A multi-omics approach to microbial nitrogen and sulfur cycling in the oxygen starved ocean

by

Alyse Kathleen Hawley

B.Sc., Biochemistry, The University of Victoria, 2004

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Microbiology and Immunology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

September 2018

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

A multi-omics approach to microbial nitrogen and sulfur cycling in the oxygen starved ocean

Submitted by Alyse K. Hawley in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in Microbiology and Immunology

Examining Committee:

Dr. Steven J. Hallam, Microbiology and Immunology

Supervisor

Dr. Philippe Tortell, Earth, Ocean and Atmospheric Sciences

Supervisory Committee Member

Dr. Julian Davies, Microbiology and Immunology

University Examiner

Dr. Rosemary J. Redfield, Zoology

University Examiner

Chair

Additional Supervisory Committee Members:

Dr. William W. Mohn, Microbiology and Immunology

Supervisory Committee Member

Dr. Lindsay Eltis, Microbiology and Immunology

Supervisory Committee Member

Dr. Leonard Foster, Biochemistry and Molecular Biology

Supervisory Committee Member

Abstract

Microbial communities mediate biogeochemical processes of Carbon (C), Nitrogen (N) and Sulfur (S) cycling in the ocean on global scales. Oxygen (O_2) availability is a key driver in these processes and shapes microbial community structure and metabolisms. As O_2 decreases, microbes utilize alternative terminal electron acceptors, nitrate (NO_3^-), nitrite, sulfate and carbon dioxide, depleting biologically available nitrogen and producing greenhouse gases nitrous oxide (N_2O) and methane (CH_4). Marine oxygen minimum zones (OMZs) are areas of O_2-depletion ($O_2 < 20\,\mu M$) in sub-surface waters due to the respiration of organic matter from the surface. In areas of acute O_2-depletion or where OMZs contact underlying sediments, hydrogen sulfide (H_2S) and CH_4 accumulate within OMZ waters, drastically altering microbial community structure and metabolism. In this thesis, I explore microbial cycles along defined gradients of O_2, NO_3^- and H_2S in Saanich Inlet, a seasonally anoxic fjord on the coast of British Columbia Canada. I develop a time-resolved multi-omic dataset consisting of small subunit ribosomal RNA amplicon sequences, single cell amplified genomes (SAGs), metagenomes, -transcriptomes and -proteomes, coupled with geochemical measurements, enabling robust microbial metabolic reconstruction at the individual, population and community levels of organization. Using metaproteomics, I construct a conceptual model of metabolic interactions involving N and S cycling, and carbon fixation, forming the basis for a collaborative effort to build a gene-centric numerical model, identifying an unrecognised niche for N_2O reduction. Using single cell amplified genomes (SAGs) from Saanich Inlet, I identify genes for N_2O reduction, (*nosZ*), within the dark matter phylum Marinimicrobia clade SHBH1141, filling the proposed niche of non-denitrifying N_2O-reducers. Using globally sourced Marinimicrobia SAGs, I further analyze energy metabolism and biogeography of several Marinimicrobia clades, revealing roles in C, N and S cycling along eco-thermodynamic gradients throughout the ocean. Finally, I chart the global abundance and distribution of *nosZ* genes and transcripts within the ocean, identifying previously unappreciated potential sinks for N_2O. As OMZs continue to expand and intensify due to climate change, defining metabolic processes and interactions along gradients of O_2-depletion becomes increasingly important. This thesis provides foundational knowledge related to the microbial communities driving coupled biogeochemical cycling in OMZs.

Lay Summary

Communities of microorganisms in the ocean are crucial for cycling carbon (C), nitrogen (N) and sulfur (S), elements essential for life. Climate change is depleting oxygen in much of the ocean, causing microbes to use nitrogen compounds instead, removing nitrogen available to other organisms for growth and producing the greenhouse gas nitrous oxide (N_2O). To better understand impacts of oxygen-depletion on C, N and S cycling, I looked at microbial genes, transcripts and proteins along gradients of oxygen-depletion in Saanich Inlet, a fjord on the British Columbia coast. With this combined dataset I constructed models for the exchange of nitrogen and sulphur based molecules between microbial groups under different oxygen conditions. I further identified microbes that consume N_2O and modeled interactions with other microbes involved in this process, improving our collective understanding of how oxygen depletion effects coupled nitrogen, sulfur and carbon cycles in the ocean.

Preface

A number of sections of this work are partly or wholly published in press or accepted. Copyright licences to all works were obtained and are listed where appropriate.

- **Chapter 1:** Text was written by Alyse Hawley. Figures were either used from other publications with permission or generated by Alyse Hawley as indicated.

- **Chapter 2:** Chapter 2 was written by Alyse Hawley with input from Steven Hallam. Saanich Inlet datasets are the result of the hard work of many students, technicians, volunteers and post-doctoral fellows. Specifically, for the generation of chemical and physical datasets: both Alyse Hawley and Mónica Torres Beltrán held the positions of Chief Scientist on board the Strickland for several years and oversaw sample collection, quality control and data curation. Chief Scientist position was also held by Steven Hallam, Elena Zaikova, Olena Shevchuk, Craig Miews and Jade Shiller during this time. Sea going technicians, Chris Payne and Laurisa Pakhomova, operated the CTD and collected samples for oxygen calibration and salinity and curated these datasets. Dissolved gas measurements and associated quality control were carried out by David Capelle.

  The generation of multi-omic datasets was the hard work of many people. Metagenomic datasets, including DNA extractions, were carried out by Alyse Hawley, Mónica Torres-Beltrán, Melanie Scofield, Payal Sipahimalani, Elena Zaikova, Olena Shevchuk and Steven Hallam. Fosmid libraries were constructed by Steven Hallam and David Walsh. Sequencing was carried out at the Joint Genome Institute (JGI) including library production and quality control. The generation of metatranscriptomic datasets included extractions with protocol designed by and carried out by Alyse Hawley. cDNA library construction and sequencing was carried out at the JGI including quality control. The generation of metaproteomic data included extractions with protocol designed by Alyse Hawley, with aid from Heather Brewer. Extractions were carried out by Alyse Hawley with aid from Jinshu Yang and Heather Brewer. Generation of peptide spectra was carried out by Heather Brewer at Environmental Molecular Science Laboratory (EMSL) at Pacific Northwest National Labs (PNNL) and spectra mapped to protein database by Angela Norbeck. Preparation of samples for small subunit ribosomal tag sequencing was carried out by Melanie Scofield and Mónica Torres Beltrán, sequencing was carried out at the JGI or at Genome Quebec. Tag sequence quality control and processing was carried out by Mónica Torres Beltrán and Kishori Konwar. The

multi-omics datasets were processed through MetaPathways, designed and built by Niels Hanson, Kishori Konwar and Steven Hallam.

➤ Portions of this text and protocols were published in the Methods in Enzymology chapter: Hawley, A. K., Kheirandish, S., Mueller, A., Leung, H. T., Norbeck, A. D., Brewer, H. M., Pasa-Tolic, L., Hallam, S. J., 2013. Molecular tools for investigating microbial community structure and function in oxygen-deficient marine waters. *Methods in Enzymology, Microbial Metagenomics, Metatranscriptomics, and Metaproteomics*. 2013;531 305-29.

➤ Datasets and descriptions were published at Scientific Data as: Hawley, A K., Torres-Beltrán, M., Bhatia, M., Zaikova, E., Walsh, D. A., *et al.* 2017. A compendium of multi-omic sequence information from the Saanich Inlet water column. *Scientific Data*.2017; 4:170160.

and

➤ Torres-Beltrán, M., Hawley, A. K., Capelle, D., Zaikova, E., Walsh, D. A., *et al.* 2017. A compendium of geochemical information from the Saanich Inlet water column. *Scientific Data* 4:170159.

➤ Additional manuscript using metagenomic and sing-cell amplified genomes is published in eLife as: Roux, S., Hawley, A. K., Torres-Beltrán, M., Scofield, M., Schwientek, P., Stepanauskas, R., Woyke, T., Hallam, S. J., Sullivan, M. B., 2014 Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell and meta-genomics. *eLife*. 3;e03125.

➤ Additional manuscript describing application of MetaPathways annotation pipeline to metagenomic datasets is published in BMC Genomics as: Hanson, N. W., Konwar, K. M., Hawley, A. K., Altman, T., Karp, P. D., Hallam, S. J., 2014. Metabolic pathways for the whole community. *BMC Genomics* 15:619

- **Chapter 3:** Chapter 3 analysis and writing was carried out by Alyse Hawley with input from Steven Hallam. Sample preparation was carried out by Alyse Hawley and Heather Brewer at EMSL. Matching of peptide spectra to protein sequences and calculation of false discovery rate was carried out by Angela Norbeck at EMSL. SSUrRNA tag sequences were sequenced at the JGI and processed by Kishori Konwar. Identification of protein taxonomy and function as well as calculation of normalised spectral abundance factor was calculated by Alyse Hawely.

  ➤ A version of this chapter is published in Proceeding of the National Academy of Sciences as: Hawley, A K. Brewer, H.M. Norbeck, A. D. Paša-Tolic, L. Hallam, S. J. 2014. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proceedings of the National Accademy of Sciences USA*. 111:31 11395-11400.

➤ Additional manuscript based on these analyses with further modeling of geochemical and multi-omic datasets is published in Proceedings of the National Academy of Sciences as: Louca, S., Hawley, A. K., Katsev, S., Torres-Beltrán, M., Bhatia, M. P., Kheirandish, S., Michiels, C., Capelle, D., Lavik, G., Doebeli, M., Crowe, S. A., Hallam, S. J. 2016. Integrating biogeochemistry with multi-omic sequence information in a model oxygen minimum zone. *Proceedings of the National Accademy of Sciences USA*. 113:40 E5925-E5933.

- **Chapter 4:** Chapter 4 analysis and writing was carried out by Alyse Hawley with contributions from Nobu Masaru and Jody Wright and input from Steven Hallam. Collection of samples for single-cell amplified genomes (SAGs) from Saanich Inlet was carried out by Alyse Hawley and Móncia Torres-Beltraán, from North Eastern Subarctic Pacific Ocean waters by Jody Wright. SAGs from other locations were collected for previous publications as indicated. Sequencing of SAGs from Saanich Inlet was carried out at the Genome Sciences Centre; sequencing of SAGs from NESAP was carried out at the JGI. Assembly and decontamination of SAGs were carried out at the JGI. Genome reduction analysis was carried out by Nobu Masaru. Phylogenetic analysis and associated figure production were carried out by Jody Wright, Brent Sage and Keith Miews. Generation of population genome bins was done by Evan Durno. Global metagenome fragment recruitment analysis was carried out by Alyse Hawley, aided by Connor Morgan-Lang. Metabolic analysis was carried out by Nobu Masaru, Alyse Hawley and Jody Wright. Expression analysis and global *nosZ* distribution was carried out by Alyse Hawley.

  ➤ A version of this chapter is published in Nature Communications as: Hawley, A. K., Nobu, M. K., Wright, J. J., Durno, W. E., Morgan-Lang, C., Sage, B., Schwientek, P., Swan, B. K., Rinke, C., Torres-Beltrán, M., Mewis, K., Liu, W., Stepanauskas, R., Woyke, T., Hallam, S. J. 2017. Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. *Nature Communications*. 8:1507.

- **Chapter 5:** Chapter 5 analysis and writing was carried out by Alyse Hawley with input from Steven Hallam. Collection of samples for single-cell amplified genomes (SAGs) from Saanich Inlet was carried out by Steven Hallam, Alyse Hawley and Móncia Torres-Beltrán. SAGs were sequenced at the Genome Sciences Centre and assembled and decontaminated by Connor Morgan-Lang. Identification of nitrogen cycling genes in SAGs and global metagenomes and expression analysis was carried out by Alyse Hawley. Phylogenetic tree for NosZ was constructed by Connor Morgan-Lang and nitrous oxide measurement done by David Capelle.

  ➤ Manuscript detailing the N_2O dynamics in Saanich Inlet is published in Limnology and Oceanography as: Capelle, D. W., Hawley, A. K., Hallam, S. J., Tortell, P. D. 2018. A Multi-year time-series of N2O dynamics in a seasonally anoxic fjord: Saanich Inlet, British

Columbia. *Limnology and Oceanography* 63:2 524-539.

None of the work encompassing this dissertation required consultation with the UBC Research Ethics Board.

Table of Contents

Appendices

List of Tables

List of Figures

Acknowledgements

First of all I would like to thank my supervisor Dr. Steven Hallam and acknowledge the chance he took in bringing me into the lab and for always pushing me to do better, think both deeply about detail and think big about global processes. Thank you for introducing me to microbial ecology, you have changed the way I see the world and I would never go back. Mostly I would like to thank you, Dr. Hallam, for a challenging and rewarding 10 years of working with you! I would like to thank my committee, Dr.'s Philippe Tortell, Bill Mohn, Leonard Foster and Lindsay Eltis for their support and input and particular Dr. Philip Tortell for my introduction to oceanography - it has begun quite a love affair! I would like to thank all of the members of the Hallam lab, past and present, for their unvaried support, both emotional and logistical. Monica Torres Beltrán, thank you for being my partner in crime and my comrade in Saanich, you have been a tremendous support and source of insight! Maya Bhatia, thank you for your advice and support and kick in the butt when I needed one. Elena Zaikova and Esther Gies, thank you for your friendship and support and willingness to go to Saanich and help with unending field work. Evan Durno and Connor Morgan-Lang thank you for writing and running so much code for me and your willingness to help me any time I needed it! Aria Hahn, thank you for consultations on data processing and visualization. I would also like to thank all of those who have worked in Saanich Inlet over the years with me, both as chief scientists and as support. I would like to thank Chris Payne and Laura Pakhomovitch for their hard work and accompaniment in Saanich. And to Captain Ken and the Strickland crew, thank you for many fond memories of Saanich Inlet. I would also like to thank my collaborators at the Joint Genome Institute for sequencing so many samples, and at Environmental Sciences Laboratory at Pacific Northwest National Labs for support in proteomics, none of this work would have been possible without you. I also thank my parents, Kathy and Jan, for endless encouragement and support and my in-laws Doreen and Ken for both financial support and child care. Lastly, I would also like to acknowledge the sacrifices of my husband, Michael Hawley, and my family for the past years as I dedicated so much time to this work. Michael, it would not have happened without all of your support.

Dedication

To Mike who reminds me to be who I am

To Avery and Riley who remind me of what is to come

To my parents who supported me to become the person I am

Chapter 1

Introduction

'When we try to pick out anything by itself, we find it hitched to everything else in the universe.'

John Muir (1911)

1.1 Global oxygen minimum zones

Oceans occupy over two-thirds of the planet's surface and host a vast diversity of microbial life. The metabolic activity of this microbial life is responsible for major global biogeochemical processes such as oxygen (O_2) production, carbon fixation (or primary production), cycling of nitrogen, sulfur, phosphorus and many other elements and nutrients. Most of the ocean is moderately oxygenated (200 - 100 µM) [1], providing a ready source of terminal electron acceptor for breakdown and respiration of large amounts of organic matter produced by primary production in the surface waters. However, $\sim$ 7% of the global ocean is depleted in oxygen [2, 3], with concentrations dropping below 20 µM. Areas of O_2-depletion, called oxygen minimum zones (OMZs) host unique microbial communities with metabolic activity adapted to thrive in an O_2-depleted environment [2, 4]. In OMZs, O_2-depletion shifts the microbial communities and their metabolisms to the use of alternative terminal electron acceptors such as nitrate (NO_3^-), nitrite (NO_2^-), and carbon dioxide (CO_2) [5], changing the biogeochemical processes and resulting in the production of greenhouse gases nitrous oxide (N_2O) and methane (CH_4) and loss of biologically available nitrogen [6] with implications for global climate and nutrient availability.

1.1.1 OMZ formation and global distribution

Oxygen minimum zones form in subsurface waters, resulting from a combination of respiration of organic matter from primary production in surface waters and restricted ventilation or mixing,

preventing re-oxygenation at the surface [2, 7]. Oxygen minimum zones most often occur on western continental margins (Figure 1.1). As winds blow north (or south in the southern hemisphere) along western coasts the coriolis effect moves the surface waters away from the coast causing deeper, nutrient (nitrate, phosphate, silicate) rich waters to move upward, in a process called upwelling. Once exposed to the sunlight, nutrient rich waters support phytoplankton to carry out rapid photosynthesis, causing a high influx of organic matter in the surface waters. As phytoplankton die off the organic matter rains down into the subsurface waters where heterotrophic bacteria degrade it, respiring the O_2 present. A lack of ventilation in the subsurface waters prevents re-oxygenation, resulting in OMZ formation [6, 7]. In the open ocean, OMZs form in a similar manner, but with little upwelling to fuel rapid photosynthesis, O_2-depletion is less intense (Figure 1.2).

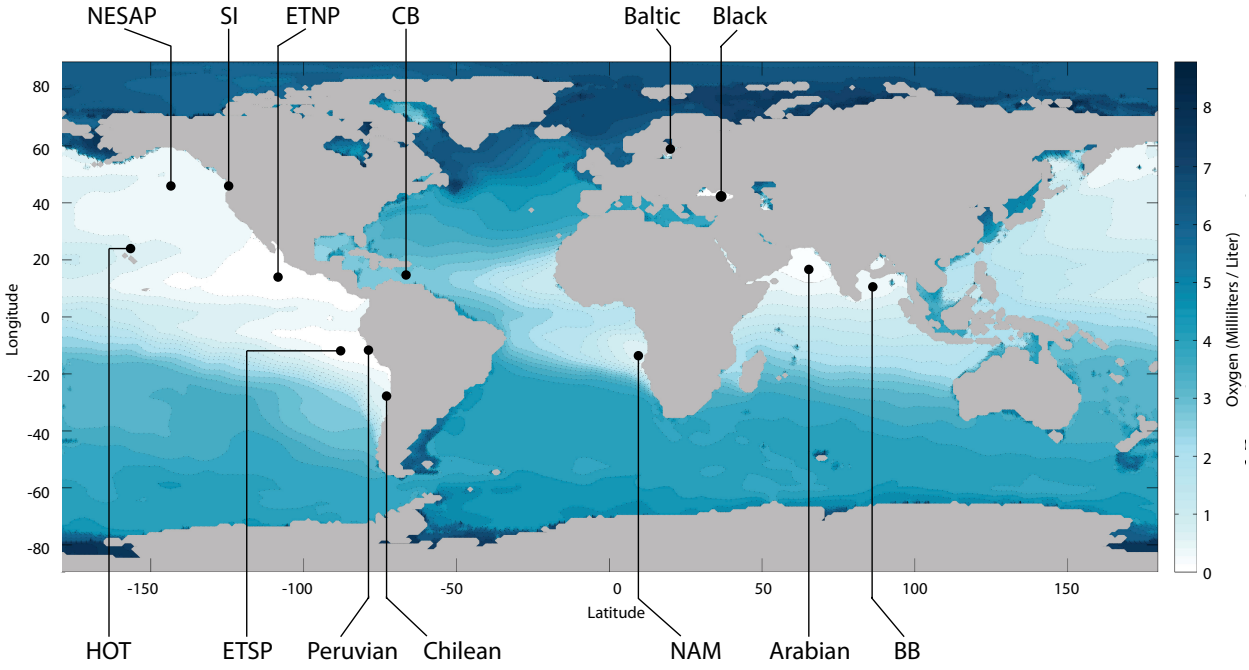


Figure 1.1: Global OMZ distribution. Global distribution of OMZs including: Northeastern Subarctic Pacific Ocean (NESAP), Saanich Inlet (SI), Eastern Tropical North Pacific (ETNP), Cariaco Basin(CB), Baltic and Black Seas, Hawaii Ocean Time-series (HOT), Eastern Tropical South Pacific (ETSP), Peruvian, Chilean, Namibian (NAM) , Arabian and Bay of Bangual (BB). Oxygen concentration shown at depth of minimum oxygen concentration. Figure was modified from Wright *et al.* 2012.

The global distribution of OMZs is substantial (Figure 1.1), with major OMZs including open ocean OMZs(Figure 1.2A): North Eastern Subarctic Pacific (NESAP), Eastern Tropical North

Pacific (ETNP), Eastern Tropical South Pacific (ETSP) and the Bay of Bengal; and coastal and semi-enclosed inlet and basin OMZs: Saanich Inlet (SI), Peruvian and Chilean upwelling systems, Baltic Sea, Black Sea, Namibian shelf and Cariaco Basin. Oxygen minimum zone formation in semi-enclosed inlets, basins and coastal areas where geography restricts mixing, promotes a stratified water column with oxygenated surface waters and O_2-depleted bottom waters [8]. Coastal systems can experience localized eutrophication (run-off of nutrients from terrestrial sources), causing results similar to upwelling with an input of large amounts of nutrients [5, 7]. Within inlets and basins, and occasionally in OMZs in close proximity to the shore, OMZ waters can contact the underlying sediments. Sufficient O_2-depletion in these areas can allow reduced compounds, such as hydrogen sulfide (H_2S) and CH_4, to efflux out of the sediments and accumulate in the overlying water column, creating a highly reduced environment termed a sulfidic OMZ (Figure 1.2C). These sulfidic OMZs are often characterized by steep gradients of O_2, NO_3^- and H_2S occurring over a few to tens of meters within the water column [4, 9].

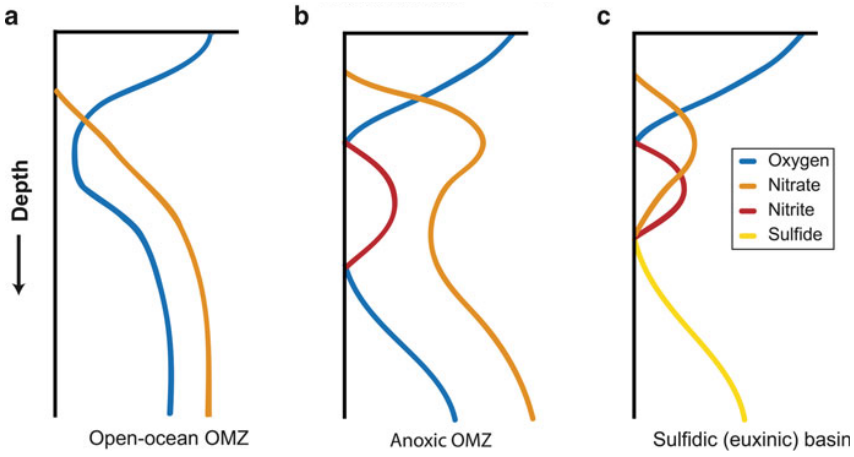


Figure 1.2: OMZ types. Characteristic chemical profiles in OMZs. **(A)** Open-ocean OMZ.. **(B)** Anoxic OMZ. **(C)** Sulfidic basin. Figure from *Pelagic Oxygen minimum zone microbial communities.* 2013. By Ulloa, O., Wright, J.J., Belmar, L., Hallam, S. J.

1.1.2 OMZ expansion and intensification

Currently, climate change and other anthropogenic forces are causing an expansion and intensification of OMZs globally [3, 7, 10, 11]. Ocean surface temperature rise causes increased

stratification of the surface water column, leading to decreased mixing and consequently an intensification of O_2-depletion as well as shallower OMZ depths [3]. Indeed, over the past 60 years O_2 concentrations in the NESAP have dropped 22% [10] (Figure1.3). Recent modeling efforts have forecast intensification of O_2-depletion globally in the coming decades [11]. As OMZs play key roles in the global nitrogen and carbon cycles [12, 13] and greenhouse gas production [6] there comes a pressing need to understand the biogeochemical cycles and contributions of the microbial metabolisms in OMZs to global processes including greenhouse gas production and carbon sequestration.

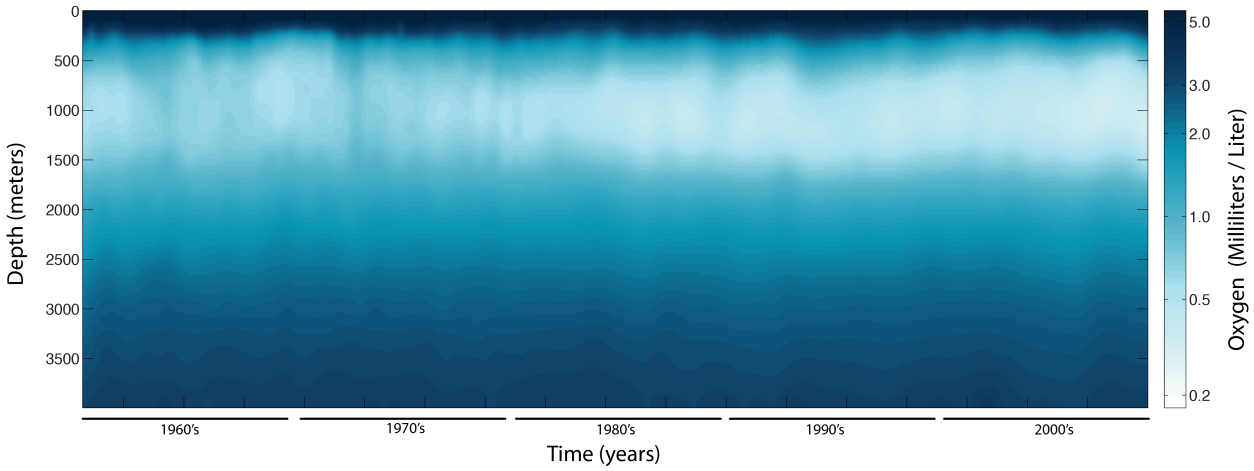


Figure 1.3: OMZ expansion in North Eastern Subarctic Pacific. Oxygen concentration at Ocean Station Papa (50.1N, 144.9W) over the past 60 years.

1.1.3 Redox gradients and redox driven niche partitioning

In OMZs, as O_2 is depleted, microbes begin to respire NO_3^-, producing NO_2^- (and potentially N_2O and N_2) [12]. Gradients of O_2, NO_3^- and NO_2^- (Figure 1.2) form an overall redox gradient that shapes the microbial community and associated metabolic activities [4]. Furthermore, in sulfidic OMZs the redox gradient extends into the more highly reduced sulfidic environment, providing additional niches for the microbial community and associated metabolic activities. As terminal electron acceptors are depleted from O_2, NO_3^-, NO_2^-, SO_4^{2-} and finally to CO_2 in respiratory processes (Figure1.4), O_2-depletion levels can be classified as oxic (>90 µM O_2), dysoxic (20-90 µM O_2), suboxic (1-20 µM O_2), anoxic (<1 µM O_2) and sulfidic. These classifications

serve to provide a framework for discussing microbial ecology within OMZ systems. Redox gradients provide conditions for redox driven niche partitioning where microbes partition at niches along the gradient according to their metabolic capabilities [2]. Notably, the presence of reduced sulfur compounds such as H_2S as well as CH_4 can provide additional electron donors and additional niches for chemotrophic metabolisms. While O_2, NO_3^-, NO_2^- and other chemical parameters can be easily measured, the microbial community and their associated metabolic activities along these gradients requires more involved sampling and analysis. While several of the more abundant microbial groups along OMZ redox gradients are known, their metabolic capabilities and thresholds (i.e. O_2 concentrations at which NO_3^- respiration takes over or N_2O production occurs) remain undetermined for the majority.

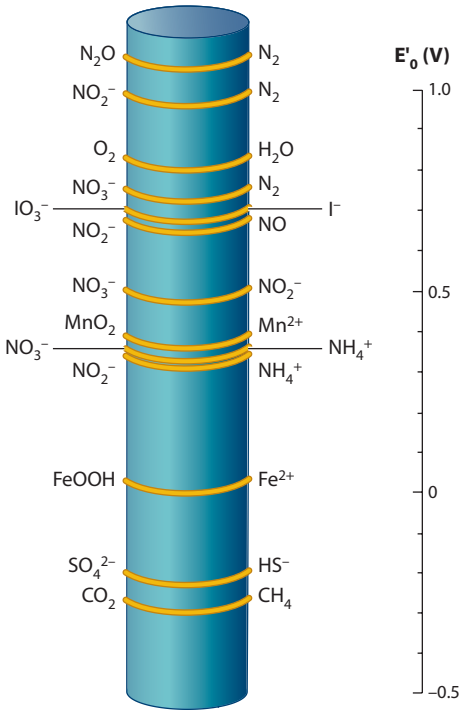


Figure 1.4: Electron tower. Electron potentials ($E^{'}\,^{\circ}$) of various redox couples possible in OMZs. Figure from Lam and Kuypers 2011.

Particle associated niches

While marine waters, including OMZs, are generally considered a homogeneous mixture, micro-environments can exist on particles formed from the dead matter of phytoplankton (or detritus) and fecal pellets from larger organisms such as copepods. These micro-environments provide additional sites of redox gradients as microbial respiration within the particle depletes O_2 and diffusion on this small scale cannot re-supply O_2 to internal spaces. As such, particles develop internal redox gradients with anoxic and even sulfidic micro-environments [2, 14]. Within OMZs, the decreased O_2 concentrations in the bulk water column provide even less O_2 available for diffusion into particles, resulting in strong redox gradients within the particles and potentially larger shifts in the particle associated microbial community and metabolic activity. While some research has been carried out on the particle associated microbial community [14–16], technical challenges around isolating particles remain substantial and the contribution of the particle associated microbial communities to biogeochemical processes on both local and global scales remains a question.

1.1.4 OMZ microbial community overview

Surveys of the small subunit ribosomal RNA (SSU rRNA) gene in OMZs have shown the partitioning of similar microbial groups (defined at varying degrees between sub-phylum and genus) along defined redox gradients [2, 9, 17, 18]. These results have been detailed in Wright *et al.* 2013 and are summarised below. As with many natural environments, numerous microbial groups remain uncultured and are known only on the basis of SSU rRNA gene sequences that have been identified. These so-called microbial dark matter (MDM) groups [19] represent unknown metabolic potential and interact with known taxa to define the metabolic networks that drive nutrient and energy cycling along OMZ redox gradients.

Prevalent sequences found in oxygenated waters overlying OMZs include: Alphaproteobacteria affiliated with SAR11, Betaprotebacteria from the order Methylophilales, Gammaproteobacteria affiliated with SAR86 and Arctic96B-1, Actinobacteria affiliated with OM1, Bacteriodetes affiliated with the genus *Polaribacter*, Arctic96A-17, and Cyanobacteria. These groups are largely

heterotrophic, remineralizing organic matter from surface waters. Members of the SAR11 clade are among the most abundant bacteria in the ocean [20] and have multiple cultured sub-clades with representatives typically exhibiting reduced genome sizes (around 1 Mb). With highly streamlined genomes, the metabolic potential of SAR11 clades can vary significantly [21], with genomic evidence for phototrophy via bacteriorhodopsin [22], as well as for one-carbon metabolism [23] and the transformation of various sulfur compounds [24].

Prevalent sequences found in dysoxic and suboxic OMZ waters include: Alphaproteobacteria affiliated with SAR11 (with distinct SAR11 clades in oxic verses dysoxic waters), Gammaproteobacteria affiliated with agg47, ZD0417, ZA3412c and Arctic96BD_19, Deltaproteobacteria affiliated with SAR324 and Nitrospina, Actinobacteria affiliated with Microthrixinea, Planctomycetes, Chloroflexi, Verrucomicrobia and Marine Group A. In addition to these bacterial groups, Archea affiliated with *Thaumarchaeota* are also present. These groups are a mixture of heterotrophs and chemoautotrophs and represent a transition into capacity for anaerobic metabolisms. For example, within dysoxic water, genomic bins affiliated with SAR11 harbour genes involved in nitrate reduction, potentially contributing to nitrogen loss within OMZs [25]. Similarly, both Arctic96BD_19 and SAR324 have been implicated in sulfur cycling processes [26, 27]. Several other groups play different roles in the nitrogen cycle. Nitrospina carry out nitrite oxidation [28], Planctomycetes carry out anammox [29], and *Thaumarchaeota* carry out ammonia oxidation [30]. Candidate phyla Marine Group A (also known as Marinimicrobia) while numerically abundant in places like the NESAP OMZ [31] remain enigmatic, although limited genomic information points to a role in sulfur cycling [32].

Prevalent sequences found in suboxic, anoxic and sulfidic OMZ waters are primarily affiliated with SUP05 within the Gammaproteobacteria. Additional sequences found in anoxic and sulfidic OMZ waters include: Deltaproteobacteria affiliated with Desulphobacteraceae, Epsilonproteobacteria affiliated with Arcobacter, Bacteriodetes affiliated with VC21_Bac22, various Gemmatimonadetes and Lantisphaerae, and the MDM phyla Marine Group A, OD1 and OP11. Genomic evidence suggests that many of these groups are chemoautotrophic with the potential to couple different aspects of the carbon, nitrogen and sulfur cycles. SUP05 has been implicated in partial denitrification coupled to sulfide oxidation, using the resulting energy to drive dark

carbon fixation [33, 34]. A recent study in the Peruvian upwelling system identified a complete denitrification pathway in a metagenome assembled metagenome (MAG) [35] assigned to SUP05, indicating the potential for functional specialization within the clade [36]. Similar to SUP05, Arcobacter have been implicated in denitrification coupled to sulfide oxidation although the two groups appear to occupy different energetic niches in OMZs [37].

1.2 Biogeochemical cycles in OMZs

Biogeochemical cycles are the biological, geological and chemical processes responsible for moving elements through both biotic and abiotic systems, recycling elements for availability to living organisms. Microorganisms play a significant role in many of these cycles, as their metabolic activities drive biogeochemical cycles [38]. Primary biogeochemical cycles of interest include carbon, nitrogen, sulfur and phosphorus as these elements are essential in biological molecules such as nucleic acids, proteins and lipids. Many biogeochemical cycles are carried out through multi-step processes involving multiple different guilds of organisms, each with a specialised role within the cycle. Within OMZs the O_2-depleted and anoxic environments provide essential niches for O_2 sensitive enzymatic reactions and reactions involving alternative electron acceptors. The microbial communities present along the redox gradients in OMZs play essential roles in biogeochemical cycles, particularly in nitrogen and sulfur cycles, as well as the carbon cycle.

1.2.1 Biogeochemical cycling of nitrogen in OMZs

Nitrogen is an essential nutrient integral in DNA, RNA and protein and thus essential for cell growth and carbon fixation. However, the majority of the Earth's nitrogen exists as inert dinitrogen (N_2) gas which must be 'fixed' into the biologically available nitrogen species such as ammonium (NH_4^+), a process carried out only by specific clades of microbes found in relatively low abundances [39] (as well as industrial chemical processes). All other organisms then obtain their nitrogen by taking up these fixed nitrogen species, NH_4^+ or NO_3^- and NO_2^-, directly or through heterotrophy of organic matter. Within OMZs, biologically available nitrogen species play additional roles of electron acceptors (NO_3^- and NO_2^-) and donors (NH_4^+) to directly

fuel energy metabolism via dissimilatory processes. Nitrogen based energy metabolisms include aerobic processes of ammonia oxidation and nitrification and anaerobic processes of denitrification, anaerobic ammonium oxidation (or anammox) and dissimilatory nitrate reduction to ammonium (DNRA) (Figure 1.5). Anaerobic processes of denitrification and anammox are termed nitrogen loss processes as they remove biologically available nitrogen from the environment, completing the cycle by returning nitrogen to the atmosphere as N_2 or N_2O. In any given system there must be enough biologically available nitrogen to support growth, while nitrogen loss is a key process in completing the nitrogen cycle, too much nitrogen loss will limit growth and primary production. Generally, it is predominantly the availability of O_2 and other electron donors and acceptors within a given niche that influences the occurrence of different nitrogen based energy metabolisms, but much remains to be understood about the balance between different nitrogen cycling processes along redox gradients in OMZs.

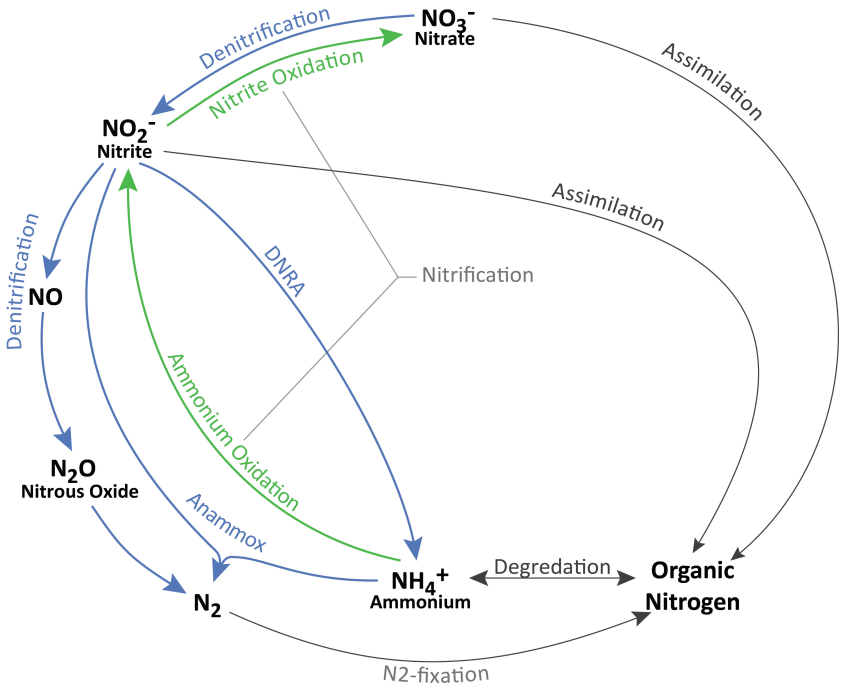


Figure 1.5: Nitrogen cycle. Biogeochemical transformation involved in the Nitrogen cycle, indication aerobic processes (green) and anaerobic processes (blue). Figure modified from *Nitrogen in the marine environment* 2008, by Capone.

Nitrogen fixation

Nitrogen fixation is not a focus of this thesis, it is however important to note its occurrence in the ocean and current state of research in OMZs. The specific enzyme responsible for nitrogen fixation is NifH. Nitrogen fixation requires significant energy and therefore is most often found coupled with photosynthetic processes in surface waters and is carried out at higher rates in warm tropical waters [39, 40]. However, more recent studies document nitrogen fixation by heterotrophic dizotrophs from many diverse taxa present throughout the ocean, including dysoxic, suboxic and sulfidic waters [41]. The potential for nitrogen fixation within O_2-deficient waters has implications for the extent of N-loss by denitrification and anammox (discussed below) as it could both fuel N-loss processes and re-supply biologically available nitrogen for growth [42, 43].

The extent of production of biologically available nitrogen from heterotrophic diazotrophs within OMZ waters appears to be variable. Studies from both the Baltic Sea [41] and the Peruvian upwelling [44] report nitrogen fixation rates ranging between 0.1 to 3.4 $\mathrm{nmol\,l^{-1}\,d^{-1}}$. Within the Peruvian upwelling, depth integrated rates measured in different years were reported ranging between 7.5 to 190 $\mu\mathrm{M\ m^{-2}d^{-1}}$ with the highest rates measured within the oxycline and OMZ core [44]. Furthermore, $NO_2{}^-$ and $PO_4{}^{3-}$ appear to be a factor in the distribution of heterotrophic diazotrophs throughout OMZ waters [45]. Studies of taxa carrying out heterotrophic diazotrophy in the Eastern Tropical South Pacific Ocean indicate a predominance of Alpha- and Gamma-proteobacteria as well as sulphate reducing Deltaproteobacteria, Clostridium and Vibrio species [46]. These smaller inputs of biologically available nitrogen into OMZs may impact the balance between different nitrogen based energy metabolisms along OMZ redox gradients with implications for global nitrogen budgets.

Nitrification

Nitrification is the processes of oxidizing $NH_4{}^+$ to $NO_3{}^-$ (Equation 1.1), as $NH_4{}^+$ is released from the degradation of organic matter. In most environments nitrification is nearly always split between two different organisms [47], ammonium oxidizers: oxidizing $NH_4{}^+$ to $NO_2{}^-$ and nitrite oxidizers: oxidizing $NO_2{}^-$ to $NO_3{}^-$. Both $NH_4{}^+$ and $NO_2{}^-$ oxidation are carried out in

subsurface waters as sunlight is inhibitory to NH_4^+ oxidation [48] (Figure 1.2). Thus, nitrification is a dominant nitrogen cycling process within the dysoxic waters of OMZs.

$$NH_4^+ + O_2 \longrightarrow NO_2^- \longrightarrow NO_3^- \tag{1.1}$$

Within marine systems ammonium oxidation is carried out nearly exclusively by ammonium oxidizing archaea (AOA) of the Thaumarcheaota lineage [49]. The first isolated archaeal ammonia oxidizer, *Nitrosopumulis maritimus*, was isolated from aquarium gravel and observed to grow chemoautotrophically on NH_4^+ [50]. The specific enzyme responsible for NH_4^+ oxidation, ammonia monooxygenase (AmoA), carries out the oxidation of NH_4^+ to hydroxylamine (NH_2OH), which is then oxidized to NO_2^- [30, 51, 52], however the specific enzymes responsible for NH_2OH oxidation to NO_2^- are undetermined in AOA [53]. Additional studies have found the archaeal *amoA* gene to be widely distributed and abundant throughout the global ocean [30, 39]. While certain bacteria are also known to harbour *amoA*, in the ocean the process of NH_4^+ oxidation is carried out predominantly by archaea [39, 40, 49].

Nitrite oxidation is carried out by a few different bacterial lineages, within marine environments they are predominately members of *Nitrococcus* [54], *Nitrospina* [28] and *Nitrospira* [55]. Members of these lineages have been isolated from OMZs and genomic evidence suggests a chemoautotrophic life style [28, 54, 55]. The enzyme responsible for NO_2^- oxidation to NO_3^- is nitrite oxido-reductase (Nxr). Activity of nitrite oxidizing bacteria has been observed in OMZs at O_2 concentrations as low as 1 µM, suggesting a broad niche for this biogeochemical process and coupling of nitrification with anaerobic nitrogen loss processes [56]. Recent studies suggest *Nitrococcus* to be numerically dominant within OMZ waters [56] and clutured *Nitrococcus* are further observed to oxidize sulfide, reduce NO_3^- and produce N_2O [54]. These metabolic capacities further expand the metabolic roles of nitrite oxidizing bacteria in O_2-deficient waters.

Denitrification

Denitrification is a nitrogen loss process, removing biologically available nitrogen in the form of N_2 or N_2O gas. Denitrification involves the successive reduction of NO_3^- to NO_2^-, to nitric oxide

(NO) to N_2O, to N_2 (Equation 1.2). All steps can be carried out by a single organism, or split across several different groups of organisms in a distributed metabolic process [57]. The individual modular steps of denitrification can be taxonomically diverse with NO_3^- reduction being carried out by multiple domains including Fungi [57, 58] and N_2O reduction observed in diverse bacterial lineages including multiple proteobacteria, as well as Verucomicrobia, Bacteriodetes, Chlorobi and halophilic archaea [59].

Initial work on denitrification in OMZs involved primarily rate measurements and gene counts for individual genes in the denitrification pathway (see below) using quantative PCR [60–62]. However, these methods yielded little information on the taxonomy of organisms carrying out these reactions [12, 63]. With the introduction of next generation sequencing and single-cell amplified genome technologies, combined with traditional culturing techniques, some taxonomic information about denitrifying organisms in OMZs has become available. Specifically, Epsilonproteobacteria *Sulfurimonas gotlandica* (isolated from the Baltic Sea) is seen to carry out complete denitrification [64]. Gammaproteobacteria of the SUP05 lineage (draft population genome from Saanich Inlet), including *Candidatus* Thioglobus autotrophicus (isolated from sulfidic basin Effingham Inlet, BC), is seen to carry out incomplete denitrification, producing N_2O [34]. However, recently metagenome assembled genome for the SUP05 group ^{U}*Thioglobus perditus*(U indicating uncultured) from the Peruvian upwelling indicates capacity for complete denitrification including N_2O reduction. The full taxonomic range of denitrifying organisms in OMZs remains to be determined and is a central focus of this thesis.

The enzyme complex responsible for the first step of denitrification (reduction of NO_3^- to NO_2^-) can be either the nitrate reductase NarDGHIJ (Active subunit NarG) or periplasmic nitrate reductase NapAB. While NarG has a wide taxonomic diversity [58] NapA has only been observed within the proteobacteria [58]. The second step of denitrification (reduction of NO_2^- to NO) can be carried out by either copper containing nitrite reductase NirK or cytochrome cd_1 containing nitrite reductase NirS. The third step of denitrification (reduction of NO to N_2O) is carried out by nitric oxide reductase NorCB, and the final step of denitrification (reduction of N_2O to N_2) is carried out by nitrous oxide reductase NosZ. Due to its modular nature, complete denitrification may not occur within a given redox niche, resulting in the accumulation of partial denitrification

products such as NO_2^- or N_2O at points along the redox gradient (the inherent instability of nitric oxide does not allow for significant accumulation within the water column). Furthermore, the diversity of organisms carrying out various steps of denitrification makes it difficult to determine the contribution of a given taxa to measured denitrification rates and the contribution of that taxa to nitrogen loss across different OMZ systems.

$$NO_3^- \longrightarrow NO_2^- \longrightarrow NO \longrightarrow N_2O \longrightarrow N_2 \tag{1.2}$$

Anammox

Anaerobic ammonium oxidation or Anammox is an additional nitrogen loss process carried out exclusively by the *Brocadia* lineage of Planctomycetes which use NO_2^- to oxidize NH_4^+ to N_2 (Equation 1.3). Members of the *Brocadia* have been isolated primarily from wastewater treatment facilities [65] and appear to be capable of a chemoautotrophic lifestyle [66, 67]. Planctomycetes of the *Scalindua* clade within the *Brocadia* appear in OMZs globally [65], indicating the global importance of this nitrogen loss process. There are three primary enzymes responsible for anammox: NirS, which reduces NO_2^- to NO, hydrazine hydrolase (HH) which converts NO to hydrazine (N_2H_2) with the addition of NH_4^+, and hydrazine dehydrogenase (HZO) which converts H_2N_2 to N_2 [66, 67]. Anammox is generally carried out under anoxic conditions but several lines of evidence suggest a broader niche for this process. For example, anammox rates have been observed in OMZs at O_2 concentrations up to 20 µM [68] and *Scalindua* sp. have been shown to be particle associated, suggesting an additional potential niche for this process [14]. Additionally, anammox-type *nirS* transcripts have been detected in water samples from the Black Sea with 2 µM H_2S [69], suggesting an additional, more reduced niche for anammox in sulfidic OMZs. In all, anammox has been expected to contribute up to 50% of nitrogen loss globally (including processes occurring in the sediments) [70, 71], but much uncertainty remains around the conditions which govern the balance between anammox and denitrification in OMZs [12, 60, 61, 72].

$$NO_2^- + NH_4^+ \longrightarrow N_2 \tag{1.3}$$

Dissimilatory nitrate reduction to ammonium

Dissimilatory nitrate reduction to ammonium (DNRA) is an additional nitrogen transformation that does not involve loss of biologically available nitrogen. As DNRA was originally thought to occur only in sediments and soils, and was only recently discovered in the water column [29], its distribution in marine systems is still being determined. Observations of DNRA in OMZs include the Peruvian [61] and Arabian [73] OMZs, but additional studies have yet to further constrain its distribution globally. Taxonomic lineages found to carry out DNRA include the gamma-, delta and epsilon-proteobacteria [12]. The enzyme responsible for DNRA, cytochrome C nitrite reductase (NrfA), carries out a six electron reduction of NO_2^- to NH_4^+ (Equation 1.4). While only sparingly observed in the marine water column thus far, the implications of DNRA on nitrogen cycling are substantial as use of NO_2^- to drive NH_4^+ production offsets or competes with nitrogen loss via denitrification and/or anammox.

$$NO_2^- \longrightarrow NH_4^+ \tag{1.4}$$

1.2.2 Biogeochemical cycling of sulfur in OMZs

Sulfur is essential for protein production and thus for cell growth. Sulfur is generally readily available in marine systems in the form of SO_4^{2-}, released upon the degradation of organic matter, and taken up directly or heterotrophically through organic matter. Within OMZs, as O_2 is depleted, SO_4^{2-} potentially becomes a viable electron acceptor, producing reduced sulfur compounds such as thiosulfate ($S_2O_3^{2-}$) and H_2S, which can then be oxidized. with both reduction and oxidation reactions fuelling chemotrophic energy metabolism. The taxonomic distribution of SO_4^{2-} reducers is generally thought to be predominantly Deltaproteobacteria in marine systems, while sulfide oxidizers have a broader taxonomic distribution including many proteobacterial lineages, Chromatiaceae and Chlorobi lineages as well as some Archaea [74]. In OMZs, potential sulfur oxidizing microbial taxa have been identified as *Candidatus* Thioglobus autotrophicus [34] and ^{U}*T. perditus* within the Gammaproteobacteria SUP05 group [33, 75, 76], Epsilonproteobacteria *Sulfurimonas gotlandica*, and Deltaproteobacteria SAR234 [27, 77].

Many of the enzymes responsible for sulfate reduction, such as dissimilatory sulfate reductase (Dsr) and adenyl-sulfate reductase (Apr) are reversible and can carry out the later steps of H_2S oxidation to SO_4^{2-} as well. Furthermore, the detection and measurement of many biologically active sulfur molecules is challenging due to extreme O_2 sensitivity and low throughput methodologies. Detection of genes involved in H_2S oxidation in non-sulfidic OMZ waters [18, 78, 79], suggests the presence of H_2S oxidizers as well as SO_4^{2-} reducers (to supply the reduced sulfur compounds) within suboxic waters. Measurements of SO_4^{2-} reduction and H_2S oxidation rates in non-sulfidic OMZs have been carried out with detection of both activities occurring concurrently within the water column [78]. However, identification of canonical SO_4^{2-} reducing bacteria in OMZs has been elusive. These results support the presence of a cryptic sulfur cycle within (non-sulfidic) OMZ waters where reduced sulfur compounds are biologically produced and oxidized so rapidly that reduced sulfur compounds are unable to accumulate within the water column [78]. These redox reactions serve to fuel chemotrophic energy metabolisms and support chemoautotrophic activities, linking the sulfur and carbon cycles within OMZs [78]. Indeed, high abundance of SUP05 Gammaproteobacteria ^{U}*T. perditus* within O_2-deficient advected shelf water off the coast of Peru coincide with high abundance of elemental sulfur, the first product of sulfide reduction, detectedin the water column [36]. Rates of carbon fixation by ^{U}*T. perditus* were highest when sulfide content of the individual cells was high as well, supporting a strong coupling of carbon fixation and sulfide oxidation [36].

1.2.3 Carbon fixation in OMZs

Most OMZs exist below the photic zone such that sunlight is not available to support carbon fixation by photosynthesis. However, the abundance of redox active molecules serve as energetic substrates to support several pathways for chemoautotrophic carbon fixation, collectively termed dark carbon fixation. Within OMZs different taxa utilize different carbon fixation pathways, however, information on taxonomic distribution of pathways is limited. Ammonia oxidizing archaea, *Thaumarcheaota* sp. use the 3-hydroxypropionate/4-hydroxybutyrate cycle [52, 80]; NO_2^- oxidizers *Nitrospira* and *Nitrospina* use the reverse tricarboxylic acid cycle ([28, 55]); Planctomycetes use the reductive acetyl-CoA pathway [66] and denitrifying sulfur oxidizing Gammaproteobacteria

of the SUP05 lineage, including *Ca.* T. autotrophicus and ^{U}*T. perditus*, use the Calvin Benson Basham cycle [33, 34, 36]. In principle, the structure of the microbial community and available energetic substrates along redox gradients dictates the amount of carbon fixed and by which pathway and taxonomic group. While rates of dark carbon fixation have been observed for several different OMZ systems [81–84], the extent to which each of these chemoautotrophic groups is actively fixing carbon and under what conditions remains to be determined.

1.3 Microbes in community

Beyond the microbes studied in the laboratory setting of isolated cultures, microbes exist in communities, with various organisms carrying out a vast diversity of metabolic processes tuned to the geochemical conditions of their environment [85]. Within microbial communities the metabolisms of individual organisms or taxa overlap, with the community as a whole having functional redundancy that promotes co-metabolic interactions where individual steps for metabolic pathways are distributed across multiple taxa [2, 38, 86–89] . These co-metabolic interactions can take the form of discrete interactions where a metabolic intermediate made by one taxa is shared only with one other taxa or can occur much more generally and widespread as some taxa share with the community as a whole [90–96].

1.3.1 Co-metabolic interactions in microbial communities

Discrete co-metabolic interactions often manifest in symbiotic associations or obligate syntrophic interactions. For example, the symbiotic association between spilltebug (*Clastoptera arizonana*) and its symbionts *Sulcia muelleri* CARI and *Candidatus* Zinderia insecticola. *Ca.* Z. insecticola makes and shares amino acids tryptophan, methionine, and histidine and no others, while *S. mulleri* CARI makes the remaining 7 of 10 essential amino acids and shares with *Ca.* Z. insecticola [97]. Discrete co-metabolic interactions are also seen in obligate syntrophic interactions such as the anaerobic methanotrophic archaea of the ANME-2 lineage and sulfate reducing Deltaproteobacteria that are suggested to directly shuttle electrons from the anaerobic oxidation of methane in the ANME-2 to reduce sulfate in the Deltaproteobacteria [98].

Co-metabolic interactions can also be much more general and widespread in the community such as in the production of public goods, where a product (a metabolite or enzyme) is made by only a subset of the community and used by all others. For example, peroxide oxidase enzyme in surface ocean waters protects not only the organisms carrying the peroxide catalase gene but also numerous *Prochlorococcus* cells lacking this important protective gene [99]. This becomes particularly relevant in environmental conditions where resources are limited and individual species cannot bear the metabolic burden of a complete pathway [91, 100]. In a strategy known as the black queen hypothesis it becomes advantageous to lose supposedly essential metabolic genes if other members of the community can provide that resource [91, 101]. Additionally, in environments with highly limited energetic conditions, such as anoxic or sulfidic environments, it may be thermodynamically unfavourable for a single cell to carry out all necessary metabolic reactions and therefore metabolic pathways may become distributed across multiple members of the community [102]. For example, in highly reduced methanogenic conditions within a bioreactor, members of the candidate phyla Marinimicrobia are suggested to degrade protein and amino acids produced from the rest of the microbial community, some of which lack amino acid degradation pathways [103].

Within OMZs, the multitude of nitrogen-based energy metabolisms (discussed above) often overlap, with one organism feeding another, such as with AOA supplying NO_2^- to nitrite oxidizing bacteria. Alternatively, organisms may compete for energetic substrates, such as NO_2^- for anammox and denitrification [61]. Additionally, metabolic pathways may be distributed across the community where different taxa carry our different steps. The nitrogen and sulfur based energy metabolisms present in OMZ microbial communities represent an excellent platform on which to investigate these co-metabolic interactions as they are separate from carbon pathways and the multitude of associated carbon-based metabolites. The gradients found in OMZs offer an opportunity to observe how these interactions may change along these gradients.

1.3.2 Using multi-omics to study microbial communities

The nature of co-metabolic interactions occurring within microbial communities remains to be fully revealed as it is only recently that technological advances have supported production

of data that would permit the identification of such interactions at the level of the microbial community. As individual cells cannot be isolated from the community (by culturing or physical mechanisms) without changing gene expression, the cultivation-independent study of microbial communities in natural (and engineered) environments utilizes a set of tools referred to as multi-omics. Multi-omics follows the flow of biological information set forth in the central dogma of biology by sequencing DNA, RNA and protein, producing metagenomes, metatranscriptomes and metaproteomes respectively. Metagenomes give information about microbial community structure and metabolic potential while metatranscriptomes and metaproteomes give information about gene and protein expression, providing information about metabolic activities occurring under a given condition. Coupled to chemical and physical data from the environment, multi-omic datasets can provide insight into biogeochemical cycles and co-metabolic interactions along redox gradients. A drawback of the multi-omic approach is the loss of taxonomic resolution, the pairing of genes, transcripts and proteins with specific taxonomic origins. While gene sequences can be mapped back to related cultured (and sequenced) representatives, precise information about taxonomic origins requires additional tools and/or methodologies that remain computationally intensive. While there are limitations to multi-omic analyses, applications involving comparisons between different conditions, along gradients or over time are particularly useful in uncovering microbial community responses to a change [104].

Additional technological and computational advances are providing ways to better link taxonomy and function by reconstructing genomes of individual cells and microbial populations respectively. Technological advances are supporting the construction of single cell amplified genomes (SAGs) by physical isolation of cells directly from the envrionment, followed by whole genome apmlification and sequencing of the product [105]. The resulting SAGs are the partial genome (as whole genome amplification has some bias) of one individual organism with all functional genes directly linked to the taxonomy of that cell. Computational advances are supporting the construction of metagenome-assembled genomes (MAGs) by assembly of metagenomic contigs with overlapping sequences and similar K-mer frequencies [35]. The resulting MAGs represent the genome of a population of phylogenetically similar cells from a given environment. Both SAGs and MAGs can be powerful tools in linking taxonomy and function and better understanding the

metabolic roles played by different taxa.

The tremendous capacity of next generation DNA sequencing, coupled to both computational and technological advances in other areas (SAGs, MAGs, proteomics, metabolomics, microscopy etc.) is ushering in a new era of microbial ecology [106] supporting the study of microbial communities and novel methods for uncovering co-metabolic interactions and their role in biogeochemical cycles on a global scale.

Within OMZs, microbial communities along redox gradients carry out different metabolic activities, providing excellent systems to study how the microbial community and potential co-metabolic interactions shift across redox gradients, informed by the availability of energetic substrates. As co-metabolic interactions are often an emergent property of the community and its environment, it is necessary to study these processes *in vivo* within the environment rather than *in vitro* in a synthetic system in the laboratory. Thus a model natural environmental system is required for the application of multi-omics datasets within OMZs, providing both redox gradients and a natural microbial community. Using multi-omics to study microbial community structure and metabolic activities along environmental gradients can help to uncover fundamental principles underlying microbial community metabolism and co-metabolic interactions.

1.4 Saanich Inlet as a model OMZ

Saanich Inlet is a seasonally anoxic fjord on the east coast of Vancouver Island, British Columbia, Canada. A shallow sill at the entrance to the inlet, at a depth of 75m, restricts circulation in the basin waters that reach a depth of ~230m (Figure 1.6). The reduced circulation results in O_2-depletion in basin waters and permits the accumulation of H_2S and CH_4 from underlying sediments into the water column [107–109], creating redox gradients of O_2, NO_3^- and H_2S, like those found in sulfidic OMZs (Figure 1.2) [4]. Saanich Inlet has long been the site of oceanographic research and houses the Institute of Ocean Science (IOS), providing shore-based resources for ongoing research.

Decades of research and environmental monitoring dating back to the 1930's [110, 111] have documented a seasonal cycle of stratification, with intensifying O_2-depletion in basin waters

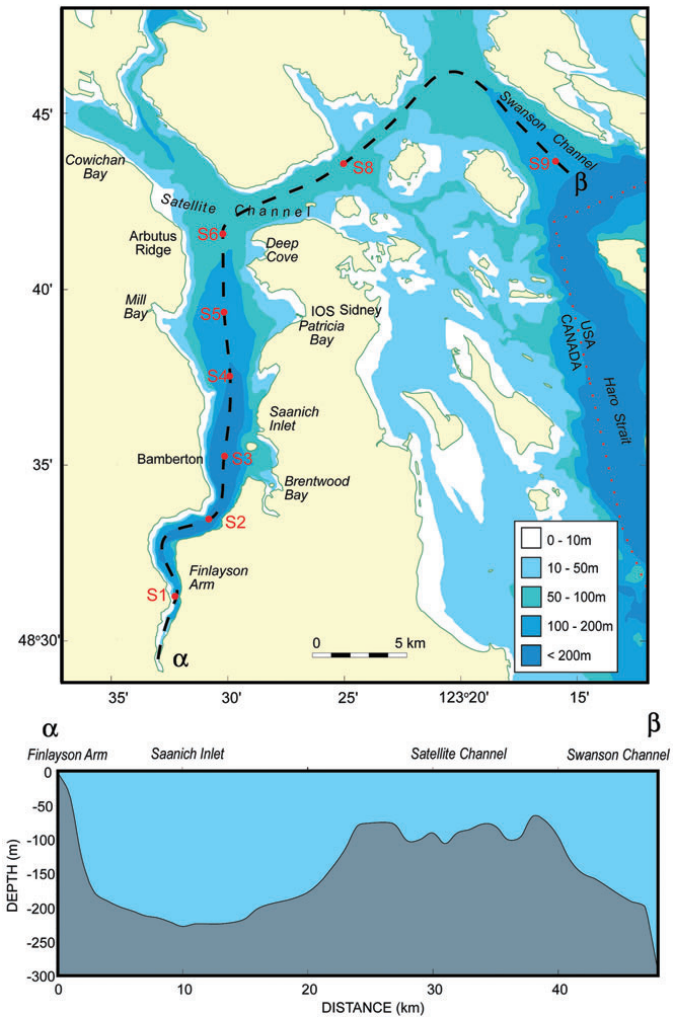


Figure 1.6: Saanich Inlet topology and bathymetry Saanich Inlet, on the east coast of Vancouver Island, indicating sampling station S3 and Institute of Ocean Science. Figure from Zaikova *et al.* 2010 *Microbial community dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia* in Environmental Microbiology.

through spring and summer, followed by deep water renewal as oxygenated waters creep into the basin during the fall months [8, 107, 111–113]. Through this cycle of stratification and renewal we see recurring gradients of O_2, NO_3^- and H_2S supporting different microbial communities along an overall redox gradient [8]. Furthermore, the O_2 profile of the Saanich Inlet water column through winter and into spring shows the gradual depletion of O_2 and NO_3^-, strongly resembling open ocean with O_2 concentrations ranging between 3 to 100 µM and average NO_3^- concentrations between 0 to 20 µM. In the later winter months through early summer the water column shows further depletion of O_2 and NO_3^- and accumulation of H_2S in the deep anoxic basin water,

resembling coastal sulfidic OMZs with H_2S concentrations ranging between 2 to 20 µM (Figure 1.2). Thus, Saanich Inlet serves as a model system for multiple OMZ types.

In efforts to understand the microbial community structure, metabolism and co-metabolic interactions in relation to biogeochemical cycles, an environmental and microbial monitoring time series has been underway in Saanich Inlet since 2006 (see Chapter 2)[114, 115]. An environmental time series can provide an extensive and robust dataset of multi-omic and environmental monitoring data such that environmental perturbations are observable in the microbial community structure and metabolic processes [92, 104]. The Saanich Inlet time series programme aims to monitor physical and chemical oceanographic characteristics and microbial community structure and metabolic activity within the context of OMZ nitrogen, sulfur and carbon cycles.

1.4.1 Saanich Inlet microbial community

The first study in the Saanich Inlet time series by Zaikova *et al.* used fosmid clone libraries and denaturing gradient gel electrophoresis of amplified small subunit ribosomal RNA gene to profile the changes in the microbial community along redox gradients and over time [8]. These results show an overall similarity to the structure of other OMZ microbial communities [2, 9, 116]. Ammonia oxidizing Archaea *Thaumarchaeota* (then termed *Crenarchaea*) were observed throughout the water column. Within the dysoxic waters at 100 m, the heterotrophic SAR11 group was found in relative abundance, along with the NO_2^- oxidizer *Nitrospina* and several other taxonomic groups. Deeper into the basin waters as H_2S began to accumulate, members of the Gammaproteobacteria SUP05 group became increasingly dominant, making up to 95% of the clone libraries in 200 m samples. Overall, the identification of patterns in microbial community structure along the Saanich Inlet redox gradient is similar to other OMZs, supporting the use of Saanich Inlet as a model oxygen minimum zone [2, 8, 9].

1.4.2 The SUP05 Gammaproteobacteria group in Saanich Inlet

The high abundance of Gammaproteobacteria SUP05 group in Saanich Inlet basin waters has supported the assembly of a draft population genome [33]. Members of the SUP05 group were originally detected in the Suyio Seamount hydrothermal plume where 88-90% of bacterial cells in

the plume layer were identified as SUP05 by fluorescence in situ hybridization [117]. Members of the SUP05 group also include sulfur oxidizing symbionts of deep sea clams and mussels living near hydrothermal vents and cold seeps [33, 117] where they act to detoxify reduced sulfur compounds and fix carbon for their hosts [118, 119]. Surveys of OMZs globally have found free-living SUP05 in many anoxic and sulfidic OMZs [2], including the Baltic Sea [76, 120], The Black Sea [121], the ETSP [18], Guaymas Basin [122], Namibian upwelling [75] and other inlets on the coast of British Columbia [34, 123]. In many of these systems, most notably in Saanich Inlet and Namibian upwelling, SUP05 is observed at highest abundance in the water column at the interface between NO_3^- and H_2S rich waters [33, 75], suggesting a redox driven energy metabolism.

The draft population genome for SUP05 from Saanich Inlet metagenomes indicated several energy metabolism pathways enabling it to thrive at the sulfide-nitrate interface in Saanich Inlet as well as other anoxic and sulfidic OMZs [18, 75, 76, 120–122, 124, 125]. Further, a recently sequenced SUP05 isolate, *Ca.* Thioglobus autotrophicus, had similar energy metabolism pathways [34]. Analysis of available SUP05 genomes revealed multiple pathways for denitrification and sulfide oxidation. With respect to denitrification, SUP05 harbours two interesting features. One is the apparent absence of the gene nitrous oxide reductase, *nosZ*, which catalyzes the reduction of N_2O to N_2, implicating SUP05 in the production of N_2O. Second is the presence of two functionally analogous enzymes mediating nitrate reduction, NarG and NapA. Membrane bound nitrate reductase NarG is common in a wide range of organisms and has been well characterized to function under anoxic, high nitrate conditions [126]. Alternatively, periplasmic nitrate reductase *napA* has so far been found only in proteobacteria, most often associated with *narG*. It is less well characterized, but studies in *E. coli* suggest that it functions under hypoxic, low nitrate conditions [127]. With respect to sulfur oxidation pathways, SUP05 harbors a reverse dissimilatory sulfate reductase pathway (rDsr) and the Sox sulfur oxidation system. The absence of *soxC* subunit implicates SUP05 in the formation of globules of elemental sulfur (S°) [74]. S° may be stored and oxidized as needed to sulfite (SO_3^-) by rDsr, allowing for an adaptive energy metabolism [74]. In addition, two enzymes mediating the initial step of H_2S oxidation to S°, Flavocytochrome/sulfidedehydrogenase (FccAB) and Sulfide:quinone oxidoreductase (sqr) were

identified. In some organisms FccAB has been observed to function under low sulfide conditions [74], while Sqr is believed to function under fully anoxic conditions [74]. These observations echo the nitrate reduction pathway in providing alternative routes to fuel SUP05 energy metabolism under changing water column conditions.

1.4.3 Saanich Inlet as a model OMZ system

The similarity of the Saanich Inlet microbial community to other OMZs [2, 4, 8] and recurring gradients of O_2, NO_3^- and H_2S provide an excellent model system to utilize a multi-omics approach for the study of OMZ biogeochemical cycles and co-metabolic interactions along redox gradients. An environmental time series can provide an extensive and robust dataset of multi-omic and environmental monitoring data such that environmental perturbations are observable in the microbial community structure and metabolic processes [92, 104]. Indeed, while Saanich Inlet is stratified throughout the summer months, annual renewal with oxygenated waters results in a chemical profile of the water column in winter and early spring similar to that of Open-ocean or Anoxic OMZs (Figure 1.2) [4, 8].

1.5 Thesis objectives and overview

Using Saanich Inlet as a model OMZ, I explore the energy metabolism and carbon fixation pathways of the microbial community along redox gradients and utilize multi-omic datasets. I identify potential co-metabolic interactions and how the availability of energetic substrates may serve to inform those interactions. I gain insight into taxonomic groups governing key processes in the nitrogen and sulfur cycles including novel taxonomic lineages carrying and expressing an otherwise unidentified N_2O reductase gene. Following these investigations in Saanich Inlet I compare selected results to other marine and OMZ systems, providing a global context for biogeochemical cycles and key microbial players.

Chapter 2: Methodologies and work flows for generating and processing multi-omic datasets

In Chapter 2 I detail the multiple datasets used throughout this thesis, including chemical and physical, SSU rRNA amplicon, metagenomic, metatranscriptomic, metaproteomic and singe-cell amplified genomes. I include both wet lab and *in silico* protocols developed through the course of the thesis project. I also include notes on the application of MetaPathways annotation pipeline to multi-omic and next generation sequencing datasets.

Chapter 3: Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes

In Chapter 3 I use metagenomic and metaproteomic workflows outlined in Chapter 2 to generate and analyse SSU rRNA tag, metagenome (specifically Sanger sequenced fosmid libraries), and metaproteomic datasets. Using these data from multiple stations and time points in Saanich Inlet I uncovered shifts in the microbial community structure and associated metabolic activities along redox gradients for dominant nitrogen and sulfur transformations and carbon fixation pathways. Additionally, I detected evidence of gene regulation within the Gammaproteobacteria SUP05 group for a combined nitrogen sulfur and carbon fixation operon and extend these results to SUP05's potential impact on global carbon budgets in OMZs globally.

Chapter 4: Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients

In Chapter 4 I explored the global distribution and metabolic capacity of the dark matter phylum *Marinimicrobia*. To do this, I use metagenomes and metatranscriptomes and associated methodologies outlined in Chapter 2 as well as 24 single-cell amplified genomes (SAGs) from seven different environments globally, representing 10 distinct clades. Comparison to over 200 environmental metagenomes globally indicates both cosmopolitan and endemic clades, as well as one clade restricted to sulfidic OMZ waters. I find the evolutionary diversification of major *Marinimicrobia* clades to be closely related to energy yields, with increased co-metabolic interactions in more deeply branching clades. Several of these clades participate in the biogeochemical cycling of sulfur

and nitrogen, filling previously unassigned niches in the ocean. Notably, two Marinimicrobia clades, occupying different energetic niches, express nitrous oxide reductase, potentially acting as a global sink for the greenhouse gas nitrous oxide.

Chapter 5: A niche for NosZ?

In Chapter 5 I take a closer look at the distribution and expression of the nitrous oxide reductase (*nosZ*) gene both in Saanich Inlet and globally. Using the collection of SAGs from Saanich Inlet to phylogenetically anchor 7 distinct *nosZ* sequences and map their global abundance and expression using available metagenomes and metatranscriptomes from other marine environments in addition to Saanich Inlet. Additionally, I explore the seasonal dynamics of these *nosZ* types in Saanich Inlet time series multi-omic datasets.

Chapter 2

Methodologies and workflows for generating and processing mulit-omic datasets[1]

This chapter describes the methodologies and workflows for generating multi-omic datasets including SSU rRNA amplicon (tag) sequencing, metagenomes, metatranscriptomes, metaproteomes, single-cell amplified genomes (SAGs) as well the as the processing of metaproteomic datasets and use of the MetaPathways annotation pipeline. These methodologies form the basis of multi-omic analyses carried out throughout this thesis in the investigation of microbial energy metabolism and co-metabolic interactions along redox gradients in oxygen-depleted waters.

2.1 Introduction

Microbial communities are the primary engines driving biogeochemical cycles on our planet, playing essential roles in carbon, nutrient and energy cycling and oxygen (O_2) production [38, 128]. Many biogeochemical cycles have key reactions occurring within the O_2-depleted waters of marine oxygen minimum zones (OMZs). OMZs are areas of O_2-depleted subsurface waters that form as a result of the combination of microbial respiration of organic matter from surface waters and decreased ventilation or mixing [2, 6]. Under O_2-depleted conditions microbes use

[1]Selected methodologies presented in this chapter have been published in Methods in Enzymology as *Molecular tools for investigating microbial community structure and function in oxygen-deficient marine waters* in 2013 by Hawley, A. K., Kheirandish, S., Mueller, A., Leung, H. T., Norbeck, A. D., Brewer, H. M., Pasa-Tolic, L. and Hallam, S. J.. Selected datasets presented in this chapter are published in part or in whole in Scientific Data as *A compendium of water column multi-omic sequence information from a seasonally anoxic fjord Saanich Inlet* in 2017 by Hawley, A K., Torres-Beltrn, M., Bhatia, M., Zaikova, E., Walsh, D. A. et al. and as *A compendium of water column chemistry from the seasonally anoxic fjord Saanich Inlet* in 2017 by Hawley, A. K. and Torres-Beltrán, M. and Bhatia, M. P. and Zaikova, E. and Walsh, D. A..

alternative terminal electron acceptors such as nitrate (NO_3^-), nitrite NO_2^-, sulfate (SO_4^{2-}) and carbon dixoide (CO_2), producing potent greenhouse gases such as nitrous oxide (N_2O) and methane (CH_4) and toxic hydrogen sulfide (H_2S). Currently, OMZs are expanding and intensifying on a global scale [10, 11, 129], making it increasingly important to characterise the microbial communities and associated metabolic activities and biogeochemical cycles.

Saanich Inlet, a seasonally anoxic fjord in British Columbia, Canada, is a model OMZ with an on-going 10 year time series monitoring geochemical parameters, microbial community structure and activity as well as viral populations [8, 33, 115, 130, 131]. Annual cycles of O_2-depletion and development of anoxic-sulfidic basin waters followed by influx of oxygenated nutrient rich waters provides recurring gradients of O_2, NO_3^- and H_2S, making Saanich Inlet a model system for studying open-ocean, anoxic and sulfidic OMZs (Figure 1.2) [4]. An overall redox gradient, created by gradients of O_2, NO_3^- and H_2S, supports microbial communities involved in biogeochemical cycles and co-metabolic interactions, reminiscent of symbiotic associations [38, 86–89, 95, 96]. Although current research efforts are increasingly focused on defining interaction networks within microbial communities [92], many open questions remain regarding the regulatory and ecological dynamics modulating microbial community structure, function and activity both along gradients of O_2-depletion and over time. Temporally resolved datasets are necessary for the development of robust metabolic and climate models and incorporating environmental sequence information that predicts future responses to OMZ expansion [3, 129, 132] as evidenced in current efforts by the Scientific Committee on Oceanic Research (SCOR) Working Group 144 Microbial Community Responses to Ocean Deoxygenation (`http://www.scor-int.org/SCOR_WGs_WG144.htm`).

Investigations into microbial community structure, biogeochemical cycling and co-metabolic interactions introduced in Chapter 1 have all benefited from the recent advances in sequencing technologies, enabling the study of microbial communities at the molecular level in a cultivation-independent manner. Amplicon or tag sequencing uses primers to amplify and subsequently sequence individual genes, often taxonomic markers such as specific regions of the small or large subunit rRNA genes (SSU or LSU). These tags provide a high-throughput method for fingerprinting the microbial community structure, which can be used to compare across samples and environments. Metagenomics, sequencing DNA from environmental microbial communities

and assembling sequence reads into contiguous regions containing multiple genes from a given population of organisms, provides an inventory of the metabolic potential present in a microbial community. However, metabolic potential does not directly correlate with metabolic activity, as different genes may be expressed by the same microbial population under different environmental conditions [133]. Metatranscriptomics, sequencing cDNA generated from RNA from environmental microbial communities and assembling either de novo or aligning to a corresponding metagenome, more closely represents the gene expression of the microbial community found in that specific environment. Moreover, additional levels of regulation occur after transcription at the level of protein production and degradation. Metaproteomics, sequencing the proteins from environmental microbial communities via tandem mass-spectroscopy, offers yet another perspective on microbial gene expression. Single-cell amplified genomes (SAGs), genomes of single cells collected from an environmental sample, can be used to directly link the taxonomy with functional genes, providing a more complete picture of co-metabolic interactions in the environment. Throughout all multi-omic analyses consistent and thorough determination of genes or open reading frames and functional annotation of those genes is paramount to the identification of biogeochemical cycles carried out by microbial communities along redox gradients present in OMZs and other environments [79, 94, 134].

2.2 Sampling and multi-omic dataset overview

Sampling was carried out monthly on-board the MSV John Strickland in Saanich Inlet at station S3 (480 35.500 N, 123 30.300 W) (Figure 1.6) and occasionally at other stations where stated. Station S3 is one of the deepest points in the inlet, providing recurring gradients of O_2, NO_3^- and H_2S along which to sample the geochemical parameters and the associated microbial community and metabolic activity. Standard physical oceanographic parameters were measured with probes via a Conductivity-Temperature-Depth (CTD) instrument attached to a line that was lowered into the inlet and provided readings for every meter. Discrete samples for chemical analysis and microbial community structure by pyrotags (amplicon sequencing of the V6-V8 region of the SSU rRNA gene) were taken at 16 high resolution (HR) depths along the oxycline (10, 20, 40, 60, 75, 80, 90,

97, 100, 110, 120, 135, 150, 165, 185 and 200 meters). Samples for multi-omics were taken at six major depths (LV) spanning the oxycline (10, 100, 120, 135, 200m) (Figure 2.1). Details of sample collection and processing are in the following text.

Eight years of monthly sampling have generated a dataset of matched multi-omic and geochemical measurements. In its entirety, Saanich Inlet time series consists of 100 time points of geochemical profiles, 412 SSU rRNA pyrotag samples, 82 SSU rRNA iTag (V4 region) samples, 90 metagenomes, 62 metatranscriptomes (including 46 unique samples and 16 replicates,) and 68 metaproteomes (64 unique samples) and 378 single-cell amplified genomes. Additional viral fraction metagenomes and fosmid libraries were also generated (Figure 2.1). Together these datasets serve as a primary resource for observing shifts in microbial community structure and metabolic activities in response to changing environmental parameters along redox gradients. These datasets of coupled multi-omic sequence information span five years of the Saanich Inlet time series and serve as a much needed resource for microbial ecology and environmental modelling efforts and contribute an significant volume of time series data, similar to other time-series datasets such as Hawaii Ocean Time Series and Bermuda Ocean Time Series.

