *Verrucomicrobia*.

Table 1.2: MAGs in Station 1 at 25m along with total SEED functions



	MAG 1	MAG 2	MAG 3	MAG 5	MAG 6	MAG 7	MAG 8	MAG 9	MAG 10	MAG 11	MAG 13	MAG 15
SEED Functions	Order: Flavobacteriales	Class: Alphaproteobacteria	Order: Opitutales	Family: Rhodobacteraceae	Family: Flavobacteriaceae	Family: Rhodobacteraceae	Family: Flavobacteriaceae	Class: Bacteroidia	Genus: Polaribacter	Family: Rhodobacteraceae	Family: Porticoccaceae	Family: Nitrincolaceae
Amino Acids and Derivatives	110	165	186	253	188	351	168	134	200	287	168	210
Carbohydrates	122	212	162	301	130	361	143	115	123	323	111	194
Cofactors, Vitamins, Prosthetic Groups, Pigments	84	87	84	119	95	153	102	69	112	141	102	111
Protein Metabolism	92	97	101	107	94	108	100	107	110	107	107	111
DNA Metabolism	42	36	64	64	48	70	52	50	54	65	57	63
Cell wall and Capsule	54	35	99	51	49	45	58	45	53	57	31	38
Fatty Acids, Lipids, and Isoprenoids	30	37	29	79	34	91	47	43	32	72	34	43
Nucleosides and Nucleotides	40	36	32	54	48	89	39	31	40	68	38	45
Strees Response	18	36	20	64	18	79	22	19	20	68	32	43
Metabolism of Aromatic Compounds	15	29	11	94	20	88	25	24	19	59	20	16
RNA Metabolism	27	30	35	40	29	43	31	29	36	39	34	39
Respiration	9	26	32	59	17	55	16	21	20	60	37	42
Clustering Based Subsystems	31	26	30	35	33	43	33	26	34	37	35	27
Membrane Transport	24	14	42	24	20	40	24	25	24	35	31	27
Cell Division and Cell Cycle	18	15	18	21	23	25	21	18	22	27	26	24
Virulence, Disease and Defense	19	16	14	30	16	26	16	12	9	18	13	21
Nitrogren Metabolism	3	11	16	23	8	37	5	5	18	31	17	23
Phosphorus Metabolsim	7	16	15	10	4	17	7	8	6	19	21	14
Regulation and Cell signaling	9	7	5	8	9	21	10	8	8	17	11	13
Iron acquisition and metabolism	9	6	17	9	9	12	9	6	9	15	9	11
Motility and Chemotaxis	4	14	7	2	4	8	4	4	4	49	4	6
Secondary Metabolism	3	7	14	5	6	9	7	1	4	40	6	5
Sulfur Metabolism	2	5	21	5	4	6	13	4	10	3	9	5
Miscellaneous	3	10	7	12	3	11	4	1	5	9	6	6
Potassium metabolism	1	1	3	5	5	0	4	7	5	3	3	1
Phages, Prophages, Transposable elements, Plasmids	3	3	3	3	3	3	3	2	3	4	2	2
Photosynthesis	0	0	0	8	0	0	0	0	0	8	0	0
Number of Genes in SEED Functions	779	977	1067	1485	917	1791	963	814	980	1661	964	1140
Genome size (bp)	1,434,525	1,533,984	2,405,150	2,343,935	1,632,131	2,868,820	2,080,800	1,693,005	1,988,966	2,941,235	1,514,297	1,754,686
Number of Genes	1,419	1,712	2,387	2,683	1,566	3,174	1,953	1,621	1,865	3,201	1,561	1,882
Completedness (%)	86.99	79.72	97.3	89.71	91.55	90.61	95.7	84.41	95.87	96.34	90.92	78.97
Contamination (%)	1.09	2.15	0.68	0.44	3.37	2.33	0.95	1.49	2.01	0.56	2.7	1.04

For functionality, five of the twelve MAGs contained genes that mainly contributed to Carbohydrate Metabolism (Figure 1.8 and Table 1.2). Three of these five MAGs were *Rhodobacteraceae*, while one of these MAGs was identified as *Flavobacteriales*. The remaining 7 MAGs contained genes that mainly contribute to degradation of Amino Acids and Derivatives, with a majority of these MAGs identified as *Flavobacteriia*. Across all MAGs, the third and fourth pathways that had the most genes were Protein Metabolism and Cofactors, Vitamins, Prosthetic Groups, and Pigments. Interestingly, two of the *Rhodobacteraceae* MAGs (5 and 11) contained genes that code for Photosynthesis. This is consistent with their known role in aerobic anoxygenic photosynthesis in other regions of the ocean (Spring et al. 2009; Pohlner et al. 2019; Pujalte et al. 2014).

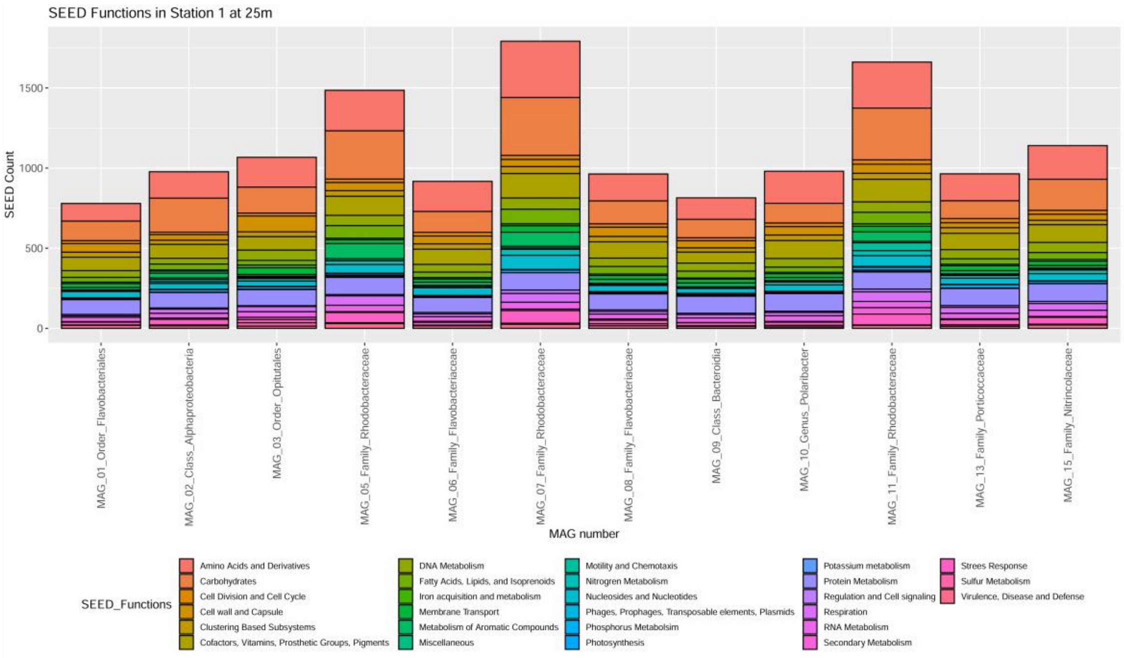


Figure 1.8: Stacked barplot showing SEED functions present in MAGs at Station 1 at 25m. Of 12 assembled MAGs, most were classified as *Flavobacteriales* or *Rhodobacteraceae* (listed below bars). SEED counts are on the y-axis of the barplot, while identified MAGs are on the x-axis. SEED functions are color coded and labeled towards the bottom.

1.5 Discussion

We examined microbial function and composition in the water column of the ETNP OMZ using shotgun metagenomics. Our results showed significant variations in metagenomic content of microbial communities between different depth regions, even within the broader OMZ—e.g., for OMZ edge vs. SCM vs. SNM samples. These variations were consistent across all metagenomic analyses—including PCA, RDA, statistical comparisons in FMAP, and clustering/heatmaps of genes and mOTUs (Figs. 1.3-1.7)—and were also strongly related to chemical gradients present within and across the ETNP in oxygen, chlorophyll, NO_2^-, NH_4^+, and other variables (Figs. 1.2-1.3). This was true even for the OMZ Edge/PNM sample station 1 at 25m, which tended to be isolated or independent from all other samples in nearly all analyses, but was unique in experiencing high NH_4^+ and chlorophyll concentrations. Our findings provide several new insights into microbial metabolism and function in biogeochemically-important OMZs/AMZs.

First, our sampling targeted specific features of known biogeochemical importance in the water column: the PNM at the base of the euphotic zone where photosynthesis occurs, and a hotspot for C respiration and nitrification; the 20 uM oxygen level that defines the edge of the OMZ; the subsurface SCM where oxygen production by specialized *Prochlorococcus* can sustain multiple other key microbial functional groups; and the SNM generated by anaerobic N cycling, where other anaerobic metabolisms should occur. Gene content of metagenomes generally closely followed this sampling (Figs. 1.4 and1. 6). This was particularly evident for SCM and SNM samples, which showed remarkable similarities across stations 1, 2, and 3—despite spanning 100s of km of distance from the near the coast of Mexico to well offshore, and despite differences in the depths of these features moving across this sampling transect. PNM samples were also similar across stations, with the notable exception of the Station 1 25 m sample (discussed in greater detail below). However, OMZ edge samples were more variable in metagenome gene content compared with PNM, SCM, and SNM samples (Figure 1.4). This was driven partly by gradients in many genes with depth through the water column, which results in a broad range of genes being present on the edge of the OMZ, as well as additional genes that were unique to the OMZ edge (Figure 1.6). A similar pattern was evident in taxonomic data, with some groups declining in abundance from the upper water column, others increasing in abundance with depth, and some particularly abundant at the OMZ edge (e.g., *Thaumarcheota*; Figure 1.7). These taxonomic patterns are consistent with high diversity at the edge of the OMZ observed previously based on 16S rRNA, and the idea that this reflects a blend of different metabolisms found in this region of the water column (Beman and Carolan 2013; Bertagnolli and Stewart 2018).

We identified genes driving significant differences with depth and these included known and new potential metabolisms of interest in OMZs/AMZs. For example, photosynthesis is of known importance in generating OMZs/AMZs through production at the surface that fuels respiration at depth, and is also critical in generating the SCM feature found in AMZs (but absent from OMZs). We found strong variations in the relative abundance of photosynthesis genes as expected throughout the ETNP water column, including high abundances in the PNM, as well as elevated abundances in the SCM. In line with this, we found significant differences in genes involved in several carbon fixation pathways. Previous work across the three major OMZs/AMZs showed consistent relative abundances of genes involved in the CBB cycle with depth, whereas genes from additional pathways were more prevalent in OMZ/AMZ waters (Ruiz-Fernández et al. 2020). We found a similar pattern for genes in the TCA cycle but also detected significant shifts in *rbcL/cbbL* genes. Consistent with the SNM serving as a diagnostic feature of AMZs, we also found that nitrate reductase genes (which produce nitrite) showed significant variations with depth. *nar* genes have been used at the global scale to reveal patterns in nutrient limitation (Ustick et al. 2021), and given their potential sensitivity, our results indicate that they could serve as useful biomarkers for anaerobic N cycling, the emergence of the SNM, and functional anoxia within OMZs that have transitioned to AMZs. Finally, we also found evidence for S cycling in the ETNP AMZ based on significant variations especially in the *aprA* gene, as well as *xsc* and *tauB* genes. While recent work in OMZs/AMZs has indicated that inorganic S transformations take place (Canfield et al. 2010; Carolan et al. 2015; Callbeck et al. 2021), our results suggest that organic S metabolism may also be relevant. Collectively these data show systematic variations in metagenomes as a whole—as well as in functional genes involved in key biogeochemical processes—across different biogeochemical features (PNM, OMZ edge, SCM, SNM) found with depth in the OMZ water column.

However, another consistent aspect of our taxonomic and functional metagenomic data was that the OMZ Edge/PNM sample collected at Station 1 at 25m was statistically distinct from the other samples. This may be attributed to the unique environmental gradients of this sample, as the PNM and OMZ edge overlapped at a comparatively shallow depth at Station 1, and there were highly elevated NH_4^+ concentrations and chlorophyll concentrations. Sampling occurred during a phytoplankton bloom, which can be beneficial for opportunistic heterotrophic prokaryotes for carbon acquisition. Due to this sample's unique set of environmental conditions and metagenomic content, we constructed Metagenome Assembled Genomes (MAGs), with 5 of 12 MAGs identified as *Bacteroidia* or *Flavobacteriales* (one *Polaribacter* and two *Flavobacteriaceae)* and 3 of 12 MAGs identified as *Rhodobacteraceae*. *Flavobacteriales* and *Rhodobacteraceae* have been identified in the ETNP, ETSP, and Arabian Sea OMZs primarily through 16S data (Beman & Carolan 2013; Guo et al. 2022; Rajpathak et al. 2018; Medina Faull et al. 2020). In these regions, both groups have a high abundance in the euphotic zone or in high light areas (Stevens & Ulloa 2008). For example, recent work in the ETNP OMZ based on 16S showed that *Rhodobacterales* and *Flavobacteriales* were abundant in the euphotic zone and were typically associated with algal blooms, elevated nutrient concentrations, and productive waters (Pajares et al. 2020). *Flavobacteriales* also occurred in regions of higher NO_2^- and NH_4^+ concentrations near the base of the euphotic zone in the ETNP OMZ (Medina Faull et al. 2020). Our metagenomic results, particularly the higher abundances we observed in the Station 1 25 m sample, are consistent with these findings. *Flavobacteriales* (specifically *Flavobacteriaceae)* and *Rhodobacteraceae* have also been observed in the ETSP OMZ off Chile, where both families are abundant in the surface, while *Rhodobacteraceae* were also detected in the deep oxycline of the OMZ (Stevens & Ulloa 2008). In the Bay of Bengal OMZ, *Flavobacteriales* and *Rhodobacteraceae* were also abundant in the euphotic zone (Gu et al. 2022; Fernandes et al. 2019), while *Rhodobacteraceae* was also detected in the deep chlorophyll maxima during an algal bloom (Gu et al. 2022). This is consistent with their known association with high productivity and phytoplankton, as both microbial taxa are heterotrophs that acquire organic matter left by phytoplankton blooms (Buchan et al. 2014). In the ETSP OMZ off central Chile, *Bacteroidetes (Flavobacteriales)* and *Roseobacter* also had a higher abundance during the upwelling season, which favors high phytoplankton activity (Aldunate et al. 2018). Our results from the ETNP are also consistent with this, given the prevalence of these groups under high ammonium and chlorophyll concentrations in the Station 1 25 m sample.

However, few if any *Flavobacteriales* and *Rhodobacteraceae* MAGs have been recovered from OMZs compared with other regions of the ocean where their metabolic potential has been examined (Priest et al 2022; Acinas et al. 2021)—yet MAGs may provide insight into their functional roles in OMZs, making our data particularly important. In our MAGs identified as *Flavobacteriales*, we found that many genes encoded SEED Functions for degradation of carbohydrates and amino acids and derivatives. In two of the four *Flavobacteriales* MAGs classified as *Flavobacteriaceae*, we identified glycoside hydrolases genes that encode for important exoenzymes for degrading polysaccharides or polymers (Lapébie et al 2019). This extends previous observations made in other areas of the ocean to OMZs: *Flavobacteriia* were previously identified during a phytoplankton bloom with CAZymes specific to the family of glycoside hydrolases including GH29, GH92, and GH90 (Teeling et al 2012), while *Bacteroidetes* MAGs from during the North Sea spring bloom possessed polysaccharide utilization loci containing genes that encode for carbohydrate-active enzymes (such as glycoside hydrolysis and polysaccharide lyases, or carbohydrate binding domains; Krüger et al. 2019). Our *Rhodobacteraceae* MAGs likewise contained many genes that encoded for degradations of carbohydrates and amino acids and derivatives. In the ETSP OMZ, *Rhodobacteraceae* genomes contained the metabolic capability of energy and carbon acquisition including degradation of polymers such as mono, di, oligo, and polysaccharides (Martínez-Pérez et al. 2018). Our results in combination with earlier work therefore indicate that *Flavobacteriaceae* and *Rhodobacteraceae* play a key role in the breakdown of organic matter in the ETNP, and may ultimately affect the formation of the OMZ and subsurface biogeochemistry. However, in addition to carbohydrate and amino acid metabolism, all *Rhodobacteraceae* MAGs present at Station 1 25m contained genes involved in photosynthesis. *Rhodobacteraceae* are known to be capable of aerobic anoxygenic photosynthesis (AAP) possibly acquired through horizontal gene transfer (Brinkmann et al. 2018). Our results indicate that AAP *Rhodobacteraceae* MAGs are prevalent in the ETNP, where they may couple energy acquisition from solar radiation to the breakdown of organic compounds.

Collectively our results underline both the fact that strong variations in subsurface microbial metabolisms and communities are found with depth in OMZs, as well as the fact that significant variations occur across upper portions of the OMZ water column. We found consistently high similarity among metagenomes collected at the same key subsurface biogeochemical features present at different stations (Figs. 1.4 and 1.6)—despite systematic differences across stations in the depth of these features, and in the biogeochemistry of the overlying water column (Fig. 1.2). Metagenomes from the upper water column above the OMZ were notably more variable, and reflect differences in microbial metabolic capabilities that tracked biogeochemical variability. Microbial communities in the upper water column may play an important role in regulating how much organic matter reaches OMZ waters, and hence the types of metabolism found there. We found that *Flavobacteria* and *Rhodobacteria* appear to be important in the ETNP above the OMZ and associated with phytoplankton blooms. How much carbon is remineralized and respired by these heterotrophs above the OMZ—or alternatively reaches deeper depths where it could contribute to oxygen consumption or anaerobic heterotrophy—should be quantified. Given that some of the abundant organism we identified possess the capability for aerobic anoxygenic photosynthesis, and that previous studies have identified some of the highest rates in this particular region of the ocean (Kolber et al. 2000), our results indicate that aerobic anoxygenic photosynthesis may be particularly critical to understand in the dynamic upper water column of the ETNP.

Acknowledgments

This work was supported by NSF CAREER Grant OCE-1555375 to J.M.B., and UCMEXUS-CONACyT Collaborative Grants Program joint awards to J.M.B. and J.Q.G.M.. We thank the officers and crew of the R/V Oceanus, Elisabet Perez-Coronel, Daniela Alonso, DeVonyo Bills, and Jorge Montiel Molina for their assistance while at sea.

1.6 References

Acinas, S. G., Sánchez, P., Salazar, G., Cornejo-Castillo, F. M., Sebastián, M., Logares, R., ... & Gasol, J. M. (2021). Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Communications Biology, 4(1), 604.

Aldunate, M., De la Iglesia, R., Bertagnolli, A. D., & Ulloa, O. (2018). Oxygen modulates bacterial community composition in the coastal upwelling waters off central Chile. Deep Sea Research Part II: Topical Studies in Oceanography, 156, 68-79.

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. Nature methods, 11(11), 1144-1146.

Beman, J. M., & Carolan, M. T. (2013). Deoxygenation alters bacterial diversity and community composition in the ocean’s largest oxygen minimum zone. Nature Communications, 4(1), 2705.

Beman, J.M., Popp, B. N., & Alford, S. E. (2012). Quantification of ammonia oxidation rates and ammonia-oxidizing archaea and bacteria at high resolution in the Gulf of California and eastern tropical North Pacific Ocean. Limnology and Oceanography, 57(3), 711-726.

Beman, J. M., Vargas, S. M., Wilson, J. M., Perez-Coronel, E., Karolewski, J. S., Vazquez, S., ... & Wankel, S. D. (2021). Substantial oxygen consumption by aerobic nitrite oxidation in oceanic oxygen minimum zones. Nature Communications, 12(1), 7043.

Bertagnolli, A. D., & Stewart, F. J. (2018). Microbial niches in marine oxygen minimum zones. Nature Reviews Microbiology, 16(12), 723-729.

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(1), 1-6.

Brinkmann, H., Göker, M., Koblížek, M., Wagner-Döbler, I., & Petersen, J. (2018). Horizontal operon transfer, plasmids, and the evolution of photosynthesis in Rhodobacteraceae. The ISME journal, 12(8), 1994-2010.

Bristow, L. A., Callbeck, C. M., Larsen, M., Altabet, M. A., Dekaezemacker, J., Forth, M., ... & Canfield, D. E. (2017). N2 production rates limited by nitrite availability in the Bay of Bengal oxygen minimum zone. Nature Geoscience, 10(1), 24-29.

Buchan, A., LeCleir, G. R., Gulvik, C. A., & González, J. M. (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. Nature Reviews Microbiology, 12(10), 686-698.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature methods, 12(1), 59-60.

Callbeck, C. M., Canfield, D. E., Kuypers, M. M., Yilmaz, P., Lavik, G., Thamdrup, B., ... & Bristow, L. A. (2021). Sulfur cycling in oceanic oxygen minimum zones. Limnology and Oceanography, 66(6), 2360-2392.

Canfield, D. E., Stewart, F. J., Thamdrup, B., De Brabandere, L., Dalsgaard, T., Delong, E. F., ... & Ulloa, O. (2010). A cryptic sulfur cycle in oxygen-minimum–zone waters off the Chilean coast. Science, 330(6009), 1375-1378.

Carolan, M. T., Smith, J. M., & Beman, J. M. (2015). Transcriptomic evidence for microbial sulfur cycling in the eastern tropical North Pacific oxygen minimum zone. *Frontiers in Microbiology*, *6*, 137342.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. Bioinformatics, 38(23), 5315-5316.

Chivian, D., Jungbluth, S. P., Dehal, P. S., Wood-Charlson, E. M., Canon, R. S., Allen, B. H., ... & Arkin, A. P. (2023). Metagenome-assembled genome extraction and analysis from microbiomes using KBase. Nature Protocols, 18(1), 208-238.

Dalsgaard, T., Stewart, F. J., Thamdrup, B., De Brabandere, L., Revsbech, N. P., Ulloa, O., ... & DeLong, E. F. (2014). Oxygen at nanomolar levels reversibly suppresses process rates and gene expression in anammox and denitrification in the oxygen minimum zone off northern Chile. MBio, 5(6), 10-1128.

Domínguez-Hernández, G., Cepeda-Morales, J., Soto-Mardones, L., Rivera-Caicedo, J. P., Romero-Rodriguez, D. A., Inda-Diaz, E. A., ... & Romero-Bañuelos, C. (2020). Semi-annual variations of chlorophyll concentration on the Eastern Tropical Pacific coast of Mexico. Advances in Space Research, 65(11), 2595-2607.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-2461.

Efremov, R. G., Baradaran, R., & Sazanov, L. A. (2010). The architecture of respiratory complex I. Nature, 465(7297), 441-445.

Ettwig, K. F., Butler, M. K., Le Paslier, D., Pelletier, E., Mangenot, S., Kuypers, M. M., ... & Strous, M. (2010). Nitrite-driven anaerobic methane oxidation by oxygenic bacteria. Nature, 464(7288), 543-548.

Fernandes, G. L., Shenoy, B. D., Menezes, L. D., Meena, R. M., & Damare, S. R. (2019). Prokaryotic diversity in oxygen depleted waters of the Bay of Bengal inferred using culture-dependent and-independent methods. Indian journal of microbiology, 59, 193-199.

Fuenzalida, R., Schneider, W., Garcés-Vargas, J., Bravo, L., & Lange, C. (2009). Vertical and horizontal extension of the oxygen minimum zone in the eastern South Pacific Ocean. Deep Sea Research Part II: Topical Studies in Oceanography, 56(16), 992-1003.

Garcia-Robledo, E., Padilla, C. C., Aldunate, M., Stewart, F. J., Ulloa, O., Paulmier, A., ... & Revsbech, N. P. (2017). Cryptic oxygen cycling in anoxic marine zones. Proceedings of the National Academy of Sciences, 114(31), 8319-8324.

Gilly, W. F., Beman, J. M., Litvin, S. Y., & Robison, B. H. (2013). Oceanographic and biological effects of shoaling of the oxygen minimum zone. Annual review of marine science, 5, 393-420.

Gu, B., Liu, J., Cheung, S., Ho, N. H. E., Tan, Y., & Xia, X. (2022). Insights into prokaryotic community and its potential functions in nitrogen metabolism in the Bay of Bengal, a pronounced oxygen minimum zone. Microbiology spectrum, 10(3), e00892-21.

Guo, R., Ma, X., Zhang, J., Liu, C., Thu, C. A., Win, T. N., ... & Wang, P. (2022). Microbial community structures and important taxa across oxygen gradients in the Andaman Sea and eastern Bay of Bengal epipelagic waters. Frontiers in Microbiology, 13, 1041521.

Hersh, L. B., Stark, M. J., Worthen, S., & Fiero, M. K. (1972). N-methylglutamate dehydrogenase: kinetic studies on the solubilized enzyme. Archives of biochemistry and biophysics, 150(1), 219-226.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

Kalvelage, T., Lavik, G., Jensen, M. M., Revsbech, N. P., Löscher, C., Schunck, H., ... & Kuypers, M. M. (2015). Aerobic microbial respiration in oceanic oxygen minimum zones. PloS one, 10(7), e0133526.

Kalvelage, T., Lavik, G., Lam, P., Contreras, S., Arteaga, L., Löscher, C. R., ... & Kuypers, M. M. (2013). Nitrogen cycling driven by organic matter export in the South Pacific oxygen minimum zone. Nature geoscience, 6(3), 228-234.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic acids research, 45(D1), D353-D361.

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, 7, e7359.

Kassambara, A. (2016). Factoextra: extract and visualize the results of multivariate data analyses. R package version, 1.

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., & Zhan, X. (2016). FMAP: functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. BMC bioinformatics, 17, 1-8.

Kolber, Z. S., Van Dover, C. L., Niederman, R. A., & Falkowski, P. G. (2000). Bacterial photosynthesis in surface waters of the open ocean. Nature, 407(6801), 177-179.

Kraft, B., Jehmlich, N., Larsen, M., Bristow, L. A., Könneke, M., Thamdrup, B., & Canfield, D. E. (2022). Oxygen and nitrogen production by an ammonia-oxidizing archaeon. Science, 375(6576), 97-100.

Krüger, K., Chafee, M., Ben Francis, T., Glavina del Rio, T., Becher, D., Schweder, T., ... & Teeling, H. (2019). In marine Bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. The ISME journal, 13(11), 2800-2816.

Kwiecinski, J. V., & Babbin, A. R. (2021). A high-resolution Atlas of the eastern tropical pacific oxygen deficient zones. Global Biogeochemical Cycles, 35(12), e2021GB007001.

Lam, P., & Kuypers, M. M. (2011). Microbial nitrogen cycling processes in oxygen minimum zones. Annual review of marine science, 3, 317-345.

Lapébie, P., Lombard, V., Drula, E., Terrapon, N., & Henrissat, B. (2019). Bacteroidetes use thousands of enzyme combinations to break down glycans. Nature communications, 10(1), 2043.

Latypova, E., Yang, S., Wang, Y. S., Wang, T., Chavkin, T. A., Hackett, M., ... & Kalyuzhnaya, M. G. (2010). Genetics of the glutamate-mediated methylamine utilization pathway in the facultative methylotrophic beta-proteobacterium Methyloversatilis universalis FAM5. Molecular microbiology, 75(2), 426-439.

Lavin, P., González, B., Santibáñez, J. F., Scanlan, D. J., & Ulloa, O. (2010). Novel lineages of Prochlorococcus thrive within the oxygen minimum zone of the eastern tropical South Pacific. Environmental microbiology reports, 2(6), 728-738.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: an R package for multivariate analysis. Journal of statistical software, 25, 1-18.

Lin, M. C., & Wagner, C. (1975). Purification and characterization of N-methylalanine dehydrogenase. Journal of Biological Chemistry, 250(10), 3746-3751.

Loescher, C. R., Bange, H. W., Schmitz, R. A., Callbeck, C. M., Engel, A., Hauss, H., ... & Wagner, H. (2016). Water column biogeochemistry of oxygen minimum zones in the eastern tropical North Atlantic and eastern tropical South Pacific oceans. Biogeosciences, 13(12), 3585-3606.

Martínez-Pérez, C., Mohr, W., Schwedt, A., Dürschlag, J., Callbeck, C. M., Schunck, H., ... & Kuypers, M. M. (2018). Metabolic versatility of a novel N2-fixing Alphaproteobacterium isolated from a marine oxygen minimum zone. Environmental Microbiology, 20(2), 755-768.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

Medina Faull, L., Mara, P., Taylor, G. T., & Edgcomb, V. P. (2020). Imprint of trace dissolved oxygen on prokaryoplankton community structure in an oxygen minimum zone. Frontiers in Marine Science, 7, 360.

Meyer, B., & Kuever, J. (2007). Molecular analysis of the diversity of sulfate-reducing and sulfur-oxidizing prokaryotes in the environment, using aprA as functional marker gene. Applied and Environmental Microbiology, 73(23), 7664-7679.

Morris, R. L., & Schmidt, T. M. (2013). Shallow breathing: bacterial life at low O2. Nature Reviews Microbiology, 11(3), 205-212.

Morris, R. M., Rappé, M. S., Connon, S. A., Vergin, K. L., Siebold, W. A., Carlson, C. A., & Giovannoni, S. J. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. *Nature*, *420*(6917), 806-810.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. BMC bioinformatics, 13, 1-7.

Naqvi, S. W. A., Jayakumar, D. A., Narvekar, P. V., Naik, H., Sarma, V. V. S. S., D'souza, W., ... & George, M. D. (2000). Increased marine production of N2O due to intensifying anoxia on the Indian continental shelf. Nature, 408(6810), 346-349.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome research, 27(5), 824-834.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., ... & Oksanen, M. J. (2013). Package 'vegan'. Community ecology package, version, 2(9), 1-295.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., ... & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research, 42(D1), D206-D214.

Pajares, S., Varona-Cordero, F., & Hernández-Becerril, D. U. (2020). Spatial distribution patterns of bacterioplankton in the oxygen minimum zone of the tropical mexican pacific. Microbial ecology, 80(3), 519-536.

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, 25(7), 1043-1055.

Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. Progress in Oceanography, 80(3-4), 113-128.

Pedersen, T. L. (2020). ggforce: Accelerating "ggplot2." R package version 0.3. 3.

Pennington, J. T., Mahoney, K. L., Kuwahara, V. S., Kolber, D. D., Calienes, R., & Chavez, F. P. (2006). Primary production in the eastern tropical Pacific: A review. Progress in oceanography, 69(2-4), 285-317.

Pohlner, M., Dlugosch, L., Wemheuer, B., Mills, H., Engelen, B., & Reese, B. K. (2019). The majority of active Rhodobacteraceae in marine sediments belong to uncultured genera: a molecular approach to link their distribution to environmental conditions. Frontiers in microbiology, 10, 443927.