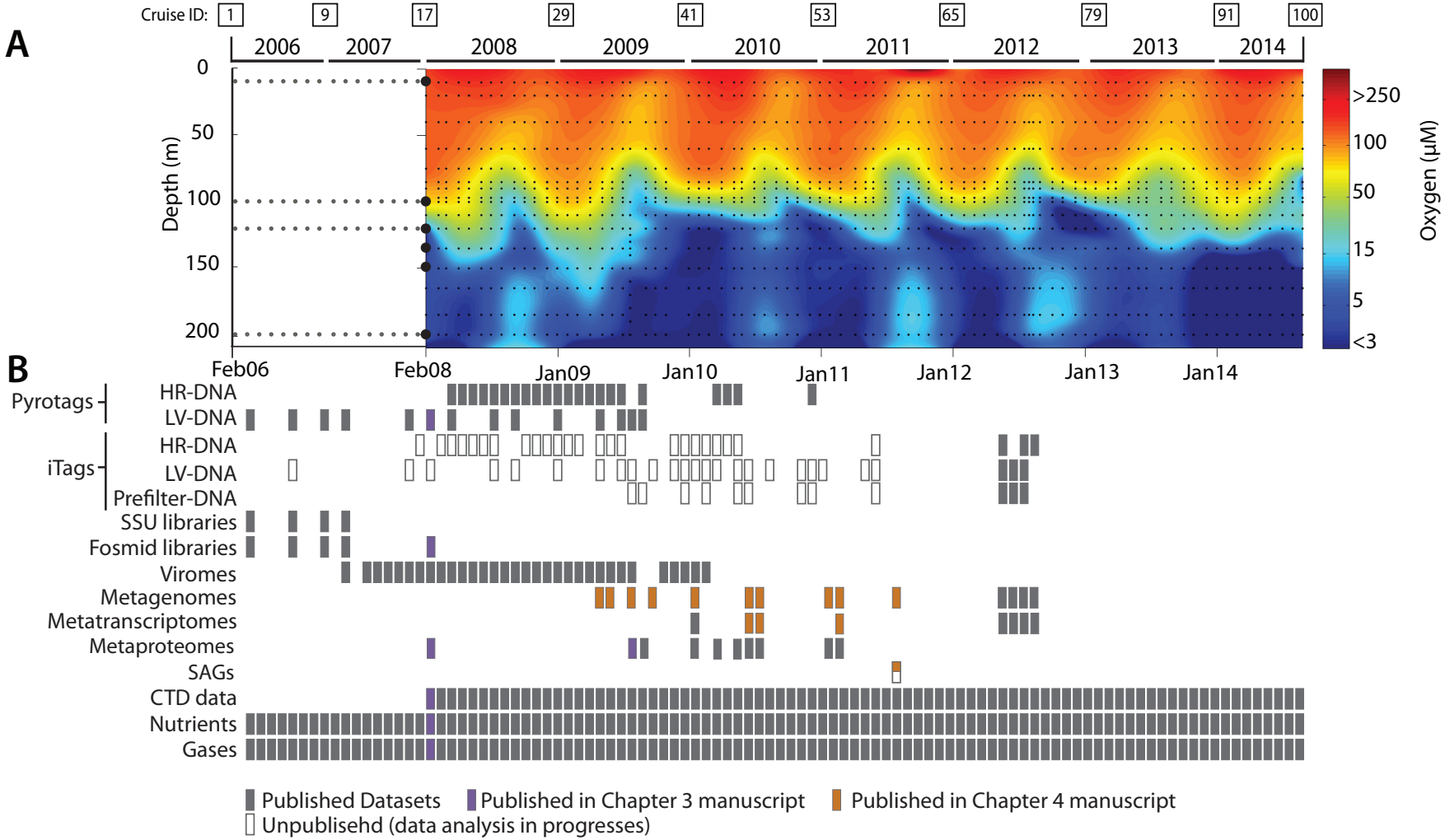


Figure 2.1: Multi-omics sampling overview. (A) Oxygen concentration contour for CTD data (February 2008 onward)8, with points for 16 sampling depths water chemistry and high-resolution (HR) DNA samples for SSU libraries (small black dots) and six major depths for large volume (LV) depths sampled for metagenomics, metatranscriptomics, metaproteomics and LV SSU libraries (depth indicated by large black dots). **(B)** Sample inventory from February 2006 to October 2014 indicating multi-molecular datasets included in this manuscript (solid black), in previous publications (gray) and accompanying datasets currently in analysis (open gray).

2.3 Establishing water column chemistry

The chemistry and redox gradients along the depth of the water column are the primary driver of microbial community structure and associated metabolic activity. Measurements of several geochemical parameters along the water column were taken in order to provide environmental context for metabolic activity of the microbial community. Standard physical oceanographic parameters were measured with probes mounted on a Conductivity-Temperature-Density (CTD). The CTD was attached to a line that was lowered into the inlet and provided readings for every meter, and samples were taken from Niskin and Go-Flow bottles attached to the same line. Samples for chemical parameters were taken at the 16 HR depths. Table 2.1 provides the measured physical and chemical parameters; the referenced methodologies and complete descriptions with data are available in Torres-Beltrán and Hawley *et al.* [115] and an on-line visualised protocol is available at `http://www.jove.com/video/1159/seawater-sampling-and-collection40`.

The recurring annual cycle of stratification and renewal in Saanich Inlet provides patterns of O_2 depletion followed by NO_3^- depletion and development of H_2S in basin waters. These recurring patterns provide a baseline for environmental monitoring as well as observations of microbial community structure and activity in addition to a vehicle for identification of environmental conditions that deviate from normal cycles. Indeed, a weak renewal in September 2009 and subsequent build up of H_2S to higher concentrations and shallower depths throughout 2010 had impacts for microbial community structure and function. It is such perturbations and associated changes in biogeochemical processes that the Saanich Inlet time series seeks to chart and utilise for understanding microbial community dynamics and future climate modeling efforts.

Table 2.1: Physical and chemical parameters and protocols.

Parameter	Instrument or Protocol	Reference
Physical Parameters		
Temperature	CTD	
Denisty	CTD	
Salinity	CTD	
Transmisivity	CTD	
Conductivity	CTD	
Dissolved Oxygen	CTD	
Chemical Parameters		
Nitrate	Bran Luebbe AutoAnalyser	Armstrong *et al.* 1967 [135]
Nitrite	Cary60 spectrometer	Armstrong *et al.* 1967 [135]
Ammonium	Varioscan FLASH (ThermoScientific)	Holmes *et al.* 1999 [136]
Hydrogen Sulfide	Hach kit reagents Varioscan FLASH (ThermoScientific)	Cline 1969 [137]
Silicic acid	Bran Luebbe AutoAnalyser	Armstrong *et al.* 1967 [135]
Phosphate	Bran Luebbe AutoAnalyser	Murphy and Riley 1962 [138]
Dissolved Oxygen	Winkler	Winkler 1888 [139]
Dissolved Gases		
Oxygen	headspace GC/MS	Zaikova *et al.* 2010 [8]
Dinitrogen	headspace GC/MS	Zaikova *et al.* 2010 [8]
Carbon Dioxide	headspace GC/MS	Zaikova *et al.* 2010 [8]
Nitrous Oxide	headspace GC/MS and purge and trap GC/MS	Zaikova *et al.* 2010 [8] and Capelle *et al.* 2015 [140]
Methane	headspace GC/MS and purge and trap GC/MS	Zaikova *et al.* 2010 [8] and Capelle *et al.* 2015 [140]

2.4 Exploring oxygen minimum zone community structure

Several methods for investigating microbial community structure have been used with varying degrees of taxonomic resolution and quantitative power in O_2-depleted waters. These include amplicon-based methods such as terminal restriction length polymorphism (T-RFLP) or denaturing gradient gel electrophoresis (DGGE) of SSU rRNA genes [8, 141–144], SSU rRNA gene clone library sequencing[8, 17, 145–149], and massively parallel tag sequencing[31, 150–152]. While T-RFLP and DGGE are inexpensive and amenable to automation, peak resolution is limited and taxonomic identification requires secondary purification and sequencing steps. Clone libraries can provide significantly more taxonomic information per read. However, quantitative power is limited by the cost of paired-end Sanger sequencing. Conversely, small subunit rRNA (SSU rRNA) tag sequencing currently provides less taxonomic information per read than clone libraries but significantly more quantitative power. Catalyzed reporter deposition (CARD) and fluorescent in situ hybridization (FISH) have also been used in community composition [141, 142, 146, 153] and group-specific studies targeting SUP05 [75, 154], Marine Group A [31] and Planctomycetes [14, 65]. While CARD-FISH provides effective group-specific quantitation, probe development and optimization can be cost prohibitive and technically demanding when profiling large numbers of samples. Finally, plurality sequencing methods (aka metagenomics and metatranscriptomics) are have also been effectively used to derive taxonomic information from O_2-depleted waters

[4, 18, 78, 155].

Quantitative polymerase chain reaction (qPCR) using dye assay chemistry such as SYBER-Green® or EvaGreen® can be a specific alternative or adjunct to sequencing and CARD-FISH methods. For example, qPCR using 5′endonuclease probe-based chemistry (Taqman) or dye assay chemistry (SYBRGreen®) have been successfully adapted for rapid and high-throughput quantification of microbial populations in seawater [156–158]. In O_2-depleted waters, qPCR has been used successfully to quantify functional gene expression of nitrogen cycling genes [60, 61]. Moreover, domain specific primers targeting total bacteria and archaea, and group specific primers targeting SUP05 and Arctic96BD-19 SSU rRNA gene copy number have been used in SYBR Green®-based qPCR assays to monitor secular changes in microbial community structure [8, 149]. The use of group specific primers provides quantitative assessments of taxon abundance needed to accurately describe and monitor population dynamics in response to changing levels of water column O_2-depletion.

2.4.1 Sample collection and DNA extraction for metagenomics and SSU rRNA tag sequencing

Multiple methods for environmental DNA (eDNA) extraction from seawater exist and no single method will unfailingly provide ultraclean nucleic acids in sufficient quantity and quality to support multiple sequencing platforms (both metagenomic and tag sequencing) and qPCR applications without prior optimization. Throughout Saanich Inlet time series, I have used methods involving a peristaltic pump to concentrate biomass from seawater onto a 0.2 µM Sterivex filter using an in-line 2.7 µm polycarbonate pre-filter to remove the bulk of the eukaryotic organisms. Samples for SSU rRNA tag sequencing were collected from 1 to 2 l of seawater at the 16 high resolution depths wtih no pre-filter. Samples for SSU rRNA and metagenomics were collected from 20 L six large volume depths and included pre-filtering. This method had been previously developed and documented in Zaikova *et al.* [8] and in the Journal of Visualized Experiments (JoVE) at `http://www.jove.com/video/1161/large-volume-20l-filtration-of-coastal-seawater-samples`[159]. Extraction of DNA from Sterivex for the Saanich Inlet time series was previously developed and documented in Zaikova *et al.* [8] and in JoVe at `http://www.jove.com/video/1352/dna-extraction-from-022\`

`-m-sterivex -filters-cesium-chloride-density23` [160].

2.4.2 SSU rRNA tag sequencing

SSU rRNA tags (Pyrotags or iTags) were generated from extracted environmental DNA. Pyrotag datasets from HR and LV samples were generated by PCR amplification using universal three-domain forward and reverse bar-coded primers targeting the V6-V8 hypervariable region of the 16S or 18S rRNA genes26: 926F (5'-AAA CTY AAA KGA ATT GRC GG- 3') and 1392R (5'-ACG GGC GGT GTG TRC- 3'). Samples were purified using the QIAquick PCR Purification Kit (Qiagen), and sequenced by 454-pyrosequencing at the Department of Energy Joint Genome Institute (JGI) in California USA, or Génome Québec Innovation Centre at McGill University. iTag datasets from HR and LV samples were generated by PCR amplification using forward and reverse bar-coded primers targeting the V4-V5 hypervariable region of the 16S rRNA bacterial gene: 515F (5'-Y GTG YCA GCM GCC GCG GTAA- 3') and 806R (5'-CCG YCA ATT YMT TTR AGT TT- 3') [161, 162]. Samples were sequenced according to the standard operating protocol on an Illumina MiSeq platform at the JGI. Quality control protocols were similar for both sequencing centers and generally followed manufactures specifications for the respective sequencing platforms). For produced 454 pyrotag datasets from both high resolution (HR) and large volume (LV) samples a histogram of raw read counts verses read length (Figure 2.2 is used to determine the success of a run. A successful run will have the majority of reads >450 base pairs. A plot of read counts versus read length for all HR and LV samples is provided in Figure 2.2.

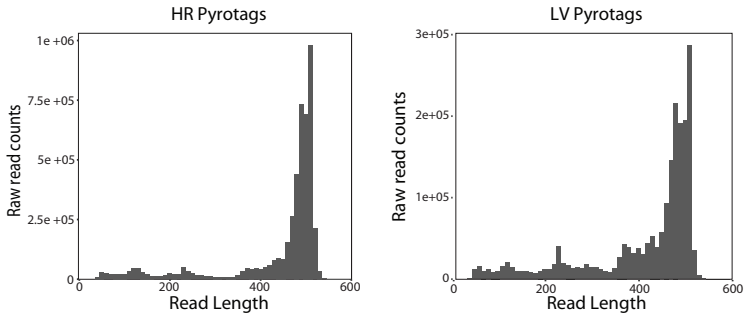


Figure 2.2: SSU rRNA tag sequencing validation 454 Pyrotags for small subunit rRNA gene showing number of raw reads versus read length for large volume samples (99 samples in total) (left) and high resolution samples (311 samples in total) (right).

2.4.3 DNA sequencing

Ilumina metagenome shotgun libraries from LV samples were generated at the JGI and paired end sequenced on the Illumina HiSeq platform. JGI quality control protocol for metagenomic sequences prior to assembly entails rolling QC, an in-house sequence QC pipeline that performs a set collection of analyses and produces a summary report for each lane of Illumina data produced by the sequencing group. This set of analyses calculates read quality, measures sequence uniqueness, and detects abnormal sequence motifs. An assembly, using Velvet, is used to measure coverage and detect contamination [163]. For individual sample assemblies the average fold coverage versus the contig length (Figure 2.3) is plotted and should have a distinct shape for different samples with peaks in contig length representing at a specific coverage represent a given closely related microbial population. Additionally, the percent GC versus the average fold coverage can be plotted, again with distinct shapes for different samples and clusters representing closely related microbial populations.

2.4.4 Metagenomic samples

A total of 90 metagenomic samples (Table A.1) were generated covering 14 time points including two renewal periods in August/September and samples from multiple stations in September 2009. These metagenomes provide means to explore metabolic pathways, taxonomic affiliations and gene abundances (see section on MetaPathways annotation pipeline below) along redox gradients.

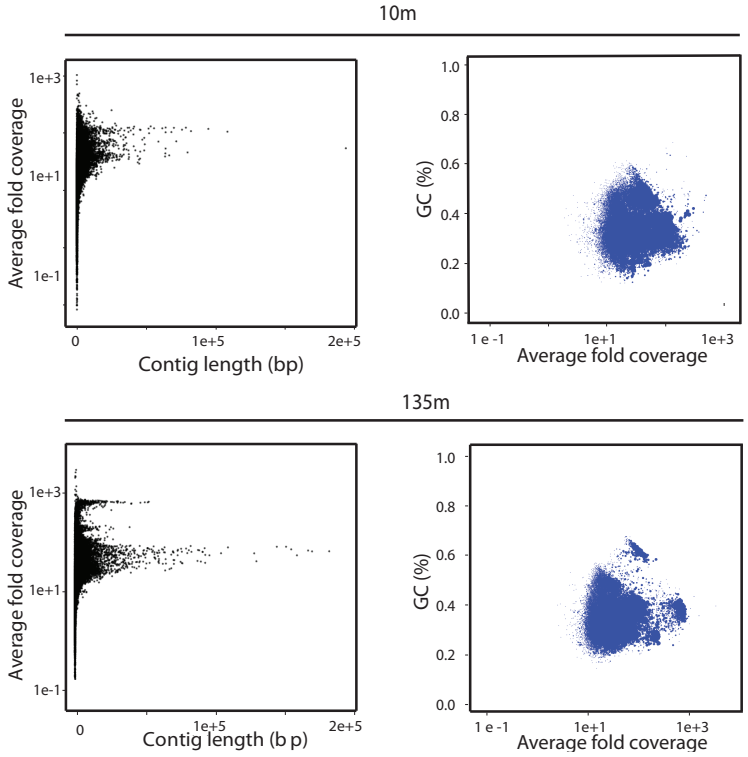


Figure 2.3: Metagenomic sequencing validation Metagenomic assemblies for two samples from different depths showing average fold coverage versus contig length and percentage GC versus average fold coverage for contigs.

2.5 Exploring oxygen mimimum zone microbial community transcriptome

Exploration of microbial community transcriptional activity has generally taken two forms, either Reverse Transcriptase-Quantative Polymerase Chain Reaction (RT-qPCR) to quantify the number of copies of a given transcript or using community wide metatranscriptomic studies. Both methods require the collection and extraction of high quality RNA from the microbial community. Within laboratory grown cells the total RNA pool in any given microbe consists predominantly of ribosomal RNA (rRNA), with low numbers of messenger (mRNA). Within cells under environmental conditions of the ocean the ratio of mRNA to rRNA is even lower, with an estimated 200 mRNA molecules per cell [133]. Combined with the short half-life of most mRNA molecules the sampling time and extraction efficiency for metatranscriptomics become critical.

Sampling and extraction protocols have been designed to maximise RNA yield while minimizing degradation and time between collection and freezing.

2.5.1 Sample collection and RNA extraction

Similar to DNA extraction, multiple methods for RNA extraction from seawater exist and are used based on the particular samples and needs of the user. Sample collection focuses on minimizing time between collection and filtration and freezing, while extraction focuses on extraction efficiency as well as limiting degradation. Sea water sampling and concentration of biomass onto Sterivex is nearly identical to metagenomic methods, including use of an in-line 2.7 µm filter to remove the bulk of eukaryotic organisms. RNA extraction protocol was based on Shi *et al.* 2009 [164] where total RNA was extracted from Sterivex using an mirVana RNA isolation kit (Ambion). Detailed protocol developed to maximise extraction efficiency and RNA quality is found in Appendix A.1 *RNA extraction and isolation protocol*.

2.5.2 RNA sequencing

Purified total RNA was used to generate paired end sequenced Illumina metatranscriptome libraries at the JGI and sequenced there on the HiSeq and MiSeq platform. The quality of purified RNA was verified on the Bioanalyzer using a RNA nano Analysis Kit (Agilent Technologies) in order to check on the RNA integrity and sample quantitation before cDNA library production and sequencing. JGI quality control protocol for metatranscriptomic sequencing preparation follows the *TruSeq Stranded Total RNA Sample Preparation Guide* (Illumina). Briefly this protocol entails the removal of ribosomal rRNA with RiboZero, followed by RNA fragmentation for first strand cDNA synthesis. This is followed by second strand synthesis and the subsequent ligation of the adapters. After PCR amplification, library quality is checked using Bioanalyzer for fragment size (260bp) and purity. Indexed (barcoded) libraries are normalized to 10nM and pooled in equal volumes. Transcriptomes are assembled de novo and or mapped to a corresponding metagenomes. For additional quality assessment of sequencing run for each sample, histograms of percentage of reads verse average read quality and of reads per percent GC are generated (Figure 2.4).

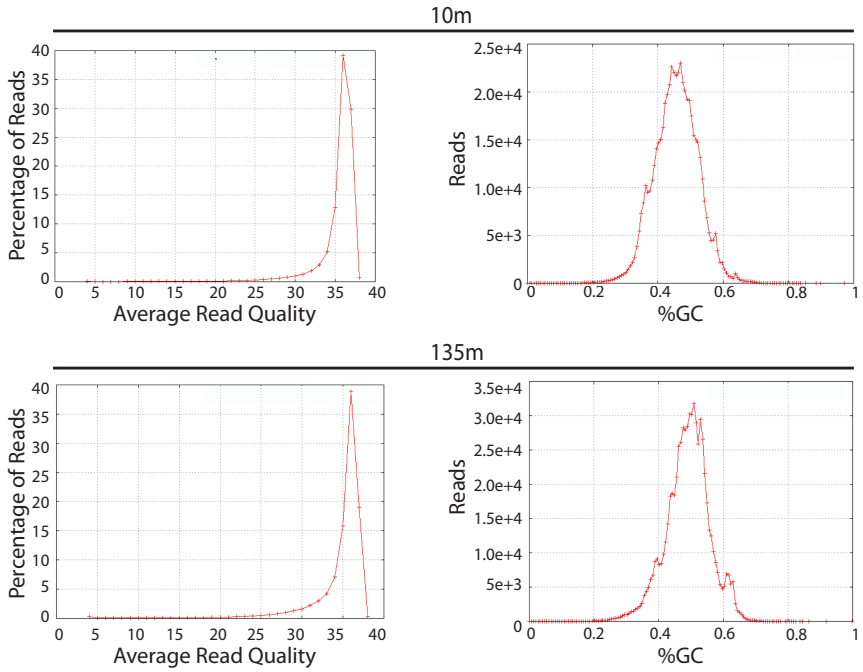


Figure 2.4: Metatranscriptomic reads for two samples from different depths showing distribution of reads over read quality (left) and percentage GC (right).

2.5.3 Metatranscriptomic Samples

A total of 62 metatranscriptomic samples (46 unique samples and 16 replicates) (Table A.2) were generated covering 8 time points including two renewal periods in August/September. These metatranscriptomes provide means to explore differential gene expression for genes involved in energy production and other metabolic activities.

2.6 Exploring oxygen minimum zone microbial community proteome

Environmental proteomics also known as metaproteomics was first used to describe microbial community gene expression in an acid mine drainage ecosystem [165]. Reduced community complexity in the acid mine milieu enabled the identification of key metabolic activities and metabolic partitioning between community members. Since that time, metaproteomic approaches have been successfully applied to a wide range of natural and human-engineered ecosystems including soils [166, 167], leaf surfaces [168], human guts [169], napthalene-degrading enrichment

cultures [170], and wastewater treatment plants [171, 172]. Although no metaproteomes for OMZ microbiota have been reported, surface ocean surveys have provided insight into microbial community responses to nutrient conditions along a coastal to open-ocean transect in the South Atlantic [173], coastal northeast Pacific Ocean upwelling [174], and winter to summer transitions off the Antarctic Peninsula [175]. Indeed, metaproteomics opens a functional window into microbial community metabolism and coupled biogeochemical cycles needed to monitor microbial community responses to changing levels of water column O_2-depletion.

2.6.1 Sample collection and protein extraction

Protein extraction yields per unit volume of seawater need to be considered prior to large-scale sample collection to ensure sufficient biomass is filtered for downstream processing and detection steps. Empirical observations suggest that a minimum of 108 cells is needed to reliably detect abundant proteins under these water column conditions using nano-high-performance liquid chromatography coupled to a Thermo Electron LTQ-Orbitrap mass spectrometer with electrospray ionization. Protocols have been designed to minimize time between sample collection and processing and freezing as not to alter the proteome of the microbial community. Detailed protocol for total protein extraction and peptide detection is provided in Appendix A.2 *Protein extraction and isolation protocol*.

2.6.2 Protein sequencing

Tandem mass spectrometry and peptide identification

While the detection and quantification of potential key microbial players in O_2-depleted waters provides insight into community structure and dynamics, additional methods for profiling environmental gene expression are needed for gene model and pathway validation. The application of high-pressure liquid chromatography (HPLC) coupled tandem mass spectrometry to identify expressed protein sequences from O_2-depleted waters offers a rapid and high-throughput profiling solution. The most effective peptide matching relies on the availability of environmental sequence information derived from the ecosystem under study although a standard reference database

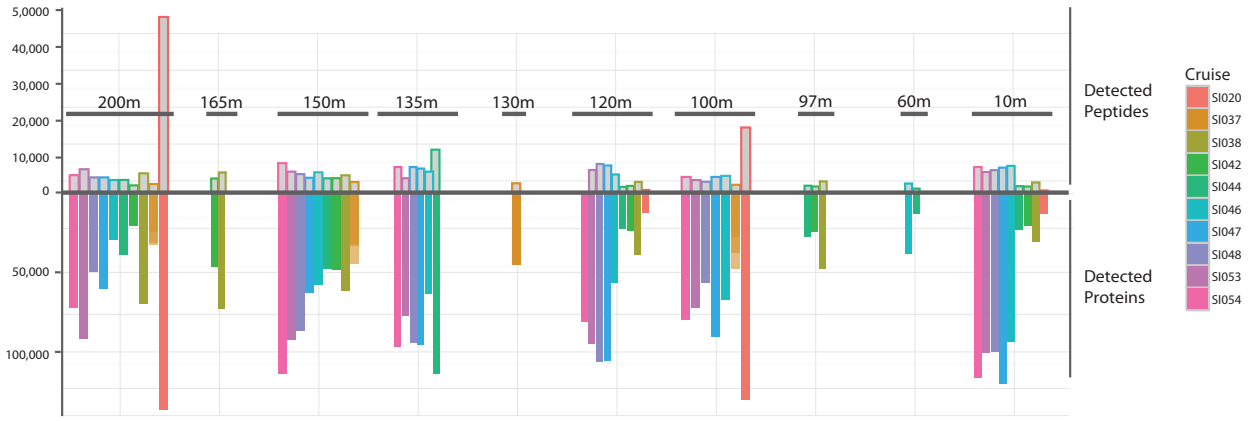


Figure 2.5: Protein validation. Metaproteome of identified peptides (top) and detected proteins (bottom) for each depth samples, colour coded by cruise ID for peptides matched to Saanich Inlet Illumnia sequenced metagenomic database. Higher number of detected proteins than peptides is due to the sequence redundancy in the metagenomic database used to identify proteins.

compiled from cultured isolates and publically available marine metagenomes can also be utilized. For analyses presented in this thesis I utilized a database of conceptually translated protein sequences from Saanich Inlet metagenomes (Chapter 3 utilizes protein sequences from paired-end fosmid and whole-genome shotgun sequences, Chapter 5 utilizes protein sequences from Illumnia platform sequenced metagenomes encompassing over 23 million protein sequences (Figure 2.1). Programs for matching peptide spectra to protein sequences varried depending on analysis (Chapter 3 utilizes the SEQUEST™ program, Chapter 5 utilizes the search tool MSGFDBPlus [176]). Detailed protocol developed to maximise protein identification from environmental samples with protein sequence database from the same environment for SEQUEST™ is found in Appendix A.3 *Protein sequencing protocol*. For MSGFDBPlus peptide mapping to full length protein sequences the False Discovery Rate was calculated using the spectra to peptide matches that resulted in reversed hits from the on-the-fly reversed database search and a filter on the MSGF value. Number of peptides and proteins detected varies between samples (Figure 2.5). Due to the large size of metagenomic dataset used and redundancy in protein sequences because of multiple sampling of the same environment in the Saanich Inlet time series most peptides map to multiple identical proteins, resulting in a greater number of proteins identified than peptides.

2.6.3 Taxonomic binning and visualization of expressed proteins

There are many ways to visualize taxonomic distributions in environmental sequence data including heat maps, histograms, bubble plots, and trees. A composite visualization method using BLAST and the least common ancestor (LCA) algorithm implemented in MEGAN [177] can be superimposed on the Interactive Tree of Life (iTOL) (`http://itol.embl.de/`) [178, 179] to visualize the taxonomic distribution of expressed proteins from the Saanich Inlet water column (Figure 2.6). Protein abundance information is mapped onto these tree structures using the normalized spectral abundance factor (NSAF) [180]. The NSAF values for a given protein can then be compared between samples for a more accurate representation of gene expression between environmental samples. A detailed protocol of NSAF calculation and taxonomic binning and visualization can be found in Appendix A.4 *Taxonomic binning and visualization of expressed proteins*.

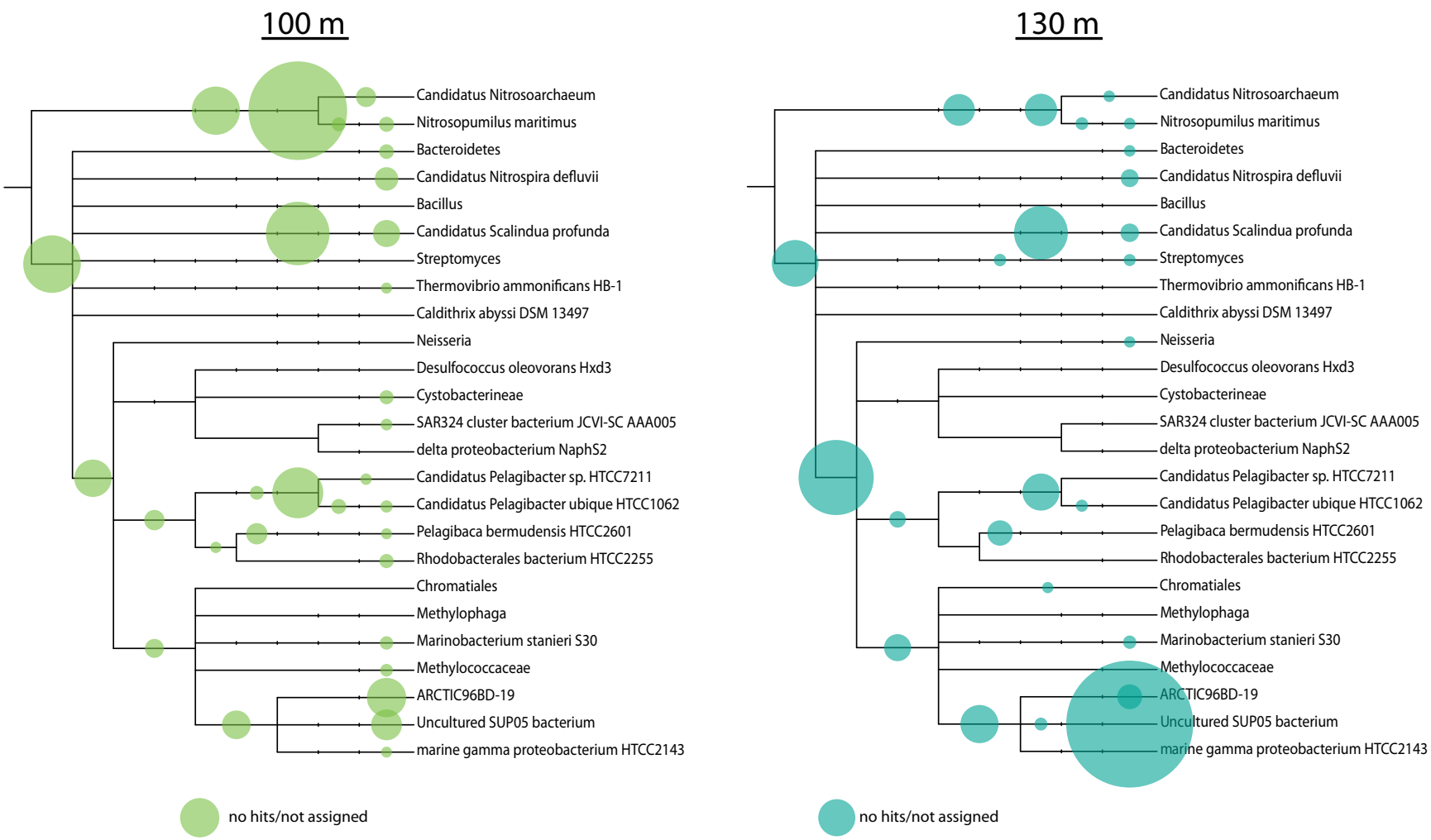


Figure 2.6: Least common ancestor distribution of detected proteins (present and preceding pages). The taxonomic distribution of detected protein sequences in Saanich Inlet Station S3 determined by MEGAN least common ancestor algorithm at 100 (green), 130 (teal), 150 (blue), and 200 (purple) m sampling intervals. Taxon abundance in the metaproteome is shown using NSAF values. The NSAF value for detected proteins not assigned to the NCBI hierarchy or with no BLAST hit/ not assigned are shown at the base of the tree for each depth interval.

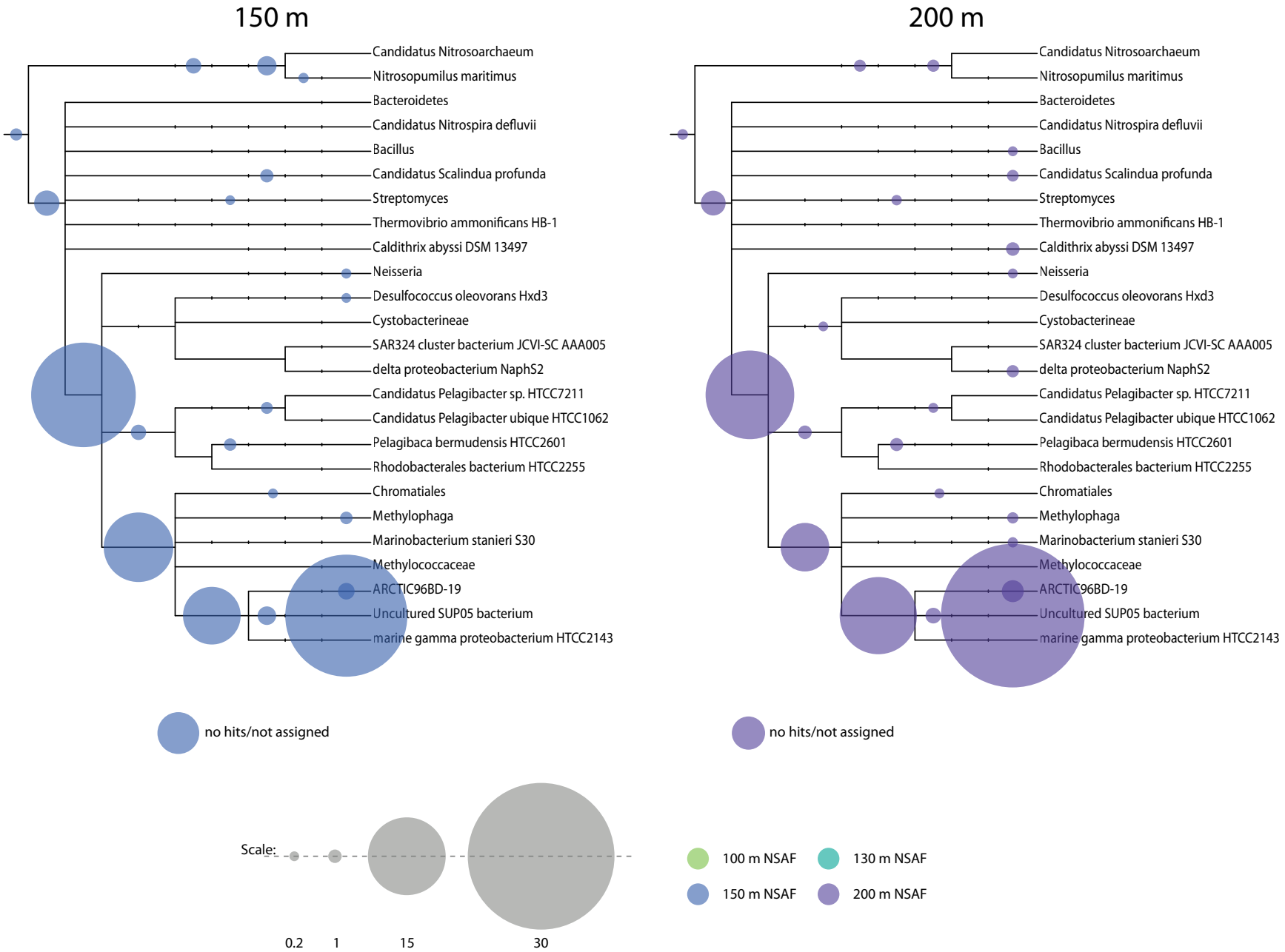


Figure 2.6 Least common ancestor distribution of detected proteins continued from previous page

2.6.4 Metaproteomic samples

A total of 68 metaproteomic samples (Table A.3) were generated covering 14 time points including two renewal periods in August/September. These metaproteomes provide a means to validate proposed metabolic pathways and energy metabolisms and co-metabolic interactions.

2.7 Single-cell amplified genomes

While multi-omic analysis is a powerful tool for exploring microbial community metabolism and expression, a concrete link between identified metabolic functions and taxonomy remains elusive with omics alone. Recent technological advances have allowed for the production of single-cell amplified genomes (SAGs) [105], generated by physically isolating individual cells from environmental samples and carrying out whole genome amplification and sequencing. Thus, genomic sequence data is generated from an individual cell, irrefutably linking the genes with the taxonomy of the cell. This is particularly useful for taxa that do not have any closely related cultured and sequenced representatives, such as microbial dark matter phyla [19]. The sorting of cells into individual wells using flow cytometery also serves to reflect the natural abundance in the environment [105], providing a way to survey microbial communities without amplification bias. Isolated cells can be screened for taxonomy by PCR amplification of the SSU rRNA and sequencing, and cells from desired taxonomy chosen for additional amplification and sequencing.

Coupling SAGs with metagenomes from the same environment by stringent alignment searches and Kmer frequency analysis [181] can phylogenetically anchor genes and whole metagenomic contigs to a given taxonomy, greatly increasing the metabolic insights into specific taxonomic groups involved in biogeochemical transformations as well as potential co-metabolic interactions. Mapping metatranscriptomic reads onto SAGs can similarly provide insight into transcriptional activity of specific taxonomic groups and individual cell populations with greater phylogenetic resolution.

In Saanich Inlet samples for SAGs were taken at a single time point, August 2012, at station S3 at three depths spanning the oxycline (100, 150 and 185 m) in efforts to capture the microbial community at particular points along the redox gradient (dysoxic, anoxic and sulfidic, respectively).

Screening of the SSU rRNA gene showed an overall decrease in diversity along the redox gradient using Silva database (Quast13). In dysoxic waters Miscellaneous Gammaproteobacteria are the most dominant at 14.8% of collected SAGs, followed by Flavobacteriales at 10.6% and several other taxa in similar abundances (Figure 2.7). In anoxic waters SUP05 Gammaproteobacterial SAGs were the dominant taxa with 54.5% of collected SAGs. In sulfidic waters Arcobacter Epsilonproteobacteria SAGs were the dominant taxa with 40.7% of detected SAGs, followed closely by SUP05 at 39.8%. After screening, SAGs were chosen for additional amplification and sequencing based primarily on efficiency of the whole genome amplification as well as desired taxonomy (Figure2.7).

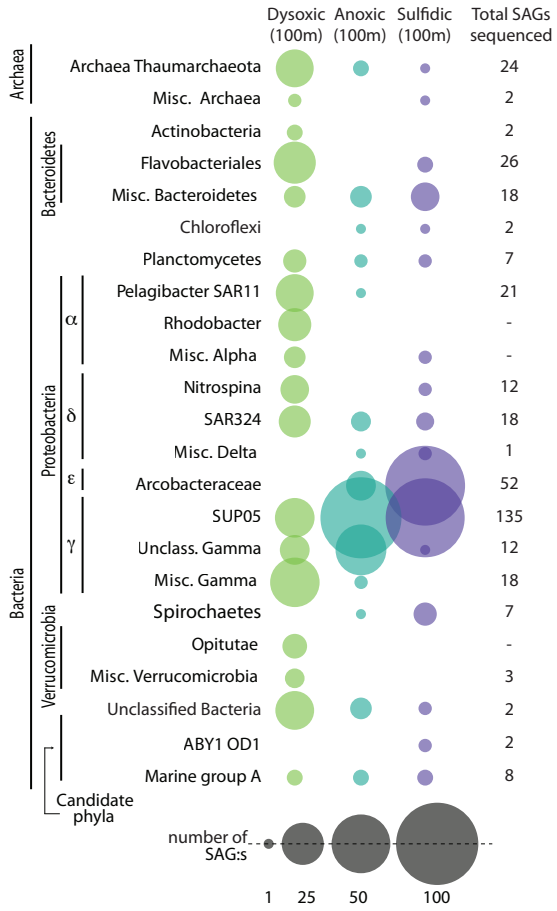


Figure 2.7: Singe-cell amplified genomes. Taxonomic distribution and number sequenced of single-cell amplified genomes (SAGs) collected from Saanich Inlet August 2012. Taxonomy assigned by screening of the small subunit rRNA gene for each SAG collected.

2.8 Annotation of meta-omics datasets by MetaPathways

Multi-omics datasets are powerful tools for investigating microbial community structure and metabolic expression and a key point to generating robust datasets is annotation. Annotation involves the identification of open reading frames (ORFs) or genes and assignment of gene function based on sequence homology searches of sequence databases. An important aspect of the multi-omics approaches applied in this thesis is the use of a consistent annotation pipeline: MetaPathways [94, 182]. Metapathways is a modular bioinformatics pipeline for multi-omic annotation developed in the Hallam lab. Input files are sequence files in the form of `.fasta` or `.fastq`, and use Programming Gene finding Algorithm (Prodigal), to identify ORFs, including incomplete or fragmented ORFs [183], thus maximizing ORF recovery for environmental sequence data types where genomic information is not exhaustively sequenced. Amino acid translated ORFs are then searched against a suite of functional databases using LAST [184] or FAST [185] algorithm and a BLAST-score ratio cut-off [186] for assignment of function from any one database. The use of multiple functional databases including KEGG [187], COG [188], RefSeq and MetaCyc [189] provides a robust functional annotation. Further, MetaPathways leverages Pathway-Tools functionality to identify metabolic pathways [189], thus providing additional insight into the metabolic potential and activity of microbial communities.

MetaPathways is also designed to scale with next generation sequencing platforms such as Illumina. Assembly of metagenomic and metatranscriptomic data into contigs results in a loss of the read-depth information, such that the number of copies of a given sequence of DNA, reflective of the number of organisms carrying that sequence, is not incorporated into the assembly information. The reads can be mapped back to the assembly using alignment algorithms [190] and accounted for using the reads per kilobase per million mapped (RPKM) (Equation 2.3). The RPKM value for an ORF reflects the number of reads mapped to an ORF while accounting for ORF length and total number of reads in a sample [190, 191] .

$$\text{RPKM} = \frac{\frac{\text{Reads Mapped to ORF}}{\text{ORF Length (bp)}}}{\frac{\text{Reads Mapped to Sample}}{10^6}} \tag{2.1}$$

Furthermore, RPKM values are additive such that for a given functional gene, *e.g.* sulfide oxidase, the RPKM for each sulfide oxidase in a metagenome (or transcriptome) can be summed to give the total relative abundance of sulfide oxidase in a given sample and compared to other samples. RPKM values can also be summed for a given taxonomy to provide the relative abundance of genes from that taxa.

Metabolic pathways of importance for microbial communities along redox gradients tend to converge around nitrogen and sulfur based reactions such as ammonia oxidation, nitrification, denitrification, anammox, and reduction and oxidation of sulfur compounds. While most of these individual reactions exist in the metaCyc and Pathway-Tools database, upstream annotation of specific proteins and known taxonomic breadth of certain pathways (i.e. anammox) are not taken into account by Pathway-Tools during identification of several nitrogen and sulfur-based pathways. For example, in a metagenome from Hawaii station ALOHA OMZ waters, the major nitrate reduction pathways (denitrification, dissimilatory nitrate reduction and intra-aerobic nitrite reduction) were identified by MetaPathways(Figure 2.8A). However, a detailed look at the enzymes present indicated that only genes involved in the denitrification pathway were present in the sample (Figure 2.8A). Further analysis of the taxonomic affiliation of the detected genes indicated the presence of genes from known nitrite and ammonia oxidizing organisms, particularly in the RNA, as well as the genes from denitrifying organisms (Figure 2.8B). Genes assigned to the nitric oxide reduction step were all annotated as regulatory proteins with no enzymatic activity to produce N_2O. These nuances can be subtle but significant to the interpretation of biogeochemical cycles occurring in OMZs. Thus, throughout the course of multi-omic analysis, annotation of functional genes was carefully scrutinised and included information about their taxonomic affiliation.

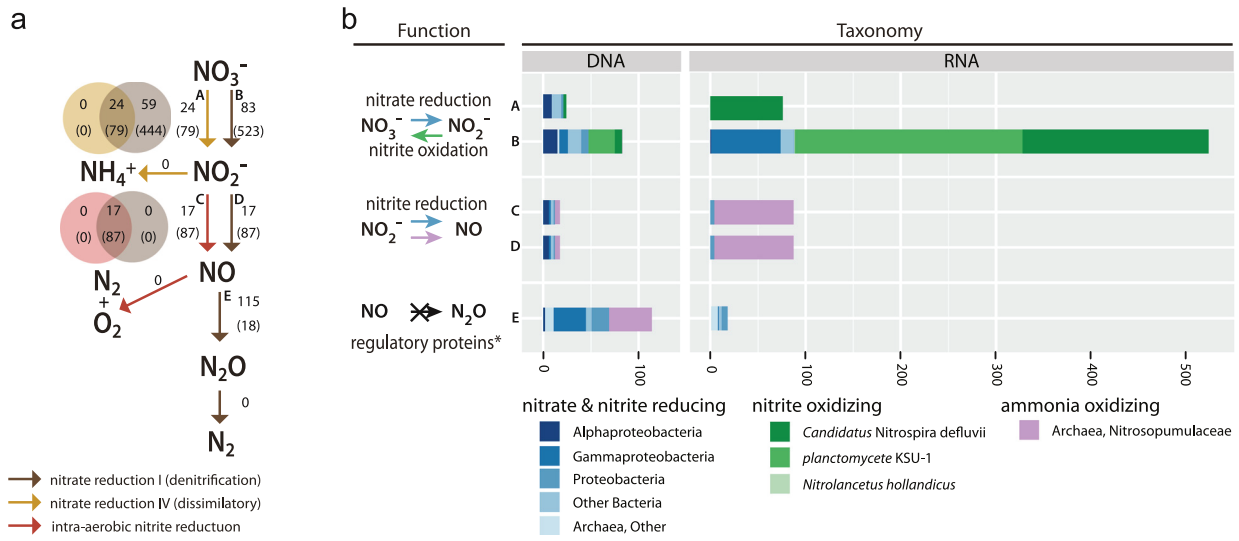


Figure 2.8: Nitrogen cycling pathways in MetaPathways. Example of Taxonomic and functional breakdown of nitrogen cycling pathways from Hawaii Station ALOHA. **(A)** Nitrogen cycling pathways and reactions assigned by PathoLogic. Arrow color indicates pathway, nitrate reduction I (denitrification) (brown), nitrate reduction IV (dissimilatory) (yellow), and intra-aerobic nitrite reduction (red). Grey numbers adjacent to arrows indicate number of reads assigned to the reaction in the DNA and RNA (RNA in parentheses). Overlapping circles indicate the distribution of reads across multiple pathways. **(B)** BLAST-based functional and taxonomic breakdown of reads assigned to reactions in given pathways as indicated by letters A-E. Function was determined by the top RefSeq BLAST hit, reported by the MetaPathways pipeline, and indicated by reaction arrows, with color corresponding to taxa or taxonomic group with known activity: taxa with nitrate and nitrite reducing activity (blue), nitrite oxidizing activity (green), and ammonia oxidizing activity (purple). Grey reactions indicate no reads for enzymatic activity were detected, only regulatory proteins that may be involved in gene expression regulation (*).

2.9 Conclusion and application

Environmental multi-omic datasets can be used to uncover biogeochemical cycling, reconstruct metabolic pathways, and identify different patterns of co-metabolic interactions along redox gradients. Within Saanich Inlet time series I focused on pathways of nitrogen, sulfur and carbon fixation along redox gradients. The utility of a time series in environmental multi-omic datasets is three-fold. Firstly, it provides a mechanism to monitor the microbial community and associated metabolic activities throughout the process of O_2 depletion and associated development of H_2S in basin waters extensible to open-ocean, anoxic and sulfidic OMZs. Secondly, it permits tracking of shifts in microbial populations and their co-metabolic interactions over time and in response to changing environmental conditions. Thirdly, it begins to address the lack of replication in

environmental omics datasets by providing pseudo-replicates, multiple samples with highly similar chemical conditions, under which to study the microbial community and associated metabolic activities. While there remain limits to using environmental multi-omics that can only be addressed through culturing and isolation, multi-omic analyses of microbial communities along redox gradients can still provide valuable knowledge and insights into biogeochemical cycling and co-metabolic interactions that will shed light on the fundamental principles shaping microbial communities and metabolisms.

Chapter 3

Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes[2]

This chapter represents the first metaproteomic analysis to chart spatial and temporal patterns of gene expression along defined redox gradients in oxygen deficient waters. Using methodologies from Chapter 2 for small subunit ribosomal RNA (SSU rRNA) tags, metagenomics and metaproteomics, I establish microbial community structure, metabolic capacity and protein expression associated with key nitrogen, sulfur and carbon biogeochemical cycles outlined in Chapter 1.2. The expression of metabolic pathway components for nitrification, anaerobic ammonium oxidation (anammox), denitrification, and inorganic carbon fixation were differentially expressed across the redoxcline and co-varied with distribution patterns of ubiquitous OMZ microbes. The numerical abundance of SUP05 proteins mediating inorganic carbon fixation under anoxic conditions suggests that SUP05 will become increasingly important in global ocean carbon and nutrient cycling as OMZs expand. The exploration of multiple stations and time points reinforces the reproducibility of the metaproteome under similar redox conditions with respect to relative abundance of energy cycling proteins. This work is a basis for interpreting microbial community function and expression and serves as a framework for co-metabolic interactions along redox

[2]A version of this chapter has been published in Proceedings of the National Academy of Sciences as *Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes* in 2014 by Hawley, A K., Brewer, H.M., Norbeck, A. D., PašaTolic, L. and Hallam, S. J..

gradients in OMZs.

3.1 Introduction

Marine oxygen (O_2) minimum zones (OMZs) are widespread and naturally occurring water column features that arise when respiratory O_2 demand during decomposition of organic matter exceeds O_2 availability in stratified waters. Operationally defined by dissolved O_2 concentrations $<$20 µM, OMZs promote the use of alternative terminal electron acceptors (TEAs) in microbial energy metabolism that results in climate active gas production including carbon dioxide (CO_2), nitrous oxide (N_2O) and methane (CH_4) [6]. Currently OMZs constitute $\sim$7% of global ocean volume [2, 10]. However, global climate change promotes conditions for OMZ expansion and intensification e.g., reduced O_2 solubility and increased stratification, with resulting feedback on the climate system, the extent to whcih have yet to be determined [10, 129].

Within OMZs, the use of nitrate (NO_3^-) and nitrite (NO_2^-) as TEAs in dissimilatory nitrate reduction (denitrification) and anaerobic ammonium oxidation (anammox) results in fixed nitrogen loss in the form of N_2O and dinitrogen gas (N_2) respectively [60, 192]. Because OMZs account for up to 50% of oceanic N_2 production, they have the potential to limit primary production in overlying surface waters [192, 193]. A recent model suggests that nitrogen fixation in proximity to OMZ waters can balance nitrogen loss processes [194], and several studies along redoxclines in the Eastern Tropical South Pacific and Baltic Sea have measured nitrogen fixation rates that support a close spatial coupling between nitrogen loss and nitrogen fixation consistent with this model [45], [44], [41]. Moreover, recent studies have begun to link the oxidation of reduced sulfur-compounds including thiosulfate ($S_2O_3^{2-}$) and hydrogen sulfide (H_2S) to nitrogen transformations in non-sulfidic OMZs providing evidence for a cryptic sulfur cycle with the potential to drive inorganic carbon fixation processes [78]. Indeed, many of the key microbial players implicated in nitrogen and sulfur transformations in OMZs, including Thaumarchaeota, Nitrospina, Nitrospira, Planctomycetes and SUP05/ARCTIC96BD-19 Gammaproteobacteria have the metabolic potential for inorganic carbon fixation [33, 52, 55, 66, 77, 80] and previous process rate measurements in OMZs point to high rates of dark primary production [81–84]. However, the relative contribution

of each player to coupled carbon (C), nitrogen (N) and sulfur (S) biogeochemistry as a function of redox zonation and in response to perturbation remains to be determined. These conributions have important implications for understanding the long-term ecological and biogeochemical impacts of OMZ expansion and intensification on carbon and nutrient cycling in the global ocean.

Here I investigate changes in microbial community structure and function in a seasonally stratified fjord, Saanich Inlet on Vancouver Island British Columbia Canada, to better understand metabolic coupling along defined redox gradients. I combine cultivation-independent molecular approaches including small subunit ribosomal RNA gene pyrosequencing, metagenomics and metaproteomics to chart the progression of microbial community structure and gene expression along the redoxcline. I then construct a conceptual model linking different modes of inorganic carbon fixation with distributed nitrogen and sulfur-based energy metabolism.

3.2 Results and Discussion

3.2.1 Water column chemistry and molecular sampling

To evaluate changes in water column redox gradients associated with different stages of stratification and renewal, samples were collected from the Saanich Inlet water column from station S3 on April 9, 2008 (Apr08) and from multiple stations along the transect from the mouth of the inlet (S4) through the midpoint (S3) and at the back (S2) on September 1, 2009 (Sep09) (Figure 3.1A and B) corresponding to metaproteomic datasets in CHapter 2. Water column chemistry profiles indicated four redox zones: upper oxycline (UO), lower oxycline (LO), sulfide nitrate transition zone (SNTZ), and sulfidic zone (SZ) generally corresponding to dysoxic (20-90 μM O_2), suboxic (1-20 μM O_2), anoxic ($<$1 μM O_2) and anoxic sulfidic water column conditions (Figure 3.2). Water column redox zonation and associated microbial community structure was consistent with other OMZs [2, 8] making Saanich Inlet a tractable model ecosystem for studying microbial community responses to changing levels of water column oxygen-deficiency.

To explore changes in microbial community structure and function along water column redox gradients, I analyzed paired metagenomic and metaproteomic datasets from Apr08 and paired small subunit ribosomal RNA gene pyrosequencing and metaproteomics datasets from

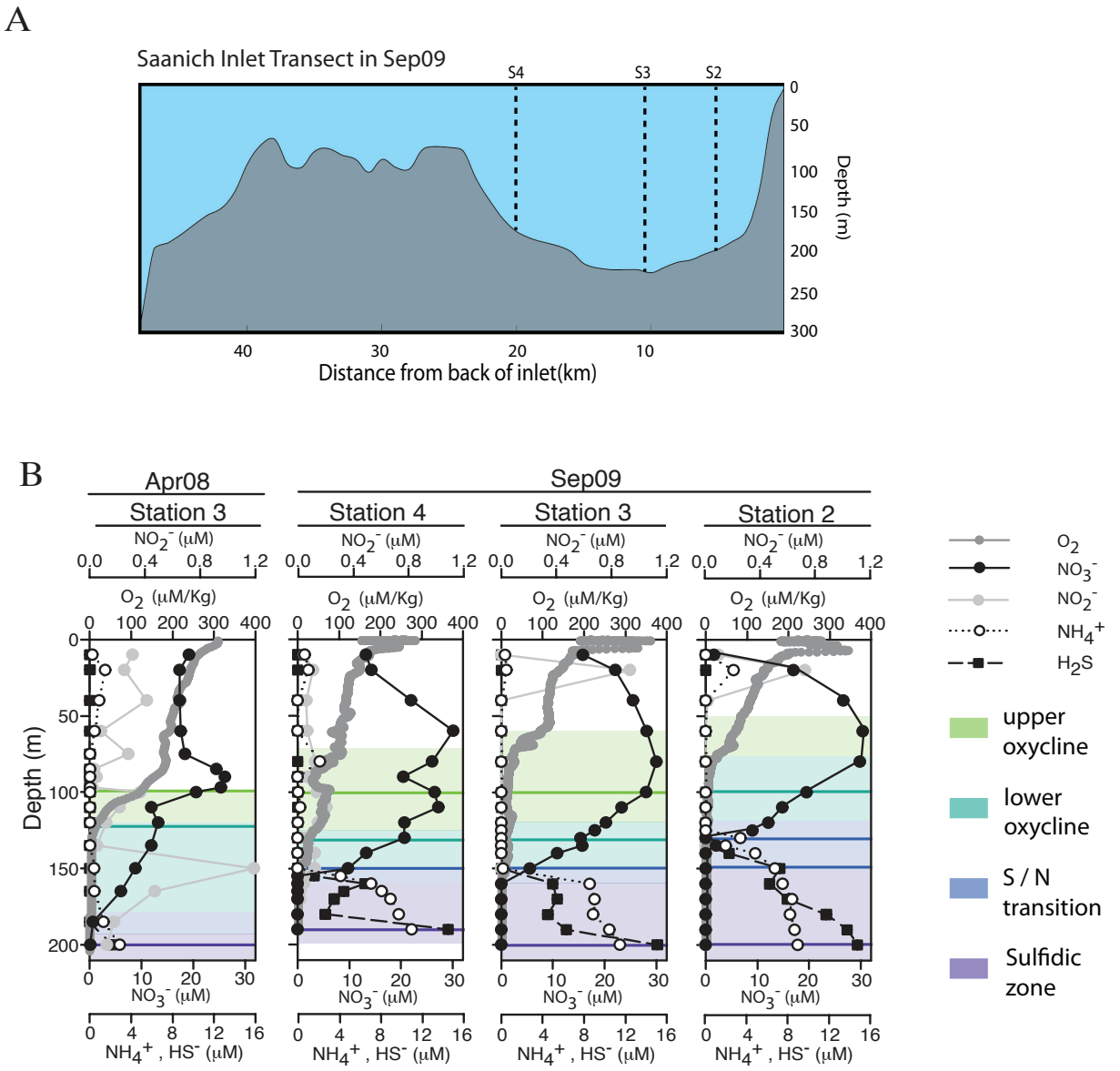


Figure 3.1: Saanich Inlet bathymetry and chemistry. **(A)** Cross section of Saanich Inlet showing sampling station locations. **(B)** Chemical profile of the water column in Saanich Inlet at indicated stations and sampling times, showing oxygen (O_2), nitrate (NO_3^-), nitrite (NO_2^-), ammonia (NH_4^+), and hydrogen sulfide (H_2S) concentrations. Colors indicate region of water column including upper oxycline (green), lower oxycline (teal), S/N transition zone (blue), and sulfidic zone (purple), and colored bars indicate sample depths for SSU rRNA gene pyrotags, metagenomics, and metaproteomics. [Reprinted with permission from Zaikova et al. 2010 (Copyright 2009, Wiley & Sons).]

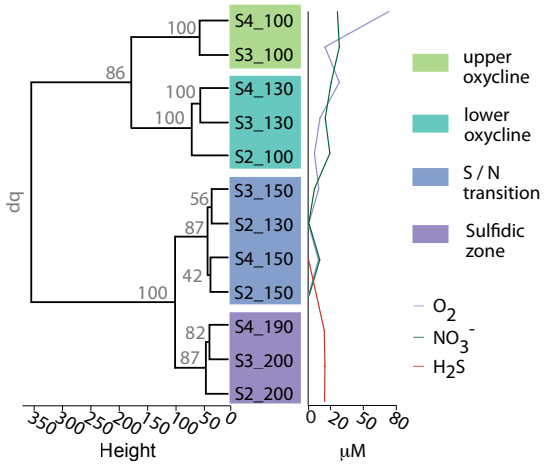


Figure 3.2: Sample hierarchical clustering. Hierarchical clustering of metaproteome by NSAF (see Methods) for detected proteins from Sep09 S2, S3, and S4 indicating compartments of the water column, with adjacent sparklines for oxygen (O_2), nitrate (NO_3^-), and hydrogen sulfide (H_2S) for each sample.

Sep09 (Figure 3.1B). Sanger end sequencing of small insert clone libraries from the three Apr08 samples yielded a total of 54,701 ORFs, with an average of 18,234 ORFs per sample. Small subunit ribosomal RNA (SSU rRNA) gene pyrosequencing of the 12 Sep09 samples yielded 87,138 sequences that clustered into 3,385 non-singleton operational taxonomic units at the 97% identity threshold. Tandem MS-coupled LC (LC/MS/MS) metaproteomic sequencing identified a total of 5,019 unique proteins (Table B.1), a number comparable to previous marine metaproteomic studies [195]. A consistent number of proteins were identified across the Sep09 samples, averaging 695 unique proteins per sample (Table B.2). Although variability in protein detection in the Apr08 samples was considerable, the high number of unique proteins detected in the Apr08 200 m sample (4,344) enabled identification of more complete metabolic pathways.

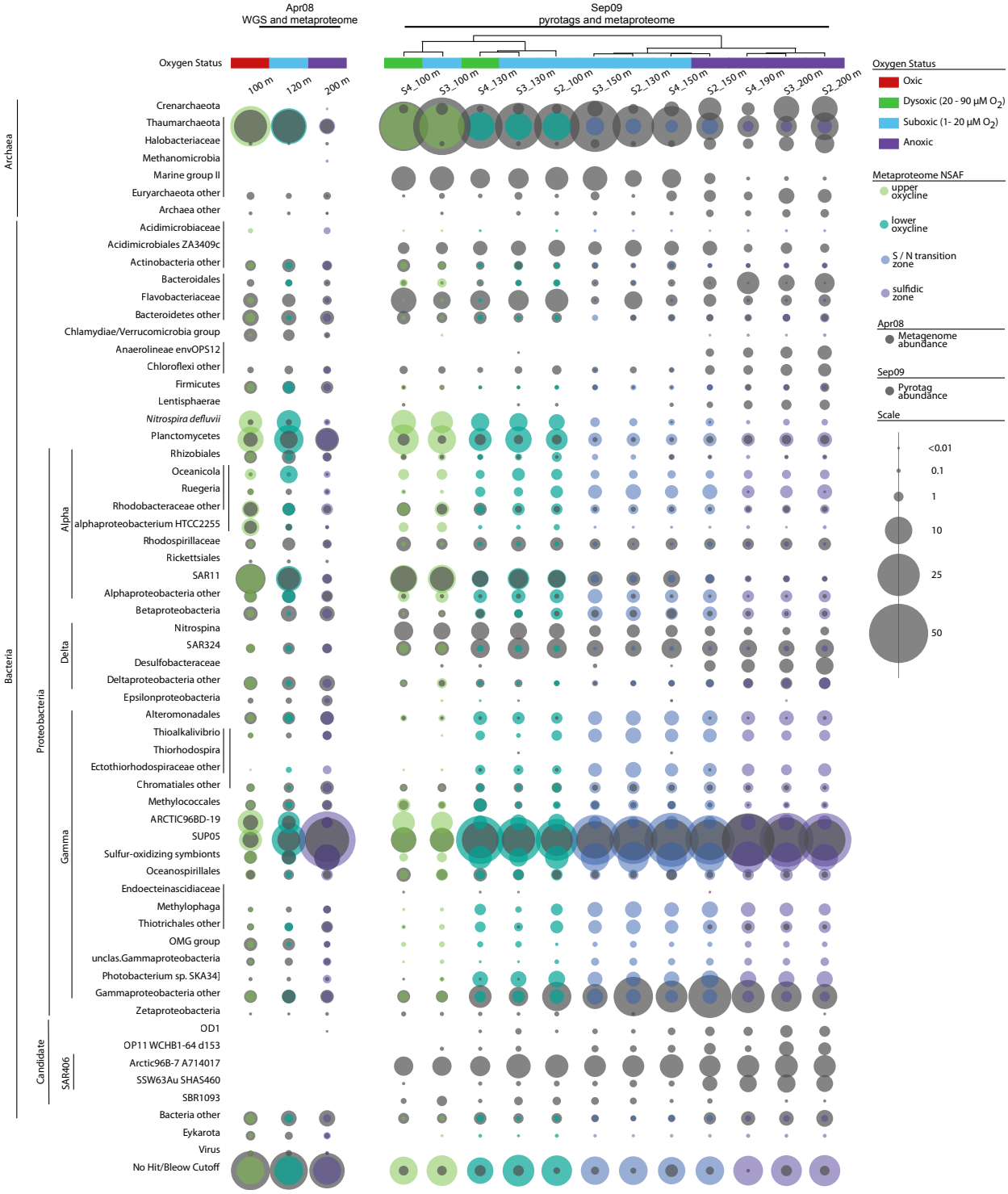

Apr08
WGS and metaproteome
Sep09
pyrotags and metaproteome
Oxygen Status
100 m
120 m
200 m
S4_100 m
S3_100 m
S4_130 m
S3_130 m
S2_100 m
S3_150 m
S2_130 m
S4_150 m
S2_150 m
S4_190 m
S3_200 m
S2_200 m
Archaea
Bacteria
Proteobacteria
Candidate
Alpha
Delta
Gamma
SAR406
Crenarchaeota
Thaumarchaeota
Halobacteriaceae
Methanomicrobia
Marine group II
Euryarchaeota other
Archaea other
Acidimicrobiaceae
Acidimicrobiales ZA3409c
Actinobacteria other
Bacteroidales
Flavobacteriaceae
Bacteroidetes other
Chlamydiae/Verrucomicrobia group
Anaerolineae envOPS12
Chloroflexi other
Firmicutes
Lentisphaerae
Nitrospira defluvii
Planctomycetes
Rhizobiales
Oceanicola
Ruegeria
Rhodobacteraceae other
alphaproteobacterium HTCC2255
Rhodospirillaceae
Rickettsiales
SAR11
Alphaproteobacteria other
Betaproteobacteria
Nitrospina
SAR324
Desulfobacteraceae
Deltaproteobacteria other
Epsilonproteobacteria
Alteromonadales
Thioalkalivibrio
Thiorhodospira
Ectothiorhodospiraceae other
Chromatiales other
Methylococcales
ARCTIC96BD-19
SUP05
Sulfur-oxidizing symbionts
Oceanospirillales
Endoecteinascidiaceae
Methylophaga
Thiotrichales other
OMG group
unclas.Gammaproteobacteria
Photobacterium sp. SKA34]
Gammaproteobacteria other
Zetaproteobacteria
OD1
OP11 WCHB1-64 d153
Arctic96B-7 A714017
SSW63Au SHAS460
SBR1093
Bacteria other
Eykarota
Virus
No Hit/Bleow Cutoff
Oxygen Status
Oxic
Dysoxic (20 - 90 μM O2)
Suboxic (1- 20 μM O2)
Anoxic
Metaproteome NSAF
upper oxycline
lower oxycline
S / N transition zone
sulfidic zone
Apr08
Metagenome abundance
Sep09
Pyrotag abundance
Scale
<0.01
0.1
1
10
25
50

Figure 3.3: Taxonomic distribution of metagenome, metaproteome and pyrotags (previous page). Taxonomic distribution and relative abundance of metagenome for Apr08 (gray), SSU rRNA pyrotag for Sep09 (gray), and metaproteome in upper oxycline (green), lower oxycline (teal), S/N transition zone (blue), and sulfidic zone (purple). Relative abundance for taxonomic groups is shown for selected groups including any representing >2% SSU rRNA gene pyrotag, metagenome, or metaproteome datasets. For metagenome percentages were determined as the sum of all ORFs in unassembled metagenomic reads hit to a given taxa, normalized to the total number of ORFs over 30 residues long. Percentage of metaproteome determined as the sum of all NSAF for all detected proteins with hit to a given taxa. Hierarchical clustering of detected protein abundance shown above with color indicating oxygen status of the water column at time of sampling.

3.2.2 Patterns of redox-driven niche partitioning

To determine patterns of redox-driven niche partitioning along redox gradients in metagenomic and metaproteomic datasets, I compared community composition with ORF counts and protein normalized spectral abundance (NSAF, see methods and Chapter 2) between UO, LO, SNTZ and SZ (Figure 3.3). Hierarchical clustering of NSAF values was consistent with redox zonation (Figure 3.2). Clear trends in protein abundance were observed in relation to redox zonation not reflected in pyrotag and metagenomic datasets, consistent with alternative forms of coupling or regulated gene expression. Ammonia oxidizing Thaumarchaeota, mediating the first step of nitrification, dominated UO and LO samples and decreased in abundance within the SNTZ and SZ. Similar trends were observed with respect to ORF counts and NSAF values (Tables B.2B.3).The nitrite oxidizing bacterium *Nitrospina gracilis* [28], mediating the second step of nitrification, was abundant in UO and LO samples and decreased in abundance within the SNTZ and SZ. A second nitrite oxidizing bacterium *Nitrospira defluvii* [55], although absent from the pyrotag datasets, exhibited high NSAF values with a similar distribution pattern as Thaumarchaeota and *N. gracilis*. Anammox bacteria affiliated with the Planctomycetes (Tables B.2 -B.4) exhibited intermediate abundance ($\sim$1%) in the UO and LO samples, decreasing in abundance within the SNTZ before increasing again in the SZ. Planctomycete ORF abundance increased along the redoxcline while protein NSAF values were high in the UO and LO, decreasing to intermediate values within the SNTZ and SZ. These patterns of protein expression confirm previous reports of coupled nitrification and anammox observed in OMZs based on process rate and functional marker gene abundance [56, 62].

In addition to known players in the nitrogen cycle summarised in section 1.2.1, taxa involved in sulfur cycling or coupled nitrogen and sulfur cycling were also abundant and active in the water column. Multiple lineages affiliated with SAR11 within the Alphaproteobacteria, mediating dimethylsulfoniopropionate (DMSP) oxidation [23], were abundant in the UO and LO samples, decreasing in abundance within the SNTZ and SZ (Figure 3.3). Similar trends were observed with respect to ORF counts and NSAF values (Tables B.2 -B.4). Multiple lineages affiliated with SUP05/ARCTIC96BD-19 and symbiont-related Gammaproteobacteria (Tables B.2 -B.4), mediating oxidation of reduced sulfur compounds using O_2 [77] or NO_3^- [33] as TEAs, were also abundant. The ORFs for ARCTIC96BD-19, SUP05 and symbionts exhibited reciprocal distribution patterns, with ARCTIC96BD-19 ORFs decreasing and SUP05 and symbiont ORFs increasing in abundance within the SNTZ and SZ. A similar pattern was observed with respect to NSAF values, with high SUP05 NSAF values in the LO, SNTZ and SZ. These distribution patterns support previous reports of ARCTIC96BD-19 and SUP05 population structure [2, 33, 77, 149, 196].

Collectively, Thaumarchaeota, Nitrospina, Nitrospira, Planctomycetes, SAR11 and SUP05/ARCTIC96BD-19 and symbiont-related Gammaproteobacteria comprised on average 48% of pyrotag, 41% of metagenomic, and 64% of metaproteomic datasets (Tables B.2, B.3). Several taxonomic groups that were abundant based on pyrotags ($\geq$1%) including Marine Group II Euryarchaea, Crenarchaeota, Acidomicrobiales, Bacteroidetes, Chloroflexi, Flavobacteria, Desulfobacteraceae and Candidate divisions OD1, OP11, Marine Group A, SBR1093 were not well represented in metagenomic or metaproteomic datasets (Figure 3.3). Lack of indigenous reference genomes likely caused many sequences originating from these groups to be classified as no hit or below cut-off (see methods). Consistent with this observation, BLAST queries against the Genomic Encyclopedia of Bacteria and Archaea Microbial Dark Matter (GEBA-MDM) single-cell genome collection [19] yielded only 23 additional protein sequences which had otherwise been classified as below cutoff or no hit. Conversely, several taxonomic groups including *N. defluvii* and ARCTIC96BD-19 that were absent in pyrotag datasets exhibited intermediate ORF counts and NSAF values. This discrepancy was likely due to incomplete taxonomic resolution for these groups within the Greengenes database. Approximately 1% of pyrotag and 10% of metagenomic and metaproteomic datasets remained unaffiliated with any taxonomic group. Taken together, these results indicate

that active nitrogen and sulfur cycling microorganisms are the primary contributors to both genetic potential and gene expression along the redoxcline in Saanich Inlet.

3.2.3 Differential gene expression patterns

To investigate patterns of gene expression driving carbon and energy metabolism along the redoxcline in Saanich Inlet, I identified nitrification, anammox, denitrification, sulfur oxidation and inorganic carbon fixation pathway components in metagenomic and metaproteomic datasets using BLAST. By summing the NSAF values for each component I observed differential patterns of gene expression and metabolic coupling along the redoxcline (Figure 3.4). Expression of these pathways was remarkably stable under similar redox conditions in space (Sep09 S2-S4) and time (Apr08 to Sep09) (Figures B.1 - B.3).

Expressed pathways for nitrogen-based energy metabolism progressed from ammonia oxidation and nitrification in the UO and LO to denitrification in the SNTZ and SZ (Figure 3.4, B.1). Proteins catalyzing the first step of nitrification, ammonia monooxygenase subunits B and C (Amo), from Thaumarchaeota, were detected in the UO and LO and decreased along the redoxcline. Proteins catalyzing the second step of nitrification, nitrite oxidase (NXR), from *N. graclilis* and *N. defluvii* followed the same pattern of expression as Amo. Moreover, the detection of both Amo and NXR from nitrifying taxa, albeit at lower NSAF values in the SNTZ and SZ, supports recent observations of NO_2^- oxidation in the Namibian OMZ with implications for NO_3^- supply for reduction via denitrification [56, 197]. Nitrite oxidase from Planctomycetes (NXR) (Figure 3.4) [66] had the highest NSAF values of any protein in the UO and LO and exhibited a similar expression profile to Amo and NXR originating from *N. gracilis* and *N. defluvii* (Figure 3.4, B.1). Conversely, proteins catalyzing anammox, including hydrazine and hydroxylamine oxidoreductases (Anx) from Planctomycetes (Figure 3.4, B.1) exhibited opposing expression patterns, with low NSAF values in the UO and LO that increased in the SNTZ and SZ. Contrasting patterns of NXR and Anx expression from Planctomycetes could reflect a metabolic response to O_2 resulting in a shift between maintenance energy production in the UO and LO to anammox for growth under more favorable redox conditions in the SNTZ and SZ. Alternatively, close sequence similarity between Planctomycetes, *N. gracilis* and *N. defluvii* NXR could confound BLAST-based taxonomic

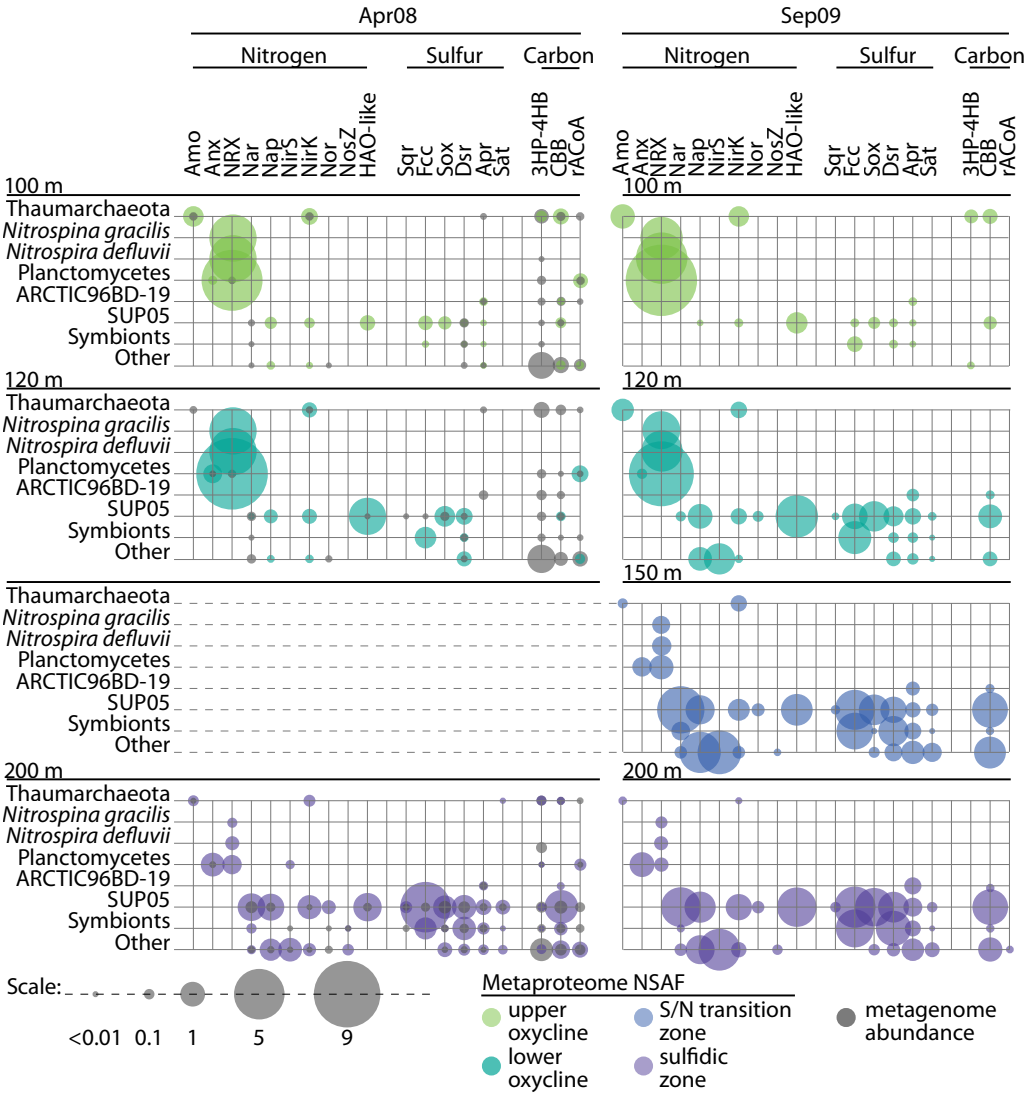


Figure 3.4: Nitrogen, sulfur and carbon cycling proteins. Distribution and NSAF value of proteins involved in nitrogen and sulfur-based energy metabolism and inorganic carbon fixation for taxa abundant in the metaproteome. For metagenome (gray, Apr08 only) and metaproteome in upper oxycline (green), lower oxycline (teal), S/N transition zone (blue), and sulfidic zone (purple). See Table C.3. for full list of protein names; Anx indicates anammox hydroxylamine oxidoreductase and hydrazine oxidoreductase proteins

assignment.

Proteins mediating the partial denitrification pathway from SUP05 including dissimilatory nitrate reductase subunits G and H (Nar), periplasmic nitrate reductase subunits A and B (Nap), and nitrite reductase (NirK) were detected in the UO and increased in abundance along the redoxcline (Figure 3.4, B.1). Protein NSAF values for SUP05 Nar increased relative to Planctomycetes NXR in the SNTZ and SZ. Additional proteins for SUP05 nitric oxide reductase subunits B and C (Nor) were detected with similar NSAF values in the LO, SNTZ and SZ. Although denitrification pathway components from other taxonomic groups were detected in the water column, SUP05 was the only group to express consecutive proteins in the denitrification pathway, making up 50% of the total NSAF value of all denitrification proteins. While SUP05 contributed 50% of the total denitrification proteins in the Saanich Inlet water column, the remaining 50% were distributed between additional taxa. In addition to SUP05, Nap associated with *Magnetococcus marinus* and *Sulfuricella denitrificans* from Alpha and Betaproteobacteria respectively was observed with high NSAF value within the LO, SNTZ and SZ. The NirS protein from other taxa including *Colwellia psychrerythraea* and *Sillicibacter lacuscaerulensis* with smaller contribution from Ruegeria sp. TW15 was also observed with high NSAF value in the LO, SNTZ and SZ. These observations point to SUP05 as the dominant player in nitrogen-based energy metabolism in the SNTZ and SZ. The detection of SUP05 Nap and NirK in the UO and LO where O_2 concentrations approached 120 µM was unexpected given that 20 µM O_2 is a commonly accepted threshold for denitrification [12] and may have implications for the O_2 threshold for nitrogen loss processes in other OMZs. Additionally, the detection of SUP05 Nor and the absence of nitrous oxide reductase (NosZ) in the LO, and low abundance in SNTZ and SZ suggest SUP05 as a source of N_2O. Recent observations of enrichment of *nor* and *nosZ* genes on particles within OMZs suggest a distributed denitrification pathway across particle and non-particle niches [16] and may account for low NSAF values observed for NorCB and NosZ.

Expressed pathways for sulfur-based energy metabolism were detected in the UO and increased in NSAF value along the redoxcline (Figure 3.4, B.2). Proteins catalyzing sulfide oxidation predominantly originated from SUP05/ARCTIC96BD-19 and symbiont-related Gammaproteobacteria. With the exception of ARCTIC96BD-19 adenylylsulfate reductase (Apr) the vast majority

of proteins originated from SUP05 and symbionts. With respect to SUP05, flavocytochrome C (Fcc), sulfide oxidation proteins (Sox), dissimilatory sulfate reductase (Dsr) and adenylylsulfate reductase (Apr) were detected in the UO and increased in NSAF value along the redoxcline. In addition, SUP05 ATP sulfurylase (Sat) and sulfide:quinone oxidoreductase (Sqr) were detected in the LO, SNTZ and SZ and SNTZ and SZ, respectively. These results are consistent with recent SUP05 protein expression profiles observed in hydrothermal plume and overlying waters [195]. With the exception of Sox, symbiont proteins catalyzing sulfide oxidation followed the same expression pattern as SUP05. The expression of sulfur oxidation pathway components from SUP05/ARCTIC96BD-19 in the UO and LO is consistent with a cryptic sulfur cycle. However, no proteins from defined sulfate (SO_4^{2-}) reducing bacteria were identified in the metaproteome [78]. This observation could reflect a bias against particle-associated microorganisms capable of SO_4^{2-} reduction during sample processing or the use of alternative electron donors including DMSP, elemental sulfur, thiosulfate or polysulfide in the UO and LO. Additionally, proteins with BLAST hits to the hydrogenase subunit HupL originating from Guaymas Basin SUP05 metagenomes [122] were detected in the SNTZ with NSAF values comparable to SUP05 NapA (Figure B.2), expanding the range of potential substrates for SUP05 energy metabolism in the Saanich Inlet water column. Expressed proteins for three inorganic carbon fixation pathways including the 3-hydroxypropionate/4-hydroxybutyrate (3HP-4HB) from Thaumarchaeota, reductive acetyl-CoA (rACoA) from Planctomycetes and Calvin Benson Basham (CBB) cycle from SUP05 cycles were differentially expressed along the redoxcline (Figure 3.4, B.3). Unlike proteins mediating nitrogen and sulfur-based energy metabolism, ORFs encoding carbon fixation pathway components were found in higher relative abundance in the metagenome (Figure 3.4).

Proteins catalyzing the 3HP-4HB pathway in Thaumarchaeota were detected predominantly in the UO including 4-hydroxybutytyl-CoA dehydratase, acetyl-CoA carboxylase and propionyl CoA carboxylase [52, 80, 198]. Similar expression patterns were observed for Amo and other ammonia oxidation pathway components, providing evidence of inorganic carbon fixation coupled to ammonia oxidation by Thaumarchaeota in the UO. Consistent with previous reports, proteins catalyzing a putative Planctomycete rACoA pathway were detected in the SZ in Apr08 along with Anx proteins providing evidence for inorganic carbon fixation coupled to anammox under sulfidic

conditions (2.1 μM) [199]. Protein NSAF values for SUP05 CBB pathway components increased relative to other bacteria in the SNTZ and SZ providing compelling evidence for inorganic carbon fixation coupled to sulfide-oxidation and partial denitrification by SUP05. Indeed, CBB pathway components had the highest ORF counts and protein NSAF values of all carbon fixation pathways, composing 47% of all carbon fixation proteins within the SNTZ and SZ. In addition to inorganic carbon fixation pathways, the abundance of SAR11 DNA and protein in the UO and LO (Figure 3.3, Tables B.2, B.3) suggest that heterotrophic remineralization of dissolved organic matter (DOM) is an active process in the UO and LO. Specifically, ABC transporter proteins for uptake of glycine betaine, spermidine/putrescine and taurine, (sources of carbon, nitrogen and sulfur, respectively) were detected with moderate NSAF values within the UO and LO. In addition to consuming molecular oxygen, remineralization of DOM by SAR11 and other heterotrophic groups in the UO and LO could act as a source of metabolic substrates including NH_4^+ SO_4^{2-} and CO_2.

3.2.4 Regulated gene expression

Given the numerical abundance of SUP05 in the Saanich Inlet water column I were able to resolve changes in protein expression originating from a metabolic island integrating nitrogen and sulfur-based energy metabolism with inorganic carbon fixation [33] (Figure 3.5). Specifically, NSAF values for SUP05 Sqr, NarH and NarG subunits appeared to vary as a function of O_2 concentration while FccAB, NapAB subunits remained relatively constant in the LO, SNTZ and SZ (Figures 3.4, B.1 - B.3). Close proximity and similar expression profiles for *napAB* and *fccAB* is consistent with regulated gene expression along the redoxcline. Indeed, two ORFs encoding Crp/Fnr transcriptional regulators implicated in redox sensing [200] are located on either side of the *nap/fcc* gene cluster with the potential to modulate gene expression and Crp/Fnr proteins (SUP05_0428) were detected in the Apr08 SZ (Figure 3.5).

Protein NSAF values for CbbM, a ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) subunit located in proximity to the *nar* gene cluster increased between the UO and LO and remained relatively constant in the SNTZ and SZ. These results provide functional evidence in support of previous genomic observations positing a highly integrated and redox-sensitive energy metabolism in SUP05 with direct implications for energy supply to inorganic carbon

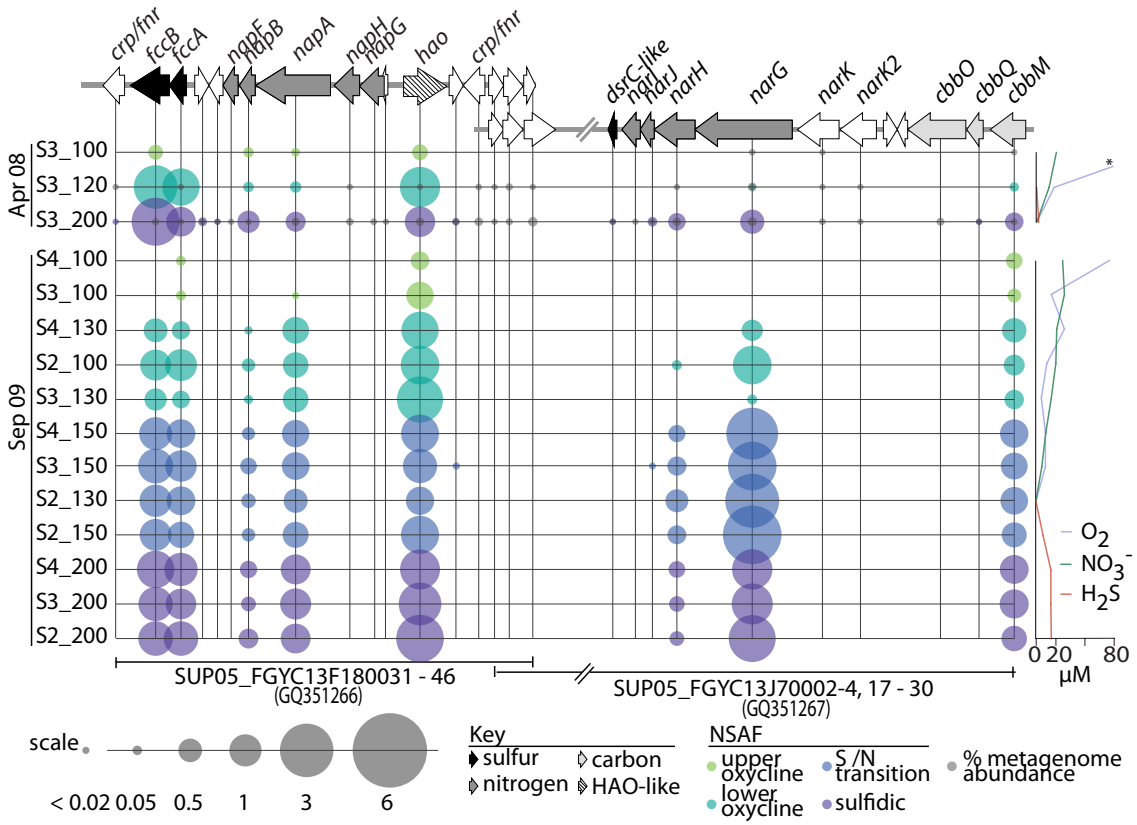


Figure 3.5: SUP05 gene expression regulation. Relative abundance of SUP05 genes and proteins in two overlapping SUP05 fosmid sequences (GQ351266 and GQ351267) [33]. Metagenome (gray, Apr08 only) and metaproteome for the upper oxycline (green), lower oxycline (teal), S/N transition zone (blue), and sulfidic zone (purple). Selected SUP05 genes involved in denitrification (dark gray shading), sulfide oxidation (black shading), and putative hydroxylamine oxidoreductase (diagonal lines) are indicated. Protein abundance shown as summed NSAF values for all detected ORFs with top hit to a given SUP05 protein. Metagenome abundances shown as percentage of ORFs with top hit to a given SUP05 gene with sparklines for oxygen (O_2), nitrate (NO_3), and hydrogen sulfide (H_2S) for each sample.

fixation. In addition to coordinated Fcc, Sqr, Nar, Nap and CbbM expression, an ORF encoding a hydroxylamine-oxidoreductase homolog (HAO-like) located in the *nap/fcc* gene cluster was among the most abundant SUP05 proteins detected in the SZ (Figure 3.4, 3.5, B.1). SUP05 *hao* is closely related to genes found in the sulfur oxidizing endosymbionts as well as the anammox bacterium *Ca.* Kuenenia stuttgartiensis. All four HAO homologs contain eight CxxCH multi-heme motifs similar to those found in NrfA, a nitrite reductase mediating dissimilatory nitrate reduction to ammonia (DNRA) [201].

3.2.5 Metabolic coupling model

Although bulk inorganic carbon fixation rates within OMZs have been measured [81–84], few studies have directly linked inorganic carbon fixation with energy metabolism of defined OMZ microbes [154]. With this linkage in mind, I construct a metabolic model describing taxonomic and metabolic networks coupling pathways of nitrogen and sulfur-based energy metabolism and inorganic carbon fixation along the redoxcline in Saanich Inlet based on metaproteomic datasets (Figure 3.6). In this model, heterotrophic remineralization of DOM releases CO_2, SO_4^{2-} and NH_4^+ within the UO and LO. Thaumarchaeota couple oxidation of $NH4^+$ to inorganic carbon fixation via the 3HP-4HB pathway within the UO producing NO_2^-, a process that has been demonstrated both in culture [50] and in situ [202]. Nitrous oxide is also produced as a by-product of ammonia oxidation [202, 203]. Nitrite produced via NH_4^+ oxidation is oxidized in turn by *N. defluvii*, *N. gracilis* and Planctomycetes in the UO [56, 66]. The extent to which this process is coupled to inorganic carbon fixation within these groups remains to be determined. Ammonia oxidation attenuates as O_2 levels decline in the LO, SNTZ and SZ accompanied by a transition to partial denitrification and anammox. In the LO, SUP05 begins to couple oxidation of reduced sulfur compounds, and possibly hydrogen, with NO_3^- reduction to N_2O to fix inorganic carbon via the CBB pathway, a trend that increases in the SNTZ and SZ [33, 75, 122, 134]. In parallel, Planctomycetes couple anammox to fix inorganic carbon via the rACoA pathway [134], although the broader occurrence of this process in the water column remains to be determined. Competition between SUP05 and Planctomycetes for oxidants could help explain variations in spatial and temporal dynamics of nitrogen loss processes observed in different OMZs. Alternatively, potential

DNRA by SUP05 could supply NH_4^+ for anammox resulting in a cometabolic linkage. Overall, the interactions described in the model are dynamic and reflect patterns of redox-driven niche partitioning regulating nitrogen loss processes and carbon flux through ubiquitous OMZ microbes.

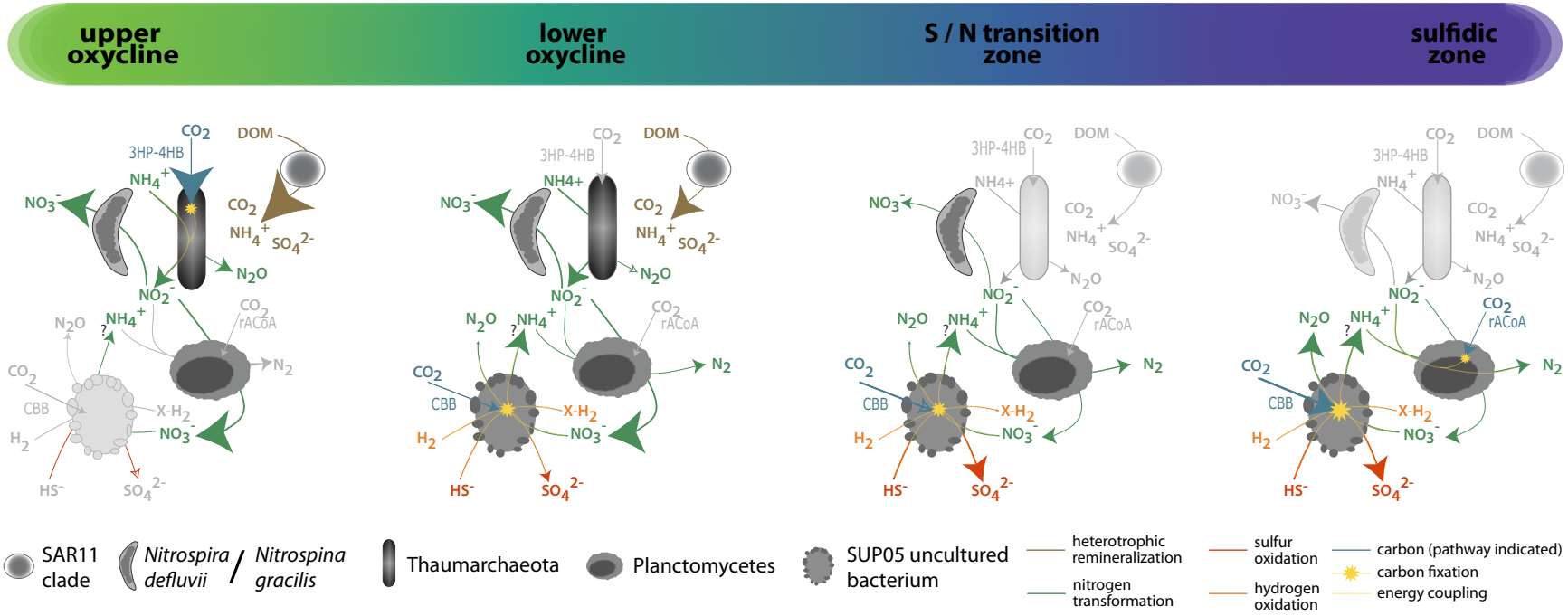


Figure 3.6: metabolic model. Proposed metabolic model based on metaproteomic observations for heterotrophic remineralization (brown) and energetic coupling (yellow dashed lines) of nitrogen (green), sulfur (red), and hydrogen (orange) based chemolithotrophic energy metabolism with carbon fixation (yellow star) for taxa abundant in the metaproteome. Line weight and arrow size indicate magnitude of metabolic activity. Gray lines, activity not occurring under given conditions; light gray taxa, reduced abundance and metabolic activity.

Energy for inorganic carbon fixation within the LO, SNTZ and SZ is derived in large part from denitrification and anammox nitrogen loss processes with the balance between these processes impacting energy flow to either SUP05 or Planctomycetes with concomitant feedback on growth rates. The numerical dominance of SUP05 DNA and protein in the LO, SNTZ and SZ relative to Planctomycetes suggests that partial denitrification outcompetes anammox from a bioenergetic perspective. Indeed, the difference in free energy yield between the two processes (denitrification coupled to sulfide oxidation yields $\sim$ 3.5 times the Gibbs free energy as anammox under standard conditions) is consistent with lower cell abundance and biomass for anammox bacteria even though anammox is observed more frequently than denitrification in many OMZs [12]. As OMZs expand, the contribution of SUP05 to inorganic carbon fixation may have significant impact on global ocean carbon cycling if sufficient energetic substrates are available. However, the fate of carbon fixed in OMZ waters is largely unknown, as the balance between carbon transport and heterotrophic remineralization processes remains to be constrained.

3.3 Conclusions and future implications

This chapter represents the first metaproteome of an O_2-deficient water column encompassing the range of redox conditions, from dysoxic to anoxic sulfidic, found in OMZs globally. Although a recent numerical model by Reed and colleagues attempted to integrate geochemical processes and functional gene markers in the Arabian Sea OMZ [204], my conceptual model uses protein expression to describe differential metabolic coupling among ubiquitous OMZ microbes. The Reed model implicitly assumes reaction rates scale linearly with gene abundance. Thus, the model does not account for biological information flow from DNA to RNA to protein, a regulated process resulting in assembly of pathways driving real world process rates. Incorporation of protein expression information into the model could be used to convert gene abundance into protein abundance or protein production rates, resulting in more accurate predictions.

Additional modeling efforts by Louca *et al.* [132] based on the conceptual model put forth here, incorporate Saanich Inlet multi-omics datasets outlined in Chapter 2 including meta-genomes, -transcritpomes and -proteomes as well as rate measurements for anammox and denitrification.

Louca's meathamatical model reproduces transcript and protein concentration profiles along the Saanich Inlet redox gradient based on fluxes of energetic substrates O_2, NO_3^-, NO_2^- and H_2S. The prediction of gene expression levels solely from geochemical fluxes suggests that for genes involved in energy metabolism, geochemistry is a robust predictor of microbial community structure and gene expression. Indeed, the Louca model predicts a metabolic niche for N_2O reduction within Saanich Inlet, and is later addressed in Chapter 4 with the addition of the dark matter phyla *Marinimicrobia*. The Louca model suggests a central role of SUP05 gammaproteobacteria in the production of NO_2^- fuelled by sulfide oxidation, a trait observed in the recently cultured *Candidatus* Thioglobus autotrophicus [34] and with the potential to feed both further denitrification and anammox driven nitrogen loss. Furthermore, the Louca model strongly supports the findings in my currently conceptual model of the central role of SUP05 as a dominant contributor to inorganic carbon fixation within OMZs, providing insight into ocean carbon and nutrient cycling. Global climate models predict future expansion and intensification of OMZs, with concomitant shoaling and stabilization of sulfidic zones [4, 199]. Such a scenario would provide an increased habitat for SUP05, supporting inorganic carbon fixation via direct oxidation of reduced sulfur compounds and cryptic sulfur cycling in oxygen-deficient waters, resulting in increased primary production and potentially increased carbon sedimentation [13]. Given an estimate of $4.61X10^{18}$ L of O_2-deficient marine waters [3] and the range of observed dark carbon fixation rates from various OMZs of 2.5 – 0.2 μM L^{-1} day^{-1} [81–84], I estimate 0.45 Pg C y^{-1} fixed in OMZs globally. This number represents up to 10% of surface primary production (using 48.5 Pg Cy^{-1}[43]) and will continue to increase with OMZ expansion. With 47% of observed carbon fixation proteins originating from SUP05, I suggest that SUP05 is responsible for 0.2 – 2.4 Pg Cy^{-1}, representing up to 5% of surface primary productivity. Although OMZ expansion is a predicted consequence of global warming, negative feedback loops may ultimately lead to increased drawdown of atmospheric CO_2 driven in large part by blooming SUP05 populations.

3.4 Methods

3.4.1 Sample collection

Sample collection was carried out as described in Chapter 2 for SSU rRNA, metagenomes and metaproteomes, on board the MSV John Strickland in Saanich Inlet April 9 2008 at station S3 (4835.30N, 12330.22W) (Apr08), and September 1, 2009 at station S2 (48°33.106 N, 123° 32.081 W), station S3 and station S4 (48° 38.310 N, 123° 30.007 W) (Sep09) (Figure 3.1A) using a combination of 12 L Niskin and 8 L Go-Flo bottles. Samples for metagenomics, metaproteomics and small subunit ribosomal RNA gene pyrosequencing were collected as described in Zaikova et al. (2009) and Hawley et al. (2016) [8, 114], with the exception the 1.0 L Apr08 metaproteomic samples where RNAlater (Ambion) was used instead of lysis buffer. Multiple depths along the oxycline at all stations and dates were sampled for NO_3^-, NO_2^-, NH_4^+ and H_2S as previously described in Zaikova et al. (2009) and Torres-Beltrán[8, 115] and a SBE O_2 sensor on CTD was used to monitor O_2 concentrations.

3.4.2 Environmental DNA extraction, sequencing and assembly

Environmental DNA extraction was carried out as previously described in Hawley et al. (2013), Zakiova et al. (2009) and in video format (`http://www.jove.com/video/1161/`). Metagenomic samples were sequenced at the Department of Energy Joint Genome Institute (Walnut Creek, CA) by Sanger shotgun sequencing. Sequences were annotated and translated into amino acid sequences using the FGENESB pipeline from Softberry (`www.softberry.com/berry.phtml`) as described in Walsh et al. (2009) supplemental material. Sanger end sequencing of small insert clone libraries from the three Apr08 samples yielded a total of 54,701 open reading frames (ORFs), with an average of 18,234 ORFs per sample. The metagenomic library used for peptide identification was assembled from DNA sequences from 16 fosmid libraries sourced from Saanich Inlet samples collected between February 2006 to February 2007 (accession: LIBGSS_03912-17) ([32]) with the addition of small insert clone libraries from April 2008 samples described in the current study. Libraries were prepared as described in Walsh 2009 supplement [33]. Assembly was carried out separately on libraries from 10 m and on libraries from 100 to 200 m using Phrap

(minmatch 30, maxmatch 55, minscore 55, max_subclone_size 50000, revise_greedy vector_bound 20) [205] to yield 5,620 scaffolds from 10 m libraries and 37,657 scaffolds from 100 - 200 m libraries with a remaining 141,349 un-assembled sequences. Scaffolds were subsequently annotated and translated with the FGENESB pipeline from Softberry (`www.softberry.com/berry.phtml`) using a cutoff of 30 amino acids for minimum protein length to yield of 56,476 ORFs from 10 m libraries, 57,674 ORFs from 100 - 200 m libraries and 112,828 ORFs from unassembled sequences.

3.4.3 PCR amplification of SSU rRNA gene for pyrotag sequencing and analysis

Pyrosequencing of Sep09 samples was carried out as described in Allers et al. (2013)[31]. Briefly, small subunit ribosomal RNA (SSU rRNA) gene pyrosequencing of the 12 Sep09 samples yielded 87,138 sequences that clustered into 3,385 non-singleton operational taxonomic units at the 97% identity threshold.

3.4.4 Environmental protein extraction and identification

Total protein was extracted from Sterivex filters as described in Hawley et al. (2013) [206] and described in Chapter 2. Briefly, BugBuster (Novagen) was added to Sterivex filters to lyse cells, and lysate was extruded. Buffer exchange with 100 mM NH_4HCO_3 was carried out using a 10K Amicon (Millipore) and purified proteins were subject to overnight Trypsin digestion followed by cleanup on C18 (Sigma-Aldrich) and strong cation exchange solid phase extraction columns. Digested protein concentration was determined by bicinchoninic acid assay. Environmental peptides were analyzed by tandem-mass spectrometry (MS/MS) at the Environmental Molecular Sciences Laboratory at Pacific Northwest National Labs (Richland, WA) as described in Hawley et al. (2013) [206] using on-line capillary liquid-chromatography-tandem mass spectrometry on a Thermo LTQ ion trap or LTQ-Orbitrap using data-dependent fragmentation. Detected peptides were identified from MS/MS using SEQUEST™with a mass spectra generating function (MS-GF) cutoff value below 10-11, corresponding to a false discovery rate of less the 2% [207]. Peptides were searched against the Saanich Inlet metagenomic database, consisting of an assembly of metagenomic sequences from previously sequenced fosmid end libraries of environmental DNA collected in Saanich Inlet from multiple depths between February 2006 and April 2007 [32] and

April 2008 metagenomic samples comprising a total of 176,978 protein sequences. Only peptides matched to protein sequences with a peptide prophet probability (PPP) score ≥ 0.95 were used in further analysis. Tandem mass-spectroscopy coupled liquid chromatography (LC-MS/MS) metaproteomic sequencing identified a total of 5,019 unique proteins, a number comparable to previous marine metaproteomic studies [208]. A consistent number of proteins were identified across the Sep09 samples, with an average of 695 unique proteins per sample Table B.1. While variability in protein detection in the Apr08 samples was considerable, the high number of unique proteins detected in the Apr08 200 m sample (4,344) was exceptional, enabling identification of more complete metabolic pathways.

3.4.5 Functional and taxonomic assignment of metagenome and metaproteome

Taxonomy and function for all metagenomic (translated) and metaproteomic amino acid sequences were assigned as the top hit from BLASTP [209] against NCBI RefSeq database (Nov. 7, 2012) augmented to include SUP05 metagenome [33], *Candidatus* Kuenenia stuttgartiensis , and *Candidatus* Scalindua profunda [208] using a BLAST score ratio of 0.4 as a cutoff [186]. Relative abundance of genes in the metagenome was reported as the sum of all sequence reads with hit to a given accession divided by the total number of sequence reads in a given sample over 30 amino acids in length. In the metaproteome the relative abundance of a detected protein was reported as the normalized spectral abundance was described in Hawley et al. 2013 [206] . Within a given sample, peptide scan counts for each protein were summed, and in cases where peptide sequences matched to multiple proteins the scan count for the peptide was divided by the number of proteins it matched to. The scan count for each protein was then divided by protein length (in amino acids) to give the spectral abundance factor (SAF), which was then divided by the sum of all SAFs for a given sample to yield the normalized spectral abundance factor (NSAF) for a given protein.

3.4.6 Hierarchical clustering of metaproteomic samples

The NSAF values for all detected proteins with a PPP ≥ 0.95 were used in the calculation of a Sorensen distance matrix using PC-ORD software, and a group average method was used for

grouping in construction of clusters.

Chapter 4

Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients[3]

As introduced in Chapter 1, microbial communities drive biogeochemical cycles through networks of metabolite exchange that are structured along redox and energy gradients. As energy yields become limiting, these networks favor co-metabolic interactions to maximize energy yield. In this chapter, I apply single-cell genomics and use metagenomic, and metatranscriptomic datasets described in Chapter 2 to study populations of the abundant microbial dark matter group Marinimicrobia along defined energy gradients. I show that evolutionary diversification of major Marinimicrobia clades appears to be closely related to energy yields, with increased co-metabolic interactions in more deeply branching clades. Several of these clades participate in the biogeochemical cycling of sulfur and nitrogen, filling previously unassigned niches in the ocean. Notably, two Marinimicrobia clades, occupying different energetic niches, express nitrous oxide reductase, potentially acting as a global sink for the greenhouse gas nitrous oxide.

4.1 Introduction

The laws of thermodynamics apply to all aspects of Life, governing energy flow in both biotic and abiotic regimes. Nicholas Georgescu-Roegen was the first to directly apply the laws of

[3]A version of this chapter has been published at Nature Communications as *Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients* in 2017 by Hawley, A. K., Nobu, M. K., Wright, J. J., Durno, W. E., Morgan-Lang, C., Sage, B., Schwientek, P., Swan, B. K., Rinke, C., Torres-Beltrán , M., Mewis, K., Liu, WT., Stepanauskas, R., Woyke, T., Hallam, S. J.

thermodynamics to economic theory, bringing to the forefront the reality of limited natural resources on sustainable growth [210]. Robert Ayers used the term eco-thermodynamics to describe the application of thermodynamics and energy flow to economic models with the controversial conclusion that future economic growth would necessitate the recycling of goods [211]. Within microbial ecology there is an emerging consensus that these same organizing principles structure microbial community interactions and growth with feedback on global nutrient and energy cycling [38, 132, 204, 212]. Indeed, recycling in the common sense may be analogous to metabolite exchange or use of public goods [213], as the goods from one production stream become available for growth of another. Microbial communities living near thermodynamic limits where high potential electron acceptors are scarce tend to utilize differential modes of metabolic coupling including obligate syntrophic interactions, maximizing any chemical disequilibria to yield energy for growth [90, 102]. Thus, the term eco-thermodynamics takes on new meaning in the context of microbial ecology where thermodynamic constraints directly shape the structure and activity of microbial interaction networks.

Eco-thermodynamic gradients are formed by the distribution of available electron donors and acceptors within the physical environment, creating metabolic niches that are occupied by diverse microbial partners playing recurring functional roles [214, 215]. Marine oxygen minimum zones (OMZs) provide a vivid example of eco-thermodynamic gradients shaping differential modes of metabolic coupling at the intersection of carbon, nitrogen and sulfur cycling in the ocean [2, 79]. For example, OMZ microbial communities manifest a modular denitrification pathway that links oxidation of reduced sulfur compounds to nitrate reduction and nitrous oxide (N_2O) production [33, 79, 134]. While many of the most abundant interaction partners are known, recent modeling efforts point to a novel metabolic niche for the terminal step in the denitrification pathway (nitrous oxide reduction to dinitrogen gas) occupied by unidentified community members [132]. By defining the interaction networks coupling microbial processes along eco-thermodynamic gradients, it becomes possible to more accurately model nutrient and energy flow at ecosystem scales.

Recent advances in sequencing technologies have opened a genomic window on uncultivated microbial diversity, illuminating the metabolic potential of numerous candidate divisions also

know as microbial dark matter (MDM) [19, 212, 216]. Many MDM organisms occupy low energy environments where high energy terminal electron acceptors are scarce, and appear to form obligate metabolic dependencies that could help explain resistance to traditional isolation methods. Marine Marinimicrobia have been previously implicated in sulfur cycling via a polysulfide reductase gene cluster [31, 32]. In studies of a methanogenic bioreactor, Marinimicrobia have also been identified to rely on syntrophic interactions with metabolic partners to accomplish degradation of amino acids [217]. The global distribution of Marinimicrobia clades indicates a much wider diversity of metabolic functions and interacting partners than currently described. Here we use shotgun metagenomics, metatranscriptomics and single-cell genomics to investigate energy metabolism within the Marinimicrobia to reveal novel modes of metabolic coupling with important implications for nutrient and energy cycling in the ocean.

4.2 Results and Discussion

4.2.1 Marinimicrobia single-cell amplified genomes and phylogeny

A total of 25 Marinimicrobia single-cell amplified genomes (SAGs) from sources along eco-thermodynamic gradients were identified globally by flow sorting, whole-genome amplification and sequencing (Table 4.1). SAG de novo assemblies ranged in size from 0.39 to 2.01 million bases (Mb) with estimated genome completeness ranging from <10% to >90% (average 45%) (Table4.1). Most Marinimicrobia SAGs manifested streamlined genomes, with high coding base percentage ($89.99 - 97.13\%$) and low cluster of orthologous group (COG) redundancy ($1.08 - 1.16$) (Figure C.1). PhyloPhlAn analysis [218] using concatenated sequences of conserved marker genes placed Marinimicrobia SAGs within the bacterial domain branching deeply from the closest cultured thermophilic representative *Caldithrix abyssi* (Figure C.2). To determine phylogenetic diversity within the Marinimicrobia, we constructed a comprehensive SSU rRNA gene tree from identified Marinimicrobia SSU rRNA genes, resolving 17 clades (Figure 4.1). SAG sequences were affiliated with 10 clades spanning the entire breadth of the Marinimicrobia tree (Figures 4.1 and 4.2 A and B), providing a broad phylogenetic range with which to assess distribution patterns and energy metabolism within the phylum.

Table 4.1: Genomic Features of SAGs

	Sampling depth (m)	Oxygen status*	Clade	Assembly size (bp)	Estimated Genome Completeness (%)	Estimated Genome Size (bp)	Number of contigs	Largest contig (bp)	Largest contig as % of assembly	GC content (%)	# predicted genes	# protein-coding genes	# RNA genes	# CRISPR
Gulf of Maine (43°84′N, 69°64W)														
AAA160-I06	1	oxic	ZA3312c-A	884,929	75.8	900,037	32	107,580	12.2	32.4	1,007	971	36	1
AAA160-C11	1	oxic	ZA3312c-A	824,595	78.0	955,155	17	190,175	23.1	32.7	883	852	31	1
AAA076-M08	1	oxic	ZA3312c-A	392,252	41.2	577,302	18	71,263	18.2	32.6	440	427	13	1
AAA160-B08	1	oxic	ZA3312c-A	922,550	78.3	1,296,752	26	21,880	2.4	33.0	995	965	30	1
HOT Station ALOHA (22°45_ N, 158°00_ W)														
AAA0298-D23	25	oxic oxic	ZA3312c-B	979,176	92.3	979,176	10	407,043	41.6	31.0	1,055	1,016	39	1
South Atlantic Gyre (12°29′S, 4°59′W)														
AAA003-E22	800	oxic	HF77OD10	888,773	33.5	1,737,273	53	67,111	7.6	37.0	788	764	24	2
AAA003-L8	800	oxic	ZA3648c	1,265,631	72.4	2,004,182	84	138,629	11.0	30.0	1,258	1,231	27	2
Saanich Inlet														
AB-746_P06AB-902	100	dysoxic	Arctic96B-7-B	1,538,315	77.6	1,848,863	69	139,459	9.1	32.0	1,547	1,522	25	1
AB-746_N13AB-902	100	dysoxic	Arctic96B-7-A	1,899,548	67.0	2,255,438	72	115,617	6.1	39.1	1,882	1,843	39	1
AB-747_F21AB-903	100	dysoxic	Arctic96B-7-A	751,298	8.6	8,976,777	55	64,402	8.6	38.9	799	789	10	1
AB-750_L13AB-904	150	anoxic	SHBH1141	2,009,596	57.1	2,659,593	112	86,252	4.3	43.0	1,921	1,885	36	1
AB-750_A02AB-903	150	anoxic	SHBH1141	1,925,097	45.8	3,701,687	136	88,465	4.6	43.1	1,762	1,741	21	1
AB-751_D09AB-904	150	anoxic	SHBH1141	569,612	11.5	3,917,216	45	50,480	8.9	43.2	528	517	11	1
AB-755_M21D07	185	sulfidic	SHAN400	999,835	TBD	2,877,074	70	53,898	5.4	37.0	971	953	18	1
AB-755_E16C12	185	sulfidic	SHBH1141	781,004	18.3	4,196,372	75	77,534	9.9	42.0	819	806	13	1
Northeast subarctic Pacific (48°52′N, 130°40′W)														
JGI 0000113-D11	2,000	dysoxic	Arctic96B-7-B	645,847	48.3	2,643,012	91	33,298	5.2	32.0	756	727	29	1
Terephthalate degrading reactor (40°6′N, 88°13′W)														
0000039-E15 (TAsludge)	na	0.0	HMTAb91-B	493,984	32.8	1,343,421	30	58,668	11.9	47.2	461	447	14	1
0000059-E23 (TAbiofilm)	na	0.0	HMTAb91-B	618,060	17.2	3,865,965	41	75,054	12.1	47.3	583	572	11	1
0000059-D20 (TAbiofilm)	na	0.0	HMTAb91-B	772,719	41.3	1,534,399	34	88,615	11.5	47.7	718	696	22	1
0000059-L03 (TAbiofilm)	na	0.0	HMTAb91-B	597,650	27.5	1,115,912	41	55,420	9.3	47.8	585	579	6	1
0000077-B04 (TAbiofilm)	na	0.0	HMTAb91-B	496,915	15.5	3,885,254	28	66,910	13.5	47.4	479	469	10	1
0000039-O11 (TAsludge)	na	0.0	HMTAb91-B	828,046	34.5	2,959,672	75	32,891	4.0	47.0	749	729	20	1
0000039-D08 (TAsludge)	na	0.0	HMTAb91-B	978,259	41.8	1,549,116	56	61,500	6.3	47.4	894	878	16	1
0000090-C20	na	0.0	HMTAb91-B	566,695	27.6	2,726,674	46	37,004	6.5	47.6	511	494	17	1
Etoliko Lagoon Sediment (38°47; N, 21°33′ E)														
AAA257-N23	na	0.0	HMTAb91-A	948,616	27.4	2,239,091	71	50,692	5.3	39.0	912	898	14	1

*oxic (>90 µM O_2), dysoxic (90 µM < O_2 > 20 µM), suboxic (20 µM< O_2 > 2 µM), anoxic (< 2 µM O_2), sulfidic.

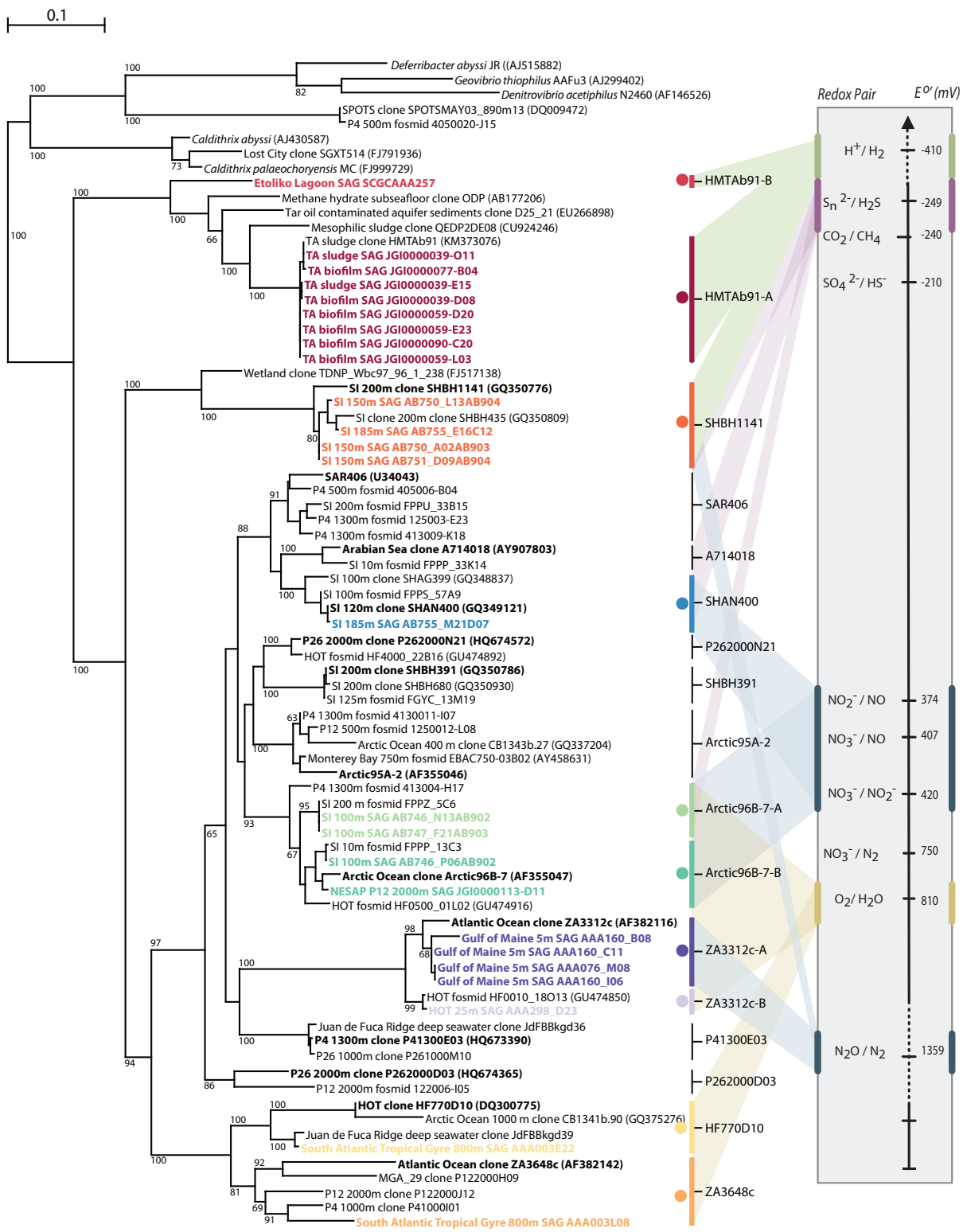

0.1
Deferribacter abyssi JR ((AJ515882)
Geovibrio thiophilus AAFu3 (AJ299402)
Denitrovibrio acetiphilus N2460 (AF146526)
SPOTS clone SPOTSMAY03_890m13 (DQ009472)
P4 500m fosmid 4050020-J15
Caldithrix abyssi (AJ430587)
Lost City clone SGXT514 (FJ791936)
Caldithrix palaeochoryensis MC (FJ999729)
Etoliko Lagoon SAG SCGCAAA257
Methane hydrate subseafloor clone ODP (AB177206)
Tar oil contaminated aquifer sediments clone D25_21 (EU266898)
Mesophilic sludge clone QEDP2DE08 (CU924246)
TA sludge clone HMTAb91 (KM373076)
TA sludge SAG JGI0000039-O11
TA biofilm SAG JGI0000077-B04
TA sludge SAG JGI0000039-E15
TA biofilm SAG JGI0000039-D08
TA biofilm SAG JGI0000059-D20
TA biofilm SAG JGI0000059-E23
TA biofilm SAG JGI0000090-C20
TA biofilm SAG JGI0000059-L03
Wetland clone TDNP_Wbc97_96_1_238 (FJ517138)
SI 200m clone SHBH1141 (GQ350776)
SI 150m SAG AB750_L13AB904
SI clone 200m clone SHBH435 (GQ350809)
SI 185m SAG AB755_E16C12
SI 150m SAG AB750_A02AB903
SI 150m SAG AB751_D09AB904
SAR406 (U34043)
P4 500m fosmid 405006-B04
SI 200m fosmid FPPU_33B15
P4 1300m fosmid 125003-E23
P4 1300m fosmid 413009-K18
Arabian Sea clone A714018 (AY907803)
SI 10m fosmid FPPP_33K14
SI 100m clone SHAG399 (GQ348837)
SI 100m fosmid FPPS_57A9
SI 120m clone SHAN400 (GQ349121)
SI 185m SAG AB755_M21D07
P26 2000m clone P262000N21 (HQ674572)
HOT fosmid HF4000_22B16 (GU474892)
SI 200m clone SHBH391 (GQ350786)
SI 200m clone SHBH680 (GQ350930)
SI 125m fosmid FGYC_13M19
P4 1300m fosmid 4130011-I07
P12 500m fosmid 1250012-L08
Arctic Ocean 400 m clone CB1343b.27 (GQ337204)
Monterey Bay 750m fosmid EBAC750-03B02 (AY458631)
Arctic95A-2 (AF355046)
P4 1300m fosmid 413004-H17
SI 200 m fosmid FPPZ_5C6
SI 100m SAG AB746_N13AB902
SI 100m SAG AB747_F21AB903
SI 10m fosmid FPPP_13C3
SI 100m SAG AB746_P06AB902
Arctic Ocean clone Arctic96B-7 (AF355047)
NESAP P12 2000m SAG JGI0000113-D11
HOT fosmid HF0500_01L02 (GU474916)
Atlantic Ocean clone ZA3312c (AF382116)
Gulf of Maine 5m SAG AAA160_B08
Gulf of Maine 5m SAG AAA160_C11
Gulf of Maine 5m SAG AAA076_M08
Gulf of Maine 5m SAG AAA160_I06
HOT fosmid HF0010_18O13 (GU474850)
HOT 25m SAG AAA298_D23
Juan de Fuca Ridge deep seawater clone JdFBBkgd36
P4 1300m clone P41300E03 (HQ673390)
P26 1000m clone P261000M10
P26 2000m clone P262000D03 (HQ674365)
P12 2000m fosmid 122006-I05
HOT clone HF770D10 (DQ300775)
Arctic Ocean 1000 m clone CB1341b.90 (GQ375276)
Juan de Fuca Ridge deep seawater clone JdFBBkgd39
South Atlantic Tropical Gyre 800m SAG AAA003E22
Atlantic Ocean clone ZA3648c (AF382142)
MGA_29 clone P122000H09
P12 2000m clone P122000J12
P4 1000m clone P41000I01
South Atlantic Tropical Gyre 800m SAG AAA003L08
HMTAb91-B
HMTAb91-A
SHBH1141
SAR406
A714018
SHAN400
P262000N21
SHBH391
Arctic95A-2
Arctic96B-7-A
Arctic96B-7-B
ZA3312c-A
ZA3312c-B
P41300E03
P262000D03
HF770D10
ZA3648c
Redox Pair
E0′ (mV)
H+/ H2
-410
Sn 2-/ H2S
-249
CO2 / CH4
-240
SO4 2-/ HS-
-210
NO2- / NO
374
NO3- / NO
407
NO3- / NO2-
420
NO3- / N2
750
O2/ H2O
810
N2O / N2
1359

Figure 4.1: Maximum likelihood small subunit rRNA tree and energy metabolism redox pairs for Marinimicrobia lineages (previous page). Maximum likelihood phylogenetic tree of small subunit ribosomal rRNA (SSU rRNA) gene from all available studies. SSU rRNA gene from SAGs used in this study are in bold and coloured to indicate there membership to population genome bins. Energy metabolism redox pairs for each lineage explored in this publication are mapped to electron tower on the right of the tree. The bar represents 1% estimated sequence divergence. Bootstrap values below 50% are not shown.

4.2.2 Biogeography of Marinimicrobia clades

Using this phylogenetic information, we determined the global biogeographic distribution of Marinimicrobia and specific SAG-affiliated clades along eco-thermodynamic gradients spanning oxic (>90 μM O_2), dysoxic (20 − 90 μM O_2), suboxic (1 −20 μM O_2), anoxic(<1 μM O_2), sulfidic and methanogenic conditions. Estimates of Marinimicrobia total abundance and clade distribution were carried out by a robust survey of 594 globally sourced metagenomes (549 assembled Illumina data sets and 45 unassembled 454 data sets) across terrestrial and marine ecosystems, including Northeastern Subarctic Pacific (NESAP, n = 43), Saanich Inlet (SI, n = 90), Eastern Tropical South Pacific (ETSP, n = 6), Peruvian (n = 17), and Guaymas Basin (n = 2) OMZs; TARA Oceans (n = 243) and several other marine (n = 141) and terrestrial sites (n = 52), (Table C.1) totalling 127 Gigabases (Gb) of sequence information. To estimate total abundance, we used a sequence similarity recruitment with a cutoff of >70% nucleotide identity over >70% of the metagenomic contig. Globally we recovered 1.3 Gb of Marinimicrobia-affiliated sequence or 1.3 million genome equivalents (assuming 1Mb average genome) representing ∼1% of surveyed data. The recovery of Marinimicrobia-affiliated sequences was highest in coastal OMZs, increasing in relation to decreasing O_2 concentration (Figure C.3A). Recovery was more variable in other marine locations and minimal in terrestrial locations. To more fully resolve this sequence information at the level of specific Marinimicrobia clades, we conducted a more stringent recruitment of ≥95% nucleotide identity across ≥200 bp intervals. On a global scale three clades constituted 75% of observed Marinimicrobia with the remaining seven clades making up the difference (Figure C.3B). Consistent with previous results, predominantly marine sites were recruited with two hits from terrestrial locations. Sakinaw Lake, a meromictic lake with high methane concentrations [216], was the only geographic location with recruitment to the HMTAb91 clade. Within marine systems,

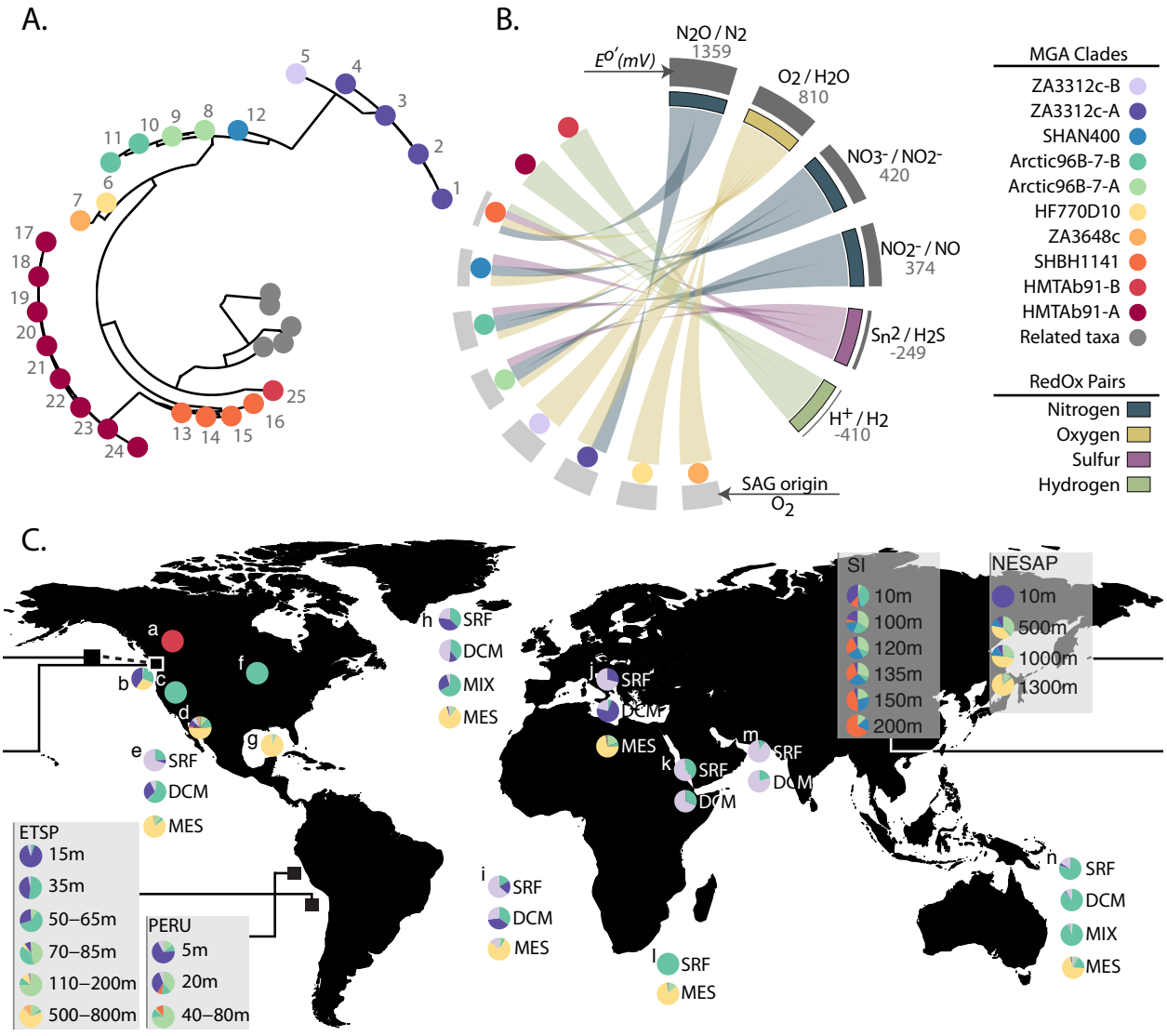


Figure 4.2: Phylogeny and electron donors of Marinimicrobia and Biogeographic distribution. **(A)** Unrooted phylogenetic tree based on small subunit rRNA (SSU rRNA) gene in single-cell amplified genomes (SAG) showing the phylogenetic affiliation of Marinimicrobia SAGs, each dot represents a SAGs in Table S1 with the corresponding number. **(B)** Circular plot indicating the terminal electron acceptors used and their respective E ° '(mV) value (right) by the different Marinimicrobia clades (left). **(C)** Global distribution of Marinimicrobia SAG-related microorganisms, as determined by metagenomic fragment recruitment using FAST(Online) with 595 global metagenomes with a threshold of $\geq$ 95% nucleotide sequence identity and alignments $\geq$ 200 bp. Recruited contig lengths were normalized by the length of each SAG assembly in mega base pairs (Mbp) and to the size of the metagenome of origin in Mbps.

SAGs recruited sequences from cognate environments and conditions consistent with observed tree branching patterns (Figure 4.2, Table C.2). Overall, trends indicated that specific clades inhabit particular energetic niches with potential for metabolic coupling within a given niche.

Population genome bin construction

To determine the energy metabolism of Marinimicrobia clades and overcome low genome completion of some SAGs, we leveraged extensive metagenomic and metatranscriptomic resources from NESAP and Saanich Inlet time series [114, 219] to construct population genome bins, improving estimated genome completion to an average of 87% (Table C.3). Metagenomic contigs >5000bp and with >95% identity to SAGs were identified followed by tetra-nucleotide frequency analysis to resolve specific clades (Figure 4.3a). A total of five population genomes for Marinimicrobia clades ZA3312c-A/B, HF770D10, Arctic96B-7-A/B, SHAN400, and SHBH1141 spanning oxic, dysoxic, suboxic, anoxic, and anoxicsulfidic conditions were resolved from Saanich Inlet and NESAP metagenomes, enabling more complete metabolic reconstruction within each clade (Figure 4.3a, b). A sixth clade (HMTAb91-A), endemic to a methanogenic bioreactor branching near the base of Marinimicrobia radiation was included in downstream comparisons of metabolic potential to encompass the complete range of electron donoracceptor pairs. Energy metabolism of Marinimicrobia population genomes was examined in relation to tree branching patterns and environmental disposition. A total of 18 metatranscriptomes from six depths and three time points (Figure 4.4 a and b) were used to explore Marinimicrobia gene expression over defined energy gradients including a deep water renewal event resulting in the influx of oxygenated nutrient rich waters in Saanich Inlet basin waters. This enabled the resolution of metabolic niches and indicted potential modes of metabolic coupling within specific Marinimicrobia clades.

4.2.3 Metabolic reconstruction and gene model validation

Marinimicrobia clades ZA3312cA/B and HF770D10 were most abundant under oxic water column conditions with extensive genome streamlining comparable to *Ca. Pelagibacter ubique* (Figure C.1A). All three clades harbored genes encoding for aerobic respiration, and heterotrophy with no indication for autotrophic CO_2 fixation. ZA3312c clades also encoded the oxidative tricarboxylic

acid (TCA) cycle (Table C.4) and proteorhodopsin, a proton-pump used to harness light energy (Figure 4.3b) [220]. ZA3312c proteorhodopsin transcripts were highly expressed in oxic surface waters of Saanich Inlet, suggesting that ZA3312c are capable of supplementing organotrophy with phototrophy in surface waters, a trait well suited to open-ocean oligotrophic environments (Figure C.6A). Interestingly, ZA3312c-A also encoded nitrous oxide reductase (*nosZ*) and associated maturation factors (*nosL*, *nosD*, and *nosY*) that drive the reduction of N_2O to N_2 in the terminal step of denitrification. Transcripts for *nosZ* were expressed throughout the Saanich Inlet water column (Figures 4.4A, C.7) and indicate potential coupling to ammonia oxidizing Thaumarchaea that produce N_2O as a byproduct of ammonia oxidation [203]. ZA3312c-A *nosZ* transcripts were also detected in suboxic waters of the NESAP, Peru, and ETSP OMZs, and four TARA oceans metagenomes contained ZA3312c-A *nosZ* sequences (>80% nucleotide identity) (Figure 4.4B) reinforcing a global distribution pattern with functional implications for marine nitrogen budgets and greenhouse gas cycling.

Marinimicrobia clades Arctic96B-7-A and B were widespread in dysoxic ocean waters. Arctic96B-7 clades harbored genes encoding for aerobic respiration, organotrophy and oxidative TCA cycle with no indication for proteorhodopsin or autotrophic CO_2 fixation (Table C.4). Arctic96B-7 clades may supplement energy generation in a similar manner to proteorhodopsin through catabolism of the common ocean compound oxalate [221], coupling a unique oxalate: formate antiporter and oxalate decarboxylase [222]. The Arctic96B 7-A clade also encoded nitrate reductase (*narG*), and polysulfide (polyS) reductase (*psrABC*) (Figures 4.3b, 4.4a) that were expressed throughout the Saanich Inlet water column. Peak expression corresponded to depths with low NO_3^- and no detectable H_2S. Interestingly, the PsrABC enzyme complex can use H_2S as an auxiliary electron donor through PsrABC-mediated H_2S oxidation to polyS and stored polyS can serve as an alternative electron sink, regenerating H_2S. The combination of *narG* and *psrABC* provides Arctic96B-7 clades with versatile energy metabolism with potential coupling to both sulfur oxidizing bacteria (ARCTIC96-BD19, SUP05) by regenerating H_2S under non-sulfidic conditions, and anaerobic ammonium (Planctomycetes) and nitrite (Nitrospina) oxidizing bacteria through the production of NO_2^- in dysoxic, suboxic, and anoxic waters (Figure 4.5A). Thus, Arctic96B-7 clades may form supportive metabolic partnerships with major of primary producers in OMZs

critical to the biogeochemical cycling of carbon, nitrogen, and sulfur [79]

Table 4.2: Genomic features of Marinimicrobia population genomes

Lineage	Singel Cell Genome Identity	Population Genome Size (Mbp)	Estimated Completeness (%)	Number of contigs	N50	GC content (%)	Number of single copy marker genes*	Strain Heterogeneity*
ZA3312c-A	AAA160-I06, AAA160-C11, AAA076-M08, AAA160-B08	11	95.8	531	35213	32.8%	56	94.33
ZA3312c-B	AAA0298-D23	1	93.4	41	236078	31.6%	147	28
HF770D10	AAA003-E22	1.4	41.2	118	15724	36.6%	104	100
Arctic96B7_A	AB-746_N13AB-902, AB-747_F21AB-903	50.9	100.0	3423	18609	39.4%	56	70.67
Arctic96B7_B	AB-746_P06AB-902, JGI 0000113-D11	6.0	96.6	583	13227	32.6%	104	63.36
SHAN400	AB-755_M21D07	32.2	87.5	2196	19252	37.4%	56	99.64
SHBH1141	AB-750_L13AB-904, AB-755_E16C12, AB-751_D09AB-904	65.6	91.7	3127	35279	43.5%	56	96.11

* Parks et al. Genome Research 2015

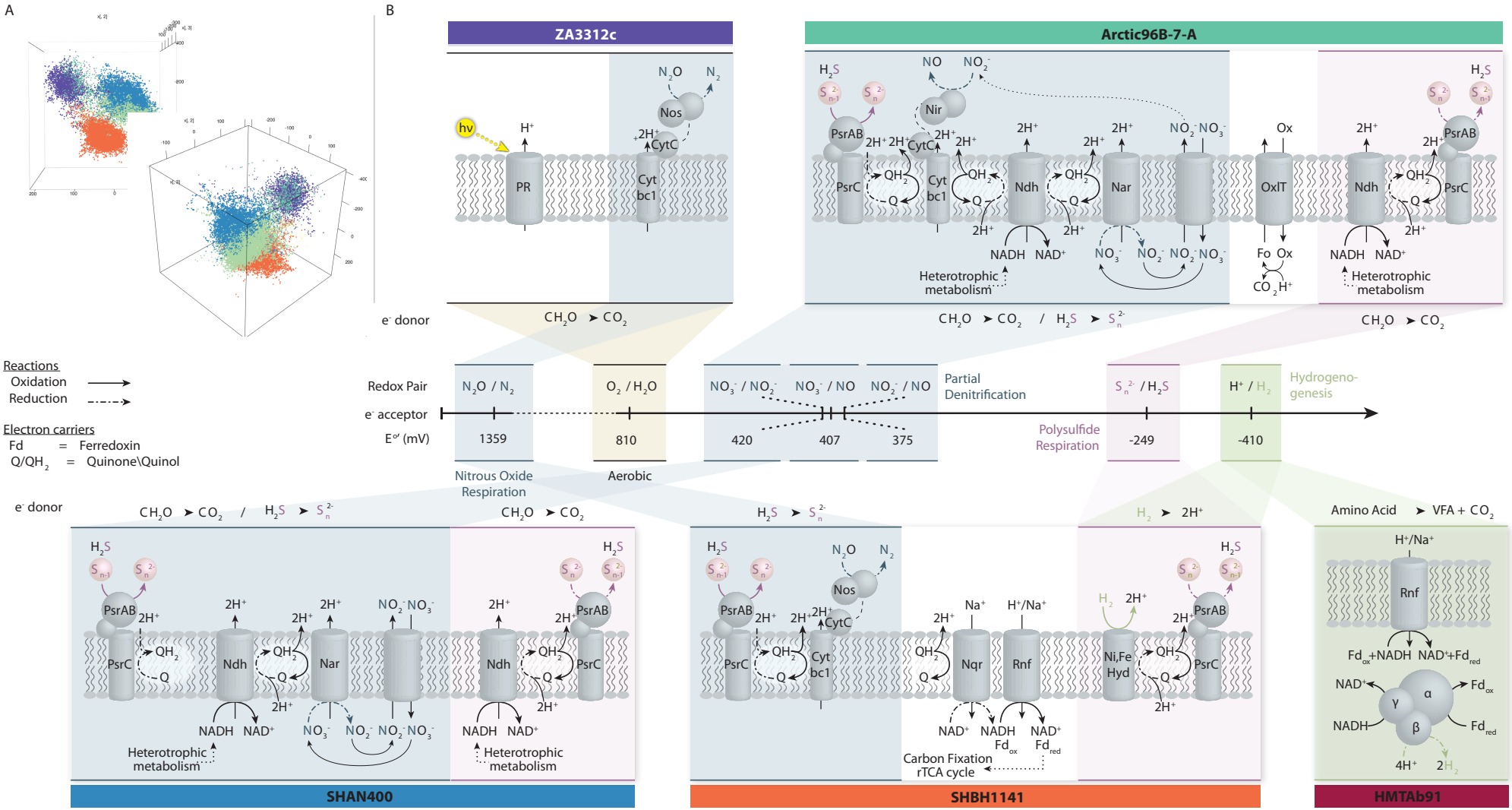

A
B
ZA3312c
Arctic96B-7-A
e- donor
CH2O ➤ CO2
CH2O ➤ CO2 / H2S ➤ Sn2-
CH2O ➤ CO2
Reactions
Oxidation
Reduction
Electron carriers
Fd = Ferredoxin
Q/QH2 = Quinone\Quinol
Redox Pair
e- acceptor
E°' (mV)
N2O / N2
O2 / H2O
NO3- / NO2-
NO3- / NO
NO2- / NO
Partial Denitrification
Sn2- / H2S
H+ / H2
Hydrogeno-genesis
1359
810
420
407
375
-249
-410
Nitrous Oxide Respiration
Aerobic
Polysulfide Respiration
Heterotrophic metabolism
Amino Acid ➤ VFA + CO2
SHAN400
SHBH1141
HMTAb91

Figure 4.3: Energy metabolism of Marinimicrobia population genome bins(Previous page). **(A)** Binning of Marinimicrobia population genomes by Kmer frequency principal component analysis, two rotations of three-dimensional plot, clouds of color coded genome bins are apparent. **(B)** Summary of co-metabolic and energy metabolism and conservation strategies of Marinimicrobia population genomes from along eco-thermodynamic gradients, for nitrogen (blue), sulfur (pink), and hydrogen (green). Enzymes include: proteorhodopsin (PR), sulfur: polysulfide reductase (PsrAB, PsrC); nitrogen: nitrite reductase (Nir), nitrate reductase Nar, nitrate/nitrite antiporter (NirK), nitrous oxide reductase (Nos); hydrogen metabolism: Ni,Fe hydrogenase (Ni,Fe Hyd), hydrogenase complex (HydBD); respiratory elements: cytochrome bc1 complex (Cytbc1), NADH dehydrogenase (Ndh), energy-conserving putative electron transfer mechanisms putative ion-translocating ferredoxin:NADH oxidoreductase (IfoAB); oxalate transporter (OxlT); *Rhodobacter* nitrogen fixation complex (Rnf). Oxidation and reduction indicated by solid or dotted arrows respectively.

Marinimicrobia clade SHAN400 appears to be endemic to Saanich Inlet where it is most abundant below the oxycline (Figure C.5). SHAN400 harbored genes encoding for aerobic and anaerobic respiration, heterotrophy and oxidative TCA cycle. SHAN400 also encoded ferredoxin, pyruvate metabolism, and NADH dehydrogenase (Figures 4.3b, C.9, C.8), potentially providing additional electron shuttles for energy metabolism under anoxic conditions. Like Arctic96B-7, SHAN400 encoded *narG* and *psrABC*, potentially linking its energy metabolism to both sulfur-oxidizing bacteria (SUP05) and anaerobic ammonium- (*Planctomycetes*) and nitrite- (*Nitrospina*) oxidizing bacteria in anoxic waters (Figures 4.3b, 4.4a, C.6ab). In contrast to Arctic96B-7, SHAN400 transcripts for heme/copper-type cytochrome and NADH dehydrogenase were most highly expressed in anoxic waters (Figure C.8). This is consistent with redox-driven niche partitioning between Arctic96B-7 and SHAN400 clades in the Saanich Inlet water column.

Marinimicrobia clade SHBH1141 was prevalent in anoxic and anoxic-sulfidic OMZ waters (Figure C.5). SHBH1141 harbored genes encoding for aerobic and anaerobic respiration, autotrophic CO_2 fixation via the reductive TCA cycle (citrate lyase and ferredoxin-dependent 2-ketoacid oxidoreductases), and the Rhodobacter nitrogen fixation (Rnf) complex to produce reduced ferredoxin to drive endergonic reductive carboxylation steps, indicating a capacity to carry out autotrophy (Figures C.9, C.8). In addition, SHBH1141 encoded *psrABC*, class I [Ni,Fe] hydrogenases (*hybOABCD*) and *nosZ* with associated maturation factors *nosL* and *nosD* (Figures 4.3b, C.6, C.9). Gene expression for *psrABC*, *hybOABCD*, and *nosZ* was elevated under anoxic to sulfidic conditions (120 m in July 2010, and 150 m in July and August 2010; Figure 4.4a).

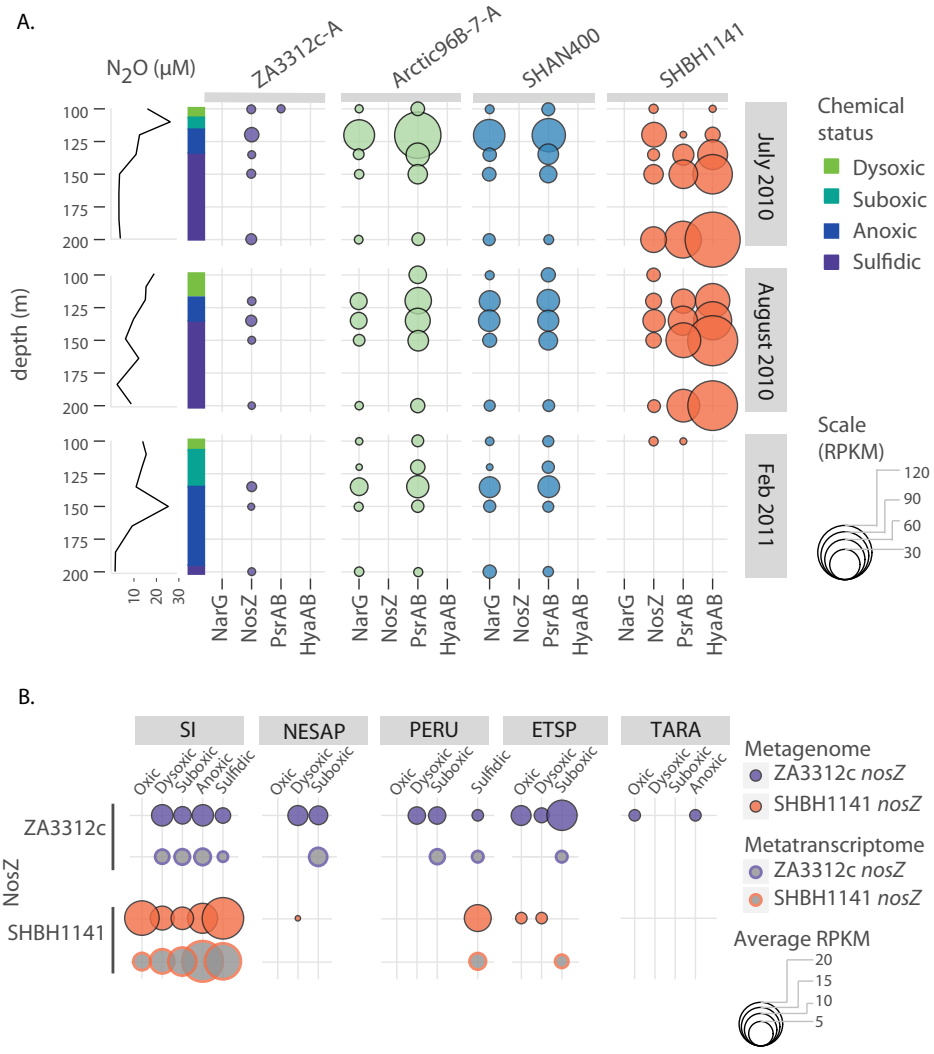


Figure 4.4: Expression of selected Marinimicrobia energy metabolism genes in Saanich Inlet. **(A)** Expression of selected genes involved in Marinimicrobia energy metabolism from Saanich Inlet station SI03 at three time points between 100 and 200 m. Size of circle represents reads per kilobase per million mapped (RPKM) (see methods) for metatranscriptomic reads mapped to the selected genes for the indicated population genomes. Water column redox status for each time point encoded on left axis and nitrous oxide concentration profile for each time point on left. Enzymes nitrate reductase (*narG*), Nitrous oxide reductase (*nosZ*), polysulfide reductase subunits A and B (*psrAB*) and Ni-Fe hydrogenase subunits A and B (*hyaAB*). **(B)** Detected genes and transcripts for Marinimicrobia ZA3312c and SHBH1141 nosZ along eco-thermodynamic gradients from oxic (>90 μM O_2), dysoxic (20-90 μM O_2), suboxic (1-20 μM O_2), anoxic (<2 μM O_2), and sulfidic conditions in Saanich Inlet time series, Northeastern Subarctic Pacific (NESAP), Peru, Eastern Tropical South Pacific (ETSP) and TARA Oceans (no transcriptomes available) data sets. For SI and ETSP dot size represents average reads per killobase per million mapped (RPKM) summed for a given nosZ type each metagenome or metatranscriptome and averaged by the total number of metagenomes or metatranscriptomes for a given water column classification. For ETSP, Peru and TARA bubble size is the number of reads (ETSP and Peru) or contigs (TARA) with nosZ averaged per number of metagenome or metatranscriptomes for a given water column classification.

SHBH1141 class I [Ni,Fe] hydrogenase is proposed to operate bidirectionally based on observations in *Escherichia coli* and *Salmonella enterica*, with proposed hydrogen production under more oxidizing conditions [223]. SHBH1141 *nosZ* was recovered on a global scale and expressed under both sulfidic conditions in Peru and suboxic conditions in the ETSP as well as Saanich Inlet (Figure 4.4b), positing a central role for SHBH1141 in OMZ N_2O reduction. The expression of these genes in nonsulfidic waters points to a new mode of dynamic metabolic mutualism in which SHBH1141 may rely on SUP05 N_2O generation in anoxic and sulfidic waters [34, 79] to store polyS and re-evolve H_2S from polyS to stimulate SUP05 N_2O production (Figure 4.5b). This would in turn support autotrophic carbon fixation in both partners and sustains N and S biogeochemical cycling under dynamic or unfavorable conditions (e.g., limited H_2S bioavailability; Figure C.4). Such mutualism would be highly dependent on either (a) migration along the eco-thermodynamic gradient or (b) seasonal/temporal changes such as renewal or upwelling events.

Marinimicrobia clades HMTAb91-A/B are prevalent in methanogenic locations at the base of the electron tower. HMTAb91-A/B did not harbor genes for aerobic respiration and had an incomplete TCA cycle. HMTAb91-A encoded the Embden-Meyerhof-Parnas pathway (Table C.4) and both HMTAb91-A/B encoded energy-conserving H^+ respiration through electron-confurcating hydrogenases (reverse electron transport), the energy-conserving (Rnf complex) and putative syntrophic amino-acid metabolism through the ion-translocating ferredoxin:NADH oxidoreductase (ifoAB) (Figure 4.3b) [217]. Within the methanogenic reactor where it was initially described, HMTAb91-A is postulated to accomplish thermodynamically unfavorable amino-acid degradation supporting methanogenesis [217]. HMTAb91-A/B clades appear restricted to methanogenic ecosystems as no metagenomic or metatranscriptomic sequences were recruited from non-methanogenic locations.