Priest, T., Heins, A., Harder, J., Amann, R., & Fuchs, B. M. (2022). Niche partitioning of the ubiquitous and ecologically relevant NS5 marine group. The ISME Journal, 16(6), 1570-1582.

Pujalte, M. J., Lucena, T., Ruvira, M. A., Arahal, D. R., & Macián, M. C. (2014). The family rhodobacteraceae. Springer.

Rajpathak, S. N., Banerjee, R., Mishra, P. G., Khedkar, A. M., Patil, Y. M., Joshi, S. R., & Deobagkar, D. D. (2018). An exploration of microbial and associated functional diversity in the OMZ and non-OMZ areas in the Bay of Bengal. Journal of biosciences, 43, 635-648.

Ruiz-Fernández, P., Ramírez-Flandes, S., Rodríguez-León, E., & Ulloa, O. (2020). Autotrophic carbon fixation pathways along the redox gradient in oxygen-depleted oceanic waters. Environmental microbiology reports, 12(3), 334-341.

Ruscheweyh, H. J., Milanese, A., Paoli, L., Karcher, N., Clayssen, Q., Keller, M. I., ... & Sunagawa, S. (2022). Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. Microbiome, 10(1), 212

Saunders, J. K., Fuchsman, C. A., McKay, C., & Rocap, G. (2019). Complete arsenic-based respiratory cycle in the marine microbial communities of pelagic oxygen-deficient zones. Proceedings of the National Academy of Sciences, 116(20), 9925-9930.

Sieber, C. M., 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.

Spring, S., Lünsdorf, H., Fuchs, B. M., & Tindall, B. J. (2009). The photosynthetic apparatus and its regulation in the aerobic gammaproteobacterium Congregibacter litoralis gen. nov., sp. nov. PloS one, 4(3), e4866.

Stevens, H., & Ulloa, O. (2008). Bacterial diversity in the oxygen minimum zone of the eastern tropical South Pacific. Environmental microbiology, 10(5), 1244-1259.

Stolper, D. A., Revsbech, N. P., & Canfield, D. E. (2010). Aerobic growth at nanomolar oxygen concentrations. Proceedings of the National Academy of Sciences, 107(44), 18755-18760.

Sun, X., Frey, C., & Ward, B. B. (2023). Nitrite oxidation across the full oxygen spectrum in the ocean. Global Biogeochemical Cycles, 37(4), e2022GB007548.

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., ... & Amann, R. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science, 336(6081), 608-611.

Tiano, L., Garcia-Robledo, E., Dalsgaard, T., Devol, A. H., Ward, B. B., Ulloa, O., ... & Revsbech, N. P. (2014). Oxygen distribution and aerobic respiration in the north and south eastern tropical Pacific oxygen minimum zones. Deep Sea Research Part I: Oceanographic Research Papers, 94, 173-183.

Thamdrup, B., Dalsgaard, T., & Revsbech, N. P. (2012). Widespread functional anoxia in the oxygen minimum zone of the Eastern South Pacific. Deep Sea Research Part I: Oceanographic Research Papers, 65, 36-45.

Tsementzi, D., Wu, J., Deutsch, S., Nath, S., Rodriguez-R, L. M., Burns, A. S., ... & Stewart, F. J. (2016). SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature, 536(7615), 179-183.

Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M., & Stewart, F. J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences, 109(40), 15996-16003.

Ustick, L. J., Larkin, A. A., Garcia, C. A., Garcia, N. S., Brock, M. L., Lee, J. A., ... & Martiny, A. C. (2021). Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science*, *372*(6539), 287-291.

Walsh, D. A., Zaikova, E., Howes, C. G., Song, Y. C., Wright, J. J., Tringe, S. G., ... & Hallam, S. J. (2009). Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. Science, 326(5952), 578-582.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.

Wickham, H. (2011). ggplot2. Wiley Interdiscplinary reviews

Wright, J. J., Konwar, K. M., & Hallam, S. J. (2012). ga of expanding oxygen minimum zones. Nature Reviews Microbiology, 10(6), 381-394.

Wyrtki, K. (1962, January). The oxygen minima in relation to ocean circulation. In Deep Sea research and oceanographic abstracts (Vol. 9, No. 1-2, pp. 11-23). Elsevier.

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.

Zakem, E. J., & Follows, M. J. (2017). A theoretical basis for a nanomolar critical oxygen concentration. Limnology and Oceanography, 62(2), 795-805.

Zakem, E. J., Mahadevan, A., Lauderdale, J. M., & Follows, M. J. (2020). Stable aerobic and anaerobic coexistence in anoxic marine zones. The ISME Journal, 14(1), 288-301.

Chapter 2: Viral Community Metagenomics in the Eastern Tropical North Pacific Oxygen Minimum Zone

2.1 Abstract

Viruses are the most abundant infecting agents on our planet, directly or indirectly affecting all organisms and ecosystems. However, through infection, viruses can also influence biogeochemical cycling by augmenting host metabolism via introduction of Auxiliary Metabolic Genes (AMGs). Viral AMGs may be particularly diverse and significant in biogeochemically important oxygen minimum zones (OMZs) yet our knowledge of viral ecology and potential influence on the biogeochemistry of these regions is still limited. We examined viral community taxonomy and function (AMGs) in the Eastern Tropical North Pacific (ETNP) Oxygen Minimum Zone (OMZ) through metagenomic sequencing and analysis that specifically targeted key regions of the water column across a range of geographical locations. Viral taxonomic identification in the ETNP OMZ consisted mainly of cyanophages, *Synechococcus* phage, and *Prochlorococcus* phage, yet we also detected eukaryotic algal viruses, as well as phages that infect heterotrophic prokaryotes, such as *Pelagibacter* phage. Cyanophages were detected in the secondary chlorophyll maxima (SCM) and *Pelagibacter* phage in the secondary nitrite maxima (SNM)—indicating a potential role in water column biogeochemistry under the unique conditions found at these depths. AMGs that contained the highest count in the ETNP OMZ contributed to photosynthesis, nucleotides biosynthesis, and carbohydrate metabolism. This included the *psbA* gene important for photosynthesis, the *cobS* gene important for production of Vitamin B12, and the Phosphoribosylglycinamide (PUR) family of genes important for purine synthesis. While phage AMGs included those involved in nitrogen and sulfur cycles, the majority contributed to photosynthesis—highlighting the potential importance of viruses in affecting primary production in the OMZ water column, including the SCM. Our results provide new insight into how viruses in biogeochemically important OMZs can influence microbial hosts to further viral production in harsh environments and indirectly influence biogeochemical cycling.

2.2 Introduction

Microorganisms are pervasive across a wide range of environments and are the most abundant cellular organisms on the planet (Whitman et al. 1998; Kallmeyer et al. 2012). However, viruses outnumber microorganisms by about tenfold (Stern & Sorek 2011; Fuhrman 1999; Wigington et al. 2016) and are found in most places on earth where microbes are present (Suttle 2007). One of the main reservoirs of viral abundance and diversity is the ocean (Suttle 2007), where marine viruses can infect a spectrum of organisms—ranging from single-celled bacteria (known as phage) and archaea (Breitbart et al. 2018) to large multicellular eukaryotes (Munn 2006)—and vary from surface to depth (Suttle 2005). Due to their ability to infect any organism, viruses also have an impact on marine food webs and can influence marine bacteria and algae population distributions and dynamics (Wommack & Colwell 2000, Rohwer & Thurber 2009). As a result, viruses are thought of primarily as predators of prokaryotes and agents of infection through lysis, containing genes that are important for viral reproduction (Weinbauer 2004).

However, viruses (specifically phages) also have direct impact beyond predation by containing genes that alter host metabolism and favor viral replication (Rohwer et al. 2009; Warwick-Dugdale et al. 2019). These genes are known as auxiliary metabolic genes (AMGs), and some of these AMGs have been identified to influence microbial metabolic pathways that contribute to biogeochemical cycling (Brum & Sullivan 2015; Breitbart et al. 2018). For example, cyanophages—phages that infect cyanobacteria such as *Prochlorococcus* and *Synechococcus* phage—have been identified with genes that encode for different proteins involved in oxygenic photosynthesis (Breitbart et al. 2007; Sullivan et al. 2005). These genes include the *petE* gene that codes for the electron carrier plastocyanin (Puxty et al. 2015), as well as *psbA*/*D* genes which are important for photosystem II (Sharon et al. 2007). In addition to photosynthesis genes, cyanophages also contain AMGs related to phosphorus acquisition and cycling—such as the *pstS* gene for phosphate-binding protein and the *phoA* gene for alkaline phosphatase (Zeng & Chisholm 2012). The presence of phosphorus may even dictate the type of phosphorus acquisition genes active in prokaryotes and phages (Zeng & Chisholm 2012). Phages have also been identified with AMGs that contribute to nitrogen cycling—such as the *amoC* gene for ammonia oxidation (Ahlgren et al. 2019)—and sulfur cycling—such as the dissimilatory sulfite reductase subunit C (*dsrC)* gene and the dissimilatory sulfite reductase (*rdsr*) gene (Roux et al. 2016; Anantharaman et al. 2014). Since viruses contain AMGs involved in major biogeochemical cycles, viral ecology is particularly important to examine and understand in regions of intense biogeochemical cycling known as oxygen minimum zones (OMZs).

OMZs are regions of the oceans where oxygen concentrations decline below 20 µM and are expanding due to global warming (Paulmier & Ruiz-Pino 2009; Stramma et al. 2008). The three major OMZs include the Eastern Tropical North Pacific (ETNP), Eastern Tropical South Pacific (ETSP), and the Arabian Sea, all of which have been detected to have regions that are fully anoxic, also termed anoxic marine zones (AMZs; Ulloa et al. 2012). OMZs are regions of substantial nitrogen loss from the ocean, as the dominant metabolic processes that are present include denitrification and anammox (Lam & Kuypers 2011). Due to low oxygen concentrations, OMZs tend to be dominated by microorganisms, and viruses have been discovered to inhabit the OMZs and infect the prokaryotes that contribute to biogeochemical cycling (Long et al. 2021; Cassman et al. 2012; Jurgensen et al. 2022). Viral communities in the ETSP OMZ have been recorded through metagenomic analysis, where phage sequences dominated the surface and anoxic core, while eukaryotic viruses occupied the oxycline (Cassman et al. 2012). Also in the ETSP OMZ, viral alpha and beta diversity was higher in surface waters compared to anoxic waters, and viruses were shown to be endemic, but shared similarities with other OMZ communities (Vik et al. 2021). In both ETSP studies, viral communities were strongly influenced by the oxygen content of the OMZ. Viral AMGs found in the ETSP correlated with physicochemical properties and contributed to biogeochemical cycling, as T4-like myovirus phages contained genes related to nitrogen cycling (Gazitúa et al. 2021). In contrast to the ETSP, fewer studies have examined viral communities in the ETNP OMZ (Fuchsman et al. 2019; Fuchsman et al. 2021; Muratore et al. 2023; Vik et al. 2017); only one study found total viral diversity increased below the oxycline and that viral AMGs contributing to nitrogen or sulfur cycling were low in anoxic regions, while those that modulated host response towards oxygen were high (Jurgensen et al. 2022).

The ETNP is notably the ocean's largest OMZ, and so plays a pivotal role in global biogeochemistry. Given a general lack of information, we investigated viral taxonomy and functionality in the ETNP OMZ to build upon previous studies of viruses inhabiting ecosystems with areas of low oxygenated waters. In this region and the other OMZs, low oxygen conditions also lead to other types of biogeochemical variability with depth. As a consequence, we analyzed metagenomes sampled from four key features present across different stations and depths in the ETNP OMZ: (i) the primary nitrite maximum (PNM) at the base of the euphotic zone, where intense biogeochemical cycling occurs (Beman et al. 2012); (ii) the edge of OMZ waters at 20 µM DO, where microbial communities are diverse (Beman and Carolan 2013; Bertagnolli and Stewart 2018); (iii) the secondary chlorophyll maximum (SCM), where cryptic biogeochemical cycling occurs between aerobic and anaerobic organisms and processes (Garcia-Robledo et al. 2017; Zakem et al. 2020); and (iv) the secondary nitrite maximum (SNM), where anaerobic organisms and metabolisms are present and active (Thamdrup et al. 2012; Ulloa et al. 2012). All of these features were present at three AMZ stations (1, 2, and 3) that extend off the coast of Mexico, but which differ in the depth of the features (Figure 2.1). For comparison, we also sampled the PNM, OMZ edge, and the single PCM present at an OMZ—but not AMZ—Station 4. With the metagenomic sequences, we used a read based and assembled contig approach to identify bacteriophages with diverse AMGs that may be present in the OMZ and play a role in global biogeochemical cycling in these pivotal regions of the ocean.

2.3 Materials and Methods

Sample collection, DNA extraction, and sequencing

Samples were collected in April 2017 aboard the *R/V Oceanus*, with samples collected in Mexican territorial waters under Instituto Nacional de Estadística y Geografía (INEGI) permit EG0062017 and Permiso de Pesca de Fomento permit PPFE/DGOPA-016/17 (Figure 2.1). At each station, conductivity/salinity, temperature, depth, pressure, chlorophyll fluorescence, and photosynthetically active radiation (PAR) were measured by a SeaBird SBE-9plus CTD, SBE-3F temperature sensor, SBE-43 DO sensor, WetLabs ECO-FLR Fluorometer, and Biospherical QCP2200 PAR sensor. Nutrient samples were analyzed for NH_4^+ and NO_2^- aboard the ship as described in Beman et al. 2021b.

Water samples were collected for DNA extraction and sequencing using sampling bottles deployed on the CTD rosette. At each depth, 2L samples were filtered through 0.22 µm filters (Millipore, Darmstadt, Germany) using a peristaltic pump, then filters were submerged in Sucrose-Tris-EDTA (STE) buffer in pre-prepped Lysis Matrix E tubes and frozen at -80°C until extraction. DNA was extracted from filters following Beman et al. (2012) and DNA samples were sent for metagenome sequencing in the Vincent J. Coates Genome Sequencing Laboratory (GSL) at the University of California, Berkeley (https://genomics.qb3.berkeley.edu/), which is supported by NIH S10 OD018174 Instrumentation Grant. For each sample, 250 ng of genomic DNA was sheared and libraries were prepared using the KAPA HyperPrep Kit (Kapa Biosystems, Wilmington, MA, USA). Samples were pooled into a single lane and sequenced via 150-cycle paired-end sequencing on the Illumina HiSeq 4000 platform (Illumina, Inc., San Diego, CA, USA). Data were demultiplexed by the GSL and reads were filtered and trimmed using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) with the following parameters: maq=8, maxns=1, minlen=40, minlenfraction=0.6, k=23, hdist=1, trimq=12, qtrim=rl. Forward and reverse reads were then merged using PANDASeq (https://github.com/neufeld/pandaseq) with default parameters.

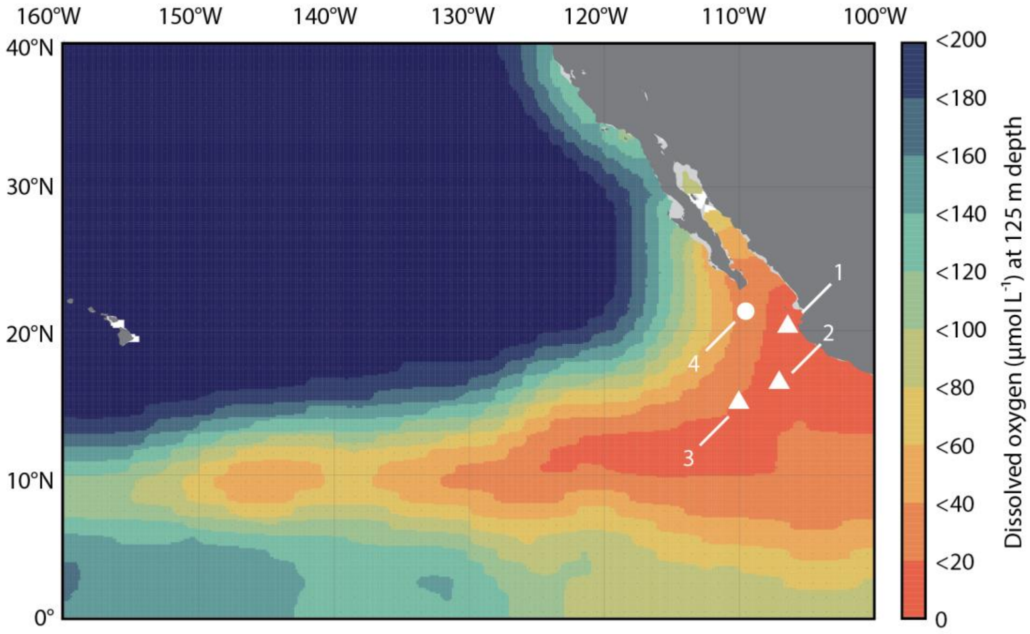


Figure 2.1: Location of sampling stations in the ETNP. Metagenomes were sequenced from key sampling depths (see text) at Stations 1-4. Stations represented by a white triangle are AMZ stations, while station 4 represented by a white circle is an OMZ station. Dissolved oxygen (DO) concentration at 125 m in the ETNP OMZ is measured in umol L and represented by a color scale on the vertical axis, where red is low DO and blue is high DO.

Overview of viral identification and annotation

We used four overall approaches to investigate viruses in metagenomic sequences sampled from the ETNP OMZ (Fig. 2.2). First, we used unassembled reads to characterize viral taxonomy using Genomic Origin Through Taxonomic CHAllenge (GOTTCHA2) (Freitas et al. 2015) implemented in Kbase (Chivian et al. 2023). We then used two approaches to provide an overview of viral AMGs and to examine patterns in specific AMGs across samples. The first of these two approaches was detecting viral AMGs using Virus Identification By iteRative ANnoTation (VIBRANT) (Kieft et al. 2020) with contigs assembled using MEGAHIT (Li et al. 2015). We used a parallel approach in Kbase where metagenomic reads were assembled into contigs using metaSPAdes (Nurk et al. 2017), and then VirSorter 2 (Guo et al. 2021) was used to detect viral contigs; viral contigs were then analyzed and annotated with Prokka (Seemann 2014), and detected for AMGs using Distilled and Refined Annotation of Metabolism-Viral (DRAM-V) (Shaffer et al. 2020; Chivian et al. 2023). Finally, a custom BLAST database was constructed to identify the AMGs in viral clusters generated by vConTACT2 (Bin Jang et al. 2019).

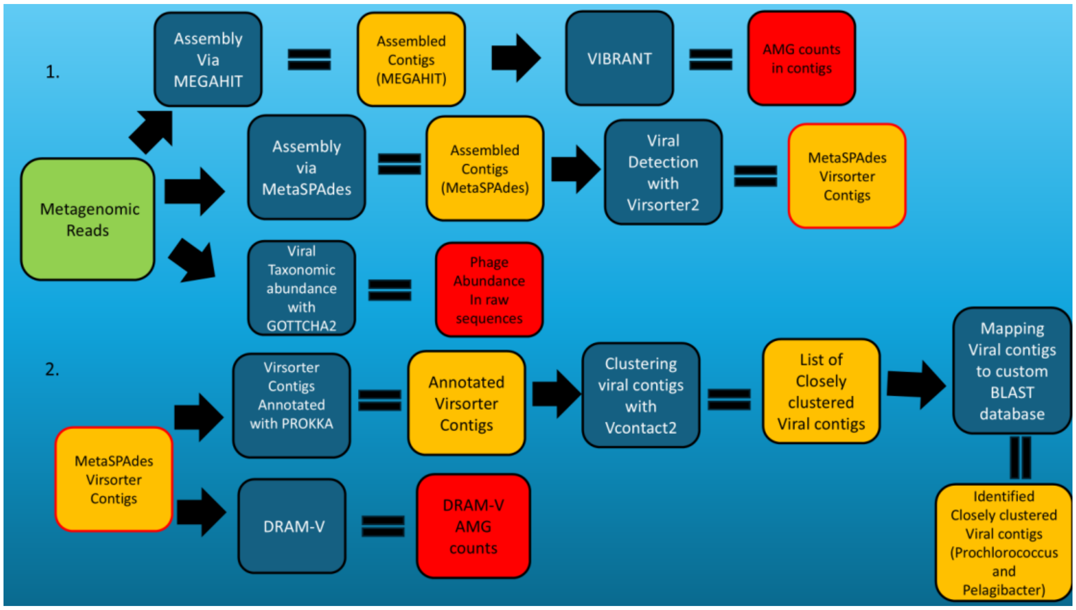

1.
Assembly Via MEGAHIT
Assembled Contigs (MEGAHIT)
VIBRANT
AMG counts in contigs
Metagenomic Reads
Assembly via MetaSPAdes
Assembled Contigs (MetaSPAdes)
Viral Detection with Virsorter2
MetaSPAdes Virsorter Contigs
Viral Taxonomic abundance with GOTTCHA2
Phage Abundance In raw sequences
2.
MetaSPAdes Virsorter Contigs
Virsorter Contigs Annotated with PROKKA
Annotated Virsorter Contigs
Clustering viral contigs with Vcontact2
List of Closely clustered Viral contigs
Mapping Viral contigs to custom BLAST database
DRAM-V
DRAM-V AMG counts
Identified Closely clustered Viral contigs (Prochlorococcus and Pelagibacter)

Figure 2.2: Schematic overview of metagenomic methods used to detect and examine viral communities and functionality in the ETNP OMZ, including VIBRANT, GOTTCHA2, DRAM-V, and VirSorter2. The top half of the schematic shows processing of metagenomic reads into contigs assembled by MEGAHIT or metaSPAdes, which were then input into VIBRANT and VirSorter2, as well as taxonomic analysis by GOTTCHA2. Bottom half shows the processing of virus sequences identified via VirSorter via DRAM-V and additional analyses.

Viral classification via GOTTCHA2

Viral identification and taxonomic classification were conducted using GOTTCHA2 (Freitas et al. 2015), which accepts shotgun metagenomic reads and provides taxonomic information for these reads. GOTTCHA2 contains two reference databases and the database that was selected for viral taxonomic classification was the RefSeq Bacteria and Viruses database (Chivian et al. 2023). Additional parameters for GOTTCHA2 included a minimum coverage of 0.004, minimum reads of 1, minimum length 40, and lastly minimum Z score of 10. GOTTCHA2 was used to produce a taxonomic abundance table with the relative abundance of viruses in each sample, including taxonomic name, level, relative abundance fraction, and depth of coverage statistics. Relative abundance was calculated by rollup depth of coverage, where depth of coverage is the number of base pairs mapped by linear length; relative abundance is rollup depth of coverage value divided by sum of abundance of given level. GOTTCHA2 also produced a dendrogram of all the organisms and higher order taxa into a phylogenetic tree, and a KRONA interactive plot showing organism content and relative abundance (Freitas et al. 2015). The taxonomic abundance files were manually filtered for viral abundance within each sample. The viral abundance from each sample was created into individual taxonomic abundance files, which were then merged into a new data frame of all the viral abundances across all samples.

Viral AMG detection using Vibrant

Viral AMGs and viral sequences were detected in the ETNP OMZ sample using VIBRANT (Kieft et al. 2020), which highlights AMGs that are used by viruses and assigns them to KEGG metabolic pathways. VIBRANT then determines which viruses are considered circular or linear using annotation methods and outputs all the data into a user-friendly folder highlighting viral and AMG identification in each assembled sample along with visualization (Kieft et al. 2020). For VIBRANT, metagenomic contigs were assembled via MEGAHIT v 1.1.3 540 (Li et al. 2015) using an initial k-mer size of 23. The assembled contigs for each sample were input into the VIBRANT workflow with default parameters.

Viral identification with VirSorter2, annotation with Prokka, and Viral AMG detection using DRAM-V

Our other main approach for viral identification and AMG annotation used VirSorter2 combined with Prokka and DRAM-V and using metaSPAdes assembled contigs. QCed fastq paired end reads for each ETNP sample were checked for read quality using FASTQC and then inputted into the genomic read assembler tool metaSPAdes (Nurk et al. 2017). Parameters for metaSPAdes to assemble contigs were set with a minimal contig length of 2000, and with kmer sizes of 21, 22, 51, 77, 99, and 127. After metaSPAdes assembly, VirSorter2 was used to detect both DNA and RNA viral genomes in our assembled dataset (Guo et al. 2021). Input sequences in VirSorter2 were annotated, and relevant features were extracted. Each sequence was classified and given a score based on their viral identity, and these scores were aggregated into a single prediction. Parameters were set for VirSorter2 to get the highest possible quality of identified viral sequences: for the adjusted parameters, the minimum score set to identify something as viral was at 0.5, the minimum sequence length was set to 3000, and the viral groups included to be identified were dsDNA and ssDNA phages. To only output high confidence viral sequences, we set a max score greater than or equal to 0.9 or with a max score greater than or equal to 0.7 with a hallmark of 1. Lastly, VirSorter2 generates a shock ID that was used for DRAM-V.

Following viral identification with VirSorter2, identified viral contigs were input into the Prokaryotic and viral annotation pipeline Prokka (Seemann 2014) to determine and label features in these genomic DNA sequences. Prokka uses outside or external feature prediction tools to identify the genomic features within contigs in two stages; Prokka predicts what genes code for in a hierarchical manner by moving the genomic DNA sequences from smaller trustworthy databases to larger databases. Finally, these sequences are curated with models of protein families. Prokka was used with default settings, except for the kingdom setting, which was set to viruses.

The assembled, sorted, and annotated viral contigs were further analyzed using the DRAM-V tool in kbase (Shaffer et al. 2020; Chivian et al. 2023), which will predict auxiliary metabolic genes (AMGs). DRAM-V accepts the ETNP viral contigs along with the shock ID that was produced from VirSorter2. The parameters for DRAM-V included a minimum contig length of 3500, Translation table of 11, Bit score threshold of 60, and a reverse search bit score threshold of 350. We used DRAM-V to produce an AMG summary file that includes relevant statistics on potential AMG genes including the number of categories (category of metabolism), header (subcategories or pathways), and AMGs in the VirSorter2 assembled and annotated contigs. The category, header, and AMG counts were pulled from the AMG summary file for each sample and an individual file was created for each group.

Viral Contigs identified using vConTACT2

The assembled, sorted, and annotated viral contigs were also input into Viral CONTigs Automatic Clustering and Taxonomy(vConTACT2) to generate viral clusters (VCs) of known taxonomy and function (Bin Jang et al. 2019). vConTACT2 uses genes shared among genomes (viral contigs), to cluster genomes together using a network approach. For parameters, ProkaryoticViralRefSeq201 was selected as the reference database (Brister et al. 2015), the protein clustering method was done with MCL, the viral clustering method was done with ClusterONE, and the minimum significance was done with less than or equal to 1 (Nepusz et al. 2012). The vConTACT2 program generates a gene2genome table for each sample that was used to identify viral contigs and determine which viral genomes from the ViralRefSEq201 database they cluster with.

vConTACT2 produces a network file for each sample that can be visualized using Cytoscape (Shannon et al. 2003). In Cytoscape, the network file displays clustered groupings of all the annotated VirSorter2 contigs and all the phages from the ViralRefSeq201 database (Fig. 2.3). We specifically analyzed network files from samples in the SCM and SNM regions at station 2 and 3, and manually searched for contigs that generated VCs with *Prochlorococcus* phages in the SCM samples and *Pelagibacter* phages in the SNM samples. Once we identified these VCs, we checked if these contigs were detected with AMGs through DRAM-V. The AMG protein sequences of these contigs were manually pulled from the FAA files produced by DRAM-V and then aligned to a custom GenomeSet using BLASTp (Altschul et al. 1990; Altschul et al. 1997; Camacho et al. 2009).

The constructed GenomeSet comprised of all public genomes of phages that were identified by GOTTCHA2, except for Ostreococcus tauri virus 2, as this genome did not appear in the Kbase Database (Chivian et al. 2023). AMG protein sequences found within the viral contigs that formed VCs with *Prochlorococcus* phage in the SCM and *Pelagibacter* phage in the SNM were extracted from FAA files produced by DRAM-V and aligned to the custom GenomeSet with BLASTp. After each gene went through the alignment process, the alignment results retained were those with the highest percent identity and lowest e values. Additionally, the Prokka output of these contigs that were clustered with *Prochlorococcus* phage and *Pelagibacter* phage were inspected for unique or interesting AMGs.

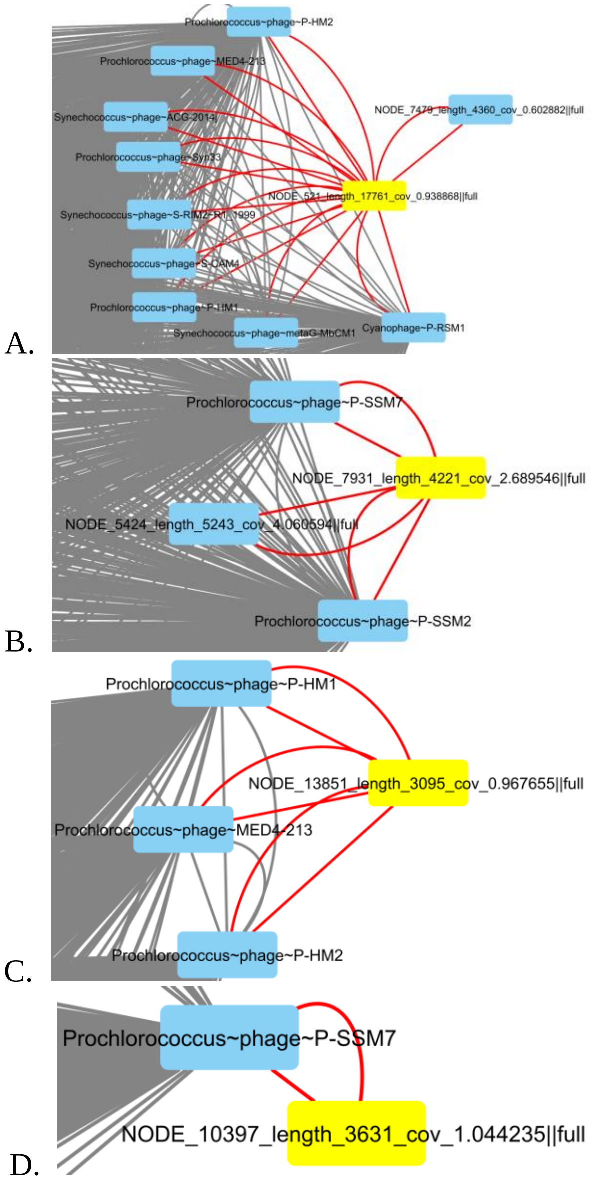

A.
Prochlorococcus~phage~P-HM2
Prochlorococcus~phage~MED4-213
NODE_7479_length_4360_cov_0.602882||full
Synechococcus~phage~ACG-2014j
Prochlorococcus~phage~Syn33
NODE_521_length_17761_cov_0.938868||full
Synechococcus~phage~S-RIM2~R1_1999
Synechococcus~phage~S-CAM4
Prochlorococcus~phage~P-HM1
Synechococcus~phage~metaG-MbCM1
Cyanophage~P-RSM1
B.
Prochlorococcus~phage~P-SSM7
NODE_7931_length_4221_cov_2.689546||full
NODE_5424_length_5243_cov_4.060594||full
Prochlorococcus~phage~P-SSM2
C.
Prochlorococcus~phage~P-HM1
NODE_13851_length_3095_cov_0.967655||full
Prochlorococcus~phage~MED4-213
Prochlorococcus~phage~P-HM2
D.
Prochlorococcus~phage~P-SSM7
NODE_10397_length_3631_cov_1.044235||full

Figure 2.3: Examples of ETNP viral clusters (VCs) (specifically from Station 2 at 130 m) clustered with vConTACT2 RefSeq genomes (specifically *Prochlorococcus* phage). A) Contig

closely clustered with *Prochlorococcus* phage and other taxa (including contigs from our data). B) Contigs closely clustered with *Prochlorococcus* phage and our data. C) Contigs closely clustered with multiple *Prochlorococcus* phage genomes. D) Contig closely clustered with a singular *Prochlorococcus* phage genome.

Data Visualization

Stacked barplots were created with ggplot using the geombar function in R studio on the metabolism (category) counts detected in VIBRANT and DRAM-V, and pathway (aka header) in DRAM-V. Total barplots were created for pathways detected in both programs (Wickham 2011). Heatmaps were created on the viral abundance generated by GOTTCHA2, along with the pathway (header) and AMG counts detected in VIBRANT and DRAM-V. Heatmaps were created using the pandas and seaborn library in python (McKinney 2011) by using the clustermap function (Waskom 2021). For hierarchical clustering, the clustering method was set to 'average' and the clustering metric to 'braycurtis.' Further customization of the heatmap was carried out using the matplotlib library (Hunter 2007).

2.4 Results

Viral Taxonomy in the ETNP OMZ

Viruses identified throughout the ETNP OMZ by GOTTCHA2 were dominated (63% of all identified viruses) by bacteriophages that infect cyanobacteria, including *Prochlorococcus*, *Synechococcus*, and cyanophages (Fig. 2.4). Of these phages, *Prochlorococcus* phage were more prevalent throughout the upper water column of multiple stations and in the SCM in station 2 and 3. Taxonomic results from GOTTCHA2 show samples divided into multiple groupings that reflect the comparatively high abundance of *Prochlorococcus* phage in these samples, as well as differences in viral composition across samples. In station 2 at 67 m, both *Prochlorococcus* phage P-SSM7 and *Prochlorococcus* phage P-SSM2 had the highest relative abundance, with respective abundance values of 0.16 and 0.14. Both these phages had the second highest abundance in station 3 at 87 m, with abundance values of 0.10 (SSM2) and 0.08 (SSM7). *Synechococcus* phages were not as abundant as *Procholrococcus* phages but were more common in deeper samples (Fig. 2.4). Lastly, cyanophages were dominant in the surface samples, where Cyanophage P-RSM6 had the highest abundance of 0.10 in station 3 at 87m, while the remaining identified cyanophages have a low abundance in deeper samples.

The remaining viruses that have a high abundance include two *Pelagibacter* phage (HTVC019P and HTVC010P) with an abundance of 0.12 and 0.03 at Station 1 at 25m. *Pelagibacter* phage HTVC010P was present in nearly all ETNP samples, while *Pelagibacter* phage HTVC008M was abundant in the SNM sample station 2 at 160m. Station 1 at 25 also contained Marine gokushovirus with an abundance of 0.17, as well as several phages not found in or at low abundance in other samples such as *Croceibacter* phage P2559Y, *Chrysochromulina ericina* virus, and *Puniceispirillum* phage HMO-2011. Finally, *Ostreococcus* virus exhibited a high abundance in station 4 at 50m. Overall, viral taxonomy displayed multiple differences between samples collected in different key regions of the OMZ water column, with *Prochlorococcus* phage particularly dominant, and with these and other phages showing the highest relative abundance in station 3 at

87m and in station 2 at 67m. Interestingly, no phages were detected in the SNM sample of station 3 at 180m by GOTTCHA2.

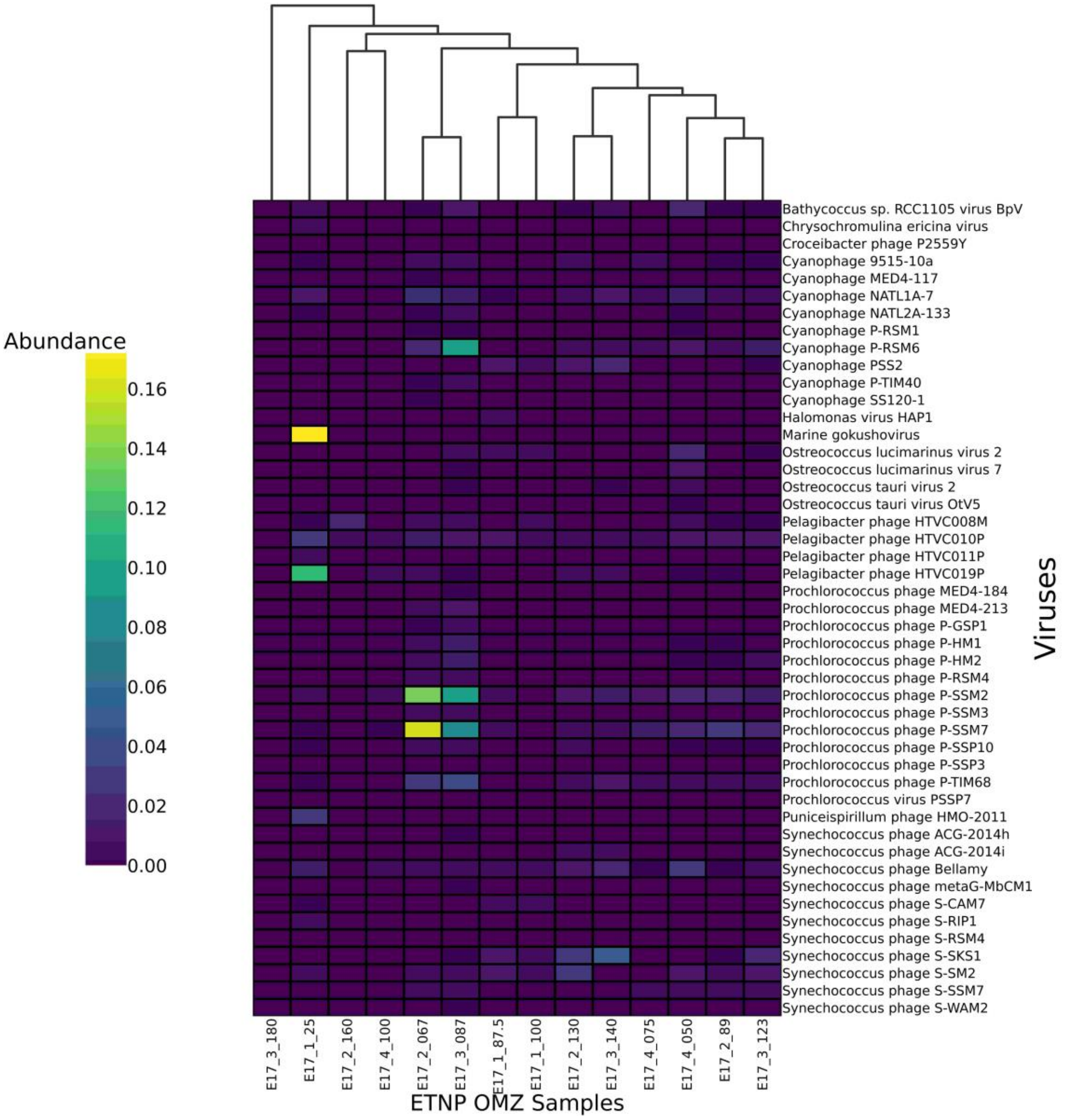


Figure 2.4: Heatmap displaying relative the abundance of different viral taxonomic groups identified within the ETNP OMZ. Abundance scale is on the left, where yellow represents high abundance and purple low abundance, and samples are listed across the bottom. Viruses identified are listed on the right side of the figure. Viruses repeatedly identified and abundant included *Synechococcus* phage, *Prochlorococcus* phage, and cyanophage. Samples were clustered by viral abundance in each sample, represented by the dendrogram at the top of the figure.

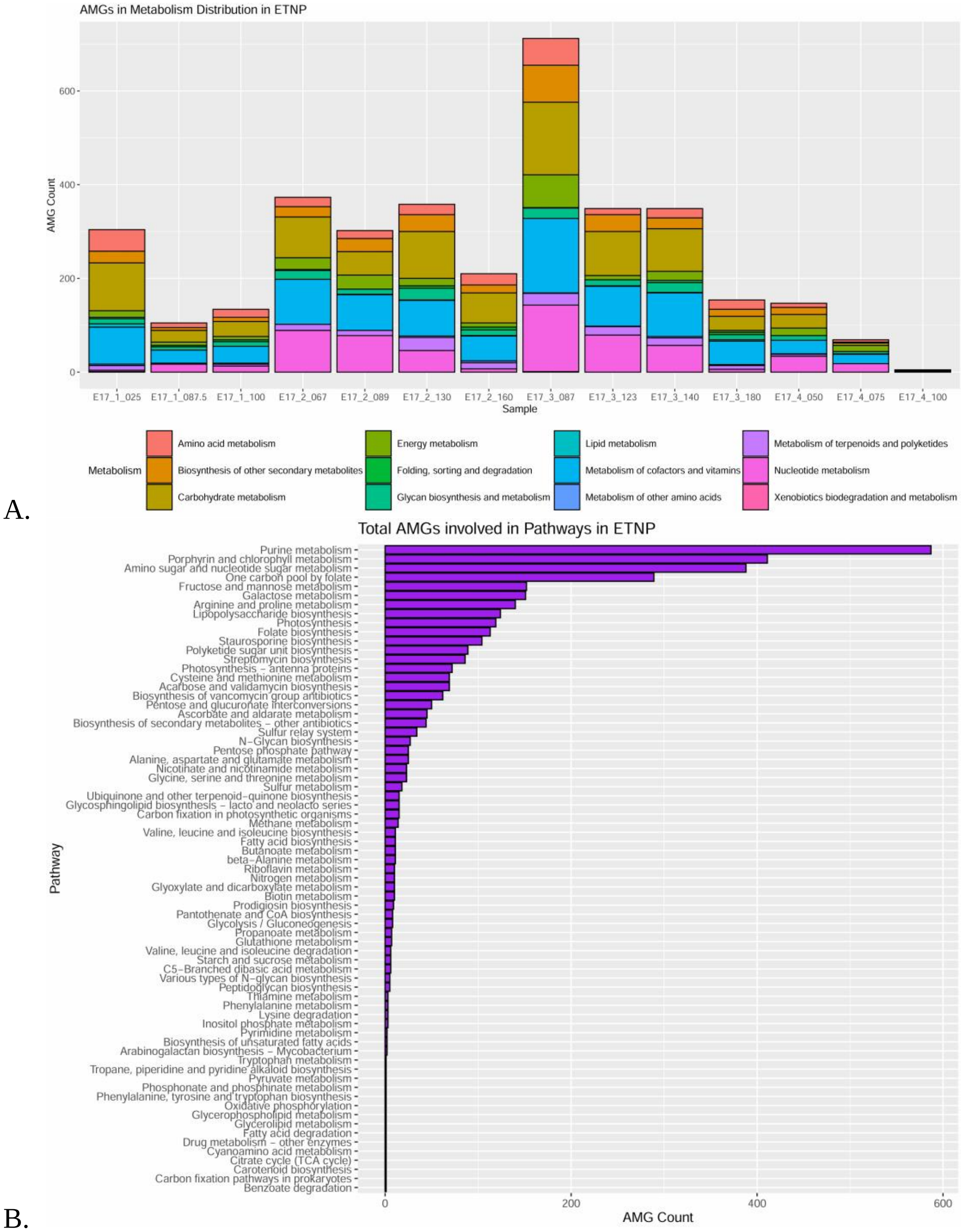
Overview of viral AMGs in the ETNP OMZ based on VIBRANT

We used VIBRANT to provide an initial overview of putative viral AMGs across our samples and then examine variations in specific AMGs throughout the ETNP OMZ

(Fig. 2.5). We found that viral AMGs fell into a total of 12 major metabolism categories, with variations across samples in terms of total numbers of putative AMGs detected and in their composition. However, hundreds of AMGs were identified in most samples. Across all samples, the metabolism with the highest AMG count was Metabolism of cofactors and vitamins with a count of 882, closely followed by Carbohydrate metabolism with a count of 864, and third was Nucleotide metabolism with an AMG count of 589. Energy metabolism was the median of all the metabolisms with an AMG count of 239. In terms of their distribution along the ETNP stations and different depth regions, samples from Stations 2 and 3 and Station 1 at 25 m had the highest AMG count compared to rest of the samples from Station 1, and especially with Station 4 as this had the lowest AMG count. Both Metabolism of cofactor and vitamins and Carbohydrate metabolism were typically present in all samples (except station 4 at 100m) and contained an appreciable count of AMGs, with both amounting to 1,746. Although Nucleotide metabolism was the third most common grouping of AMGs, it was not as dominant in Station 1 or 4—especially Station 1 at 25m—and was instead more common at Stations 2 and 3.

Within these broad metabolic categories, three pathways had particularly high AMGs counts, several moderate, and many with low counts. The pathway with the highest AMG count detected by VIBRANT was Purine metabolism with an AMG count of 587, followed by Porphyrin and chlorophyll metabolism with an AMG count of 411, and Amino sugar and nucleotide sugar metabolism with an AMG count of 388. Multiple viral AMGs also contributed to Energy metabolism; interestingly, the dominant pathway in this metabolism were AMGs contributing to autotrophic oxygenic Photosynthesis. The remaining pathways that contributed to Energy metabolism included Sulfur metabolism (AMG count of 18), Carbon fixation in photosynthetic organisms (AMG count of 15), Methane metabolism (AMG count of 15), and Nitrogen metabolism (AMG count of 10) (Fig. 2.5).

The three pathways with the highest AMG counts were found in multiple samples and contained a high count throughout much of the ETNP water column. Purine metabolism, Porphyrin and chlorophyll metabolism, and Amino sugar and nucleotide sugar metabolism all displayed a similar pattern of higher relative abundance at Stations 2 and 3 in the PNM, OMZ edge, and SCM, while in the SNM samples (station 2 at 160 and station 3 at 180m) these pathways were less dominant. Other pathways following this pattern were One carbon pool and folate and Folate biosynthesis. Although Purine metabolism had the highest AMG count, it was less common than Porphyrin and chlorophyll metabolism and Amino sugar and nucleotide sugar metabolism in the SNM samples and station 1 at 25m. Other pathways with a high count in SNM samples included Fructose and mannose metabolism, Galactose metabolism, and Lipopolysaccharide metabolism. Finally, the photosynthesis-related pathways obviously showed higher abundance in OMZ Edge and PNM samples, but were still prevalent in the SCM (Fig. 2.5).


AMGs in Metabolism Distribution in ETNP
AMG Count
0
200
400
600
E17_1_025
E17_1_087.5
E17_1_100
E17_2_067
E17_2_089
E17_2_130
E17_2_160
E17_3_087
E17_3_123
E17_3_140
E17_3_180
E17_4_050
E17_4_075
E17_4_100
Sample
Metabolism
Amino acid metabolism
Biosynthesis of other secondary metabolites
Carbohydrate metabolism
Energy metabolism
Folding, sorting and degradation
Glycan biosynthesis and metabolism
Lipid metabolism
Metabolism of cofactors and vitamins
Metabolism of other amino acids
Metabolism of terpenoids and polyketides
Nucleotide metabolism
Xenobiotics biodegradation and metabolism
A.
Total AMGs involved in Pathways in ETNP
Pathway
Purine metabolism
Porphyrin and chlorophyll metabolism
Amino sugar and nucleotide sugar metabolism
One carbon pool by folate
Fructose and mannose metabolism
Galactose metabolism
Arginine and proline metabolism
Lipopolysaccharide biosynthesis
Photosynthesis
Folate biosynthesis
Staurosporine biosynthesis
Polyketide sugar unit biosynthesis
Streptomycin biosynthesis
Photosynthesis - antenna proteins
Cysteine and methionine metabolism
Acarbose and validamycin biosynthesis
Biosynthesis of vancomycin group antibiotics
Pentose and glucuronate interconversions
Ascorbate and aldarate metabolism
Biosynthesis of secondary metabolites - other antibiotics
Sulfur relay system
N-Glycan biosynthesis
Pentose phosphate pathway
Alanine, aspartate and glutamate metabolism
Nicotinate and nicotinamide metabolism
Glycine, serine and threonine metabolism
Sulfur metabolism
Ubiquinone and other terpenoid-quinone biosynthesis
Glycosphingolipid biosynthesis - lacto and neolacto series
Carbon fixation in photosynthetic organisms
Methane metabolism
Valine, leucine and isoleucine biosynthesis
Fatty acid biosynthesis
Butanoate metabolism
beta-Alanine metabolism
Riboflavin metabolism
Nitrogen metabolism
Glyoxylate and dicarboxylate metabolism
Biotin metabolism
Prodigiosin biosynthesis
Pantothenate and CoA biosynthesis
Glycolysis / Gluconeogenesis
Propanoate metabolism
Glutathione metabolism
Valine, leucine and isoleucine degradation
Starch and sucrose metabolism
C5-Branched dibasic acid metabolism
Various types of N-glycan biosynthesis
Peptidoglycan biosynthesis
Thiamine metabolism
Phenylalanine metabolism
Lysine degradation
Inositol phosphate metabolism
Pyrimidine metabolism
Biosynthesis of unsaturated fatty acids
Arabinogalactan biosynthesis - Mycobacterium
Tryptophan metabolism
Tropane, piperidine and pyridine alkaloid biosynthesis
Pyruvate metabolism
Phosphonate and phosphinate metabolism
Phenylalanine, tyrosine and tryptophan biosynthesis
Oxidative phosphorylation
Glycerophospholipid metabolism
Glycerolipid metabolism
Fatty acid degradation
Drug metabolism - other enzymes
Cyanoamino acid metabolism
Citrate cycle (TCA cycle)
Carotenoid biosynthesis
Carbon fixation pathways in prokaryotes
Benzoate degradation
0
200
400
600
AMG Count
B.

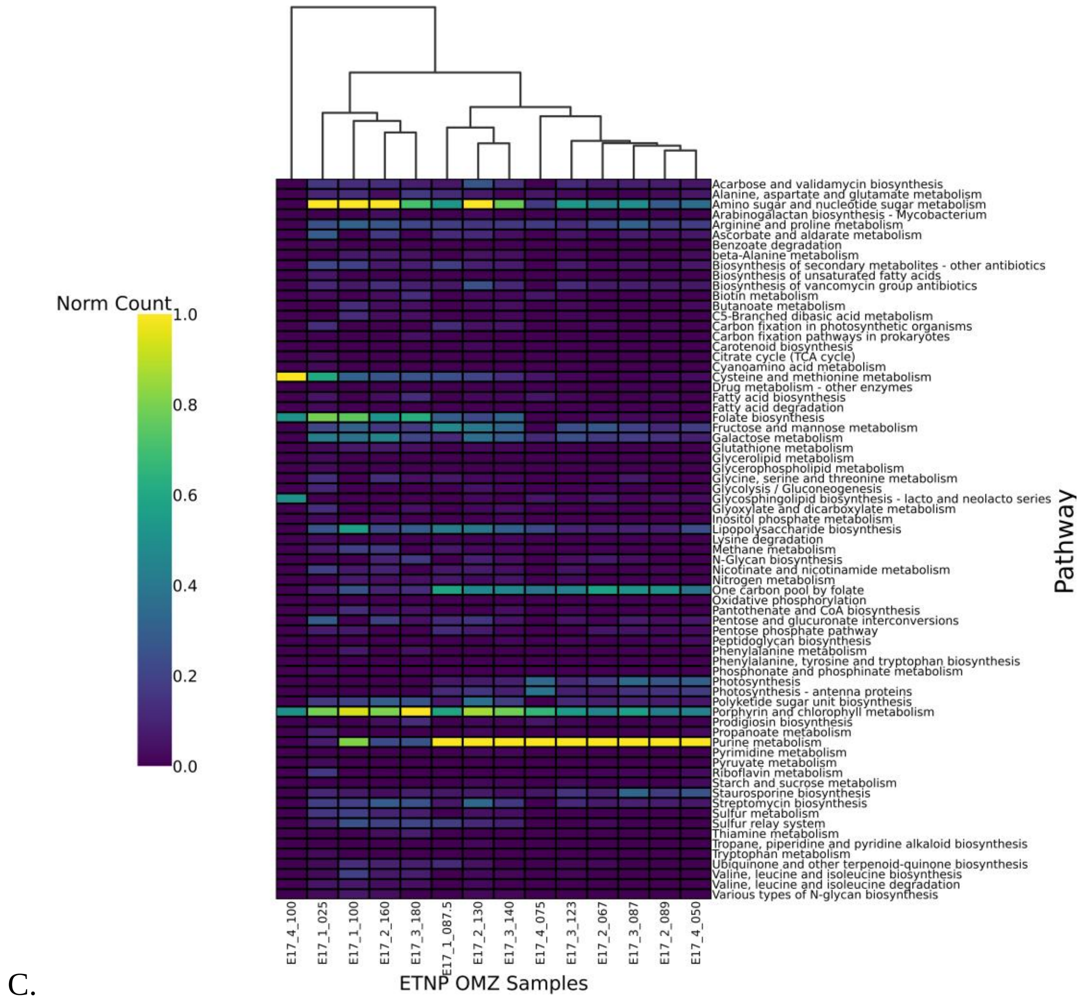


Figure 2.5: General overview of viral AMGs identified by VIBRANT in the ETNP OMZ. A) Stacked barplot of viral AMG counts within different metabolisms for each sample. Samples are listed along the bottom of the barplot and identified metabolisms are indicated by the different colors. Number of AMG counts are shown on the left side of the barplot. B) Barplot of viral AMG counts in identified pathways. Pathways are listed on the vertical axis in descending order, and AMG counts are on the horizontal axis. C) Heatmap displaying normalized counts of different viral AMG pathways. Identified pathways are listed on the right vertical axis in alphabetical order. Normalized count color scale is on the left, where yellow represents high abundance and purple low abundance, and samples are listed across the bottom. Within each sample, the pathways have been normalized with max-min normalization, where pathways are scaled from 0 to 1. Samples are listed on the bottom and are ordered by AMG pathway clustered groups which are represented by the dendrogram at the top.

Distribution of specific viral AMGs in the ETNP OMZ

These variations were driven by multiple genes of known important function towards viral replication and biogeochemical cycling (Fig. 2.6). Station 3 at 087m consistently had the highest AMG count out of all samples, and it contained the AMG with the highest count, phosphoribosyl glycinamide formyltransferase 1 (*purN*), which is involved in multiple pathways. The *purN* gene also had a relatively high AMG count across other samples—including all of station 2 and 3 in the PNM, OMZ edge, and SCM. However, the deeper SNM samples (such as Station 2 at 160m and Station 3 at 180 m) contained a lower *purN* count, as did the Station 1 and 4 samples Additional genes belonging to the phosphoribosyl (PUR) family of genes also showed high counts and similar patterns, including the *purM*, *purH*, and *purC* genes.

Several AMGs followed the distribution pattern of the PUR family of genes and also contained a high count, including: heme oxygenase (*hmuO*) and UDP-glucose 4-epimerase (*galE*). The cobaltochelatase (*cobS)* gene also showed a high count in PNM, OMZ Edge, and SCM samples, but also had a higher count in the SNM samples, and so were much more evenly distributed in the ETNP OMZ. An additional grouping of AMGs included the prolyl 4-hydroxylase (*P4HA)*, tryptophan 7-halogenase (*prnA)*, and photosystem II P680 reaction center D1 protein (*psbA)* genes, which shared a similar count pattern among the samples. All these genes contained a high count in the PNM and OMZ Edge samples of Stations 2 and 3, but had a low count in the SCM and SNM. In addition to *psbA*, other AMGs involved in photosynthesis included photosystem II P680 reaction center D2 (*psbD*) and CpeT protein (*cpeT*)—which exhibited a high count of 16 at 25m depth at Station 1. Overall, patterns for individual viral AMGs were consistent with those for identified metabolism and pathways. For example, samples clustering on the VIBRANT AMG heatmap shows that SCM, OMZ Edge, and PNM samples from Station 2 and 3 all fall within one cluster driven by these dominant genes (Fig. 2.6). Notably, most of these genes were also relatively uncommon at Station 1 at 25m.

Station 1 at 25m had a relatively low AMG count in total but was distinguished by many genes that were exclusive to this station and/or had a relatively high count compared to other samples. AMGs exclusively found in station 1 at 25m included DNA (cytosine-5)-methyltransferase 3A (*DNMT3A)*, long-chain acyl-CoA synthetase (*ACSL)*, and 3-oxoacyl-[acyl-carrier-protein] synthase II (*fabF*), while AMGs with a higher count in this sample compared to other samples included UDP-glucose 6-dehydrogenase (*UGDH*) and DNA (cytosine-5)-methyltransferase 1 (*DNMT1)*.