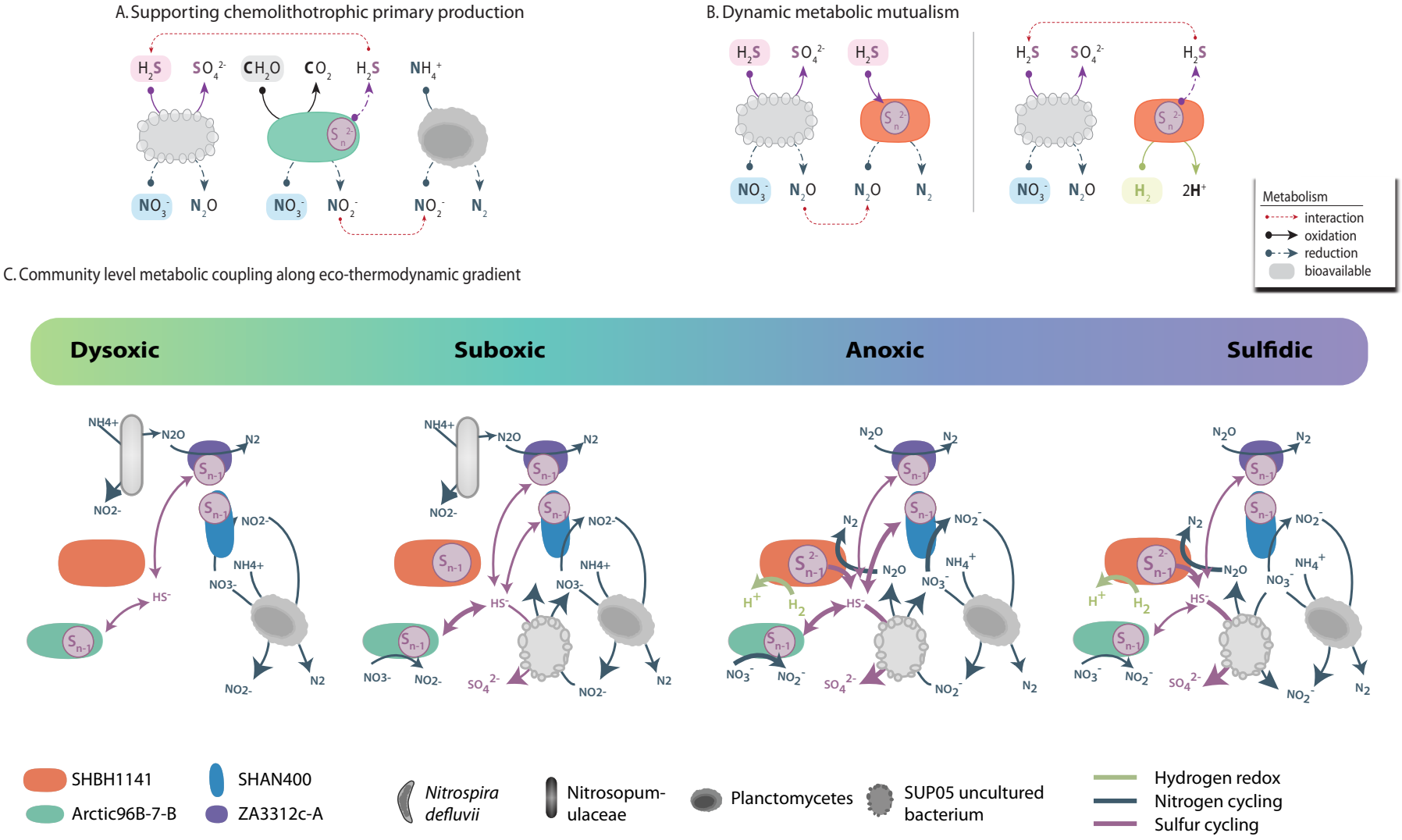


Figure 4.5: Proposed co-metabolic model along eco-thermodynamic gradients in Saanich Inlet. **(A)** Proposed coupling where ARCTIC96B-1 clades support the chemolithotrophic primary production of SUP05 and Planctomycetes. **(B)** Proposed dynamic metabolic mutualism between SUP05 and SHBH1141. **(C)** Overall proposed model for Marinimicrobia co-metabolic activity with other dominant microbial groups in Saanich Inlet along eco-thermodynamic gradients. Interactions based on expression data for sulfur (pink), nitrogen (blue) and hydrogen (green) for dominant Marinimicrobia lineages in Saanich Inlet as well as metabolic partners *Nitrosopumulaceae* sp.,*Planctomycetes*, and SUP05 uncultured bacterium.

4.3 Discussion

Co-metabolic functions encoded and expressed within globally distributed Marinimicrobia clades would fill several hitherto unassigned niches in the nitrogen and sulfur cycles and support recent modeling efforts integrating biogeochemical and multi-omic sequence information in the Saanich Inlet water column [113–115]. The N_2O reductase expressed on a global basis by ZA3312c-A and SHBH1141 clades has the potential to act as a biological filter for N_2O produced by the ubiquitous marine processes of ammonia oxidation (e.g., Thaumarchaeota) [203] and partial denitrification (e.g., SUP05) [34, 79]. In contrast, nitrate reduction to NO_2^- by other Marinimicrobia clades (i.e., Arctic96B-7-A and SHAN400) has potential to provide NO_2^- to anaerobic ammonium-oxidizing (*Planctomycetes*) and nitrite-oxidizing (*Nitrospina*) bacteria in dysoxic, suboxic, and anoxic waters. The polysulfide reductase expressed by multiple Marinimicrobia clades (e.g., Arctic96B-7, SHAN400, and SHBH1141) has potential to provide an energy storage mechanism via accumulation of polyS that can be reduced or oxidized under changing water column redox conditions and support both cooperative and dynamic interactions including cryptic sulfur and cycling and dark carbon fixation [78].

The application of eco-thermodynamics principles to microbial ecology provides perspective on how thermodynamic constraints serve to shape microbial community structure and the nature of co-metabolic interactions along energy gradients. Indeed, phylo- genetic branching patterns often coincided with energy yields of redox pairs for identified clade energy metabolism, with deeper branching clades near the base of the electron tower where lower energy yields would increase potential for metabolic coupling. Additionally, many Marinimicrobia clades encoded enzyme systems tied to both nitrogen- and sulfur-cycling, suggesting extensive specialization for metabolic cooperation bridging within and between biogeochemical cycles. Such dependencies likely confound isolation efforts within the phylum and point to an ancestral state primed for co-existence. The extent to which this reflects the diversification of other phyla, particularly MDM across the Tree of Life is an interesting area of research with implications for understanding and directing the evolution of metabolic networks driving Earths biogeochemical cycles.

4.4 Methods

4.4.1 SAG collection, sequencing, assembly, and decontamination

SAGs from Gulf of Maine, HOT station ALOHA, South Atlantic Gyre, the Terephthalate degrading bioreactor and Etoliko Lagoon Sediment were included in Rinke *et al.* [19], and collection, assembly and decontamination follows accordingly. See Supplementary Data 1 for details on SAG genomics. SAGs from Northeast subarctic Pacific (NESAP) and Saanich Inlet followed the following protocol. Replicate 1 mL aliquots of sea water collected for single-cell analyses were cryopreserved with 6% glycine betaine (Sigma-Aldrich), frozen on dry ice and stored at -80°C. Single-cell sorting, whole-genome amplification, real-time PCR screens, and PCR product sequence analyses were performed at the Bigelow Laboratory for Ocean Sciences Single Cell Genomics Center (`www.bigelow.org/scgc`), as described by Stepanauskas and Sieracki [105]. SAGs from the NESAP were generated at the DOE Joint Genome Institute (JGI) using the Illumina platform as described in Rinke *et al.* [19]. SAGs from Saanich Inlet were sequenced at the Genome Sciences Centre, Vancouver BC, Canada, as described in Roux *et al.* [224]. All SAGs were assembled at JGI as described in Rinke *et al.* [19, 224].

The following steps were performed for SAG assembly: (1) filtered Illumina reads were assembled using Velvet version 1.1.0437 using the VelvetOptimiser script (version 2.1.7) with parameters: (–v –s 51 –e 71 –i 4 –t 1 –o -ins_length 250 -min_contig_lgth 500) 2) wgsim (-e 0 -1 100 -2 100 -r 0 -R 0 -X 0) 3) Allpaths-LG (prepareAllpathsParams: PHRED_64 = 1 PLOIDY = 1 FRAG_COVERAGE = 125 JUMP_COVERAGE = 25 LONG_JUMP_COV = 50, runAllpathsParams: THREADS = 8 RUN = std_pairs TARGETS = standard VAPI_WARN_ONLY = True OVERWRITE = True). SAG prediction analysis and functional annotation was performed within the Integrated Microbial Genomes (IMG) platform [225] (`http://img.jgi.doe.gov`) developed by the Joint Genome Institute, Walnut Creek, CA, USA.

4.4.2 Phylogenomic analysis of SAGs

The PhyloPhlAn pipeline was used to determine relationships among Marinimicrobia SAGs [218] (Figure C.3) as well as the phylogenetic placement of Marinimicrobia within the bacterial

domain (Figure C.2) . In both cases, fasta files for the 25 SAGs and related genomes were passed to PhyloPhlAn and resulting trees were visualized and drawn using GraPhlAn. The 25 Marinimicrobia SAGs and related genomes were inserted into the already built PhyloPhlAn microbial Tree of Life containing 3737 genomes using the insert functionality, and a de novo phylogenetic tree was created using the user functionality based solely on the 25 Marinimicrobia SAG and related genome fasta files. Default parameters were used in each case with the exception of a custom annotation file used in GraPhlAn to colour the leaves based on phylum in the microbial Tree of Life, and subgroup in the de novo phylogenetic tree.

4.4.3 Metagenome fragment recruitment

The proportion of Marinimicrobia represented in the 594 globally distributed metagenomes (Figure C.3) was determined by SAG nucleotide sequence alignment to individual metagenomes using FAST [185]. Parameters of 70% nucleotide identity cutoff over 70% of the contig length (or 454-read, where applicable) were employed to encompass the Marinimicrobia phylum [226]. The small subunit ribosomal RNA (SSU rRNA) gene was removed from SAG sequences before alignment searches to prevent cross-recruitment to non-Marinimicrobia sequences. The total length of contigs passing the cutoff for a given metagenome was summed and divided by the total contig length for that metagenome to calculate percentage of Marinimicrobia. Where data on O_2 concentration was available, for Saanich Inlet, NESAP, ETSP [134], and Peruvian upwelling [199], O_2 status of the sample was used as indicated. Data on O_2 concentration were unavailable for Marine-Misc. and terrestrial samples.

Biogeography of Marinimicrobia SAG-affiliated clades was similarly determined using alignment parameters of 95% identity cutoff and >200 base pairs (bp) alignment length to ensure only contigs with high sequence similarity while maintaining clade resolution. Metagenomic contigs mapping to more than one Marinimicrobia clade were assigned to the clade with greatest percent identity and in the event of a tie were assigned to the clade with the greatest alignment length. Overall abundance was calculated for each metagenome by summing the total lengths of all contigs with hits to a given Marinimicrobia clade divided by the total size of the SAG and the total size of the assembled metagenome in base pairs. Results by metagenome and clade were then

summed in Figure 4.2 and itemized in Table C.2. Global relative abundance of Marinimicrobia clades shown in Figure C.3 was calculated similarly by summing the total lengths of all contigs with hits to a given Marinimicrobia clade divided by the total size of the SAG and the total size of the assembled metagenome in base pairs and then summing for all hits to a given clade.

4.4.4 Saanich Inlet and NESAP metagenomes and metatranscriptomes

Saanich Inlet metagenomes and metatranscriptomes were collected, sequenced, and assembled as described in Hawley *et al.* [114] and cognate chemical and physical measurements can be found in and Torres-Beltrán *et al.* [115] . Briefly, Saanich Inlet samples for metagenomic and metatranscriptomic sequencing were collected by Niskin or Go-Flow on line with CTD. Samples for metatranscriptomics, 2 L, were filtered by peristaltic pump with in-line 2.7 µm prefilter onto a sterivex filter with 1.8 mL RNALater added and frozen on dry ice within 20 min of bottle on-deck. Metagenomic samples, 20 L, were filtered within 8 h of collection by peristaltic pump with in-line 2.7 µm prefilter onto a sterivex filter with 1.8 µL lysis buffer added and frozen at -80°C. Metagenomic and metatranscriptomic samples were processed, sequenced, and assembled according to Hawley *et al.*[114] at the JGI using the Illumina HiSeq platform.

Sampling in the NESAP was conducted via multiple hydrocasts using a Conductivity, Temperature, Depth (CTD) rosette water sampler aboard the CCGS John P. Tully during three Line P cruises: 2009-09 [June 2009, major stations P4 (48°39.0 N, 126°4.0 W, 7 June), P12 (48°58.2 N, 130°40.0 W, 9 June), and P26 (50°N, 145°W, 14 June), 2009-10 [August 2009, major stations P4 (21 August), P12 (23 August) and P26 (27 August)], and 2010-01 [February 2010, major stations P4 (4 February) and P12 (11 February)]. At these stations, large volume (20 L) samples for DNA isolation were collected from the surface (10 m), while 120 L samples were taken from three depths spanning the OMZ core and upper and deep oxyclines (500, 1000, 1300 m at station P4; 500, 1000, 2000 m at station P12). Sequencing and assembly was carried out as described above for Saanich Inlet and accession numbers are available in C.1.

Construction and validation of population genome bins. Marinimicrobia population genome bins were constructed by identifying metagenomic contigs from Saanich Inlet, and NESAP metagenomes mapping to specific SAG(s) using a supervised binning method based in part on

methodologies developed by Dodsworth *et al.* [181] in the construction of OP9 population genome bins. Initially, determination of membership of individual SAGs to SAG-clusters making up a given phylogenetic clade was conducted. SAG tetranucleotide frequencies were then calculated and converted to z-scores with TETRA (`http://www.megx.net/tetra`) [227, 228]. Z-scores were reduced to three dimensions with principal component analysis (PCA) using PRIMER v6.1.13 [229] and hierarchical cluster analysis of the z-score PCA with Euclidian distance (also performed in PRIMER) was carried out to generate SAG-clusters. These SAG-clusters reflected phylogenetic placement of the SAGs by SSU rRNA gene analysis. For construction of population genome bins, metagenomic contigs from NESAP and SI data sets were aligned to SAG contigs with >95% nucleotide identity using BLAST [209] and a minimum of 5 kilobase pairs alignment length, Tetranucleotide frequencies of all metagenomic contigs passing this identity and length threshold were calculated and converted to z-scores. SAG–upervised binning as described in Dodsworth *et al.* using linear discriminant analysis was carried out using all z-scores with the SAG-bins as training data to classify the metagenomic conigs as making up a given population-genome bin.

Individual SAGs and population genome bins were analyzed for completeness and strain heterogeneity using CheckM v1.0.5 [230]. Specifically, the lineage_wf workflow was used with default parameters. The lineage_wf workflow includes determination of the probable phylogenetic lineage based on detected marker genes. The determined lineage then dictates the sets of marker genes that is most relevant for estimating a given genomes completeness and other statistics. The strain heterogeneity metric is highly informative for population genome bins as it is essentially the average amino-acid identity for pairwise comparisons of the (lineage appropriate) redundant single-copy marker genes within a population genome bin (Table 4.2). For population genome bins the higher the strain heterogeneity value, the more similar the amino acid identity of the redundant maker genes indicating the sequences in the bin originate from a closely related, if not identical, phylogenetic source.

4.4.5 Marinimicrobia genome streamlining

Gene-coding bases and COG-based gene redundancy shown in Figure C.1 were calculated using cluster of orthologous group (COG)-based genome redundancy as described in Rinke *et al.* [19].

Each genes COG category was predicted through the JGI IMG pipeline. COG redundancy was calculated by averaging the occurrence of each COG in the genome. The percentage of gene-coding bases was calculated by dividing the number of bases contributing to protein and RNA-coding genes by the total genome size. For SAGs, the length of the assembled genome was used rather than the estimated genome size.

4.4.6 Annotation and identification of metabolic genes of interest

Genes of interest were identified in the SAGs and in IMG/M (`https://img.jgi.doe.gov/cgi-bin/m/main.cgi`) [231] for the metagenomic contigs which made up the population genome bins. Contigs making up Marinimicrobia population genome bins were run through MetaPathways 2.5 [94, 182] to annotate open reading frames (ORFs) and reconstruct metabolic pathways. As the population genome bins were constructed from multiple metagenomes they contained redundant sequence information, BLASTp [209] (amino-acid identity cutoff >75%) was used to identify all copies of a given gene of interest in each population genome bin, which was then used in gene model validation and expression mapping.

4.4.7 Gene expression mapping

Metatranscriptomes from three time points in Saanich Inlet time series [114] were used to investigate changes in gene expression along water column redox gradients over time for selected ORFs involved in energy metabolism and electron shuttling. Quality controlled reads from metatranscriptomes were mapped to identified ORFs of interest using bwa -mem [190] and reads per kilobase per million mapped (RPKM) per ORF was calculated using RPKM calculation in MetaPathways 2.5 [191]. For each population genome bin RPKM values for a given sample were summed for ORFs with the same functional annotation to yield an RPKM for a given functional gene. For other taxonomic groups in Saanich Inlet shown in Supplementary Figure C.6B, genes were identified by sequence alignment searches of Saanich Inlet metatranscriptomes (bioSample indicated above) assembled and conceptually translated using BLASTp against selected nitrogen and sulfur cycling genes from Hawley *et al.* [79] and RPKM values calculated as described above.

4.4.8 Global distribution and expression of nosZ

Further analysis was carried out to determine the global distribution of Marinimicrobia *nosZ* in 594 metagenomes. The *nosZ* nucleotide sequences from SHBH1141 and ZA3312c, which exhibited a 65% nucleotide identity to each other by BLAST, were clustered at 95% identity using the USEARCH cluster fast algorithm [232], resulting in three clusters, two SHBH1141 and one ZA3312c. Nucleotide sequence alignment was carried out using FAST [185], with parameters of >80% nucleotide identity and >60 bp alignment length against 594 metagenomes. For Saanich Inlet and NESAP data sets, abundance of *nosZ* in a given metagenome or metatranscriptome was determined by summing the RPKM value for ORF hits to either SHBH1141 or ZA3312c for a given metagenome or metatranscriptome. For 454 sequenced metagenomes and metatranscriptomes [134, 199], the number of reads which hit to either SHBH1141 or ZA3312c were summed for a given metagenome. For the TARA Oceans data set [233], the number of genes identified in an assembled metagenome was summed. Metatranscriptomic data for Tara was unavailable at this time.

Chapter 5

A niche for NosZ? [4]

5.1 Introduction

Nitrous oxide (N_2O) is a potent green house gas with 298 times the atmospheric heat-trapping activity of carbon dioxide and is also currently the most dominant ozone depleting substance [234, 235]. Marine ecosystems are estimated to account for one third of total global N_2O production [234], with lower oxygen (O_2) concentrations correlating with increased N_2O production [113]. Oxygen minimum zones (OMZs) and coastal upwelling regions are seen as the dominant marine sources of N_2O, where it is produced by microorganisms under low O_2 or anoxic conditions [236, 237]. However, the organisms responsible for N_2O production and consumption and their dynamics within the marine environment have yet to be constrained. As concentrations of O_2 in the Global Ocean are expected to decrease significantly in the coming decades [11], identification of marine N_2O sources and sinks are a pressing area of interest.

The only known biological sink for N_2O within both marine and terrestrial environments is the nitrous oxide reductase enzyme NosZ (encoded by the *nosZ* gene). Nitrous oxide reductase is typically found as the final step in the denitrification pathway which reduces nitrate (NO_3^-) to nitrite (NO_2^-), to nitric oxide (NO), then N_2O and finally to nitrogen gas (N_2), a process that occurs largely under O_2-depleted to anoxic conditions (i.e. $<20\,\mu M$). However, approximately 40% of organisms carrying genes for NO_3^- reduction do not carry the *nosZ* gene [238] resulting in incomplete denitrification and the production of N_2O [239], as in the case of the prevalent OMZ Gammaproteobacteria, SUP05 [33, 34]. Furthermore, the high sensitivity of NosZ to O_2, causing inhibition of N_2O reduction, can result in N_2O production from complete denitrifiers

[4] Portions of this chapter, namely N_2O concentrations in Saanich Inlet, have been previously published in Limnology and Oceanography as *A Multi-year time-series of N_2O dynamics in a seasonally anoxic fjord: Saanich Inlet, British Columbia* in 2017 by Capelle, D. W. and Hawley, A. K. and Hallam, S. J. and Tortell, P. D..

under microaerophilic conditions [240]. Nitrous oxide is also produced by several other processes including ammonia oxidation, nitrifier denitrification (where nitric oxide reductase reduces NO formed upon NO_2^- oxidation) and commomox [241]. With these abundant and diverse sources of environmental N_2O, recent studies including pure cultures [240, 242, 243], enrichments [244], molecular [245], genomic and metagenomic surveys [59, 246] and multi-omic models [132] suggest multiple niches for non-denitrifying N_2O-reducing organisms within O_2-depleted environments. While organisms capable of surviving on N_2O as the sole electron acceptor have been known for decades [243], two clades of *nosZ* have only been recently observed [59] and much remains to be understood regarding the dynamics of different N_2O-reducing organisms within the Global Ocean.

Recent work has established two different types of *nosZ* [59]. Type I are typically found in Alpha- Beta- and Gammaproteobacteria, 83% of which carry some genes for the complete denitrification pathway [238]. Type II are found with a broad taxonomic distribution of organisms, 51% of which do not carrying any other genes for the denitrification pathway [59, 238, 246, 247]. While Type I was thought to be the dominant NosZ, work in terrestrial systems has found Type II to be up to an order of magnitude more abundant [238]. Differences in the structure of the *nos* gene cluster (NGC) indicate possible differences in energy metabolism between the two types. Type I *nosZ* NGC, with a twin-arginine signal sequence, typically occure in a *nosRZDFYL* cluster, encoding proteins responsible for passing electrons from quinol to NosZ via cytochrome bc_1 and cytochrome *c* complex and NosR, a periplasmic flavin mononucleotide binding protein [115]. Type II *nosZ* NGC, with a sec signal sequence, typically occurs in a *nosZBDFYL* cluster, though greater variation exists than in Type I NGC and additional *nosGH* may also be present [115]. The Type II NGC found in Epsilonproteobacteria, for example, encodes proteins proposed to pass electrons from menaquinone/menaquinol to NosZ via cytochrome bc_1 complexes, NosGH or NosB proteins. Notably, NosGH and NosB may be capable of generating a proton motive force by acting as a menaquinol-reactive proton pump [115], supplying additional energy for Type II N_2O-reducing organisms.

Differences in the abundance and distribution of Type I and II *nosZ* as well as complete vs incomplete denitrification have been the focus of recent studies, with results pointing to several

possible environmental factors that may be mediating these dynamics. Ample carbon sources seem to support complete denitrifiers and organisms with Type I *nosZ*. With no limits on carbon source, complete denitrifiers grow faster than organisms which only respire N_2O [244]. However, with ample carbon supply and sufficient source of electron acceptors, denitrifying organisms may carry out incomplete denitrification, reducing NO_3^- and NO_2^- but halting at NO reduction and producing N_2O. Indeed, NOx^- limitation is seen to favour complete denitrifiers, as the Type I *nosZ* generally has a higher μ_{max}/K_s and affinity for N_2O to maximise energy yield from dwindling NOx^- and making them resistant to scavenging by Type II [242, 244, 245]. Finally, O_2 concentrations also appear to shape the abundance and distribution of Type I vs Type II *nosZ* as O_2 sensitivity may inhibit N_2O reduction. However, recent research indicates some Type II appear to recover more quickly from O_2 exposure and may operate under microaerophilic conditions [240], presenting a unique niche for *nosZ*. Indeed, recent studies have found evidence of association of organisms carrying Type II *nosZ* with incomplete denitrifiers within agricultural soils and marine sediments [245, 248]. Similar association has also been proposed in Chapter 4 between incomplete denitrifier SUP05 and N_2O-reducer Marinimicrobia SHBH1141 [249]. Much of the research on the Type I and Type II *nosZ* abundance and distribution has been carried out in agricultural soils and waster water treatment with the aim to mitigate N_2O release, however, comprehensive research in marine systems has been lacking.

To date, identification of N_2O-reducing organisms within marine environments has been limited to complete denitrifiers: *Sulfuramonas gotlandica*, isolated from the Baltic Sea [37] and SUP05 group member ^{U}*Thioglobus perditus* (U to indicate uncultured), metagenome assembled genome from the Peruvian upwelling [36]; and non-denitrifying N_2O-reducers: Marinimicrobia SHBH1141, single cell amplified genome from Saanich Inlet, and Marinimicrobia ZA3312c, metagenomic population genome bin from Saanich Inlet and Gulf of Maine [249]. In order to better phylogenetically constrain *nosZ* genes within OMZs and the Global Ocean and survey the relative abundance and distribution of different *nosZ* clades, I utilize single cell genomes from Saanich Inlet, a seasonally anoxic fjord and model OMZ system on the coast of British Columbia. I define 13 NosZ clades encompassing both Type I and TypeII and phylogenetically anchor additional six *nosZ* genes from OMZs. I further chart the distribution of these clades in Global

Ocean metagenomes and explore metatranscriptomes in the OMZs of Peru, Eastern Tropical South Pacific and Saanich Inlet (metaproteomes on for Saanich Inlet). I explore the seasonal dynamics of *nosZ* clades in Saanich Inlet over time and with accompanying rate measurements of nitrogen loss processes denitrification and anammox. Finally, I explore abundance and expression of *nosZ* clades along gradients of O_2, NO_3^- and hydrogen sulfide (H_2S) exploring possible niches for respective clades.

5.2 Results

5.2.1 Inventory of single-cell amplified genomes

In order to better identify the taxonomic origins of *nosZ* genes within the Gobal Ocean, samples for single cell amplified genomes (SAGs) [105] from Saanich Inlet were collected, sequenced and *nosZ* genes identified. Samples for SAGs were collected in August 2011 from station S3 in Saanich Inlet (48°35.500 N, 123°30.300W) at three depths to capture different chemical conditions: dysoxic (20-90 µM O_2)at 100 m, suboxic (2-20 µM O_2) at 120 m and sulfidic at 185 m. From the collected samples, three 96 well plates were sorted for each depth, resulting in 864 sorted wells, of those wells, a total of 645 had a small subunit ribosonal RNA gene (SSUrRNA) that was able to be amplified and assembled as described in Stepanauskas and Sieracki, 2007 and Swan *et al.*, 2013 [77, 105], resulting in SSU rRNA genes from a range of Bacteria and a few Archaea. From that a total of 371 SAGs were chosen for sequencing based on taxonomy and aplification efficiency (Figure 5.1). Annotation of sequenced SAGs by Metapathways [182] revealed 51 *nosZ* gene sequences in six different taxonomic groups (21 Arcobacteraceae, 12 Bacteroidales VC21, three Ectothiorhodospirales, two Marinimicrobia SHBH1141, two SAR324 and 10 SUP05_1a (See Figure D.1 for SUP05 phylogenetic tree). On average, SAGs containing *nosZ* were 62% complete with 2.98% contamination (see TableD.1 for CheckM statistics on SAG completion and contamination [230]).

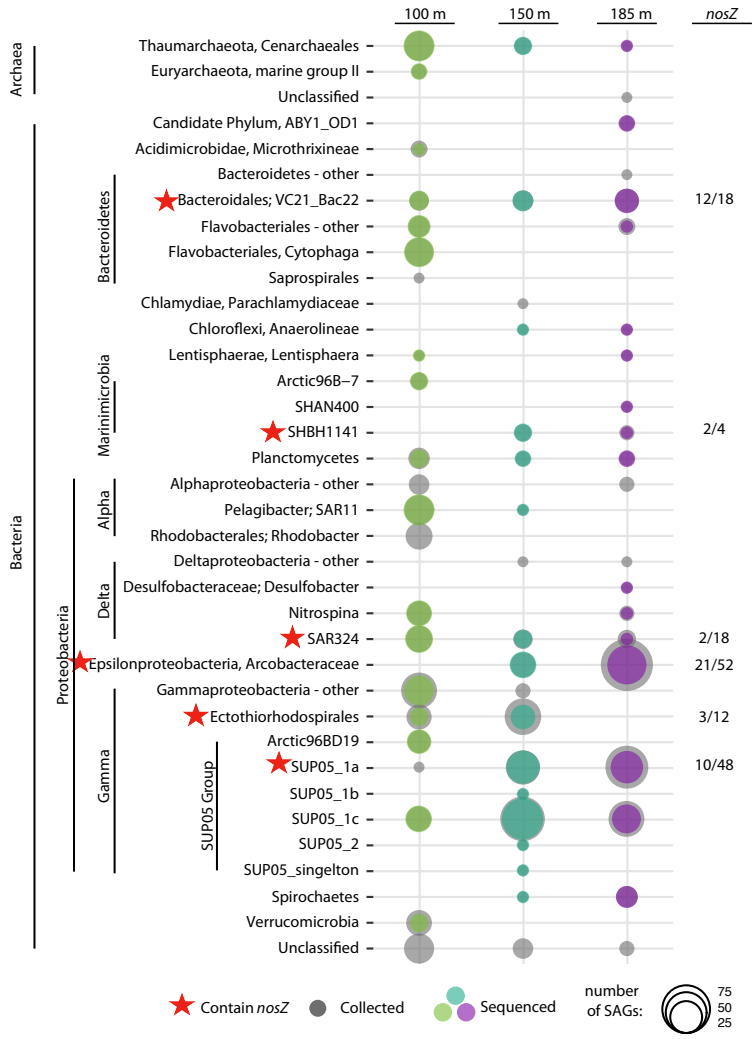


Figure 5.1: Saanich Inlet SAG inventory and taxonomy. Taxonomy of Saanich Inlet SAGs collected (grey) and sequenced (coloured) from 100 m (green), 150 m (teal) and 185 m (purple) on August 10th 2011. Red star indicates SAGs containing *nosZ* gene and number of SAGs with *nosZ* out of the total number of SAGs for that taxa are show on the right (*nosZ*/total). Only SAGs with an amplified and assembled small subunit ribosomal RNA gene are shown.

5.2.2 Clustering & genomic neighbourhood analysis

In order to confirm that the SAG *nosZ* sequences were consistent within a given taxonomy, the SAG *nosZ* nucleotide sequences as well as *nosZ* previously identified as Marinimicrobia ZA3312c [249] were clustered at 95% ID. Sequences from the same taxonomic group clustered together with the exception of the Arcobacteracea that had two additional singletons (which clustered together at 90% identity). Within each taxonomy the genomic area surrounding *nosZ* (as compared among SAG contigs) was consistent within the taxonomy and differed across taxonomic groups (Figure 5.2). The Arcobacteracea, Ectothiorhodospirales and SUP05_1a all carried additional genes in the denitrification cycle on the same contig, often interspersed within the *nosZ* gene cluster. Notably, *nosR* was found in the Type I *nosZ* NGC but not the Type II, though many genes were annotated as hypothetical, better annotation may reveal more consistencies or differences across NGCs of different clades.

5.2.3 NosZ phylogeny and abundance

To phylogenetically place the *nosZ* genes from Saanich Inlet SAGs within the context of isolated N_2O-reducing organisms, NosZ protein sequences from Sanford *et al.* 2012, Saanich Inlet SAGs and Marinimicrobia ZA3312c, were clustered at 95% identity and a maximum-likelihood tree was constructed (Figure 5.3). The resulting tree retained previously observed topology of Type I and Type II NosZ. Of the Saanich Inlet SAGs only SUP05_1a NosZ clustered in the Type I portion of the tree. Thirteen clades were resolved as indicated by clustering patterns. Saanich Inlet SAG *nosZ* sequences clustered into five pre-existing clades: Bacteroidales VC21 and Marinimicrobia ZA3312c NosZ sequences clustered closely together in Clade 5 consisting primarily of other members of the Fibrobacteres-Chlorobi-Bacteroidetes superphylum (FCB); Marinimicrobia SHBH1141 NosZ sequences clustered in Clade 6 also with members of the FCB and Aquificae; SAR324 and Ectothiorhodospirales NosZ sequences clustered in Clade 9 with several Proteobacterial lineages, as did Arcobacteracea NosZ in Clade 10. The SUP05_1a NosZ sequence clustered with Alphaproteobacteria in Clade 13.

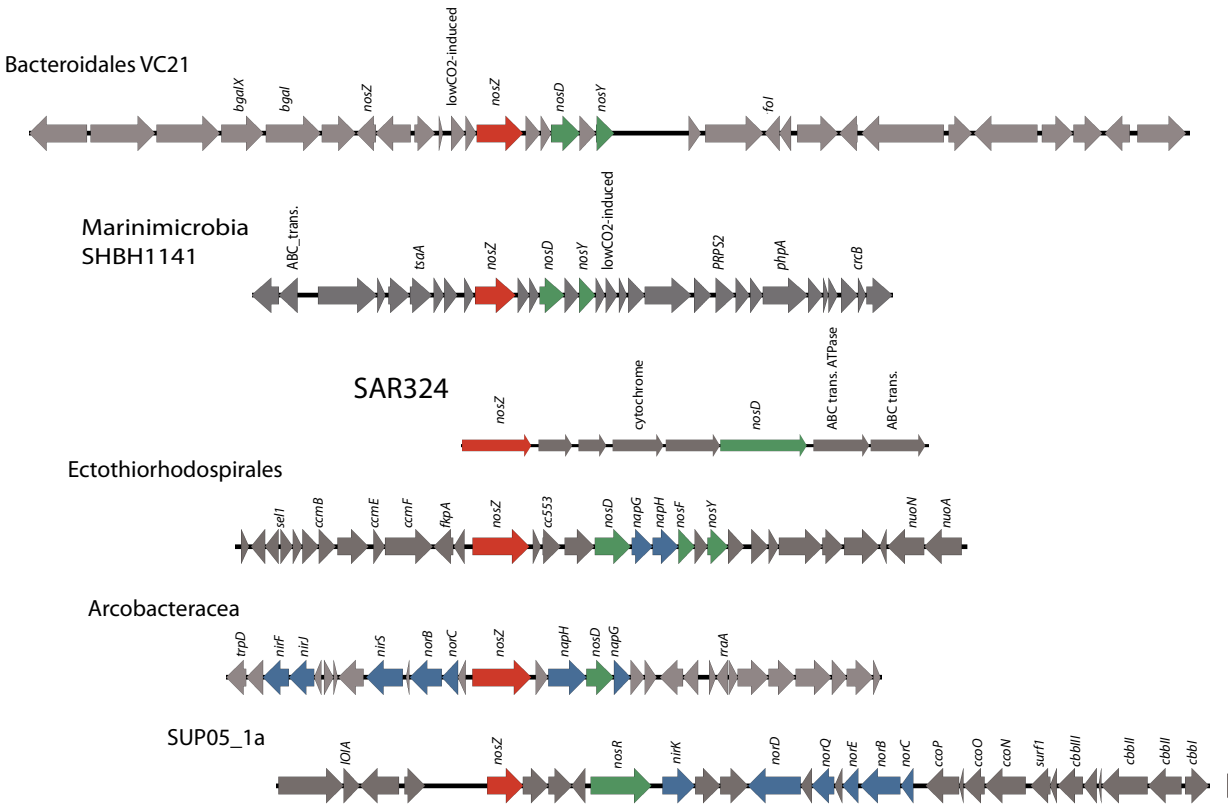


Figure 5.2: Genome neighbourhood analysis for *nosZ* SAGs. Genomic neighbourhood for Saanich Inlet SAG contigs containing the *nosZ* gene (red), other nos gene cluster genes (green) and other genes in the denitrification pathway (blue).

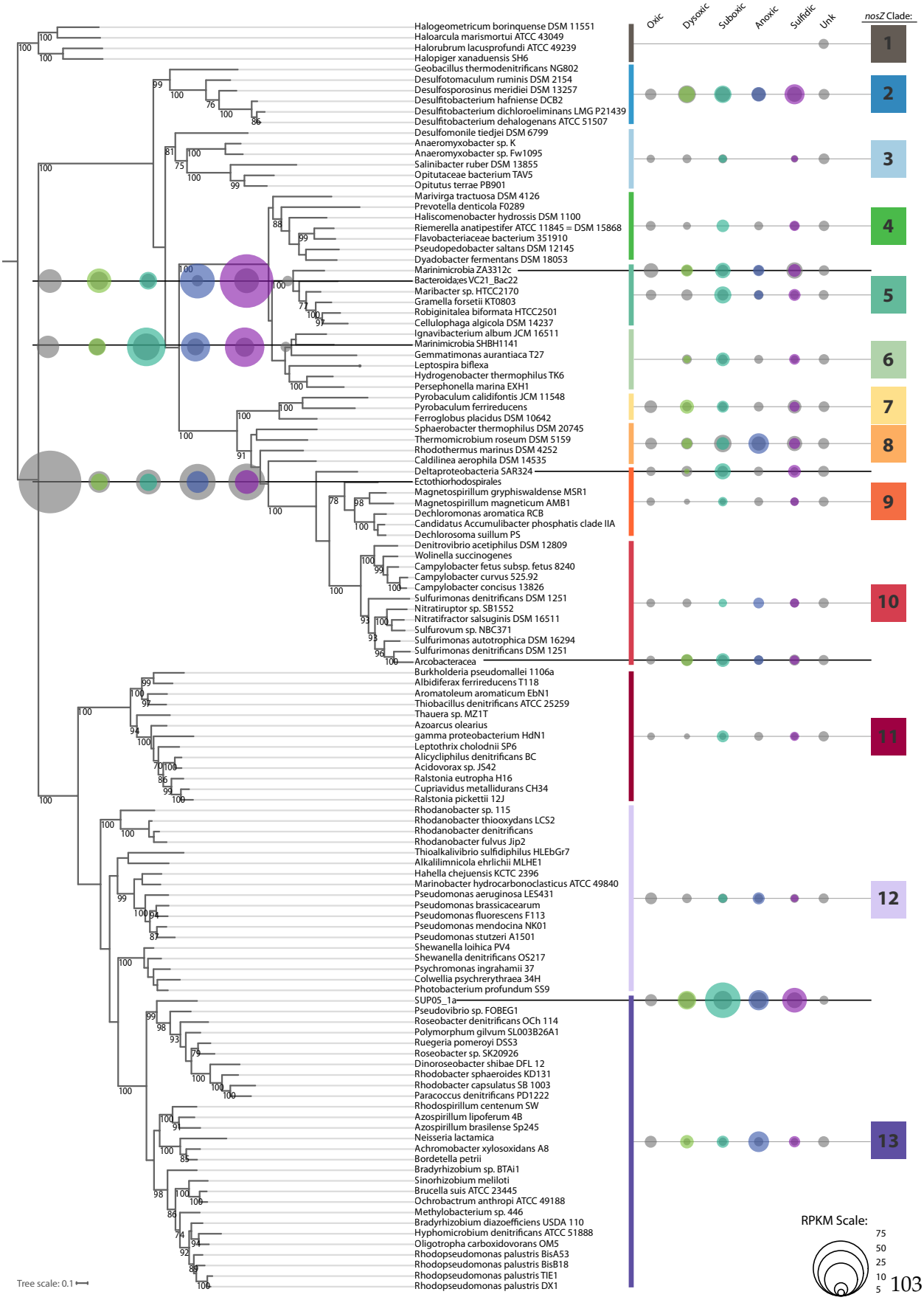

Oxic
Dysoxic
Suboxic
Anoxic
Sulfidic
Unk
nosZ Clade:
Halogeometricum borinquense DSM 11551
Haloarcula marismortui ATCC 43049
Halorubrum lacusprofundi ATCC 49239
Halopiger xanaduensis SH6
Geobacillus thermodenitrificans NG802
Desulfotomaculum ruminis DSM 2154
Desulfosporosinus meridiei DSM 13257
Desulfitobacterium hafniense DCB2
Desulfitobacterium dichloroeliminans LMG P21439
Desulfitobacterium dehalogenans ATCC 51507
Desulfomonile tiedjei DSM 6799
Anaeromyxobacter sp. K
Anaeromyxobacter sp. Fw1095
Salinibacter ruber DSM 13855
Opitutaceae bacterium TAV5
Opitutus terrae PB901
Marivirga tractuosa DSM 4126
Prevotella denticola F0289
Haliscomenobacter hydrossis DSM 1100
Riemerella anatipestifer ATCC 11845 = DSM 15868
Flavobacteriaceae bacterium 351910
Pseudopedobacter saltans DSM 12145
Dyadobacter fermentans DSM 18053
Marinimicrobia ZA3312c
Bacteroida;es VC21_Bac22
Maribacter sp. HTCC2170
Gramella forsetii KT0803
Robiginitalea biformata HTCC2501
Cellulophaga algicola DSM 14237
Ignavibacterium album JCM 16511
Marinimicrobia SHBH1141
Gemmatimonas aurantiaca T27
Leptospira biflexa
Hydrogenobacter thermophilus TK6
Persephonella marina EXH1
Pyrobaculum calidifontis JCM 11548
Pyrobaculum ferrireducens
Ferroglobus placidus DSM 10642
Sphaerobacter thermophilus DSM 20745
Thermomicrobium roseum DSM 5159
Rhodothermus marinus DSM 4252
Caldilinea aerophila DSM 14535
Deltaproteobacteria SAR324
Ectothiorhodospirales
Magnetospirillum gryphiswaldense MSR1
Magnetospirillum magneticum AMB1
Dechloromonas aromatica RCB
Candidatus Accumulibacter phosphatis clade IIA
Dechlorosoma suillum PS
Denitrovibrio acetiphilus DSM 12809
Wolinella succinogenes
Campylobacter fetus subsp. fetus 8240
Campylobacter curvus 525.92
Campylobacter concisus 13826
Sulfurimonas denitrificans DSM 1251
Nitratiruptor sp. SB1552
Nitratifractor salsuginis DSM 16511
Sulfurovum sp. NBC371
Sulfurimonas autotrophica DSM 16294
Sulfurimonas denitrificans DSM 1251
Arcobacteracea
Burkholderia pseudomallei 1106a
Albidiferax ferrireducens T118
Aromatoleum aromaticum EbN1
Thiobacillus denitrificans ATCC 25259
Thauera sp. MZ1T
Azoarcus olearius
gamma proteobacterium HdN1
Leptothrix cholodnii SP6
Alicycliphilus denitrificans BC
Acidovorax sp. JS42
Ralstonia eutropha H16
Cupriavidus metallidurans CH34
Ralstonia pickettii 12J
Rhodanobacter sp. 115
Rhodanobacter thiooxydans LCS2
Rhodanobacter denitrificans
Rhodanobacter fulvus Jip2
Thioalkalivibrio sulfidiphilus HLEbGr7
Alkalilimnicola ehrlichii MLHE1
Hahella chejuensis KCTC 2396
Marinobacter hydrocarbonoclasticus ATCC 49840
Pseudomonas aeruginosa LES431
Pseudomonas brassicacearum
Pseudomonas fluorescens F113
Pseudomonas mendocina NK01
Pseudomonas stutzeri A1501
Shewanella loihica PV4
Shewanella denitrificans OS217
Psychromonas ingrahamii 37
Colwellia psychrerythraea 34H
Photobacterium profundum SS9
SUP05_1a
Pseudovibrio sp. FOBEG1
Roseobacter denitrificans OCh 114
Polymorphum gilvum SL003B26A1
Ruegeria pomeroyi DSS3
Roseobacter sp. SK20926
Dinoroseobacter shibae DFL 12
Rhodobacter sphaeroides KD131
Rhodobacter capsulatus SB 1003
Paracoccus denitrificans PD1222
Rhodospirillum centenum SW
Azospirillum lipoferum 4B
Azospirillum brasilense Sp245
Neisseria lactamica
Achromobacter xylosoxidans A8
Bordetella petrii
Bradyrhizobium sp. BTAi1
Sinorhizobium meliloti
Brucella suis ATCC 23445
Ochrobactrum anthropi ATCC 49188
Methylobacterium sp. 446
Bradyrhizobium diazoefficiens USDA 110
Hyphomicrobium denitrificans ATCC 51888
Oligotropha carboxidovorans OM5
Rhodopseudomonas palustris BisA53
Rhodopseudomonas palustris BisB18
Rhodopseudomonas palustris TIE1
Rhodopseudomonas palustris DX1
1
2
3
4
5
6
7
8
9
10
11
12
13
RPKM Scale:
75
50
25
10
5
Tree scale: 0.1

Figure 5.3: NosZ phylogenetic tree with global abundance and expression (previous page). Phylogenetic tree of *nosZ* gene for cultured isolates [59] and SAGs from Saanich Inlet with clade labels (far right). Abundance (by RPKM) of clades (or specific SAG sequences within a clade where connected by a thick black line) in globally found metagenomes (grey) and metatranscriptomes (coloured) for indicated chemical conditions: dysoxic (20-90 µM, green), suboxic (2-20 µM, teal), anoxic (0-2 µM, blue), and sulfidic (purple) shown to the right. Bubble size represents the RPKM averaged for the number of samples from each chemical condition. Sequences not mapping to leaves, but to internal nodes made up 2.2% and 0.6% of metagenomic and metatranscriptomic sequences respectively.

With the inclusion of OMZ origin NosZ sequences in the phylogenetic tree with isolated organisms, I explored the abundance and expression of *nosZ* clades in Saanich Inlet and the Global Ocean under different O_2 regimes: oxic (>90 µM), dysoxic (20-90 µM), suboxic (2-20 µM), anoxic (0-2 µM) and sulfidic. Metagenomic datasets were searched by protein sequence similarity to NosZ SAG sequences using LAST+ [185]. Sequences were identified in Saanich Inlet [114], other OMZs (Peru [199] and Eastern tropical south pacific (ETSP) [134], North Eastern Subarctic Pacific (NESAP), Eastern South Atlantic (Knorr cruise), TARA Global Oceans survey [233] and a collection of >200 marine metagenomes sourced globally (see Table C.1 for list). Identified NosZ sequences were then mapped to the phylogenetic tree using MLTreeMap [250] and calculated RPKM values (or relative abundance for 454-sequenced datasets) were summed for hits to individual SAG sequences or clusters.

The top most abundant *nosZ* clades in the surveyed metagenomes were Clade 6, containing Marinimicrobia SHBH1141, Clade 9 containing the Ectothiorhodospirales and SAR324, Clade 5 containing Bacteroidales VC21 and Marinimicrobia ZA3312c, followed by Clade 13 containing SUP05_1a (Figure 5.3, Table D.2). Within other OMZ environments, particularly the ETSP [134] and the Peruvian upwelling [199], the distribution of *nosZ* clades in the metagenomes were similar to Saanich Inlet. While the most highly abundant clades in Saanich Inlet correspond to the collected SAGs, non-OMZ environments, represented in the TARA, Knorr and Global datasets showed broader distribution among the different clades (Figure 5.4).

Where available, *nosZ* gene expression in metatranscriptomic datasets (de-novo assemblies for Saanich Inlet, NESAP and Knorr and 454-sequenced reads for Peru) were also searched by protein sequence homology to Saanich Inlet NosZ SAG sequences using LAST+ [185]. It is noted that low abundance sequences in the transcriptome may not assemble and as such would not be

detected with this method, unassembled reads may still be present but not detected here. Top most abundantly expressed clades in the metatranscriptomes were similarly dominated by Type II NosZ sequences and were similar to the metagenomes however, in slightly different order (Table D.3). Clade 6 Marinimicrobia SHBH1141 showed highest levels of expression followed by Clade 5 containing Marinimicrobia ZA3312c and Bacteroidales V21. Further differences in *nosZ* gene expression are explored later in the text.

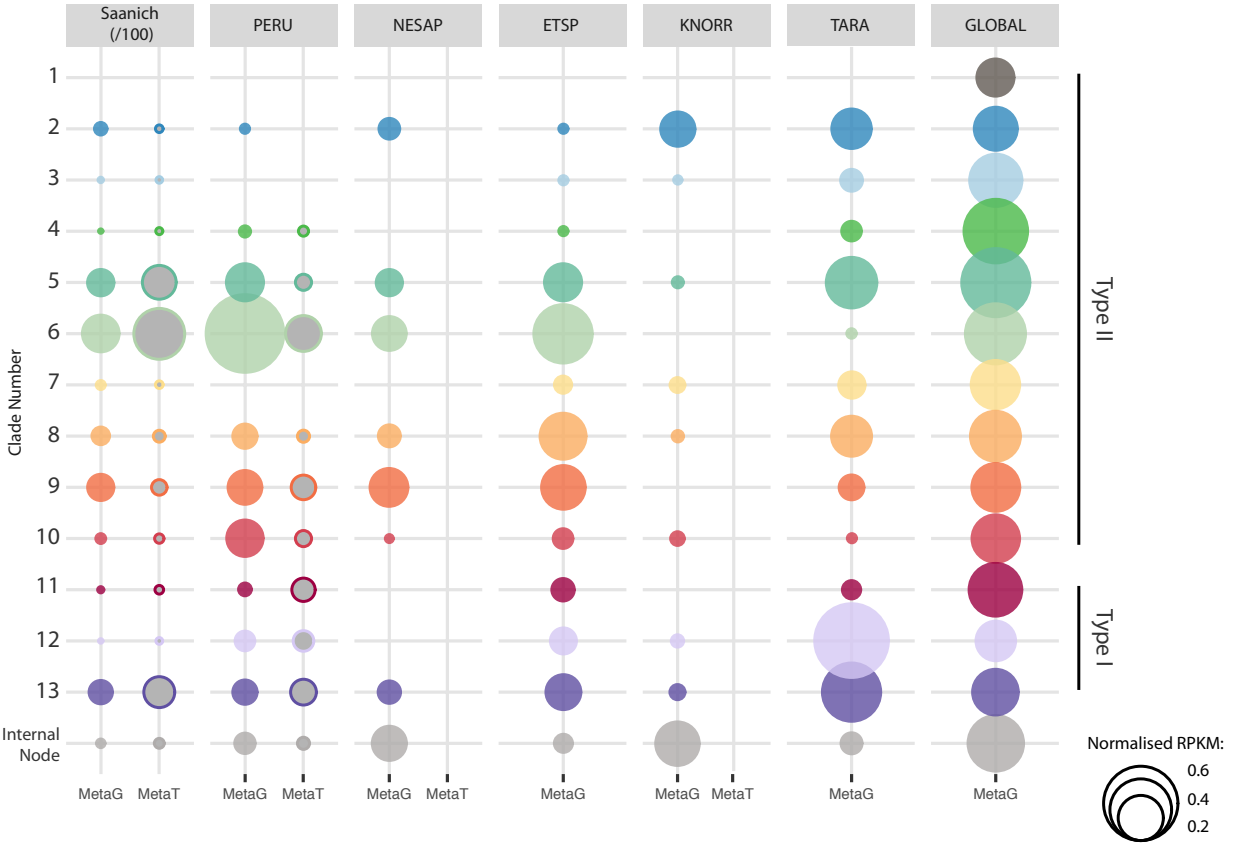


Figure 5.4: Abundance and expression of *nosZ* in global systems. Abundance (in RPKM) of *nosZ* clades in various datasets globally, for metagenomes (MetaG) and metatranscriptomes (MetaT)(where available, indicated by a vertical line with MetaT at the base) normalized to the total number of samples containing *nosZ* in a given dataset. Saanich Inlet abundance is shown decreased by factor of 100 for the purpose of visualization. Internal Node indicates *nosZ* RPKM mapped to internal nodes of the tree, rather then specific clades. Global refers to collection of >300 globally sourced metagenomes.

5.2.4 *nosZ* global distribution

The abundance and distribution of *nosZ* clades in other OMZs and in global ocean metagenomes showed substantial diversity of *nosZ* clades with biogeographic patterns emerging for some specific regions such as the Eastern South Atlantic (Knorr) (Figure D.4) as well as regions within the TARA Oceans global cruise track (Figure D.5). Notably, Clade 2 was more abundant in the Knorr samples and also apparent in the deep chlorophyll maximum (DCM) and mesopelagic in the TARA samples. Knorr deep water samples (>4000 m) showed an abundance of sequences which mapped to internal nodes, further placement of these sequences in the NosZ phylogenetic tree may resolve yet unidentified clades. Points where Tara cruise track passes through OMZs (as indicated by coloured dot at the base of the stacked bar) generally indicated a shift in clade structure within the DCM and mesopelagic samples, notably the inclusion of Clades 9 and 8.

Within other OMZs such as ETSP and Peru, where multi-omic samples were available, there was a similar clade distribution to Saanich Inlet (Figures 5.5, 5.6) with an overall dominance of Clade 6 in the metagenome. Indeed, Clade 6 appears to be primarily constrained to sulfidic systems and does not appear in Tara or Knorr samples. This is consistent with the distribution of the Marinimicrobia SHBH1141 clade in Chapter 4. Interestingly, there was a notable inconsistency in the metatranscriptome for Peru at 20 m where Clades 11 and 12 were more highly expressed. Both clades 11 and 12, found within the Type I portion of the NosZ tree, likely belong to complete denitrifiers. The chemistry of the water column at 20 m indicated a depletion of both NO_3^- and NO_2^- [199] consistent with denitrification, pointing to a possible niche for complete denitrifiers within OMZs as NOx^- become scarce.

5.2.5 NosZ time resolved multi-omic dynamics in Saanich Inlet

To further identify specific niches for *nosZ* clades I explored clade dynamics in the Saanich Inlet multi-omics dataset outlined in Chapter 2 [114, 115]. By charting multiple levels of information flow at the DNA, RNA and protein level patterns of expression along gradients of O_2, NO_3^- and H_2S were observed. In general, *nosZ* abundance increased with depth, generally showing peak abundance at 200 m but also at 120 m and 135 m following renewal in cruise SI075, consistent

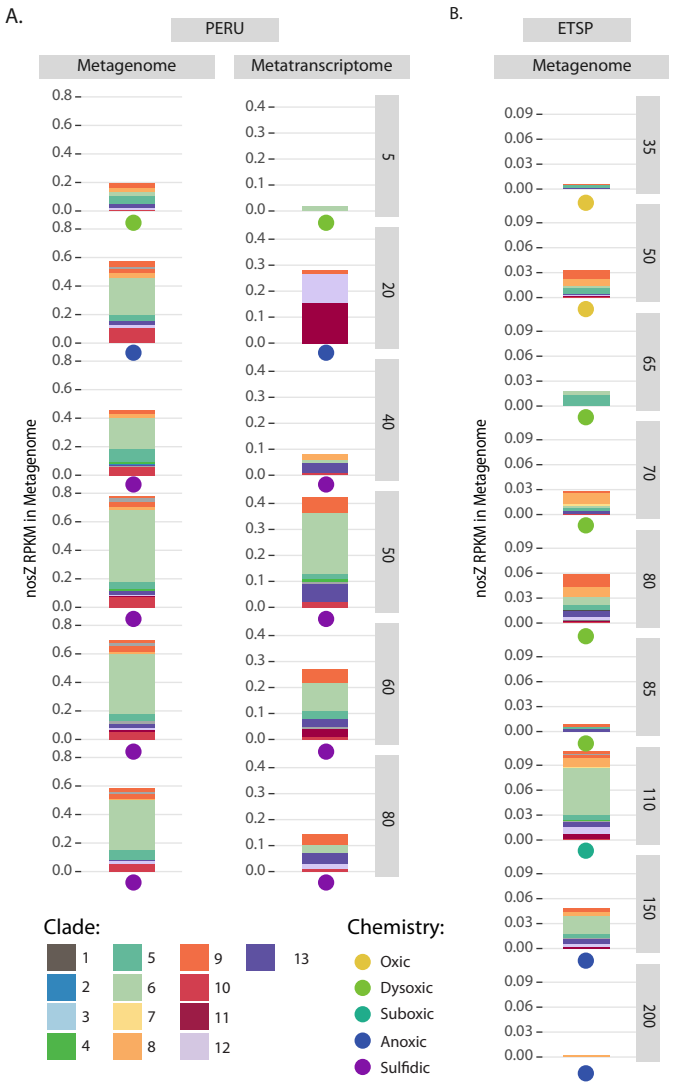


Figure 5.5: Peruvian and ETSP *nosZ* clade distribution. Distribution of *nosZ* clades in metagenome and mor metaranscriptome over depth in the Peruvian upwelling (PERU)(12°21.88'S to 77°0.00'W, [199]) and Eastern Tropical South Pacific (ETSP)(20°07'S, 70°23'W, [18]). Abundance in RPKM of *nosZ* clades in metagenome and metatranscriptome (Peru only) at indicated depths as found in previous studies. Chemical condition of individual sample are indicated in the coloured dot at the base of each stacked bar; oxic (>90 µM yellow), dysoxic (20-90 µM, green), suboxic (2-20 µM, teal) and anoxic (0-2 µM, blue).

with shoaling of 200 m waters in the renewal process (Figures 5.6, D.2) [8]. Clade abundances showed consistency with global analysis in the most abundant Clades, 6, 5, 13, 9, 8 with Clade 2 also showing intermittent low levels of abundance. Generally Clades 9 and 8 were more abundant in 100 m samples and Clade 2 more abundant at 100 and 120 m and during renewal at 135 m and 150 m in samples SI072, SI073 and SI074.

Saanich Inlet metatranscriptomes showed lower diversity compared to the metagenomes (Figure 5.6). The most abundant clades in the metatranscriptome included 6, 5 and 13 (Figure D.2). Total *nosZ* expression was highest during renewal in August 2010, cruise SI048, from 120 m down, peaking at 150 m. Expression from cruise SI048 varied from different clades, with Caldes 6 and 13 dominating at 120 and 135m, Clade 6 dominating at 150 m and Clade 5 dominating at 200 m. Interestingly, different taxa within some clades showed differential expression both over time and in the water column. Within Clade 5, Bacteroidales VC21 dominated expression in sulfidic waters, while Marinimicrobia ZA3312c was more often expressed in anoxic and suboxic waters (Figure 5.7). Clade 9 Ectothiorhodospirales was more highly expressed in sulfidic waters and following renewal at 135 m in cruise SI075, possibly due to shoaling of sulfidic basin waters, while SAR324 was expressed predominantly in dysoxic waters. Other members of Clade 9 also showed expression in predominantly sulfidic waters for cruises SI047, SI048 and SI054, time when the water column was highly stratified and sulfide accumulated at shallower depths than usual. Clade 13 was primarily dominated by SUP05_1a expression in anoxic waters.

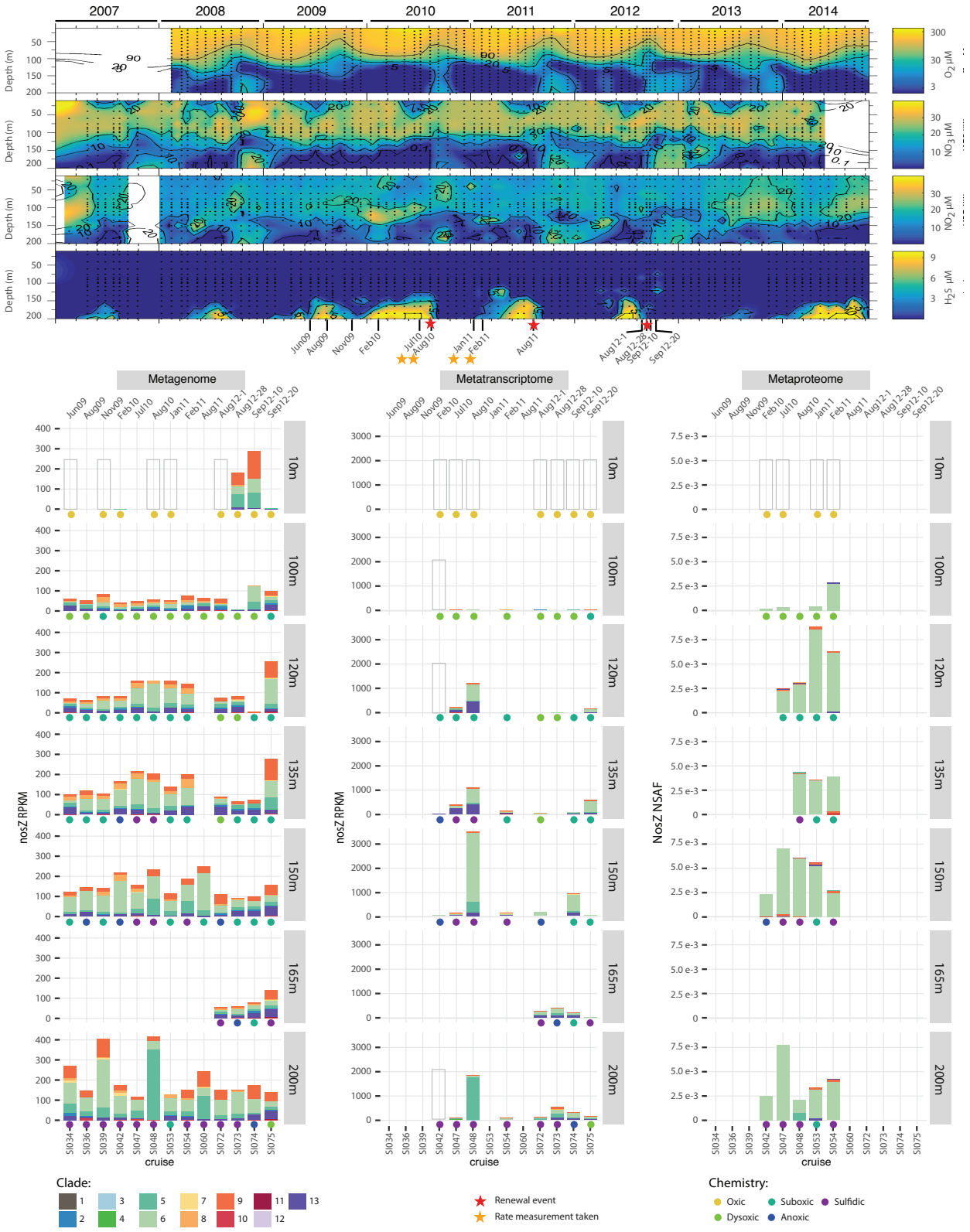

2007
2008
2009
2010
2011
2012
2013
2014
Depth (m)
Metagenome
Metatranscriptome
Metaproteome
nosZ RPKM
NosZ NSAF
cruise
10m
100m
120m
135m
150m
165m
200m
Clade:
Renewal event
Rate measurement taken
Chemistry:
Oxic
Suboxic
Sulfidic
Dysoxic
Anoxic

Figure 5.6: Saanich Inlet time series chemical profiles and *nosZ* multi-omic dynamics(previous page). Saanich Inlet water column chemistry for oxygen (O_2), nitrate (NO_3^-), nitrous oxide (N_2O) and hydrogen sulfide (SO_4^{2-}) over seven years. Abundance of *nosZ* clade in the metagenome, metatranscriptome and metaproteome for multiple time points (Month and year on top x axis, cruise ID on bottom x axis) and depths (Y-axis) in Saanich Inlet. Chemical condition of individual sample are indicated in the coloured dot at the base of each stacked bar; oxic (>90 µM yellow), dysoxic (20-90 µM, green), suboxic (2-20 µM, teal), anoxic (0-2 µM, blue), and sulfidic (purple). Absence of coloured bar and chemistry dot indicate no sample was available for analysis. Outline of a bar with the chemistry dot indicate no *nosZ* as detected. (The chemical profiles portion of this figure was previously published in Capelle *et al.* 2016, used with permission.)

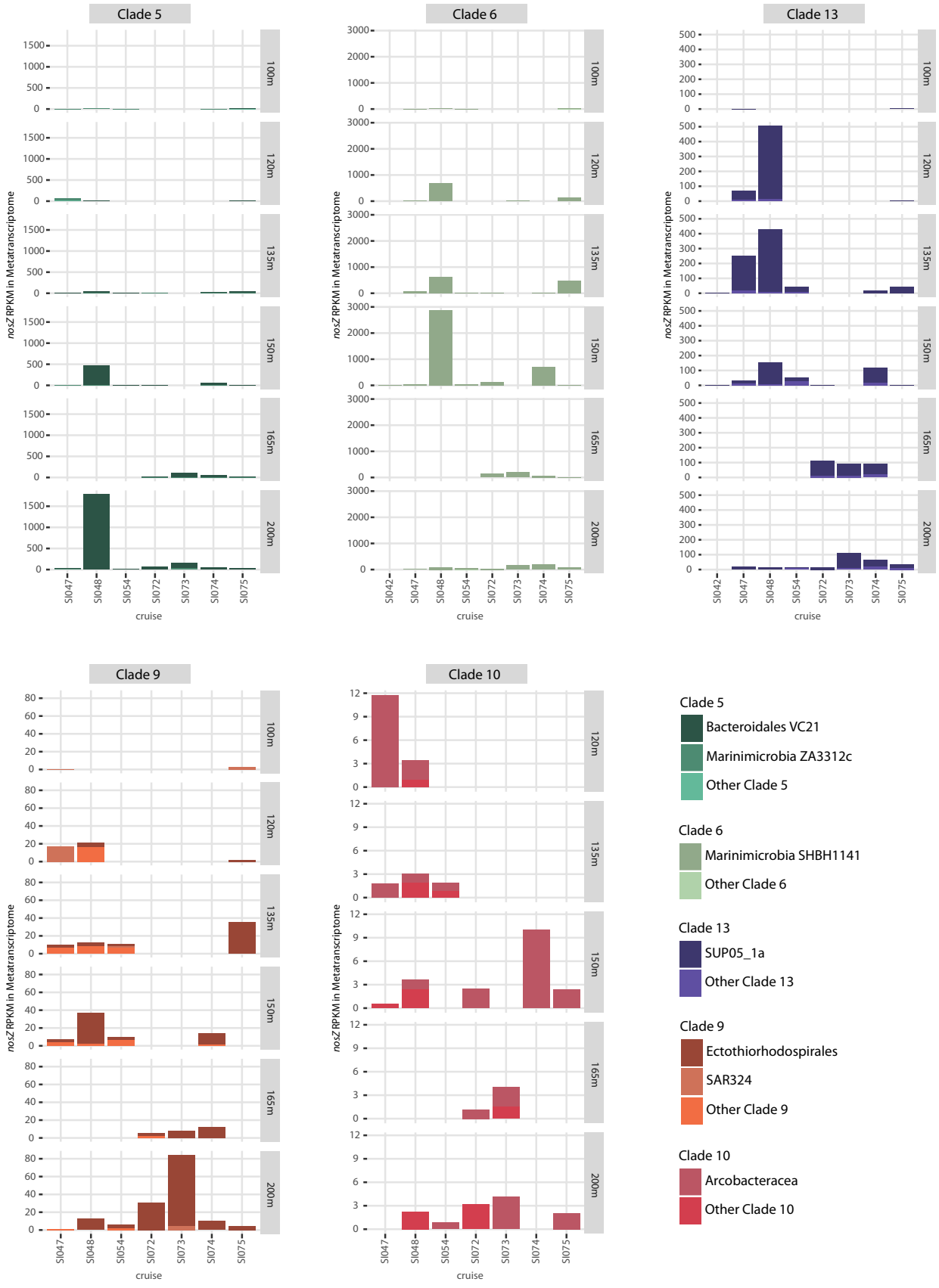

Clade 5
Clade 6
Clade 13
Clade 9
Clade 10
100m
120m
135m
150m
165m
200m
nosZ RPKM in Metatranscriptome
cruise
SI042
SI047
SI048
SI054
SI072
SI073
SI074
SI075
Clade 5
Bacteroidales VC21
Marinimicrobia ZA3312c
Other Clade 5
Clade 6
Marinimicrobia SHBH1141
Other Clade 6
Clade 13
SUP05_1a
Other Clade 13
Clade 9
Ectothiorhodospirales
SAR324
Other Clade 9
Clade 10
Arcobacteracea
Other Clade 10

Figure 5.7: Metatranscriptome expression dynamics of *nosZ* subclades in Saanich Inlet (previous page). Metatranscriptome RPKM for indicated *nosZ* subclades in Saanich Inlet over multiple time points and depths. No Expression from 10m samples was detected and this thus not included in this figure.

Saanich Inlet metaproteome NosZ was nearly completely dominated by Clade 6, affiliated with Marinimicrobia SHBH1141. NosZ protein was most abundant in cruise SI047, when the water column had not been renewed in over two years and was highly stratified. Interestingly, *nosZ* expression in the metatranscriptome was highest the following month in cruise SI048 following the influx of renewal waters (as indicated by drop in H_2S concentration in basin waters (see Figure 5.8)). Additionally, SI048 200 m showed protein expression from Clade 5 in relatively high abundance, consistent with Clade 5 expression in the metatranscriptome for the same time point and depth. In general, Expression of NosZ in the proteome did not correspond to potential rates or concentrations of N_2O in the water column (Figure 5.8). However, abundance of NosZ protein would likely result in low N_2O concentrations, confounding the expected correlation between N_2O and NosZ expression.

To further investigate the activity of NosZ and nitrogen-loss pathways including denitrification and anammox in Saanich Inlet, I carried out processes rate measurements with $^{15}NO_3^-$ and and $^{15}NH_4^+$ respectively. Overall, anammox appeared to be the dominant nitrogen-loss processes with high potential rates of denitrification (N_2 production) measured only during renewal in August 2010 (SI048) (Figure 5.8). While high potential rates of denitrification coincided with maximal expression of *nosZ* in the metatrancriptome at cruise SI048 they did not coincide with peak NosZ protein expression. In fact, peak protein expression of NosZ was at 120 m at SI053 (unfortunately no metatranscriptome was available for that time point). NosZ protein was detected in medial amounts for all samples were denitrification was measured. Interestingly, N_2O potential production was observed under anoxic, sulfidic conditions, with no detected NO_3^- or NO_2^- (SI047 200 m, SI048 135 m, 150 m, 200 m), though at SI048 the observation of N_2O production did correspond to high rates of denitrification.

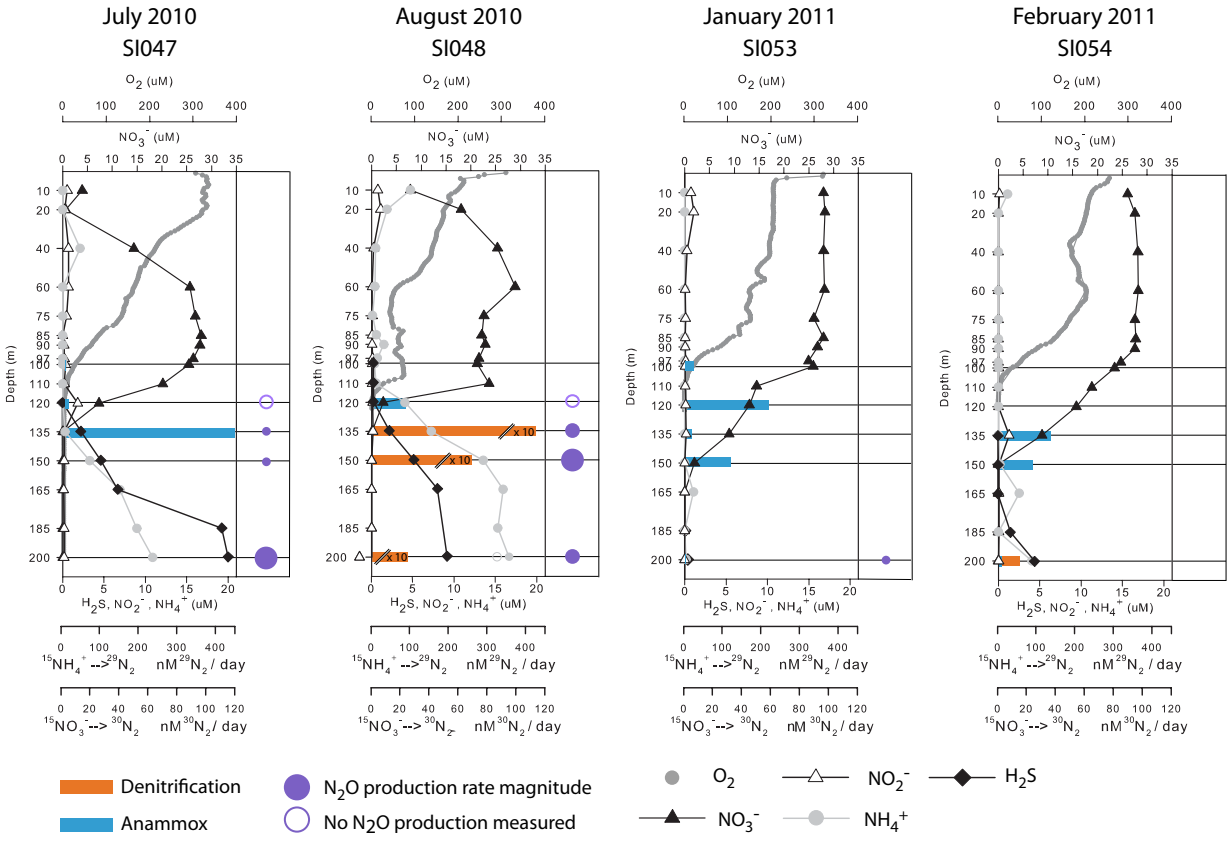


Figure 5.8: Saanich Inlet denitrification and anammox rates. Potential rates for denitrification (orange) and anammox (blue) taken as production of $^{30}N_2$ from addition of added $^{15}NO_3^-$ and $^{29}N_2$ from $^{15}NH_4^+$ addition respectively. Relative magnitude of N_2O production shown by purple circle, empty circle indicates no N_2O production measured and no circle indicates no measurement taken. Horizontal black line indicates depths at which rate measurements were taken. Absence of coloured bar indicates no rate detected.

5.2.6 *nosZ* global niches

Mapping NosZ clades and abundance in available metagenomes and metatranscriptomes onto chemical parameters of O_2, H_2S and NO_3^- revealed patterns of distribution and expression along environmental gradients. Within the metagenome, Clades 2, 4, 5, 7, 8, 9, 12 and 13 showed a distribution that included higher O_2 concentrations and a range of NO_3^- concentrations. Clades 3, 6, 10 and 11 showed presence with O_2 concentrations much closer to zero, though a range of NO_3^- was still is apparent. All clades were apparent under sulfidic conditions, though some appeared in higher abundance in anoxic waters. Expression was not observed to a large extent under higher O_2 conditions and *nosZ* was most abundantly expressed under sulfidic conditions (with zero NO_3^-) form Clades 5, 6 and 13. Some expression was seen at low to 0 µM O_2 concentrations and higher NO_3^- conditions. Overall, Clade 6 appeared to be most dominantly expressed in the metatranscriptome and primarily under sulfidic conditions.

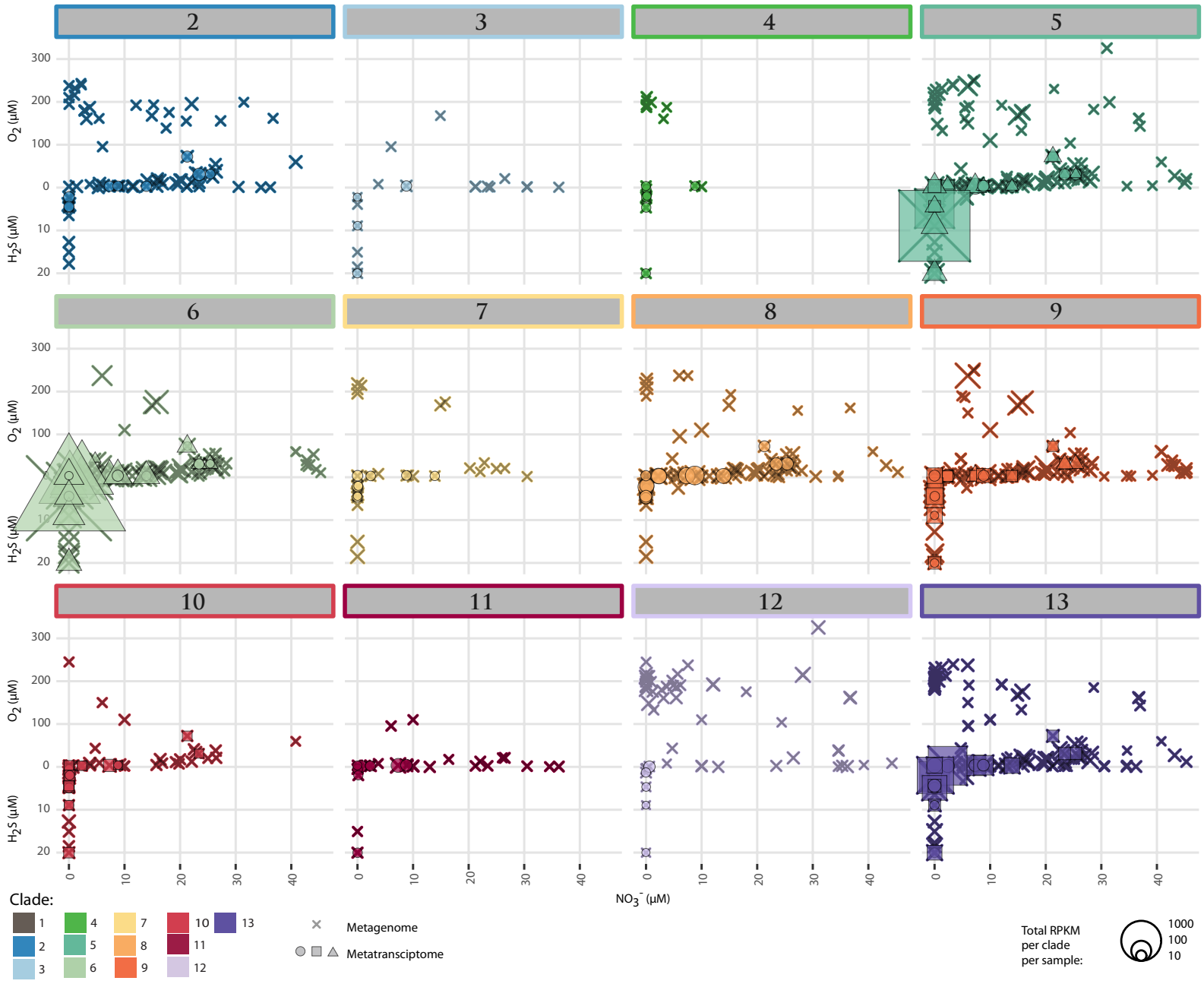

2
3
4
5
6
7
8
9
10
11
12
13
O2 (μM)
H2S (μM)
300
200
100
0
10
20
0
10
20
30
40
NO3− (μM)
Clade:
1
2
3
4
5
6
7
8
9
10
11
12
13
Metagenome
Metatransciptome
Total RPKM per clade per sample:
1000
100
10

Figure 5.9: *nosZ* clade distribution and expression along chemical gradients (previous page). Distribution and expression of *nosZ* clades along oxygen (O_2), hydrogen sulfide (H_2S) and nitrate (NO_3^-) from all datasets where size of symbol corresponds to RPKM in metagenome (X) and metatranscriptomes (coloured symbols); Clade 5: □ Bacteriodetes, △ Marinimicrobia Za3312c; Clade 6: △ Marinimicrobia SHBH1141; Clade 9: □ Ectothiorhodospirales, △ SAR324; Clade 10: □ Arcobacter; Clade 13: □ SUP05_1a.