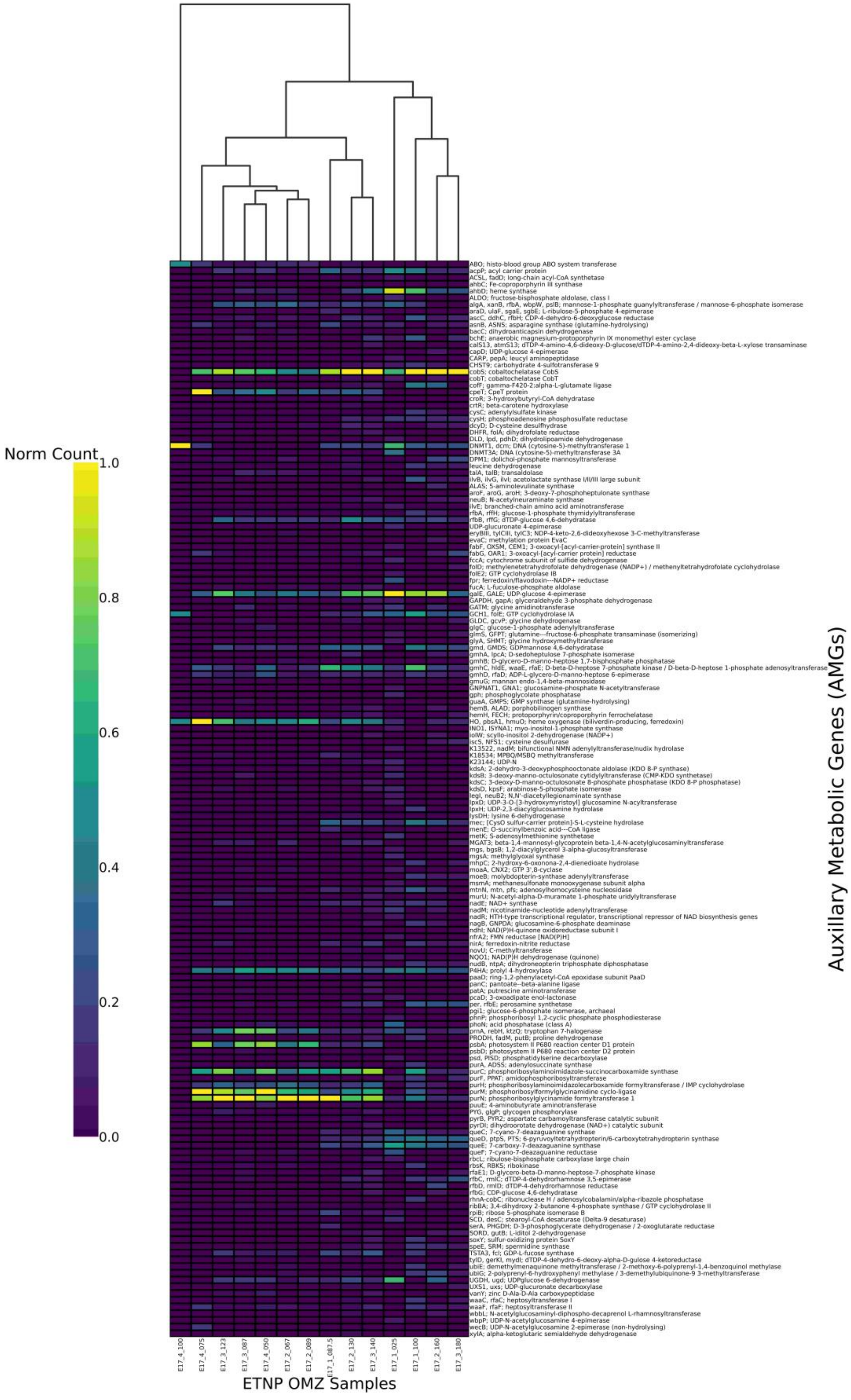

Norm Count
1.0
0.8
0.6
0.4
0.2
0.0
ABO; histo-blood group ABO system transferase
acpP; acyl carrier protein
ACSL, fadD; long-chain acyl-CoA synthetase
ahbC; Fe-coproporphyrin III synthase
ahbD; heme synthase
ALDO; fructose-bisphosphate aldolase, class I
algA, xanB, rfbA, wbpW, pslB; mannose-1-phosphate guanylyltransferase / mannose-6-phosphate isomerase
araD, ulaF, sgaE, sgbE; L-ribulose-5-phosphate 4-epimerase
ascC, ddhC, rfbH; CDP-4-dehydro-6-deoxyglucose reductase
asnB, ASNS; asparagine synthase (glutamine-hydrolysing)
bacC; dihydroanticapsin dehydrogenase
bchE; anaerobic magnesium-protoporphyrin IX monomethyl ester cyclase
calS13, atmS13; dTDP-4-amino-4,6-dideoxy-D-glucose/dTDP-4-amino-2,4-dideoxy-beta-L-xylose transaminase
capD; UDP-glucose 4-epimerase
CARP, pepA; leucyl aminopeptidase
CHST9; carbohydrate 4-sulfotransferase 9
cobS; cobaltochelatase CobS
cobT; cobaltochelatase CobT
cofF; gamma-F420-2:alpha-L-glutamate ligase
cpeT; CpeT protein
croR; 3-hydroxybutyryl-CoA dehydratase
crtR; beta-carotene hydroxylase
cysC; adenylylsulfate kinase
cysH; phosphoadenosine phosphosulfate reductase
dcyD; D-cysteine desulfhydrase
DHFR, folA; dihydrofolate reductase
DLD, lpd, pdhD; dihydrolipoamide dehydrogenase
DNMT1, dcm; DNA (cytosine-5)-methyltransferase 1
DNMT3A; DNA (cytosine-5)-methyltransferase 3A
DPM1; dolichol-phosphate mannosyltransferase
leucine dehydrogenase
talA, talB; transaldolase
ilvB, ilvG, ilvI; acetolactate synthase I/II/III large subunit
ALAS; 5-aminolevulinate synthase
aroF, aroG, aroH; 3-deoxy-7-phosphoheptulonate synthase
neuB; N-acetylneuraminate synthase
ilvE; branched-chain amino acid aminotransferase
rfbA, rffH; glucose-1-phosphate thymidylyltransferase
rfbB, rffG; dTDP-glucose 4,6-dehydratase
UDP-glucuronate 4-epimerase
eryBIII, tylCIII, tylC3; NDP-4-keto-2,6-dideoxyhexose 3-C-methyltransferase
evaC; methylation protein EvaC
fabF, OXSM, CEM1; 3-oxoacyl-[acyl-carrier-protein] synthase II
fabG, OAR1; 3-oxoacyl-[acyl-carrier protein] reductase
fccA; cytochrome subunit of sulfide dehydrogenase
folD; methylenetetrahydrofolate dehydrogenase (NADP+) / methenyltetrahydrofolate cyclohydrolase
folE2; GTP cyclohydrolase IB
fpr; ferredoxin/flavodoxin---NADP+ reductase
fucA; L-fuculose-phosphate aldolase
galE, GALE; UDP-glucose 4-epimerase
GAPDH, gapA; glyceraldehyde 3-phosphate dehydrogenase
GATM; glycine amidinotransferase
GCH1, folE; GTP cyclohydrolase IA
GLDC, gcvP; glycine dehydrogenase
glgC; glucose-1-phosphate adenylyltransferase
glmS, GFPT; glutamine---fructose-6-phosphate transaminase (isomerizing)
glyA, SHMT; glycine hydroxymethyltransferase
gmd, GMDS; GDPmannose 4,6-dehydratase
gmhA, lpcA; D-sedoheptulose 7-phosphate isomerase
gmhB; D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase
gmhC, hldE, waaE, rfaE; D-beta-D-heptose 7-phosphate kinase / D-beta-D-heptose 1-phosphate adenosyltransferase
gmhD, rfaD; ADP-L-glycero-D-manno-heptose 6-epimerase
gmuG; mannan endo-1,4-beta-mannosidase
GNPNAT1, GNA1; glucosamine-phosphate N-acetyltransferase
gph; phosphoglycolate phosphatase
guaA, GMPS; GMP synthase (glutamine-hydrolysing)
hemB, ALAD; porphobilinogen synthase
hemH, FECH; protoporphyrin/coproporphyrin ferrochelatase
HO, pbsA1, hmuO; heme oxygenase (biliverdin-producing, ferredoxin)
INO1, ISYNA1; myo-inositol-1-phosphate synthase
iolW; scyllo-inositol 2-dehydrogenase (NADP+)
iscS, NFS1; cysteine desulfurase
K13522, nadM; bifunctional NMN adenylyltransferase/nudix hydrolase
K18534; MPBQ/MSBQ methyltransferase
K23144; UDP-N
kdsA; 2-dehydro-3-deoxyphosphooctonate aldolase (KDO 8-P synthase)
kdsB; 3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase)
kdsC; 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase (KDO 8-P phosphatase)
kdsD, kpsF; arabinose-5-phosphate isomerase
legI, neuB2; N,N'-diacetyllegionaminate synthase
lpxD; UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase
lpxH; UDP-2,3-diacylglucosamine hydrolase
lysDH; lysine 6-dehydrogenase
mec; [CysO sulfur-carrier protein]-S-L-cysteine hydrolase
menE; O-succinylbenzoic acid---CoA ligase
metK; S-adenosylmethionine synthetase
MGAT3; beta-1,4-mannosyl-glycoprotein beta-1,4-N-acetylglucosaminyltransferase
mgs, bgsB; 1,2-diacylglycerol 3-alpha-glucosyltransferase
mgsA; methylglyoxal synthase
mhpC; 2-hydroxy-6-oxonona-2,4-dienedioate hydrolase
moaA, CNX2; GTP 3',8-cyclase
moeB; molybdopterin-synthase adenylyltransferase
msmA; methanesulfonate monooxygenase subunit alpha
mtnN, mtn, pfs; adenosylhomocysteine nucleosidase
murU; N-acetyl-alpha-D-muramate 1-phosphate uridylyltransferase
nadE; NAD+ synthase
nadM; nicotinamide-nucleotide adenylyltransferase
nadR; HTH-type transcriptional regulator, transcriptional repressor of NAD biosynthesis genes
nagB, GNPDA; glucosamine-6-phosphate deaminase
ndhI; NAD(P)H-quinone oxidoreductase subunit I
nfrA2; FMN reductase [NAD(P)H]
nirA; ferredoxin-nitrite reductase
novU; C-methyltransferase
NQO1; NAD(P)H dehydrogenase (quinone)
nudB, ntpA; dihydroneopterin triphosphate diphosphatase
P4HA; prolyl 4-hydroxylase
paaD; ring-1,2-phenylacetyl-CoA epoxidase subunit PaaD
panC; pantoate--beta-alanine ligase
patA; putrescine aminotransferase
pcaD; 3-oxoadipate enol-lactonase
per, rfbE; perosamine synthetase
pgi1; glucose-6-phosphate isomerase, archaeal
phnP; phosphoribosyl 1,2-cyclic phosphate phosphodiesterase
phoN; acid phosphatase (class A)
prnA, rebH, ktzQ; tryptophan 7-halogenase
PRODH, fadM, putB; proline dehydrogenase
psbA; photosystem II P680 reaction center D1 protein
psbD; photosystem II P680 reaction center D2 protein
psd, PISD; phosphatidylserine decarboxylase
purA, ADSS; adenylosuccinate synthase
purC; phosphoribosylaminoimidazole-succinocarboxamide synthase
purF, PPAT; amidophosphoribosyltransferase
purH; phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase
purM; phosphoribosylformylglycinamidine cyclo-ligase
purN; phosphoribosylglycinamide formyltransferase 1
puuE; 4-aminobutyrate aminotransferase
PYG, glgP; glycogen phosphorylase
pyrB, PYR2; aspartate carbamoyltransferase catalytic subunit
pyrDI; dihydroorotate dehydrogenase (NAD+) catalytic subunit
queC; 7-cyano-7-deazaguanine synthase
queD, ptpS, PTS; 6-pyruvoyltetrahydropterin/6-carboxytetrahydropterin synthase
queE; 7-carboxy-7-deazaguanine synthase
queF; 7-cyano-7-deazaguanine reductase
rbcL; ribulose-bisphosphate carboxylase large chain
rbsK, RBKS; ribokinase
rfaE1; D-glycero-beta-D-manno-heptose-7-phosphate kinase
rfbC, rmlC; dTDP-4-dehydrorhamnose 3,5-epimerase
rfbD, rmlD; dTDP-4-dehydrorhamnose reductase
rfbG; CDP-glucose 4,6-dehydratase
rhnA-cobC; ribonuclease H / adenosylcobalamin/alpha-ribazole phosphatase
ribBA; 3,4-dihydroxy 2-butanone 4-phosphate synthase / GTP cyclohydrolase II
rpiB; ribose 5-phosphate isomerase B
SCD, desC; stearoyl-CoA desaturase (Delta-9 desaturase)
serA, PHGDH; D-3-phosphoglycerate dehydrogenase / 2-oxoglutarate reductase
SORD, gutB; L-iditol 2-dehydrogenase
soxY; sulfur-oxidizing protein SoxY
speE, SRM; spermidine synthase
TSTA3, fcl; GDP-L-fucose synthase
tylD, gerKI, mydI; dTDP-4-dehydro-6-deoxy-alpha-D-gulose 4-ketoreductase
ubiE; demethylmenaquinone methyltransferase / 2-methoxy-6-polyprenyl-1,4-benzoquinol methylase
ubiG; 2-polyprenyl-6-hydroxyphenyl methylase / 3-demethylubiquinone-9 3-methyltransferase
UGDH, ugd; UDPglucose 6-dehydrogenase
UXS1, uxs; UDP-glucuronate decarboxylase
vanY; zinc D-Ala-D-Ala carboxypeptidase
waaC, rfaC; heptosyltransferase I
waaF, rfaF; heptosyltransferase II
wbbL; N-acetylglucosaminyl-diphospho-decaprenol L-rhamnosyltransferase
wbpP; UDP-N-acetylglucosamine 4-epimerase
wecB; UDP-N-acetylglucosamine 2-epimerase (non-hydrolysing)
xylA; alpha-ketoglutaric semialdehyde dehydrogenase
Auxillary Metabolic Genes (AMGs)
E17_4_100
E17_4_075
E17_3_123
E17_3_087
E17_4_050
E17_2_067
E17_2_089
E17_1_087.5
E17_2_130
E17_3_140
E17_1_025
E17_1_100
E17_2_160
E17_3_180
ETNP OMZ Samples

Figure 2.6: Heatmap displaying the normalized counts of viral AMGs in the ETNP OMZ detected through VIBRANT. Identified AMGs are listed on the right vertical axis in alphabetical order. Abundance Normalized count color scale is on the left, where yellow represents high abundance and purple low abundance, and samples are listed across the bottom. Within each sample, the AMG counts have been normalized with max-min normalization, where AMG counts are scaled from 0 to 1. Samples are listed on the bottom, and are ordered by AMG clustered groups which are represented by the dendrogram at the top.

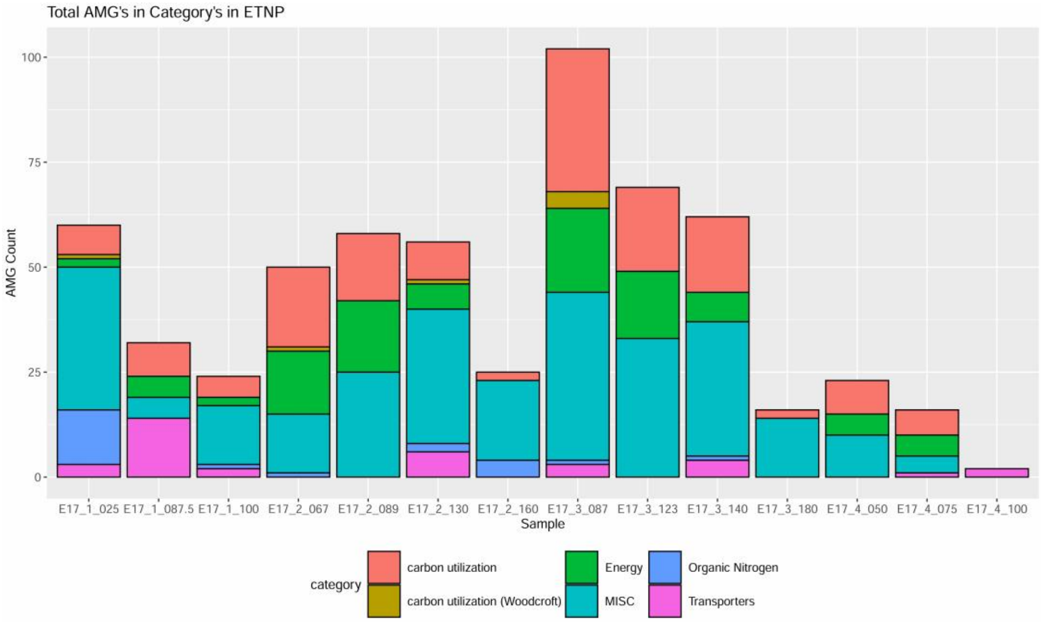
Overview of viral AMGs in ETNP OMZ based on DRAM-V

In addition to VIBRANT, we also examined putative viral AMGs using DRAM-V to give a complete representation of viral genes that may be present in the ETNP OMZ. DRAM-V detected viral AMGs belonging to 6 major categories of metabolism that included Carbon utilization, Carbon utilization (woodcroft), Energy, ‘Miscellaneous’, Organic Nitrogen, and Transporters. The ‘Miscellaneous’ (MISC) category in DRAM-V includes Information Systems (processes involved in biosynthesis, DNA synthesis, and transcription), antibiotic resistance, and CRISPR. The vast majority of the AMGs and pathways in the ETNP samples that fell within MISC are Information Systems. Out of all the DRAM-V categories, MISC contained the most AMGs with a count of 276, followed by Carbon Utilization with a count of 154, Energy with a count of 100, Transporters with a count of 35, Organic Nitrogen with a count of 23, and Carbon utilization (woodcroft) with a count of 7. The sample with the most AMGs was Station 3 at 87m, and the sample with the lowest number of AMGs was Station 4 at 100m (and which only contained one category, which was transporters; Fig. 2.7).

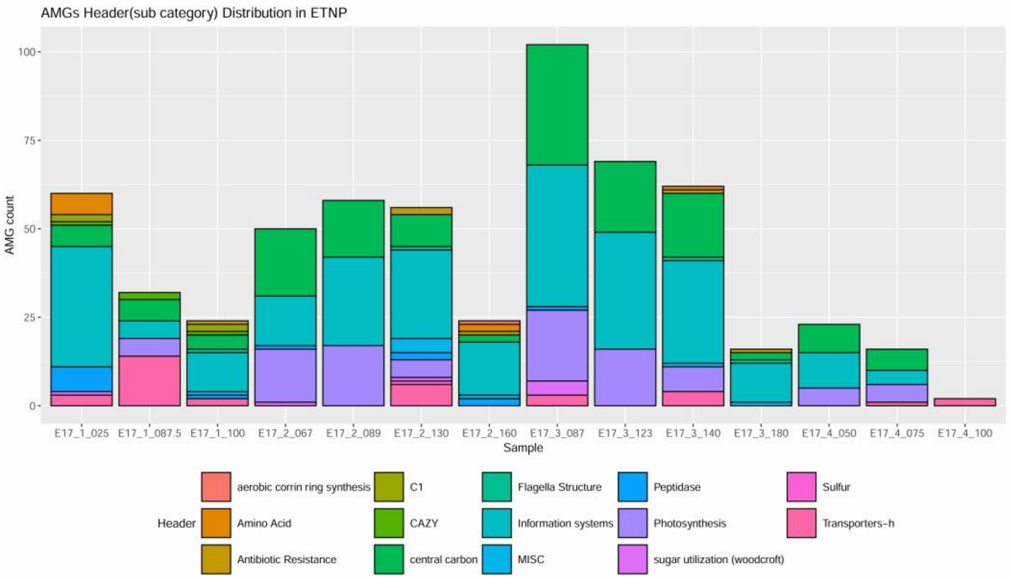
The category of metabolism MISC, Carbon utilization, and Energy were the most dominant across all the ETNP OMZ samples—except in station 1 at 87m where the metabolism of Transporters had a higher AMG count. In addition, Carbon utilization and Energy contained higher AMG counts in the OMZ edge and the PNM in station 2 and 3. However, transporters, Organic nitrogen, and Carbon utilization were also present in only half of the samples, and no category of metabolism was present in all samples. Interestingly, AMGs that contribute to Organic Nitrogen were high in station 1 at 25m the PNM/OMZ edge and in the SNM sample station 2 at 160m. Although the total known AMG counts detected by DRAM-V was much lower compared to the number of AMGs detected by VIBRANT, the AMG distribution in DRAM-V and VIBRANT were similar, as the samples retrieved from station 2, 3, and 1 at 25m have the highest AMG counts (Fig. 2.7). The disparity in AMG detection between VIBRANT and DRAM-V has been observed in other distinct environments (Busse et al. 2022, Chu et al. 2022), indicating a common underlying cause. Notably, VIBRANT operates with assembled contigs that have been sorted internally, as it performs both viral sequence detection and AMG detection within the same pipeline, whereas DRAM-V operates with contigs that are first screened using VirSorter2. Different methods were also used to assemble and predict viral sequences, namely MEGAHIT and VIBRANT on one side, and metaSPAdes and VirSorter2 on the other.

Within these categories of metabolism, the subsection of categories that were identified by DRAM-V is the header (pathway) or subcategory, and a total of 14 pathways were detected by DRAM-V across all samples. There were three pathways—Information Systems, Central Carbon, and Photosynthesis—detected with the highest

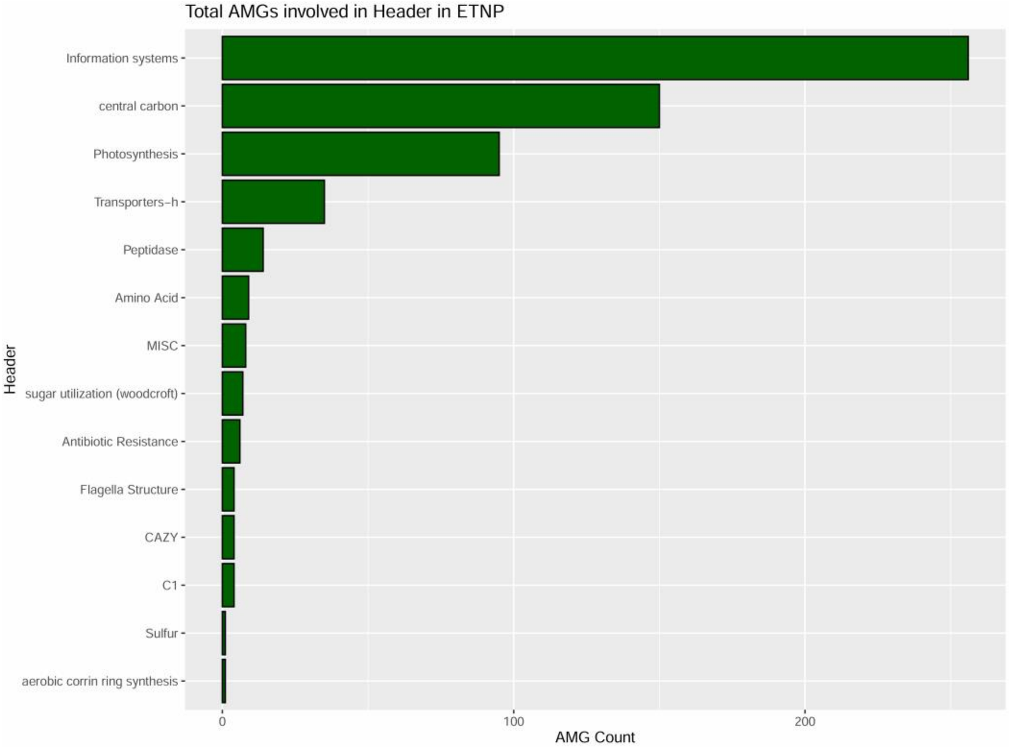
AMG counts; since the three dominant pathways contained the most AMGs, they were distributed throughout the OMZ samples. Based on AMG count alone, Information Systems was the most dominant pathway. However, there were instances where this pathway contained a lower AMG count compared to the other pathways in some samples. This was the case in the PNM sample station 2 at 67m and station 4 at 50m, for example, where Photosynthesis and Central carbon had the highest counts. Photosynthesis was well represented in most samples and had a higher count in the OMZ edge and PNM samples of station 2 and 3. Interestingly, AMGs that contribute to Photosynthesis were detected in the SCM of station 2 and 3, as well as below the SCM within the SNM. Some samples contained fewer pathways overall—including station 2 at 89m, station 2 at 123m, and all the station 4 samples—whereas station 1 at 25m surprisingly had a high AMG pathway count. A consistent pattern across both viral AMG detection programs was low AMG counts in metabolic categories and pathways for samples collected at Station 4 and within the SNM (Fig. 2.7).



A.



B.



C.

Figure 2.7: Viral metabolic categories and pathways (header) identified with DRAM-V in the ETNP OMZ. Stacked barplots of viral AMGs within different (A) metabolic categories and (B) pathways (header). In both barplots, samples are listed along the bottom and identified categories or pathways are indicated by the different colors. Number of AMG counts in a category, or pathways is shown on the left side of the barplot. C) Barplot of viral AMG counts in identified pathways (header). Pathways are listed on the vertical axis in descending order, and AMG counts are on the horizontal axis.

Specific viral AMGs in the ETNP OMZ detected by DRAM-V

DRAM-V detected a total of 54 putative viral AMGs with known function. Of these identified AMGs, the AMG with the highest count was transaldolase in the station 3 at 87m sample (Fig. 2.8). The transaldolase (*talA*) gene also had a relatively high abundance in a majority of the ETNP samples—and specifically in the PNM, OMZ Edge, and SCM of station 2 and 3, and the PCM and PNM of station 4. Additional AMGs with high abundance in the ETNP OMZ detected by DRAM-V also included phosphoribosylglycinamide formyltransferase (*purN*), phosphoribosylformylglycinamidine cyclo-ligase (*purM*), and orotate phosphoribosyltransferase (*pyeR*), which are involved in MISC and information systems. This is similar to what was detected by VIBRANT, as the phosphoribosyl (PUR) family of genes were also found across a majority of samples with high AMG counts within the water column, specifically in the PNM, OMZ Edge and SCM of station 2 and 3. DRAM-V results also indicate the PUR family of genes as being second the dominant AMG group after transaldolase (*talA*) (Fig. 2.8).

Aside from the transaldolase (*talA*) and the PUR family of genes, additional viral AMGs of interest included those involved in energy and photosynthesis. AMGs involved in this metabolic category and pathway with a relatively high count included photosystem II reaction center D1 protein (*psbA),* plastocyanin (*petE*), and ferredoxin (*petF*). Ferredoxin (*petF*) had the highest count out of these three genes, and was mainly found in the PNM, OMZ edge and SCM of station 2 and 3. *psbA* and plastocyanin (*petE*) had a lower count and were primarily found in the PNM and OMZ Edge. However, a clear difference in the AMGs that were detected by DRAM-V versus VIBRANT was that in DRAM-V, cobaltochelatase (*cobS*) genes were much rarer and far less prevalent in the ETNP OMZ, with the highest count in the SNM station 2 at 160 m. Another AMG of interest was small subunit ribosomal protein S21, which showed a different pattern compared to the previous AMGs detected by DRAM-V, as it has a high abundance in deep samples, specifically in the SNM of station 2 and 3.

There were also some AMGs that were exclusive to specific samples, including the maltose/maltodextrin transport system ATP-binding protein gene (*msmX*) in station 1 at 87.5 and thymidylate synthase (*thyA*) in station 1 at 25m. Furthermore, most of the viral AMGs in station 1 at 25m had a higher abundance in this sample relative to the other samples, or were only exclusive to station 1 at 25 m. This included AMGs such as the xyloglucosyltransferase (*GH16*) gene and the ABC-2 type transport system permease protein (*ABC-2.P*). Essentially, the pattern of AMG counts for both DRAM-V and VIBRANT was similar for station 1 at 25m, as it was unique compared to the other samples in station 1 with the higher count and contained AMGs unique to this sample.

Finally, these differences across samples led to consistent patterns in sample clustering in the ETNP OMZ. In particular, the sample from station 4 at 100 m was distinct from the other samples—likely due to the uniqueness of station 4 at 100m, as it has the lowest total AMG count of all the ETNP samples (Fig. 2.8)—while other samples typically clustered based on their position within the water column. SNM samples clustered together (to the right side of both Fig. 2.6 and 2.8), as did SCM samples.

Samples from the PNM and OMZ edge also clustered together, with the exception of the Station 1 25 m sample, which was more similar to but distinct from SNM samples.

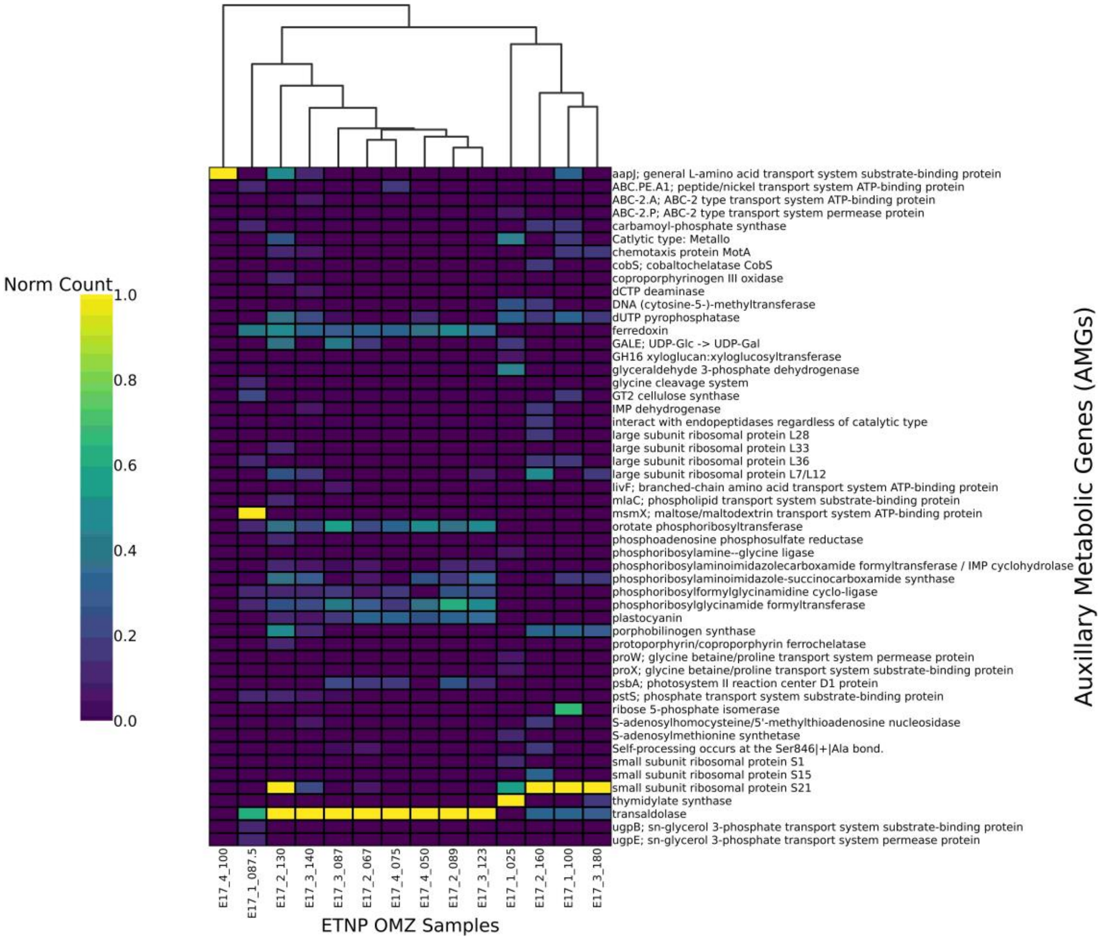


Figure 2.8: Heatmap displaying viral AMG normalized count in the ETNP OMZ detected through DRAM-V. Identified AMGs are listed on the right vertical axis in alphabetical order. Normalized count scale is on the left, where yellow represents high abundance and purple low abundance, and samples are listed across the bottom. Within each sample, the AMG counts have been normalized with max-min normalization, where AMG counts are scaled from 0 to 1. Samples are listed on the bottom and are ordered by AMG clustered groups which are represented by the dendrogram at the top.

Linking viral AMGs and virus host taxonomy

Our results demonstrate consistent patterns in viral taxonomy and AMG composition between different biogeochemical regimes in the ETNP OMZ. In particular, we found that viruses inhabiting the pivotal SCM and SNM depths likely infect the dominant and biogeochemically important organisms found there—specifically low-light adapted *Prochlorococcus* in the SCM (Beman et al. 2021b) and *Pelagibacter* in the SNM (which may have the capability for nitrate reduction; Tsementzi et al. 2016). We used several approaches to determine which AMGs were present on contigs recovered from these phages within the SCM and SNM (Table 2.1; Table 2.2).

We used vConTACT2 to generate viral clusters (VCs) via a multi-gene networking approach (Bin Jang et al. 2019), and specifically focused on *Prochlorococcus* phage VCs in the SCM and *Pelagibacter* phage VCs in the SNM. Across SCM samples, between 54-144 *Prochlorococcus* phage or *Prochlorococcus* phage + other taxa phage VCs were detected via this approach (Table 2.1). Most of the VCs clustered with *Prochlorococcus* phage as well as those infecting other taxa, and so we conservatively focused on those contigs that formed VCs only with *Prochlorococcus* phage. We used blastp (see Methods) to identify AMGs present on these *Prochlorococcus* phage contigs and assign function. Five putative AMGs were located on multiple *Prochlorococcus* phage VC contigs across multiple samples. These viral AMGs included plastocyanin (*petE*), ferredoxin (*petF*), phosphoribosylaminoimidazolecarboxamide formyltransferase (*purH*), and phosphoribosylaminoimidazole-succinocarboxamide synthase (*purC*) (Table 2.2). These genes are involved in information systems, transporters, and especially photosynthesis. In other words, *Prochlorococcus* phage present in the biogeochemically-important SCM of the ETNP OMZ—where they presumably actively infect *Prochlorococcus* cells—contain genes involved in photosynthesis.

Table 2.1: Number of *Prochlorococcus* phages and *Pelagibacter* phages identified by vConTACT2 in the SNM and SCM samples from station 1, 2 and 3

Station	Depth (m)	Region	# Prochlorococcus/ Pelagibacter phage taxa only	# Prochlorococcus/ Pelagibacter phage with our samples	# Prochlorococcus/ Pelagibacter phage with other taxa
1	87.5	SCM	1	1	52
2	130	SCM	10	4	110
3	140	SCM	4	14	126
2	160	SNM	2	4	28
3	180	SNM	1	25	20

We applied the same overall approach to *Pelagibacter* phage in the SNM and found fewer overall VCs (including none recovered from the station 1 100 m SNM sample). The SNM sample at station 2 at 160m had a higher viral contig count that reflects VCs with *Pelagibacter* phage and other taxa, similar to the SCM sample pattern in contig gene clustering. However, the last SNM sample from Station 3 at 180m contained more VCs that clustered with *Pelagibacter* phage and our data. AMGs detected by DRAM-V found within the SNM sample contigs included small subunit ribosomal protein S21, UDP-glucose family protein, and putative lysozyme (Table 2.2).