5.2.7 Completing the tree with additional clades

Several NosZ sequences could not accurately be mapped to specific taxa and were consequently mapped to internal nodes by MLTreeMAP (Figure 5.10) [250]. The tree was rebuilt including these sequences to determine if they were new nodes or variants of existing leaves in the tree. After removing sequence that were below 30 amino acids long, 13 new sequences were added to the tree, resulting in two new clades and the re-assortment of another.

A new clade was formed from three sequences branching deeply in between the halophilic Archaea making up Clade 1 and the remainder of the Type II portion of the tree. These sequences were assigned a taxonomy of Bacteria using the TreeSAP algorithm to determine likely taxonomy based on the surrounding sequences in the tree [251]. These sequences were from different locations geographically, the Eastern tropical north pacific OMZ, the Arabian up-welling OMZ (branching together) and a third from Juan DeFuca hydrothermal vent.

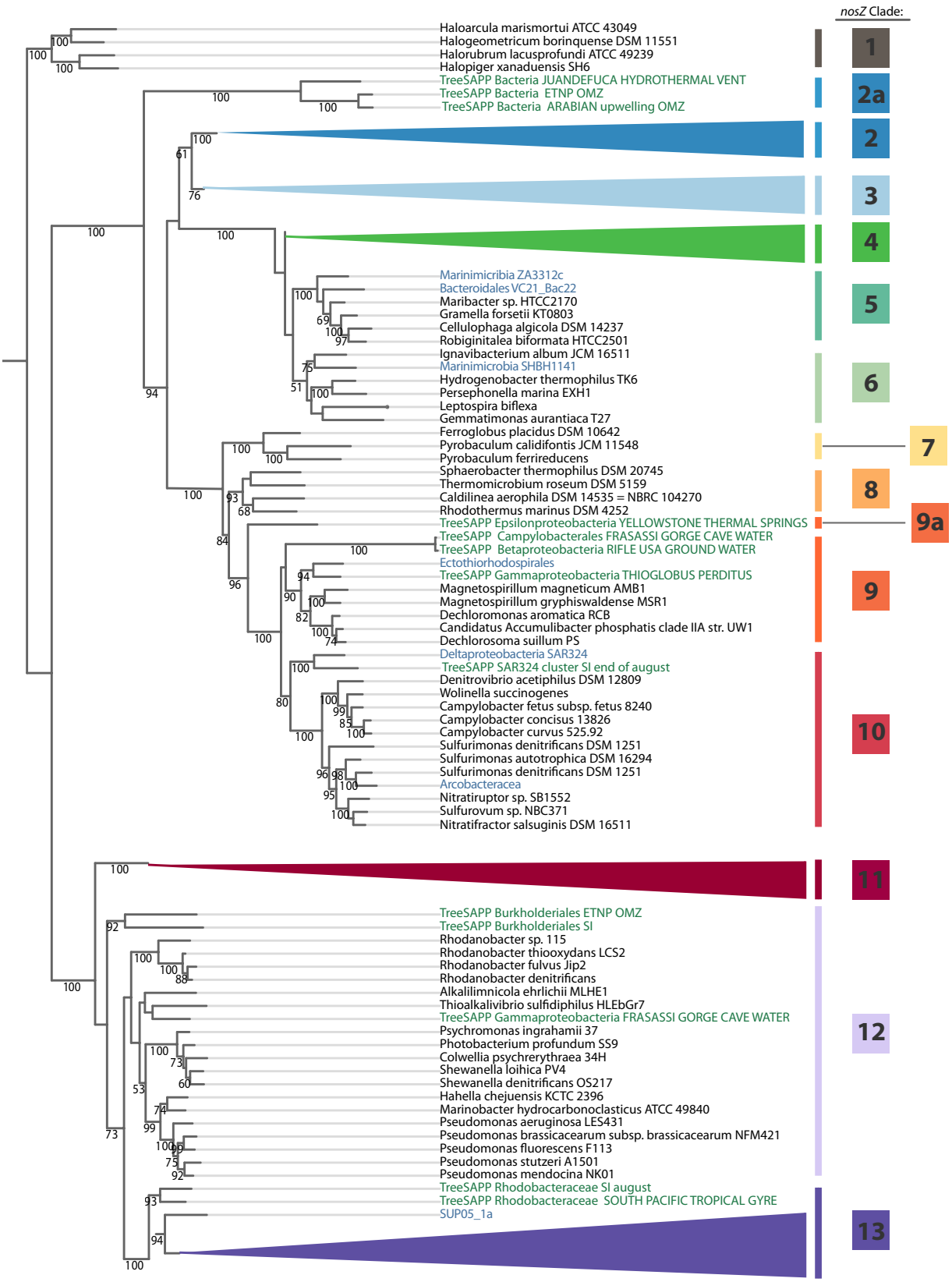

nosZ Clade:
Haloarcula marismortui ATCC 43049
Halogeometricum borinquense DSM 11551
Halorubrum lacusprofundi ATCC 49239
Halopiger xanaduensis SH6
TreeSAPP Bacteria JUANDEFUCA HYDROTHERMAL VENT
TreeSAPP Bacteria ETNP OMZ
TreeSAPP Bacteria ARABIAN upwelling OMZ
Marinimicribia ZA3312c
Bacteroidales VC21_Bac22
Maribacter sp. HTCC2170
Gramella forsetii KT0803
Cellulophaga algicola DSM 14237
Robiginitalea biformata HTCC2501
Ignavibacterium album JCM 16511
Marinimicrobia SHBH1141
Hydrogenobacter thermophilus TK6
Persephonella marina EXH1
Leptospira biflexa
Gemmatimonas aurantiaca T27
Ferroglobus placidus DSM 10642
Pyrobaculum calidifontis JCM 11548
Pyrobaculum ferrireducens
Sphaerobacter thermophilus DSM 20745
Thermomicrobium roseum DSM 5159
Caldilinea aerophila DSM 14535 = NBRC 104270
Rhodothermus marinus DSM 4252
TreeSAPP Epsilonproteobacteria YELLOWSTONE THERMAL SPRINGS
TreeSAPP Campylobacterales FRASASSI GORGE CAVE WATER
TreeSAPP Betaproteobacteria RIFLE USA GROUND WATER
Ectothiorhodospirales
TreeSAPP Gammaproteobacteria THIOGLOBUS PERDITUS
Magnetospirillum magneticum AMB1
Magnetospirillum gryphiswaldense MSR1
Dechloromonas aromatica RCB
Candidatus Accumulibacter phosphatis clade IIA str. UW1
Dechlorosoma suillum PS
Deltaproteobacteria SAR324
TreeSAPP SAR324 cluster SI end of august
Denitrovibrio acetiphilus DSM 12809
Wolinella succinogenes
Campylobacter fetus subsp. fetus 8240
Campylobacter concisus 13826
Campylobacter curvus 525.92
Sulfurimonas denitrificans DSM 1251
Sulfurimonas autotrophica DSM 16294
Sulfurimonas denitrificans DSM 1251
Arcobacteracea
Nitratiruptor sp. SB1552
Sulfurovum sp. NBC371
Nitratifractor salsuginis DSM 16511
TreeSAPP Burkholderiales ETNP OMZ
TreeSAPP Burkholderiales SI
Rhodanobacter sp. 115
Rhodanobacter thiooxydans LCS2
Rhodanobacter fulvus Jip2
Rhodanobacter denitrificans
Alkalilimnicola ehrlichii MLHE1
Thioalkalivibrio sulfidiphilus HLEbGr7
TreeSAPP Gammaproteobacteria FRASASSI GORGE CAVE WATER
Psychromonas ingrahamii 37
Photobacterium profundum SS9
Colwellia psychrerythraea 34H
Shewanella loihica PV4
Shewanella denitrificans OS217
Hahella chejuensis KCTC 2396
Marinobacter hydrocarbonoclasticus ATCC 49840
Pseudomonas aeruginosa LES431
Pseudomonas brassicacearum subsp. brassicacearum NFM421
Pseudomonas fluorescens F113
Pseudomonas stutzeri A1501
Pseudomonas mendocina NK01
TreeSAPP Rhodobacteraceae SI august
TreeSAPP Rhodobacteraceae SOUTH PACIFIC TROPICAL GYRE
SUP05_1a
1
2a
2
3
4
5
6
7
8
9a
9
10
11
12
13

Figure 5.10: NosZ tree with additional environmental sequences (previous page). Phylogenetic tree of NosZ protein sequences from figure 5.3 including metagenomic sequences which were previously mapped to internal nodes. Clades which were unchanged and do not harbour SAG or metagenomic sequences were collapsed. Blue sequence names indicate sequences from SAGs, green sequence names indicate new metagenomic sequences.

The second new clade is formed from a single sequence branching between Clades 8 and 9 and is taxonomically assigned to Epsilonproteobacteia and was from Yellowstone thermal springs. Two additional sequences were also added to Clade 9, a Campylobacterales from a fresh water gorge in Farassi Italy and a Betaproteobacteria from ground water Colorado USA. These taxonomic assignments are consistent with Clades 9 and 10 containing NosZ from various proteobacteria. Clade 9 is further re-arranged with the addition of NosZ from ^{U}*Thioglobus perditus* a metagenome assembled genome from the Peruvian upwelling [36], which is part of the SUP05 clade. Interestingly, the ^{U}*T. perditus* NosZ does not cluster with the Saanich Inlet SAG NosZ in the Type I portion of the tree. These additions to the tree further rearrange Clade 9 and 10, with SAR324 and an additional sequence assigned to SAR324 from Saanich Inlet, moving to Clade 10.

Within the Type I portion of the tree, a deep branching node at the base of Clade 12 is formed by two sequences taxonomically assigned to Burkholderiales (Betaproteobacteria). One sequence is from the Eastern Tropical North Pacific (ETNP) and one from Saanich Inlet. Also within Clade 12, an additional sequence was added branching with Thioalkalivibrio and assigned to Gammaproteobacteria from fresh water sulfidic gorge in Frasassi Italy. Within Clade 13, an additional, deeper branching node, was formed by two sequences taxonomically assigned to Rhodobacteraceae. One sequence is from the South Pacific Tropical Gyre (SPTG) and another from Saanich Inlet.

5.3 Discussion

Using SAGs, I was able to phylogenetically anchor environmental metagenomic NosZ sequences, identify previous unknown taxonomic groups with the potential metabolic capacity to reduce N_2O and appended additional clades to the NosZ reference tree. Furthermore, assessment of the abundance of identified NosZ clades confirmed the recently found global distribution

of Marinimicrobia SHBH1141 *nosZ* and established it as the dominant *nosZ* in OMZs waters. Taxonomies not previously known to carry *nosZ* such as Bacteroidales VC21, SAR324 and the Gammaproteobacterial Ectothiorhodospirales were identified. Saanich Inlet SUP05_1a SAGs appear to be similar to the SUP05 group ^{U}*T. perditus* (unclutured), recently binned from Peruvian OMZ metagenomes, as both are seen to carry *nosZ* gene cluster [36]. However, the different location of these two SUP05 NosZ sequences in the phylogenetic tree raises several questions about SUP05 metabolic flexibility and convergent evolution. While little is known about SAR324, SAGs from the deep ocean [77] and metagenome assembled genome from the Guaymas Basin [27] indicate capacity for C1 metabolism and sulfur oxidation, as well as genes for initial steps of denitrification. Further investigation of SAR324 Saanich Inlet SAGs should determine SAR324 to be a complete denitrifier or non-denitrifying N_2O reducer and also identify linkages between proposed nitrogen, sulphur and carbon metabolisms in OMZs. No other genomic information is available for Bacteroidales VC21, a group often found in OMZs [2] and hydrothermal vents [252], further metabolic analysis of the Saanich Inlet Bacteroidales VC21 SAGs may provide important information with respect to nitrogen cycling in the Global Ocean.

Samples from the deep South Atlantic (>4000 m) from the Knorr Cruise and globally sourced metagenomes showed several sequences mapping to internal nodes on the NosZ tree, which were then incorporated into a new tree, adding a few new leaves and two new clades. The addition of new clades from relatively unexplored environments suggests that metagenomic analysis of environments such as deep sea sediments and hydrothermal vents may yield additional novel *nosZ* clades and unanticipated potential N_2O sinks.

Some trends are apparent in the biogeography of individual *nosZ* clades. Clade 2, consisting predominantly of Firmicutes, appears to represent an open ocean clade given its consistent abundance within the TARA and Knorr datasets and relatively low abundance in OMZ datasets. The apparent restriction of Clade 6 to systems with sulfide (Saanich Inlet, Peruvian upwelling) or active cryptic sulfur cycling (ETSP) [78] confirm Clade 6 Marinimicrobia SHBH1141 to be the primarily active N_2O-reducer within OMZs. The appearance of Clade 5 in higher abundances in more oxygenated waters both in the open ocean and OMZ samples as well as in Saanich Inlet renewal samples, suggests a higher O_2 tolerance for this clade. Clade 5 also shows interesting

expression dynamics between clade members Bacteroidales VC21 and Marinimicrobia ZA3312c along the Saanich Inlet water column over time including a higher abundance of Bacteroidales VC21 during renewal and may indicate its presence in oxygenated renewal waters. Assessment of microbial community and *nosZ* clade structure within renewal waters would be necessary to confirm this hypothesis.

Interestingly, inconsistencies in expression dynamics within certain clades in Saanich Inlet suggests that organisms within a given clade may not behave similarly and thus clade membership may not be the best predictor of expression patterns or global distribution. This points to regulatory factors, likely within the NGC, which may differ within clades. Further analysis of elements within the NGC, including phylogenetic trees, may help to identify regulatory elements and possibly identify horizontal gene transfer either within or between clades leading to variation in expression patterns. Ultimately, understanding the environmental conditions leading to expression of NosZ and N_2O reduction will aid in understanding of N_2O consumption within the ocean.

The prevalence of *nosZ* within the surface waters of TARA Global Oceans Survey as well as throughout the water column in Knorr metagenomes is puzzling, as N_2O reduction is an anaerobic process. Presence in oxygenated waters may reflect particle association of *nosZ* carrying organisms, as observed in the ETSP subsurface OMZ waters [15]. The formation of anoxic micro-niches within particles may serve to support anaerobic processes in otherwise oxygenated bulk water [253], providing an additional niche for N_2O production and consumption. Recently, N_2O production was observed from *Nitrococcus* [54], a nitrite oxidizing bacteria isolated from OMZ waters and points to a potential source of N_2O for N_2O-reducers in dysoxic and oxygenated waters. Additional work on rate measurements of N_2O production and consumption within open ocean surface waters and on particles may identify a previously unaccounted for N_2O reducing capacity.

The dynamics of the NosZ proteins in Saanich Inlet, coupled to rate measurements, indicate a discrepancy between RNA and protein expression. Reasons behind this most likely lie in sampling methodology. Although care was taken to reduce wait times between sample collection and filtering for RNA and protein, the microbial community may respond to trace amounts of O_2 introduced during collection, altering expression profiles. Lack of detected $^{30}N_2$ production

in many samples may also indicate reduction of NOx^- to ammonia (NH_4^+) via dissimilatory nitrate reduction to ammonium (DNRA) rather than to N_2 via denitrification. Future rate measurements including DNRA and N_2O production as well as analysis of the metatranscriptomes and metaproteomes for genes involved in other transformations in the nitrogen cycle and house keeping genes/proteins from various groups of interest may provide additional insight into nitrogen cycling processes at work in Saanich Inlet.

It is intriguing that Saanich Inlet showed a low abundance of Type I *nosZ* that are generally associated with complete denitrifiers. Type I *nosZ* were seen more abundantly at specific depths in the Peru, ETSP and TARA data sets. Low occurrence of Type I *nosZ* in Saanich Inlet suggests that the beneficial mutualism between incomplete denitrifying type SUP05 and Marinimicrobia SHBH1141 proposed in Chapter 4 [249] may serve to outcompete individual Type I complete denitrifiers. Particularly within Saanich Inlet, the fluctuations in NOx^- and H_2S may select for organisms with robust metabolisms such as SUP05 with multiple NO_3^- reductases and multiple H_2S oxidases [33], as well as Marinimicrobia capable of storing polysulfide for reduction/oxidation under energy stress [249]. The proposed mutualism between these organisms would further their resilience beyond an individual complete denitrifier such as *S. gotlandica*, similar to Arcobacteraceae found in Saanich Inlet. Interestingly, the presence of *nosZ* in SUP05_1a also posits the possibility of a similar mutualism between incomplete and complete denitrifying SUP05 clades which could also account for the success of SUP05 in Saanich Inlet. The extent to which these partnerships occur and the ramifications for N_2O production and consumption in Global Ocean OMZs have yet to be determined.

The presence of different *nosZ* sequences in the Saanich Inlet SUP05 SAGs and in the Peruvian upwelling ^{U}*T. perditus* metagenome assembled genome (MAG) brings up questions about SUP05 metabolic flexibility and convergent evolution. It also may call into question the validity of MAGs to reflect natural populations. Metagenome assembled genomes have recently come into wide use in the field of metagenomics as a method for automated generation of population genomes. However, the methods of validating their completeness and contamination have yet to be thoroughly vetted. When considering reasons for different *nosZ* sequences in these two SUP05 genomes, the possibility that the *nosZ* in ^{U}*T. perditus* is a contaminant is raised, however,

its location on a long contig (>10,000 base pairs) makes it unlikely that the ^{U}*T. perditus nosZ* is a contaminant. The presence of *nosZ* on 10 out of 48 SUP05_1a SAGs (several of which also reside on contigs >10,000) is fairly conclusive that the SUP05 Type I *nosZ* is contained in at least some of the SUP05_1a population in Saanich Inlet. Interestingly, the Ectothiorhodospirales-type *nosZ* found on a single SUP05_1c SAG (contig size >45,000 base pairs), thought to be a contaminant, may reflect the ability of SUP05 to pick up a *nosZ* NGC from the environment when it is advantageous to do so. This supports an idea of convergent evolution of two different SUP05 groups, gaining the same function, i.e. N_2O-reduction, from different sources to occupy a similar niche. Further investigation of the structure and similarity of the NGC from the Saanich Inlet SAGs for SUP05_1a, SUP05_1c and Ectothiorhodospirales, and ^{U}*T. perditus* would be very informative as to the location and method of potential gene transfer. Further analysis of Saanich Inlet population dynamics of SUP05 groups carrying respective NGCs, as well as the dynamics of the ^{U}*T. perditus* NGC within Peruvian upwelling metagenomes, may highlight conditions under which SUP05 N_2O-reducing capacity would be advantageous. Such information would be valuable for understanding the contribution of N_2O production/consumption by this abundant OMZ group.

Furthermore, the potential of both complete and incomplete denitrifying SUP05 clades and interactions between N_2O-reducing Marinimicrobia SHBH1141 brings up questions regarding gene loss/acquisition and fitness of individual organisms vs fitness of metabolically coupled partners/communities. Incomplete denitrifying SUP05 has apparently made a living by creating the *nosZ* niche for other organism by producing N_2O. Further investigation of SUP05 genomes both from Saanich Inlet and other OMZs would help to address fundamental questions about fitness and gene loss within the context of microbial communities and distributed metabolisms.

5.4 Conclusions

Overall, marine systems show a significantly greater abundance of Type II *nosZ* over Type I, similar to surveyed terrestrial systems. The abundance of Type II *nosZ*, associated with non-denitrifying N_2O-reducers, indicates globally distributed niches for non-denitrifying N_2O-reducing organisms along varying concentrations of O_2, NO_3^- and H_2S. The distribution of the nitrous oxide reductase

gene *nosZ* in the Global Oceans, with broad taxonomic associations, points to unappreciated sources and sinks of the potent greenhouse gas N_2O. While coastal areas and OMZs are stronger sources for N_2O production [236], the expression of *nosZ* in OMZ metatranscriptomes and metaproteomes also point to N_2O consumption, but the factors regulating the balance between roduction and consumption have yet to be constrained. The presence of *nosZ* within oxic surface waters suggests potential N_2O sources in the surface ocean. However, the flux of N_2O in many areas of the ocean has yet to be measured and thus the balance of N_2O production vs. consumption has yet to be constrained. As Global Ocean O_2 concentrations continue to decline globally, increasing the production of N_2O, the capacity of marine waters to consume N_2O is of critical importance for future climate models.

5.5 Methods

5.5.1 Single-cell Amplified genome collection, sequencing and annotation

SAGs were collected and sequenced from Saanich Inlet as described in Hawley *et al.* 2017 and in Chapter 4, and sequenced on Illumnia hi-seq. Samples for SAGs were collected from Saanich Station S3 on August 10th, 2011 at 100 m, 150 m, 185 m, using 1 mL of sample water into 143 µL 48% Betaine, frozen on dry ice and stored at -80°C. Sequences were assembled in SPAdes3.9 [254] and functional annotation carreid out in the Metapathways pipeline [191]. SAGs containing the nitrous oxide reductase gene *nosZ* were identified and used in further analysis. SAGs containing *nosZ* were manually decontaminated based on visual analysis of Kmer frequency principle component analysis of contigs containing genes of interest in the denitrification pathway, ensuring all contigs containing genes of interest shared Kmer space with the majority of the SAG Kmer space.

5.5.2 Multi-omics datasets

Saanich Inlet and NESAP metagenomes and metatranscriptomes and metaproteomes were generated as described in Chapter 2. Metagenomes and metatranscriptomes from other geographic areas were from other publications detailed in Table C.1.

5.5.3 Identification and clustering of *nosZ* sequences

Nitrous oxide reductase genes were identified in the SAGs based on functional annotation in Metapathways [94, 182] and validated by sequence homology searches to RefSeq. SAG NosZ amino acid sequences were then clustered at 95% id and representative sequences used in sequence homology comparisons against environmental metagenomes and metatranscriptomes using FAST algorithm [185] with a cutoff of 30% identity and alignment length of over 50 amino acids. For each metagenomic dataset (i.e. Saanich Inlet/NESAP, ETSP, PERU, TARA, KNORR) identified NosZ amino acid sequences were clustered at 85% identity using USEARCH `cluster_fast` with default settings and removing sequences <90 amino acids long. Representative sequences from each clustered dataset were then clustered together at 85% identity and mapped to phylogenetic reference tree using MLTreemap [250].

5.5.4 Generation of NosZ phylogenetic tree

Reference tree of NosZ protein sequences was generated using sequences from Sanford *et al.* 2012 clustered at 95% (USEARCH `cluster_fast`) to collapsed sequence redundancy, totalling 113 representative sequences from an original 136. RaxML was used to construct the tree with MLTreemap [250] was used to add SAG NosZ sequences. Metagenomic and metatranscriptomic sequences were then mapped on using MLTreemap.

5.5.5 Gene, transcript and protein abundance mapping

For metagenomes and metatranscriptoms from Saanich Inlet, NESAP, KNORR and TARA reads per kilobase mapped per million were (RPKM) were calculated through the Metapathways pipeline [191] and summed for a given *nosZ* clade and or SAG sequence such as clade 5 Marinimicrobia ZA3312c and Clade 5 Bacteroidales VC21. For all other datasets the number of genes found for a given dataset and cluster were summed for a given *nosZ* clade. For metaproteome from Saanich Inlet Normalized spectral abundance factor (NSAF) were calculated as described in Chapter 2)

5.5.6 Denitrification and Anammox rate measurements

Rates of denitrification and anammox were taken as described in Holtappels *et al.* 2011 [255].

Chapter 6

Conclusions

Throughout this thesis, multi-omics approaches were used to chart microbial community structure, identify nitrogen (N) and sulphur (S)-based energy metabolisms and carbon (C) fixation pathways as well as to propose metabolic interactions along redox gradients in marine oxygen minimum zones (OMZs). Metaproteomics was used in the development of a conceptual model for C, N and S-based metabolic interactions between dominant taxa along the redox gradient in Saanich Inlet, illustrating energetic coupling of denitrification and anammox to sulfur oxidation and carbon fixation within these taxa. These findings formed the basis for a collaborative work in Louca *et al.* to build a steady state multi-omic mathematical model, confirming the conceptual model and realising a previously unrecognized niche for nitrous oxide (N_2O) reduction [132]. Genes for N_2O reduction (*nosZ*) were identified within single cell amplified genomes (SAGs) from the dark matter phylum Marinimicrobia SHBH1141 clade collected from Saanich Inlet sulfidic basin waters, filling the niche of non-denitrifying N_2O-reducers proposed by Louca *et al.*. Analysis of energy metabolism and biogeography of additional Marinimicrobia clades, defined by globally sourced SAGs, found several clades to play additional roles in C, N and S cycling along redox gradients in the Global Ocean. Finally, using SAGs from Saanich Inlet, I phylogenetically anchored multiple *nosZ* genes, placing the Marinimicrobia SHBH1141 as the dominant N_2O-reducing group within anoxic and sulfidic OMZs globally and further explored the distribution and abundance of different *nosZ* clades in the Global Ocean. As OMZs continue to expand and intensify due to climate change the metabolic processes and interactions involved in N-loss and greenhouse gas production/consumption within O_2-depleted systems becomes increasingly important to nutrient and energy flow and ecosystem services. The findings in this thesis add knowledge about key microbial taxa involved in N and S-cycling and carbon fixation, adding insights into metabolic

interactions on the level of the microbial community in OMZs and the Global Ocean.

This final chapter discusses some of the advantages and limitations of multi-omic approaches and identifies improvements to expand these approaches both in the field and *in-silico*. This chapter also looks at expanding findings from Saanich Inlet to other global OMZs with respect to the ecology of dominant organisms across multiple OMZs and further explores the themes in eco-thermodynamics in relation to microbial community structure and metabolic interactions.

6.1 Advantages and limitations with multi-omics approaches

Multi-omic approaches facilitate the culture independent study of microbial communities within natural and engineered environments and support the construction of hypotheses about metabolic networks and interactions. However, reconstructing microbial interactions from environmental samples is far from a perfect practice with many assumptions and limitations. While next generation sequencing has greatly reduced sampling bias compared to traditional methods involving preparation of Sanger sequencing libraries, assembly of contiguous genomic sequences for a given taxonomic group from millions of short reads generated by sequencing platforms such as Illumnia may lead to construction of chimeras, confounding taxonomic assignment and metabolic reconstruction efforts. Once assembled, computational advances have been developed to predict open reading frames and assign function but ultimately functional annotation inherently relies on sequence similarity to existing databases, essentially making such functional annotations hypotheses. While manually checking annotation against multiple hits and databases to address miss-annotation is possible for selected genes of interest, it is not feasible for entire metagenomes. Additional manual annotation techniques, such as building phylogenetic trees, as done for *nosZ*, and exploring protein structures, while highly informative, are possible for only a limited number of genes. Furthermore, functional annotation does not take into account enzyme activity and/or regulation, which cannot be uncovered without further laboratory experimentation. Despite these inherent limitations, multi-omic analysis remains the best approach currently available to assess the overall metabolic potential of a microbial community at the individual, population and community level.

6.2 Methodological and analytical developments

6.2.1 Field work and sampling

Oxygen measurements

Throughout this thesis, gradients of O_2 and NO_3^- are seen to impact processes of H_2S oxidation, nitrogen loss and production/consumption of N_2O, particularly relating to O_2 inhibition of NosZ and other nitrogen cycling processes. Accurate measurement of O_2 concentrations directly within the sampling environment is critical to understand the impact of these processes on metabolic processes. Furthermore, several recent studies have uncovered evidence of cryptic oxygen cycling where nano-molar concentrations of oxygen (O_2) are generated within the OMZ [256, 257], with implications for nitrogen loss and greenhouse gas production as many enzymes involved in denitrification and anammox are oxygen sensitive. The ability to detect nano-molar concentrations of O_2 requires a switchable trace oxygen (STOX) sensor [258] and adaptations to sampling protocols. Currently, the Saanich Inlet time series is limited to the 2-3 µM detection range provided by an O_2 optode and coupled Winkler titration method. Implementation of the STOX sensor could address some of the inconsistencies seen in the Saanich Inlet time series data sets such as the presence of proteins involved in aerobic processes, i.e. nitrification, in anoxic waters seen in Chapters 3 and 5 or expression of different *nosZ* clades in the proteome under varying conditions. Incorporation of more precise O_2 measurements into the Saanich Inlet time series could contribute to more meaningful interpretations of multi-omic datasets and more accurate modeling of N and S energy metabolism and dark carbon fixation extensible to other O_2-depleted systems.

6.2.2 Analysis and Methodologies

Linking taxonomy and function

One of the main challenges in multi-omics analysis is the ability to match taxonomy to function, identifying the microbial groups within a community responsible for specific metabolic processes. Within this thesis, I used sequence homology to previously isolated and sequenced organisms

or manually curated metagenome assembled genome (MAG) and single cell amplified genomes (SAGs), for SUP05 and Marinimicrobia respectively. Advances in both sequencing technologies and computational analyses are making matching taxonomy and function more accurate and more accessible for high-through-put metagenomic applications. On the technological front, SAGs and long-read technologies such as Oxford-nanopore, produce sequence data from a single organism, ideally with enough sequence data to provide phylogenetic and taxonomic anchors (such as ribosomal genes and single copy marker genes [230]). Both SAG sequences and long-reads can serve as nucleation points for further recruiting reads and/or contigs from short-read metagenomes (e.g. Illumnia) and provide information on abundance of sequenced organisms within an environment. On the analytical front, there is a growing collection of binning algorithms and software to produce MAGs en-mass from metagenomic sequence data [35]. While MAGs may provide insight into matching function to taxonomy, the phylogenetic level at which they operate is not yet well defined in that if a bin reflects a strain, species, genus seems to be variable. Phylogenetic anchoring of key functional genes, such as the N_2O reductase *nosZ* and other genes involved in the nitrogen cycle, is key to building comprehensive understanding of microbial interactions as well as extending that understanding to global processes and models.

SAG-EXtrapolator

Mapping metagenomic contigs to long reads (including Sanger-sequenced full fosmids) and SAGs with high stringency may provide an exploitation of these phylogenetically anchored sequences to build population level genomic bins. SAGExtrapolator (SAGEX), put forth by previous Hallam Lab Masters student W. Evan Durno [259], which automates an approach originally conceived by Dodsworth *et at.*, 2013 to recruit metagenomic contigs to microbial dark matter phyla OP9 SAG t [181]. SAGEX recruits metagenomic contigs to a SAG at high sequence similarity and validates by tetranucleotide frequency [227], effectively extrapolating the genomic sequence data beyond the SAG. SAGEX offers a high-throughput supervised binning algorithm based on the approach used in Chapter 4 to bind metagenomic contigs to Marinimicrobia SAGs. SAGEX-generated population genome bins have an advantage over purely algorithm-generated MAGs in that SAGEX basses binning on sequence data directly from an individual organism opposed to only assembled contigs

within the metagenome.

Including abundance information such as RPKM for recruited contigs can additionally provide valuable information about the abundance of various microbial populations. Used across geographic areas, SAGEX can facilitate comparative genomics for various taxa, addressing questions of endemism vs cosmopolitanism and correlation of specific groups with environmental conditions such as O_2 concentrations, providing a more complete picture of metabolic processes and coupling in O_2-depleted waters.

6.3 Expanding the Saanich Inlet model to Global OMZs

Saanich Inlet serves as a model OMZ but the extent to which the information that is uncovered in this model informs what is known about other OMZ systems, both coastal and open ocean, in many cases remains to be explored and validated. Ongoing research both in Saanich Inlet and other OMZs continually adds new knowledge and insights to microbial processes operating under O_2-depletion. Integration of this knowledge, such as using genomic information for SAGs in Saanich to phylogenetically anchor functional genes in other OMZs, can address large scale ecological questions.

6.3.1 Questions about ecology and global implications

Fundamental questions about microbial community organization arise when comparing Saanich Inlet and other OMZs on a global scale. For example, the question of endemism vs cosmopolitanism; which species or clades may be found only in Saanich Inlet and which ones are found in other OMZs? Are the endemic groups perhaps key-stone members, carrying out functions essential and specific to that environment (e.g. trace metal redox reactions or detoxification). Do cosmopolitan taxa shift in abundance in the same manner along environmental gradients in different systems? While these questions hold across all ecosystems, the unique redox gradients found in OMZs (and O_2-deficient waters) support investigation of a range of micro-niches within a larger environment. The assembly of microbial communities and associated metabolic interactions along these gradients within different geographic locations offer an opportunity to address specific

questions of community organization and population dynamics within specific clades. Within taxa contributing to global nitrogen, sulphur and carbon cycles such as SUP05 and Marinimicrobia SHBH1141, understanding population dynamics are key for modeling global processes such as N_2O production/consumption.

6.3.2 SUP05 sub-clade metabolism, population dynamics and biogeography

The SUP05 group of denitrifying, sulfur oxidizing Gammaproteobacteria is an obvious group for investigation of population dynamics both in Saanich Inlet and other OMZs. The SAGs collected in Saanich Inlet indicate the presence of two predominate clades, SUP05_1a and SUP05_1c (Figure D.1), but questions remain about their metabolic capacities and niche differentiation, specifically with respect to N_2O reduction. The Saanich Inlet time series from Chapter 2 offers a unique opportunity to address time-resolved population dynamics and changes in metabolic capacity along defined redox gradients by mapping sequencing reads from the time series metagenomes to the SAGs within the respective clades. Expression from the respective clades could be further observed by mapping reads from the metatranscriptomes in the same way. Clade abundance or expression over time and along the redox gradient could possibly identify correlations of clades with specific environmental factors (O_2, NO_3^-, NO_2^- etc.) and aid in the definition of niches for the two clades.

Further to SUP05 population dynamics in Saanich Inlet would be a comparative genomics study across SUP05 genomes from multiple environments including SUP05 species that have been recently sequenced from Effingham Inlet Canada, *Candidatus* Thioglobus autotrophicus (cultured isolate) [34] and the Peruvian upwelling system ^{U}*Thioglobus perditus* (MAG) [36]. Together, population dynamics and comparative genomics could provide a greater understanding of the role different clades of the abundant SUP05 group with respect to N_2O production/consumption, sulfide oxidation and detoxification [75] and carbon fixation [36, 79] in different OMZs.

6.4 Themes in microbial interactions along eco-thermodynamic gradients

Using an eco-thermodynamic approach to microbial ecology, I explore how energy available to a microbial community flows through community members in the form of metabolite exchange, shaping community structure under a given energy regime. Eco-thermodynamic gradients, the availability of electron donors (e.g. reduced sulphur compounds) and acceptors (e.g. NO_3^-, NO_2^-, N_2O) within the physical environment, such as those observed along redox gradients in Saanich Inlet, appear to govern N and S-based metabolic interactions. While direct evidence of metabolic coupling is difficult to obtain in un-cultured systems, particularly so for dissimilatory processes, both conceptual and mathematical models support such interactions [79, 132, 204, 249] and suggest energy availability to be a strong organizing principal for microbial community structure and metabolic processes.

The exploration of metabolic interactions along eco-thermodynamic gradients within this thesis suggests different motifs for metabolic interactions exist under different energy regimes (Figure 6.1). Within the energy replete, highly oxidized, surface ocean, a model of public-goods is proposed where one species makes a product that is released into the environment and used by many other species. For example, cyanobacteria release vitamin-B12 that is used by many other bacteria as an enzyme co-factor [91]. Within energy deficient, reduced, deep-sea sediments a model of discrete exchange is proposed where metabolites are passed directly between two species. For example, methane oxidizing Archaea pass electrons to sulfate reducing bacteria [260, 261]. Within moderately reduced environments, such as marine oxygen minimum zones and tidal sediments [262], a model of selective exchange is proposed where metabolites are shared among a limited number of taxonomic groups. For example NO_2^- produced by SUP05-group taken up by *Planctomycetes* and *Nitrospira* for anammox and nitrification respectively [132]. Taken together, these observations point to a continuum of metabolite exchange from public goods through selective exchange to discrete exchange along an eco-thermodynamic gradient from oxic to reduced environments. I further construct a hypothesis where, as energy available to the community decreases (in the form of Gibbs free energy between electron donors and acceptors)

metabolic couplings become increasingly more discrete as the specificity of each interaction is under greater and greater selective pressure to optimize energetic gain from individual metabolite exchanges over welfare of the community as a whole.

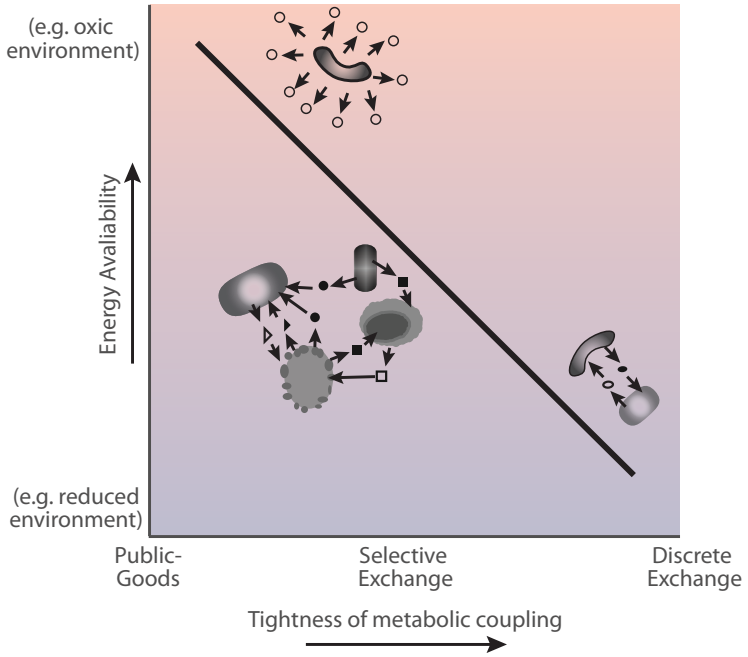


Figure 6.1: Smotifs for metabolic interactions. Diagram showing various motifs for metabolic interactions along an energy gradient.

6.5 Closing

As technological and analytical tools have advanced to the point of providing sequence information with taxonomic association from a diverse range of environments, it is apparent that microbial communities are truly the engines that drive Earth's biogeochemical cycles [38]. Within OMZs, the microbial community and associated biogeochemical cycles play key roles in nitrogen, sulfur and carbon cycling on a global scale. The redox gradients within OMZs offer an opportunity to chart how decreases in energy availability, in the form of high-energy electron acceptors, shape the microbial community and metabolic interactions carrying out these processes. Throughout this thesis, Saanich Inlet serves as a model OMZ to study the microbial communities along

these redox gradients and provides a crucial framework for the development of multi-omics approaches to studying microbial communities over spatial, temporal and energetic gradients. Findings from Saanich Inlet and Global Ocean genomic surveys in this thesis reveal the impacts of energy availability on processes that are critical to climate change on our planet, including greenhouse gas production/consumption, loss of biologically available nitrogen, and dark carbon fixation. As O_2 concentrations in the Global Ocean continue to decease, causing OMZ expansion and intensification, this thesis provides an essential knowledge base of the dominant microbial players and processes, providing important information for global modeling and environmental monitoring efforts. Further, this thesis builds hypotheses about the influence of energy availability on microbial metabolic interactions, offering important insights into factors that may shape the nature of interactions within microbial communities along eco-thermodynamic gradients. As more information about microbial communities and metabolism is revealed, we come to see a profound connectedness at the smallest levels of life, bringing home the necessity of community interactions in order to carry out global processes.

Bibliography

[1] P. K. Weyl. On the oxygen supply of the deep pacific ocean. *Limnology and Oceanography*, 10(2):215–219, 1965.

[2] J. J. Wright, K. M. Konwar, and S. J. Hallam. Microbial ecology of expanding oxygen minimum zones. *Nat Rev Microbiol*, 10(6):381–94, 2012.

[3] R. E. Keeling, A. Kortzinger, and N. Gruber. Ocean deoxygenation in a warming world. *Ann Rev Mar Sci*, 2:199–229, 2010.

[4] O. Ulloa, D. E. Canfield, E. F. DeLong, R. M. Letelier, and F. J. Stewart. Microbial oceanography of anoxic oxygen minimum zones. *Proc Natl Acad Sci U S A*, 109(40):15996–6003, 2012.

[5] R. J. Diaz and R. Rosenberg. Spreading dead zones and consequences for marine ecosystems. *Science*, 321(5891):926–9, 2008.

[6] A. Paulmier and D. Ruiz-Pino. Oxygen minimum zones (omzs) in the modern ocean. *Progress in Oceanography*, 80(3-4):113–128, 2009.

[7] Douglas G. Capone and David A. Hutchins. Microbial biogeochemistry of coastal upwelling regimes in a changing ocean. *Nature Geoscience*, 6, 2013.

[8] E. Zaikova, D. A. Walsh, C. P. Stilwell, *et al.* Microbial community dynamics in a seasonally anoxic fjord: Saanich inlet, british columbia. *Environ Microbiol*, 12(1):172–91, 2010.

[9] Osvaldo Ulloa, Jody J. Wright, Lucy Belmar, and Steven J. Hallam. Pelagic oxygen minimum zone microbial communities. pages 113–122, 2013.

[10] F. A. Whitney, H. J. Freeland, and M. Robert. Persistently declining oxygen levels in the interior waters of the eastern subarctic pacific. *Progress in Oceanography*, 75(2):179–199, 2007.

[11] M. C. Long, C. Deutsch, and T. Ito. Finding forced trends in oceanic oxygen. *Global Biogeochemical Cycles*, 30(2):381–397, 2016.

[12] P. Lam and M. M. Kuypers. Microbial nitrogen cycling processes in oxygen minimum zones. *Ann Rev Mar Sci*, 3:317–45, 2011.

[13] A. H. Devol and H. E. Hartnett. Role of the oxygen-deficient zone in transfer of organic carbon to the deep ocean. *Limnology and Oceanography*, 46(7):1684–1690, 2001.

[14] D. Woebken, B. M. Fuchs, M. M. M. Kuypers, and R. Amann. Potential interactions of particle-associated anammox bacteria with bacterial and archaeal partners in the namibian upwelling system. *Applied and Environmental Microbiology*, 73(14):4648–4657, 2007.

[15] S. Ganesh, L. A. Bristow, M. Larsen, *et al.* Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. *ISME J*, 2015.

[16] S. Ganesh, D. J. Parris, E. F. DeLong, and F. J. Stewart. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *ISME J*, 8(1):187–211, 2014.

[17] H. Stevens and O. Ulloa. Bacterial diversity in the oxygen minimum zone of the eastern tropical south pacific. *Environmental Microbiology*, 10(5):1244–1259, 2008.

[18] Frank J. Stewart, Osvaldo Ulloa, and Edward F. DeLong. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environmental Microbiology*, pages no–no, 2011.

[19] C. Rinke, P. Schwientek, A. Sczyrba, *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–7, 2013.

[20] R. M. Morris, C. D. Frazar, and C. A. Carlson. Basin-scale patterns in the abundance of sar11 subclades, marine actinobacteria (om1), members of the roseobacter clade and ocs116 in the south atlantic. *Environ Microbiol*, 14(5):1133–44, 2012.

[21] J. Viklund, T. J. Ettema, and S. G. Andersson. Independent genome reduction and phylogenetic reclassification of the oceanic sar11 clade. *Mol Biol Evol*, 29(2):599–615, 2012.

[22] S. J. Giovannoni, L. Bibbs, J. C. Cho, *et al.* Proteorhodopsin in the ubiquitous marine bacterium sar11. *Nature*, 438(7064):82–5, 2004.

[23] J. Sun, L. Steindler, J. C. Thrash, *et al.* One carbon metabolism in sar11 pelagic marine bacteria. *PLoS One*, 6(8):e23973, 2011.

[24] R. R. Malmstrom, R. P. Kiene, M. T. Cottrell, and D. L. Kirchman. Contribution of sar11 bacteria to dissolved dimethylsulfoniopropionate and amino acid uptake in the north atlantic ocean. *Appl Environ Microbiol*, 70(7):4129–35, 2004.

[25] D. Tsementzi, J. Wu, S. Deutsch, *et al.* Sar11 bacteria linked to ocean anoxia and nitrogen loss. *Nature*, 536:179–183, 2016.

[26] K. T. Marshall and R. M. Morris. Isolation of an aerobic sulfur oxidizer from the sup05/arctic96bd-19 clade. *ISME J*, 7(2):452–5, 2013.

[27] C. S. Sheik, S. Jain, and G. J. Dick. Metabolic flexibility of enigmatic sar324 revealed through metagenomics and metatranscriptomics. *Environ Microbiol*, 16(1):304–17, 2014.

[28] S. Lücker, B. Nowka, T. Rattei, E. Spieck, and H. Daims. The genome of nitrospina gracilis illuminates the metabolism and evolution of the major marine nitrite oxidizer. *Front Microbiol*, 4:27, 2013.

[29] B. Kartal, M. M. Kuypers, G. Lavik, *et al.* Anammox bacteria disguised as denitrifiers: nitrate reduction to dinitrogen gas via nitrite and ammonium. *Environ Microbiol*, 9(3):635–42, 2007.

[30] C. A. Francis, K. J. Roberts, J. M. Beman, A. E. Santoro, and B. B. Oakley. Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci U S A*, 102(41):14683–8, 2005.

[31] E. Allers, J. J. Wright, K. M. Konwar, *et al.* Diversity and population structure of marine group a bacteria in the northeast subarctic pacific ocean. *ISME J*, 7(2):256–68, 2013.

[32] J. J. Wright, K. Mewis, N. W. Hanson, *et al.* Genomic properties of marine group a bacteria indicate a role in the marine sulfur cycle. *ISME J*, 8(2):455–68, 2014.

[33] D. A. Walsh, E. Zaikova, C. G. Howes, *et al.* Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science*, 326(5952):578–82, 2009.

[34] V. Shah, B. X. Chang, and R. M. Morris. Cultivation of a chemoautotroph from the sup05 clade of marine bacteria that produces nitrite and consumes ammonium. *ISME J*, 2016.

[35] R. Knight, A. Vrbanac, B. C. Taylor, *et al.* Best practices for analysing microbiomes. *Nature Reviews Microbiology*, 16(7):410–422, 2018.

[36] C. M. Callbeck, G. Lavik, T. G. Ferdelman, *et al.* Oxygen minimum zone cryptic sulfur cycling sustained by offshore transport of key sulfur oxidizing bacteria. *Nature Communications*, 9(1), 2018.

[37] M. Labrenz, J. Grote, K. Mammitzsch, *et al.* Sulfurimonas gotlandica sp. nov., a chemoautotrophic and psychrotolerant epsilonproteobacterium isolated from a pelagic redoxcline, and an emended description of the genus sulfurimonas. *Int J Syst Evol Microbiol*, 63(Pt 11):4141–8, 2013.

[38] P. G. Falkowski, T. Fenchel, and E. F. Delong. The microbial engines that drive earth's biogeochemical cycles. *Science*, 320(5879):1034–9, 2008.

[39] J. P. Zehr and R. M. Kudela. Nitrogen cycle of the open ocean: from genes to ecosystems. *Ann Rev Mar Sci*, 3:197–225, 2011.

[40] J. A. Sohm, E. A. Webb, and D. G. Capone. Emerging patterns of marine nitrogen fixation. *Nat Rev Microbiol*, 9(7):499–508, 2011.

[41] H. Farnelid, M. Bentzon-Tilia, A. F. Andersson, *et al.* Active nitrogen-fixing heterotrophic bacteria at and below the chemocline of the central baltic sea. *ISME J*, 7(7):1413–23, 2013.

[42] T. Grosskopf, W. Mohr, T. Baustian, *et al.* Doubling of marine dinitrogen-fixation rates based on direct measurements. *Nature*, 488(7411):361–4, 2012.

[43] C. B. Field, M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science*, 281(5374):237–240, 1998.

[44] C. Fernandez, L. Farías, and O. Ulloa. Nitrogen fixation in denitrified marine waters. *PlosOne*, 6(6):e20539, 2011.

[45] C. R. Loescher, T. Grosskopf, F. D. Desai, *et al.* Facets of diazotrophy in the oxygen minimum zone waters off peru. *ISME J*, 8(11):2180–92, 2014.

[46] S. Bonnet, J. Dekaezemacker, K. A. Turk-Kubo, *et al.* Aphotic n2 fixation in the eastern tropical south pacific ocean. *PLOS ONE*, 8(12):1–14, 12 2013.

[47] E. Costa, J. Perez, and J. U. Kreft. Why is metabolic labour divided in nitrification? *Trends Microbiol*, 14(5):213–9, 2006.

[48] S. N. Merbt, D. A. Stahl, E. O. Casamayor, *et al.* Differential photoinhibition of bacterial and archaeal ammonia oxidation. *FEMS Microbiol Lett*, 327(1):41–6, 2012.

[49] C. Wuchter, B. Abbas, M. J. Coolen, *et al.* Archaeal nitrification in the ocean. *Proc Natl Acad Sci U S A*, 103(33):12317–22, 2006.

[50] M. Konneke, A. E. Bernhard, J. R. de la Torre, *et al.* Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature*, 437(7058):543–6, 2005.

[51] C. A. Schleper, R. V. Swanson, E. J. Mathur, and E. F. DeLong. Characterization of a dna polymerase from the uncultivated psychrophilic archaeon cenarchaeum symbiosum. *J Bacteriol*, 179(24):7803–7811, 1997.

[52] S. J. Hallam, T. J. Mincer, C. Schleper, *et al.* Pathways of carbon assimilation and ammonia oxidation suggested by environmental genomic analyses of marine crenarchaeota. *PLoS Biol*, 4(4):e95, 2006.

[53] N. Vajrala, W. Martens-Habbena, L. A. Sayavedra-Soto, *et al.* Hydroxylamine as an intermediate in ammonia oxidation by globally abundant marine archaea. *Proc Natl Acad Sci U S A*, 110(3):1006–1011, 2013.

[54] J. Füssel, S. Lücker, P. Yilmaz, *et al.* Adaptability as the key to success for the ubiquitous marine nitrite oxidizer nitrococcus. *Science Advances*, 3(11), 2017.

[55] S. Lucker, M. Wagner, F. Maixner, *et al.* A nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci U S A*, 107(30):13479–84, 2010.

[56] J. Füssel, P. Lam, G. Lavik, *et al.* Nitrite oxidation in the namibian oxygen minimum zone. *ISME J*, 6(6):1200–9, 2012.

[57] D. E. Canfield, A. N. Glazer, and P. G. Falkowski. The evolution and future of earth's nitrogen cycle. *Science*, 330(6001):192–6, 2010.

[58] J. Simon and M. G. Klotz. Diversity and evolution of bioenergetic systems involved in microbial nitrogen compound transformations. *Biochim Biophys Acta*, 1827(2):114–35, 2013.

[59] R. A. Sanford, D. D. Wagner, Q. Wu, *et al.* Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils. *Proc Natl Acad Sci U S A*, 109(48):19709–14, 2012.

[60] B. B. Ward, A. H. Devol, J. J. Rich, *et al.* Denitrification as the dominant nitrogen loss process in the arabian sea. *Nature*, 461(7260):78–81, 2009.

[61] P. Lam, G. Lavik, M. M. Jensen, *et al.* Revising the nitrogen cycle in the peruvian oxygen minimum zone. *Proceedings of the National Academy of Sciences*, 106(12):4752–4757, 2009.

[62] P. Lam, M. M. Jensen, G. Lavik, *et al.* Linking crenarchaeal and bacterial nitrification to anammox in the black sea. *Proc Natl Acad Sci U S A*, 104(17):7104–9, 2007.

[63] A. Jayakumar, G. D. OMullan, S. W. A. Naqvi, and B. B. Ward. Bacterial community composition changes associated with stages of denitrification in oxygen minimum zones. *Microbial Ecology*, 52(2):350–626, 2009.

[64] C. G. Bruckner, K. Mammitzsch, G. Jost, *et al.* Chemolithoautotrophic denitrification of epsilonproteobacteria in marine pelagic redox gradients. *Environ Microbiol*, 15(5):1505–13, 2013.

[65] D. Woebken, P. Lam, M. M. M. Kuypers, *et al.* A microdiversity study of anammox bacteria reveals a novel candidatusscalindua phylotype in marine oxygen minimum zones. *Environmental Microbiology*, 10(11):3106–3119, 2008.

[66] M. Strous, E. Pelletier, S. Mangenot, *et al.* Deciphering the evolution and metabolism of an anammox bacterium from a community genome. *Nature*, 440(7085):790–4, 2006.

[67] M. S. Jetten, Lv Niftrik, M. Strous, *et al.* Biochemistry and molecular biology of anammox bacteria. *Crit Rev Biochem Mol Biol*, 44(2-3):65–84, 2009.

[68] T. Kalvelage, M. M. Jensen, S. Contreras, *et al.* Oxygen sensitivity of anammox and coupled n-cycle processes in oxygen minimum zones. *PLoS One*, 6(12):e29299, 2011.

[69] J. B. Kirkpatrick, C. A. Fuchsman, E. Yakushev, J. T. Staley, and J. W. Murray. Concurrent activity of anammox and denitrifying bacteria in the black sea. *Front Microbiol*, 3:256, 2012.

[70] B. Thamdrup and T. Dalsgaard. Production of n2 through anaerobic ammonium oxidation coupled to nitrate reduction in marine sediments. *Applied and Environmental Microbiology*, 68(3):1312–1318, 2002.

[71] M. A. Azhar, D. E. Canfield, K. Fennel, B. Thamdrup, and C. J. Bjerrum. A model-based insight into the coupling of nitrogen and sulfur cycles in a coastal upwelling system. *Journal of Geophysical Research: Biogeosciences*, 119:264–285, 2014.

[72] M. Voss and J. P. Montoya. Nitrogen cycle: Oceans apart. *Nature*, 461:49–50, 2009.

[73] P. Lam, M. M. Jensen, A. Kock, *et al.* Origin and fate of the secondary nitrite maximum in the arabian sea. *Biogeosciences*, 8(6):1565–1577, 2011.

[74] W. Ghosh and B. Dam. Biochemistry and molecular biology of lithotrophic sulfur oxidation by taxonomically and ecologically diverse bacteria and archaea. *FEMS Microbiol Rev*, 33(6):999–1043, 2009.

[75] G. Lavik, T. Stuhrmann, V. Bruchert, *et al.* Detoxification of sulphidic african shelf waters by blooming chemolithotrophs. *Nature*, 457(7229):581–4, 2009.

[76] S. Glaubitz, K. Kiesslich, C. Meeske, M. Labrenz, and K. Jurgens. Sup05 dominates the gammaproteobacterial sulfur oxidizer assemblages in pelagic redoxclines of the central baltic and black seas. *Appl Environ Microbiol*, 79(8):2767–76, 2013.

[77] B. K. Swan, M. Martinez-Garcia, C. M. Preston, *et al.* Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science*, 333(6047):1296–300, 2011.

[78] D. E. Canfield, F. J. Stewart, B. Thamdrup, *et al.* A cryptic sulfur cycle in oxygen-minimum-zone waters off the chilean coast. *Science*, 330(6009):1375–8, 2010.

[79] A K. Hawley, H.M. Brewer, A. D. Norbeck, L. Paa-Toli c, and S. J. Hallam. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proc Natl Acad Sci U S A*, 11(31):11395–11400, 2014.

[80] C. B. Walker, J. R. de la Torre, M. G. Klotz, *et al.* Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci U S A*, 107(19):8818–8823, 2010.

[81] Y. I. Sorokin, P. Y Sorokin, V. A. Avdeev, D. Y. Sorokin, and S. V. Ilchenkol. Biomass, production and activity of bacteria in the black sea, with special reference to chemosynthesis and the sulfur cycle. *Hydrobiologia*, 308(1):61–76, 1995.

[82] G. Jost, M. V. Zubkov, E. Yakushev, M. Labrenz, and K. Jrgens. High abundance and dark co2 fixation of chemolithoautotrophic prokaryotes in anoxic waters of the baltic sea. *Limnology and Oceanography*, 53(1):14–22, 2008.

[83] B. B. Ward, H. E. Glover, and F. Lipschultz. Chemoautotrophic activity and nitrification in the oxygen minimum zone off peru. *Deep Sea Research*, 36(7):1031–1051, 1989.

[84] G. T. Taylor, M. Iabichella, T. Ho, *et al.* Chemoautotrophy in the redox transition zone of the cariaco basin: A significant midwater source of organic carbon production. *Limnology and Oceanography*, 46(1):149–163, 2001.

[85] S. Louca, M. P. Polz, F. Mazel, *et al.* Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, pages 2397–334X, 2018.

[86] E. F. DeLong. Microbial community genomics in the ocean. *Nat Rev Microbiol*, 3(6):459–69, 2005.

[87] DeLongE. F., C. M. Preston, T. Mincer, *et al.* Community genomics amoung stratified microbial assemblages in the ocean's interior. *Science*, 331, 2006.

[88] J. A. Fuhrman. Microbial community structure and its functional implications. *Nature*, 459(7244):193–9, 2009.

[89] S. L. Strom. Microbial ecology of ocean biogeochemistry: a community of perspective. *Science*, 320(5879):1043–4045, 2008.

[90] B. E. Morris, R. Henneberger, H. Huber, and C. Moissl-Eichinger. Microbial syntrophy: interaction for the common good. *FEMS Microbiol Rev*, 37(3):384–406, 2013.

[91] S. J. Giovannoni. Vitamins in the sea. *Proc Natl Acad Sci U S A*, 109(35):13888–9, 2012.

[92] J. A. Fuhrman, J. A. Cram, and D. M. Needham. Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol*, 13(3):133–46, 2015.

[93] M. T. Mee, J. J. Collins, G. M. Church, and H. H. Wang. Syntrophic exchange in synthetic microbial communities. *Proc Natl Acad Sci U S A*, 111(20):E2149–E2156, 2014.

[94] N. W. Hanson, K. M. Konwar, A. K. Hawley, *et al.* Metabolic pathways for the whole community. *BMC Genomics*, 15(619), 2014.

[95] S. J. Hallam and J. P. McCutcheon. Microbes don't play solitaire: how cooperation trumps isolation in the microbial world. *Environ Microbiol Rep*, 7(1):26–8, 2015.

[96] K. Anantharaman, C. T. Brown, L.A. Hug, *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, 7:13219, 2016.

[97] J. P. McCutcheon and N. A. Moran. Functional convergence in reduced genomes of bacterial symbionts spanning 200 my of evolution. *Genome Biol Evol*, 2, 2010.

[98] S. Scheller, H. Yu, G. L. Chadwick, S. E. McGlynn, and V. J. Orphan. Artificial electron acceptors decouple archaeal methane oxidation from sulfate reduction. *Science*, 351(6274):703–707, 2016.

[99] J. J. Morris, Z. I. Johnson, M. J. Szul, M. Keller, and E. R. Zinser. Dependence of the cyanobacterium prochlorococcus on hydrogen peroxide scavenging microbes for growth at the ocean's surface. *PLoS One*, 6(2):e16805, 2011.

[100] S. J. Giovannoni, H. J. Tripp, S. Givan, *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–5, 2005.

[101] J. J. Morris, R. E. Lenski, and E. R. Zinserc. The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *MBio*, 3(2):e00036–12, 2012.

[102] E. F. DeLong. Life on the thermodynamic edge. *Science*, 317:327–328, 2007.

[103] M. K. Nobu, H. Tamaki, K. Kubota, and W. T. Liu. Metagenomic characterization of 'candidatus defluviicoccus tetraformis strain tfo71', a tetrad-forming organism, predominant in an anaerobic-aerobic membrane bioreactor with deteriorated biological phosphorus removal. *Environ Microbiol*, 16(9):2739–51, 2014.

[104] S. J. Giovannoni and K. L. Vergin. Seasonality in ocean microbial communities. *Science*, 335(6069):671–676, 2012.

[105] R. Stepanauskas and M. E. Sieracki. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci U S A*, 104(21):9052–7, 2007.

[106] A. S. Hahn, K. M. Konwar, S. Louca, N. W. Hanson, and S. J. Hallam. The information science of microbial ecology. *Curr Opin Microbiol*, 31:209–16, 2016.

[107] R. H. Herlinveaux. Oceanography of saanich inlet in vancouver island, british columbia. *Journal of the Fisheries Research Board of Canada*, 19:1–37, 1962.

[108] M. D. Lilley, J. A. Baross, and L. I. Gordon. Dissolved hydrogen and methane in saanich inlet, british columbia. *Deep Sea Research Part A. Oceanographic Research Papers*, 29(1471):1484, 1982.

[109] B. B. Ward and K. A. Kilpatrick. Relationship between substrate concentration and oxidation of ammonium and methane in a stratified water column. *Continental Shelf Research*, 10:1193–1208, 1990.

[110] N. M. Carter. The oceanography of the fjords of southern british columbia. *Fish. Res. Bd. Canada Prog. Rept. Pacific Coast Sta.*, 12:7–11, 1932.

[111] N. M. Carter. Physiography and oceanography of some british columbia fjords. *Proc. Fifth. Pacific Sci. Cong.*, 1:721, 1934.

[112] J. J. Anderson and A. H. Devol. Deep water renewal in saanich inlet, an intermittently anoxic basin. *Estuarine and Coastal and Marine Science*, 1:1–10, 1973.

[113] D. W. Capelle, A. K. Hawley, S. J. Hallam, and P. D. Tortell. A multi-year time-series of n2o dynamics in a seasonally anoxic fjord: Saanich inlet, british columbia. *Limnology and Oceanography*, 63(2):524–539, 2017.

[114] A K. Hawley, M. Torres Beltrán, M. P. Bhatia, *et al.* A compendium of water column multi-omic sequence information from a seasonally anoxic fjord saanich inlet. *Scientific Data*, submitted, 2017.

[115] M. Torres-Beltrán, A. K. Hawley, D. Capelle, *et al.* A compendium of water column chemistry from the seasonally anoxic fjord saanich inlet. *Scientific Data*, Submitted, 2017.

[116] Osvaldo Ulloa and Silvio Pantoja. The oxygen minimum zone of the eastern south pacific. *Deep Sea Research Part II: Topical Studies in Oceanography*, 56(16):987–991, 2009.

[117] M. Sunamura, Y. Higashi, C. Miyako, J. Ishibashi, and A. Maruyama. Phylotypes are predominant in the suiyo seamount hydrothermal plume. *Appl Environ Microbiol*, 70(2):1190–1198, 2004.

[118] I. L. Newton, T. Woyke, T. A. Auchtung, *et al.* The calyptogena magnifica chemoautotrophic symbiont genome. *Science*, 315(5814):998–1000, 2007.

[119] M. Harada, T. Yoshida, H. Kuwahara, *et al.* Expression of genes for sulfur oxidation in the intracellular chemoautotrophic symbiont of the deep-sea bivalve calyptogena okutanii. *Extremophiles*, 13(6):895–903, 2009.

[120] S. Glaubitz, M. Labrenz, G. Jost, and K. Jurgens. Diversity of active chemolithoautotrophic prokaryotes in the sulfidic zone of a black sea pelagic redoxcline as determined by rrna-based stable isotope probing. *FEMS Microbiol Ecol*, 74(1):32–41, 2010.

[121] C. A. Fuchsman, J. B. Kirkpatrick, W. J. Brazelton, J. W. Murray, and J. T. Staley. Metabolic strategies of free-living and aggregate-associated bacterial communities inferred from biologic and chemical profiles in the black sea suboxic zone. *FEMS Microbiol Ecol*, 78(3):586–603, 2011.

[122] K. Anantharaman, J. A. Breier, C. S. Sheik, and G. J. Dick. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. *Proc Natl Acad Sci U S A*, 110(1):330–335, 2012.

[123] J. Schmidtova, S. J. Hallam, and S. A. Baldwin. Phylogenetic diversity of transition and anoxic zone bacterial communities within a near-shore anoxic basin: Nitinat lake. *Environ Microbiol*, 11(12):3233–51, 2009.

[124] R. A. Lesniewski, S. Jain, K. Anantharaman, P. D. Schloss, and G. J. Dick. The metatranscriptome of a deep-sea hydrothermal plume is dominated by water column methanotrophs and lithotrophs. *ISME J*, 6(12):2257–68, 2012.

[125] B. J. Baker, C. S. Sheik, C. A. Taylor, *et al.* Community transcriptomic assembly reveals microbes that contribute to deep-sea carbon and nitrogen cycling. *ISME J*, 7(10):1962–73, 2013.

[126] D. J. Richardson, B. C. Berks, D. A. Russell, S. Spiro, and C. J. Taylor. Functional, biochemical and genetic diversity of prokaryotic nitrate reductase. *Cellular and Molecular Life Sciences*, 58:165–178, 2001.

[127] V. Stewart, Y. Lu, and A. J. Darwin. Periplasmic nitrate reductase (napabc enzyme) supports anaerobic respiration by escherichia coli k-12. *Journal of Bacteriology*, 184:1314–1323, 2002.

[128] M.A. Moran. The global ocean microbiome. *Science*, 350(6266):aac8455, 2015.

[129] P. G. Falkowski, T. Algeo, L. Codispoti, *et al.* Ocean deoxygenation: Past, present, and future. *EOS, Trans AGU*, 92(46):409–410, 2011.

[130] J. M. Labonte, S. J. Hallam, and C. A. Suttle. Previously unknown evolutionary groups dominate the ssdna gokushoviruses in oxic and anoxic waters of a coastal marine environment. *Front Microbiol*, 6:315, 2015.

[131] C. E. Chow, D. M. Winget, 3rd White, R. A., S. J. Hallam, and C. A. Suttle. Combining genomic sequencing methods to explore viral diversity and reveal potential virus-host interactions. *Front Microbiol*, 6:265, 2015.

[132] S. Louca, A. K. Hawley, S. Katsev, *et al.* Integrating biogeochemistry with multi-omic sequence information in a model oxygen minimum zone. *Proc Natl Acad Sci U S A*, In press, 2016.

[133] M. A. Moran, B. Satinsky, S. M. Gifford, *et al.* Sizing up metatranscriptomics. *ISME J*, 7(2):237–43, 2013.

[134] F. J. Stewart, O. Ulloa, and E. F. DeLong. Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol*, 14(1):23–40, 2012.

[135] F. A. J. Armstrong, C. R. Stearns, and J. D. H. Strickland. The measurement of upwelling and subsequent biological process by means of the technicon autoanalyzer and associated equipment. *Deep Sea Research and Oceanographic Abstracts*, 14:381–389, 1967.

[136] R. M. Holmes, A. Aminot, R. Krouel, B. A. Hooker, and B. J. Peterson. A simple and precise method for measuring ammonium in marine and freshwater ecosystems. *Canadian Journal of Fisheries and Aquatic Sciences*, 59(10):1801–1808, 1999.

[137] J. D. Cline. Spectrophotometric determination of hydrogen sulfide in natural waters. *Limnology and Oceanography*, 14:454–458, 1969.

[138] J. Murphy and J. P. Riley. A modified single solution method for the determination of phosphate in natural waters. *Analytica Chemica Acta*, 27:31–36, 1962.

[139] L. W. Winkler. Die bestimmung des im wasser gelsten sauerstoffes. *Berichte der deutschen chemischen Gesellschaft*, 21:2843–2854, 1888.

[140] D. Capelle, J. Dacey, and P. D Tortell. An automated, high-throughput method for accurate and precise measurements of dissolved nitrous-oxide and methane concentrations in natural seawaters. *Limnology and Oceanography: Methods*, in review, 2015.

[141] X. J. Lin, M. I. Scranton, A. Y. Chistoserdov, R. Varela, and G. T Taylor. Spatiotemporal dynamics of bacterial populations in the anoxic cariaco basin. *Limnology and Oceanography*, 53(1):37–51, 2008.

[142] X. J. Lin, M. I Scranton, R. Varela, A. Chistoserdov, and G. T. Taylor. Compositional responses of bacterial communities to redox gradients and grazing in the anoxic cariaco basin. *Aquatic Microbial Ecology*, 47(1):57–72, 2007.

[143] M. J. Rodriguez-Mora, M. I. Scranton, G. T. Taylor, and A. Y. Chistoserdov. Bacterial community composition in a large marine anoxic basin: a cariaco basin time-series survey. *FEMS Microbiol Ecol*, 84(3):625–39, 2013.

[144] C. Vetriani, H. V. Tran, and L. J. Kerkhof. Fingerprinting microbial assemblages from the oxic/anoxic chemocline of the black sea. *Applied and environmental microbiology*, 69(11):6481–6488, 2003.

[145] V. Edgcomb, W. Orsi, C. Leslin, *et al.* Protistan community patterns within the brine and halocline of deep hypersaline anoxic basins in the eastern mediterranean sea. *SExtremophiles*, 13(1):151–167, 2009.

[146] B. M. Fuchs, D. Woebken, M. V. Zubkov, P. Burkill, and R. Amann. Molecular identification of picoplankton populations in contrasting waters of the arabian sea. *Aquatic Microbial Ecology*, 39(2):145–157, 2005.

[147] W. Orsi, Y. C. Song, S. Hallam, and V. Edgcomb. Effect of oxygen minimum zone formation on communities of marine protists. *The ISME journal*, 6(8):1586–601, 2012.

[148] T. Stoeck, B. Hayward, G. T. Taylor, R. Varela, and S. S. Epstein. A multiple pcr-primer approach to access the microeukaryotic diversity in environmental samples. *Protist*, 157(1):31–43, 2006.

[149] D. A. Walsh and S. J. Hallam. *Bacterial community structure and dynamics in a sea- sonally anoxic fjord: Saanich Inlet, British Columbia*, pages 253–267. Wiley-Blackwell, Hoboken, NJ, 2011.

[150] D. P. Herlemann, M. Labrenz, K. Jurgens, *et al.* Transitions in bacterial communities along the 2000 km salinity gradient of the baltic sea. *ISME J*, 5(10):1571–1579, 2011.

[151] T. Stoeck, D. Bass, M. Nebel, *et al.* Multiple marker parallel tag environmental dna sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol*, 19, 2010.

[152] T. Stoeck, A. Behnke, R. Christen, *et al.* Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biol*, 7:72, 2009.

[153] S. Wakeham, R. Amann, K. Freeman, *et al.* Microbial ecology of the stratified water column of the black sea as revealed by a comprehensive biomarker study. *Organic Geochemistry*, 38(12):2070–2097, 2007.

[154] S. Glaubitz, T. Lueders, W. R. Abraham, *et al.* 13c-isotope analyses reveal that chemolithoautotrophic gamma- and epsilonproteobacteria feed a microbial food web in a pelagic redoxcline of the central baltic sea. *Environ Microbiol*, 11(2):326–37, 2009.

[155] J. A. Bryant, F. J. Stewart, J. M. Eppley, and E. F. DeLong. Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. *Ecology*, 93(7):1659–1673, 2012.

[156] T. J. Mincer, M. J. Church, L. T. Taylor, *et al.* Quantitative distribution of presumptive archaeal and bacterial nitrifiers in monterey bay and the north pacific subtropical gyre. *Environmental Microbiology*, 9(5):1162–1175, 2007.

[157] M. T. Suzuki, L. T. Taylor, and E. F. DeLong. Quantitative analysis of. *Appl Environ Microbiol*, 66(11):4605–4614, 2000.

[158] K. Takai and K. Horikoshi. Rapid detection and quantification of members of the archaeal community by quantitative pcr using fluorogenic probes. *Applied and environmental microbiology*, 2000(66):11, 2000.

[159] D. A. Walsh, E. Zaikova, and S. J. Hallam. Small volume (1-3l) filtration of coastal seawater samples. *JoVE*, e1163, 2009.

[160] J. J. Wright, S. Lee, E. Zaikova, D. A. Walsh, and S. J. Hallam. Dna extraction from 0.22 micron sterivex filters and cesium chloride density gradient centrifugation. *Journal of Vissualized Experiments*, page e1352, 2009.

[161] J. A. Cram, C. E. Chow, R. Sachdeva, *et al.* Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *ISME J*, 9(3):563–80, 2015.

[162] J. Tremblay, K. Singh, A. Fern, *et al.* Primer and platform effects on 16s rrna tag sequencing. *Front Microbiol*, 6:771, 2015.

[163] C. Daum, J. Han, M. Zane, *et al.* Illumina ga iix & hiseq 2000 production sequencing and qc analysis pipelines at the doe joint genome institute (advances of genome biology and technology meeting 2011), 2011.

[164] Y. Shi, G. W. Tyson, and E. F. DeLong. Metatranscriptomics reveals unique microbial small rnas in the ocean's water column. *Nature*, 459(7244):266–9, 2009.

[165] Rachna J. Ram, Nathan C. VerBerkmoes, Michael P. Thelen, *et al.* Community proteomics of a natural microbial biofilm. *Science*, 208:1915–1920, 2005.

[166] K. M. Keiblinger, I. C. Wilhartitz, T. Schneider, *et al.* Soil metaproteomics - comparative evaluation of protein extraction protocols. *Soil Biol Biochem*, 54(150-10):14024, 2012.

[167] T. Schneider, K. M. Keiblinger, E. Schmid, *et al.* Who is who in litter decomposition? metaproteomics reveals major microbial players and their biogeochemical functions. *The ISME journal*, 6(9):1749–62, 2012.

[168] N. Delmotte, C. Knief, S. Chaffron, *et al.* Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc Natl Acad Sci U S A*, 106(38):16428–33, 2009.

[169] N. C. Verberkmoes, A. L. Russell, M. Shah, *et al.* Shotgun metaproteomics of the human distal gut microbiota. *The ISME journal*, 3(2):179–89, 2009.

[170] M. E. Guazzaroni, F. A.' Herbst, I. Lores, *et al.* Metaproteogenomic insights beyond bacterial response to naphthalene exposure and bio-stimulation. *ISME J*, 7(1):122–136, 2013.

[171] R. Kuhn, D. Benndorf, E. Rapp, *et al.* Metaproteome analysis of sewage sludge from membrane bioreactors. *Proteomics*, 11(13):2738–2744, 2011.

[172] P. Wilmes, M. Wexler, and P. L. Bond. Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLosObe*, 3(3):e1778, 2008.

[173] R. M. Morris, B. L. Nunn, C. Frazar, *et al.* Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J*, 4(5):673–85, 2010.

[174] S. M. Sowell, P.E. Abraham, M. Shah, *et al.* Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *The ISME Journal*, 5(5):856–865, 2011.

[175] T. J. Williams, E. Long, F. Evans, *et al.* A metaproteomic assessment of winter and summer bacterioplankton from antarctic peninsula coastal surface waters. *The ISME Journal*, 2012.

[176] S. Kim, N. Mischerikow, N. Bandeira, *et al.* The generating function of cid, etd and cid/etd pairs of tandem mass spectra: Applications to database search. *Molecular & Cellular Proteomics*, 9:2840–2852, 2010.

[177] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. Megan analysis of metagenomic data. *Genome Research*, 17(3):377–386, 2007.

[178] I. Letunic and P. Bork. Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–8, 2007.

[179] I. Letunic and P. Bork. Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*, 39(Web Server issue):W475–8, 2011.

[180] B. Zybailov, A. L. Mosley, M. E. Sardiu, *et al.* Statistical analysis of membrane proteome expression changes in saccharomyces cerevisiae. *Journal of proteome research*, 5:2339–2347, 2006.

[181] J. A. Dodsworth, P. C. Blainey, S. K. Murugapiran, *et al.* Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the op9 lineage. *Nature Communications*, 4:1854, 2013.

[182] K. M. Konwar, N. W. Hanson, A. P. Page, and S. J. Hallam. Metapathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics*, 14(202), 2013.

[183] D. Hyatt, G-L Chen, M. L. LoCascio, P.F.and Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(11):119, 2010.

[184] S. M. Kiełbasa, R. Wan, K Sato, P. Horton, and M. C. Frith. Adaptive seeds tame genomic sequence comparison. *Genome Research*, 21:487–493, 2011.

[185] Dongjae Kim, Aria S Hahn, Kishori M Hanson, Niels Wand Konwar, and Steven J Hallam. Fast: Fast annotation with synchronized threads. *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*, in press, 2016.

[186] D. A. Rasko, G. S. Myers, and Ravel J. Visualization of comparative genomic analyses by blast score ratio. *BMC Bioinformatics*, 6(2), 2005.

[187] S Okuda, T Yamada, M. Hamajima, *et al.* Kegg atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Research*, 36:W423, 2008.

[188] R. L. Tatusov, D. A. Natale, I. V. Garkavtsev, *et al.* The cog database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29(1):22, 2001.

[189] R. Caspi, T. Altman, K. Dreher, *et al.* The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 40(D1):D742–D753, 2012.

[190] H. Li and R. Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.

[191] K. M. Konwar, N. W. Hanson, M. P. Bhatia, *et al.* Metapathways v2.5: quantitative functional, taxonomic and usability improvements. *Bioinformatics*, 31(20):3345–7, 2015.

[192] N. Gruber and J. L. Sarmiento. Global patterns of marine nitrogen fixation and denitrification. *Global Biogeochemical Cycles*, 11(2):235–266, 1997.

[193] L. A. Codispoti, J. A. Brandes, J. P. Christensen, *et al.* The oceanic fixed nitrogen and nitrous oxide budgets: Moving targets as we enter the anthropocene? *Scientia Marina*, 65(2):85–105, 2001.

[194] C. Deutsch, J. L. Sarmiento, D. M. Sigman, N. Gruber, and J. P. Dunne. Spatial coupling of nitrogen inputs and losses in the ocean. *Nature*, 445(7124):163–7, 2007.

[195] T. E. Mattes, B. L. Nunn, K. T. Marshall, *et al.* Sulfur oxidizers dominate carbon fixation at a biogeochemical hot spot in the dark ocean. *ISME J*, 7(12):2349–60, 2013.

[196] R. E. Anderson, M. T. Beltran, S. J. Hallam, and J. A. Baross. Microbial community structure across fluid gradients in the juan de fuca ridge hydrothermal system. *FEMS Microbiol Ecol*, 83(2):324–39, 2013.

[197] J. M. Beman, J. Leilei Shih, and B. N. Popp. Nitrite oxidation in the upper water column and oxygen minimum zone of the eastern tropical north pacific ocean. *ISME J*, 7(11):2192–205, 2013.

[198] M. Hugler and S. M. Sievert. Beyond the calvin cycle: autotrophic carbon fixation in the ocean. *Ann Rev Mar Sci*, 3:261–89, 2011.

[199] H. Schunck, G. Lavik, D. K. Desai, *et al.* Giant hydrogen sulfide plume in the oxygen minimum zone off peru supports chemolithoautotrophy. *PLoS One*, 8(8):e68661, 2013.

[200] H. Körner, H. J. Sofia, and W. G. Zumft. Phylogeny of the bacterial superfamily of crp-fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol Rev*, 27(5):559–592, 2003.

[201] M. G. Klotz, M. C. Schmid, M. Strous, *et al.* Evolution of an octahaem cytochrome c protein family that is key to aerobic and anaerobic ammonia oxidation by bacteria. *Environ Microbiol*, 10(11):3150–63, 2008.

[202] L. Farías, C. Fernández, J. Fau Faúndez, M. Cornejo, and M. E. Alcaman. Chemolithoautotrophic production mediating the cycling of the greenhouse gases n2o and ch4 in an upwelling ecosystem. *Biogeosciences*, 6(3053-3069), 2009.

[203] A. E. Santoro, C. Buchwald, M. R. McIlvin, and K. L. Casciotti. Isotopic signature of n2o produced by marine ammonia-oxidizing archaea. *Science*, 333, 2011.

[204] D. C. Reed, C. K. Algar, J. A. Huber, and G. J. Dick. Gene-centric approach to integrating environmental genomics and biogeochemical models. *Proc Natl Acad Sci U S A*, 111(5):1879–84, 2014.

[205] K. Mavromatis, N. Ivanova, K. Barry, *et al.* Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods*, 4(6):495–500, 2007.

[206] A. K. Hawley, S. Kheirandish, A. Mueller, *et al.* Molecular tools for investigating microbial community structure and function in oxygen-deficient marine waters. *Methods Enzymol*, 531:305–29, 2013.

[207] S. Kim, N. Gupta, and P.A. Pevzner. Spectral probabilities and generating functions of tandem mass spectra: A strike against decoy databases. *J Proteome Res*, 7(8):33543363, 2008.

[208] J. Van de Vossenberg, D. Woebken, W. J. Maalcke, *et al.* The metagenome of the marine anammox bacterium candidatus scalindua profunda illustrates the versatility of this globally important nitrogen cycle bacterium. *Environmental Microbiology*, 15(5):12751289, 2013.

[209] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 216:403–410, 1990.

[210] N. Georgescu-Roegen. *The Entropy Law and the Economic Process*. Harvard University Press, Cambridge, MA, 1971.

[211] R. U. Ayres. Eco-thermodynamics: economics and the second law. *Ecological Economics*, 26(189-209), 1997.

[212] L. A. Hug, B. C. Thomas, I. Sharon, *et al.* Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ Microbiol*, 18(1):159–73, 2016.

[213] H. J. Tripp, J. B. Kitner, M. S. Schwalbach, *et al.* Sar11 marine bacteria require exogenous reduced sulphur for growth. *Nature*, 452(7188):741–4, 2008.

[214] F. O. Aylwarda, J. A M. Eppley, J. M. Smith, *et al.* Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proc Natl Acad Sci U S A*, 112(17):5443–5448, 2015.

[215] S. Louca, L. Wegener Parfrey, and M. Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(1272-1277), 2016.

[216] E. A. Gies, K. M. Konwar, J. T. Beatty, and S. J. Hallam. Illuminating microbial dark matter in meromictic sakinaw lake. *AEM*, 80(21):6807–6018, 2014.

[217] M. K. Nobu, T. Narihiro, C. Rinke, *et al.* Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. *ISME J*, 2015.

[218] N. Segata, D. Bornigen, X. C. Morgan, and C. Huttenhower. Phylophlan is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun*, 4:2304, 2013.

[219] S.J. Hallam, M. Torres Beltrán, and A.K. Hawley. Monitoring microbial responses to ocean deoxygenation in a model oxygen minimum zone. *Sci. Data*, 4:170158, 2017.

[220] O. Béjà, L. Aravind, E. V. Koonin, *et al.* Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science*, 289:1902–1906, 2000.

[221] S. M. Steinberg and J. L. Badal. Oxalic, glyoxalic and pyruvic acids in eastern pacific ocean waters. *Journal of Marine Research*, 42:697–708, 1984.

[222] V. Anantharam, M. J. Allison, and P. C. Maloney. 0xalate:formate exchange. *Journal of Biological Chemistry*, 264(13):7244–7250, 1989.

[223] C. Greening, A. Biswas, C. R. Carere, *et al.* Genomic and metagenomic surveys of hydrogenase distribution indicate h2 is a widely utilised energy source for microbial growth and survival. *ISME J.*, 10:761–777, 2016.

[224] S. Roux, A. K. Hawley, M. Torres Beltrán, *et al.* Ecology and evolution of viruses infecting uncultivated sup05 bacteria as revealed by single-cell and meta-genomics. *Elife*, 3:e03125, 2014.

[225] V. M. Markowitz, I. M. Chen, K. Palaniappan, *et al.* Img: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res*, 40(Database issue):D115–22, 2012.

[226] N. J. Varghese, S. Mukherjee, N. Ivanova, *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Res*, 43:6761–71, 2015.

[227] H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, and F. O. Glockner. Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol*, 6(9):938–47, 2004.

[228] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glockner. Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5:163, 2004.

[229] K. R. Clarke and R. N. Gorley. Primer v6: User manual/tutorial. 2006.

[230] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*, 25(7):1043–55, 2015.

[231] V. M. Markowitz, K. Mavromatis, N. N. Ivanova, *et al.* Img er: a system for microbial genome annotation expert review and curation. *Bioinformatics*, 25:2271–2278, 2009.

[232] R. C Edgar. Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26:2460246, 2010.

[233] S. Pesant, F. Not, M. Picheral, *et al.* Open science resources for the discovery and analysis of tara oceans data. *Sci Data*, 2:150023, 2015.

[234] IPCC. *Climate Change 2013: The Physical Sciences Basis. Contribution of Working Group I to the Fith Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2013.

[235] A. R. Ravishankara, J. S. Daniel, and R. W. Portmann. Nitrous oxide (n2o): The dominant ozone-depleting substance emitted in the 21st century. *Science*, 326(5949):123–125, 2009.

[236] S. W. A. Naqvi, H. W. Bange, L. Faras, *et al.* Marine hypoxia/anoxia as a source of ch4 and n2o. *Biogeosciences*, 7(7):2159–2190, 2010.

[237] S. W. Naqvi, D. A. Jayakumar, P. V. Narvekar, *et al.* Increased marine production of n2o due to intensifying anoxia on the indian continental shelf. *Nature*, 408(6810):346–9, 2000.

[238] S. Hallin, L. Philippot, F. E. Loffler, R. A. Sanford, and C. M. Jones. Genomics and ecology of novel n2o-reducing microorganisms. *Trends Microbiol*, 26(1):43–55, 2018.

[239] W. G. Zumft. Cell biology and molecular basis of denitrification. *Microbiol. Mol. Biol. Rev.*, 61:533616, 1997.

[240] T. Suenaga, S. Riya, M. Hosomi, and A. Terada. Biokinetic characterization and activities of n2o-reducing bacteria in response to various oxygen levels. *Front Microbiol*, 9:697, 2018.

[241] M. M. M. Kuypers, H. K. Marchant, and B. Kartal. The microbial nitrogen-cycling network. *Nat Rev Microbiol*, 16(5):263–276, 2018.

[242] I. Koike and A. Hattori. Energy yield of denitrification: An estimate from growth yield in continuous cultures of pseudomonas denitrijicans under nitrate=,nitrite- and nitrous oxide-limited conditions. *Journal of General Microbiology*, 88:11–19, 1975.

[243] T. Yoshinari. N2o reduction by vibrio succinogenes. *Appl Environ Microbiol.*, 39(1):81–84, 1980.

[244] M. Conthe, L. Wittorf, J. G. Kuenen, *et al.* Life on n2o: deciphering the ecophysiology of n2o respiring bacterial communities in a continuous culture. *IMSEJ*, 12:11421153, 2018.

[245] J. Juhanson, S. Hallin, M. Söderström, M. Stenberg, and C. M. Jones. Spatial and phyloecological analyses of nosz genes underscore niche differentiation amongst terrestrial n2o reducing communities. *Soil Biology and Biochemistry*, 115:82 – 91, 2017.

[246] C. M. Jones, D. R. Graf, D. Bru, L. Philippot, and S. Hallin. The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. *ISME J*, 7(2):417–26, 2013.

[247] Daniel R. H. Graf, Christopher M. Jones, and Sara Hallin. Intergenomic comparisons highlight modularity of the denitrification pathway and underpin the importance of community structure for n2o emissions. *PLOS ONE*, 9(12):1–20, 12 2014.

[248] B. J. Baker, C. S. Lazar, A. P. Teske, and G. J. Dick. Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome*, 3, 2015.

[249] A. K. Hawley, M. K. Nobu, J. J. Wright, *et al.* Diverse marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. *Nature Communications*, 8(1507), 2017.

[250] M. Stark, S. A. Berger, A. Stamatakis, and C. von Mering. Mltreemap - accurate maximum likelihood placement of environmental dna sequences into taxonomic and functional reference phylogenies. *BMC Genomics*, 11(1):461, 2010.

[251] C. Morgan Lang, A. K. Hawley, J. Anistad, and S. J. Hallam. Treesapp: Tree-based sensitive and accurate protein profiler. *BMC Genomics*, In Progress, 2018.

[252] A. Oulas, P. N. Polymenakou, R. Seshadri, *et al.* Metagenomic investigation of the geologically unique hellenic volcanic arc reveals a distinctive ecosystem with unexpected physiology. *Environ Microbiol*, 18(4):1122–36, 2016.

[253] A. L. Alldredge and Y. Cohen. Can microscale chemical patches persist in the sea? microelectrode study of marine snow, fecal pellets. *Science*, 235(4789):689–691, 1987.

[254] A. Bankevich, S. Nurk, D. Antipov, *et al.* Spades: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.

[255] M Holtappels, G Lavik, MM Jensen, and MMM Kuypers. 15n-labeling experiments to dissect the contributions of heterotrophic denitrification and anammox to nitrogen removal in the omz waters of the ocean. *Methods in Enzymology*, 486:223–251, 2011.

[256] E. Garcia-Robledo, C. C. Padilla, M. Aldunate, *et al.* Cryptic oxygen cycling in anoxic marine zones. *Proc Natl Acad Sci U S A*, 114(31):8319–8324, 2017.

[257] C. C. Padilla, L. A. Bristow, N. Sarode, *et al.* Nc10 bacteria in marine oxygen minimum zones. *ISME J*, 10(8):2067–71, 2016.

[258] N. P. Revsbech, B. Thamdrup, T. Dalsgaard, and D. E. Canfield. Chapter fourteen - construction of stox oxygen sensors and their application for determination of o2 concentrations in oxygen minimum zones. In Martin G. Klotz, editor, *Research on Nitrification and Related Processes, Part A*, volume 486 of *Methods in Enzymology*, pages 325 – 341. Academic Press, 2011.

[259] W. E. Durno. Precise correlation and metagenomic binning uncovers fine microbial community structure. Master's thesis, University of British Columbia, 2017. Retrieved from https://circle.ubc.ca/.

[260] V. J. Orphan, C. H. House, K. U. Hinrichs, K. D. McKeegan, and E. F. DeLong. Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. *Science*, 293(5529):484–7, 2001.

[261] S. E. McGlynn, G. L. Chadwick, C. P. Kempes, and V. J. Orphan. Single cell activity reveals direct electron transfer in methanotrophic consortia. *Nature*, 526(7574):531–5, 2015.

[262] J. Chen, A. Hanke, H. E. Tegetmeyer, *et al.* Impacts of chemical gradients on microbial community structure. *ISME J*, 11(4):920–931, 2017.

[263] T. C. Walther and M. Mann. Mass spectrometry-based proteomics in cell biology. *J Cell Biol*, 190(4):491–500, 2010.

[264] S Sunagawa, L Pedro Coelho, and S... et al Chaffron. Structure and function of the global ocean microbiome. *Science*, 348(6237), 2015.

Appendix A

Chapter 2: Supplementary material

Table A.1: Metagenome inventory Inventory of metagenomic datasets and accession numbers

SampleID	Cruise ID	Year	Month	Station	Depth (m)	MetaG IMG/M Genome ID	MetaG BioSample Accession
SI034_S3_10	34	2009	Jun	SI03	10	3300000224	SAMN05224402
SI034_S3_100	34	2009	Jun	SI03	100	3300000254	SAMN05224404
SI034_S3_120	34	2009	Jun	SI03	120	3300000225	SAMN05224407
SI034_S3_135	34	2009	Jun	SI03	135	3300000226	SAMN05224408
SI034_S3_150	34	2009	Jun	SI03	150	3300000237	SAMN05224411
SI034_S3_200	34	2009	Jun	SI03	200	3300000172	SAMN05224484
SI036_S3_100	36	2009	Aug	SI03	100	3300000238	SAMN05224405
SI036_S3_120	36	2009	Aug	SI03	120	3300000239	SAMN05224472
SI036_S3_135	36	2009	Aug	SI03	135	3300000170	SAMN05224406
SI036_S3_150	36	2009	Aug	SI03	150	3300000204	SAMN05224409
SI036_S3_200	36	2009	Aug	SI03	200	3300000155	SAMN05224412
SI037_S2_100	37	2009	Sep	SI02	100	3300004110	SAMN05224479
SI037_S2_150	37	2009	Sep	SI02	150	3300004109	SAMN05224480
SI037_S2_200	37	2009	Sep	SI02	200	3300004111	SAMN05224481
SI037_S3_10	37	2009	Sep	SI03	10	3300003599	SAMN05224521
SI037_S3_100	37	2009	Sep	SI03	100	3300003478	SAMN05224482
SI037_S3_110	37	2009	Sep	SI03	110	3300003615	SAMN05224526
SI037_S3_120	37	2009	Sep	SI03	120	3300003600	SAMN05224527
SI037_S3_125	37	2009	Sep	SI03	125	3300003620	SAMN05224532
SI037_S3_130	37	2009	Sep	SI03	130	3300003498	SAMN05224483
SI037_S3_150	37	2009	Sep	SI03	150	3300003494	SAMN05224486
SI037_S3_200	37	2009	Sep	SI03	200	3300003496	SAMN05224487
SI037_S4_100	37	2009	Sep	SI04	100	3300003500	SAMN05224429
SI037_S4_130	37	2009	Sep	SI04	130	3300003501	SAMN05224434
SI037_S4_150	37	2009	Sep	SI04	150	3300003495	SAMN05224435
SI039_S3_10	39	2009	Nov	SI03	10	4096421	SAMN05224416
SI039_S3_100	39	2009	Nov	SI03	100	4096422	SAMN05224417
SI039_S3_120	39	2009	Nov	SI03	120	4096423	SAMN05224422
SI039_S3_135	39	2009	Nov	SI03	135	4096424	SAMN05224423
SI039_S3_150	39	2009	Nov	SI03	150	4096425	SAMN05224428
SI039_S3_200	39	2009	Nov	SI03	200	4096426	SAMN05224477
SI042_S3_10	42	2010	Feb	SI03	10	*	SAMN05224451
SI042_S3_100	42	2010	Feb	SI03	100	*	SAMN05224447
SI042_S3_120	42	2010	Feb	SI03	120	*	SAMN05224436
SI042_S3_135	42	2010	Feb	SI03	135	*	SAMN05224437
SI042_S3_150	42	2010	Feb	SI03	150	*	SAMN05224442
SI042_S3_200	42	2010	Feb	SI03	200	*	SAMN05224443
SI047_S3_100	47	2010	July	SI03	100	3300000148	SAMN05224454
SI047_S3_120	47	2010	July	SI03	120	3300000212	SAMN05224455
SI047_S3_135	47	2010	July	SI03	135	3300000193	SAMN05224458
SI047_S3_150	47	2010	July	SI03	150	3300000154	SAMN05224459
SI047_S3_200	47	2010	July	SI03	200	3300000171	SAMN05224463
SI048_S3_10	48	2010	Aug	SI03	10	3300000207	SAMN05224462
SI048_S3_100	48	2010	Aug	SI03	100	3300000324	SAMN05224393
SI048_S3_120	48	2010	Aug	SI03	120	3300000150	SAMN05224394
SI048_S3_135	48	2010	Aug	SI03	135	3300000160	SAMN05224397
SI048_S3_150	48	2010	Aug	SI03	150	3300000200	SAMN05224398
SI048_S3_200	48	2010	Aug	SI03	200	3300000166	SAMN05224401
SI053_S3_10	53	2011	Jan	SI03	10	3300000143	SAMN05224489
SI053_S3_100	53	2011	Jan	SI03	100	3300000187	SAMN05224490
SI053_S3_120	53	2011	Jan	SI03	120	3300000215	SAMN05224491
SI053_S3_135	53	2011	Jan	SI03	135	3300000211	SAMN05224492
SI053_S3_150	53	2011	Jan	SI03	150	3300000216	SAMN05224466
SI053_S3_200	53	2011	Jan	SI03	200	3300000151	SAMN05224467
SI054_S3_100	54	2011	Feb	SI03	100	3300000158	SAMN05224410
SI054_S3_120	54	2011	Feb	SI03	120	3300000146	SAMN05224433
SI054_S3_135	54	2011	Feb	SI03	135	3300000201	SAMN05224473
SI054_S3_150	54	2011	Feb	SI03	150	3300000147	SAMN05224478
SI054_S3_200	54	2011	Feb	SI03	200	3300000214	SAMN05224438
SI060_S3_100	60	2011	Aug	SI03	100	3300000192	SAMN05224439
SI060_S3_150	60	2011	Aug	SI03	150	3300000188	SAMN05224444
SI060_S3_200	60	2011	Aug	SI03	200	3300000174	SAMN05224474
SI072_S3_10	72	2012	Aug-1	SI03	10	3300003592	SAMN05224440
SI072_S3_100	72	2012	Aug-1	SI03	100	3300003588	SAMN05224441
SI072_S3_120	72	2012	Aug-1	SI03	120	3300003589	SAMN05224512
SI072_S3_135	72	2012	Aug-1	SI03	135	3300003585	SAMN05224513
SI072_S3_150	72	2012	Aug-1	SI03	150	3300003591	SAMN05224518
SI072_S3_165	72	2012	Aug-1	SI03	165	3300003619	SAMN05224523
SI072_S3_200	72	2012	Aug-1	SI03	200	3300003590	SAMN05224519
SI073_S3_10	73	2012	Aug-28	SI03	10	3300003582	SAMN05224534
SI073_S3_100	73	2012	Aug-28	SI03	100	3300003583	SAMN05224524
SI073_S3_120	73	2012	Aug-28	SI03	120	3300003584	SAMN05224525
SI073_S3_135	73	2012	Aug-28	SI03	135	3300003596	SAMN05224530
SI073_S3_150	73	2012	Aug-28	SI03	150	3300003587	SAMN05224531
SI073_S3_165	73	2012	Aug-28	SI03	165	3300003618	SAMN05224533
SI073_S3_200	73	2012	Aug-28	SI03	200	3300003581	SAMN05224508

A.1 Metagenome inventory continued from previous page

SampleID	Cruise ID	Year	Month	Station	Depth (m)	MetaG IMG/M Genome ID	MetaG BioSample Accession
SI074_S3_10	74	2012	Sep-10	SI03	10	3300003594	SAMN05224529
SI074_S3_100	74	2012	Sep-10	SI03	100	3300003593	SAMN05224509
SI074_S3_120	74	2012	Sep-10	SI03	120	3300003580	SAMN05224514
SI074_S3_135	74	2012	Sep-10	SI03	135	3300003586	SAMN05224515
SI074_S3_150	74	2012	Sep-10	SI03	150	3300003602	SAMN05224528
SI074_S3_165	74	2012	Sep-10	SI03	165	3300003601	SAMN05224535
SI074_S3_200	74	2012	Sep-10	SI03	200	3300003595	SAMN05224520
SI075_S3_10	75	2012	Sep-20	SI03	10	3300004279	SAMN05224536
SI075_S3_100	75	2012	Sep-20	SI03	100	3300004280	SAMN05224522
SI075_S3_120	75	2012	Sep-20	SI03	120	3300004274	SAMN05224493
SI075_S3_150	75	2012	Sep-20	SI03	135	3300004278	SAMN05224495
SI075_S3_165	75	2012	Sep-20	SI03	165	3300004276	SAMN05224496
SI075_S3_200	75	2012	Sep-20	SI03	200	3300004277	SAMN05224497
SI075_S3_135	75	2012	Sep-20	SI03	135	3300004273	SAMN05224494

* SI042 samples are not currently in IMG/M database, but are available in the NCBI sequence read Archive with the indicated BioSample

Table A.2: Metatranscriptome inventory Inventory of metatranscriptomic datasets and accession numbers

SampleID	Cruise ID	Year	Month	Station	Depth (m)	MetaT IMG/M JGI project ID	MetaT BioSample Accession
SI042_S3_10	42	2010	Feb	SI03	10	1001537	SAMN05238748
SI042_S3_100	42	2010	Feb	SI03	100	1001540	SAMN05238739
SI042_S3_120	42	2010	Feb	SI03	120	1001543	SAMN05238743
SI042_S3_135	42	2010	Feb	SI03	135	1001546	SAMN05238741
SI042_S3_150	42	2010	Feb	SI03	150	1001549	SAMN05238745
SI042_S3_200	42	2010	Feb	SI03	200	1001552	SAMN05238751
SI047_S3_10	47	2010	July	SI03	10	3300004642	SAMN05224517
SI047_S3_100	47	2010	July	SI03	100	3300005234	SAMN05224498
SI047_S3_120	47	2010	July	SI03	120	3300004958	SAMN05224499
SI047_S3_135	47	2010	July	SI03	135	3300004640	SAMN05224500
SI047_S3_150	47	2010	July	SI03	150	3300004637	SAMN05224516
SI047_S3_200	47	2010	July	SI03	200	3300004974	SAMN05224501
SI048_S3_10	48	2010	Aug	SI03	10	3300004960	SAMN05224502
SI048_S3_100	48	2010	Aug	SI03	100	3300004962	SAMN05224503
SI048_S3_120	48	2010	Aug	SI03	120	3300004639	SAMN05224504
SI048_S3_135	48	2010	Aug	SI03	135	3300004638	SAMN05224505
SI048_S3_150	48	2010	Aug	SI03	150	3300004636	SAMN05224511
SI048_S3_200	48	2010	Aug	SI03	200	3300004641	SAMN05223291
SI054_S3_10	54	2011	Feb	SI03	10	3300004957	SAMN05223292
SI054_S3_100	54	2011	Feb	SI03	100	3300004975	SAMN05223293
SI054_S3_120	54	2011	Feb	SI03	120	3300004954	SAMN05224510
SI054_S3_135	54	2011	Feb	SI03	135	3300005233	SAMN05236416
SI054_S3_150	54	2011	Feb	SI03	150	3300004968	SAMN05224506
SI054_S3_200	54	2011	Feb	SI03	200	3300004627	SAMN05224507
SI072_S3_10	72	2012	Aug 1	SI03	10	1024556	SAMN05238753
SI072_S3_100	72	2012	Aug 1	SI03	100	1024559 1024562	SAMN05238755 SAMN05236417
SI072_S3_135	72	2012	Aug 1	SI03	135	1024571 1024574	SAMN05238757 SAMN05238759
SI072_S3_150	72	2012	Aug 1	SI03	150	1024577 1024580	SAMN05238761 SAMN05238729
SI072_S3_165	72	2012	Aug 1	SI03	165	1024583 1024586	SAMN05238731 SAMN05238732
SI072_S3_200	72	2012	Aug 1	SI03	200	1024589 1024592	SAMN05238733 SAMN05238734
SI073_S3_10	73	2012	Aug 28	SI03	10	1024595	SAMN05238721
SI073_S3_165	73	2012	Aug 28	SI03	165	1024622 1024625	SAMN05238722 SAMN05238723
SI073_S3_200	73	2012	Aug 28	SI03	200	1024628	SAMN05238724
SI074_S3_10	74	2012	Sep 10	SI03	10	1024634	SAMN05238725
SI074_S3_100	74	2012	Sep 10	SI03	100	1024637 1024640	SAMN05238726 SAMN05238727
SI074_S3_120	74	2012	Sep 10	SI03	120	1024643	SAMN05238728
SI074_S3_135	74	2012	Sep 10	SI03	135	1024649 1024652	SAMN05238730 SAMN05238763
SI074_S3_150	74	2012	Sep 10	SI03	150	1024655 1024658	SAMN05238765 SAMN05238736
SI074_S3_165	74	2012	Sep 10	SI03	165	1024661 1024664	SAMN05238738 SAMN05238740
SI074_S3_200	74	2012	Sep 10	SI03	200	1024667 1024670	SAMN05238742 SAMN05238744
SI075_S3_10	75	2012	Sep 20	SI03	10	1024673	SAMN05238746
SI075_S3_100	75	2012	Sep 20	SI03	100	1024676 1024679	SAMN05238749 SAMN05238747
SI075_S3_120	75	2012	Sep 20	SI03	120	1024682 1024685	SAMN05238750 SAMN05238752
SI075_S3_150	75	2012	Sep 20	SI03	135	1024694 1024697	SAMN05238758 SAMN05238760
SI075_S3_200	75	2012	Sep 20	SI03	200	1024706 1024709	SAMN05236415 SAMN05238735
SI075_S3_135	75	2012	Sep 20	SI03	135	1024688 1024691	SAMN05238754 SAMN05238756

* SI042 samples are not currently in IMG/M database but are avaliable in the NCBI sequence read Archive with the indicated BioSample

Table A.3: Metaproteome inventory Inventory of metaproteomic datasets and accession numbers

SampleID	Cruise ID	Year	Month	Station	Depth (m)	MetaP Pride File Prefix
SI020_S3_100	20	2008	Apr	SI03	100	SH_SBI_02, SH_SBI_03, SH_SBI_19
SI020_S3_200	20	2008	Apr	SI03	200	SH_SBI_04, SH_SBI_05, SH_SBI_21
SI020_S3_10	20	2008	Apr	SI03	10	SH_SBI_18
SI020_S3_120	20	2008	Apr	SI03	120	SH_SBI_20
SI037_S2_100	37	2009	Sep	SI02	100	SH_SBI_06
SI037_S2_130	37	2009	Sep	SI02	130	SH_SBI_09
SI037_S2_150	37	2009	Sep	SI02	150	SH_SBI_12
SI037_S2_200	37	2009	Sep	SI02	200	SH_SBI_15
SI037_S3_100	37	2009	Sep	SI03	100	SH_SBI_07
SI037_S3_130	37	2009	Sep	SI03	130	SH_SBI_10
SI037_S3_150	37	2009	Sep	SI03	150	SH_SBI_13
SI037_S3_200	37	2009	Sep	SI03	200	SH_SBI_16
SI037_S4_100	37	2009	Sep	SI04	100	SH_SBI_08
SI037_S4_130	37	2009	Sep	SI04	130	SH_SBI_11
SI037_S4_150	37	2009	Sep	SI04	150	SH_SBI_14
SI037_S4_190	37	2009	Sep	SI04	190	SH_SBI_17
SI038_S3_10	38	2009	Oct	SI03	10	SH_SBI_TC_01
SI038_S3_97	38	2009	Oct	SI03	97	SH_SBI_TC_02
SI038_S3_120	38	2009	Oct	SI03	120	SH_SBI_TC_03
SI038_S3_150	38	2009	Oct	SI03	150	SH_SBI_TC_04
SI038_S3_165	38	2009	Oct	SI03	165	SH_SBI_TC_05
SI038_S3_200	38	2009	Oct	SI03	200	SH_SBI_TC_06
SI042_S3_10	42	2010	Feb	SI03	10	SH_SBI_TC_07
SI042_S3_120	42	2010	Feb	SI03	120	SH_SBI_TC_09
SI042_S3_150	42	2010	Feb	SI03	150	SH_SBI_TC_10
SI042_S3_200	42	2010	Feb	SI03	200	SH_SBI_TC_12
SI044_S3_10	44	2010	Apr	SI03	10	SH_SBI_TC_13
SI044_S3_60	44	2010	Apr	SI03	60	SH_SBI_TC_14
SI044_S3_97	44	2010	Apr	SI03	67	SH_SBI_TC_15
SI044_S3_120	44	2010	Apr	SI03	120	SH_SBI_TC_16
SI044_S3_135	44	2010	Apr	SI03	135	SH_SBI_TC_17
SI044_S3_150	44	2010	Apr	SI03	150	SH_SBI_TC_18
SI044_S3_200	44	2010	Apr	SI03	200	SH_SBI_TC_19
SI046_S3_10	46	2010	Jun	SI03	10	SH_SBI_TC_20
SI046_S3_60	46	2010	Jun	SI03	60	SH_SBI_TC_21
SI046_S3_100	46	2010	Jun	SI03	100	SH_SBI_TC_22
SI046_S3_120	46	2010	Jun	SI03	120	SH_SBI_TC_23
SI046_S3_135	46	2010	Jun	SI03	135	SH_SBI_TC_24
SI046_S3_150	46	2010	Jun	SI03	150	SH_SBI_TC_25
SI046_S3_200	46	2010	Jun	SI03	200	SH_SBI_TC_26
SI047_S3_10	47	2010	July	SI03	10	SH_SBI_TC2_SI047_10m
SI047_S3_100	47	2010	July	SI03	100	SH_SBI_TC2_SI047_100m
SI047_S3_120	47	2010	July	SI03	120	SH_SBI_TC2_SI047_120m
SI047_S3_135	47	2010	July	SI03	135	SH_SBI_TC2_SI047_135m
SI047_S3_150	47	2010	July	SI03	150	SH_SBI_TC2_SI047_150m
SI047_S3_200	47	2010	July	SI03	200	SH_SBI_TC2_SI047_200m
SI048_S3_10	48	2010	Aug	SI03	10	SH_SBI_TC2_SI048_10m
SI048_S3_100	48	2010	Aug	SI03	100	SH_SBI_TC2_SI048_100m
SI048_S3_120	48	2010	Aug	SI03	120	SH_SBI_TC2_SI048_120m
SI048_S3_135	48	2010	Aug	SI03	135	SH_SBI_TC2_SI048_135m
SI048_S3_150	48	2010	Aug	SI03	150	SH_SBI_TC2_SI048_150m
SI048_S3_200	48	2010	Aug	SI03	200	SH_SBI_TC2_SI048_200m

Table A.3 Metaproteome inventory continued from previous page.

SampleID	Cruise ID	Year	Month	Station	Depth (m)	MetaP Pride File Prefix
SI053_S3_10	53	2011	Jan	SI03	10	SH_SBI_TC2_SI053_10m
SI053_S3_100	53	2011	Jan	SI03	100	SH_SBI_TC2_SI053_100m
SI053_S3_120	53	2011	Jan	SI03	120	SH_SBI_TC2_SI053_120m
SI053_S3_135	53	2011	Jan	SI03	135	SH_SBI_TC2_SI053_135m
SI053_S3_150	53	2011	Jan	SI03	150	SH_SBI_TC2_SI053_150m
SI053_S3_200	53	2011	Jan	SI03	200	SH_SBI_TC2_SI053_200m
SI054_S3_10	54	2011	Feb	SI03	10	SH_SBI_TC2_SI054_10m
SI054_S3_100	54	2011	Feb	SI03	100	SH_SBI_TC2_SI054_100m
SI054_S3_120	54	2011	Feb	SI03	120	SH_SBI_TC2_SI054_120m
SI054_S3_135	54	2011	Feb	SI03	135	SH_SBI_TC2_SI054_135m
SI054_S3_150	54	2011	Feb	SI03	150	SH_SBI_TC2_SI054_150m
SI054_S3_200	54	2011	Feb	SI03	200	SH_SBI_TC2_SI054_200m

A.1 RNA extraction and isolation protocol

Here I detail the protocol developed from Shi *et al.* 2009 to maximise extraction efficiency and RNA quality. As with all RNA extractions clean your work area with RNAse away or other RNAse cleaner to neutralise any RNAse that may degrade RNA in your samples. Use only RNase free tips, tubes and buffers.

RNA Extration

1. Using a peristaltic pump (Cole-Parmer), seawater is filtered through a 2.7-mm GF/D prefilter to reduce particle and eukaryotic cell loading. Flow through biomass is concentrated in-line onto a 0.2 mm Sterivex filter (Millipore). Filter volumes will vary with cell densities ranging between 1 to 5 L as greater volume take longer and microbial expression may change. Following biomass concentration, a syringe is used to purge remaining seawater from the filter cartridge prior to addition of RNALater.
2. 1.8 mL of RNALater (Ambion) is added to the Sterivex filter, sealed at both ends with parafilm, placed in steril bag, frozen on dry ice, and stored at -80°C until extraction.
3. Prior to RNA extraction, the Sterivex filter is thawed on ice.
4. Remove RNA later using a sterile 10 mm syringe to slowly push RNAlater out of Sterivex into two nuclease-free 1.5 ml tubes. Store on ice until extraction is complete and presence of RNA is validated.
5. Wash Stervix with Ringer′s Solution by adding 1.8 mL of Ringer′s (prepared with RNAse-free water) solution to Sterivex. Invert several times to mix and incubate with rolling at room temperature for 20 min.
6. Remove Ringer′s Solution from sterivex using a steril 10 mL syringe to slowly push Ringer′s solution out of Sterivex into two nuclease-free 1.5 ml tubes. Store on ice until extraction is complete and presence of RNA is validated.
7. Lyse cells in Sterivex by adding 1.8 mm of Lysis/Binding from mirVana kit to Sterivex, add 100 µL lysozyme to Sterivex (62.5 mg lysozyme in 500 µL nuclease-free TE. Invert several times to mix and incubate at 37°C for 30 min. with rolling.
8. Remove lysate from Sterivex with sterile 10 mm syringe, collect into a 15 mm falcon tube. Wash Sterivex with 1 mL Lysis/Binding buffer and add to lysate. Total lysate volume will be ∼3.5 mL
9. Organic extraction of RNA; add $^1/_{10}$ lysate volume miRNA Homogenate Additive form mirVana kit ∼350 µL. Mix well by inverting several times, incubate on ice for 10 min.

10. Add 1 lysate volume of Acid-Phenol:Chloroform (take Acid-Phenol:Chloroform from bottom of bottle, as the top is aqueous buffer), invert several times to mix. Centrifuge 10 min at 6,000 rpm in swinging bucket rotor at ∼5°, remove aqueous layer to new falcon tube and record volume.
11. Add 1.25 volumes of room temperature 100% ethanol to aqueous phase (∼350 mL).
12. Pass sample through filter cartridge. Place filter cartridge into collection tube, pipet lysate/ethanol mixture onto filter cartridge 700 µL at a time, spin 10,000 x g for ∼15 s, after each addition and apply more ethanol/lysate until all is loaded onto column.
13. Wash filter by adding 700 µL miRNA Wash Solution 1 from mirVana kit to filter cartridge. Centrifuge for 5 -10 s and discard flow-through.
14. Wash filter by adding 500 µL Wash Solution 2/3 from mirVana kit. Centrifuge 5 - 10 seconds and discard flow-through. Repeat with additional 500 µL Wash Solution 2/3 and spin. Centrifuge an additional 1 min and discard flow-through.
15. Elute RNA by transfering filter cartridge into fresh collection tube, apply 60 µL pre-heated 95°C elution solution (or nuclease free water and spin for 20-30 seconds at max. Store total RNA at -80 °C or -20°C or continue on to cleaning procedures. Aliquot two 1.0 µL aliquots into two 5 µL PCR tubes before freezing for quality control analysis on bioannalyzer.

DNA removal using TURBO DNA-free kit

1. Add 0.1 volume 10X TURBO DNase Buffer to extracted RNA solution and mix gently.
2. Incubate in heating block at 37°C for 30 min.
3. Add 6 µL resuspended DNase Inactivation Reagent and mix well.
4. Centrifuge at 10,000 x g for 1.5 min.
5. Transfer supernatant to clean tube. The supernatant contains total RNA and mo DNA.

Clean totalRNA using RNeasy MiniElute Cleanup Kit

1. Adjust sample volume to 100 µL with RNase-free water.
2. Add 350 µL Buffer RLT and mix well.
3. Add 250 µL 100% Ethanol and well by pipetting. Proceed immediately to next step.

4. Trasfer samples (700 µL) to RNeasy MinElute spin column in 2 mL collection tube and centrifuge for 15 s at $>$ 8000 x g and discard flow-through.

5. Place column in fresh 2 mL collection tube and add 500 µL Buffer RPE to spin column. Centrifuge for 15 s at $>$ 8000 x g and discard flow-through.

6. Add 500 µL of 80% Ethanol to column. Centrifuge for 2 min at $>$ 8000 x g and discard flow-through and collection tube

7. Place column in new 2 mL collection tube. Centrifuge at full speed with lid of column open for 5 min (easiest to cut the lid off at this point). Discard flow-through and collection tube.

8. Place column in new 1.5 mL collection tube and add 14 µL of RNase-free water directly to centre of column membrane. Centrifuge for 1 min at full speed to collect clean total RNA.

9. Aliquot two 1.0 µL aliquots into two 5 µL PCR tubes before freezing for quality control analysis on bioannalyzer.

A.2 Protein extraction and isolation protocol

Here I detail the protocol developed to effeciently extract total protein from Sterivex filters and peptide detection optimized for community gene expression profiling in O_2-deficient marine waters.

Sample Processing and Protein Extraction

1. Using a peristaltic pump (Cole-Parmer), seawater is filtered through a 2.7 mm GF/D prefilter to reduce particle and eukaryotic cell loading. Flow through biomass is concentrated in-line onto a 0.2 mm Sterivex filter (Millipore). Filter volumes required will vary with corresponding cell densities and typically range between 1 L in surface ocean waters and up to 200 L in dark ocean waters. Following biomass concentration, a syringe is used to purge remaining seawater from the filter cartridge prior to lysis buffer addition.

2. Add 1.8 mL of lysis buffer (0.75 M sucrose, 40 mM EDTA, 50 mM Tris, pH 8.3) to the Sterivex filter, sealed at both ends with parafilm, frozen on dry ice, and stored at -80°C until extraction.

3. Prior to protein extraction, the Sterivex filter is thawed on ice followed by the addition of 200 mM of 10X Bugbuster (Novagen). The Sterivex filter is then incubated at room temperature with rocking or rolling for 20-30 min to lyse cells.

4. The lysate is extruded from the filter into a 15 mL tube using a 10 mL syringe and put on ice prior to centrifugation at 3500 x g for 10 min at 4°C to pellet cellular debris. Rinse the filter with 1 mL of lysis buffer, extrude, and combine with lysate.

5. For buffer exchange, transfer aqueous layer to Amicon(need circledR) Ultra filter with 10K nominal molecular weight limit cutoff (Millipore), increase volume to 4 mL with 100 µM urea NH_4HCO_3, and centrifuge at 3500 x g for 10 min at 4°C or until there is less the 1 mL remaining in the Amicon filter. Keep samples on ice during buffer exchange steps. Buffer exchange two more times with 1 - 3 mL of 100 mM NH_4HCO_3.

6. In the final spin, bring volume down to to 200 mL–500 mL. Record the final extraction volume and transfer to 1.5 mL tube.

7. Protein concentration is determined with 2-(4-carboxyquinolin-2-yl) quinoline-4-carboxylic acid (Bicinchoninic acid or BCA) assay.

8. Add powdered urea to a final concentration of 8 M (780 mg/mL). NOTE: each mg of Urea added will add 0.8 mL of volume.

9. A 50 mM working stock of the reducing agent Dithiothreitol (DTT) is added to a final concentration of 5 mM and the sample is incubated at 60° for 30 min.
10. Following DTT incubation, the sample is diluted 10-fold with 100 mM NH_4HCO_3 and 1 M $CaCl_2$ is added to a final concentration of 1 mM.
11. The sample can now be flash-frozen in liquid nitrogen and stored at -80°C until trypsin digestion.

Protein digestion and sample clean-up

To remove residual salts from seawater samples as well as detergents used in protein extraction both a C18 column and strong cation exchange column are used following trypsin digest.

1. Trypsin digest is carried out using 1 unit of mass spectrometry grade trypsin to 50 units protein at 37°C for 6 hours.
2. A 1 mL/50 mg bed volume C18 Solid Phase Extraction (SPE) (Sigma-Aldrich, Supelco Supelclean) column is conditioned with 3 ml of methanol and rinsed with 2 ml 0.1% Trifluoroacetic acid (TFA) using a vacuum manifold.
3. After conditioning,the sample is added to the column and washed with 4 mL of 95:5 0.1%TFA:Acetonitrile (ACN) and allowed to dry. Peptides are eluted with 1 mL of 80:20 0.1%TFA:ACN using vacuum and concentrated to 50 mL–100 mL in a speed-vac.
4. A 1 mL/50 mg bed volume SCX SPE column is used to clean remaining detergents from the sample. Condition the column by following steps 5 - 10 on a vacuum manifold. (Note: A SCX SPE 1 mL/50 mg tube is sufficient for up to 400 mg of protein, use a 1 mL/100 mg tube for larger protein amounts.)
5. Condition column with 2 mL of methanol.
6. Rinse column with 2 mL 10 mM ammonium formate (NH_4HCO_2),25% ACN, pH 3.0.
7. Rinse column with 2 mL of 500 mM NH_4HCO_2, 25%ACN, pH 6.8.
8. Rinse column with 2 mL of 10 mM NH_4HCO_2, 25% ACN, pH 3.0.
9. Rinse column with 2 mL of Nanopure water.
10. Rinse column with 4 mL of 10 mM NH_4HCO_2, 25% ACN, pH 3.0.
11. Acidify sample by adding 10%TFA in Nanopure water to a final sample concentration of 1% and centrifuge for 5 min at 15,000 x g at room temperature to pellet any precipitates. Slowly pass the supernatant through column.

12. Wash the column with 4 mL of 10 mM NH_4HCO_2 25%ACN, pH 3.0, and elute to dryness. Blot ends of manifold tubing below columns dry.

13. Place fresh 2.0 mL microcentrifuge tubes below columns, and with the vacuum turned off, add 1.0 mL MeOH:H_2O:NH_4OH (80:15:5) to each column.

14. Turn on vacuum, slowly elute sample from columns, and when columns are dry add an additional 500 mL of MeOH:H_2O:NH_4OH (80:15:5) for a total elution volume of 1.5 mL. Concentrate the sample in a speed-vac to a final volume of 50 mL–100 mL, adding small volumes of H2O to dissolve particulate matter on the side of the tube (if needed). Perform BCA protein assay.

15. The sample can now be flash-frozen in liquid nitrogen and stored at 80°C until needed for MS analysis.

A.3 Protein sequencing protocol

Here I detail the protocol developed to effeciently match peptide sequences to a protein sequence database composed of metagenomic sequences (translated into protein sequences) from Saanich Inlet.

Tandem mass spectrometry and peptide identification

1. Aliquots containing 5 mg of protein are analyzed by online capillary liquid-chromatography?tandem mass spectrometry (Thermo, LTQ ion trap mass spectrometer or Thermo LTQ-Orbitrap mass spectrometer) using data-dependent fragmentation on the top 10 ions per duty cycle and a 100-min LC gradient from 0.1% formic acid in water to 0.1% formic acid acetonitrile. (Note: Reverse-phase capillary HPLC column used was made in-house at Environmental Molecular Sciences Laboratory at Pacific Northwest National Laboratories by slurry packing 3 mm Jupiter C_{18} stationary phase into a 60 cm length of 360 mm o. d. 75 mm i.d. fused silica capillary tubing using a 1 cm sol-gel frit for retention of the packing material.)

2. Peptides are identified from MS/MS spectra using SEQUEST™allowing for a potential oxidation of the methionine residues. Each search is performed using an environmental database of predicted protein sequences generated from the source location. For ion trap data, a mass error window of 3 m/z units is used for the precursor mass. A mass error window of 1 m/z unit is used for Orbitrap data, given the higher resolution of the instrument. In both cases, a 0 m/z tolerance is used for the fragmentation mass. Peptides identifications are permitted if they have a mass spectra

generating functions value of less than 10^{11}, which corresponds to a false discovery rate below 2% [207]. Identifications are allowed for all possible peptide termini, that is, not limited by tryptic-only termini.

3. The number of peptide observations(scans matching to a peptide) from each protein is used as a rough measure of relative abundance and multiple charge states of a single peptide are considered as individual observations, as are the same peptides detected in different mass spectral analyses. However, it is important to note that while abundant proteins tend to produce more spectra, not all peptides ionize equally well. The most accurate quantitation requires some form of metabolic, isotopic, or isobaric tagging, or in the case of targeted proteomics, selected reaction monitoring or multiple reaction monitoring using stable isotope-labeled synthetic peptides [263].

A.4 Taxonomic binning and visualization of expressed proteins

Here I detail the protocol developed to calculate normalised spectral abundance factors (NSAF) for metaproteomic samples to estaimate the quantity of peptides detected for a given protein in a metaproteome and detail process of taxonomic binning and vissualization of expressed proteins.

1. Amino acid sequences of detected proteins are compared to known protein sequences in genomic databases such as NCBI RefSeq via BLAST. A bit score ratio of 0.4 is used as a cutoff for confidence and the top hit is assigned to that protein sequence.

2. To increase taxonomic resolution and include genomes or metagenomes that are not yet in public database (such as RefSeq), the target database can be amended with the user-defined sequence information. For example, protein sequences for SUP05 uncultured bacterium and *Candidatus* Kuenenia stuttgartiensis were included in the BLAST against the NCBI RefSeq database by amending the RefSeq database with the additional genomic sequence information from desired organisms.

3. Peptide scan counts are summed for each protein (with PPP>0.95). For peptides mapping to more than one protein, scan counts are divided between the total number of identified proteins.

4. The spectral abundance factor (SAF) (Equation A.1) is calculated using the sum of all scan counts for a given protein divided by the number of amino acids making up the protein sequence. The NSAF (Equation A.2) is the SAF for a given protein divided by the sum of all SAFs for a given sample.

$$\text{SAF} = \frac{\text{Sum of scan counts for a given protein}}{\text{Length of a given protein}} \tag{A.1}$$

$$\text{NSAF} = \frac{(\text{SAF})}{\text{Sum of all SAF in a given sample}} \tag{A.2}$$

5. MEGAN is run using the BLAST output for all identified protein sequences in a given sample by using the `Import from BLAST` option in the `File` menu. In the `Import` tab of the import dialogue box, select the BLAST output file, in the `Content` tab, deselect SEED and KEGG options, in the `LCA Params` tab, change `Min Support` to `1` and deselect `Use Min-Complexity Filter`. (Note: These param- eters are user defined and can be altered based on user preferences or specific data requirements.)

6. Users can include taxa missing from the NCBI taxonomy or alter the structure of the NCBI taxonomy, for example, SUP05 uncultured bacterium by downloading the NCBI taxonomy structure from `http://ab.inf.uni-tuebingen.de/data/software/megan4/download/welcome.html` under `Updates of NCBI taxonomy`. Unzip the file in the `MEGAN/class/resources/files` directory. Open the `names.dmp` file in a text editor, append an unused taxon ID number (left most field) for the new `species` and enter `scientific name` in the far right field maintaining the syntax present in the rest of the file. Repeat this process at the genus or family level (parent nodes) as needed. For example, in the case of SUP05 uncultured bacterium, the `names.dmp` file was amended with the following lines:

```
805819 |SUP05 cluster| |scientific name|
```

```
805820 |uncultured SUP05 cluster bacteriumr| |scientific name|
```

To place SUP05 in context with existing parent nodes for higher order taxonomic structure, determine the taxon ID number present in the `names.dmp` file, for example, the Gammaproteobacteria taxon ID number is 1236. Open the `nodes.dmp` file, locate the position of your new taxon ID number in the far left field, and enter the new taxon ID number in the far left field and the taxon ID number for the parent node in the second field position, for the remainder of the fields copy from an existing line (the NCBI download site explains all these fields in `taxdump_readme.txt` at `ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/`). For example, in the case of SUP05 uncultured bacterium, the `nodes. dmp` file was amended with the following lines:

```
805819 | 135619 | order | | 1 | 1 | 1 | 1 | 5 | 1 | 0 | 0 | |
```

```
805820 | 805819 | species | | 1 | 1 | 1 | 1 | 5 | 1 | 0 | 0 | |
```

Save both of these files and relaunch MEGAN to use the newly assigned taxonomy and import your BLAST output file as described in step 1.

7. In the MEGAN tree view, open all the desired nodes by selecting aleaf on the far right, and selecting `Uncollapse Subtree` from the `Tree` menu to open the internal nodes for that leaf, repeat for all leaves. From the `Select` menu, select `All Nodes` and then from the `Export` menu select `Reads` in the `File` menu to export a list of sequences belonging to each taxon level.

8. On your desktop, open a terminal window. Using the `cd` command, enter the directory containing the MEGAN output files. Combine these files into a single file with sequence name and taxon ID using the following gawk scripts:

```
user$  gawk     ?{f=FILENAME".new";sub(" ^>",?",$0);print $0
       ?\t? FILENAME> f}? *.fasta
user$  cat *.new >combine_taxanames.txt
user$ gawk ?! a[$0]  ? combine_taxanames.txt >
       combine_taxonnames_clean.txt
```

9. To sum the NSAF values for all proteins assigned to agiven taxon node, copy the `combine_taxonnames_clean.txt` file into Excel and sum the NSAF values for all sequences with a unique taxon ID using the `SUMIF` function.

10. Tree generation using iTOL requires the NCBI taxon ID assigned by MEGAN. A user account in iTOL is required to enter and store created trees. Enter a list of all taxon IDs (obtained from the `names.dmp` file for identified taxa) into iTOL at `other trees` (no number below 1 is permitted) to generate a Newick formatted output tree. Copy the tree and upload it to a new project in iTOL using `advanced options` with internal node IDs selected. View tree in iTOL to evaluate information content and visual balance. If there are more nodes than are possible to view in a single figure, consider adjusting the MEGAN parameters, or use a cutoff of minimum NSAF value, excluding all underrepresented nodes, to reduce to total number of nodes and repeat the import process. Once satisfied with your tree, export in Newick format including internal nodes.

11. Compare the taxon names in the exported tree to the taxon names of the nodes from MEGAN. It is important that names match exactly. Change spaces and '|' into '_'.

12. For each sample, make a .csv file in the format `taxa value value value`. Use the same value in the first two `value` columns with an `R` in front of the first value and a zero for the third value (e.g., Gammaproteobacteria, R28, 28, 0). See iTOL `Uploading and working with your own trees` for more information on file headers. To provide a scale for your tree, add values to nodes which are '0' and manipulate the graphics file later to move scale bubbles into tree positions in the figure.

13. The.csv file is up loaded with a new tree in iTOL using the `Multivalue bar or pie chart` data type. Additional settings such as the minimum and maximum radii are user defined. Occasionally, the tree or the data may not be displayed in the iTOL user interface but can still be exported. Use the export function to generate an `.svg` file.

14. The output of multiple `.svg` files can be composited together using graphic design software (e.g., Adobe Illustrator) to visualize NSAF values, sequence read count, and taxonomic structure in a unified perspective (Figure 2.6).

Appendix B

Chapter 3: Supplementary material

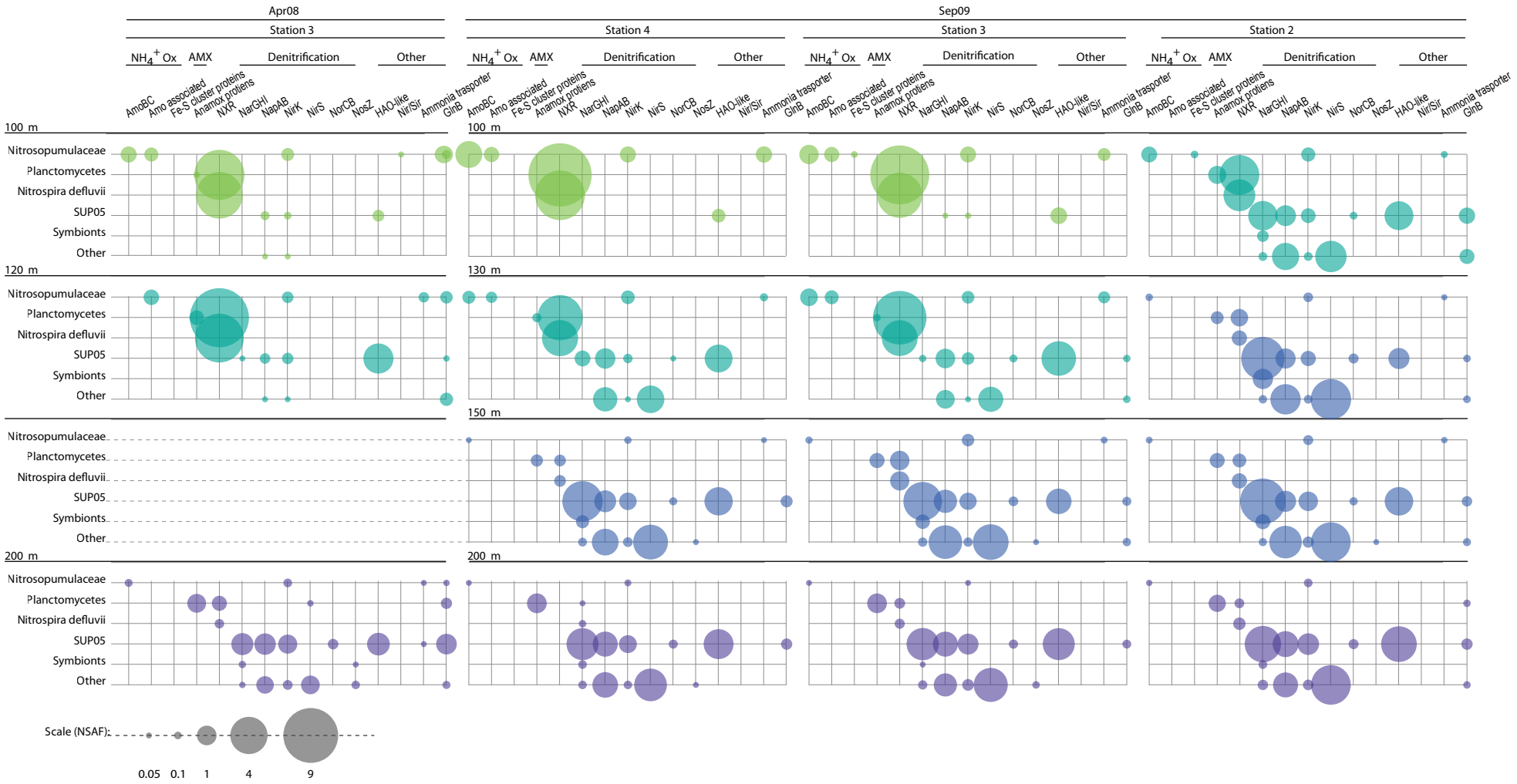


Figure B.1: Detected nitrogen cycling proteins. Detected protein NSAF values in the Nitrogen cycle for April 2008 and September 2009 showing highly similar protein expression profiles within water column compartments across multiple stations. Ammonia monooxygenase (AmoBC), protein product of Nmar_1501 adjacent to genes for AmoA and AmoB (Amo associated), 4Fe-4S binding domain proteins from Nitrosopumulacea apparently co-expressed with Amo proteins (Fe-S cluster proteins), Anammox proteins hydroxylamine oxidoreductase, hydrazine oxidoreductase, Nitrate reductase (NarGHI), periplasmic nitrate reductase (NapAB), copper containing nitrite reductase (NirK), nitrite reductase (NirS), nitric oxide reductase (NorCB), nitrous oxide reductase (NosZ), potential hydroxylamine oxidoreductase (HAO-like), nitrite/sulfite reductase proteins (Nir/Sir), Ammonium transporters, nitrogen PII regulatory proteins (GlnB).

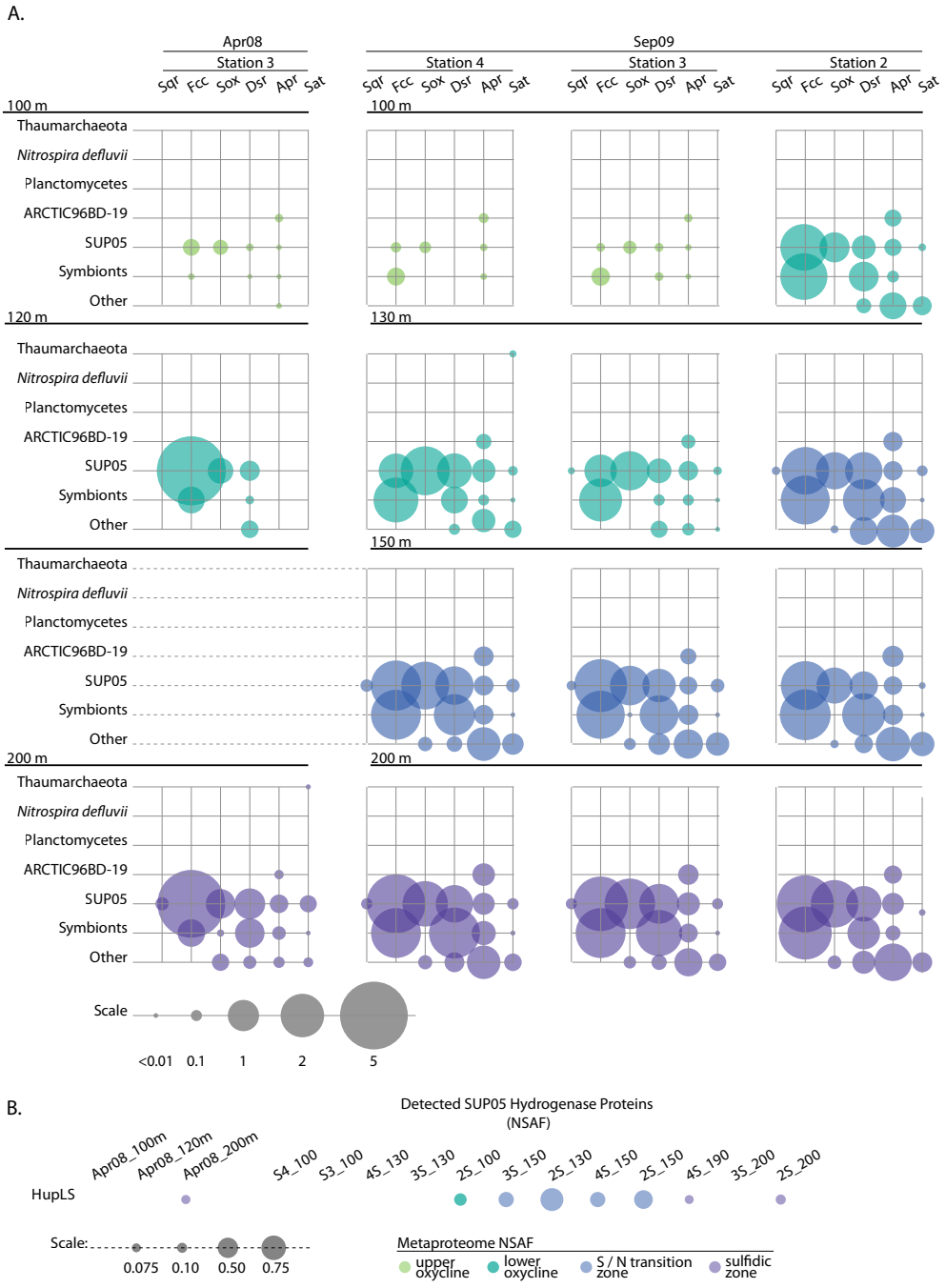


Figure B.2: Detected sulfur and hydrogen cycling proteins. (A) Detected protein NSAF values in the sulfur cycle for April 2008 and September 2009 showing highly similar protein expression profiles within water column compartments across multiple stations. Sulfide:quinone reductase (Sqr), Fcc flavocytochrome C (Fcc), Sox sulfide oxidation protein complex (Sox), dissimilatory sulfate reductase pathway (Dsr), adenylylsulfate reductase (Apr), ATP sulfurylase (Sat). **(B)** Detected protein NSAF values for hydrogen cycling gene Hydrogenase (HupLS).

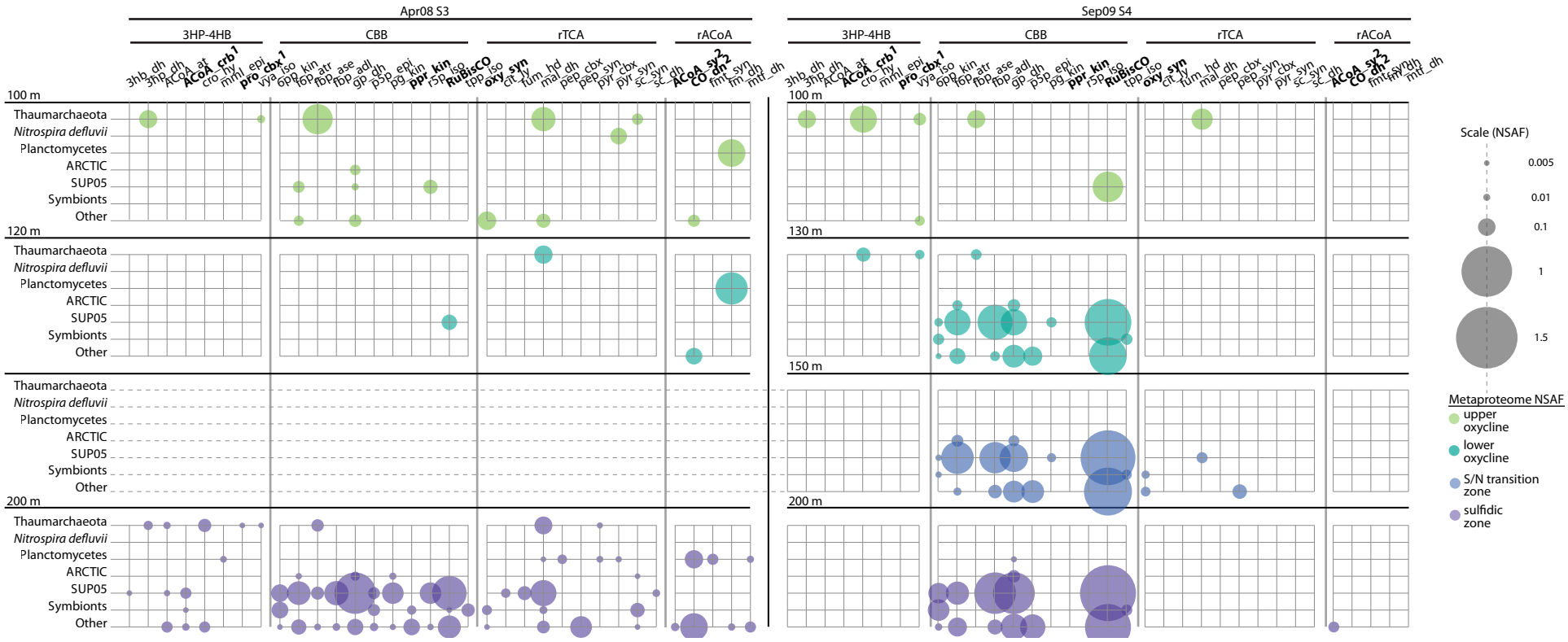


Figure B.3: Detected proteins in carbon fixation pathways. Detected protein NSAF values in inorganic carbon fixation pathways for April 2008 and September 2009. See Table B.4 for a list of protein names.

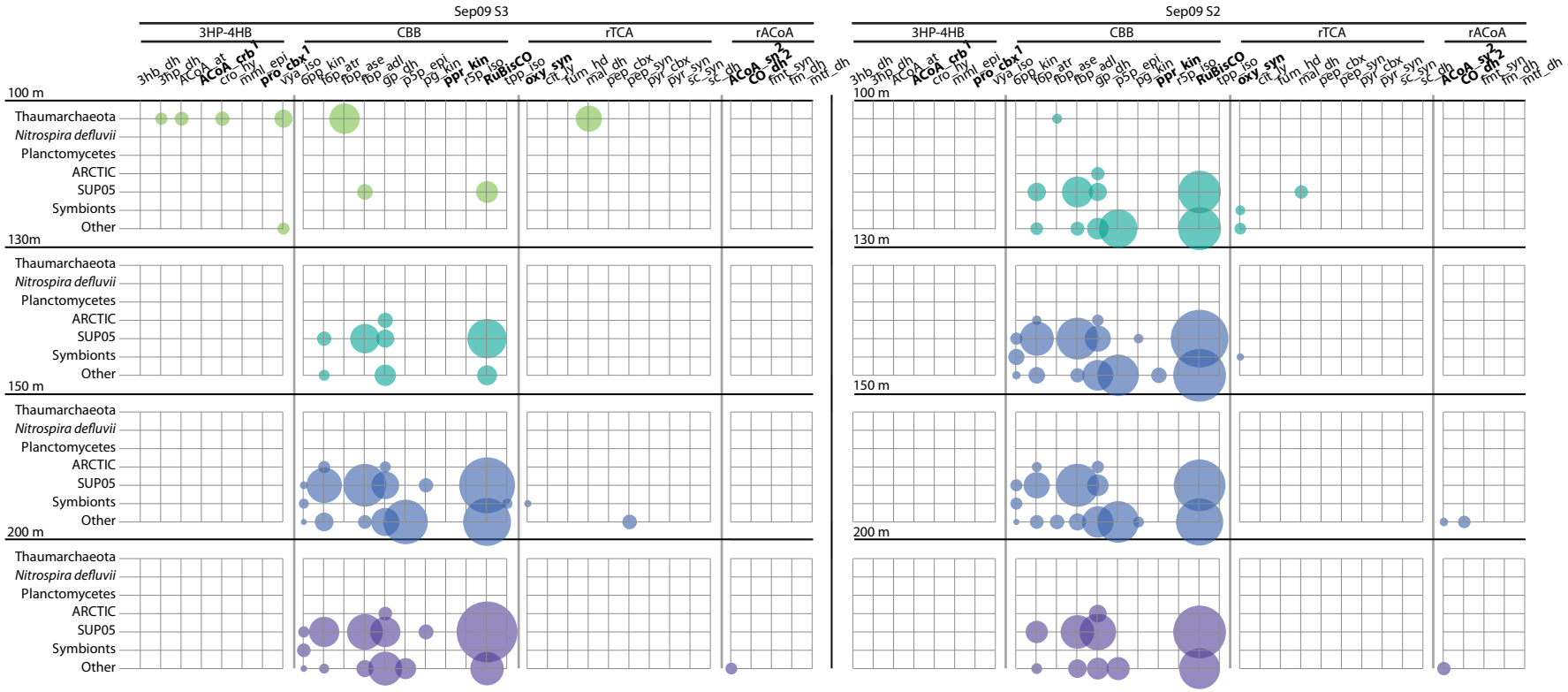


Figure B.3 continued

Table B.1: Number of detected peptides and proteins. Detected peptides and proteins for samples from April 2008 and September 2009, showing before and after peptide prophet probability (PPP) score cutoff of ≥ 0.95

			Total peptides detected	Unique peptides detected	Proteins detected
Apr-08	100m 10 L	total	829	549	811
		PPP $\geq$ .95	719	448	699
	100m 1.0 L	total	1480	771	815
		PPP $\geq$ .95	738	281	270
	120m 1.0 L	total	1514	784	763
		PPP $\geq$ .95	727	292	234
	200m 10L	total	17089	7651	6032
		PPP $\geq$ .95	14850	5796	4344
	200m 1.0 L	total	3344	1571	1436
		PPP $\geq$ .95	1883	692	487
Sep-09	2S_100 9.5 L	total	4145	2123	2046
		PPP $\geq$ .95	2220	887	708
	3S_100 9.4 L	total	2542	1307	1467
		PPP $\geq$ .95	1447	584	654
	S4_100 9.4 L	total	1709	881	1031
		PPP $\geq$ .95	978	395	439
	2S_130 10 L	total	6292	2952	2796
		PPP $\geq$ .95	3321	1148	847
	3S_130 10 L	total	2800	1520	1562
		PPP $\geq$ .95	1542	666	557
	4S_130 7.4 L	total	3550	1798	1812
		PPP $\geq$ .95	2136	844	758
	2S_150 8.6 L	total	5490	2659	2501
		PPP $\geq$ .95	2860	1008	713
	3S_150 9.0 L	total	6638	3120	2967
		PPP $\geq$ .955	3686	1253	924
	4S_150 9.0 L	total	5649	2777	2589
		PPP $\geq$ .95	3088	1117	772
	2S_200 9.9 L	total	3565	1970	1958
		PPP $\geq$ .95	1734	719	574
	3S_200 9.0 L	total	4010	2111	2027
		PPP $\geq$ .95	2162	866	678
	4S_200 8.0 L	total	4872	2557	2443
		PPP $\geq$ .95	2461	982	715

Table B.2: Taxonomic breakdown for April 2008 metagenome. Taxonomic breakdown for abundant groups for metagenomic reads from April 9, 2008 samples. Percentage of metagenome indicates the percentage of metagenomic reads (above 30 amino acids) which had top BLAST hit to indicated taxa. Number of unique genes detected is the number of unique reference sequences for indicated taxa that were recovered in top BLAST hits. Percentage genome covered is derived by the number of unique references recovered for a taxa divided by the total number of protein coding genes in the genome of that taxa.

	% of Metagenome			Total number of genes detected			Number of unique genes detected			% genome covered		
	Apr08_ 100	Apr08_ 120	Apr08_ 200	Apr08_ 100	Apr08_ 120	Apr08_ 200	Apr08_ 100	Apr08_ 120	Apr08_ 200	Apr08_100	Apr08_120	Apr08_200
Nitrosopumilaceae	14.834	14.903	2.357	1875	1720	361	-	-	-	-	-	-
Ca. Nitrosoarchaeum koreensis MY1	1.108	0.988	0.183	140	114	28	117	96	26	6.02	4.94	1.34
Ca. Nitrosoarchaeum limnia BG20	0.649	0.771	0.085	82	89	13	69	75	13	2.99	3.26	0.56
Ca. Nitrosoarchaeum limnia SFB1	0.815	1.178	0.157	103	136	24	87	106	20	4.27	5.2	0.98
Ca. Nitrosopumilus salaria BD31	6.036	5.138	0.986	763	593	151	485	402	136	22.52	18.66	6.31
Nitrosopumilus maritimus SCM1	5.965	6.629	0.92	754	765	141	506	482	127	28.17	26.84	7.07
Cenarchaeum symbiosum A	0.261	0.199	0.026	33	23	4	22	19	4	1.09	0.94	0.2
Candidatus Nitrospira defluvii	0.253	0.286	0.072	32	33	11	28	26	9	0.66	0.61	0.21
Planctomycetaceae	2.445	3.89	7.338	309	449	1124	-	-	-	-	-	-
Ca. Kuenenia stuttgartiensis	0.055	0.104	0.209	7	12	32	6	12	24	0.13	0.26	0.51
*Ca. Scalindua profunda**	0.672	2.192	6.763	85	253	1036	67	192	742			
Blastopirellula marina DSM 3645	0.245	0.173	0.046	31	20	7	27	17	7	0.45	0.28	0.12
Planctomyces brasiliensis DSM 5305	0.309	0.208	0.013	39	24	2	31	22	2	0.65	0.46	0.04
Planctomyces limnophilus DSM 3776	0.079	0.078	0.026	10	9	4	8	9	4	0.19	0.21	0.09
Planctomyces maris DSM 8797	0.348	0.364	0.085	44	42	13	38	40	9	0.59	0.62	0.14
Singulisphaera acidiphila DSM 18658	0.127	0.225	0.026	16	26	4	14	24	4	0.18	0.31	0.05
planctomycete KSU-1	0.111	0.147	0.091	14	17	14	14	14	13	0.39	0.39	0.36
Gemmata obscuriglobus UQM 2246	0.087	0.052	0.033	11	6	5	8	6	4	0.1	0.08	0.05
Isosphaera pallida ATCC 43644	0.055	0.026	0	7	3	0	7	3	0	0.19	0.08	0
Pirellula staleyi DSM 6068	0.19	0.173	0.033	24	20	5	23	18	5	0.49	0.38	0.11
Rhodopirellula baltica SH 1	0.166	0.147	0.013	21	17	2	19	16	2	0.26	0.22	0.03
SAR11 Cluster	12.033	6.533	1.025	1521	754	157	-	-	-	-	-	-
alpha proteobacterium HIMB114	0.095	0.069	0.007	12	8	1	12	7	1	0.84	0.49	0.07
Ca. Pelagibacter sp. HTCC7211	4.881	3.726	0.607	617	430	93	435	315	82	30.06	21.77	5.67
Ca. Pelagibacter sp. IMCC9063	0.285	0.139	0.052	36	16	8	34	13	7	2.35	0.9	0.48
Ca. Pelagibacter ubique HTCC1002	2.492	1.022	0.118	315	118	18	247	94	17	17.73	6.75	1.22
Ca. Pelagibacter ubique HTCC1062	4.28	1.577	0.242	541	182	37	397	156	34	29.32	11.52	2.51
ARCTIC96BD-19	2.745	2.227	1.234	347	257	189	229	180	75	25.79	20.27	8.45
SUP05	3.125	6.776	27.216	395	782	4169	259	499	1097	20.08	38.68	85.04
Symbionts	1.891	2.27	8.8	239	262	1348	-	-	-	-	-	-
Ca. Vesicomyosocius okutanii HA	0.459	0.442	1.528	58	51	234	45	45	136	4.8	4.8	14.51
Ca. Ruthia magnifica str. Cm (Calyptogena magnifica)	0.728	0.927	3.551	92	107	544	73	85	241	7.48	8.71	24.69
Endoriftia persephone 'Hot96_1+Hot96_2'	0.016	0.026	0.091	2	3	14	2	2	13	0.03	0.03	0.2
endosymbiont of Riftia pachyptila (vent Ph05)	0.055	0.087	0.405	7	10	62	6	10	53	0.19	0.31	1.67
endosymbiont of Tevnia jerichonana (vent Tica)	0.119	0.165	0.947	15	19	145	14	18	102	0.43	0.56	3.16
*endosymbiont of Bathymodiolus sp.**	0.514	0.624	2.278	65	72	349	41	57	131	-	-	-

* complete genome not avaliable

Table B.3: Taxonomic breakdown for Sepetmber 2009 metaproteome. Taxonomic break down of abundant groups for metaproteome samples from September 1, 2009. Protein NSAF shows the total value for indicated group or taxa. Total proteins detected shows the total number of detected proteins originating from indicated taxa. Unique proteins detected shows the number of unique reference sequences for indicated taxa which were recovered in top BLAST hit. Percent genome coverage in proteome is derived by the number of unique references recovered for a taxa divided by the total number of protein coding genes in the genome of that taxa.