Table 2.2: AMGs in *Prochlorococcus* phage and *Pelagibacter* phage contigs in the SNM and SCM samples from station 2 and 3.

Station	Depth (m)	Region	Contig	Gene ID	Gene description	Category	Header (Pathway)
2	130	SCM	NODE_13830_length_3098_cov_0.811511||full_7	K02638	plastocyanin	Energy	Photosynthesis
2	130	SCM	NODE_9863_length_3738_cov_1.202714||full_3	K02639	ferredoxin	Energy	Photosynthesis
3	140	SCM	NODE_6579_length_4328_cov_1.035706||full_3	K00602	phosphoribosylaminoimidazolecarboxamide formyltransferase / IMP cyclohydrolase [EC:2.1.2.3 3.5.4.10] [RN:R04560 R01127]	MISC	Information systems
3	140	SCM	NODE_8351_length_3792_cov_1.056753||full_3	K02639	ferredoxin	Energy	Photosynthesis
3	140	SCM	NODE_696_length_14499_cov_1.334122||full_12	K01923	phosphoribosylaminoimidazole-succinocarboxamide synthase [EC:6.3.2.6] [RN:R04591]	MISC	Information systems
2	160	SNM	NODE_3431_length_6274_cov_0.922401||full_12	K02970	small subunit ribosomal protein S21	MISC	Information systems
2	160	SNM	NODE_9581_length_3623_cov_1.280892||full_2	YP_007518028.1	YP_007518028.1 NAD-binding, UDP-glucose/GDP-mannose dehydrogenase family protein [Pelagibacter phage HTVC008M]		
3	180	SNM	NODE_4286_length_5737_cov_0.769519||full_4	K02970	small subunit ribosomal protein S21	MISC	Information systems
3	180	SNM	NODE_2697_length_7225_cov_1.135108||full_6	YP_007518053.1	YP_007518053.1 putative lysozyme [Pelagibacter phage HTVC008M]		

2.5 Discussion

Viruses are globally important for marine microbial ecology and biogeochemistry, and particularly within OMZs, where they may regulate the important microbial processes found near the surface, along oxygen gradients, and within low oxygen waters (Jurgensen et al. 2022; Cassman et al. 2012; Gazitúa et al. 2021). Sampling along oxygen gradients at multiple stations in the ocean's largest OMZ, we sequenced metagenomes and examined putative viral sequences to identify viral taxonomic groups and auxiliary metabolic genes (AMGs) in the ETNP OMZ.

Based on these overall findings, the dominant phages present in our samples include cyanophages, *Synechococcus* phage and *Prochlorococcus* phage. All these phages infect bacterial photosynthesizers that inhabit a broad range of marine environments. In our data, the distribution of the phage sequences follows the population distribution patterns of their hosts (Beman et al. 2021a), in that they were typically more common in samples collected at the base of the euphotic zone. However, we also recovered cyanobacterial-infecting phage sequences within the secondary chlorophyll maximum (SCM) present at station 1 at 87.5m, station 2 at 130m, and station 3 at 140m (Fig. 2.4). The SCM is a unique feature of AMZs/ODZs inhabited primarily by low-light and low-oxygen adapted strains of *Prochlorococcus* important for carbon and oxygen cycling, as well as coupled nitrogen cycling (Lavin et al. 2010; Goericke et al. 2000; Aldunate et al. 2022; Beman et al. 2021a). The presence of these phages is consistent with previous research in the ETNP OMZ, especially Prochlorococcus phage in the ODZ (Fuchsman et al. 2019; Fuchsman et al. 2021). Phages that infect *Prochlorococcus* include both the P-SSM2 and P-SSM4 T-4 like myoviruses, which are similar morphologically and genomically (Sullivan et al. 2005) and both were recovered in our samples. While *Prochlorococcus* phage P-SSM2 and P-SSM4 were abundant in the ETNP OMZ upper water column, P-SSM2 was more abundant in the SCM of the AMZ core compared to P-SSM4. P-SSM2 *Prochlorococcus* phage contain the transport genes coding for plastocyanin (*petE*) and ferredoxin (*petF*) and these genes can play a role in photosynthesis (Lindell et al. 2004). This indicates that *Prochlorococcu*s phages found in the SCM can withstand anoxic conditions while also indirectly contributing to photosynthesis in deeper depths of the ETNP OMZ.

We also recovered *Pelagibacter* phage viral sequences from multiple samples, and these were most prevalent in station 1 at 25 m (Fig. 2.4). The presence of these phages likely indicates that they are infecting the heterotrophic bacteria SAR11, which are highly abundant throughout the ocean (Zhao et al. 2013). There are many types of *Pelagibacter* phages, with HTVC010P considered as a major group and have been found to contain two subgroups. These phages are mainly inhabiting epipelagic waters, and they tend to decrease in abundance with a depth past 1000m but are found in different regions ranging from polar to more tropical (Du et al. 2021). Our metagenomic data from the ETNP are consistent with this pattern, as HTVC010P were uncommon in the offshore SNM samples. *Pelagibacter* phages HTVC010P and HTVC019P have also been found in productive coastal waters of the Black Sea, which has an epipelagic surface and anoxic deep layer like the ETNP OMZ (Jaiani et al. 2020). This is consistent with their abundance in the near-coastal station 1 25m sample, which was characterized by high chlorophyll and ammonium concentrations due to active upwelling and an accompanying phytoplankton bloom. However, specially adapted low-oxygen ecotypes of SAR11 are known to inhabit OMZs and have been shown to be capable of nitrate reduction (Tsementzi et al. 2016). In line with this, we recovered *Pelagibacter* phage sequences from SNM sample station 2. Interestingly, *Pelagibacter* phage HTVC008M were most prevalent in the SNM, whereas HTVC010P were more common in the PNM and OMZ edge. *Pelagibacter* phages have not been recovered from the biogeochemically-important SNM previously, and our results indicate that the *Pelagibacter* phage HTVC008M are present in this region of the OMZ water column where they could potentially affect the specialized SAR11 ecotypes that contribute to the formation of this feature.

In addition to the high abundance of *Pelagibacter* phages, Marine gokushovirus sequences were also prevalent at 25 m at Station 1 and were unique to this sample (Fig. 2.4). Gokushoviruses are lytic viruses composed of ssDNA that are part of the *Microviridae* family and have been found in a wide range of environments—such as freshwater and marine ecosystems— (Labonté and Suttle 2013; Labonté et al. 2015; Roux et al. 2012)—and were first identified from the Sargasso Sea (Angly et al. 2006). Although known to infect parasitic bacteria, they have been recently found to infect SUP05 (Roux et al. 2014), a group of sulfur oxidizing bacteria that commonly inhabit OMZs (Labonté et al. 2015). Gokushoviruses have also been shown to have higher richness in surface waters compared to anoxic and deeper depths (Labonté et al. 2015), indicating that Marine gokushoviruses may have a wide range of hosts, and consistent with their presence in the Station 1 25 m sample in the ETNP OMZ.

The last viral group with a high abundance detected in our samples is interesting in that it is not a bacteriophage, but instead a virus that infects eukaryotic phytoplankton, the *Ostreococcus* virus (Fig. 2.4). Compared to other eukaryotic phytoplankton, *Ostreococcus* cells are small (Palenik et al. 2007) and are diverse, divided into ecotypes that are adapted to low-light warm oligotrophic waters or high-light mesotrophic regions (Worden 2006, Demir-Hilton et al. 2011). Consistent with expected host distribution, we recovered *Ostreococcus* virus sequences (specifically *Ostreococcus lucimarinus* virus) primarily from station 4 at 50 m. However, some *Ostreococcus* virus sequence sequences were retrieved from station 1 at 100 m and station 3 at 123 m, which are much deeper samples and fall below the euphotic zone. This is interesting as Ostreococcus are aerobic photosynthetic organisms, however there are ecotypes of *Ostreococcus tauri* that can survive in deeper depths (Cardol et al. 2008), such as those within our samples. *Ostreococcus* viruses have not been previously detected in the ETNP OMZ, although a study conducted in the ETSP suggested that *Ostreococcus tuari* virus can possibly assist in nitrogen assimilation at the oxycline, as they contain genes for ammonia transporters (Monier et al 2017).

Across these and all other viral sequences, we recovered a wide range of putative AMGs. Based on VIBRANT, many of the AMGs found in our samples contributed to Metabolism of cofactors and vitamins, Carbohydrate metabolism, and Nucleotide metabolism—with abundant pathways within these categories including Purine metabolism, Porphyrin and chlorophyll metabolism, and Amino sugar and nucleotide sugar metabolism (Fig. 2.5). AMGs detected with DRAM-V were classified into 6 categories where the highest counts included Miscellaneous (MISC), Carbon utilization, and Energy (Fig. 2.7); within these metabolic categories, the most prevalent pathways included Information systems, Central carbon, and Photosynthesis. These broad patterns are consistent with previous viral metagenomic studies in low-oxygen regions of the ocean. For example, DRAM-V phage metabolic categories were also identified in the Eastern Tropical South Pacific (ETSP), and modules involved in Information systems (nucleotide synthesis) were steady and abundant throughout the water column, whereas modules involved in photosynthesis were abundant in the SCM (Vik et al. 2021). Our findings further solidify that viral AMGs contributing to photosynthesis can be found in low and well-lit oxygenated and anoxic waters. AMGs involved in Amino acid metabolism and Carbohydrate metabolism have been previously found in the ETNP OMZ, with Amino acid metabolism found in well mixed and low oxygenated water, while Carbohydrate metabolism was more abundant in low-oxygen waters (Jurgensen et al. 2022). Our sampling of additional features in the water column across multiple sampling stations shows a more complicated pattern, as genes involved in Carbohydrate metabolism and Amino acid metabolism were more frequently detected in the upper region and well oxygenated areas of the ETNP OMZ. However, at Station 2, Carbohydrate metabolism contained a high count in deeper depths such as the SCM and SNM, while Amino acid metabolism had a high count in the SNM.

Within these metabolisms and pathways, there was a clear pattern of AMGs with a high count involved in DNA synthesis and replication. A group of AMGs that stood out in our samples were a part of the phosphoribosyl (PUR) family of genes (Fig. 2.6; Fig. 2.8). The PUR family of genes code for purine biosynthetic enzymes and contribute to purine synthesis, which may indicate that the phages in the ETNP OMZ are performing viral DNA replication (Zhang et al. 2008, Coutinho et al. 2020). The PUR gene with the highest count in our samples was phosphoribosylglycinamide formyltransferase (*purN*), which was found to be a part of Nucleotide metabolism and Purine metabolism pathway within the ETNP OMZ. Another AMG involved in replication detected in our samples was the gene that encoded for the cobaltochelatase subunit (*cobS*) which is a part of the cobaltochelatase complex. The *cobS* gene is critically important as it activates the final step of Cobalamin (vitamin B12) synthesis (Heyerhoff, et al. 2022), as vitamin B12 can act as a limiting nutrient in the ocean (Sañudo-Wilhelmy et al. 2014). Certain marine prokaryotes are sole producers of vitamin B12 and can exert control over marine productivity, as auxotrophs interact with these cobalamin producers to fulfill their vitamin needs (Heal et al. 2017, Croft et al. 2005). Vitamin B12 has been measured with higher concentrations in coastal waters as opposed to open ocean, however it can be undetectable in large areas of the coastal ocean (Panzeca et al. 2009; Sañudo-Wilhelmy et al. 2012). Cobalamin is also a cofactor that plays a role in reducing ribonucleotides to deoxyribonucleotides for viral replication. (Zimmerman et al. 2020). Interestingly, the *cobS* gene has been found in tailed viruses that infect members of both bacterial and archaeal members (Liu et al. 2021). We found that *cobS* had the highest count at 87 m at Station 3 in the PNM of the OMZ.

Among viral replication genes, some were interestingly only found or abundant in the Station 1 25m sample. At the time of our sampling, Station 1 had notably high chlorophyll and ammonium concentrations and a shallow OMZ at 25 m indicative of upwelling-driven productivity. We previously showed a unique microbial community composition through 16s amplicon sequencing (Beman et al. 2021a). Unique or highly abundant viral AMGs in this sample included DNA (cytosine-5)-methyltransferase 3A (*DNMT3A*) and DNA (cytosine-5)-methyltransferase 1 (*DNMT1*), which are a part of Amino acid metabolism and Cysteine and methionine metabolism and may be part of the virus-host arms race (Gao et al. 2022). Both *DNMT3A* and *DNMT1* have not been previously detected in the ETNP OMZ. In addition, the 3-oxoacyl-[acyl-carrier-protein] synthase II (*fabF)* gene was unique to Station 1 at 25m, which is important for the synthesis of fatty acids and phospholipids (Forcone et al. 2021). Although the *fabF* gene has not been found in previous OMZ studies, it has been detected in the Prophages of Roseobacters that were sampled during an algal bloom (Forcone et al. 2021). Other AMGs found in Station 1 at 25m included the UDP-glucose 6-dehydrogenase (*UGDH*) gene, part of Carbohydrate metabolism and Amino sugar and nucleotide sugar metabolism, Ascorbate and aldarate metabolism, and Pentose and glucuronate interconversions. The *UGDH* gene has previously been detected in phages important in galactose metabolism towards energy production (Heyerhoff et al. 2022; Zhao et al. 2022). Another gene only found in station 1 at 25m is the xyloglucan:xyloglucosyltransferase(*GH16)* gene, which is found in all land plants and encodes for the building blocks of plant cell walls, but subgroups of this gene have been found in Gammaproteobacteria and used to degrade plant cell walls (Edwards et al. 2010; Strohmeier et al. 2004). The *GH16* AMGs has been detected in lytic viruses that are coded for cell wall degradation (Luo et al. 2022).

Outside of Station 1 at 25m, other genes that appeared in our samples with a high count included the tryptophan 7-halogenase (*prnA*) gene. This gene is involved in the initial reaction of pyrrolnitrin formation, a strong antifungal antibiotic, and has been found in phages found in aquatic environments when detected by VIBRANT (Kieft et al. 2020). Fungi do have a presence in the OMZ, with mainly Basidiomycota and Ascomycota contributing to biogeochemical cycling by producing nitrous oxide (Peng et al. 2021). In our dataset, the *prnA* gene had the highest count at the PNM sample station 3 at 87m, The heme oxygenase (*hmuO)* gene also displayed a high count in the upper water column and participates in phycobilin pigment biosynthesis, and is important as phycobilin is a precursor to phycobilin-bearing phycobiliproteins (PBPs) to create phycobilisomes (PBSs) which are light harvesting complexes (Ledermann et al. 2017; Okada 2009; Puxty et al. 2015; Shan et al. 2008; Waldbauer et al. 2019). The heme oxygenase gene has been detected in Cyanophage S-SM1 that infect Synechococcus WH8102, and also in the genomes of several *Prochlorococcus* phages and a single *Synechococcus* phage S-SM7 (Crummett et al. 2016; Waldbauer et al. 2019). Phycobilisomes are found in photoautotrophic picocyanobacteria (specifically *Synechococcus*) in the Baltic Sea, which contains similar conditions as OMZs (Larsson et al. 2014). Although phycobilisomes are not typically associated with *Prochlorococcus*, they have been found in specific ecotypes of *Prochlorococcus* inhabit hypoxic and low-light waters (Ulloa et al. 2021) and are the main groups of cyanobacteria that inhabit the ODZ (Aldunate et al. 2022). This could explain the presence of these genes in the SCM in station 2 and 3. The last interesting gene that contained a high count in the ETNP OMZ samples was the AMG transaldolase (*talA*). The viral version of transaldolase is *talC* and has been identified in cyanophages; Sullivan et al. 2005 suggests that the role of *talC* is in metabolizing carbon substrates for biosynthesis and energy production during phage infection of the cyanobacterial host. With the viral transaldolase gene *talC*, cyanophages may increase host pentose phosphate reactions by boosting pentose phosphate pathway activity to produce ribose 5-phosphate and NADPH for viral replication (Thompson et al. 2011).

Finally, critically important AMGs consistently present in marine phage are those involved in photosynthesis (Lindell et al. 2004; Millard et al. 2004; Clokie et al. 2010). Viral transfer of photosynthesis genes may be particularly relevant in OMZs/AMZs because surface primary production is critical in the generation of OMZs/AMZs (Gilly et al. 2013), while the subsurface SCM is also biogeochemically important (Beman et al. 2021b). In our samples, one of the viral AMGs with the highest overall count was the *psbA* gene; we also detected the *psbD* viral AMG in our samples. Both *psbA* and *psbD* genes are important because they encode for core proteins, D1 and D2, involved in photosystem II for oxygenic photosynthesis in cyanobacteria (Millard et al. 2004). Viral AMGs also included the *petE* and *petF* genes, with *petE* coding for the photosynthetic electron transport protein plastocyanin, which transfers electrons from the cytochrome b6f complex of photosystem II to P700+ of photosystem I—it has been found in T-4 cyanophages and phages that infect marine *Prochlorococcus* (Lindell et al. 2004; Clokie et al. 2010). The *petF* gene encodes for ferredoxin iron containing proteins involved in electron transfer of photosystem I (Lindell et al. 2004; Clokie et al. 2010). Finally, the *cpeT* gene that codes for the *CpeT* protein, which also had a high count in the ETNP OMZ, with the highest count in station 3 at 87m (Fig. 2.6). The *cpeT* gene has been detected previously in cyanophages and is similar to the *cpcT* gene in cyanobacteria; the viral *cpeT* gene encodes for a putative phycobiliprotein lyase that helps with the assembly of light-harvesting phycobiliproteins through the attachment of chromophores (Gasper et al. 2017). The presence of these putative viral AMGs in the ETNP suggests that phages indirectly play a role in primary production in this biogeochemically important environment by influencing cyanobacterial host photosynthesis during infection.

2.6 References

Ahlgren, N. A., Fuchsman, C. A., Rocap, G., & Fuhrman, J. A. (2019). Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. The ISME journal, 13(3), 618-631.

Aldunate, M., Von Dassow, P., Vargas, C. A., & Ulloa, O. (2022). Carbon assimilation by the picoplanktonic community inhabiting the secondary chlorophyll maximum of the anoxic marine zones of the eastern tropical north and south pacific. *Frontiers in Marine Science*, *9*, 858308.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

Anantharaman, K., Duhaime, M. B., Breier, J. A., Wendt, K. A., Toner, B. M., & Dick, G. J. (2014). Sulfur oxidation genes in diverse deep-sea viruses. Science, 344(6185), 757-760.

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., ... & Rohwer, F. (2006). The marine viromes of four oceanic regions. PLoS biology, 4(11), e368.

Beman, J.M., Popp, B. N., & Alford, S. E. (2012). Quantification of ammonia oxidation rates and ammonia-oxidizing archaea and bacteria at high resolution in the Gulf of California and eastern tropical North Pacific Ocean. Limnology and Oceanography, 57(3), 711-726.

Beman, J. M., & Carolan, M. T. (2013). Deoxygenation alters bacterial diversity and community composition in the ocean’s largest oxygen minimum zone. Nature Communications, 4(1), 2705.

Beman, J. M., Vargas, S. M., Vazquez, S., Wilson, J. M., Yu, A., Cairo, A., & Perez-Coronel, E. (2021a). Biogeochemistry and hydrography shape microbial community assembly and activity in the eastern tropical North Pacific Ocean oxygen minimum zone. Environmental Microbiology, 23(6), 2765-2781.

Beman, J. M., Vargas, S. M., Wilson, J. M., Perez-Coronel, E., Karolewski, J. S., Vazquez, S., ... & Wankel, S. D. (2021b). Substantial oxygen consumption by aerobic nitrite oxidation in oceanic oxygen minimum zones. *Nature Communications*, *12*(1), 7043.

Bertagnolli, A. D., & Stewart, F. J. (2018). Microbial niches in marine oxygen minimum zones. Nature Reviews Microbiology, 16(12), 723-729.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., ... & Sullivan, M. B. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nature biotechnology, 37(6), 632-639.

Breitbart, M. Y. A., Thompson, L. R., Suttle, C. A., & Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. Oceanography, 20(2), 135-139

Breitbart, M., Bonnain, C., Malki, K., & Sawaya, N. A. (2018). Phage puppet masters of the marine microbial realm. Nature microbiology, 3(7), 754-766.

Brister, J. R., Ako-Adjei, D., Bao, Y., & Blinkova, O. (2015). NCBI viral genomes resource. Nucleic acids research, 43(D1), D571-D577.

Brum, J. R., & Sullivan, M. B. (2015). Rising to the challenge: accelerated pace of discovery transforms marine virology. Nature Reviews Microbiology, 13(3), 147-159.

Busse, L., Tisza, M., & DiRuggiero, J. (2022). Viruses ubiquity and diversity in Atacama Desert endolithic communities. Viruses, 14(9), 1983.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC bioinformatics, 10, 1-9.

Cardol, P., Bailleul, B., Rappaport, F., Derelle, E., Béal, D., Breyton, C., ... & Finazzi, G. (2008). An original adaptation of photosynthesis in the marine green alga Ostreococcus. *Proceedings of the National Academy of Sciences, 105*(22), 7881-7886.

Cassman, N., Prieto-Davó, A., Walsh, K., Silva, G. G., Angly, F., Akhter, S., ... & Dinsdale, E. A. (2012). Oxygen minimum zones harbour novel viral communities with low diversity. Environmental microbiology, 14(11), 3043-3065.

Chivian, D., Jungbluth, S. P., Dehal, P. S., Wood-Charlson, E. M., Canon, R. S., Allen, B. H., ... & Arkin, A. P. (2023). Metagenome-assembled genome extraction and analysis from microbiomes using KBase. Nature Protocols, 18(1), 208-238.

Chu, Y., Zhao, Z., Cai, L., & Zhang, G. (2022). Viral diversity and biogeochemical potential revealed in different prawn-culture sediments by virus-enriched metagenome analysis. Environmental Research, 210, 112901.

Clokie, M. R., Millard, A. D., & Mann, N. H. (2010). T4 genes in the marine ecosystem: studies of the T4-like cyanophages and their role in marine ecology. Virology Journal, 7, 1-19.

Coutinho, F. H., Cabello-Yeves, P. J., Gonzalez-Serrano, R., Rosselli, R., López-Pérez, M., Zemskaya, T. I., ... & Rodriguez-Valera, F. (2020). New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. Microbiome, 8(1), 1-15.

Croft, M. T., Lawrence, A. D., Raux-Deery, E., Warren, M. J., & Smith, A. G. (2005). Algae acquire vitamin B12 through a symbiotic relationship with bacteria. Nature, 438(7064), 90-93.

Crummett, L. T., Puxty, R. J., Weihe, C., Marston, M. F., & Martiny, J. B. (2016). The genomic content and context of auxiliary metabolic genes in marine cyanomyoviruses. Virology, 499, 219-229.

Demir-Hilton, E., Sudek, S., Cuvelier, M. L., Gentemann, C. L., Zehr, J. P., & Worden, A. Z. (2011). Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. The ISME journal, 5(7), 1095-1107.

Du, S., Qin, F., Zhang, Z., Tian, Z., Yang, M., Liu, X., ... & Zhao, Y. (2021). Genomic diversity, life strategies and ecology of marine HTVC010P-type pelagiphages. Microbial Genomics, 7(7).

Edwards, J. L., Smith, D. L., Connolly, J., McDonald, J. E., Cox, M. J., Joint, I., ... & McCarthy, A. J. (2010). Identification of carbohydrate metabolism genes in the Metagenome of a marine biofilm community shown to be dominated by Gammaproteobacteria, Bacteroidetes. Genes, 1(3), 371-384.

Forcone, K., Coutinho, F. H., Cavalcanti, G. S., & Silveira, C. B. (2021). Prophage genomics and ecology in the family Rhodobacteraceae. Microorganisms, 9(6), 1115.

Freitas, T. A. K., Li, P. E., Scholz, M. B., & Chain, P. S. (2015). Accurate read-based metagenome characterization using a hierarchical suite of unique signatures. Nucleic acids research, 43(10), e69-e69.

Fuchsman, C. A., Carlson, M. C., Garcia Prieto, D., Hays, M. D., & Rocap, G. (2021). Cyanophage host-derived genes reflect contrasting selective pressures with depth in the oxic and anoxic water column of the Eastern Tropical North Pacific. Environmental Microbiology, 23(6), 2782-2800.

Fuchsman, C. A., Palevsky, H. I., Widner, B., Duffy, M., Carlson, M. C., Neibauer, J. A., ... & Rocap, G. (2019). Cyanobacteria and cyanophage contributions to carbon and nitrogen cycling in an oligotrophic oxygen-deficient zone. The ISME Journal, 13(11), 2714-2726.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, *399*(6736), 541-548.

Gao, C., Liang, Y., Jiang, Y., Paez-Espino, D., Han, M., Gu, C., ... & Wang, M. (2022). Virioplankton assemblages from challenger deep, the deepest place in the oceans. Iscience, 25(8).

Garcia-Robledo, E., Padilla, C. C., Aldunate, M., Stewart, F. J., Ulloa, O., Paulmier, A., ... & Revsbech, N. P. (2017). Cryptic oxygen cycling in anoxic marine zones. Proceedings of the National Academy of Sciences, 114(31), 8319-8324.

Gasper, R., Schwach, J., Hartmann, J., Holtkamp, A., Wiethaus, J., Riedel, N., ... & Frankenberg-Dinkel, N. (2017). Distinct features of cyanophage-encoded T-type phycobiliprotein lyase ΦCpeT: The role of auxiliary metabolic genes. Journal of Biological Chemistry, 292(8), 3089-3098.

Gazitúa, M. C., Vik, D. R., Roux, S., Gregory, A. C., Bolduc, B., Widner, B., ... & Sullivan, M. B. (2021). Potential virus-mediated nitrogen cycling in oxygen-depleted oceanic waters. The ISME Journal, 15(4), 981-998.

Gilly, W. F., Beman, J. M., Litvin, S. Y., & Robison, B. H. (2013). Oceanographic and biological effects of shoaling of the oxygen minimum zone. Annual review of marine science, 5(1), 393-420.

Goericke, R., Olson, R. J., & Shalapyonok, A. J. D. S. R. P. I. O. R. P. (2000). A novel niche for Prochlorococcus sp. in low-light suboxic environments in the Arabian Sea and the Eastern Tropical North Pacific. Deep Sea Research Part I: Oceanographic Research Papers, 47(7), 1183-1205.

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., ... & Roux, S. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome, 9, 1-13.

Heal, K. R., Qin, W., Ribalet, F., Bertagnolli, A. D., Coyote-Maestas, W., Hmelo, L. R., ... & Ingalls, A. E. (2017). Two distinct pools of B12 analogs reveal community interdependencies in the ocean. Proceedings of the National Academy of Sciences, 114(2), 364-369.

Heyerhoff, B., Engelen, B., & Bunse, C. (2022). Auxiliary metabolic gene functions in pelagic and benthic viruses of the Baltic Sea. Frontiers in microbiology, 13, 863620.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in science & engineering, 9(03), 90-95.

Jaiani, E., Kusradze, I., Kokashvili, T., Geliashvili, N., Janelidze, N., Kotorashvili, A., ... & Prangishvili, D. (2020). Microbial Diversity and Phage–Host Interactions in the Georgian Coastal Area of the Black Sea Revealed by Whole Genome Metagenomic Sequencing. Marine Drugs, 18(11), 558.

Jurgensen, S. K., Roux, S., Schwenck, S. M., Stewart, F. J., Sullivan, M. B., & Brum, J. R. (2022). Viral community analysis in a marine oxygen minimum zone indicates increased potential for viral manipulation of microbial physiological state. The ISME journal, 16(4), 972-982.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D’Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. Proceedings of the National Academy of Sciences, 109(40), 16213-16216.

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome, 8(1), 1-23.

Lam, P., & Kuypers, M. M. (2011). Microbial nitrogen cycling processes in oxygen minimum zones. Annual review of marine science, 3, 317-345.

Labonté, J. M., & Suttle, C. A. (2013). Metagenomic and whole-genome analysis reveals new lineages of gokushoviruses and biogeographic separation in the sea. *Frontiers in microbiology*, *4*, 404

Labonté, J. M., Hallam, S. J., & Suttle, C. A. (2015). Previously unknown evolutionary groups dominate the ssDNA gokushoviruses in oxic and anoxic waters of a coastal marine environment. Frontiers in Microbiology, 6, 315.

Larsson, J., Celepli, N., Ininbergs, K., Dupont, C. L., Yooseph, S., Bergman, B., & Ekman, M. (2014). Picocyanobacteria containing a novel pigment gene cluster dominate the brackish water Baltic Sea. The ISME Journal, 8(9), 1892-1903.

Lavin, P., González, B., Santibáñez, J. F., Scanlan, D. J., & Ulloa, O. (2010). Novel lineages of Prochlorococcus thrive within the oxygen minimum zone of the eastern tropical South Pacific. Environmental microbiology reports, 2(6), 728-738.

Ledermann, B., Aras, M., & Frankenberg-Dinkel, N. (2017). Biosynthesis of cyanobacterial light-harvesting pigments and their assembly into phycobiliproteins. Modern topics in the phototrophic prokaryotes: metabolism, bioenergetics, and omics, 305-340.

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10), 1674-1676.

Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F., & Chisholm, S. W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proceedings of the National Academy of Sciences, 101(30), 11013-11018.

Liu, Y., Demina, T. A., Roux, S., Aiewsakun, P., Kazlauskas, D., Simmonds, P., ... & Krupovic, M. (2021). Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. PLoS biology, 19(11), e3001442.

Long, A. M., Jurgensen, S. K., Petchel, A. R., Savoie, E. R., & Brum, J. R. (2021). Microbial ecology of oxygen minimum zones amidst ocean deoxygenation. Frontiers in Microbiology, 12, 748961.

Luo, X. Q., Wang, P., Li, J. L., Ahmad, M., Duan, L., Yin, L. Z., ... & Li, W. J. (2022). Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts. Microbiome, 10(1), 190.

McKinney, W. (2011). pandas: a foundational Python library for data analysis and statistics. Python for high performance and scientific computing, 14(9), 1-9.

Millard, A., Clokie, M. R., Shub, D. A., & Mann, N. H. (2004). Genetic organization of the psbAD region in phages infecting marine Synechococcus strains. Proceedings of the National Academy of Sciences, 101(30), 11007-11012.

Monier, A., Chambouvet, A., Milner, D. S., Attah, V., Terrado, R., Lovejoy, C., ... & Richards, T. A. (2017). Host-derived viral transporter protein for nitrogen uptake in infected marine phytoplankton. Proceedings of the National Academy of Sciences, 114(36), E7489-E7498.