	NSAF														
	3S 100	S4 100	2S 100	3S 130	4S 130	2S 130	2S 150	3S 150	4S 150	2S 200	3S 200	4S 200	Apr08 100	Apr08 120	Apr08 200
Thaumarchaeota	28.591	30.035	10.738	9.7	10.983	3.018	3.282	4.15	1.8	2.484	1.44	1.224	22.797	17.668	3.113
Ca. Nitrosoarchaeum koreensis MY1	2.669	3.903	0.81	1.008	0.753	0.104	0.089	0.129	0.04	0.023	0.019	0.017	1.876	0.341	0.278
Ca. Nitrosoarchaeum limnia BG20	2.948	2.442	0.635	0.891	0.971	0.037	0	0.146	0	0	0	0	2.296	2.497	0.131
Ca. Nitrosoarchaeum limnia SFB1	6.822	9.89	3.214	3.275	3.704	1.347	1.399	1.355	0.805	1.292	0.7	0.686	5.145	6.147	0.384
Ca. Nitrosopumilus salaria BD31	7.134	6.288	1.002	1.133	1.64	0.166	0.12	0.245	0.087	0.099	0.026	0.072	4.797	0.697	0.959
Nitrosopumilus maritimus SCM1	8.943	7.496	5.069	3.392	3.891	1.364	1.674	2.276	0.867	1.07	0.696	0.449	8.566	7.986	1.353
Cenarchaeum symbiosum A	0.075	0.016	0.008	0	0.024	0	0	0	0	0	0	0	0.116	0	0.008
Ca Nitrospira defluvii	6.285	7.995	3.085	4.096	4.136	0.624	0.689	0.954	0.36	0.35	0.372	0.096	7.105	7.57	0.362
Planctomycetia	10.68	12.36	6.1	9.186	6.752	2.06	2.036	2.158	1.423	2.54	2.311	2.045	8.874	11.573	5.872
Ca. Kuenenia stuttgartiensis	0.081	0.116	0.216	0.256	0.269	0.118	0.306	0.188	0.277	0.222	0.348	0.237	0.414	0	0.286
Ca. Scalindua profunda	0.142	0.171	1.291	0.379	0.459	1.143	1.249	0.964	0.843	2.172	1.795	1.781	0.371	0.675	5.256
Blastopirellula marina DSM 3645	0	0	0	0	0	0	0	0	0	0	0	0	0.328	0	0
Planctomyces brasiliensis DSM 5305	0.014	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001
Planctomyces limnophilus DSM 3776	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Planctomyces maris DSM 8797	0	0	0	0	0	0	0	0	0	0	0	0	0.276	0.385	0
Singulisphaera acidiphila DSM 18658	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.022
planctomycete KSU-1	10.45	12.08	4.594	8.551	6.024	0.799	0.48	1.006	0.303	0.145	0.168	0.027	7.484	10.513	0.306
SAR11 cluster	10.536	10.322	4.281	5.936	3.532	0.565	0.607	0.704	0.552	0.151	0.29	0.415	10.356	8.742	0.703
alpha proteobacterium HIMB114	0	0	0	0	0	0	0	0	0	0	0	0	0.034	0	0.015
Ca. Pelagibacter sp. HTCC7211	0.853	0.764	0.263	0.251	0.131	0.071	0.073	0.088	0.11	0.044	0.08	0.117	0.823	0.104	0.183
Ca. Pelagibacter sp. IMCC9063	0.374	0.189	0	0	0	0	0	0	0	0	0	0	0.376	0	0
Ca. Pelagibacter ubique HTCC1002	3.664	4.942	1.233	2.077	1.688	0.07	0.102	0.121	0	0	0	0	5.492	3.798	0.164
Ca. Pelagibacter ubique HTCC1062	5.645	4.426	2.785	3.608	1.714	0.425	0.432	0.494	0.442	0.107	0.209	0.298	3.631	4.839	0.34
ARCTIC96BD-19	6.009	6.505	3.686	3.145	2.273	1.475	1.829	1.916	2.067	2.357	3.205	2.747	8.717	6.156	1.396
SUP05	7.779	7.948	24.45	28.64	32.78	32.69	34.2	35.52	41.53	40.49	41.53	40.07	6.824	15.975	46.307
sulfur-oxidizing symbionts	1.054	0.726	7.247	4.39	7.005	12.86	11.27	10.91	12.26	9.33	11.83	13.73	1.688	2.586	8.754
Ca. Vesicomyosocius okutanii HA	0.007	0.006	0.32	0.158	0.551	0.752	0.618	0.821	1.204	0.875	1.273	1.548	0.098	0	1.448
Ca. Ruthia magnifica str. Cm (Calyptogena magnifica)	0.583	0.322	0.631	0.664	1.551	2.061	1.941	2.083	2.425	1.728	3.744	2.949	1.217	0.52	3.379
Endoriftia persephone 'Hot96_1+Hot96_2'	0	0	0.016	0	0	0	0.151	0.06	0.012	0	0	0	0	0	0.039
endosymbiont of Riftia pachyptila (vent Ph05)	0.323	0.306	4.22	2.579	3.192	5.766	5.377	4.974	5.144	4.839	4.288	5.046	0.048	0.784	1.778
endosymbiont of Tevnia jerichonana (vent Tica)	0	0	1.432	0.613	0.77	2.947	2.016	2.063	2.067	0.886	0.859	1.895	0.062	0	0.678
endosymbiont of Bathymodiolus sp.	0.141	0.094	0.627	0.373	0.942	1.335	1.168	0.917	1.411	1.002	1.669	2.299	0.262	1.282	1.432

Table B.3 Taxonomic breakdown for Sepetmber 2009 metaproteome continued

	Total Proteins Detected														
	3S 100	S4 100	2S 100	3S 130	4S 130	2S 130	2S 150	3S 150	4S 150	2S 200	3S 200	4S 200	Apr08 100	Apr08 120	Apr08 200
Thaumarchaeota	241	145	75	82	117	37	28	51	24	16	16	13	245	44	308
Ca. Nitrosoarchaeum koreensis MY1	26	22	15	14	16	7	6	8	5	2	2	2	21	6	30
Ca. Nitrosoarchaeum limnia BG20	13	10	7	9	9	4	0	5	0	0	0	0	14	8	9
Ca. Nitrosoarchaeum limnia SFB1	14	11	4	5	10	4	3	4	3	3	3	3	16	4	9
Ca. Nitrosopumilus salaria BD31	98	52	25	25	37	8	8	13	6	3	3	3	94	6	130
Nitrosopumilus maritimus SCM1	88	49	23	29	44	14	11	21	10	8	8	5	98	20	128
Cenarchaeum symbiosum A	2	1	1	0	1	0	0	0	0	0	0	0	2	0	2
Ca Nitrospira defluvii	9	6	4	4	4	4	4	3	4	3	4	3	7	6	21
Planctomycetia	19	17	33	25	37	36	31	40	29	36	38	33	39	15	284
Ca. Kuenenia stuttgartiensis	4	2	12	5	13	11	8	11	10	10	14	10	13	0	29
Ca. Scalindua profunda	5	6	12	11	15	17	16	19	14	23	21	21	15	6	239
Blastopirellula marina DSM 3645	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Planctomyces brasiliensis DSM 5305	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Planctomyces limnophilus DSM 3776	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Planctomyces maris DSM 8797	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
Singulisphaera acidiphila DSM 18658	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
planctomycete KSU-1	9	9	9	9	9	8	7	10	5	3	3	2	9	8	12
SAR11 cluster	35	30	29	24	29	10	11	21	9	7	8	7	50	18	61
alpha proteobacterium HIMB114	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
Ca. Pelagibacter sp. HTCC7211	9	4	6	4	5	4	3	5	5	3	4	3	12	1	25
Ca. Pelagibacter sp. IMCC9063	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
Ca. Pelagibacter ubique HTCC1002	10	9	9	8	8	2	2	4	0	0	0	0	16	8	10
Ca. Pelagibacter ubique HTCC1062	15	16	14	12	16	4	6	12	4	4	4	4	20	9	24
ARCTIC96BD-19	20	15	25	19	21	22	20	23	22	16	18	28	27	8	99
SUP05	94	66	204	151	216	301	251	324	297	205	261	280	104	47	1529
sulfur-oxidizing symbionts	25	12	57	36	71	100	79	102	106	68	86	90	29	8	463
Ca. Vesicomyosocius okutanii HA	4	1	11	6	12	16	13	16	19	15	15	15	3	0	82
Ca. Ruthia magnifica str. Cm (Calyptogena magnifica)	11	8	18	14	24	38	26	41	39	22	33	32	19	3	188
Endoriftia persephone 'Hot96_1+Hot96_2'	0	0	1	0	0	0	1	1	1	0	0	0	0	0	3
endosymbiont of Riftia pachyptila (vent Ph05)	2	1	8	6	8	13	8	13	10	8	8	11	2	3	22
endosymbiont of Tevnia jerichonana (vent Tica)	2	0	10	1	7	13	13	16	12	10	11	13	1	0	40
endosymbiont of Bathymodiolus sp.	6	2	9	9	20	20	18	15	25	13	19	19	4	2	128

Table B.3 Taxonomic breakdown for Sepetmber 2009 metaproteome continued

	Unique proteins detected														
	3S 100	S4 100	2S 100	3S 130	4S 130	2S 130	2S 150	3S 150	4S 150	2S 200	3S 200	4S 200	Apr08 100	Apr08 120	Apr08 200
Thaumarchaeota	120	75	42	44	64	23	18	29	16	10	11	9	128	27	161
Ca. Nitrosoarchaeum koreensis MY1	15	11	9	7	9	4	3	5	2	1	1	1	11	4	18
Ca. Nitrosoarchaeum limnia BG20	9	6	4	5	5	1	0	2	0	0	0	0	9	4	8
Ca. Nitrosoarchaeum limnia SFB1	10	7	4	5	7	4	3	4	3	3	3	3	11	4	8
Ca. Nitrosopumilus salaria BD31	41	23	12	11	18	6	5	7	5	2	2	2	47	5	57
Nitrosopumilus maritimus SCM1	43	27	12	16	24	8	7	11	6	4	5	3	49	10	68
Cenarchaeum symbiosum A	2	1	1	0	1	0	0	0	0	0	0	0	1	0	2
Ca Nitrospira defluvii	5	3	2	2	2	2	2	1	2	1	2	1	4	3	9
Planctomycetia	9	6	13	11	15	22	17	20	18	19	18	17	16	8	175
Ca. Kuenenia stuttgartiensis	2	1	4	2	4	6	4	4	5	4	5	3	5	0	15
Ca. Scalindua profunda	4	3	7	7	9	13	10	13	11	14	12	13	7	5	152
Blastopirellula marina DSM 3645	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Planctomyces brasiliensis DSM 5305	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Planctomyces limnophilus DSM 3776	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Planctomyces maris DSM 8797	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
Singulisphaera acidiphila DSM 18658	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
planctomycete KSU-1	2	2	2	2	2	3	3	3	2	1	1	1	2	2	4
SAR11 cluster	24	19	15	12	16	9	9	13	8	5	6	5	36	7	43
alpha proteobacterium HIMB114	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
Ca. Pelagibacter sp. HTCC7211	8	4	6	4	5	3	3	4	5	3	4	3	11	1	19
Ca. Pelagibacter sp. IMCC9063	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
Ca. Pelagibacter ubique HTCC1002	5	4	4	3	3	2	2	3	0	0	0	0	10	3	7
Ca. Pelagibacter ubique HTCC1062	10	10	5	5	8	4	4	6	3	2	2	2	13	3	15
ARCTIC96BD-19	9	7	10	7	7	9	7	9	9	6	6	12	12	2	44
SUP05	32	24	71	58	73	103	87	108	97	67	83	90	46	23	517
sulfur-oxidizing symbionts	14	8	40	23	43	60	51	61	60	41	50	53	17	7	264
Ca. Vesicomyosocius okutanii HA	3	1	7	3	7	9	7	9	10	7	7	7	3	0	56
Ca. Ruthia magnifica str. Cm (Calyptogena magnifica)	5	4	12	8	14	20	14	20	19	13	18	17	8	3	101
Endoriftia persephone 'Hot96_1+Hot96_2'	0	0	1	0	0	0	1	1	1	0	0	0	0	0	2
endosymbiont of Riftia pachyptila (vent Ph05)	1	1	5	4	4	8	5	8	5	4	4	6	2	2	13
endosymbiont of Tevnia jerichonana (vent Tica)	1	0	8	1	5	11	11	14	10	8	9	11	1	0	31
endosymbiont of Bathymodiolus sp.	4	2	7	7	13	12	13	9	15	9	12	12	3	2	61

Table B.3 Taxonomic breakdown for Sepetmber 2009 metaproteome continued

	% genome coverage in proteome														
	3S 100	S4 100	2S 100	3S 130	4S 130	2S 130	2S 150	3S 150	4S 150	2S 200	3S 200	4S 200	Apr08 100	Apr08 120	Apr08 200
Thaumarchaeota	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ca. Nitrosoarchaeum koreensis MY1	0.771	0.566	0.463	0.36	0.463	0.206	0.154	0.257	0.103	0.051	0.051	0.051	0.566	0.206	0.925
Ca. Nitrosoarchaeum limnia BG20	0.391	0.26	0.174	0.217	0.217	0.043	0	0.087	0	0	0	0	0.391	0.174	0.347
Ca. Nitrosoarchaeum limnia SFB1	0.491	0.343	0.196	0.245	0.343	0.196	0.147	0.196	0.147	0.147	0.147	0.147	0.54	0.196	0.393
Ca. Nitrosopumilus salaria BD31	1.903	1.068	0.557	0.511	0.836	0.279	0.232	0.325	0.232	0.093	0.093	0.093	2.182	0.232	2.646
Nitrosopumilus maritimus SCM1	2.394	1.503	0.668	0.891	1.336	0.445	0.39	0.612	0.334	0.223	0.278	0.167	2.728	0.557	3.786
Cenarchaeum symbiosum A	0.099	0.05	0.05	0	0.05	0	0	0	0	0	0	0	0.05	0	0.099
Ca Nitrospira defluvii	0.117	0.07	0.047	0.047	0.047	0.047	0.047	0.023	0.047	0.023	0.047	0.023	0.094	0.07	0.211
Planctomycetia	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ca. Kuenenia stuttgartiensis	0.043	0.021	0.086	0.043	0.086	0.129	0.086	0.086	0.107	0.086	0.107	0.064	0.107	0	0.322
Ca. Scalindua profunda	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Blastopirellula marina DSM 3645	0	0	0	0	0	0	0	0	0	0	0	0	0.017	0	0
Planctomyces brasiliensis DSM 5305	0.021	0	0	0	0	0	0	0	0	0	0	0	0	0	0.021
Planctomyces limnophilus DSM 3776	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.023
Planctomyces maris DSM 8797	0	0	0	0	0	0	0	0	0	0	0	0	0.015	0.015	0
Singulisphaera acidiphila DSM 18658	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.026
planctomycete KSU-1	0.056	0.056	0.056	0.056	0.056	0.083	0.083	0.083	0.056	0.028	0.028	0.028	0.056	0.056	0.111
SAR11 cluster	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
alpha proteobacterium HIMB114	0	0	0	0	0	0	0	0	0	0	0	0	0.07	0	0.14
Ca. Pelagibacter sp. HTCC7211	0.553	0.276	0.415	0.276	0.346	0.207	0.207	0.276	0.346	0.207	0.276	0.207	0.76	0.069	1.313
Ca. Pelagibacter sp. IMCC9063	0.069	0.069	0	0	0	0	0	0	0	0	0	0	0.069	0	0
Ca. Pelagibacter ubique HTCC1002	0.359	0.287	0.287	0.215	0.215	0.144	0.144	0.215	0	0	0	0	0.718	0.215	0.503
Ca. Pelagibacter ubique HTCC1062	0.739	0.739	0.369	0.369	0.591	0.295	0.295	0.443	0.222	0.148	0.148	0.148	0.96	0.222	1.108
ARCTIC96BD-19	1.01	0.79	1.13	0.79	0.79	1.01	0.79	1.01	1.01	0.68	0.68	1.35	1.35	0.23	4.95
SUP05	2.48	1.86	5.5	4.5	5.66	7.98	6.74	8.37	7.52	5.19	6.43	6.98	3.57	1.78	40.08
sulfur-oxidizing symbionts	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ca. Vesicomyosocius okutanii HA	0.32	0.107	0.747	0.32	0.747	0.961	0.747	0.961	1.067	0.747	0.747	0.747	0.32	0	5.977
Ca. Ruthia magnifica str. Cm (Calyptogena magnifica)	0.512	0.41	1.23	0.82	1.434	2.049	1.434	2.049	1.947	1.332	1.844	1.742	0.82	0.307	10.348
Endoriftia persephone 'Hot96_1+Hot96_2'	0	0	0.016	0	0	0	0.016	0.016	0.016	0	0	0	0	0	0.031
endosymbiont of Riftia pachyptila (vent Ph05)	0.031	0.031	0.157	0.126	0.126	0.251	0.157	0.251	0.157	0.126	0.126	0.189	0.063	0.063	0.409
endosymbiont of Tevnia jerichonana (vent Tica)	0.031	0	0.248	0.031	0.155	0.341	0.341	0.433	0.31	0.248	0.279	0.341	0.031	0	0.96
endosymbiont of Bathymodiolus sp.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

* completed genome not avaliable

Table B.4: Protein naming key. Table of full names of detected proteins for nitrogen and sulfur-based energy metabolism and inorganic carbon fixation pathways depicted in Figures 3.4 and B.1, B.2, B.3. Bold indicates enzyme is essential for carbon fixation pathway to be occurring

Pathway	Abreviation	Function
Nitrogen energy metabolism		
Nitrification	Amo	Ammonia monooxygenase
Anammox	Anx	hydroxylamine oxidoreductase, hydrazine oxidoreductases
Nitrification/Denitrification	Nar	Nitrate reductase
Denitrification	Nap	periplasmic nitrate reductase
Denitrification	NirS	nitrite reductase
Denitrification	NirK	copper containing nitrite reductase
Denitrification	Nor	nitric oxide reductase
Denitrification	NosZ	nitrous oxide reductase
Possible Dissimilatory nitrate reduction	HAO	Hydroxylamine oxidoreductase-like protein
Sulfur energy metabolism		
sulfide oxidation	Sqr	sulfide:quinone reductase
sulfide oxidation	Fcc	Fcc flavocytochrome C
sulfide oxidation	Sox	Sox sulfide oxidation protein complex
sulfide oxidation	Dsr	dissimilatory sulfate reductase pathway
sulfide oxidation	Apr	adenylylsulfate reductase
sulfide oxidation	Sat	ATP sulfurylase
Inorganic Carbon Fixation		
3-hydroxypropionate/4-hydroxybutyrate (3HP-4HB)		
3HP-4HB	3hb_dh	3-hydroxybutryryl-CoA dehydrogenase
3HP-4HB	3hp-dh	3-hydroxypropionyl-CoA dehydratase
3HP-4HB	ACoA_at	acetyl-CoA acetyltransferase
3HP-4HB	**ACoA_crb**	**acetyl-CoA carboxylase**[1]
3HP-4HB	cro_hy	crotonyl-CoA hydratase
3HP-4HB	mml_epi	methylmalonyl-CoA epimerase
3HP-4HB	**pro_cbx**	**propionyl CoA carboxylase**[1]
3HP-4HB	vya_iso	vinylacetyl-CoA isomerase
Calvin Benson Basham (CBB)		
CBB	6pp_kin	6-phosphofructokinase
CBB	f6p_atr	transketolase
CBB	fbp_ase	fructose-1,6-bisphosphatase
CBB	fbp_adl	fructose-bisphosphate aldolase
CBB	gp_dh	glyceraldehyde-3-phosphate dehydrogenase
CBB	p5p_epi	ribulose-5-phosphate 3-epimerase
CBB	pg_kin	3-phosphoglycerate kinase
CBB	**ppr_kin**	**phosphoribulokinase**
CBB	r5p_iso	ribose 5-phosphate isomerase
CBB	**RuBisCO**	**Ribulose 1,5-bisphosphate carboxylase**
CBB	tpp_iso	triosephosphate isomerase
Reductive tricarboxylic acid cycle		
rTCA	**oxy_syn**	**2-oxoglutarate synthase**
rTCA	cit_ly	aconitase B
rTCA	fum_hd	fumarate hydratase
rTCA	mal_dh	malate dehydrogenase
rTCA	pep_cbx	Phosphoenolpyruvate carboxykinase
rTCA	pep_syn	phosphoenolpyruvate synthase
rTCA	pyr_cbx	pyruvate carboxylase
rTCA	pyr_syn	pyruvate synthase
rTCA	sc_syn	succinyl-CoA synthetase
rTCA	sc_dh	Succinate dehydrogenase
Reductive Acetyl-CoA pathway		
rACoA	**ACoA_syn**	**acetyl-CoA synthetase**[2]
rACoA	**CO_dh**	**carbon-monoxide dehydrogenase**[2]
rACoA	fmt_syn	formyltetrahydrofolate synthetase
rACoA	fm_dh	formate dehydrogenase
rACoA	mtf_dh	methylene-tetrahydrofolate dehydrogenase

1. Within Nitrosopumulis maritimus these steps are proposed to be catalyzed by the same enzyme (Walker et al 2010).
2. These enzymes are often found as a single protein with dual functions.

Appendix C

Chapter 4: Supplementary material

Table C.1: Metagenome inventory for global fragment recruitment analysis. Inventory of environmental metagenomes with hits to Marinimicrobia SAGs in biogeography fragment recruitment analysis.

File Name	Accession	Data Repository	Location/sampleID	Specific Ecosystem	Metagenome Group	Latitude	Longitude	Depth (m)	Size (bp)	Sequencing Platform	Reference / Contact
CENF	SAMEA2619399	ENA	DCM_004_0.22-1.6	North Atlantic Subtropical Gyre	DCM_North Atlantic Ocean	36.5533	-6.5669	40m	4.27E+08	Illumina	Sunagawa et al. 2015
CENG	SAMEA2591057	ENA	SRF_007_0.22-1.6	Mediterranean Sea, Black Sea	SRF_Mediterranean Sea	37.051	1.9378	5m	2.25E+08	Illumina	Sunagawa et al. 2015
CENI	SAMEA2591107	ENA	DCM_023_0.22	Mediterranean Sea, Black Sea	DCM_Mediterranean Sea	42.1735	17.7252	55m	9.62E+07	Illumina	Sunagawa et al. 2015
CENJ	SAMEA2619531	ENA	SRF_009_0.22-1.6	Mediterranean Sea, Black Sea	SRF_Mediterranean Sea	39.1633	5.916	5m	4.36E+08	Illumina	Sunagawa et al. 2015
CENN	SAMEA2591084	ENA	SRF_023_0.22-1.6	Mediterranean Sea, Black Sea	SRF_Mediterranean Sea	42.2038	17.715	5m	1.97E+08	Illumina	Sunagawa et al. 2015
CENO	SAMEA2619857	ENA	SRF_033_0.22-1.6	Red Sea, Persian Gulf	SRF_Red Sea	21.9467	38.2517	5m	2.28E+08	Illumina	Sunagawa et al. 2015
CENP	SAMEA2591122	ENA	DCM_030_0.22-1.6	Mediterranean Sea, Black Sea	DCM_Mediterranean Sea	33.9235	32.8118	70m	5.23E+08	Illumina	Sunagawa et al. 2015
CENQ	SAMEA2619802	ENA	SRF_031_0.22-1.6	Red Sea, Persian Gulf	SRF_Red Sea	27.16	34.835	5m	2.90E+08	Illumina	Sunagawa et al. 2015
CENT	SAMEA2591074	ENA	DCM_007_0.22-1.6	Mediterranean Sea, Black Sea	DCM_Mediterranean Sea	37.0541	1.9478	42m	2.12E+08	Illumina	Sunagawa et al. 2015
CENU	SAMEA2619766	ENA	SRF_025_0.22-1.6	Mediterranean Sea, Black Sea	SRF_Mediterranean Sea	39.3888	19.3905	5m	4.67E+08	Illumina	Sunagawa et al. 2015
CENW	SAMEA2619818	ENA	SRF_032_0.22-1.6	Red Sea, Persian Gulf	SRF_Red Sea	23.36	37.2183	5m	2.90E+08	Illumina	Sunagawa et al. 2015
CENX	SAMEA2619678	ENA	DCM_018_0.22-1.6	Mediterranean Sea, Black Sea	DCM_Mediterranean Sea	35.7528	14.2765	60m	4.65E+08	Illumina	Sunagawa et al. 2015
CENY	SAMEA2619840	ENA	DCM_032_0.22-1.6	Red Sea, Persian Gulf	DCM_Red Sea	23.4183	37.245	80m	4.01E+08	Illumina	Sunagawa et al. 2015
CENZ	SAMEA2619782	ENA	DCM_025_0.22-1.6	Mediterranean Sea, Black Sea	DCM_Mediterranean Sea	39.3991	19.3997	50m	4.28E+08	Illumina	Sunagawa et al. 2015
CEOC	SAMEA2619952	ENA	DCM_036_0.22-1.6	Northwest Arabian Sea Upwelling	DCM_Indian Ocean	20.8222	63.5133	17m	3.30E+08	Illumina	Sunagawa et al. 2015
CEOF	SAMEA2619667	ENA	SRF_018_0.22-1.6	Mediterranean Sea, Black Sea	SRF_Mediterranean Sea	35.759	14.2574	5m	4.70E+08	Illumina	Sunagawa et al. 2015
CEOJ	SAMEA2619879	ENA	SRF_034_0.22-1.6	Red Sea, Persian Gulf	SRF_Red Sea	18.3967	39.875	5m	2.25E+08	Illumina	Sunagawa et al. 2015
CEOK	SAMEA2591108	ENA	SRF_030_0.22-1.6	Mediterranean Sea, Black Sea	SRF_Mediterranean Sea	33.9179	32.898	5m	4.77E+08	Illumina	Sunagawa et al. 2015
CEOM	SAMEA2619376	ENA	SRF_004_0.22-1.6	North Atlantic Subtropical Gyral	SRF_North Atlantic Ocean	36.5533	-6.5669	5m	3.91E+08	Illumina	Sunagawa et al. 2015
CEOO	SAMEA2619927	ENA	SRF_036_0.22-1.6	Northwest Arabian Sea Upwelling	SRF_Indian Ocean	20.8183	63.5047	5m	2.16E+08	Illumina	Sunagawa et al. 2015
CEOP	SAMEA2619548	ENA	DCM_009_0.22-1.6	Mediterranean Sea, Black Sea	DCM_Mediterranean Sea	39.0609	5.9422	55m	4.91E+08	Illumina	Sunagawa et al. 2015
CEOQ	SAMEA2619907	ENA	DCM_034_0.22-1.6	Red Sea, Persian Gulf	DCM_Red Sea	18.4417	39.8567	60m	6.06E+08	Illumina	Sunagawa et al. 2015
CEOR	SAMEA2620836	ENA	DCM_064_0.1-0.22	Eastern Africa Coastal	DCM_Indian Ocean	-29.5333	37.9117	65m	1.16E+08	Illumina	Sunagawa et al. 2015
CEOS	SAMEA2620666	ENA	MES_056_0.22-3	Eastern Africa Coastal	MES_Indian Ocean	-15.3379	43.2948	1000m	2.53E+08	Illumina	Sunagawa et al. 2015
CEOV	SAMEA2620756	ENA	SRF_062_0.22-3	Eastern Africa Coastal	SRF_Indian Ocean	-22.3368	40.3412	5m	2.33E+08	Illumina	Sunagawa et al. 2015
CEOW	SAMEA2620734	ENA	DCM_058_0.22-3	Eastern Africa Coastal	DCM_Indian Ocean	-17.2855	42.2866	66m	2.64E+08	Illumina	Sunagawa et al. 2015
CEOX	SAMEA2620890	ENA	DCM_065_0.22-3	Eastern Africa Coastal	DCM_Indian Ocean	-35.2421	26.3048	30m	2.05E+08	Illumina	Sunagawa et al. 2015
CEOY	SAMEA2620651	ENA	SRF_056_0.22-3	Eastern Africa Coastal	SRF_Indian Ocean	-15.3424	43.2965	5m	1.99E+08	Illumina	Sunagawa et al. 2015
CEOZ	SAMEA2620786	ENA	SRF_064_0.22-3	Eastern Africa Coastal	SRF_Indian Ocean	-29.5019	37.9889	5m	4.11E+08	Illumina	Sunagawa et al. 2015
CEPA	SAMEA2620542	ENA	SRF_052_0.22-1.6	Indian South Subtropical Gyre	SRF_Indian Ocean	-16.957	53.9801	5m	3.11E+08	Illumina	Sunagawa et al. 2015
CEPB	SAMEA2620672	ENA	SRF_057_0.22-3	Eastern Africa Coastal	SRF_Indian Ocean	-17.0248	42.7401	5m	2.20E+08	Illumina	Sunagawa et al. 2015
CEPC	SAMEA2620828	ENA	DCM_064_0.22-3	Eastern Africa Coastal	DCM_Indian Ocean	-29.5333	37.9117	65m	2.58E+08	Illumina	Sunagawa et al. 2015
CEPD	SAMEA2620815	ENA	MES_064_0.22-3	Eastern Africa Coastal	MES_Indian Ocean	-29.5046	37.9599	1000m	1.36E+08	Illumina	Sunagawa et al. 2015
CEPJ	SAMEA2620404	ENA	SRF_048_0.22-1.6	Indian South Subtropical Gyre	SRF_Indian Ocean	-9.3921	66.4228	5m	2.84E+08	Illumina	Sunagawa et al. 2015
CEPK	SAMEA2620259	ENA	DCM_042_0.22-1.6	Indian Monsoon Gyres	DCM_Indian Ocean	5.9998	73.9067	80m	3.65E+08	Illumina	Sunagawa et al. 2015
CEPL	SAMEA2620339	ENA	SRF_045_0.22-1.6	Indian Monsoon Gyres	SRF_Indian Ocean	0.0033	71.6428	5m	2.37E+08	Illumina	Sunagawa et al. 2015
CEPS	SAMEA2620035	ENA	MES_038_0.22-1.6	Indian Monsoon Gyres	MES_Indian Ocean	19.0351	64.5638	340m	2.19E+08	Illumina	Sunagawa et al. 2015
CEPT	SAMEA2620081	ENA	DCM_039_0.22-1.6	Indian Monsoon Gyres	DCM_Indian Ocean	18.5839	66.4727	25m	3.04E+08	Illumina	Sunagawa et al. 2015
CEPU	SAMEA2620194	ENA	SRF_041_0.22-1.6	Indian Monsoon Gyres	SRF_Indian Ocean	14.6059	69.9776	5m	2.74E+08	Illumina	Sunagawa et al. 2015
CEPV	SAMEA2620097	ENA	MES_039_0.22-1.6	Indian Monsoon Gyres	MES_Indian Ocean	18.7341	66.3896	270m	2.61E+08	Illumina	Sunagawa et al. 2015
CEPW	SAMEA2620021	ENA	DCM_038_0.22-1.6	Indian Monsoon Gyres	DCM_Indian Ocean	19.0284	64.5126	25m	3.38E+08	Illumina	Sunagawa et al. 2015
CEPX	SAMEA2619974	ENA	MES_037_0.1-0.22	Northwest Arabian Sea Upwelling	MES_Indian Ocean	20.8457	63.5851	600m	5.00E+08	Illumina	Sunagawa et al. 2015
CEPZ	SAMEA2620106	ENA	MES_039_0.1-0.22	Indian Monsoon Gyres	MES_Indian Ocean	18.7341	66.3896	270m	2.52E+08	Illumina	Sunagawa et al. 2015
CEQA	SAMEA2620227	ENA	DCM_041_0.22	Indian Monsoon Gyres	DCM_Indian Ocean	14.5536	70.0128	60m	1.38E+08	Illumina	Sunagawa et al. 2015
CEQD	SAMEA2620230	ENA	SRF_042_0.22-1.6	Indian Monsoon Gyres	SRF_Indian Ocean	6.0001	73.8955	5m	2.42E+08	Illumina	Sunagawa et al. 2015
CEQE	SAMEA2619970	ENA	MES_037_0.22-1.6	Northwest Arabian Sea Upwelling	MES_Indian Ocean	20.8457	63.5851	600m	3.84E+08	Illumina	Sunagawa et al. 2015
CEQF	SAMEA2621003	ENA	MES_068_0.45-0.8	South Atlantic Gyral	MES_South Atlantic Ocean	-31.0198	4.6685	700m	2.59E+08	Illumina	Sunagawa et al. 2015
CEQG	SAMEA2621242	ENA	MES_076_0.45-0.8	South Atlantic Gyral	MES_South Atlantic Ocean	-20.9315	-35.1794	800m	3.21E+08	Illumina	Sunagawa et al. 2015
CEQH	SAMEA2621232	ENA	MES_076_0.22-3	South Atlantic Gyral	MES_South Atlantic Ocean	-20.9315	-35.1794	800m	2.91E+08	Illumina	Sunagawa et al. 2015
CEQI	SAMEA2621204	ENA	SRF_076_0.45-0.8	South Atlantic Gyral	SRF_South Atlantic Ocean	-20.9354	-35.1803	5m	2.00E+08	Illumina	Sunagawa et al. 2015

C.1 Metagenome inventory for global fragment recruitment continued from previous page

File Name	Accession	Data Repository	Location/sampleID	Specific Ecosystem	Metagenome Group	Latitude	Longitude	Depth (m)	Size (bp)	Sequencing Platform	Reference / Contact
CEQJ	SAMEA2621198	ENA	SRF_076_0.22-3	South Atlantic Gyral	SRF_South Atlantic Ocean	-20.9354	-35.1803	5m	1.10E+08	Illumina	Sunagawa et al. 2015
CEQL	SAMEA2621203	ENA	SRF_076_0.22-0.45	South Atlantic Gyral	SRF_South Atlantic Ocean	-20.9354	-35.1803	5m	3.28E+08	Illumina	Sunagawa et al. 2015
CEQM	SAMEA2621278	ENA	DCM_078_0.45-0.8	South Atlantic Gyral	DCM_South Atlantic Ocean	-30.1484	-43.2705	120m	2.69E+08	Illumina	Sunagawa et al. 2015
CEQN	SAMEA2621277	ENA	DCM_078_0.22-0.45	South Atlantic Gyral	DCM_South Atlantic Ocean	-30.1484	-43.2705	120m	3.75E+08	Illumina	Sunagawa et al. 2015
CEQO	SAMEA2621272	ENA	DCM_078_0.22-3	South Atlantic Gyral	DCM_South Atlantic Ocean	-30.1484	-43.2705	120m	2.82E+08	Illumina	Sunagawa et al. 2015
CEQP	SAMEA2621221	ENA	DCM_076_0.22-0.45	South Atlantic Gyral	DCM_South Atlantic Ocean	-21.0292	-35.3498	150m	3.07E+08	Illumina	Sunagawa et al. 2015
CEQQ	SAMEA2621222	ENA	DCM_076_0.45-0.8	South Atlantic Gyral	DCM_South Atlantic Ocean	-21.0292	-35.3498	150m	2.50E+08	Illumina	Sunagawa et al. 2015
CEQR	SAMEA2621085	ENA	SRF_070_0.22	South Atlantic Gyral	SRF_South Atlantic Ocean	-20.4091	-3.1759	5m	1.13E+08	Illumina	Sunagawa et al. 2015
CEQS	SAMEA2621176	ENA	MES_072_0.22-3	South Atlantic Gyral	MES_South Atlantic Ocean	-8.7986	-17.9034	800m	1.65E+08	Illumina	Sunagawa et al. 2015
CEQT	SAMEA2621132	ENA	SRF_072_0.22-3	South Atlantic Gyral	SRF_South Atlantic Ocean	-8.7789	-17.9099	5m	2.77E+08	Illumina	Sunagawa et al. 2015
CEQU	SAMEA2621216	ENA	DCM_076_0.22-3	South Atlantic Gyral	DCM_South Atlantic Ocean	-21.0292	-35.3498	150m	2.77E+08	Illumina	Sunagawa et al. 2015
CEQW	SAMEA2621066	ENA	SRF_070_0.22-3	South Atlantic Gyral	SRF_South Atlantic Ocean	-20.4091	-3.1759	5m	1.31E+08	Illumina	Sunagawa et al. 2015
CEQX	SAMEA2621076	ENA	SRF_070_0.45-0.8	South Atlantic Gyral	SRF_South Atlantic Ocean	-20.4091	-3.1759	5m	3.69E+08	Illumina	Sunagawa et al. 2015
CEQY	SAMEA2621075	ENA	SRF_070_0.22-0.45	South Atlantic Gyral	SRF_South Atlantic Ocean	-20.4091	-3.1759	5m	4.03E+08	Illumina	Sunagawa et al. 2015
CERB	SAMEA2621155	ENA	DCM_072_0.22-3	South Atlantic Gyral	DCM_South Atlantic Ocean	-8.7296	-17.9604	100m	3.43E+08	Illumina	Sunagawa et al. 2015
CERC	SAMEA2621099	ENA	MES_070_0.22-0.45	South Atlantic Gyral	MES_South Atlantic Ocean	-20.4075	-3.1641	800m	1.86E+08	Illumina	Sunagawa et al. 2015
CERD	SAMEA2621101	ENA	MES_070_0.45-0.8	South Atlantic Gyral	MES_South Atlantic Ocean	-20.4075	-3.1641	800m	2.81E+08	Illumina	Sunagawa et al. 2015
CERE	SAMEA2621092	ENA	MES_070_0.22-3	South Atlantic Gyral	MES_South Atlantic Ocean	-20.4075	-3.1641	800m	2.26E+08	Illumina	Sunagawa et al. 2015
CERG	SAMEA2621037	ENA	DCM_068_0.22-3	South Atlantic Gyral	DCM_South Atlantic Ocean	-31.027	4.6802	50m	1.85E+08	Illumina	Sunagawa et al. 2015
CERI	SAMEA2620979	ENA	SRF_067_0.22-0.45	Benguela Current Coastal	SRF_South Atlantic Ocean	-32.2401	17.7103	5m	3.27E+08	Illumina	Sunagawa et al. 2015
CERJ	SAMEA2620967	ENA	DCM_066_0.22	Benguela Current Coastal	DCM_South Atlantic Ocean	-34.8901	18.0459	30m	8.73E+07	Illumina	Sunagawa et al. 2015
CERK	SAMEA2620950	ENA	DCM_066_0.22-3	Benguela Current Coastal	DCM_South Atlantic Ocean	-34.8901	18.0459	30m	1.29E+08	Illumina	Sunagawa et al. 2015
CERL	SAMEA2620995	ENA	MES_068_0.22-3	South Atlantic Gyral	MES_South Atlantic Ocean	-31.0198	4.6685	700m	2.38E+08	Illumina	Sunagawa et al. 2015
CERM	SAMEA2620970	ENA	SRF_067_0.22-3	Benguela Current Coastal	SRF_South Atlantic Ocean	-32.2401	17.7103	5m	2.55E+08	Illumina	Sunagawa et al. 2015
CERN	SAMEA2620947	ENA	SRF_066_0.22	Benguela Current Coastal	SRF_South Atlantic Ocean	-34.9449	17.9189	5m	1.03E+08	Illumina	Sunagawa et al. 2015
CERO	SAMEA2621033	ENA	SRF_068_0.22	South Atlantic Gyral	SRF_South Atlantic Ocean	-31.0266	4.665	5m	1.00E+08	Illumina	Sunagawa et al. 2015
CERP	SAMEA2620991	ENA	SRF_067_0.22	Benguela Current Coastal	SRF_South Atlantic Ocean	-32.2401	17.7103	5m	1.10E+08	Illumina	Sunagawa et al. 2015
CERQ	SAMEA2620882	ENA	MES_065_0.22-3	Eastern Africa Coastal	MES_Indian Ocean	-35.1889	26.2905	850m	2.46E+08	Illumina	Sunagawa et al. 2015
CERR	SAMEA2621021	ENA	SRF_068_0.45-0.8	South Atlantic Gyral	SRF_South Atlantic Ocean	-31.0266	4.665	5m	2.91E+08	Illumina	Sunagawa et al. 2015
CERS	SAMEA2621020	ENA	SRF_068_0.22-0.45	South Atlantic Gyral	SRF_South Atlantic Ocean	-31.0266	4.665	5m	3.16E+08	Illumina	Sunagawa et al. 2015
CERU	SAMEA2620925	ENA	DCM_065_0.22	Eastern Africa Coastal	DCM_Indian Ocean	-35.2421	26.3048	30m	1.24E+08	Illumina	Sunagawa et al. 2015
CERV	SAMEA2621045	ENA	DCM_068_0.45-0.8	South Atlantic Gyral	DCM_South Atlantic Ocean	-31.027	4.6802	50m	2.67E+08	Illumina	Sunagawa et al. 2015
CERW	SAMEA2620929	ENA	SRF_066_0.22-3	Benguela Current Coastal	SRF_South Atlantic Ocean	-34.9449	17.9189	5m	2.59E+08	Illumina	Sunagawa et al. 2015
CERX	SAMEA2621044	ENA	DCM_068_0.22-0.45	South Atlantic Gyral	DCM_South Atlantic Ocean	-31.027	4.6802	50m	2.96E+08	Illumina	Sunagawa et al. 2015
CERZ	SAMEA2621448	ENA	DCM_082_0.22	Southwest Atlantic Shelves	DCM_South Atlantic Ocean	-47.2007	-57.9446	40m	3.41E+07	Illumina	Sunagawa et al. 2015
CESA	SAMEA2621254	ENA	SRF_078_0.22-3	South Atlantic Gyral	SRF_South Atlantic Ocean	-30.1367	-43.2899	5m	2.11E+08	Illumina	Sunagawa et al. 2015
CESB	SAMEA2621259	ENA	SRF_078_0.22-0.45	South Atlantic Gyral	SRF_South Atlantic Ocean	-30.1367	-43.2899	5m	2.70E+08	Illumina	Sunagawa et al. 2015
CESC	SAMEA2621295	ENA	MES_078_0.45-0.8	South Atlantic Gyral	MES_South Atlantic Ocean	-30.1471	-43.2915	800m	2.07E+08	Illumina	Sunagawa et al. 2015
CESD	SAMEA2622021	ENA	DCM_098_0.22-3	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-25.826	-111.7294	188m	2.60E+08	Illumina	Sunagawa et al. 2015
CESF	SAMEA2621779	ENA	SRF_093_0.22-3	Chile-Peru Current Coastal	SRF_South Pacific Ocean	-34.0614	-73.1066	5m	3.45E+08	Illumina	Sunagawa et al. 2015
CESG	SAMEA2621287	ENA	MES_078_0.22-3	South Atlantic Gyral	MES_South Atlantic Ocean	-30.1471	-43.2915	800m	2.38E+08	Illumina	Sunagawa et al. 2015
CESI	SAMEA2621423	ENA	DCM_082_0.22-3	Southwest Atlantic Shelves	DCM_South Atlantic Ocean	-47.2007	-57.9446	40m	3.54E+08	Illumina	Sunagawa et al. 2015
CESJ	SAMEA2621487	ENA	SRF_084_0.22-3	Antarctic	SRF_Southern Ocean	-60.2287	-60.6476	5m	2.96E+08	Illumina	Sunagawa et al. 2015
CESL	SAMEA2622197	ENA	MES_102_0.22-3	Pacific Equatorial Divergence	MES_South Pacific Ocean	-5.261	-85.1678	480m	2.66E+08	Illumina	Sunagawa et al. 2015
CESM	SAMEA2622219	ENA	DCM_102_0.22-3	Pacific Equatorial Divergence	DCM_South Pacific Ocean	-5.2669	-85.2732	40m	5.38E+08	Illumina	Sunagawa et al. 2015
CESN	SAMEA2621990	ENA	SRF_098_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-25.8051	-111.7202	5m	1.89E+08	Illumina	Sunagawa et al. 2015
CESO	SAMEA2621812	ENA	DCM_093_0.22-3	Chile-Peru Current Coastal	DCM_South Pacific Ocean	-33.9116	-73.0537	35m	3.71E+08	Illumina	Sunagawa et al. 2015
CESP	SAMEA2622074	ENA	SRF_099_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-21.146	-104.787	5m	#VALUE!	Illumina	Sunagawa et al. 2015
CESQ	SAMEA2622119	ENA	DCM_100_0.22-3	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-12.9723	-96.0122	50m	4.51E+08	Illumina	Sunagawa et al. 2015
CESR	SAMEA2622149	ENA	MES_100_0.22-3	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-12.9794	-96.0232	177m	4.19E+08	Illumina	Sunagawa et al. 2015

C.1 Metagenome inventory for global fragment recruitment continued from previous page

File Name	Accession	Data Repository	Location/sampleID	Specific Ecosystem	Metagenome Group	Latitude	Longitude	Depth (m)	Size (bp)	Sequencing Platform	Reference / Contact
CESS	SAMEA2591098	ENA	DCM_023_0.22-1.6	Mediterranean Sea, Black Sea	DCM_Mediterranean Sea	42.1735	17.7252	55m	2.15E+08	Illumina	Sunagawa et al. 2015
CEST	SAMEA2621509	ENA	SRF_085_0.22-3	Antarctic	SRF_Southern Ocean	-62.0385	-49.529	5m	1.93E+08	Illumina	Sunagawa et al. 2015
CESV	SAMEA2622097	ENA	SRF_100_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-13.0023	-95.9759	5m	4.17E+08	Illumina	Sunagawa et al. 2015
CESW	SAMEA2621859	ENA	SRF_096_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-29.7238	-101.1604	5m	2.81E+08	Illumina	Sunagawa et al. 2015
CESY	SAMEA2622048	ENA	MES_098_0.22-3	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-25.8076	-111.6906	488m	2.60E+08	Illumina	Sunagawa et al. 2015
CESZ	SAMEA2621260	ENA	SRF_078_0.45-0.8	South Atlantic Gyral	SRF_South Atlantic Ocean	-30.1367	-43.2899	5m	2.06E+08	Illumina	Sunagawa et al. 2015
CETA	SAMEA2622362	ENA	MES_109_0.22-3	Chile-Peru Current Coastal	MES_North Pacific Ocean	2.0649	-84.5546	380m	2.65E+08	Illumina	Sunagawa et al. 2015
CETB	SAMEA2622545	ENA	DCM_112_0.22-3	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-23.2189	-129.4997	155m	3.66E+08	Illumina	Sunagawa et al. 2015
CETC	SAMEA2622694	ENA	DCM_122_0.1-0.22	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-9.0063	-139.1394	115m	2.02E+08	Illumina	Sunagawa et al. 2015
CETD	SAMEA2622402	ENA	DCM_110_0.22-3	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-1.9002	-84.6265	50m	3.71E+08	Illumina	Sunagawa et al. 2015
CETE	SAMEA2622336	ENA	DCM_109_0.22-3	Chile-Peru Current Coastal	DCM_North Pacific Ocean	2.0299	-84.5546	30m	2.80E+08	Illumina	Sunagawa et al. 2015
CETG	SAMEA2622696	ENA	DCM_122_0.45-0.8	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-9.0063	-139.1394	115m	4.76E+08	Illumina	Sunagawa et al. 2015
CETH	SAMEA2622695	ENA	DCM_122_0.22-0.45	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-9.0063	-139.1394	115m	5.53E+08	Illumina	Sunagawa et al. 2015
CETI	SAMEA2622478	ENA	DCM_111_0.22-3	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-16.9587	-100.6751	90m	5.02E+08	Illumina	Sunagawa et al. 2015
CETJ	SAMEA2622677	ENA	MES_122_0.22-0.45	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-8.9729	-139.2393	600m	2.00E+08	Illumina	Sunagawa et al. 2015
CETK	SAMEA2622452	ENA	SRF_111_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-16.9601	-100.6335	5m	3.91E+08	Illumina	Sunagawa et al. 2015
CETL	SAMEA2622518	ENA	SRF_112_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-23.2811	-129.3947	5m	3.13E+08	Illumina	Sunagawa et al. 2015
CETM	SAMEA2622690	ENA	DCM_122_0.22-3	South Pacific Subtropical Gyre	DCM_South Pacific Ocean	-9.0063	-139.1394	115m	5.46E+08	Illumina	Sunagawa et al. 2015
CETO	SAMEA2622678	ENA	MES_122_0.45-0.8	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-8.9729	-139.2393	600m	1.98E+08	Illumina	Sunagawa et al. 2015
CETP	SAMEA2622657	ENA	SRF_122_0.22-0.45	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9971	-139.1963	5m	1.98E+08	Illumina	Sunagawa et al. 2015
CETQ	SAMEA2622376	ENA	SRF_110_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-2.0133	-84.589	5m	2.76E+08	Illumina	Sunagawa et al. 2015
CETR	SAMEA2622568	ENA	MES_112_0.22-3	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-23.2232	-129.5986	696m	2.87E+08	Illumina	Sunagawa et al. 2015
CETS	SAMEA2622673	ENA	MES_122_0.22-3	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-8.9729	-139.2393	600m	3.08E+08	Illumina	Sunagawa et al. 2015
CETT	SAMEA2622429	ENA	MES_110_0.22-3	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-1.8902	-84.6141	380m	3.07E+08	Illumina	Sunagawa et al. 2015
CETU	SAMEA2623488	ENA	DCM_142_0.22-3	Caribbean	DCM_North Atlantic Ocean	25.6168	-88.4532	125m	3.58E+08	Illumina	Sunagawa et al. 2015
CETV	SAMEA2623155	ENA	MES_133_0.22-3	North Pacific Subtropical and Polar Fronts	MES_North Pacific Ocean	35.2698	-127.7268	650m	2.74E+08	Illumina	Sunagawa et al. 2015
CETW	SAMEA2620570	ENA	DCM_052_0.22-1.6	Indian South Subtropical Gyre	DCM_Indian Ocean	-16.9534	53.9601	75m	3.48E+08	Illumina	Sunagawa et al. 2015
CETX	SAMEA2623390	ENA	MES_138_0.22-3	North Pacific Equatorial Countercurrent	MES_North Pacific Ocean	6.3559	-103.0598	450m	3.57E+08	Illumina	Sunagawa et al. 2015
CETY	SAMEA2623135	ENA	DCM_133_0.22-3	North Pacific Subtropical and Polar Fronts	DCM_North Pacific Ocean	35.4002	-127.7499	45m	4.87E+08	Illumina	Sunagawa et al. 2015
CETZ	SAMEA2623098	ENA	MES_132_0.22-3	North Pacific Subtropical and Polar Fronts	MES_North Pacific Ocean	31.528	-159.0224	550m	2.60E+08	Illumina	Sunagawa et al. 2015
CEUA	SAMEA2623079	ENA	DCM_132_0.22-3	North Pacific Subtropical and Polar Fronts	DCM_North Pacific Ocean	31.5168	-159.046	115m	4.79E+08	Illumina	Sunagawa et al. 2015
CEUB	SAMEA2623370	ENA	DCM_138_0.22-3	North Pacific Equatorial Countercurrent	DCM_North Pacific Ocean	6.3378	-102.9538	60m	3.51E+08	Illumina	Sunagawa et al. 2015
CEUC	SAMEA2623295	ENA	DCM_137_0.22-3	North Pacific Equatorial Countercurrent	DCM_North Pacific Ocean	14.2075	-116.6468	40m	4.09E+08	Illumina	Sunagawa et al. 2015
CEUD	SAMEA2623314	ENA	MES_137_0.22-3	North Pacific Equatorial Countercurrent	MES_North Pacific Ocean	14.2025	-116.6433	375m	3.80E+08	Illumina	Sunagawa et al. 2015
CEUE	SAMEA2623116	ENA	SRF_133_0.22-3	North Pacific Subtropical and Polar Fronts	SRF_North Pacific Ocean	35.3671	-127.7422	5m	6.25E+08	Illumina	Sunagawa et al. 2015
CEUF	SAMEA2623350	ENA	SRF_138_0.22-3	North Pacific Equatorial Countercurrent	SRF_North Pacific Ocean	6.3332	-102.9432	5m	2.75E+08	Illumina	Sunagawa et al. 2015
CEUG	SAMEA2622901	ENA	SRF_128_0.22-3	Pacific Equatorial Divergence	SRF_South Pacific Ocean	0.0003	-153.6759	5m	2.67E+08	Illumina	Sunagawa et al. 2015
CEUH	SAMEA2623275	ENA	SRF_137_0.22-3	North Pacific Equatorial Countercurrent	SRF_North Pacific Ocean	14.2035	-116.6261	5m	3.13E+08	Illumina	Sunagawa et al. 2015
CEUJ	SAMEA2622842	ENA	MIX_125_0.22-0.45	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-8.8999	-142.5461	140m	4.16E+08	Illumina	Sunagawa et al. 2015
CEUK	SAMEA2622843	ENA	MIX_125_0.45-0.8	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-8.8999	-142.5461	140m	4.15E+08	Illumina	Sunagawa et al. 2015
CEUL	SAMEA2622800	ENA	MIX_124_0.22-0.45	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-9.0714	-140.5973	120m	4.15E+08	Illumina	Sunagawa et al. 2015
CEUM	SAMEA2622716	ENA	SRF_123_0.45-0.8	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9068	-140.283	5m	3.99E+08	Illumina	Sunagawa et al. 2015
CEUN	SAMEA2622801	ENA	MIX_124_0.45-0.8	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-9.0714	-140.5973	120m	5.25E+08	Illumina	Sunagawa et al. 2015
CEUO	SAMEA2622799	ENA	MIX_124_0.1-0.22	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-9.0714	-140.5973	120m	2.52E+08	Illumina	Sunagawa et al. 2015
CEUP	SAMEA2622738	ENA	MIX_123_0.45-0.8	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-8.9109	-140.2845	150m	1.27E+08	Illumina	Sunagawa et al. 2015
CEUQ	SAMEA2622796	ENA	MIX_124_0.22-3	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-9.0714	-140.5973	120m	5.25E+08	Illumina	Sunagawa et al. 2015
CEUR	SAMEA2622715	ENA	SRF_123_0.22-0.45	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9068	-140.283	5m	3.30E+08	Illumina	Sunagawa et al. 2015
CEUS	SAMEA2622837	ENA	MIX_125_0.22-3	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-8.8999	-142.5461	140m	6.69E+08	Illumina	Sunagawa et al. 2015
CEUT	SAMEA2622821	ENA	SRF_125_0.1-0.22	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9111	-142.5571	5m	1.94E+08	Illumina	Sunagawa et al. 2015
CEUU	SAMEA2622658	ENA	SRF_122_0.45-0.8	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9971	-139.1963	5m	2.67E+08	Illumina	Sunagawa et al. 2015

C.1 Metagenome inventory for global fragment recruitment continued from previous page

File Name	Accession	Data Repository	Location/sampleID	Specific Ecosystem	Metagenome Group	Latitude	Longitude	Depth (m)	Size (bp)	Sequencing Platform	Reference / Contact
CEUV	SAMEA2622710	ENA	SRF_123_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9068	-140.283	5m	3.58E+08	Illumina	Sunagawa et al. 2015
CEUW	SAMEA2622652	ENA	SRF_122_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9971	-139.1963	5m	2.47E+08	Illumina	Sunagawa et al. 2015
CEUY	SAMEA2622737	ENA	MIX_123_0.22-0.45	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-8.9109	-140.2845	150m	5.50E+08	Illumina	Sunagawa et al. 2015
CEVB	SAMEA2622763	ENA	SRF_124_0.1-0.22	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-9.1504	-140.5216	5m	1.98E+08	Illumina	Sunagawa et al. 2015
CEVC	SAMEA2622817	ENA	SRF_125_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-8.9111	-142.5571	5m	3.72E+08	Illumina	Sunagawa et al. 2015
CEVD	SAMEA2622733	ENA	MIX_123_0.22-3	South Pacific Subtropical Gyre	MIX_South Pacific Ocean	-8.9109	-140.2845	150m	6.12E+08	Illumina	Sunagawa et al. 2015
CEVE	SAMEA2622764	ENA	SRF_124_0.22-0.45	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-9.1504	-140.5216	5m	3.02E+08	Illumina	Sunagawa et al. 2015
CEVF	SAMEA2622765	ENA	SRF_124_0.45-0.8	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-9.1504	-140.5216	5m	3.71E+08	Illumina	Sunagawa et al. 2015
CEVG	SAMEA2623426	ENA	SRF_140_0.22-3	Central American Coastal	SRF_North Pacific Ocean	7.4122	-79.3017	5m	3.43E+08	Illumina	Sunagawa et al. 2015
CEVH	SAMEA2623446	ENA	SRF_141_0.22-3	Guianas Coastal	SRF_North Atlantic Ocean	9.8481	-80.0454	5m	3.53E+08	Illumina	Sunagawa et al. 2015
CEVI	SAMEA2623513	ENA	MES_142_0.22-3	Caribbean	MES_North Atlantic Ocean	25.6236	-88.45	640m	2.30E+08	Illumina	Sunagawa et al. 2015
CEVJ	SAMEA2623693	ENA	MES_146_0.22-3	North Atlantic Subtropical Gyral	MES_North Atlantic Ocean	34.6663	-71.2907	640m	1.80E+08	Illumina	Sunagawa et al. 2015
CEVK	SAMEA2623794	ENA	MES_149_0.22-3	North Atlantic Subtropical Gyral	MES_North Atlantic Ocean	34.0771	-49.8233	740m	2.04E+08	Illumina	Sunagawa et al. 2015
CEVL	SAMEA2623673	ENA	SRF_146_0.22-3	North Atlantic Subtropical Gyral	SRF_North Atlantic Ocean	34.6712	-71.3093	5m	3.91E+08	Illumina	Sunagawa et al. 2015
CEVM	SAMEA2623649	ENA	MES_145_0.22-3	Gulf Stream	MES_North Atlantic Ocean	39.2392	-70.0343	590m	2.80E+08	Illumina	Sunagawa et al. 2015
CEVN	SAMEA2623463	ENA	SRF_142_0.22-3	Caribbean	SRF_North Atlantic Ocean	25.5264	-88.394	5m	3.89E+08	Illumina	Sunagawa et al. 2015
CEVO	SAMEA2623907	ENA	MES_152_0.22-3	North Atlantic Subtropical Gyral	MES_North Atlantic Ocean	43.7182	-16.8714	800m	2.49E+08	Illumina	Sunagawa et al. 2015
CEVP	SAMEA2623808	ENA	SRF_150_0.22-3	North Atlantic Subtropical Gyral	SRF_North Atlantic Ocean	35.9346	-37.3032	5m	3.44E+08	Illumina	Sunagawa et al. 2015
CEVQ	SAMEA2623756	ENA	MES_148b_0.22-3	North Atlantic Subtropical Gyral	MES_North Atlantic Ocean	34.1504	-56.9684	250m	3.75E+08	Illumina	Sunagawa et al. 2015
CEVR	SAMEA2623734	ENA	SRF_148_0.22-3	North Atlantic Subtropical Gyral	SRF_North Atlantic Ocean	31.6948	-64.2489	5m	3.51E+08	Illumina	Sunagawa et al. 2015
CEVS	SAMEA2623850	ENA	SRF_151_0.22-3	North Atlantic Subtropical Gyral	SRF_North Atlantic Ocean	36.1715	-29.023	5m	4.00E+08	Illumina	Sunagawa et al. 2015
CEVT	SAMEA2623826	ENA	DCM_150_0.22-3	North Atlantic Subtropical Gyral	DCM_North Atlantic Ocean	35.8427	-37.1526	40m	3.72E+08	Illumina	Sunagawa et al. 2015
CEVU	SAMEA2623868	ENA	DCM_151_0.22-3	North Atlantic Subtropical Gyral	DCM_North Atlantic Ocean	36.1811	-28.9373	80m	3.63E+08	Illumina	Sunagawa et al. 2015
CEVV	SAMEA2623919	ENA	MIX_152_0.22-3	North Atlantic Subtropical Gyral	MIX_North Atlantic Ocean	43.7056	-16.8794	25m	3.81E+08	Illumina	Sunagawa et al. 2015
CEVW	SAMEA2623774	ENA	SRF_149_0.22-3	North Atlantic Subtropical Gyral	SRF_North Atlantic Ocean	34.1132	-49.9181	5m	3.73E+08	Illumina	Sunagawa et al. 2015
CEVX	SAMEA2623886	ENA	SRF_152_0.22-3	North Atlantic Subtropical Gyral	SRF_North Atlantic Ocean	43.6792	-16.8344	5m	3.43E+08	Illumina	Sunagawa et al. 2015
CEVY	SAMEA2620855	ENA	SRF_065_0.22-3	Eastern Africa Coastal	SRF_Indian Ocean	-35.1728	26.2868	5m	1.85E+08	Illumina	Sunagawa et al. 2015
CEVZ	SAMEA2620217	ENA	DCM_041_0.22-1.6	Indian Monsoon Gyres	DCM_Indian Ocean	14.5536	70.0128	60m	4.64E+08	Illumina	Sunagawa et al. 2015
CEWA	SAMEA2620000	ENA	SRF_038_0.22-1.6	Indian Monsoon Gyres	SRF_Indian Ocean	19.0393	64.4913	5m	2.00E+08	Illumina	Sunagawa et al. 2015
CEWB	SAMEA2621013	ENA	SRF_068_0.22-3	South Atlantic Gyral	SRF_South Atlantic Ocean	-31.0266	4.665	5m	1.97E+08	Illumina	Sunagawa et al. 2015
CEWE	SAMEA2622759	ENA	SRF_124_0.22-3	South Pacific Subtropical Gyre	SRF_South Pacific Ocean	-9.1504	-140.5216	5m	6.14E+08	Illumina	Sunagawa et al. 2015
CEWG	SAMEA2622173	ENA	SRF_102_0.22-3	Pacific Equatorial Divergence	SRF_South Pacific Ocean	-5.2529	-85.1545	5m	4.27E+08	Illumina	Sunagawa et al. 2015
CEWH	SAMEA2622316	ENA	SRF_109_0.22-3	Chile-Peru Current Coastal	SRF_North Pacific Ocean	1.9928	-84.5766	5m	2.82E+08	Illumina	Sunagawa et al. 2015
CEWI	SAMEA2622499	ENA	MES_111_0.22-3	South Pacific Subtropical Gyre	MES_South Pacific Ocean	-16.9486	-100.6715	350m	2.63E+08	Illumina	Sunagawa et al. 2015
CEWJ	SAMEA2622923	ENA	DCM_128_0.22-3	Pacific Equatorial Divergence	DCM_South Pacific Ocean	0.0222	-153.6858	40m	3.11E+08	Illumina	Sunagawa et al. 2015
CEWK	SAMEA2623059	ENA	SRF_132_0.22-3	North Pacific Subtropical and Polar	SRF_North Pacific Ocean	31.5213	-158.9958	5m	2.78E+08	Illumina	Sunagawa et al. 2015
CEWO	SAMEA2623627	ENA	SRF_145_0.22-3	Gulf Stream	SRF_North Atlantic Ocean	39.2305	-70.0377	5m	4.23E+08	Illumina	Sunagawa et al. 2015
CEWP	SAMEA2620980	ENA	SRF_067_0.45-0.8	Benguela Current Coastal	SRF_South Atlantic Ocean	-32.2401	17.7103	5m	4.14E+08	Illumina	Sunagawa et al. 2015
CEWR	SAMEA2621551	ENA	MES_085_0.22-3	Antarctic	MES_Southern Ocean	-61.9689	-49.5017	790m	3.59E+08	Illumina	Sunagawa et al. 2015
engcyc_2081372001	2081372001	IMG/M taxon oid	Deepwater Horizon Oil Spill	Oil-contaminated	Gulf of Mexico	28.672222	-88.4375	unknown	2.38E+07	Illumina	Mason et al. 2012 Janet Jansson
engcyc_2088090017	2088090017	IMG/M taxon oid	Deepwater Horizon Oil Spill	Oil-contaminated	Gulf of Mexico	28.672222	-88.4375	unknown	2.53E+07	Illumina	Mason et al. 2012 Janet Jansson
engcyc_2149837025	2149837025	IMG/M taxon oid	Gulf of Mexico	Black smokers	Gulf of Mexico	28.716667	-88.466667	1250m	5.25E+07	Illumina	Adam R. Rivers
engcyc_2149837027	2149837027	IMG/M taxon oid	Gulf of Mexico	Black smokers	Gulf of Mexico	28.716667	-88.466667	1300m	4.36E+07	Illumina	Adam R. Rivers
engcyc_2149837028	2149837028	IMG/M taxon oid	Gulf of Mexico	Black smokers	Gulf of Mexico	28.716667	-88.466667	1210m	5.26E+07	Illumina	Adam R. Rivers
engcyc_2236347000	2236347000	IMG/M taxon oid	Guaymas Basin	Hydrothermal vent plume	Guaymas Basin	27.823	-111.4	1996m	2.11E+07	Illumina HiSeq	Gregory Dick
engcyc_2263328000	2263328000	IMG/M taxon oid	Sakinaw Lake, BC, Canada	meromitic lake	Sakinaw Lake	49.682207	-124.005217	120m	4.37E+08	454	Rinke et al. 2013 / Tanja Woyke
engcyc_3300001139	3300001139	IMG/M taxon oid	Iowa, USA	Grasslands	Iowa	43.303333	-89.3325	NA	1.90E+09	454-GS-FLX, Illumina GAii	Janet Jansson
engcyc_3300001683	3300001683	IMG/M taxon oid	Guyams Basin	Hydrothermal vent plume	Guaymas Basin	27.015833	-111.425	1993 m	5.72E+08	Illumina HiSeq	Gregory Dick
engcyc_3300002908	3300002908	IMG/M taxon oid	Angelo Coastal Reserve	Grasslands	Angelo Coastal Reserve	39.718176	-123.652732	NA	9.67E+08	Illumina HiSeq	Jill Banfield
engcyc_3300004069	3300004069	IMG/M taxon oid	Juan de Fuca Ridge flank	Hydrothermal vents	Juan de Fuca Ridge flank	47.76	-127.76	2667m	3.90E+08	Illumina HiSeq	Ramunas Stepanauskas

C.1 Metagenome inventory for global fragment recruitment continued from previous page

File Name	Accession	Data Repository	Location/sampleID	Specific Ecosystem	Metagenome Group	Latitude	Longitude	Depth (m)	Size (bp)	Sequencing Platform	Reference / Contact
G0K4BGD03	4461588.3	MG-RAST	mgm4461588.3	Coastal waters off Lima	PERU_40-80m	-12.37444	-77	60m	1.39E+08	pyrosequencing	Schunck et al. 2013
GJCUZF103	4450891.3	MG-RAST	mgm4450891.3	Coastal waters off Lima	PERU_40-80m	-12.37444	-77	40m	1.19E+08	pyrosequencing	Schunck et al. 2013
GJCUZF104	4450892.3	MG-RAST	mgm4450892.3	Coastal waters off Lima	PERU_20m	-12.37444	-77	20m	1.01E+08	pyrosequencing	Schunck et al. 2013
GXED30K01	4460676.3	MG-RAST	mgm4460676.3	Coastal waters off Lima	PERU_40-80m	-12.37444	-77	80m	1.54E+08	pyrosequencing	Schunck et al. 2013
GXED30K02	4460677.3	MG-RAST	mgm4460677.3	Coastal waters off Lima	PERU_5m	-12.37444	-77	5m	1.22E+08	pyrosequencing	Schunck et al. 2013
GZ2L4FS03	4460736.3	MG-RAST	mgm4460736.3	Coastal waters off Lima	PERU_40-80m	-12.37444	-77	50m	1.24E+08	pyrosequencing	Schunck et al. 2013
SI_MetaG_1039680	SAMN05224482	NCBI bioproject	SI037_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.70E+09	Illumnia	Steven Hallam
SI_MetaG_1039686	SAMN05224486	NCBI bioproject	SI037_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	1.25E+09	Illumnia	Steven Hallam
SI_MetaG_1039689	SAMN05224487	NCBI bioproject	SI037_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	1.48E+09	Illumnia	Steven Hallam
SI_MetaG_1039704	SAMN05224440	NCBI bioproject	SI072_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	1.50E+09	Illumnia	Steven Hallam
SI_MetaG_1039707	SAMN05224441	NCBI bioproject	SI072_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.98E+09	Illumnia	Steven Hallam
SI_MetaG_1039713	SAMN05224513	NCBI bioproject	SI072_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.70E+09	Illumnia	Steven Hallam
SI_MetaG_1039716	SAMN05224518	NCBI bioproject	SI072_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	1.39E+09	Illumnia	Steven Hallam
SI_MetaG_1039719	SAMN05224519	NCBI bioproject	SI072_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	7.57E+08	Illumnia	Steven Hallam
SI_MetaG_1039722	SAMN05224534	NCBI bioproject	SI073_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	1.33E+09	Illumnia	Steven Hallam
SI_MetaG_1039725	SAMN05224524	NCBI bioproject	SI073_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.62E+09	Illumnia	Steven Hallam
SI_MetaG_1039728	SAMN05224525	NCBI bioproject	SI073_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	1.66E+09	Illumnia	Steven Hallam
SI_MetaG_1039731	SAMN05224530	NCBI bioproject	SI073_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.79E+09	Illumnia	Steven Hallam
SI_MetaG_1039734	SAMN05224531	NCBI bioproject	SI073_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	1.69E+09	Illumnia	Steven Hallam
SI_MetaG_1039737	SAMN05224508	NCBI bioproject	SI073_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	7.87E+08	Illumnia	Steven Hallam
SI_MetaG_1039740	SAMN05224529	NCBI bioproject	SI074_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	8.44E+08	Illumnia	Steven Hallam
SI_MetaG_1039743	SAMN05224509	NCBI bioproject	SI074_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	7.16E+08	Illumnia	Steven Hallam
SI_MetaG_1039746	SAMN05224514	NCBI bioproject	SI074_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	9.69E+08	Illumnia	Steven Hallam
SI_MetaG_1039749	SAMN05224515	NCBI bioproject	SI074_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.74E+09	Illumnia	Steven Hallam
SI_MetaG_1039752	SAMN05224528	NCBI bioproject	SI074_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	1.91E+09	Illumnia	Steven Hallam
SI_MetaG_1039755	SAMN05224520	NCBI bioproject	SI074_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	1.18E+09	Illumnia	Steven Hallam
SI_MetaG_1040232	SAMN05224521	NCBI bioproject	SI037_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	1.03E+09	Illumnia	Steven Hallam
SI_MetaG_1040238	SAMN05224527	NCBI bioproject	SI037_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	1.07E+09	Illumnia	Steven Hallam
SI_MetaG_1057018	SAMN05224536	NCBI bioproject	SI075_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	1.54E+09	Illumnia	Steven Hallam
SI_MetaG_1057019	SAMN05224522	NCBI bioproject	SI075_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.53E+09	Illumnia	Steven Hallam
SI_MetaG_1057020	SAMN05224493	NCBI bioproject	SI075_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	8.17E+08	Illumnia	Steven Hallam
SI_MetaG_1057021	SAMN05224494	NCBI bioproject	SI075_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	7.86E+08	Illumnia	Steven Hallam
SI_MetaG_1057022	SAMN05224495	NCBI bioproject	SI075_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	1.34E+09	Illumnia	Steven Hallam
SI4093112	SAMN05224413	NCBI bioproject	LP_J09_P20_500m	North eastern subartic pacific	LP_500m	49.567	-138.664	500m	6.80E+07	Illumnia	Steven Hallam
SI4093113	SAMN05224425	NCBI bioproject	LP_A09_P04_10m	North eastern subartic pacific	LP_10m	48.651	-126.667	10m	1.62E+08	Illumnia	Steven Hallam
SI4093125	SAMN05224430	NCBI bioproject	LP_A09_P04_500m	North eastern subartic pacific	LP_500m	48.651	-126.667	500m	1.00E+08	Illumnia	Steven Hallam
SI4093127	SAMN05224418	NCBI bioproject	LP_A09_P04_1000m	North eastern subartic pacific	LP_1000m	48.651	-126.667	1000m	5.74E+07	Illumnia	Steven Hallam
SI4093128	SAMN05224419	NCBI bioproject	LP_A09_P04_1300m	North eastern subartic pacific	LP_1300m	48.651	-126.667	1300m	1.31E+08	Illumnia	Steven Hallam
SI4093129	SAMN05224424	NCBI bioproject	LP_A09_P20_1000m	North eastern subartic pacific	LP_1000m	49.567	-138.664	1000m	1.01E+08	Illumnia	Steven Hallam
SI4093130	SAMN05224446	NCBI bioproject	LP_A09_P20_500m	North eastern subartic pacific	LP_500m	49.567	-138.664	500m	1.17E+08	Illumnia	Steven Hallam
SI4093131	SAMN05224427	NCBI bioproject	LP_J08_P16_500m	North eastern subartic pacific	LP_500m	49.283	-134.666	500m	9.79E+07	Illumnia	Steven Hallam
SI4093132	SAMN05224450	NCBI bioproject	LP_J09_P20_1000m	North eastern subartic pacific	LP_1000m	49.567	-138.664	1000m	9.26E+07	Illumnia	Steven Hallam
SI4093144	SAMN05224451	NCBI bioproject	SI042_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	1.46E+08	Illumnia	Steven Hallam
SI4093145	SAMN05224447	NCBI bioproject	SI042_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.95E+08	Illumnia	Steven Hallam
SI4093146	SAMN05224436	NCBI bioproject	SI042_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	1.94E+08	Illumnia	Steven Hallam
SI4093147	SAMN05224437	NCBI bioproject	SI042_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.68E+08	Illumnia	Steven Hallam
SI4093148	SAMN05224442	NCBI bioproject	SI042_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	1.39E+08	Illumnia	Steven Hallam
SI4093149	SAMN05224443	NCBI bioproject	SI042_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	1.57E+08	Illumnia	Steven Hallam
SI4096364	SAMN05224469	NCBI bioproject	LP_A08_P12_1000m	North eastern subartic pacific	LP_1000m	48.97	-130.666	1000m	9.31E+07	Illumnia	Steven Hallam
SI4096365	SAMN05213796	NCBI bioproject	LP_A08_P12_2000m	North eastern subartic pacific	LP_2000m	48.97	-130.666	2000m	1.48E+08	Illumnia	Steven Hallam

C.1 Metagenome inventory for global fragment recruitment continued from previous page

File Name	Accession	Data Repository	Location/sampleID	Specific Ecosystem	Metagenome Group	Latitude	Longitude	Depth (m)	Size (bp)	Sequencing Platform	Reference / Contact
SI4096367	SAMN05213798	NCBI bioproject	LP_A08_P20_500m	North eastern subartic pacific	LP_500m	49.567	-138.664	500m	4.77E+07	Illumnia	Steven Hallam
SI4096368	SAMN05224403	NCBI bioproject	LP_J09_P12_10m	North eastern subartic pacific	LP_10m	48.97	-130.666	10m	5.09E+07	Illumnia	Steven Hallam
SI4096369	SAMN05213797	NCBI bioproject	LP_A08_P20_2000m	North eastern subartic pacific	LP_2000m	49.567	-138.664	2000m	6.44E+07	Illumnia	Steven Hallam
SI4096370	SAMN05224414	NCBI bioproject	LP_A09_P16_10m	North eastern subartic pacific	LP_10m	49.283	-134.666	10m	1.01E+08	Illumnia	Steven Hallam
SI4096371	SAMN05224420	NCBI bioproject	LP_A09_P16_500m	North eastern subartic pacific	LP_500m	49.283	-134.666	500m	1.48E+08	Illumnia	Steven Hallam
SI4096373	SAMN05224415	NCBI bioproject	LP_A09_P16_2000m	North eastern subartic pacific	LP_2000m	49.283	-134.666	2000m	4.85E+07	Illumnia	Steven Hallam
SI4096375	SAMN05224449	NCBI bioproject	LP_A09_P20_2000m	North eastern subartic pacific	LP_2000m	49.567	-138.664	2000m	1.32E+08	Illumnia	Steven Hallam
SI4096377	SAMN05224426	NCBI bioproject	LP_F10_P16_500m	North eastern subartic pacific	LP_500m	49.283	-134.666	500m	5.55E+07	Illumnia	Steven Hallam
SI4096378	SAMN05224432	NCBI bioproject	LP_F10_P16_1000m	North eastern subartic pacific	LP_1000m	49.283	-134.666	1000m	1.69E+08	Illumnia	Steven Hallam
SI4096379	SAMN05224421	NCBI bioproject	LP_F10_P16_2000m	North eastern subartic pacific	LP_2000m	49.283	-134.666	2000m	3.17E+07	Illumnia	Steven Hallam
SI4096381	SAMN05224400	NCBI bioproject	LP_J08_P04_500m	North eastern subartic pacific	LP_500m	48.651	-126.667	500m	1.11E+08	Illumnia	Steven Hallam
SI4096382	SAMN05224396	NCBI bioproject	LP_J08_P04_1000m	North eastern subartic pacific	LP_1000m	48.651	-126.667	1000m	1.80E+07	Illumnia	Steven Hallam
SI4096383	SAMN05224468	NCBI bioproject	LP_J08_P04_1300m	North eastern subartic pacific	LP_1300m	48.651	-126.667	1300m	1.15E+08	Illumnia	Steven Hallam
SI4096385	SAMN05224448	NCBI bioproject	LP_J09_P16_500m	North eastern subartic pacific	LP_500m	49.283	-134.666	500m	1.26E+08	Illumnia	Steven Hallam
SI4096386	SAMN05224475	NCBI bioproject	LP_J09_P16_1000m	North eastern subartic pacific	LP_1000m	49.283	-134.666	1000m	1.42E+08	Illumnia	Steven Hallam
SI4096387	SAMN05224445	NCBI bioproject	LP_J09_P16_2000m	North eastern subartic pacific	LP_2000m	49.283	-134.666	2000m	1.31E+08	Illumnia	Steven Hallam
SI4096389	SAMN05224452	NCBI bioproject	LP_J09_P20_2000m	North eastern subartic pacific	LP_2000m	49.567	-138.664	2000m	4.20E+07	Illumnia	Steven Hallam
SI4096390	SAMN05224470	NCBI bioproject	LP_A08_P26_10m	North eastern subartic pacific	LP_10m	50	-145	10m	8.19E+07	Illumnia	Steven Hallam
SI4096391	SAMN05224471	NCBI bioproject	LP_A08_P26_500m	North eastern subartic pacific	LP_500m	50	-145	500m	1.72E+07	Illumnia	Steven Hallam
SI4096392	SAMN05224488	NCBI bioproject	LP_A08_P26_1000m	North eastern subartic pacific	LP_1000m	50	-145	1000m	1.13E+08	Illumnia	Steven Hallam
SI4096394	SAMN05224453	NCBI bioproject	LP_A09_P26_500m	North eastern subartic pacific	LP_500m	50	-145	500m	1.17E+08	Illumnia	Steven Hallam
SI4096395	SAMN05224461	NCBI bioproject	LP_F09_P12_500m	North eastern subartic pacific	LP_500m	48.97	-130.666	500m	9.20E+07	Illumnia	Steven Hallam
SI4096396	SAMN05224460	NCBI bioproject	LP_F09_P12_1000m	North eastern subartic pacific	LP_1000m	48.97	-130.666	1000m	9.63E+06	Illumnia	Steven Hallam
SI4096398	SAMN05224457	NCBI bioproject	LP_F09_P26_500m	North eastern subartic pacific	LP_500m	50	-145	500m	8.23E+07	Illumnia	Steven Hallam
SI4096399	SAMN05224456	NCBI bioproject	LP_F09_P26_1000m	North eastern subartic pacific	LP_1000m	50	-145	1000m	1.38E+08	Illumnia	Steven Hallam
SI4096400	SAMN05224431	NCBI bioproject	LP_J08_P26_500m	North eastern subartic pacific	LP_500m	50	-145	500m	1.18E+08	Illumnia	Steven Hallam
SI4096401	SAMN05224485	NCBI bioproject	LP_J08_P16_1000m	North eastern subartic pacific	LP_1000m	49.283	-134.666	1000m	4.28E+07	Illumnia	Steven Hallam
SI4096402	SAMN05224395	NCBI bioproject	LP_J08_P12_500m	North eastern subartic pacific	LP_500m	48.97	-130.666	500m	1.33E+08	Illumnia	Steven Hallam
SI4096403	SAMN05224464	NCBI bioproject	LP_J08_P12_1000m	North eastern subartic pacific	LP_1000m	48.97	-130.666	1000m	1.77E+07	Illumnia	Steven Hallam
SI4096404	SAMN05224465	NCBI bioproject	LP_J08_P12_2000m	North eastern subartic pacific	LP_2000m	48.97	-130.666	2000m	4.07E+07	Illumnia	Steven Hallam
SI4096405	SAMN05224476	NCBI bioproject	LP_J09_P12_500m	North eastern subartic pacific	LP_500m	48.97	-130.666	500m	1.74E+08	Illumnia	Steven Hallam
SI4096409	SAMN05224402	NCBI bioproject	SI034_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	1.48E+08	Illumnia	Steven Hallam
SI4096410	SAMN05224404	NCBI bioproject	SI034_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.93E+08	Illumnia	Steven Hallam
SI4096411	SAMN05224407	NCBI bioproject	SI034_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	2.13E+08	Illumnia	Steven Hallam
SI4096412	SAMN05224408	NCBI bioproject	SI034_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	2.38E+08	Illumnia	Steven Hallam
SI4096413	SAMN05224411	NCBI bioproject	SI034_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	7.78E+07	Illumnia	Steven Hallam
SI4096414	SAMN05224484	NCBI bioproject	SI034_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	2.01E+08	Illumnia	Steven Hallam
SI4096416	SAMN05224405	NCBI bioproject	SI036_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.03E+08	Illumnia	Steven Hallam
SI4096417	SAMN05224472	NCBI bioproject	SI036_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	2.15E+08	Illumnia	Steven Hallam
SI4096418	SAMN05224406	NCBI bioproject	SI036_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.36E+08	Illumnia	Steven Hallam
SI4096419	SAMN05224409	NCBI bioproject	SI036_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	4.33E+07	Illumnia	Steven Hallam
SI4096420	SAMN05224412	NCBI bioproject	SI036_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	6.19E+07	Illumnia	Steven Hallam
SI4096421	SAMN05224416	NCBI bioproject	SI039_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	6.03E+07	Illumnia	Steven Hallam
SI4096422	SAMN05224417	NCBI bioproject	SI039_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.95E+08	Illumnia	Steven Hallam
SI4096423	SAMN05224422	NCBI bioproject	SI039_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	2.34E+08	Illumnia	Steven Hallam
SI4096424	SAMN05224423	NCBI bioproject	SI039_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.62E+08	Illumnia	Steven Hallam
SI4096425	SAMN05224428	NCBI bioproject	SI039_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	7.10E+07	Illumnia	Steven Hallam
SI4096426	SAMN05224477	NCBI bioproject	SI039_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	2.05E+08	Illumnia	Steven Hallam
SI4096428	SAMN05224454	NCBI bioproject	SI047_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.27E+08	Illumnia	Steven Hallam
SI4096429	SAMN05224455	NCBI bioproject	SI047_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	1.77E+08	Illumnia	Steven Hallam