Munn, C. B. (2006). Viruses as pathogens of marine organisms—from bacteria to whales. Journal of the Marine Biological Association of the United Kingdom, 86(3), 453-467.

Muratore, D., Bertagnolli, A. D., Bristow, L. A., Thamdrup, B., Weitz, J. S., & Stewart, F. J. (2023). Microbial and viral genome and proteome nitrogen demand varies across multiple spatial scales within a marine oxygen minimum zone. Msystems, 8(2), e01095-22.

Nepusz, T., Yu, H., & Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nature methods, 9(5), 471-472.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome research, 27(5), 824-834.

Okada, K. (2009). HO1 and PcyA proteins involved in phycobilin biosynthesis form a 1: 2 complex with ferredoxin-1 required for photosynthesis. FEBS letters, 583(8), 1251-1256.

Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., ... & Grigoriev, I. V. (2007). The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proceedings of the National Academy of Sciences, 104*(18), 7705-7710.

Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. Progress in Oceanography, 80(3-4), 113-128.

Panzeca, C., Beck, A. J., Tovar-Sanchez, A., Segovia-Zavala, J., Taylor, G. T., Gobler, C. J., & Sanudo-Wilhelmy, S. A. (2009). Distributions of dissolved vitamin B12 and Co in coastal and open-ocean environments. Estuarine, Coastal and Shelf Science, 85(2), 223-230.

Peng, X., & Valentine, D. L. (2021). Diversity and N2O production potential of fungi in an oceanic oxygen minimum zone. Journal of Fungi, 7(3), 218.

Puxty, R. J., Millard, A. D., Evans, D. J., & Scanlan, D. J. (2015). Shedding new light on viral photosynthesis. Photosynthesis research, 126, 71-97.

Rohwer, F., & Thurber, R. V. (2009). Viruses manipulate the marine environment. Nature, 459(7244), 207-212

Rohwer, F., Prangishvili, D., & Lindell, D. (2009). Roles of viruses in the environment. Environmental microbiology, 11(11), 2771-2774.

Roux, S., Brum, J. R., Dutilh, B. E., Sunagawa, S., Duhaime, M. B., Loy, A., ... & Sullivan, M. B. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature, 537(7622), 689-693.

Roux, S., Krupovic, M., Poulet, A., Debroas, D., & Enault, F. (2012). Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads. PloS one, 7(7), e40418.

Roux, S., Hawley, A. K., Torres Beltran, M., Scofield, M., Schwientek, P., Stepanauskas, R., ... & Sullivan, M. B. (2014). Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell-and meta-genomics. elife, 3, e03125.

Sañudo-Wilhelmy, S. A., Cutter, L. S., Durazo, R., Smail, E. A., Gómez-Consarnau, L., Webb, E. A., ... & Karl, D. M. (2012). Multiple B-vitamin depletion in large areas of the coastal ocean. Proceedings of the National Academy of Sciences, 109(35), 14041-14045.

Sañudo-Wilhelmy, S. A., Gómez-Consarnau, L., Suffridge, C., & Webb, E. A. (2014). The role of B vitamins in marine biogeochemistry. *Annual review of marine science*, *6*(1), 339-367.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics, 30*(14), 2068-2069.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., ... & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic acids research, 48(16), 8883-8900.

Shan, J., Jia, Y., Clokie, M. R., & Mann, N. H. (2008). Infection by the ‘photosynthetic’phage S-PM2 induces increased synthesis of phycoerythrin in Synechococcus sp. WH7803. FEMS microbiology letters, 283(2), 154-161.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504.

Sharon, I., Tzahor, S., Williamson, S., Shmoish, M., Man-Aharonovich, D., Rusch, D. B., ... & Béja, O. (2007). Viral photosynthetic reaction center genes and transcripts in the marine environment. The ISME journal, 1(6), 492-501.

Stern, A., & Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. Bioessays, 33(1), 43-51

Strohmeier, M., Hrmova, M., Fischer, M., Harvey, A. J., Fincher, G. B., & Pleiss, J. (2004). Molecular modeling of family GH16 glycoside hydrolases: potential roles for xyloglucan transglucosylases/hydrolases in cell wall modification in the poaceae. Protein Science, 13(12), 3200-3213.

Stramma, L., Johnson, G. C., Sprintall, J., & Mohrholz, V. (2008). Expanding oxygen-minimum zones in the tropical oceans. science, 320(5876), 655-658..

Sullivan, M. B., Coleman, M. L., Weigele, P., Rohwer, F., & Chisholm, S. W. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS biology, 3(5), e144.

Suttle, C. A. (2007). Marine viruses—major players in the global ecosystem. Nature reviews microbiology, 5(10), 801-812.

Suttle, C. A. (2005). Viruses in the sea. Nature, 437(7057), 356-361.

Thamdrup, B., Dalsgaard, T., & Revsbech, N. P. (2012). Widespread functional anoxia in the oxygen minimum zone of the Eastern South Pacific. Deep Sea Research Part I: Oceanographic Research Papers, 65, 36-45.

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., & Chisholm, S. W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proceedings of the National Academy of Sciences, 108(39), E757-E764.

Tsementzi, D., Wu, J., Deutsch, S., Nath, S., Rodriguez-R, L. M., Burns, A. S., ... & Stewart, F. J. (2016). SAR11 bacteria linked to ocean anoxia and nitrogen loss. Nature, 536(7615), 179-183.

Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M., & Stewart, F. J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences, 109(40), 15996-16003.

Ulloa, O., Henríquez-Castillo, C., Ramírez-Flandes, S., Plominsky, A. M., Murillo, A. A., Morgan-Lang, C., ... & Stepanauskas, R. (2021). The cyanobacterium Prochlorococcus has divergent light-harvesting antennae and may have evolved in a low-oxygen ocean. Proceedings of the National Academy of Sciences, 118(11), e2025638118.

Vik, D., Gazitúa, M. C., Sun, C. L., Zayed, A. A., Aldunate, M., Mulholland, M. R., ... & Sullivan, M. B. (2021). Genome-resolved viral ecology in a marine oxygen minimum zone. Environmental Microbiology, 23(6), 2858-2874.

Vik, D. R., Roux, S., Brum, J. R., Bolduc, B., Emerson, J. B., Padilla, C. C., ... & Sullivan, M. B. (2017). Putative archaeal viruses from the mesopelagic ocean. PeerJ, 5, e3428.

Waldbauer, J. R., Coleman, M. L., Rizzo, A. I., Campbell, K. L., Lotus, J., & Zhang, L. (2019). Nitrogen sourcing during viral infection of marine cyanobacteria. Proceedings of the National Academy of Sciences, 116(31), 15590-15595.

Warwick-Dugdale, J., Buchholz, H. H., Allen, M. J., & Temperton, B. (2019). Host-hijacking and planktonic piracy: how phages command the microbial high seas. Virology journal, 16, 1-13.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. FEMS microbiology reviews, 28(2), 127-181.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. Proceedings of the National Academy of Sciences, 95(12), 6578-6583.

Wickham, H. (2011). ggplot2. Wiley interdisciplinary reviews: computational statistics, 3(2), 180-185.

Wigington, C. H., Sonderegger, D., Brussaard, C. P., Buchan, A., Finke, J. F., Fuhrman, J. A., ... & Weitz, J. S. (2016). Re-examination of the relationship between marine virus and microbial cell abundances. Nature microbiology, 1(3), 1-9.

Wommack, K. E., & Colwell, R. R. (2000). Virioplankton: viruses in aquatic ecosystems. Microbiology and molecular biology reviews, 64(1), 69-114.

Worden, A. Z. (2006). Picoeukaryote diversity in coastal waters of the Pacific Ocean. Aquatic Microbial Ecology, 43(2), 165-175.

Zakem, E. J., Mahadevan, A., Lauderdale, J. M., & Follows, M. J. (2020). Stable aerobic and anaerobic coexistence in anoxic marine zones. The ISME Journal, 14(1), 288-301.

Zeng, Q., & Chisholm, S. W. (2012). Marine viruses exploit their host's two-component regulatory system in response to resource limitation. Current Biology, 22(2), 124-128.

Zhang, Y., Morar, M., & Ealick, S. E. (2008). Structural biology of the purine biosynthetic pathway. Cellular and molecular life sciences, 65, 3699-3724.

Zhao, J., Jing, H., Wang, Z., Wang, L., Jian, H., Zhang, R., ... & Zhang, Y. (2022). Novel viral communities potentially assisting in carbon, nitrogen, and sulfur metabolism in the upper slope sediments of Mariana Trench. MSystems, 7(1), e01358-21.

Zhao, Y., Temperton, B., Thrash, J. C., Schwalbach, M. S., Vergin, K. L., Landry, Z. C., ... & Giovannoni, S. J. (2013). Abundant SAR11 viruses in the ocean. Nature, 494(7437), 357-360.

Zimmerman, A. E., Howard-Varona, C., Needham, D. M., John, S. G., Worden, A. Z., Sullivan, M. B., ... & Coleman, M. L. (2020). Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. Nature Reviews Microbiology, 18(1), 21-34.

Chapter 3: Ammonia-Oxidizing Archaea in Oxygen Minimum Zones Exhibit Functional and Genomic Differences

3.1 Abstract

Ammonia oxidizing Archaea (AOA) are one of the most abundant microbes in our oceans and are key players to the nitrogen cycle, as they help convert ammonia to nitrite. Although AOA are considered an aerobic group, there have been cases where AOA have been detected in hypoxic environments. Oxygen minimum zones (OMZs) are regions in the ocean where dissolved oxygen levels are <20 µM and contribute greatly to the global nitrogen cycle. AOA have previously been identified in the Eastern Tropical North Pacific (ETNP) OMZ though the 16s rRNA and the *amoA* gene. In this study we revisited AOA community analysis through metagenome assembled genomes (MAGs) to explore the functionality of AOA that inhabit anoxic regions of the ETNP OMZ. Throughout the water column we were able to recover 8 AOA MAGs which have been identified as *Nitrosopelagicus* and *Nitrosopumilus*. Interestingly the MAGs were mainly recovered from the Secondary Chlorophyll Maxima within the Anoxic Marine Zone (AMZ) core and the OMZ Edge sample station 3 at 123m. Phylogenetic analysis of the MAGs led to 3 separate clades that are mainly separated by identification, however the AOA MAG from station 3 at 123m are present in all clades. This study shows that AOA are not fully aerobic as they are found in anoxic conditions and still being able to perform ammonia oxidation. The MAGs also revealed that AOA are diverse and contain unique metabolic capabilities while inhabiting these harsh conditions.

3.2 Introduction

Marine *Thaumarchaeota* are pervasive throughout the deep ocean below the euphotic zone and are one of Earth's single most abundant groups of organisms (Karner et al. 2001). The majority of these organisms are able to oxidize ammonia using oxygen to generate energy (Francis et al. 2005; Könneke et al. 2005), while fixing organic carbon chemoautotrophically via the efficient 3-hydroxypriopionate/4-hydroxybutyrate pathway (Berg et al. 2007; Berg et al. 2010); some from the pSL12-like clade may be able to function heterotrophically based on metagenome assembled genomes (MAGs) (Aylward and Santoro 2020; Reji & Francis 2020). Within these extremes, *Thaumarchaeota* acquire and use comparatively simple organic N compounds like urea (Kitzinger et al. 2019; Qin et al. 2024), take up amino acids (Ouverney & Fuhrman 2000), and use organic carbon when bound in N-containing substrates (Parada et al. 2023). However, all marine *Thaumarchaeota* are understood to be obligately aerobic and require oxygen (Ren et al. 2019; Santoro et al. 2019; Baker et al. 2020). The fact that *Thaumarchaeota* are commonly detected within low-oxygen ecosystems of all kinds (Francis et al. 2007; Santoro et al. 2019; Vuillemin et al. 2019) therefore raises multiple questions about their persistence and activity under such conditions.

In the ocean where *Thaumarchaeota* are highly abundant, they are commonly detected and exhibit high levels of gene expression (Stewart et al. 2012) even within oxygen minimum zones (OMZs) defined by dissolved oxygen concentrations <20 µM (Paulmier and Ruiz-Pino 2009). Within this broader OMZ category, a subset of these regions are functionally anoxic and known as oxygen deficient zones (ODZs) or anoxic marine zones (AMZs; Ulloa et al. 2012). These regions include the Arabian Sea, the eastern tropical South Pacific (ETSP), and the largest such region, the eastern tropical North Pacific (ETNP). Other regions of the ocean (e.g., the Bay of Bengal) sit at a tipping point between OMZ and AMZ (Bristow et al. 2016a), while OMZ/AMZ regions are expanding and intensifying due to climate change (Stramma et al. 2008; Gilly et al. 2013).

AMZs are distinguished from OMZs by active anaerobic N cycling and related processes (Ulloa et al. 2012)—particularly the presence of low-light adapted *Prochlorococcus* that form a prominent secondary chlorophyll maximum (SCM;Garcia-Robledo et al. 2017), as well as high rates of nitrite oxidation (especially within the SCM; Beman et al. 2021). Nitrite oxidation is sustained by specific *Nitrospina* ecotypes with an exceptionally high affinity for oxygen (Bristow et al. 2016b; Beman et al. 2021), and marine *Thaumarchaeota* also exhibit high oxygen affinities in experiments (Bristow et al. 2016b). Ammonia oxidation and nitrite oxidation are often tightly coupled throughout the ocean; however, a major exception occurs in OMZs, where ammonia oxidation rates are substantially lower than nitrite oxidation rates (Beman et al. 2013 and 2021; Füssel et al. 2012)—despite the presence of AOA and their potentially high affinity for oxygen. This suggests fundamental differences between the two main groups of oceanic nitrifiers within OMZs, but it remains unclear as to why AOA are much less active. Interestingly, AOA cultures were recently shown to produce oxygen in its absence (Kraft et al. 2022). This finding indicates that, under specialized circumstances, AOA might actively produce oxygen via a still unidentified pathway. This hypothesis is obviously deserving of additional study in low oxygen regions of the ocean given the fundamental implications and applications.

To examine AOA in the ETNP OMZ, we assembled MAGs from samples collected within four key features present across different stations and depths: (i) the primary nitrite maximum (PNM) at the base of the euphotic zone, where we expect high rates nitrification to occur (Beman et al. 2012); (ii) the edge of OMZ waters at 20 µM DO, where microbial communities are diverse and include AOA (Beman and Carolan 2013; Bertagnolli and Stewart 2018); (iii) the secondary chlorophyll maximum (SCM), where oxygen production could potentially support ammonia oxidation by AOA (Garcia-Robledo et al. 2017; Zakem et al. 2020); and (iv) the secondary nitrite maximum (SNM), where anaerobic organisms and metabolisms are present and active (Thamdrup et al. 2012; Ulloa et al. 2012). All of these features were present at three AMZ stations (1, 2, and 3) that extend off the coast of Mexico, but which differ in the depth of the features (Figure 2.1). For comparison, we also sampled the PNM, OMZ edge, and the single PCM present at an OMZ—but not AMZ—Station 4.

3.3 Materials and Methods

Sample collection, DNA extraction, and sequencing

Samples were collected in April 2017 aboard the *R/V Oceanus*, with samples collected in Mexican territorial waters under Instituto Nacional de Estadística y Geografía (INEGI) permit EG0062017 and Permiso de Pesca de Fomento permit PPFE/DGOPA-016/17. At each station, conductivity/salinity, temperature, depth, pressure, chlorophyll fluorescence, and photosynthetically active radiation (PAR) were measured by a SeaBird SBE-9plus CTD, SBE-3F temperature sensor, SBE-43 DO sensor, WetLabs ECO-FLR Fluorometer, and Biospherical QCP2200 PAR sensor. Nutrient samples were analyzed for NH_4^+ and NO_2^- aboard the ship as described in Beman et al. (2021 *Nature Communications*).

Water samples were collected for DNA extraction and sequencing using sampling bottles deployed on the CTD rosette. At each depth, 2L samples were filtered through 0.22 µm filters (Millipore, Darmstadt, Germany) using a peristaltic pump, then filters were submerged in Sucrose-Tris-EDTA (STE) buffer in pre-prepped Lysis Matrix E tubes and frozen at -80°C until extraction. DNA was extracted from filters following Beman et al. (2012) and DNA samples were sent for metagenome sequencing in the Vincent J. Coates Genome Sequencing Laboratory (GSL) at the University of California, Berkeley (https://genomics.qb3.berkeley.edu/), which is supported by NIH S10 OD018174 Instrumentation Grant. For each sample, 250 ng of genomic DNA was sheared and libraries were prepared using the KAPA HyperPrep Kit (Kapa Biosystems, Wilmington, MA, USA). Samples were pooled into a single lane and sequenced via 150-cycle paired-end sequencing on the Illumina HiSeq 4000 platform (Illumina, Inc., San Diego, CA, USA). Data were demultiplexed by the GSL and reads were filtered and trimmed using BBDuk (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) with the following parameters: maq=8, maxns=1, minlen=40, minlenfraction=0.6, k=23, hdist=1, trimq=12, qtrim=rl. Forward and reverse reads were then merged using PANDASeq (https://github.com/neufeld/pandaseq; Masella et al. 2012) with default parameters.

AOA MAG Assembly

After DNA extraction and metagenome sequencing, reads were imported to kbase and scanned for quality using FastQC (Arkin et al. 2018; Andrews 2010) and then assembled using MetaSPAdes (Nurk et al. 2017). MetaSPAdes was used with default parameters and with a contig length of 2000. Assembled contigs were then input into 3 individual binning tools/apps to create draft metagenome-assembled genomes (MAGs) following the approach of Chivian et al. 2023. The first binning tool was MaxBin2, which was used with default settings where probability threshold was at 0.8, “marker set” was at 107 bacterial marker geneset, and minimum contig length was set to 1000 (Wu et al. 2016). The next binning tool was CONCOCT, which was also used with default parameters, where the read mapping tool is bowtie2, the minimum contig length was 2500, contig split size was 10000, contig split overlap was 0, kmer length was 4, Maximum Number of Clusters for VGMM was 400, Maximum Number of Iterations for VGMM was 500, and lastly the Percent of Total PCA Used was 90 (Alneberg et al. 2013). The last binning tool used was MetaBAT2, also used with default parameters where the minimum contig length was 2500 (Kang et al. 2015). Once all contigs were binned with the different binning tools, putative MAGs were input into dereplication, aggregation and scoring strategy, DAS Tool (Sieber et al. 2018).

DAS Tool took the “bins” that were generated from the 3 binning tools on the contigs that were assembled with metaSPAdes and created a single aggregate, optimized, non-redundant set of bins from a single assembly (Sieber et al. 2018). DAS Tool was used with default parameters, where the gene identification tool was Diamond (Buchfink et al. 2015), score threshold was 0.5, duplicate penalty was 0.6, Megabin penalty was 0.5, and bin evals were written. Once all binned samples were input, an individual single aggregated optimized set of bins were created for each sample. The optimized bin sets from all samples were then input into CheckM for bin quality control, which produced completeness and contamination percentages (Parks et al. 2015). We further used CheckM to filter bins with a completion greater than 70 percent and contamination less than 5 percent for further analysis. These filtered bins were then extracted into assembly objects using the kbase app “BinUtil” or “Extract Bins as Assemblies from BinnedContigs” (Arkin et al. 2018; Chivian et al. 2023) and annotated using Rapid Annotations using Subsystems Technology toolkit (RASTtk; Aziz et al. 2008), which generated MAGs genome type objects with both coding and non-coding features. MAGs were then classified with GTDB-Tk v2.3.2, a software toolkit for assigning objective taxonomic classifications to bacterial and archaeal genomes (Chaumeil et al. 2020).

To identify putative AOA MAGs within the full dataset, archaeal MAGs were input into Distilled and Refined Annotation of Metabolism (DRAM; Shaffer et al. 2020) to generate a metabolic summary. DRAM was used with default parameters, where the bit score threshold was 60 and the Reverse search bit score threshold was 350, and used to identify archaeal MAGs containing the diagnostic archaeal ammonia monooxygenase gene *amoA*. Once these putative AOA MAGs were identified, they were then used for subsequent phylogenomic and pangenome analysis.

Phylogenomics

For phylogenomic analysis, all identified AOA genomes were input into the kbase application “Insert Set of Genomes Into SpeciesTree - v2.2.0,” where the neighbor public genome count was set to 21, and these 21 public genome where then copied into the kbase workspace (Price et al. 2010). We generated phylogeny on all identified AOA genomes along with the 21 public genomes that were closely related to the ETNP AOA based on 49 core universal genes defined by Clusters of Orthologous Groups (COGs) gene families (Tatusov et al. 2000). The genomes in this phylogeniy were grouped together into a GenomeSet and input into the “BLASTp prot-prot Search - v2.13.0+” app or the BLASTp app, and then used as a reference to search for protein coding sequences (Altschul et al. 1997; Camacho et al. 2009). The protein sequence that we used to search within the ETNP AOA genomes and the 21 closely related public genomes was the *amoA* gene, specifically from MAG 21 at station 3 at 140m, which was the most complete MAG in our dataset.

The BLASTp tool was used with default parameters, with an E value threshold of 0.001 and Sequence Identity Threshold (%) of 20%. Once this program was run on the GenomeSet, the *amoA* sequences from all the genomes were combined into a single feature set and aligned using MUSCLE (Edgar 2004). Then alignment file of the aligned *amoA* sequences of the AOA MAGs and 21 public genomes are inputted into “GBLOCKS Trim Multiple Sequence Alignment (MSA) - v0.91b”, which trimmed the *amoA* sequence alignment file to only conserve blocks with more reliable regions for evolutionary comparison (Castresana 2000). Once the *amoA* alignment file has been trimmed with GBLOCKs, it was inputted into the kbase app “Build Phylogenetic Tree from MSA using FastTree2 - v2.1.11” (Price et al. 2010). FastTree created a phylogenetic tree based on the *amoA* trimmed multiple sequence alignment file of the AOA genomes and the 21 public genomes. The newick files generated on these phylogenies were downloaded and modified in Rstudio with the package APE (Paradis et al. 2004).

We also compared the AOA MAGs within the ETNP OMZ to AOA Genomes detected in the Eastern Tropical South Pacific (ETSP) OMZ, specifically from the Sun & Ward 2021 dataset. The AOA MAGs from the ETSP dataset were downloaded from: https://figshare.com/articles/dataset/MAGs_from_ETSP_OMZ/12291281. These MAGs were then inputted into Kbase and are processed identically to our AOA MAGs from the ETNP.

AOA Genome Comparison and Pangenome Analysis

After phylogenomic analysis, all ETNP AOA genomes were grouped together and reannotated with protein domains from different domain libraries that included COGs (Clusters of Orthologous Groups) from the NCBI conserved domains database (CDD) version 3.19 and TIGRFAMs version 15.0 hidden Markov models from the J. Craig Venter Institute (Tatusov et al. 2000; Haft et al. 2012; Selengut et al. 2007; Eddy 2011). These newly annotated AOA MAGs were then input into the kbase app “View Function Profile for Genomes - v1.4.0”, which produced a heatmap of the percentages of the COG, TIGRFAM, and SEED functions of all the AOA MAGs. This app also created a table of the percentages and raw numbers of the COG, TIGRFAM, and SEED functions of all AOA MAGs (Tatusov et al. 2000; Haft et al. 2012; Selengut et al. 2007; Eddy 2011).

Each heatmap and table came with a key or list of genes that contributed to the functions that appeared in the figure.

We conducted a series of pangenome analyses on AOA MAGs. Four GenomeSets were used for Pangenome analysis: one GenomeSet included only the 8 AOA MAGs from the ETNP, and three new GenomeSets were created based on the placement of the AOA MAGs on the constructed phylogenetic trees. The first MAG GenomeSet was comprised of MAG 32 at station 1 at 87.5m, MAG 3 at station 3 at 123m, and MAG 21 at station 3 at 140m, and was called Clade1. The second GenomeSet, Clade2, consisted of MAG 14 at station 1 at 87.5m, MAG 13 at station 2 at 130m, MAG 9 at station 3 at 123m, and MAG 18 at station 3 at 140m. The last GenomeSet, Clade 3, consisted only of MAG 12 at station 3 at 123m.

Each GenomeSet was individually inputted into the Kbase app “Build Pangenome with OrthoMCL - v2.0” with default parameters, where a pangenome object of these three GenomeSets were produced (Arkin et al. 2018; Li et al. 2003). The pangenome objects included a pangenome summary, shared homolog families amongst the genomes, and a list of homologous protein clusters that were predicted by MCL.

AOA Pangenome Comparisons

The last program used was “Compare Genomes from Pangenome”, where all the MAGs within each AOA GenomeSet were compared to each other, and the proteins present in an Pangenome object were compared based on their function and sequence (Medini et al. 2005; Rasko et al. 2008). We conducted four separate comparisons, where one comparison was on the first pangenome object on the 8 AOA MAGs. The other three comparisons were on newly made pangenome objects created based on the placement of the AOA MAGs on the phylogenetic tree that was created on the 8 AOA MAGs along with the 21 closely related species.

Each individual pangenome object was then inputted into the kbase app “Compare Genomes from Pangenome”, where all homologous gene families for all genomes within a Pangenome were compared to one another (Medini et al. 2005; Rasko et al. 2008). After each pangenome has been put through the app, the output from this program provides an overview for each pangenome object which includes the genome comparisons, gene families, and gene functions found in each MAG. Each pangenome object comparison is manually inspected where we selected gene functions that are found in all genomes within a pangenome object. The gene functions found in all genomes of a pangenome object of one clade are compared to the functions that were found in another pangenome clade. This is to determine which gene functions are similar or unique between pangenome object clades.

We manually compared pangenome objects to one another in the following comparisons: Clade1 and Clade2, Clade2 and Clade3, Clade1 and Clade3. Gene functions in the Pangenome of all 8 AOA MAGs were then compared to these three pangenome comparison groups. The number of gene functions that were found in these comparison groups were then inputted into R to draw a triple venn diagram which displayed the number of gene functions that were found between and within each specific pangenome. The R package used to construct the triple venn diagram was “VennDiagram" where we used the command draw.triple.venn (Chen & Boutros 2011).

3.4 Results

AOA MAG Distribution, Characteristics, and Phylogenetic Relationships in the ETNP

We identified 8 AOA MAGs (Table 3.1) out of a total of 157 MAGs that were assembled from samples collected in different regions of the water column and across four sampling stations. With one exception (Station 3 at 123 m), all of the AOA MAGs were recovered from the secondary chlorophyll maximum (SCM). No AOA MAGs were recovered at Station 4 as this is a non-AMZ/ODZ station and so lacks an SCM. Two AOA MAGs were recovered from the Station 1 SCM at 87m, a single AOA MAG from the Station 2 SCM at 130m, and two AOA MAGs from the Station 3 SCM at 140m. Three AOA MAGs were recovered from the Station 3 OMZ edge sample at 123 m.

MAG 14 in station 1 at 87m, MAG 13 at station 2 at 130m, MAG 9 and 12 from station 3 at 123m, and MAG 18 at station 3 at 140m were identifiable at the genus level, while the remaining MAGs—MAG 32 in station 1 at 87m, MAG 3 at station 2 at 130m, and MAG 21 at station 3 140m—were identifiable at the species level. The highest genus level identified for all AOA MAGs was *Nitrosopelagicus,* while the highest species level of 3 of the AOA MAGs is *Nitrosopumilus* sp002730325. The most complete MAG was MAG 21 at station 3 at 140m, with a genome completeness of 98 percent and a contamination of 0 percent. MAG 21 also had a GC content of 0.32 and a genome size of 1,176,053 base pairs. The MAG with the lowest completion was MAG 14 with a completion of 77.5 percent and a contamination of 0.97 percent.

Table 3.1: AOA MAG general overview

Station	Depth (m)	Sample Type	MAG bin #	Completeness (%)	Contamination (%)	GC content	Size (bp)	QTDB Taxonomy
1	87.5	SCM	bin.014	77.35	0.97	0.3395	881,315	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumilaceae;g__Nitrosopelagicus;s__
1	87.5	SCM	bin.032	97.09	1.78	0.3198	1,172,122	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumilaceae;g__Nitrosopumilus;s__Nitrosopumilus sp002730325
2	130	SCM	bin.013	91.67	0	0.3386	1,015,864	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumilaceae;g__Nitrosopelagicus;s__
3	123	OMZ Edge	bin.003	90.94	0.97	0.3196	1,115,814	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumilaceae;g__Nitrosopumilus;s__Nitrosopumilus sp002730325
3	123	OMZ Edge	bin.009	86.97	1.46	0.3398	1,095,377	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumila

								ceae;g__Nitrosopelagicus;s__
3	123	OMZ Edge	bin.012	80.37	0	0.3262	835,247	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumilaceae;g__Nitrosopelagicus;s__
3	140	SCM	bin.018	91.26	0	0.3382	1,019,889	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumilaceae;g__Nitrosopelagicus;s__
3	140	SCM	bin.021	98.06	0	0.3190	1,176,053	d__Archaea;p__Thermoproteota;c__Nitrososphaeria;o__Nitrososphaerales;f__Nitrosopumilaceae;g__Nitrosopumilus;s__Nitrosopumilus sp002730325

We compared the phylogenetic relationships among these MAGs in three ways. First, we compared only our 8 ETNP AOA MAGs using 49 COGs; second, we compared our AOA MAGs with 21 extant archaeal genomes based on 49 COGs; and third, we compared our 8 AOA MAGs with 12 existing AOA genomes based only on the *amoA* gene. All of these approaches yielded the same consistent pattern, with three major AOA clades recovered from our ETNP dataset (Figure 3.1). Clade1 contained MAG 3 at station 3 at 123m, MAG 21 at station 3 at 140m, and MAG 32 in station 1 at 87m, while Clade2 consisted of MAG 18 at station 3 at 140m, MAG 14 at station 1 at 87m, MAG 9 at station 3 at 123m, and MAG 13 at station 2 at 130m. The last clade, Clade3, comprised only MAG 12 at station 3 at 123m.

Comparison with available genomes showed that this single MAG Clade3 was most closely related to the AOA *Nitrosopelagicus brevius*—which is characterized by a particularly streamlined genome (Santoro et al. 2015). Clade1 was most closely related to *Nitrosopumilus* genomes based on 49 COGs and *amoA*. Clade2 was more closely related to *Nitrososphaera* and *Nitrosoarchaeum* genomes based on *amoA* alone, but formed a distinct clade based especially on 49 COGs and comparison with 21 genomes. This approach based on 49 COGs is obviously more detailed and robust than reliance on a single gene.

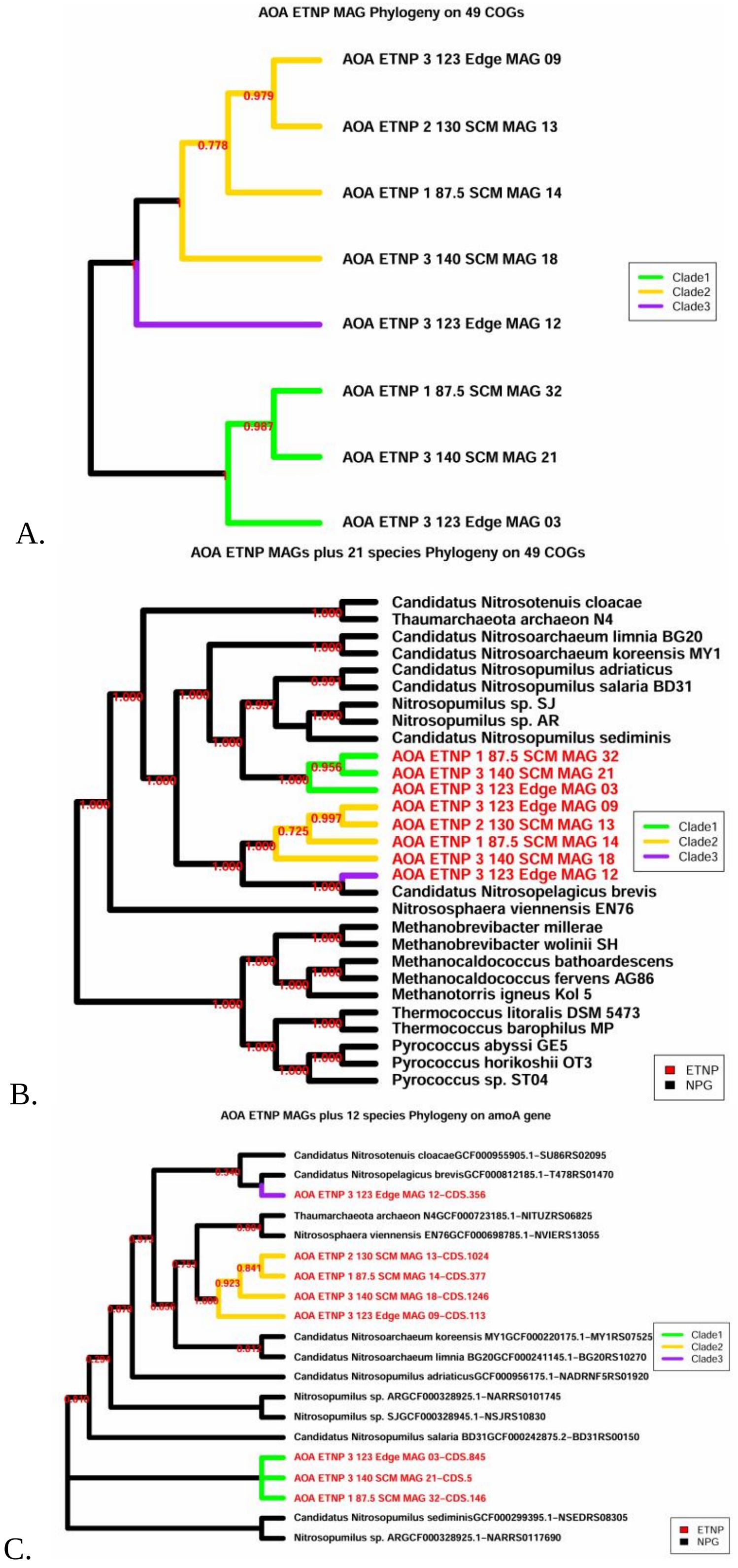
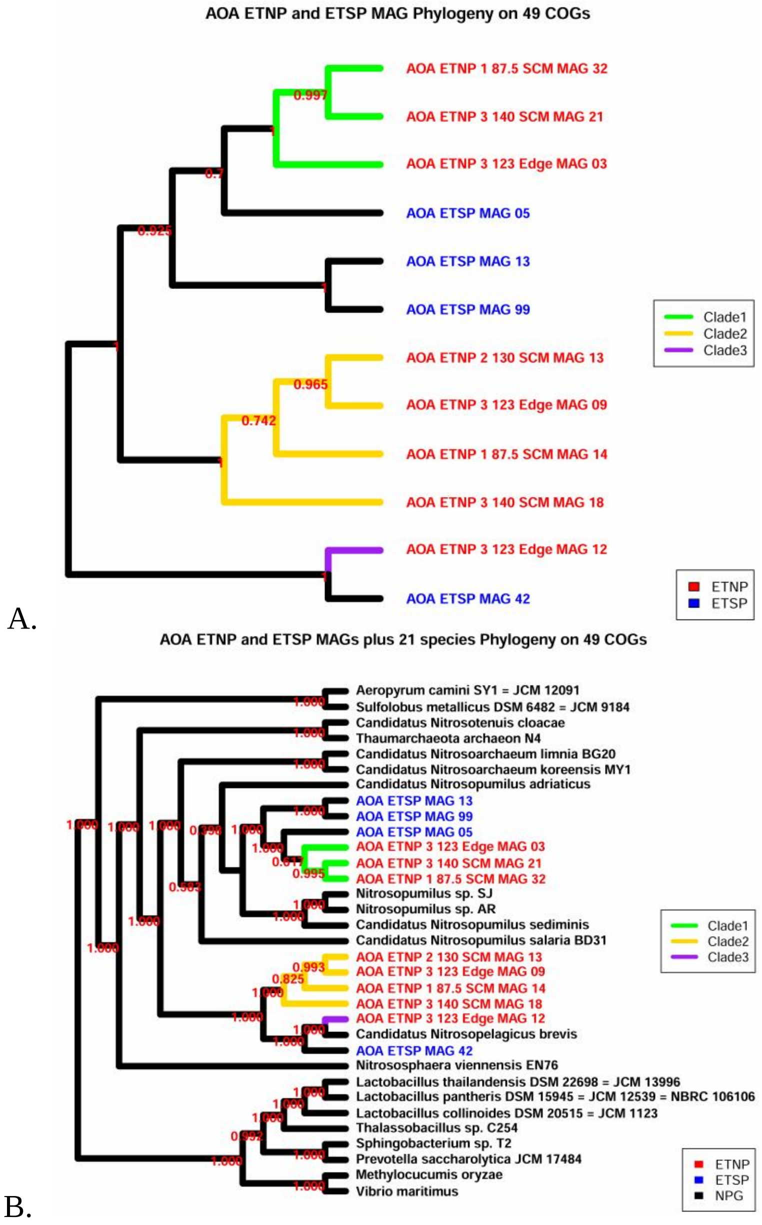
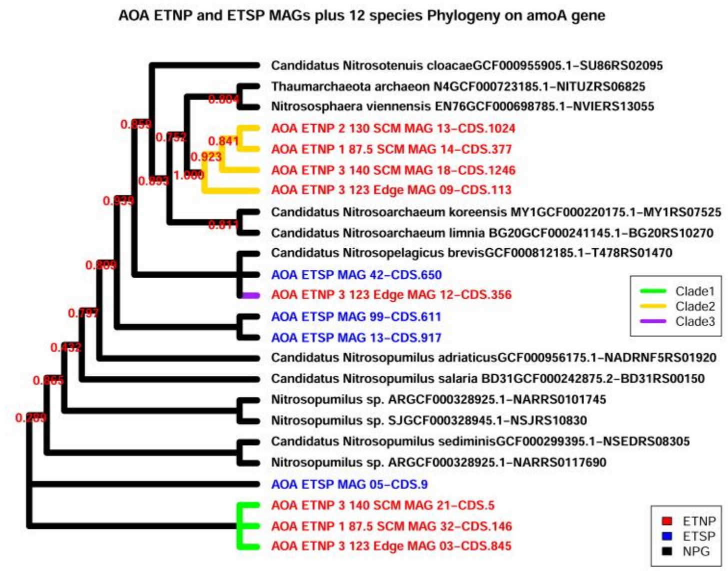


Figure 3.1: Phylogenies on AOA MAGs in the ETNP OMZ. A) Phylogeny of the 8 AOA MAGs constructed on 49 COGs B) Phylogeny of the 8 AOA MAGs and 21 public archaeal genomes

constructed on 49 COGs. C) Phylogeny of the AOA MAGs along with 12 AOA genomes built on the *amoA* gene. In each phylogeny the AOA ETNP MAGs form 3 distinct clades, with Clade1 shown in green, Clade2 in gold, and Clade3 in purple. Phylogeny tip labels are color coded, AOA ETNP MAG tip labels are black in A, and red in B and C. Neighbor Public Genomes (NPG) tip labels are black.


AOA ETNP and ETSP MAG Phylogeny on 49 COGs
AOA ETNP 1 87.5 SCM MAG 32
AOA ETNP 3 140 SCM MAG 21
AOA ETNP 3 123 Edge MAG 03
AOA ETSP MAG 05
AOA ETSP MAG 13
AOA ETSP MAG 99
AOA ETNP 2 130 SCM MAG 13
AOA ETNP 3 123 Edge MAG 09
AOA ETNP 1 87.5 SCM MAG 14
AOA ETNP 3 140 SCM MAG 18
AOA ETNP 3 123 Edge MAG 12
AOA ETSP MAG 42
0.997
0.7
0.925
0.965
0.742
1
Clade1
Clade2
Clade3
ETNP
ETSP
A.
AOA ETNP and ETSP MAGs plus 21 species Phylogeny on 49 COGs
Aeropyrum camini SY1 = JCM 12091
Sulfolobus metallicus DSM 6482 = JCM 9184
Candidatus Nitrosotenuis cloacae
Thaumarchaeota archaeon N4
Candidatus Nitrosoarchaeum limnia BG20
Candidatus Nitrosoarchaeum koreensis MY1
Candidatus Nitrosopumilus adriaticus
AOA ETSP MAG 13
AOA ETSP MAG 99
AOA ETSP MAG 05
AOA ETNP 3 123 Edge MAG 03
AOA ETNP 3 140 SCM MAG 21
AOA ETNP 1 87.5 SCM MAG 32
Nitrosopumilus sp. SJ
Nitrosopumilus sp. AR
Candidatus Nitrosopumilus sediminis
Candidatus Nitrosopumilus salaria BD31
AOA ETNP 2 130 SCM MAG 13
AOA ETNP 3 123 Edge MAG 09
AOA ETNP 1 87.5 SCM MAG 14
AOA ETNP 3 140 SCM MAG 18
AOA ETNP 3 123 Edge MAG 12
Candidatus Nitrosopelagicus brevis
AOA ETSP MAG 42
Nitrososphaera viennensis EN76
Lactobacillus thailandensis DSM 22698 = JCM 13996
Lactobacillus pantheris DSM 15945 = JCM 12539 = NBRC 106106
Lactobacillus collinoides DSM 20515 = JCM 1123
Thalassobacillus sp. C254
Sphingobacterium sp. T2
Prevotella saccharolytica JCM 17484
Methylocucumis oryzae
Vibrio maritimus
1.000
0.398
0.617
0.995
0.583
0.993
0.825
0.992
Clade1
Clade2
Clade3
ETNP
ETSP
NPG
B.


AOA ETNP and ETSP MAGs plus 12 species Phylogeny on amoA gene
Candidatus Nitrosotenuis cloacaeGCF000955905.1-SU86RS02095
Thaumarchaeota archaeon N4GCF000723185.1-NITUZRS06825
Nitrososphaera viennensis EN76GCF000698785.1-NVIERS13055
AOA ETNP 2 130 SCM MAG 13-CDS.1024
AOA ETNP 1 87.5 SCM MAG 14-CDS.377
AOA ETNP 3 140 SCM MAG 18-CDS.1246
AOA ETNP 3 123 Edge MAG 09-CDS.113
Candidatus Nitrosoarchaeum koreensis MY1GCF000220175.1-MY1RS07525
Candidatus Nitrosoarchaeum limnia BG20GCF000241145.1-BG20RS10270
Candidatus Nitrosopelagicus brevisGCF000812185.1-T478RS01470
AOA ETSP MAG 42-CDS.650
AOA ETNP 3 123 Edge MAG 12-CDS.356
AOA ETSP MAG 99-CDS.611
AOA ETSP MAG 13-CDS.917
Candidatus Nitrosopumilus adriaticusGCF000956175.1-NADRNF5RS01920
Candidatus Nitrosopumilus salaria BD31GCF000242875.2-BD31RS00150
Nitrosopumilus sp. ARGCF000328925.1-NARRS0101745
Nitrosopumilus sp. SJGCF000328945.1-NSJRS10830
Candidatus Nitrosopumilus sediminisGCF000299395.1-NSEDRS08305
Nitrosopumilus sp. ARGCF000328925.1-NARRS0117690
AOA ETSP MAG 05-CDS.9
AOA ETNP 3 140 SCM MAG 21-CDS.5
AOA ETNP 1 87.5 SCM MAG 32-CDS.146
AOA ETNP 3 123 Edge MAG 03-CDS.845
0.804
0.859
0.752
0.841
0.923
0.893
1.000
0.939
0.811
0.809
0.797
0.432
0.865
0.289
Clade1
Clade2
Clade3
ETNP
ETSP
NPG

C.

Figure 3.2: Phylogenies on AOA MAGs in ETNP and from the ETSP (Sun and Ward 2021). A) Phylogeny of only the 12 AOA MAGs from ETNP and ETSP constructed on 49 COGs. B) Phylogeny of the 12 AOA MAGs from the ETNP and ESTP and 21 public archaeal genomes constructed on 49 COGs. C) Phylogeny of the ETNP and ETSP AOA MAGs along with 12 AOA genomes built on the *amoA* gene. ETNP AOA MAG Clade1 shown in green, Clade2 in gold, and Clade3 in purple, while ETSP AOA MAGs are shown in blue. Phylogeny tip labels are color coded, AOA ETNP MAGs are red, AOA ETSP MAGs are blue, Neighbor Public Genomes (NPG) are black.

We further compared our ETNP AOA MAGs with 4 AOA MAGs previously recovered from the ETSP (Sun & Ward 2021). The same set of three comparisons recapitulated some of the previous patterns and identified several new insights (Figure 3.2). First, ETSP MAG 42 was most similar to our Clade3 and *Nitrosopelagius brevius*, providing support for the presence of this group in the major low oxygen regions of the ocean. Based on 49 COGs, the remaining ETSP MAGs were clustered with our ETNP Clade1 and *Nitrosopumilus* genomes, indicating further similarities between the ETNP and ETSP. However, none of the previously recovered ETSP MAGs clustered with our Clade2, represented by 4 different ETNP AOA MAGs.

Functional Differences Encoded in ETNP AOA MAGs

Phylogenetic differences among ETNP AOA MAGs may also be indicative of functional differences, which we evaluated in several ways. We first generated DRAM metabolic summaries of the 8 AOA MAGs (Figure 3.3). As expected, AOA MAGs did not contain CAZys, sulfur metabolism, photosynthesis, and other reductases, but did contain genes for nitrogen metabolism—especially and obviously genes coding ammonia oxidation. Our ETNP AOA MAGs also contained the *nirK* gene that helps convert nitrite to nitric oxide, and which has been widely observed in other AOA (Francis et al. 2007; Lund et al. 2012; Kraft et al. 2022). All the AOA genomes contained genes for alcohol production, but only MAG 12 from station 3 at 123m, MAG 13 at station 2 at 130m, and MAG 14 in station 1 at 87m dids not contain genes for converting pyruvate (pyruvate =>

acetyl CoA v2). AOA MAGs also contained several modules with a high percentage including the Reductive pentose phosphate cycle (Calvin cycle), Citrate cycle (TCA cycle, Krebs cycle), Pentose phosphate pathway (Pentose phosphate cycle) and Glycolysis (Embden-Meyerhof pathway glucose => pyruvate) which all hover around 50 percent. In terms of the electron transport chain, the Complex V: V/A-type ATPase stand out and were mostly complete in MAG 21 at station 3 at 140m and MAG 32 in station 1 at 87m, with a completion of 88 percent.

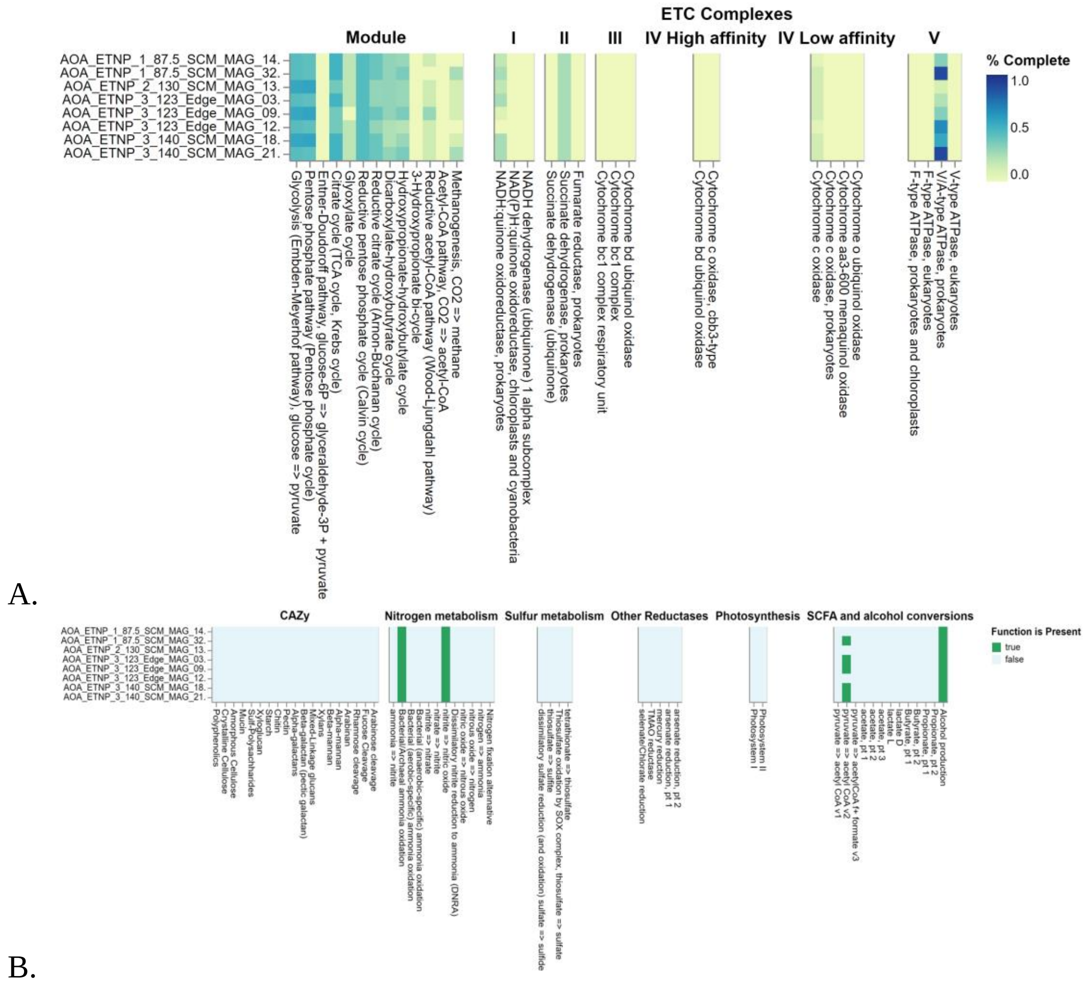


Figure 3.3: DRAM Metabolic summary on all 8 AOA MAGs. A) Metabolic summary on Electron Transport Chain complex modules present within each AOA MAG represented by multiple small heatmaps. Heatmap scaling is based on percent completion, where dark blue is higher percentage and light green is low percentage. B) The bottom half of the figure displays the presence or absence of genes in different metabolic processes among the AOA MAGs. Green indicates genes are present, whereas light blue represents genes that are absent.

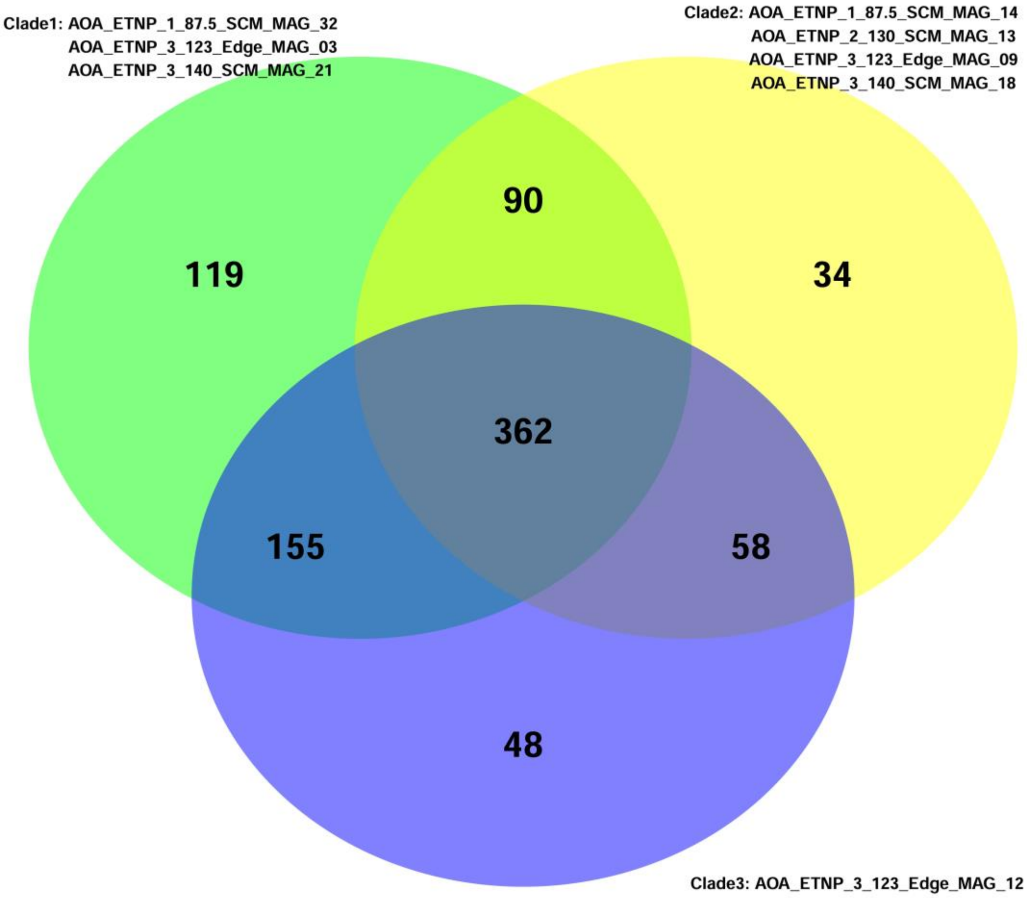

Clade1: AOA_ETNP_1_87.5_SCM_MAG_32
AOA_ETNP_3_123_Edge_MAG_03
AOA_ETNP_3_140_SCM_MAG_21
Clade2: AOA_ETNP_1_87.5_SCM_MAG_14
AOA_ETNP_2_130_SCM_MAG_13
AOA_ETNP_3_123_Edge_MAG_09
AOA_ETNP_3_140_SCM_MAG_18
90
119
34
362
155
58
48
Clade3: AOA_ETNP_3_123_Edge_MAG_12

Figure 3.4: Triple Venn Diagram on gene functions compared amongst the AOA pangenomes. MAGs are grouped into three clades based on phylogenetic similarity/relatedness. Clade1 is green, Clade2 is yellow, and Clade3 is purple.

We also examined gene functions that are shared among the AOA MAGs and unique to the MAG Pangenome groups. The pangenome objects are MAGs grouped based on their consistent placement on the previous phylogenetics trees. All three clades shared a total of 362 gene functions (Figure 3.4). Clade1 contained 119 unique gene functions, Clade2 contained 34 unique gene functions, and Clade3 contained 48 unique gene functions. 90 gene functions were shared among Clades 1 and 2, 155 between Clades 1 and 3, and 58 functions between Clades 2 and 3. Clade1 contained the highest number of gene functions overall, while the clade with the lowest number of gene functions was Clade2. Core Genes shared across all genomes included Cobalamin synthase, Cob(I)alamin adenosyltransferase, 3-hydroxyacyl-CoA dehydrogenase, Ammonia monooxygenase, Nitric oxide reductase activation protein NorQ, Putative sodium:solute symporter, urea transporter DUR3, and Pyruvate,phosphate dikinase. Core gene functions shared in Clades 1 & 2 included Menaquinone via futalosine polyprenyltransferase. For Clades 1 & 3, shared gene functions included Arsenate reductase (EC 1.20.4.4), thioredoxin-coupled, LMWP family and Phage tail fiber protein.

Finally, a key gene function shared in Clades 2 & 3 included Urease alpha subunit, while a number of gene functions were obviously unique to each clade (Supplementary Information). Pangenome comparison therefore indicates genomic differences among the AOA genomes, as some core gene functions were only shared within specific clades.

3.5 Discussion

Distribution of and Phylogenetic Relationships Among ETNP AOA MAGs

We assembled 8 AOA MAGs from the ocean's largest OMZ in the ETNP. These MAGs were recovered from SCM samples collected at three different sampling stations, as well as from a sample located on the edge of the OMZ at Station 3. The presence of AOA MAGs in the SCM is consistent with the idea that they may utilize oxygen produced via photosynthesis by specialized *Prochlorococcus* groups that form this feature (Garcia-Robledo et al. 2017), as well as the expectation of high oxygen affinity among AOA (Bristow et al. 2016b). We did not recover MAGs from PNM samples, where AOA may be expected to be active, although this may reflect the diversity of organisms found at this depth in the ETNP. This contrasts with the ETSP, where AOA MAGs were recovered in the upper regions of the water and dominate the oxic regions, while having a near absent abundance in deeper depths and anoxic core (Sun and Ward 2021).

In comparison with other archaeal genomes and AOA MAGs from the ETSP (Sun and Ward 2021), we recovered at least three distinct clades in the ETNP. While one of these clades (Clade3) was represented by a single AOA MAG from the ETNP, an ETSP AOA MAG was closely related—indicating that Clade3 is present in both tropical Pacific OMZs. Clade1 also appears to be common in both the ETNP and ETSP based on the similarity among MAGs and the number recovered in both locations. However, our Clade2 appears to be a distinct group present in the ETNP. Whether this reflects a fundamental biogeographical or biogeochemical difference between the ETNP and ETSP, or simply methodological factors (e.g., depth of sampling or sequencing), remains to be determined. Given its evident uniqueness, our ETNP Clade2 deserves additional investigation.

The AOA MAGs identified in the ETNP OMZ and the relationships between these groups are consistent with current metagenomic analysis on AOA across different environments in the water column. For example, the placement of these AOA grouping is consistent with current metagenomic studies on the AOA MAGs across all marine environments, where both *Nitrosopelagicus* and *Nitrosopumilus* form different clades, and *Nitrosopumilus* is more closely related to *Nitrosoarchaeum* (Qin et al. 2020). *Nitrosopelagicus* and N*itrosopumilus* both have also been detected in the OMZ off the coast of Namibia, where in the ETNP both AOA groups were quite abundant in dyoxic waters of the OMZ (Vuillemin 2023). Vuillemin (2023) also constructed a phylogeny which demonstrated that most of the AOA that resided in the pelagic waters consisted of both *Nitrosopelagicus* and *Nitrosopumilus*. Previous work in the ETNP has shown through 16S amplicon sequencing that *Nitrospelagicus* to be more common in the oxic layer instead of the OMZ core (Medina Faull et al. 2020), whereas our MAGs indicate that *Nitrosopelagicus* are present in hypoxic and the anoxic regions of the ETNP OMZ. Collectively these data indicate that these AOA groups are diverse and based on the variable regions of the water column they inhabit, and can survive hypoxic conditions. Interestingly, we found that all three clades co-occurred in the Station 3 OMZ edge sample, while Clades 1 and 2 co-occurred in additional samples (and notably Clade3 was only represented by the one MAG from the Station 3 OMZ edge sample from 123 m depth).

Common Metabolic Functions Encoded in ETNP AOA MAGs

In addition to co-occurrence, analysis of our AOA MAGs showed multiple common genes, pathways, and capabilities across the ETNP AOA MAG Clades. First, all MAGs contained another gene involved in the nitrogen cycle (in addition to the defining presence of the *amoA* gene), the nitrite reductase (*nirK*) gene important for nitrite to nitric oxide (NO) production. *Nitrosopumilus* have been detected with the *nirK* gene in previous studies, although the function of this *nirK* has remained elusive (Francis et al. 2007; Lund et al. 2012). However, recent work has suggested that the *nirK* may function to indirectly produce oxygen in order to continue aerobic ammonia oxidation under anoxic conditions (Kraft et al. 2022). Kraft et al. (2022) hypothesized that AOA reduce nitrite to nitric oxide via the *nirK* gene, that nitric oxide is further reduced into oxygen and nitrous oxide by another enzyme still to be characterized, and that the oxygen produced is then used for aerobic ammonia oxidation (Kraft et al. 2022; Martens-Habbena & Qin 2022). Although this requires that several key biogeochemical conditions are met (Kraft et al. 2022), our data confirm the presence of a *nirK* in all AOA MAGs from the ETNP and could be further queried for the pivotal unknown enzyme involved in the hypothetical oxygen production pathway.

Our MAGs also contained genes involved in the well-known citric acid cycle (TCA cycle, aka Krebs cycle). In previous work, *Nitrosopumilus* have been detected with genes in the TCA cycle, but not enough to be considered complete and used only for biosynthetic purposes (Walker et al. 2010). Another set of genes that our AOA MAGs possess are involved in the Pentose phosphate pathway (Pentose phosphate cycle)—which has been proposed to be heterotrophically convert CO2 and rubisco to create 3-phosphoglycerate to enter the central carbon cycle (Reji & Francis 2020; Sato et al. 2007). AOA MAGs also contained genes involved in Glycolysis (Embden-Meyerhof pathway, glucose => pyruvate) and archaea have been identified with an alternate form of the Embden-Meyerhof pathway shown to be beneficial for anaerobic organisms (Bräsen et al. 2014).

In addition to these pathways, AOA are expected to grow primarily autotrophically and possess a unique carbon fixation pathway, the hydroxypropionate–hydroxybutyrate cycle, as they have been identified with genes encoding all key enzymes of this cycle and absent in other autotrophic pathways (Berg et al. 2010). All of our MAGs included genes involved in this carbon fixation pathway. In relation to OMZs, *Thaumarchaeota* of the Marine Group I have genes associated with the 3-hydroxypropionate/4-hydroxybutyrate cycle found in the upper and lower oxyclines of OMZs (Ruiz-Fernández et al. 2020), which is consistent with our study as these genes have been identified in hydroxypropionate–hydroxybutyrate cycle with low abundance in the AMZ core. However, in the Tropical Atlantic, the accA gene acetyl-CoA/propionyl-CoA carboxylase, important for the 3-hydroxypropionate/4-hydroxybutyrate cycle increases in abundance towards the OMZ (Bergauer et al. 2013). Overall, AOA utilize hydroxypropionate–hydroxybutyrate cycle as it is more efficient and requires less energy compared to the traditional Reductive pentose phosphate (Calvin) cycle (Könneke et al. 2014).

Another common set of genes found in all ETNP AOA MAGs were those involved in cobalamin synthesis. *Thaumarchaeta* have been shown to contain cobalamin synthesis genes such as cob/cbi and are one of the most dominant cobalamin and Vitamin B12 producers globally (Doxey et al. 2015). Cobalamin biosynthesis genes are also a part of the core genome of *Thaumarchaeota*, which comprises 344 genes (Qin et al. 2020). Previous work has suggested that AOA detected in the ETNP ODZ produced nitro-cobalamin under hypoxic and anoxic conditions (Heal et al. 2018). AOA seem to produce nitro-cobalamin under oxygen stress, but also under low amounts of ammonia and copper (Heal et al. 2018). Since AOA in previous studies have been shown to produce some amount of cobalamin, it is consistent with this current study as all 8 AOA genome contain gene related to cobalamin, especially as these are AOA that inhabit anoxic regions of the water column of the ETNP OMZ in the SCM and OMZ edge. This also further reinstates the importance of cobalamin, as it is important vitamin found in all organisms (Heal et al. 2018)

Additional core genes shared across MAGs included Putative sodium:solute symporter, similarity with yeast urea transporter DUR3 and Pyruvate,phosphate dikinase. A previous study examined complete genomes of AOA that contained the dur3 gene in both *Nitrosopumilus* and *Nitrosopelagicus* (Liu et al. 2023), and which were collected from extreme marine environments and other aquatic locations. The reason why the AOA MAGs contain these genes is because they help transfer urea to satisfy nitrogen demands (Liu et al. 2023; Kitzinger et al. 2020). Pyruvate phosphate dikinase have also been previously identified in AOA, where they may be used in the transformation of pyruvate to phosphoenolpyruvate for gluconeogenesis (Park et al. 2014). Finally, nitric oxide reductase activation protein NorQ has been identified previously in Nitrosopelgaicus, but these genomes were retrieved from the ocean surface instead of the ODZ (Santoro et al. 2015). This gene is important as it helps activate the norB gene, which reduces NO to nitrous oxide (N_2O), producing a potent greenhouse (Santoro et al. 2015; Braker & Tiedje 2003). Our data indicate that this capability is present in all AOA MAGs in the ETNP.

In addition to these core genes, pangenome analysis also indicated that pairs of clades shared genes there were not detected in the other clade (with the caveat that only one MAG was recovered from Clade3). For example, genes in common between Clades 1 & 2 but absent in Clade3 included Menaquinone via futalosine polyprenyltransferase (MenA homolog), Nitrosopumilus/Caldivirga type. Futalosine pathway is important as it helps in menaquinone synthesis, an important vitamin (Zhi et al. 2014). Archaea have acquired this gene in menaquinone through horizontal gene transfer. This is understandable as menaquinone are involved in anaerobic electron transport systems and serves as an electron carrier (Unden et al. 2014; Suvarna et al. 1998). An interesting commonality between Clades 1 and 3 but absent in Clade2 was an arsenate reductase thioredoxin-coupled gene. To our knowledge, no known AOA genomes have been identified with arsenate reductase genes. However, arsenic resistance genes have been identified in both archaea and bacteria (Jackson et al. 2003), and the role of the arsC gene is for arsenate resistance, by transforming arsenate into arsenite (Ranganathan, et al. 2023). Once they transform arsenate, they can then expel it. Interestingly, the ETNP was recently identified as a site of active arsenic cycling, with genes encoding both subunits of the respiratory arsenite oxidase (*AioA*) and the dissimilatory arsenate reductase (*ArrA)* (Saunders et al. 2019). These genes were not found or detected in AOA, but our MAGs raise this possibility.

Finally, we found gene functions involved in nitrogen cycling that were present in Clade 2 & 3 MAGs which includes urease alpha subunit. *Thaumarchaeota* previously have been identified with *ureC* gene (and contained other subunits of the urease enzyme), although this does not appear to be universal across all AOA (Kitzinger et al. 2019, Qin et al. 2024). The *ureC* gene is often accompanied with the *dur3* gene, since urease breaks down the transported urea into ammonia and carbon dioxide as a source for nitrogen and carbon (Liu et al. 2023; KItzinger et al. 2020). Our data indicate that different coexisting clades of AOA may have varying abilities to use urea.

In conclusion, we recovered *Nitrosopelagicus* and *Nitrosopumilus* MAGs from hypoxic and anoxic regions of the ETNP OMZ. The presence of these AOA groups under these conditions indicates that AOA are potentially adapted to low oxygen, and still able to perform ammonia oxidation along with other metabolic functions to continue autotrophy. Both phylogenomic and pangenome analysis showed that AOA are also diverse in core genes that are unique to certain MAG groups, and within 3 distinct clades that are formed amongst the MAGs. What seems to be the most diverse groups of MAGs are AOA found in station 3 at 123, as they are present in all three clades, and contain core genes unique in all clades. This study also shows that AOA are genomically unique to the habitat in which they reside, as MAGs from the ETNP are often their own clades that are separate from the MAGs in the ETSP OMZ. AOA are quite diverse, and future studies will further reveal how AOA are able to survive these harsh environmental conditions.

3.6 References

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., ... & Quince, C. (2013). CONCOCT: clustering contigs on coverage and composition. arXiv preprint arXiv:1312.4038.

Aylward, F. O., & Santoro, A. E. (2020). Heterotrophic Thaumarchaea with small genomes are widespread in the dark ocean. Msystems, 5(3), 10-1128.

Andrews, S. (2010, April). *FastQC: a quality control tool for high throughput sequence data*.

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., ... & Yu, D. (2018). KBase: the United States department of energy systems biology knowledgebase. Nature biotechnology, 36(7), 566-569.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., ... & Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, *9*, 1-15.

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.

Beman, J. M., & Carolan, M. T. (2013). Deoxygenation alters bacterial diversity and community composition in the ocean’s largest oxygen minimum zone. Nature Communications, 4(1), 2705.

Beman, J. M., Leilei Shih, J., & Popp, B. N. (2013). Nitrite oxidation in the upper water column and oxygen minimum zone of the eastern tropical North Pacific Ocean. The ISME journal, 7(11), 2192-2205.

Beman, J. M., Popp, B. N., & Alford, S. E. (2012). Quantification of ammonia oxidation rates and ammonia-oxidizing archaea and bacteria at high resolution in the Gulf of California and eastern tropical North Pacific Ocean. Limnology and Oceanography, 57(3), 711-726.

Beman, J. M., Vargas, S. M., Wilson, J. M., Perez-Coronel, E., Karolewski, J. S., Vazquez, S., ... & Wankel, S. D. (2021). Substantial oxygen consumption by aerobic nitrite oxidation in oceanic oxygen minimum zones. Nature Communications, 12(1), 7043.

Berg, I. A., Kockelkorn, D., Buckel, W., & Fuchs, G. (2007). A 3-hydroxypropionate/4-hydroxybutyrate autotrophic carbon dioxide assimilation pathway in Archaea. Science, 318(5857), 1782-1786.

Berg, I. A., Kockelkorn, D., Ramos-Vera, W. H., Say, R. F., Zarzycki, J., Hügler, M., ... & Fuchs, G. (2010). Autotrophic carbon fixation in archaea. Nature Reviews Microbiology, 8(6), 447-460.

Bergauer, K., Sintes, E., van Bleijswijk, J., Witte, H., & Herndl, G. J. (2013). Abundance and distribution of archaeal acetyl-CoA/propionyl-CoA carboxylase genes indicative for putatively chemoautotrophic Archaea in the tropical Atlantic's interior. FEMS microbiology ecology, 84(3), 461-473.

Bertagnolli, A. D., & Stewart, F. J. (2018). Microbial niches in marine oxygen minimum zones. Nature Reviews Microbiology, 16(12), 723-729.

Braker, G., & Tiedje, J. M. (2003). Nitric oxide reductase (norB) genes from pure cultures and environmental samples. Applied and environmental microbiology, 69(6), 3476-3483.

Bräsen, C., Esser, D., Rauch, B., & Siebers, B. (2014). Carbohydrate metabolism in Archaea: current insights into unusual enzymes and pathways and their regulation. Microbiology and Molecular Biology Reviews, 78(1), 89-175.

Bristow, L., Callbeck, C., Larsen, M. et al. N2 production rates limited by nitrite availability in the Bay of Bengal oxygen minimum zone. Nature Geosci 10, 24–29 (2016a).

Bristow, L. A., Dalsgaard, T., Tiano, L., Mills, D. B., Bertagnolli, A. D., Wright, J. J., ... & Thamdrup, B. (2016b). Ammonium and nitrite oxidation at nanomolar oxygen concentrations in oxygen minimum zone waters. Proceedings of the National Academy of Sciences, 113(38), 10601-10606.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. Nature methods, 12(1), 59-60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. BMC bioinformatics, 10, 1-9.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution, 17(4), 540-552.

Chaumeil, P. A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database.

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC bioinformatics, 12, 1-7.

Chivian, D., Jungbluth, S. P., Dehal, P. S., Wood-Charlson, E. M., Canon, R. S., Allen, B. H., ... & Arkin, A. P. (2023). Metagenome-assembled genome extraction and analysis from microbiomes using KBase. Nature Protocols, 18(1), 208-238.

Doxey, A. C., Kurtz, D. A., Lynch, M. D., Sauder, L. A., & Neufeld, J. D. (2015). Aquatic metagenomes implicate Thaumarchaeota in global cobalamin production. The ISME journal, 9(2), 461-471.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS computational biology*, *7*(10), e1002195.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5), 1792-1797.

Francis, C. A., Roberts, K. J., Beman, J. M., Santoro, A. E., & Oakley, B. B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proceedings of the National Academy of Sciences, 102(41), 14683-14688.

Francis, C. A., Beman, J. M., & Kuypers, M. M. (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. The ISME journal, 1(1), 19-27.

Füssel, J., Lam, P., Lavik, G., Jensen, M. M., Holtappels, M., Günter, M., & Kuypers, M. M. (2012). Nitrite oxidation in the Namibian oxygen minimum zone. The ISME journal, 6(6), 1200-1209.

Garcia-Robledo, E., Padilla, C. C., Aldunate, M., Stewart, F. J., Ulloa, O., Paulmier, A., ... & Revsbech, N. P. (2017). Cryptic oxygen cycling in anoxic marine zones. Proceedings of the National Academy of Sciences, 114(31), 8319-8324.

Gilly, W. F., Beman, J. M., Litvin, S. Y., & Robison, B. H. (2013). Oceanographic and biological effects of shoaling of the oxygen minimum zone. Annual review of marine science, 5(1), 393-420.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., & Beck, E. (2012). TIGRFAMs and genome properties in 2013. Nucleic acids research, 41(D1), D387-D395.

Heal, K. R., Qin, W., Amin, S. A., Devol, A. H., Moffett, J. W., Armbrust, E. V., ... & Ingalls, A. E. (2018). Accumulation of NO2-cobalamin in nutrient-stressed ammonia-oxidizing archaea and in the oxygen deficient zone of the eastern tropical North Pacific. Environmental microbiology reports, 10(4), 453-457.

Jackson, C. R., & Dugas, S. L. (2003). Phylogenetic analysis of bacterial and archaeal arsC gene sequences suggests an ancient, common origin for arsenate reductase. BMC evolutionary biology, 3, 1-10.

Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ, 3, e1165.

Karner, M. B., DeLong, E. F., & Karl, D. M. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature, 409(6819), 507-510.

Kitzinger, K., Marchant, H. K., Bristow, L. A., Herbold, C. W., Padilla, C. C., Kidane, A. T., ... & Kuypers, M. M. (2020). Single cell analyses reveal contrasting life strategies of the two main nitrifiers in the ocean. Nature Communications, 11(1), 767.

Kitzinger, K., Padilla, C. C., Marchant, H. K., Hach, P. F., Herbold, C. W., Kidane, A. T., ... & Bristow, L. A. (2019). Cyanate and urea are substrates for nitrification by Thaumarchaeota in the marine environment. Nature microbiology, 4(2), 234-243.

Kraft, B., Jehmlich, N., Larsen, M., Bristow, L. A., Könneke, M., Thamdrup, B., & Canfield, D. E. (2022). Oxygen and nitrogen production by an ammonia-oxidizing archaeon. Science, 375(6576), 97-100.

Könneke, M., Bernhard, A. E., de La Torre, J. R., Walker, C. B., Waterbury, J. B., & Stahl, D. A. (2005). Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature, 437(7058), 543-546.

Könneke, M., Schubert, D. M., Brown, P. C., Hügler, M., Standfest, S., Schwander, T., ... & Berg, I. A. (2014). Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. Proceedings of the National Academy of Sciences, 111(22), 8239-8244.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), 2178-2189.

Liu, Q., Chen, Y., & Xu, X. W. (2023). Genomic insight into strategy, interaction and evolution of nitrifiers in metabolizing key labile-dissolved organic nitrogen in different environmental niches. Frontiers in Microbiology, 14, 1273211.

Lund, M. B., Smith, J. M., & Francis, C. A. (2012). Diversity, abundance and expression of nitrite reductase (nirK)-like genes in marine thaumarchaea. The ISME Journal, 6(10), 1966-1977.

Martens-Habbena, W., & Qin, W. (2022). Archaeal nitrification without oxygen. Science, 375(6576), 27-28.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: paired-end assembler for illumina sequences. BMC bioinformatics, 13, 1-7.

Medina Faull, L., Mara, P., Taylor, G. T., & Edgcomb, V. P. (2020). Imprint of trace dissolved oxygen on prokaryoplankton community structure in an oxygen minimum zone. Frontiers in Marine Science, 7, 360.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. Current opinion in genetics & development, 15(6), 589-594.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome research, 27(5), 824-834.

Ouverney, C. C., & Fuhrman, J. A. (2000). Marine planktonic archaea take up amino acids. Applied and environmental microbiology, 66(11), 4829-4833.

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., ... & Stevens, R. (2014). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic acids research, 42(D1), D206-D214.

Parada, A. E., Mayali, X., Weber, P. K., Wollard, J., Santoro, A. E., Fuhrman, J. A., ... & Dekas, A. E. (2023). Constraining the composition and quantity of organic matter used by abundant marine Thaumarchaeota. Environmental Microbiology, 25(3), 689-704.

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics, 20(2), 289-290.

Park, S. J., Ghai, R., Martín-Cuadrado, A. B., Rodríguez-Valera, F., Chung, W. H., Kwon, K., ... & Rhee, S. K. (2014). Genomes of two new ammonia-oxidizing archaea enriched from deep marine sediments. PLoS One, 9(5), e96449.

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, 25(7), 1043-1055.

Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. Progress in oceanography, 80(3-4), 113-128.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one, 5(3), e9490.

Qin, W., Zheng, Y., Zhao, F., Wang, Y., Urakawa, H., Martens-Habbena, W., ... & Ingalls, A. E. (2020). Alternative strategies of nutrient acquisition and energy conservation map to the biogeography of marine ammonia-oxidizing archaea. The ISME Journal, 14(10), 2595-2609.

Qin, W., Wei, S. P., Zheng, Y., Choi, E., Li, X., Johnston, J., ... & Winkler, M. K. H. (2024). Ammonia-oxidizing bacteria and archaea exhibit differential nitrogen source preferences. Nature Microbiology, 9(2), 524-536.

Ranganathan, S., Sethi, D., Kasivisweswaran, S., Ramya, L., Priyadarshini, R., & Yennamalli, R. M. (2023). Structural and functional mapping of ars gene cluster in Deinococcus indicus DR1. Computational and Structural Biotechnology Journal, 21, 519-534.

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., ... & Ravel, J. (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. Journal of bacteriology, 190(20), 6881-6893.

Reji, L., & Francis, C. A. (2020). Metagenome-assembled genomes reveal unique metabolic adaptations of a basal marine Thaumarchaeota lineage. The ISME Journal, 14(8), 2105-2115.

Ren, M., Feng, X., Huang, Y., Wang, H., Hu, Z., Clingenpeel, S., ... & Luo, H. (2019). Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. The ISME journal, 13(9), 2150-2161.

Ruiz-Fernández, P., Ramírez-Flandes, S., Rodríguez-León, E., & Ulloa, O. (2020). Autotrophic carbon fixation pathways along the redox gradient in oxygen-depleted oceanic waters. Environmental microbiology reports, 12(3), 334-341.

Santoro, A. E., Dupont, C. L., Richter, R. A., Craig, M. T., Carini, P., McIlvin, M. R., ... & Saito, M. A. (2015). Genomic and proteomic characterization of "Candidatus Nitrosopelagicus brevis": an ammonia-oxidizing archaeon from the open ocean. Proceedings of the National Academy of Sciences, 112(4), 1173-1178.

Santoro, A. E., Richter, R. A., & Dupont, C. L. (2019). Planktonic marine archaea. Annual review of marine science, 11(1), 131-158.

Sato, T., Atomi, H., & Imanaka, T. (2007). Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science, 315(5814), 1003-1006.

Saunders, J. K., Fuchsman, C. A., McKay, C., & Rocap, G. (2019). Complete arsenic-based respiratory cycle in the marine microbial communities of pelagic oxygen-deficient zones. Proceedings of the National Academy of Sciences, 116(20), 9925-9930.

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., ... & White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic acids research, 35(suppl_1), D260-D264.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., ... & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic acids research, 48(16), 8883-8900.

Sieber, C. M., 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.

Stewart, F. J., Ulloa, O., & DeLong, E. F. (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. Environmental microbiology, 14(1), 23-40.

Stramma, L., Johnson, G. C., Sprintall, J., & Mohrholz, V. (2008). Expanding oxygen-minimum zones in the tropical oceans. science, 320(5876), 655-658.

Sun, X., & Ward, B. B. (2021). Novel metagenome-assembled genomes involved in the nitrogen cycle from a Pacific oxygen minimum zone. ISME communications, 1(1), 26.

Suvarna, K., Stevenson, D., Meganathan, R., & Hudspeth, M. E. S. (1998). Menaquinone (vitamin K2) biosynthesis: localization and characterization of the menA gene from Escherichia coli. Journal of bacteriology, 180(10), 2782-2787.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., & Koonin, E. V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic acids research, 28(1), 33-36.

Thamdrup, B., Dalsgaard, T., & Revsbech, N. P. (2012). Widespread functional anoxia in the oxygen minimum zone of the Eastern South Pacific. Deep Sea Research Part I: Oceanographic Research Papers, 65, 36-45.

Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M., & Stewart, F. J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences, 109(40), 15996-16003.

Unden, G., Steinmetz, P. A., & Degreif-Dünnwald, P. (2014). The aerobic and anaerobic respiratory chain of Escherichia coli and Salmonella enterica: enzymes and energetics. EcoSal Plus, 6(1), 10-1128.

Vuillemin, A. (2023). Nitrogen cycling activities during decreased stratification in the coastal oxygen minimum zone off Namibia. Frontiers in Microbiology, 14, 1101902.

Vuillemin, A., Wankel, S. D., Coskun, Ö. K., Magritsch, T., Vargas, S., Estes, E. R., ... & Orsi, W. D. (2019). Archaea dominate oxic subseafloor communities over multimillion-year time scales. Science Advances, 5(6), eaaw4108.

Walker, C. B., de la Torre, J. R., Klotz, M. G., Urakawa, H., Pinel, N., Arp, D. J., ... & Stahl, D. (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proceedings of the National Academy of Sciences, 107(19), 8818-8823.

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.

Zakem, E. J., Mahadevan, A., Lauderdale, J. M., & Follows, M. J. (2020). Stable aerobic and anaerobic coexistence in anoxic marine zones. The ISME Journal, 14(1), 288-301.

Zhi, X. Y., Yao, J. C., Tang, S. K., Huang, Y., Li, H. W., & Li, W. J. (2014). The futalosine pathway played an important role in menaquinone biosynthesis during early prokaryote evolution. Genome biology and evolution, 6(1), 149-160.

Conclusion and Future Directions

In conclusion, my dissertation showed that both microbes and viruses are diverse, thrive in extreme anoxic conditions, and contain different metabolic capabilities in the biogeochemically important ETNP OMZ (Ulloa et al. 2012; Long et al. 2021; Wright et al. 2012). OMZs are complex environments (Paulmier & Ruiz-Pino 2009) and in my first chapter we provided updated measurements of the physiochemical properties and environmental factors in the ETNP OMZ. We were also able to identify which OMZs were classified as AMZs, and measured variation of SNM, SCM, PNM, PCM, and OMZ Edge in all OMZ/AMZ stations. In the ETNP OMZ microbial communities and microbial genes abundances were identified with shotgun metagenomic reads (Ruscheweyh et al. 2022; Kim et al. 2016) and built upon previous research that targeted 16S rRNA or functional genes (Beman & Carolan 2013; Beman et al. 2021). The ETNP OMZ revealed that the biogeochemical properties influenced specific genes and their abundances patterns, as these genes had higher abundance in specific regions of the water column and correlated with key environmental factors. Lastly, in chapter 1, I have assembled 12 new MAGs which mainly consisted of heterotrophs that inhabited the most productive station. These newly constructed MAGs will allow researchers to delve into these genomes and provide a reference for comparison of similar heterotrophs that live in similar or different marine regions. This is how chapter 1 uncovered microbial ecology and functionality through metagenomics, and chapter 2 continued with viruses that also inhabit the ETNP OMZ.

In Chapter 2, our results also showed that viruses mainly consisted of bacteriophages that infect marine prokaryotes and viral groups that infect photosynthetic algae. Many of the bacteriophages that were identified mainly consisted of cyanophages, *Prochlorococcus* phage, *Synechococcus* phage, and *Pelagibacter* phage. In chapter 2 viral AMGs were detected using two different metagenomic programs to give a complete representation of the viral functionality. VIBRANT has detected more AMGs compared to DRAM-V, and the AMGs also showed a higher count (Kieft et al. 2020; Shaffer et al. 2020). However, both programs showed similarities in distribution of these AMGs. The AMG counts were much higher in the PNM, OMZ Edge, and SCM regions of the OMZ in station 2 and 3, and also in the PNM/OMZ Edge sample station 1 at 25m. Interestingly both programs detected genes that contained high counts in purine synthesis, represented by the PUR family of genes, and genes that contribute to photosynthesis. The presence of these AMG shows that viruses are actively participating in primary production and infecting their hosts for optimal performance in viral DNA replication, to further viral production. In chapter 2 we tested metagenomic bioinformatic methods to identify viruses in the ETNP OMZ, which provided an overview of the communities and different functionalities throughout these variable conditions and showing that viruses are integral part to biogeochemical cycling such as microbial communities, ecologically important, and more than agents of infection (Fuhrman 1999; Breitbart et al. 2007; Sullivan et al. 2006).

The previous 2 chapters I investigated microbial and viral communities through metagenomics in the ETNP OMZ. In chapter 3 we scaled down from the community level to focusing on AOA at the genomic level in the ETNP OMZ. In chapter 3 we have constructed 8 AOA MAG that have been detected in the AMZ core and OMZ Edge. These MAGs have been identified with a completion of greater than 70 percent and a contamination of less than 5 percent (Chivian et al. 2023). In chapter 3 we also utilized tools that helped provide metabolic summaries on the AOA (Shaffer et al. 2020), which ultimately highlighted genes involved in respiration. This contributes to the metabolic diversity of AOA, where certain respiratory pathways are more abundant than others. This showed that AOA plays multiple roles in nitrogen cycling, but ultimately aids in ammonia oxidation. Phylogenomic analyses on the first 8 AOA genomes determined how phylogenetically related they are to each other (Price et al. 2010), which led to 3 distinct clades across all phylogenies. For the second set of phylogenies, it included AOA MAGs constructed and retrieved from the ETSP OMZ. These phylogenies showed that AOA MAGs from the ETNP OMZ are distinct compared to the AOA from the ETSP OMZ. Lastly, pangenome analysis on the AOA MAGs further highlighted variations of core functional genes between AOA (Li et al. 2003; Medini et al. 2005; Arkin et al. 2018). As there were genes that were unique to specific AOA groups, genes shared amongst groups, and genes that are shared across all groups. Overall, chapter 3 builds on previous research of AOA in the ETNP OMZ showing that they are much more complex and have shown that autotrophic ammonia oxidizers can thrive in harsh environments and are genomically variable.

For future work I think it is important to highlight and continue the metagenomic techniques that were conducted on all three chapters and elevate our methods to aid in further analysis of microbial and viral communities that reside in the ETNP OMZ. Future work on chapter 1 will involve research on the MAGs that were constructed from all samples in chapter 3. In chapter 1 we primarily focused on MAGs that were present in station 1 at 25m, due to the high productivity and the unique environmental conditions of these samples as both the PNM and OMZ Edge overlapped. Looking into these newly constructed MAGs from the ETNP OMZ will provide much more information on major microbial groups that were identified from the raw metagenomic reads, such as exploring other metabolic capabilities in microbial groups across all stations and regions of the water column. Lastly for future work we plan to analyze microbial functionality through RNA sequences (Bashiardes et al. 2016). In chapter 1 we focused on DNA sequences which tells us gene abundance, but it doesn't tell us if the genes are being used. With RNA sequencing, these metatranscriptomes will tell us which genes in the ETNP are being transcribed, expressed, and active at the different environmental conditions highlighting microbial activity.

For chapter 2, we provided a snapshot of all the viruses that are present in the ETNP OMZ and the functional AMGs distributed throughout the water column. However, there is still more analysis to be conducted as viral taxonomic and AMGs identification were separate entities. For future work we aim to refine viral and AMGs detection by determining what AMGs these viruses possess at the genomic level. Many viral sequences that were assembled into contigs were also considered incomplete to be identified as viral genomes. Higher quality of viral sequences will allow us to uncover more genetic and metabolic content of these viral groups that were unretrievable with short read analysis. This will give us more of a refined idea of what the *Prochlorococcus* and *Peglaibacter* phages are doing in key regions of the ETNP OMZ, but also discover potential AMGs related to biogeochemical cycles. We want to extend this process to phages that are infecting other microbial groups that contribute to biogeochemical cycling in the ETNP OMZ. This will include microbial communities that contribute to sulfur and nitrogen cycling and give us a complete picture of phage ecology in these important regions of the ocean.

Lastly for chapter 3, we were able to detect and construct AOA MAGs from the intense regions of the ETNP OMZ. With these AOA MAGs we will continue to compare them to AOA that have been identified in other global OMZs and AOA than live in less extreme environments. We will also try to construct MAGs with lenient parameters such as 50 percent competition and less than 10 percent contamination. Our last focus for future directions is to investigate all the core genes that were identified during pan genomic analysis. Although a total of 362 genes were shared amongst all AOA genomes, in future work we will manually research each individual gene function that has been detected within each unique clade and determine its role and function. We will also investigate more gene functions that are shared among the AOA grouping and genes shared across clades. This investigation will paint a clearer picture of the genomic diversity of these AOA in the ETNP OMZ and how they are able to thrive in low oxygen marine environments.

References:

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., ... & Yu, D. (2018). KBase: the United States department of energy systems biology knowledgebase. Nature biotechnology, 36(7), 566-569.

Bashiardes, S., Zilberman-Schapira, G., & Elinav, E. (2016). Use of metatranscriptomics in microbiome research. Bioinformatics and biology insights, 10, BBI-S34610.

Beman, J. M., & Carolan, M. T. (2013). Deoxygenation alters bacterial diversity and community composition in the ocean’s largest oxygen minimum zone. *Nature Communications*, *4*(1), 2705.

Beman, J. M., Vargas, S. M., Vazquez, S., Wilson, J. M., Yu, A., Cairo, A., & Perez-Coronel, E. (2021). Biogeochemistry and hydrography shape microbial community assembly and activity in the eastern tropical North Pacific Ocean oxygen minimum zone. *Environmental Microbiology*, *23*(6), 2765-2781.

Breitbart, M. Y. A., Thompson, L. R., Suttle, C. A., & Sullivan, M. B. (2007). Exploring the vast diversity of marine viruses. Oceanography, 20(2), 135-139.

Chivian, D., Jungbluth, S. P., Dehal, P. S., Wood-Charlson, E. M., Canon, R. S., Allen, B. H., ... & Arkin, A. P. (2023). Metagenome-assembled genome extraction and analysis from microbiomes using KBase. Nature Protocols, 18(1), 208-238.

Fuhrman, J. A. (1999). Marine viruses and their biogeochemical and ecological effects. Nature, 399(6736), 541-548.

Kieft, K., Zhou, Z., & Anantharaman, K. (2020). VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome, 8(1), 1-23.

Kim, J., Kim, M. S., Koh, A. Y., Xie, Y., & Zhan, X. (2016). FMAP: functional mapping and analysis pipeline for metagenomics and metatranscriptomics studies. BMC bioinformatics, 17, 1-8.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome research, 13(9), 2178-2189.

Long, A. M., Jurgensen, S. K., Petchel, A. R., Savoie, E. R., & Brum, J. R. (2021). Microbial ecology of oxygen minimum zones amidst ocean deoxygenation. Frontiers in Microbiology, 12, 748961.

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. Current opinion in genetics & development, 15(6), 589-594.

Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. Progress in Oceanography, 80(3-4), 113-128.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. PloS one, 5(3), e9490.

Ruscheweyh, H. J., Milanese, A., Paoli, L., Karcher, N., Clayssen, Q., Keller, M. I., ... & Sunagawa, S. (2022). Cultivation-independent genomes greatly expand taxonomic-profiling capabilities of mOTUs across various environments. Microbiome, 10(1), 212

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., ... & Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic acids research, 48(16), 8883-8900.

Sullivan, M. B., Lindell, D., Lee, J. A., Thompson, L. R., Bielawski, J. P., & Chisholm, S. W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS biology, 4(8), e234.

Wright, J. J., Konwar, K. M., & Hallam, S. J. (2012). Microbial ecology of expanding oxygen minimum zones. Nature Reviews Microbiology, 10(6), 381-394.

Ulloa, O., Canfield, D. E., DeLong, E. F., Letelier, R. M., & Stewart, F. J. (2012). Microbial oceanography of anoxic oxygen minimum zones. Proceedings of the National Academy of Sciences, 109(40), 15996-16003.