C.1 Metagenome inventory for global fragment recruitment continued from previous page

File Name	Accession	Data Repository	Location/sampleID	Specific Ecosystem	Metagenome Group	Latitude	Longitude	Depth (m)	Size (bp)	Sequencing Platform	Reference / Contact
SI4096430	SAMN05224458	NCBI bioproject	SI047_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.81E+08	Illumnia	Steven Hallam
SI4096431	SAMN05224459	NCBI bioproject	SI047_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	1.83E+08	Illumnia	Steven Hallam
SI4096432	SAMN05224463	NCBI bioproject	SI047_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	1.00E+08	Illumnia	Steven Hallam
SI4096433	SAMN05224462	NCBI bioproject	SI048_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	5.85E+07	Illumnia	Steven Hallam
SI4096434	SAMN05224393	NCBI bioproject	SI048_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.75E+08	Illumnia	Steven Hallam
SI4096435	SAMN05224394	NCBI bioproject	SI048_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	1.12E+08	Illumnia	Steven Hallam
SI4096436	SAMN05224397	NCBI bioproject	SI048_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.36E+08	Illumnia	Steven Hallam
SI4096437	SAMN05224398	NCBI bioproject	SI048_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	6.47E+07	Illumnia	Steven Hallam
SI4096438	SAMN05224401	NCBI bioproject	SI048_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	4.66E+07	Illumnia	Steven Hallam
SI4096439	SAMN05224489	NCBI bioproject	SI053_SI3_10m	Saanich Inlet	SI_10m	48.588	-123.504	10m	1.38E+07	Illumnia	Steven Hallam
SI4096440	SAMN05224490	NCBI bioproject	SI053_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.35E+08	Illumnia	Steven Hallam
SI4096441	SAMN05224491	NCBI bioproject	SI053_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	1.67E+08	Illumnia	Steven Hallam
SI4096442	SAMN05224492	NCBI bioproject	SI053_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	1.07E+08	Illumnia	Steven Hallam
SI4096443	SAMN05224466	NCBI bioproject	SI053_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	2.07E+08	Illumnia	Steven Hallam
SI4096444	SAMN05224467	NCBI bioproject	SI053_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	1.65E+08	Illumnia	Steven Hallam
SI4096446	SAMN05224410	NCBI bioproject	SI054_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.48E+08	Illumnia	Steven Hallam
SI4096447	SAMN05224433	NCBI bioproject	SI054_SI3_120m	Saanich Inlet	SI_120m	48.588	-123.504	120m	8.58E+07	Illumnia	Steven Hallam
SI4096448	SAMN05224473	NCBI bioproject	SI054_SI3_135m	Saanich Inlet	SI_135m	48.588	-123.504	135m	7.67E+07	Illumnia	Steven Hallam
SI4096449	SAMN05224478	NCBI bioproject	SI054_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	4.10E+07	Illumnia	Steven Hallam
SI4096450	SAMN05224438	NCBI bioproject	SI054_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	9.80E+07	Illumnia	Steven Hallam
SI4096451	SAMN05224439	NCBI bioproject	SI060_SI3_100m	Saanich Inlet	SI_100m	48.588	-123.504	100m	1.73E+08	Illumnia	Steven Hallam
SI4096452	SAMN05224444	NCBI bioproject	SI060_SI3_150m	Saanich Inlet	SI_150m	48.588	-123.504	150m	9.75E+07	Illumnia	Steven Hallam
SI4096453	SAMN05224474	NCBI bioproject	SI060_SI3_200m	Saanich Inlet	SI_200m	48.588	-123.504	200m	1.90E+08	Illumnia	Steven Hallam
SRR064444	SRS113624	NCBI SRA	MOOMZ1 50m DNA	Coastal waters off Iquique	ETSP_50-65m	-20.07	-70.23	50m	1.08E+08	454 GS FLX	Stewart et al.2012
SRR064446	SRS113625	NCBI SRA	MOOMZ1 85m DNA	Coastal waters off Iquique	ETSP_70-85m	-20.07	-70.23	85m	1.64E+08	454 GS FLX	Stewart et al.2012
SRR064448	SRS113626	NCBI SRA	MOOMZ1 110m DNA	Coastal waters off Iquique	ETSP_110-200m	-20.07	-70.23	110m	1.11E+08	454 GS FLX	Stewart et al.2012
SRR064450	SRS113627	NCBI SRA	MOOMZ1 200m DNA	Coastal waters off Iquique	ETSP_110-200m	-20.07	-70.23	200m	1.42E+08	454 GS FLX	Stewart et al.2012
SRR070081	SRS118145	NCBI SRA	MOOMZ2_70m_DNA	Coastal waters off Iquique	ETSP_70-85m	-20.07	-70.23	70m	6.30E+08	454 GS FLX Titanium	Stewart et al.2012
SRR070082	SRS118146	NCBI SRA	MOOMZ2 200m DNA	Coastal waters off Iquique	ETSP_110-200m	-20.07	-70.23	200m	5.84E+08	454 GS FLX Titanium	Stewart et al.2012
SRR070083	SRS118147	NCBI SRA	MOOMZ3 80m DNA	Coastal waters off Iquique	ETSP_70-85m	-20.07	-70.23	80m	7.45E+08	454 GS FLX Titanium	Stewart et al.2012
SRR070084	SRS118148	NCBI SRA	MOOMZ3 150m DNA	Coastal waters off Iquique	ETSP_110-200m	-20.07	-70.23	150m	7.17E+08	454 GS FLX Titanium	Stewart et al.2012
SRR304656	SRS213611	NCBI SRA	Moomz1 65m DNA	Coastal waters off Iquique	ETSP_50-65m	-20.07	-70.23	65m	1.16E+08	454 GS FLX	Stewart et al.2012
SRR304668	SRS213613	NCBI SRA	Moomz1 500m DNA	Coastal waters off Iquique	ETSP_500-800m	-20.07	-70.23	500m	1.46E+08	454 GS FLX	Stewart et al.2012
SRR304671	SRS213616	NCBI SRA	Moomz2 35m DNA	Coastal waters off Iquique	ETSP_35m	-20.07	-70.23	35m	5.59E+08	454 GS FLX Titanium	Stewart et al.2012
SRR304672	SRS213617	NCBI SRA	Moomz2 50m DNA	Coastal waters off Iquique	ETSP_50-65m	-20.07	-70.23	50m	6.11E+08	454 GS FLX Titanium	Stewart et al.2012
SRR304673	SRS213618	NCBI SRA	Moomz2 110m DNA	Coastal waters off Iquique	ETSP_110-200m	-20.07	-70.23	110m	5.03E+08	454 GS FLX Titanium	Stewart et al.2012
SRR304674	SRS213619	NCBI SRA	Moomz3 50m DNA	Coastal waters off Iquique	ETSP_50-65m	-20.07	-70.23	50m	8.79E+08	454 GS FLX Titanium	Stewart et al.2012
SRR304680	SRS213623	NCBI SRA	Moomz3 110m DNA	Coastal waters off Iquique	ETSP_110-200m	-20.07	-70.23	110m	8.07E+08	454 GS FLX Titanium	Stewart et al.2012
SRR304683	SRS213614	NCBI SRA	Moomz1 800m DNA	Coastal waters off Iquique	ETSP_500-800m	-20.07	-70.23	800m	5.50E+07	454 GS FLX	Stewart et al.2012
SRR304684	SRS213624	NCBI SRA	Moomz1 15m DNA	Coastal waters off Iquique	ETSP_15m	-20.07	-70.23	15m	1.56E+08	454 GS FLX	Stewart et al. 2012

Table C.2: Summary of recruited fragments to metagenome groups in global fragment recruitment analysis

Metagenome Study	Abreviation	Metagenome group	# metagenomes in group	Marinimicrobia lineage									
				ZA3312c-B	ZA3312c-A	Arctic96B-7-A	Arctic96B-7-B	HF770D10	ZA3648c	SHAN400	SHBH1141	HMTAb91-B	Group total
Saanich Inlet	SI	SI_10m	10	0	3287	3328	3917	0	0	978	9948	0	21458
		SI_100m	14	2	4622	18597	21844	4	0	5931	7997	0	58997
		SI_120m	12	0	2811	12709	10918	3	0	4277	14561	0	45279
		SI_135m	12	0	1954	10153	11818	1	0	4903	21423	0	50252
		SI_150m	14	0	5432	18304	7663	6	0	9097	32848	0	73350
		SI_200m	13	1	1554	14080	996	2	0	6361	29807	0	52801
North Eastern Sub Arctic Pacific	NESAP	NESAP_10m	4	1	1272	0	1	0	0	0	0	0	1274
		NESAP_1000m	12	50	1186	3624	459	4335	0	1868	55	0	11577
		NESAP_500m	16	16	869	2701	130	3374	0	1181	59	0	8330
		NESAP_1300m	2	12	7	102	10	977	0	2	0	0	1110
		NESAP_2000m	8	35	768	674	1111	3215	0	387	38	0	6228
Eastern Tropical South Pacific	ETSP	ETSP_15m	1	27	367	2	26	1	0	0	0	0	423
		ETSP_35m	1	7	89	1	174	0	0	0	0	0	271
		ETSP_50-65m	4	13	326	350	1519	4	0	0	0	0	2212
		ETSP_70-85m	3	9	81	1067	745	86	8	6	0	0	2002
		ETSP_110-200m	6	17	24	2165	244	266	9	8	4	0	2737
		ETSP_500-800m	2	2	1	77	18	192	38	1	0	0	329
Peru	PERU	PERU_5m	1	2	11	3	2	0	0	0	0	0	18
		PERU_20m	1	2	10	15	7	0	0	0	5	0	39
		PERU_40-80m	4	0	2	351	46	6	1	1	79	0	486
TARA Oceans	m	SRF_Indian Ocean	12	2042	1	0	5	0	0	0	0	0	2048
	j	SRF_Mediterranean Sea	6	179	44	0	1	0	0	0	0	0	224
	h	SRF_North Atlantic Ocean	10	626	206	5	313	0	0	0	0	0	1150
	e	SRF_North Pacific Ocean	6	468	54	0	4	0	0	0	0	0	526
	k	SRF_Red Sea	4	220	0	0	3	1	0	0	0	0	224
	i	SRF_South Atlantic Ocean	21	1685	431	2	122	0	0	0	0	0	2240
	n	SRF_South Pacific Ocean	22	114	27	0	19	0	0	0	0	0	160
	l	SRF_Southern Ocean	2	0	0	0	2	0	0	0	0	0	2
	m	DCM_Indian Ocean	12	606	2	0	17	0	1	0	0	0	626
	j	DCM_Mediterranean Sea	7	123	126	0	22	0	1	0	0	0	272
	h	DCM_North Atlantic Ocean	4	236	40	0	26	1	0	0	0	0	303
	e	DCM_North Pacific Ocean	5	26	146	13	31	3	0	0	0	0	219
	k	DCM_Red Sea	2	132	0	0	3	0	0	0	0	0	135
	i	DCM_South Atlantic Ocean	14	689	259	1	136	1	2	0	0	0	1088
	n	DCM_South Pacific Ocean	12	57	11	1	20	0	2	0	0	0	91
	h	MIX_North Atlantic Ocean	1	7	35	0	25	0	0	0	0	0	67
	e	MIX_South Pacific Ocean	10	8	4	2	10	1	0	0	0	0	25
	m	MES_Indian Ocean	8	9	5	46	7	298	4	0	0	0	369
	h	MES_North Atlantic Ocean	6	3	7	19	9	257	3	0	0	0	298
	e	MES_North Pacific Ocean	5	0	3	49	11	341	4	0	0	0	408
	i	MES_South Atlantic Ocean	10	243	11	79	191	1418	8	0	0	0	1950
	n	MES_South Pacific Ocean	9	13	11	42	13	435	1	0	0	0	515
	l	MES_Southern Ocean	1	0	1	9	0	56	0	0	0	0	66
Hydro-thermal	d	Guaymas Basin	2	11	5	59	16	189	8	2	5	0	295
	g	Gulf of Mexico	5	2	0	16	4	286	1	0	0	0	309
Strat. Lake	b	Juan de Fuca Ridge flank	1	0	9	1	5	7	0	0	0	0	22
TR	a	Sakinaw Lake	1	0	0	0	0	0	0	0	0	1	1
	f	Iowa	1	0	0	0	1	0	0	0	0	0	1
	c	Angelo Coastal Reserve	1	0	0	0	1	0	0	0	0	0	1
			Total:	7695	26111	88647	62665	15766	91	35003	116829	1	

*Abbreviations: SRF - surface; DCM - deep chlorophil max; MIX - mix layer; MES - mesopelagic; Strat. Lake - stratified lake; TR - trrestereal.

Table C.3: Genomic features of Mrinimicrobia population genome bins.

Clade	Singel Cell Genome Identity	Population Genome Size (Mbp)	Estimated Completeness (%)	Number of Contigs	N50	GC Content (%)	Single Copy Marker Genes	Strain Heterogeneity	Marker Lineage
ZA3312c-A	AAA160-I06, AAA160-C11, AAA076-M08, AAA160-B08	11	95.8	531	35213	32.8%	56	94.33	root
ZA3312c-B	AAA0298-D23	1	93.4	41	236078	31.6%	147	28	Bacteria
HF770D10	AAA003-E22	1.4	41.2	118	15724	36.6%	104	100	Bacteria
Arctic96B7_A	AB-746_N13AB-902, AB-747_F21AB-903	50.9	100.0	3423	18609	39.4%	56	70.67	root
Arctic96B7_B	AB-746_P06AB-902, JGI 0000113-D11	6.0	96.6	583	13227	32.6%	104	63.36	Bacteria
SHAN400	AB-755_M21D07	32.2	87.5	2196	19252	37.4%	56	99.64	root
SHBH1141	AB-750_L13AB-904, AB-755_E16C12, AB-751_D09AB-904	65.6	91.7	3127	35279	43.5%	56	96.11	root

Table C.4: Summary of central metabolism in Marinimicrobia lineages by SAGs and population genomes.Mirinimicrobia lineage is listed in top row, use of SAGs inidates multiple SAGs in a given lineages as per figure 1A. Abbrevation 'pop. Genome indicates the population genome including recruited etagenomic contigs, if no pop. Genome is listed no metagenomic contigs were recruited.

Metabolism	Pathway or Ggene	ZA3312c-A		ZA3312c-B		HF770D10		ZA3648c	Arctic96-B-7-A		Arctic96-B-7-B		SHAN400		SHBH1141		HMTAb91-A	HMTAb91-B
		SAGs	pop. Genome	SAG	pop. Genome	SAG	pop. Genome	SAG	SAGs	pop. Genome	SAGs	pop. Genome	SAG	pop. Genome	SAGs	pop. Genome	SAGs	SAG
Sugar Metabolism	Entner Doudoroff	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Non-Oxidative Pentose PhosphatePpathway	Y	Y	-	-	-	-	-	-	-	-	-	-	-	Y	Y	-	-
	Pentose Phosphate Pathway	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Embden-Meyerhof-Parnar	-	-	-	-	-	-	-	-	Y	-	-	-	Y	Y	Y	Y	-
	Pyruvate Kinase	-	Y	-	-	-	-	-	-	-	-	Y	Y	Y	-	-	-	-
Gluconeogenesis	Phosphoenolpyruvate Carboxylase/Carboxykinase	Y	Y	Y	Y	-	-	Y	Y		-	Y	Y	Y	Y	Y	Y	Y
	Fructose-1,6-bisphosphatase	Y	Y	Y	-	-	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
TCA	Pyr	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	RHox	Y	Y	Y	-	-	-	-	-	Y	-	Y	-	Y	Y	Y	-	-
	LHox	Y	Y	Y	-	-	-	-	-	Y	Y	Y	-	Y	Y	Y	-	-
	Sdh	Y	Y	Y	-	-	-	Y	-	Y	Y	-	Y	Y	Y	Y	Y	-
Carbon Fixation	Reductive-TCA (citrate-lyase)	-	-	-	-	-	-	-	-	-	-	-	-	-	Y	Y	-	-
	WoodLjungdahl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Hydroxypropionate Bicycle	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Hydroxy-ropionate/4-Hydroxybutyrate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Calvin Benson Bassham	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Motility	Flagellum	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Pillin	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cell Wall	Lipopolysaccharide A	Y	Y	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	Peptidoglycan	-	-	-		-	-	-	-	-	-	-	-	-	-	-	-	-

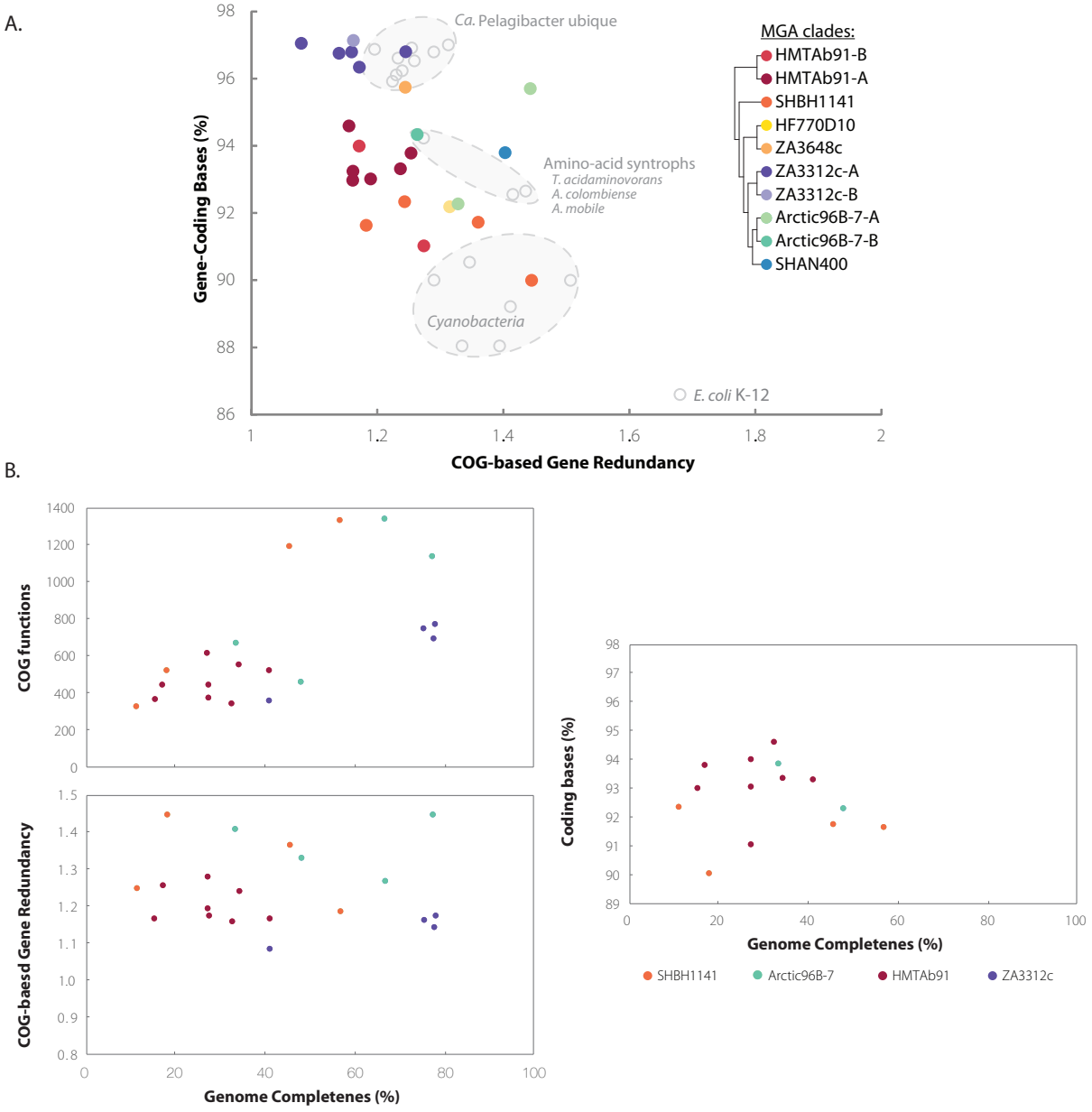


Figure C.1: Genomic streamlining in Marinimicrobia clades. **(A)** Comparison of genome reduction between Marinimicrobia clades and selected reference organisms based on estimated gene redundancy using clusters of orthologous groups (COG) annotation and frequency of gene-coding bases. (B) Benchmarking of genome reduction showing COG functions recovered, COG-based gene redundancy and percentage of coding bases as a function of estimated genome completeness.

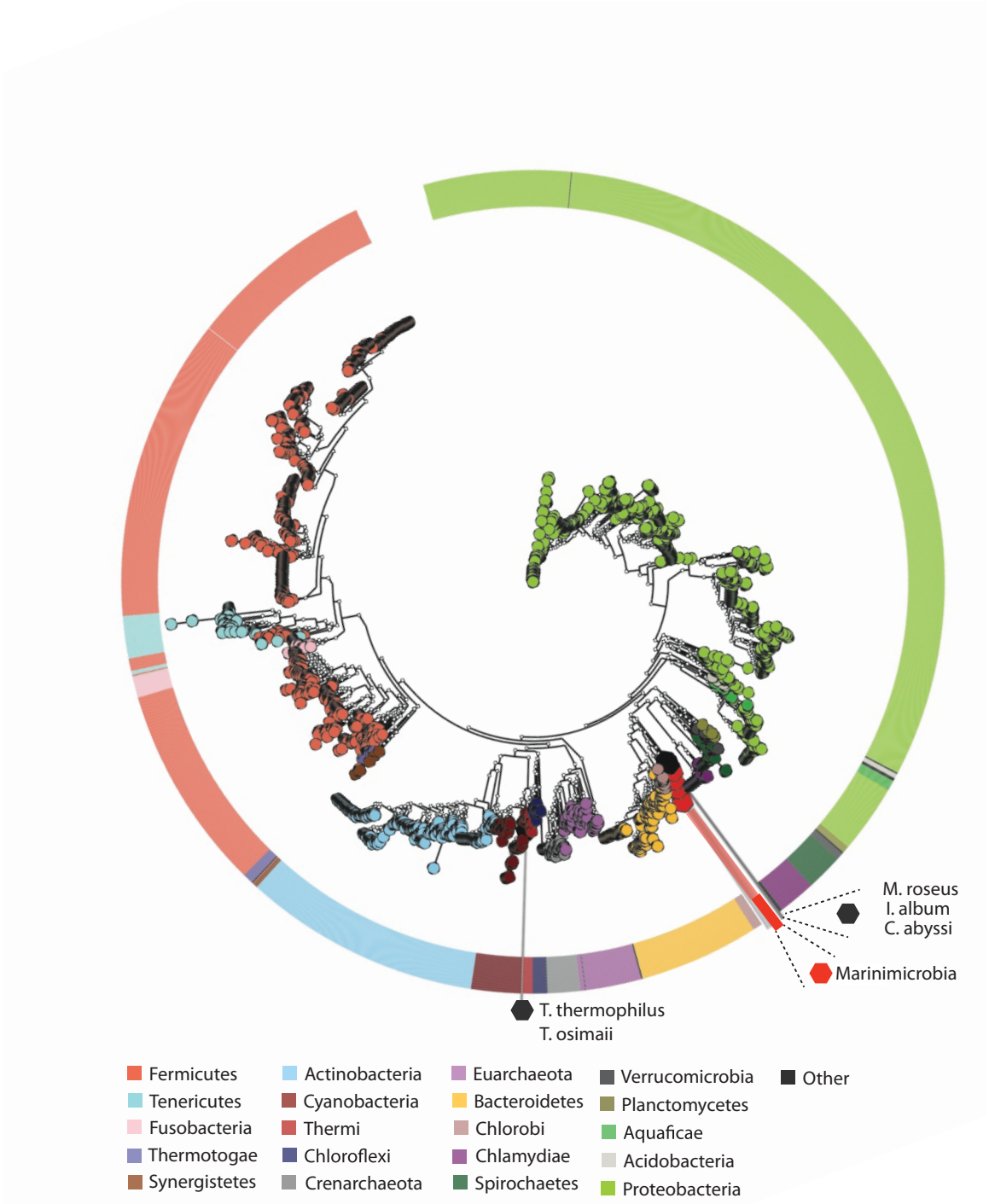


Figure C.2: Phylogenomic tree placing Marinimicrobia in bacterial phylum.Phylogenetic relationship of Marinimicrobia SAGs and related genomes within the microbial tree of life as determined by sequence alignment of 400 conserved protein sequences. The tree was generated using the PhyloPhlAn pipeline, placing Marinimicrobia SAG sequences within a phylogeny of 3,737 curated microbial genomes (colored by phylum). 25 MGA SAGs are shown as red hexagons, and 5 genomes previously identified as being highly similar to Marinimicrobia based on small subunit rRNA gene sequence shown as black hexagons.

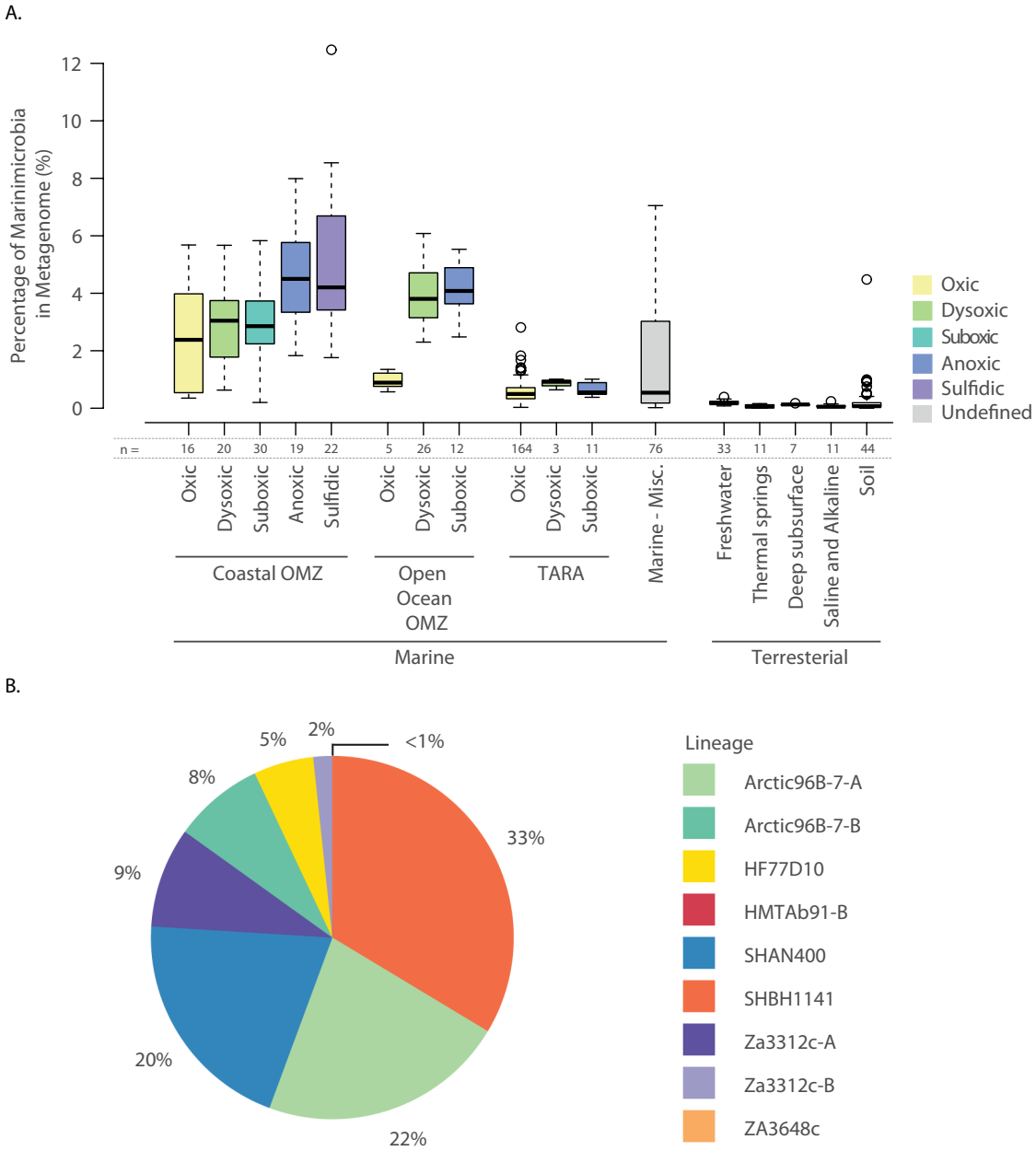


Figure C.3: Global prevalence of Marinimicrobia in surveyed metagenomes(A) Box and whisker plot showing the distribution of percentage of Marinimicrobia in surveyed metagenomes by region and redox status including Coastal OMZs (Saanich Inlet, Eastern Tropical South Pacific and the Peruvian upwelling system), Open Ocean OMZ (Northeastern subarctic Pacific), TARA oceans survey, miscellaneous marine and terrestrial environments (see table A3.2) with the number of samples with Marinimicrobia present indicated as n =. Where environmental data was available the metagenenomic sample was categorized as oxic (>90 µM O_2; yellow), dysoxic (20-90 µM O_2; teal), suboxic (1-20 µM O_2; blue), anoxic (<1), sulfidic (purple). **(B)** Global abundance of Marinimicrobia clades.

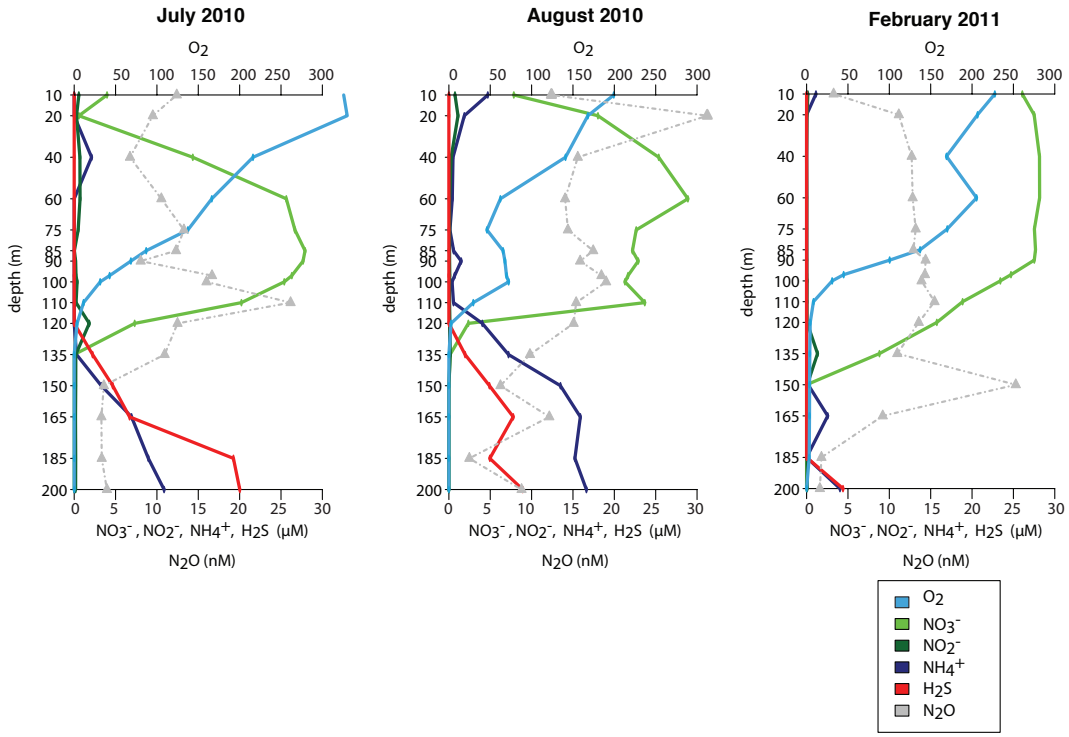


Figure C.4: Saanich Inlet water column chemistry for metatranscriptome samples.Plots of Saanich Inlet water column chemistry for time points used for metatranscriptomic expression analysis.

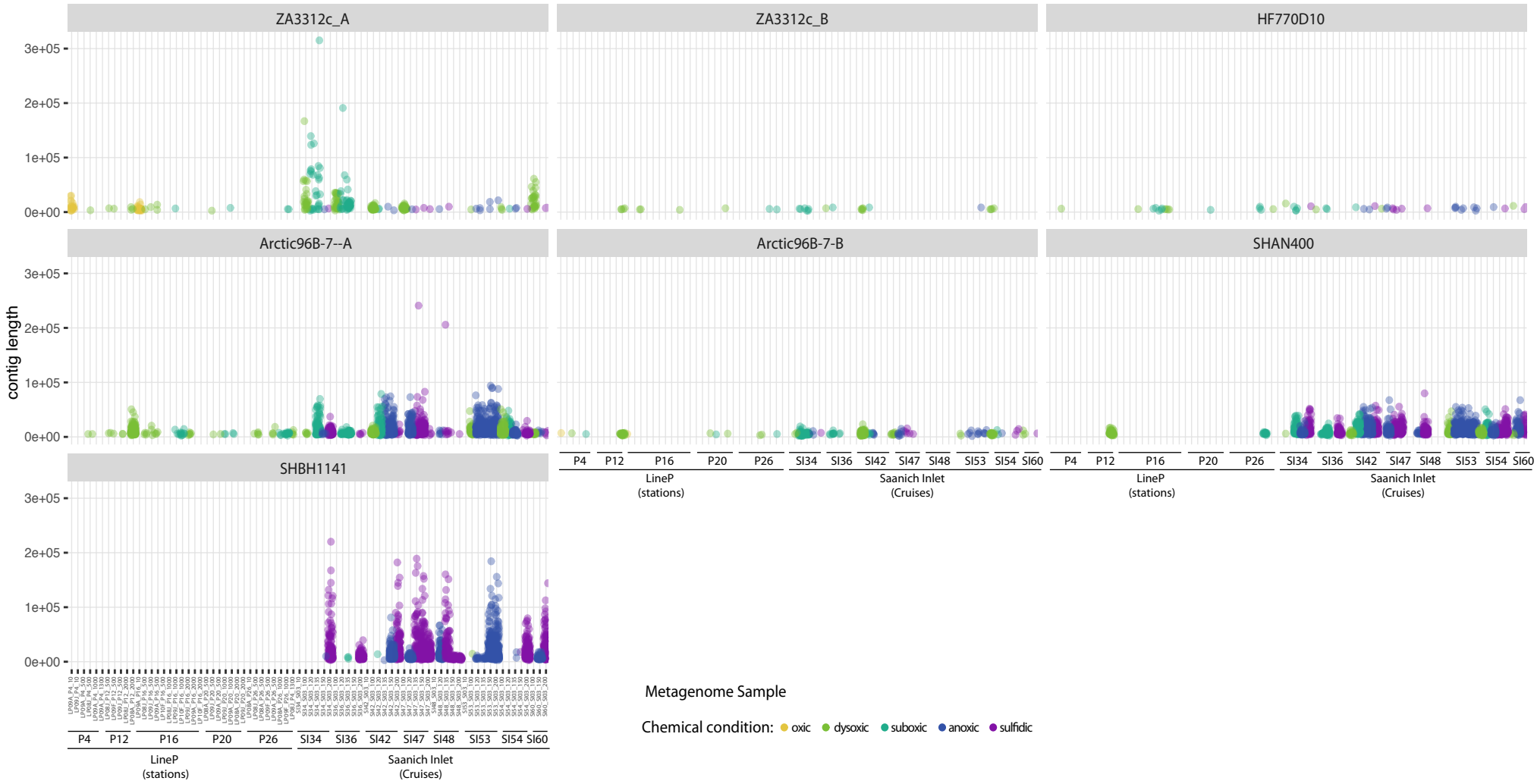


Figure C.5: Origin, length and abundance of contigs in population genomes. Distribution of metagenomic contigs from North eastern subarctic Pacific (NESAP) Ocean and Saanich Inlet metagenomic samples making up indicated SAG population genomes. Each dot represents a contig, sample origin is shown on the X-axis as indicated for NESAP and Saanich Inlet and contig length of recruited contigs is shown on the Y-axis. Colours represent the chemical condition of the water column at time of sampling for metagenoms: oxic (>90 µM O_2; yellow), dysoxic (90 µM < O_2 > 20 µM; green), suboxic (20 µM< O_2 > 2 µM; teal), anoxic (< 2 µM O_2; blue) and sulfidic (purple).

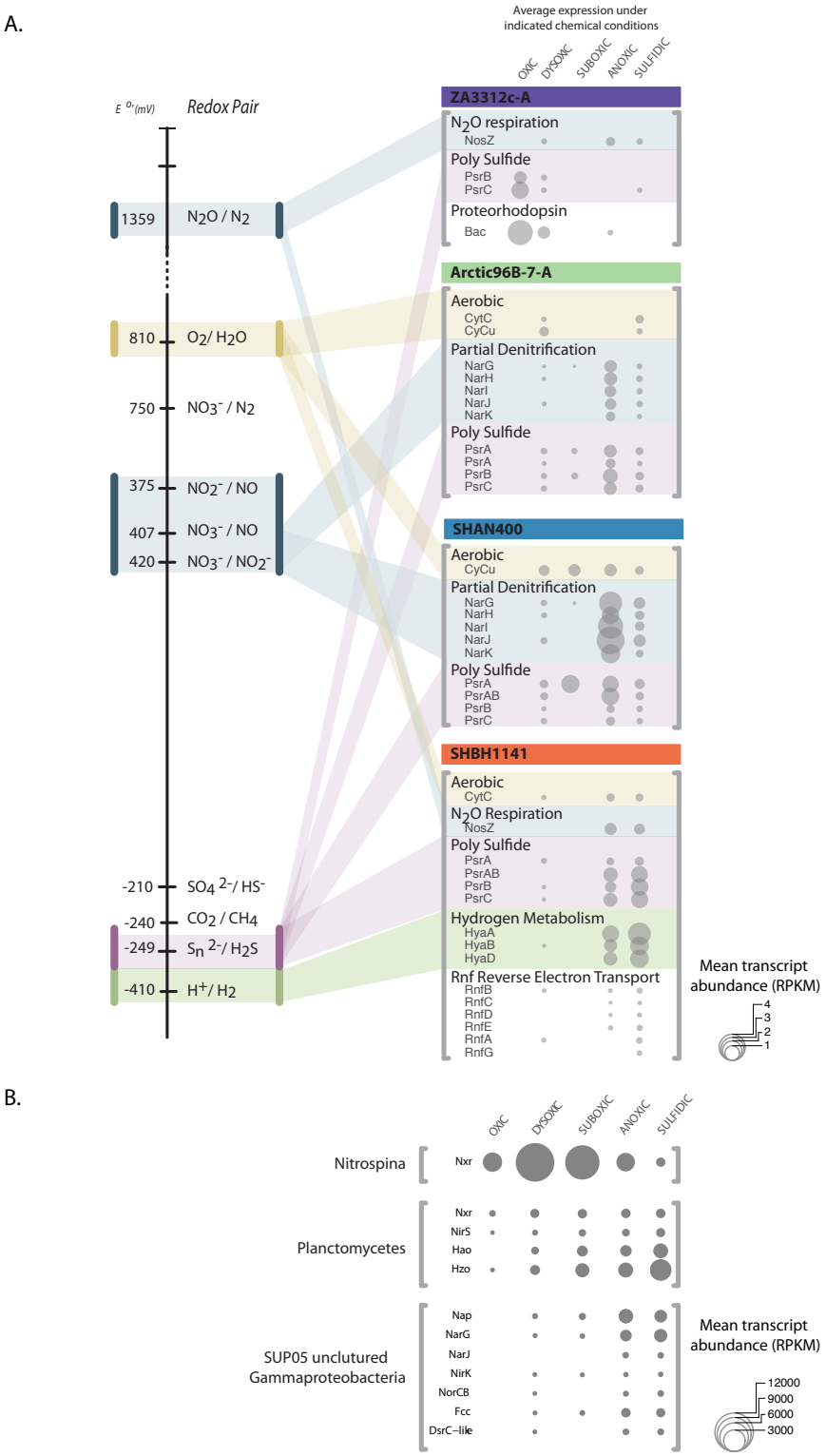


Figure C.6: Expression of energy metabolism enzyme subunits from Marinimicrobia and co-metabolic partners. **(A)** Expression of energy metabolism gene subunits average over water column redox regimes (oxic, dysoxic, suboxic, anoxic, sulfidic), mapped to redox pairs on the electron tower. **(B)** Expression selected energy metabolism genes for proposed co-metabolic partners in Saanich Inlet.

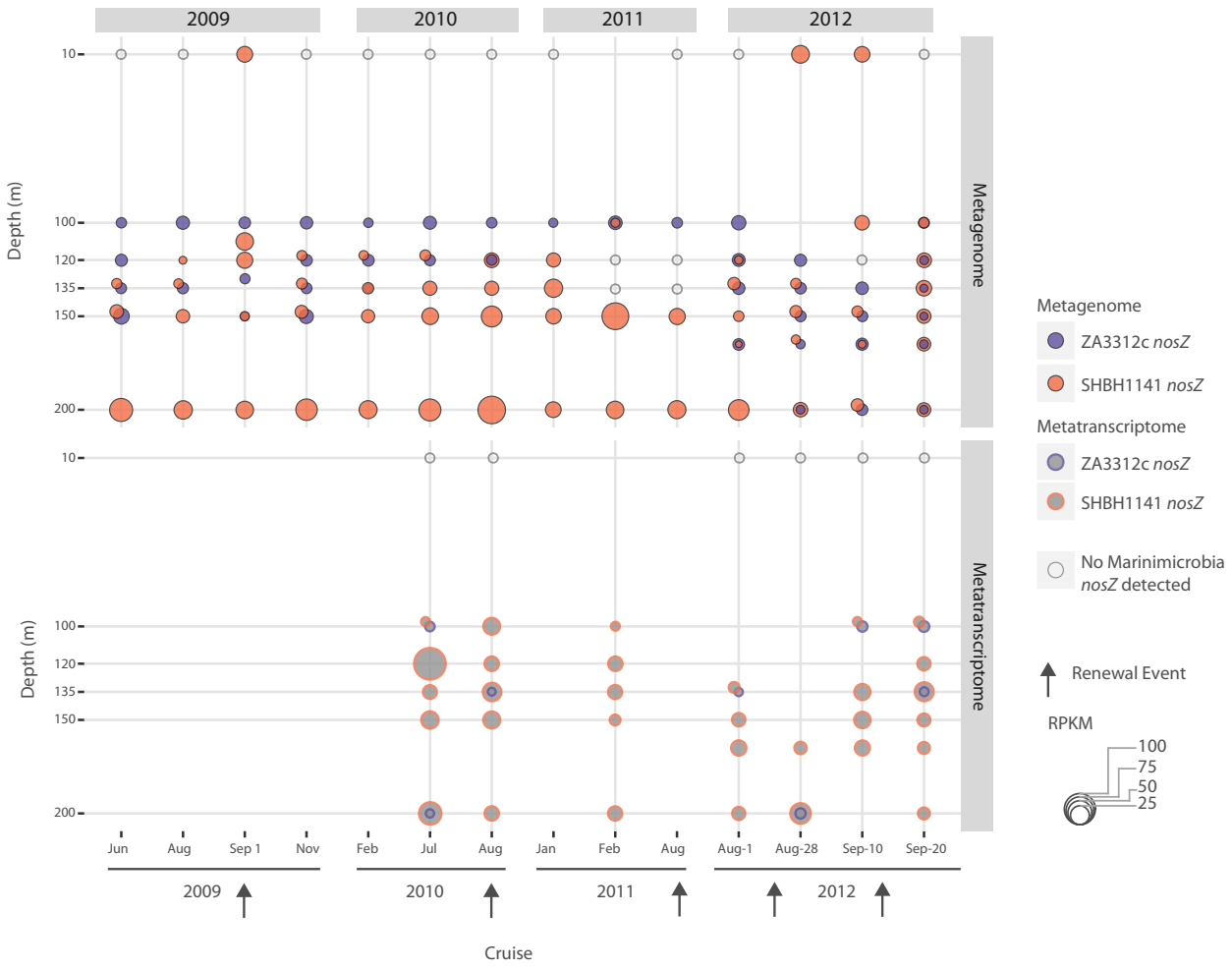
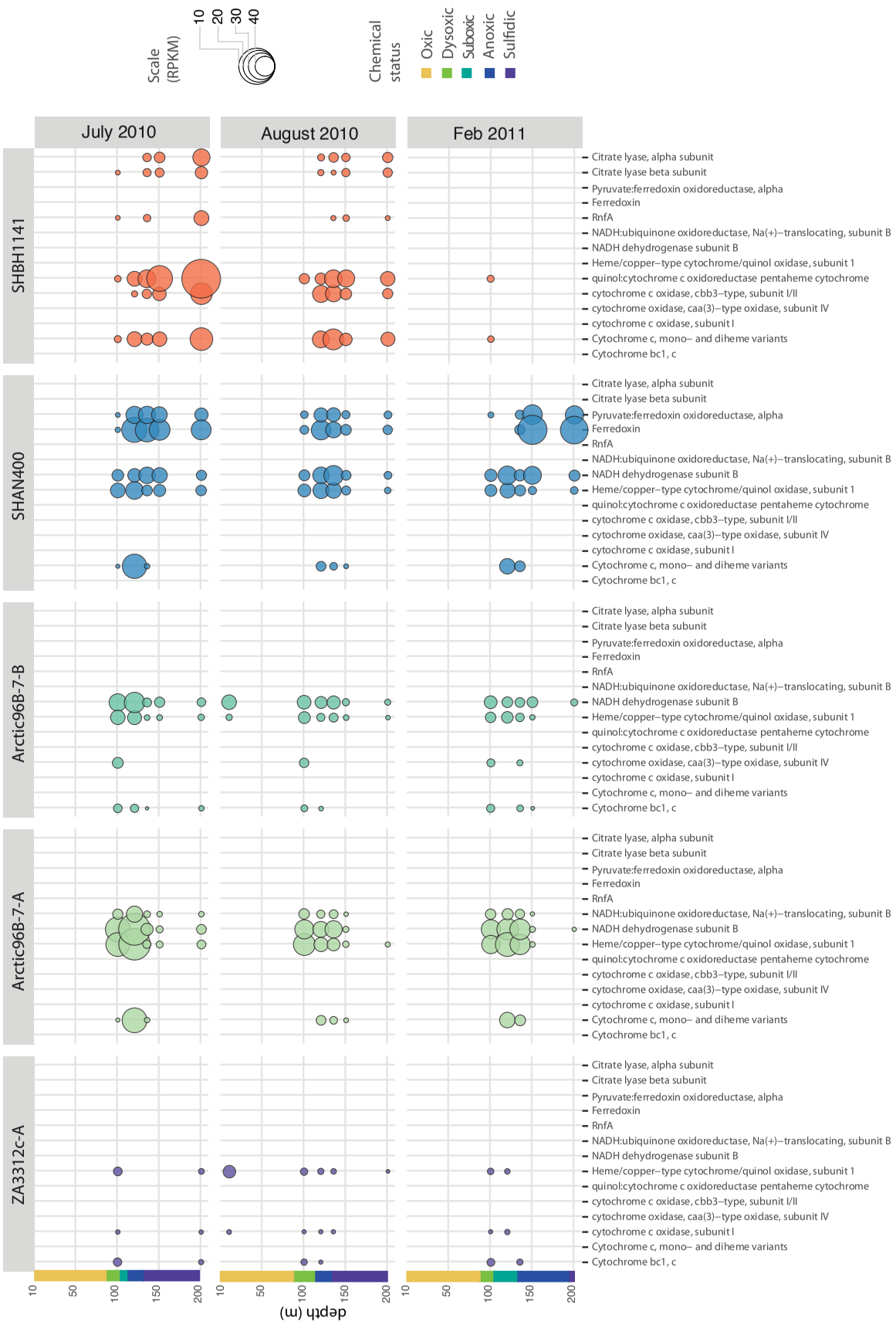


Figure C.7: Marinimicrobia *nosZ* genes and expression in Saanich Inlet Time Series Marinimicrobia nosZ abundance in Saanich Inlet time series metagenomes and metatranscriptomes. Dot size represents summed RPKM for each *nosZ* type in a given metagenome or metatranscriptome.


Scale (RPKM)
10
20
30
40
Chemical status
Oxic
Dysoxic
Suboxic
Anoxic
Sulfidic
July 2010
August 2010
Feb 2011
SHBH1141
SHAN400
Arctic96B-7-B
Arctic96B-7-A
ZA3312c-A
Citrate lyase, alpha subunit
Citrate lyase beta subunit
Pyruvate:ferredoxin oxidoreductase, alpha
Ferredoxin
RnfA
NADH:ubiquinone oxidoreductase, Na(+)−translocating, subunit B
NADH dehydrogenase subunit B
Heme/copper−type cytochrome/quinol oxidase, subunit 1
quinol:cytochrome c oxidoreductase pentaheme cytochrome
cytochrome c oxidase, cbb3−type, subunit I/II
cytochrome oxidase, caa(3)−type oxidase, subunit IV
cytochrome c oxidase, subunit I
Cytochrome c, mono− and diheme variants
Cytochrome bc1, c
10
50
100
150
200
depth (m)

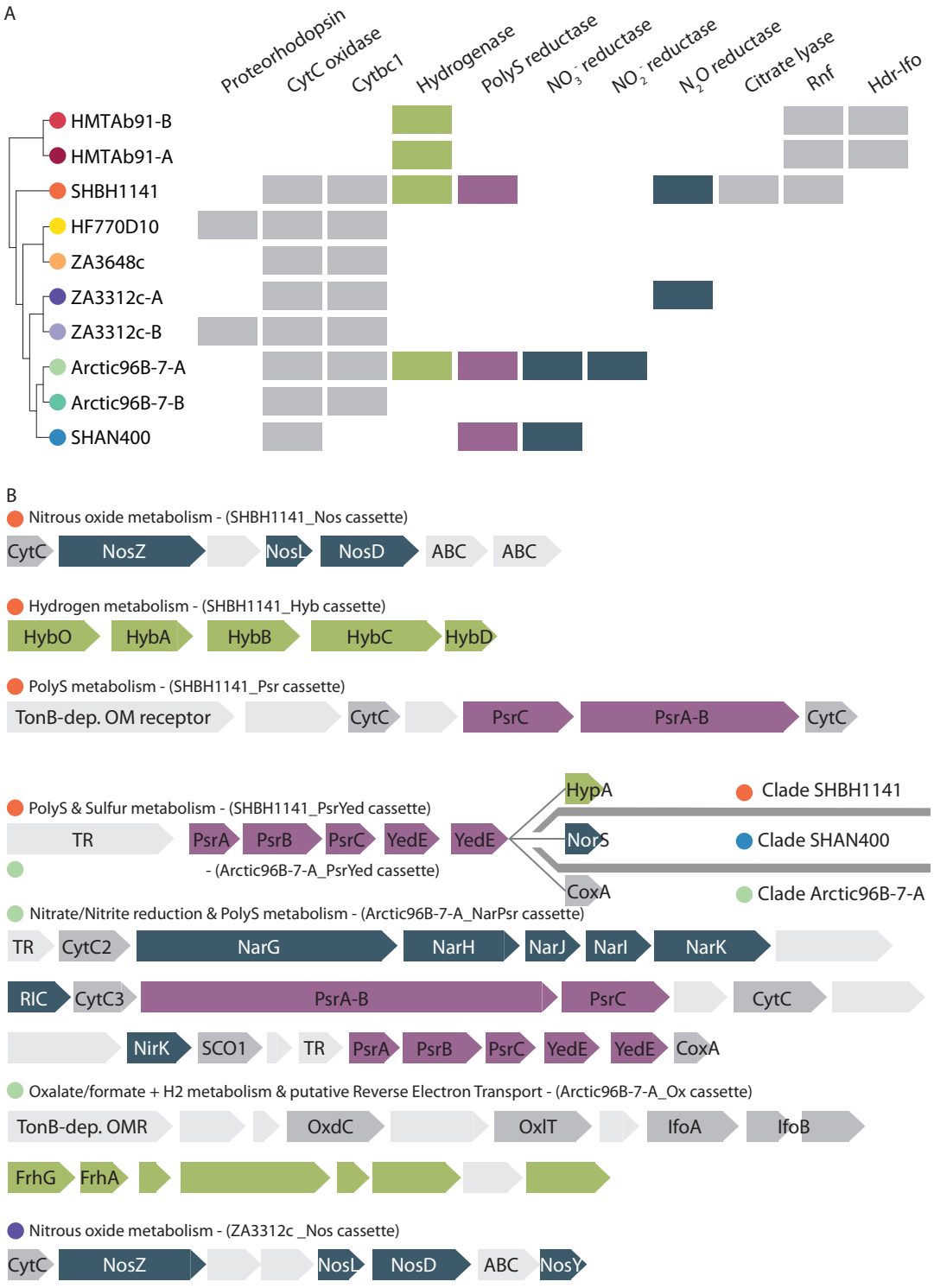


Figure C.9: Energy metabolism summary and operons across Marinimicrobia clades. **(A)** summary of energy metabolism, carbon fixation and co-metabolic interdependency (Rnf and Hdr-Ifo) for Marinimicrobia lineages. **(B)** Operons in Marinimicrobia SAGs showing different gene arrangement in different lineages.

Appendix D

Chapter 5: Supplementary material

Table D.1: Summary of CheckM statistics for SAGs with taxonomies containing *nosZ*

Taxonomy	Completeness	Contamination	Strain Heterogeneity
Bacteroidales	59.64	3.59	13.61
Marinimicrobia	69.79	1.36	2.78
SAR324	63.92	2.73	18.29
Arcobacteraceae	47.70	3.33	25.12
SUP05_1c	38.49	1.64	18.75
SUP05_1a	40.72	0.35	18.40
Ectothiorhodospirales	61.41	6.65	10.08

Table D.2: Metagenome and metatranscriptome RPKM for clades and nodes by chemistry

Chemistry	Clade	Metagenome RPKM	Metatranscriptome RPKM
Oxic	2	1.789	0.000
Oxic	3	0.864	0.000
Oxic	4	1.247	0.000
Oxic	5	62.503	0.000
Oxic	6	44.001	0.000
Oxic	7	3.554	0.000
Oxic	8	4.880	0.000
Oxic	9	101.691	0.000
Oxic	10	2.962	0.000
Oxic	11	1.115	0.000
Oxic	12	4.457	0.000
Oxic	13	6.629	0.000
Oxic	2-10	2.279	0.000
Oxic	4-6	2.035	0.000
Dysoxic	2	6.760	4.200
Dysoxic	3	1.000	0.000
Dysoxic	4	0.781	0.000
Dysoxic	5	22.801	15.557
Dysoxic	6	13.464	7.322
Dysoxic	7	1.671	3.022
Dysoxic	8	7.899	1.430
Dysoxic	9	22.055	5.256
Dysoxic	10	3.642	2.067
Dysoxic	11	1.146	0.000
Dysoxic	12	2.776	0.000
Dysoxic	13	16.068	14.826
Dysoxic	2-10	1.736	3.689
Dysoxic	2-13	1.000	0.000
Dysoxic	4-6	3.105	0.000
Dysoxic	11-13	1.000	0.000
Suboxic	2	4.918	5.093
Suboxic	3	1.564	4.814
Suboxic	4	0.000	2.312
Suboxic	5	20.910	22.387
Suboxic	6	53.310	98.863
Suboxic	7	1.720	3.040
Suboxic	8	11.306	11.707
Suboxic	9	21.076	20.517
Suboxic	10	3.211	5.519
Suboxic	11	1.248	7.200
Suboxic	12	2.097	8.000
Suboxic	13	21.053	86.593
Suboxic	2-10	1.127	6.305

Metagenome and metatranscriptome RPKM for clades and nodes by chemistry (continued from previous page)

Chemistry	Clade	Metagenome RPKM	Metatranscriptome RPKM
Suboxic	4-6	6.732	0.000
Suboxic	9-10	1.000	1.521
Suboxic	11-13	0.434	0.661
Anoxic	2	3.757	3.304
Anoxic	4	1.000	0.000
Anoxic	5	10.916	29.115
Anoxic	6	63.105	30.994
Anoxic	7	3.030	0.000
Anoxic	8	12.176	7.684
Anoxic	9	35.197	8.986
Anoxic	10	1.691	2.834
Anoxic	11	3.010	0.000
Anoxic	12	3.000	2.058
Anoxic	13	14.373	16.798
Anoxic	2-10	1.000	2.036
Anoxic	4-6	2.000	0.000
Anoxic	9-10	1.000	2.379
Anoxic	11-13	1.000	0.000
Sulfidic	2	4.234	7.283
Sulfidic	3	1.066	0.970
Sulfidic	4	1.000	1.657
Sulfidic	5	55.338	193.119
Sulfidic	6	81.584	247.016
Sulfidic	7	7.583	2.996
Sulfidic	8	9.301	6.642
Sulfidic	9	38.678	22.478
Sulfidic	10	5.695	2.905
Sulfidic	11	1.208	2.147
Sulfidic	12	2.000	0.911
Sulfidic	13	16.960	72.401
Sulfidic	2-10	6.394	2.854
Sulfidic	4-6	1.000	0.000
Sulfidic	7-10	0.000	0.654
Sulfidic	9-10	1.732	3.311
Sulfidic	11-13	0.680	1.092
Unknown	1	9.200	0.000
Unknown	2	3.484	0.000
Unknown	3	4.291	0.000
Unknown	4	8.824	0.000
Unknown	5	10.813	0.000
Unknown	6	7.905	0.000
Unknown	7	4.534	0.000
Unknown	8	5.062	0.000

Metagenome and metatranscriptome RPKM for clades and nodes by chemistry (continued from previous page)

Chemistry	Clade	Metagenome RPKM	Metatranscriptome RPKM
Unknown	9	8.464	0.000
Unknown	10	9.056	0.000
Unknown	11	7.692	0.000
Unknown	12	2.531	0.000
Unknown	13	5.221	0.000
Unknown	2-10	2.364	0.000
Unknown	2-13	2.000	0.000
Unknown	4-6	4.543	0.000
Unknown	4-10	1.600	0.000
Unknown	5-6	1.500	0.000
Unknown	7-10	1.000	0.000
Unknown	9-10	2.444	0.000
Unknown	11-13	1.500	0.000
Unknown	12-13	1.750	0.000

Table D.3: Total clade abundance and expression

Clade Number	Metagenome RPKM	Metatranscriptome RPKM
1	9.200	0.000
2	24.941	19.880
3	8.785	5.784
4	12.852	3.969
5	183.281	260.178
6	263.368	384.195
7	22.092	9.057
8	50.624	27.463
9	227.162	57.237
10	26.257	13.325
11	15.420	9.347
12	16.860	10.969
13	80.303	190.618
Internal Node Range:		
2-10	14.900	14.884
2-13	3.000	0.000
4-6	19.415	0.000
4-10	1.600	0.000
5-6	1.500	0.000
7-10	1.000	0.654
9-10	6.176	7.211
11-13	4.614	1.752
12-13	1.750	0.000

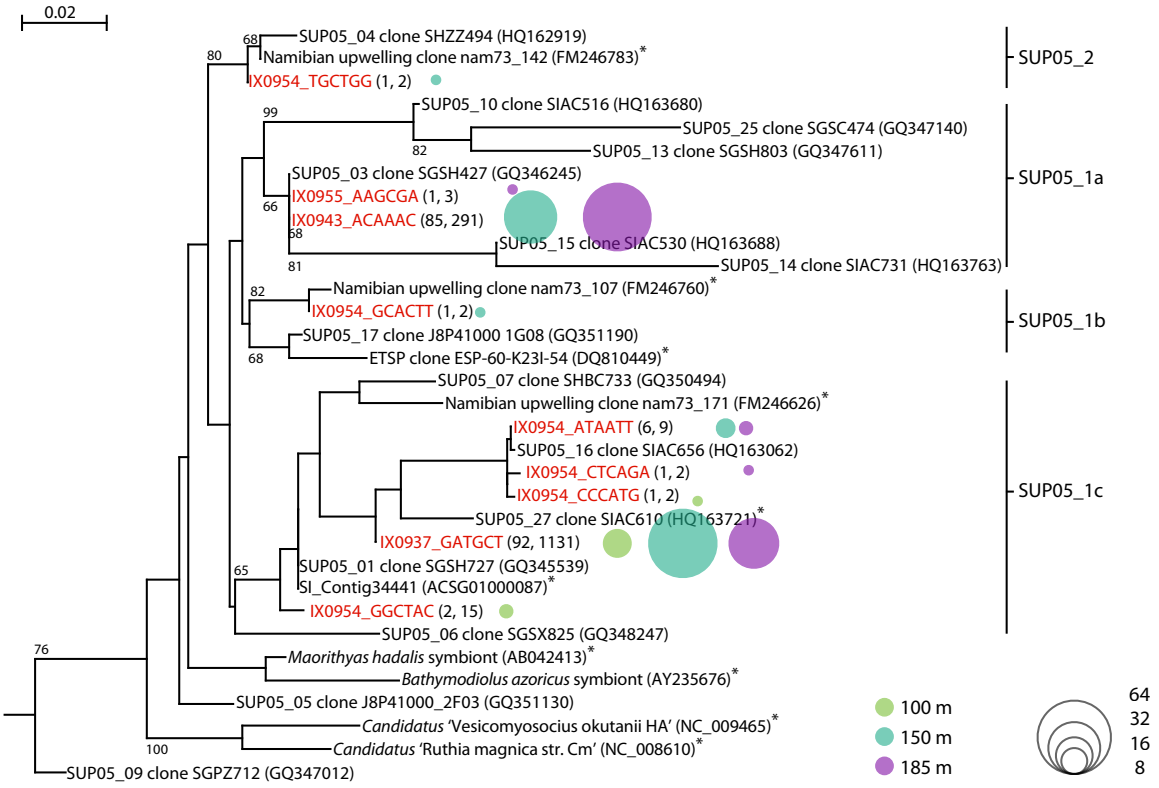


Figure D.1: SUP05 phylogenetic tree. Maximum likelihood tree for SUP05 small subunit ribosomal RNA sequences from Saanich Inlet single cell amplified genomes (SAGs) and publically available datasets. SUP05 clades are indicated to the right of the tree. Number of collected SAGs indicated with bubbles size for 100 m (green) 120 m (blue) and 185m (purple).

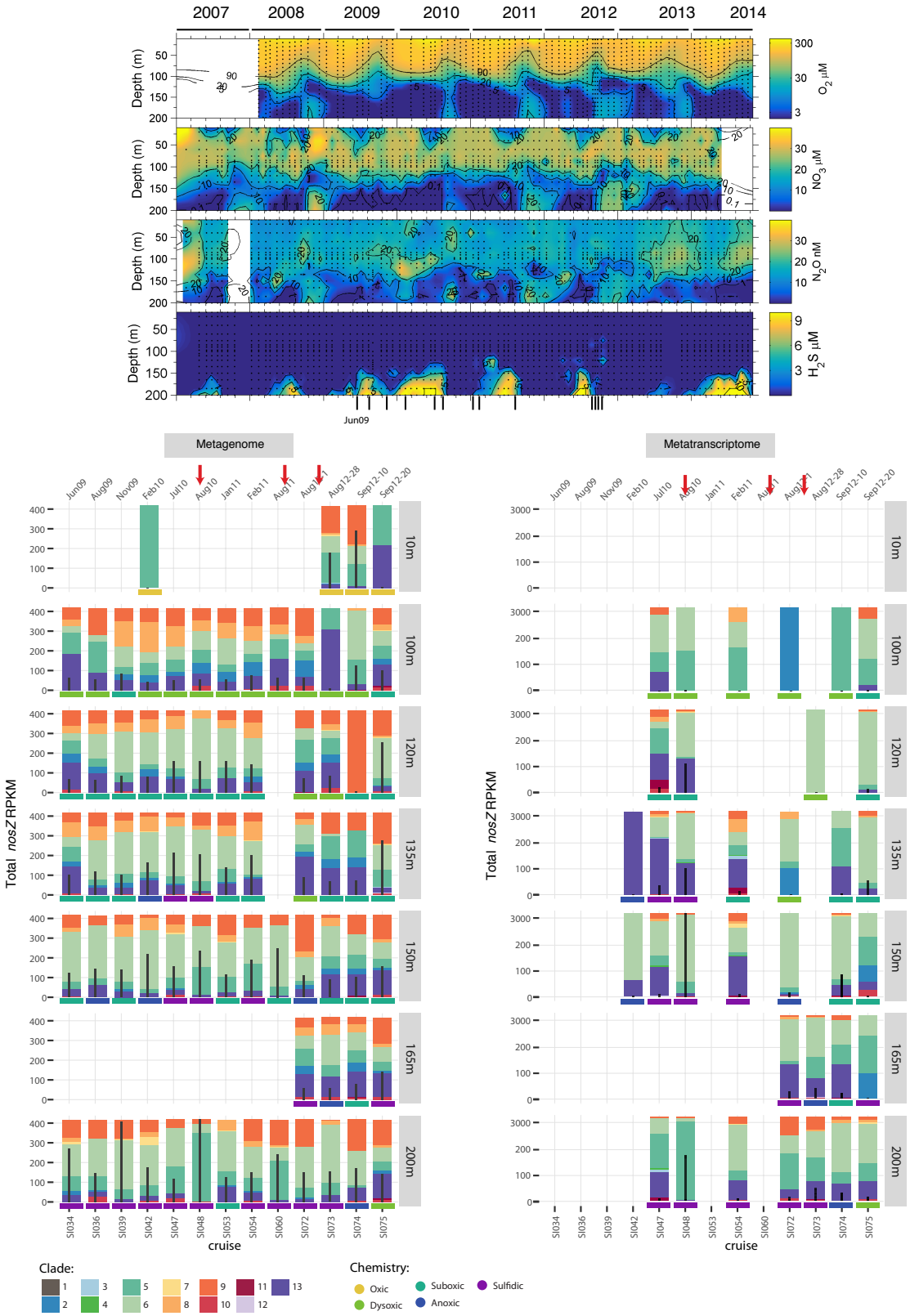


Figure D.2: Proportions of *nosZ* clades in Saanich Inlet metagenome and metatranscriptome. Proportion of *nosZ* clades with total RPKM indicated by black bar. Sample chemistry is indicated by coloured dot below each stacked bar.

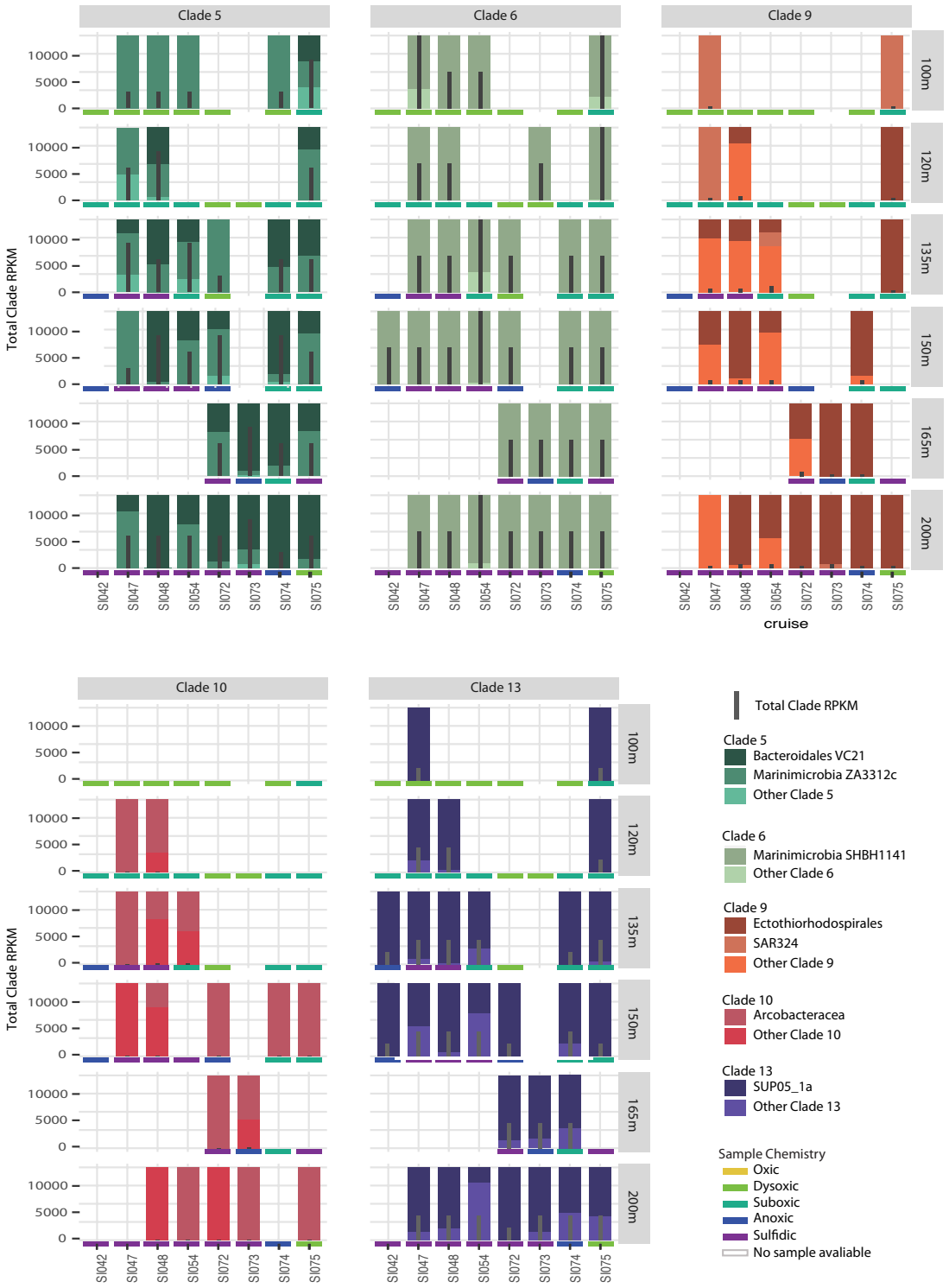


Figure D.3: Proportions of *nosZ* subclades in metatranscriptome. Proportion of RPKM for indicated subclades of *nosZ* with total *nosZ* clade expression indicated as black bar. Sample chemistry is indicated by coloured dot below stacked bar for each sample.

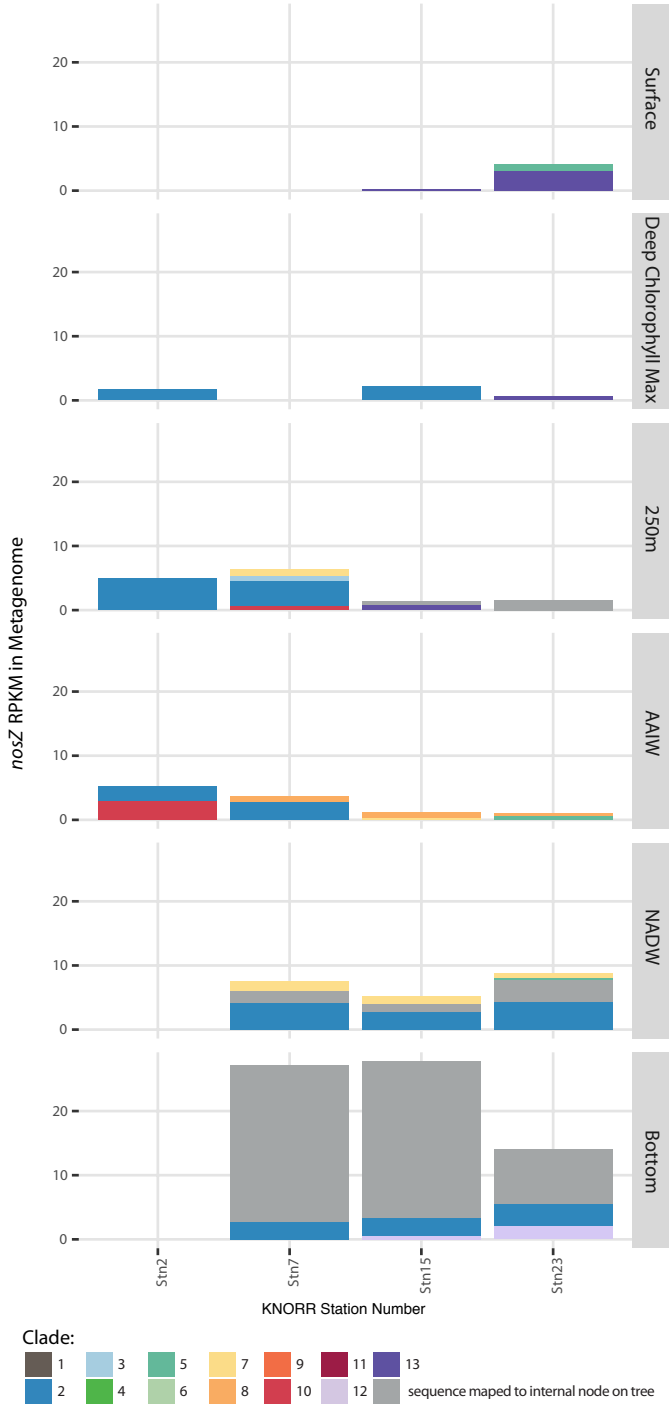


Figure D.4: Abundance of *nosZ* clades for Knorr Cruise. Abundance of *nosZ* in the metagenome from the Knorr cruise in the mid western Atlantic ocean at Station 2 (-38°N, -45°W), Station 7(-22.5°N, -33°W), Station 15 (-2.7°N, -28.5°W) and Station 23 (9.75°N, -55.3°W) at depths including Surface, Deep Chlorophyll Maximum (DCM), 250 m, Antarctic Intermediate Water (AAIW, ~800 m), North Atlantic Deep Water (NADW, ~2500 m), Antarctic Bottom Water (AABW, >4000 m).

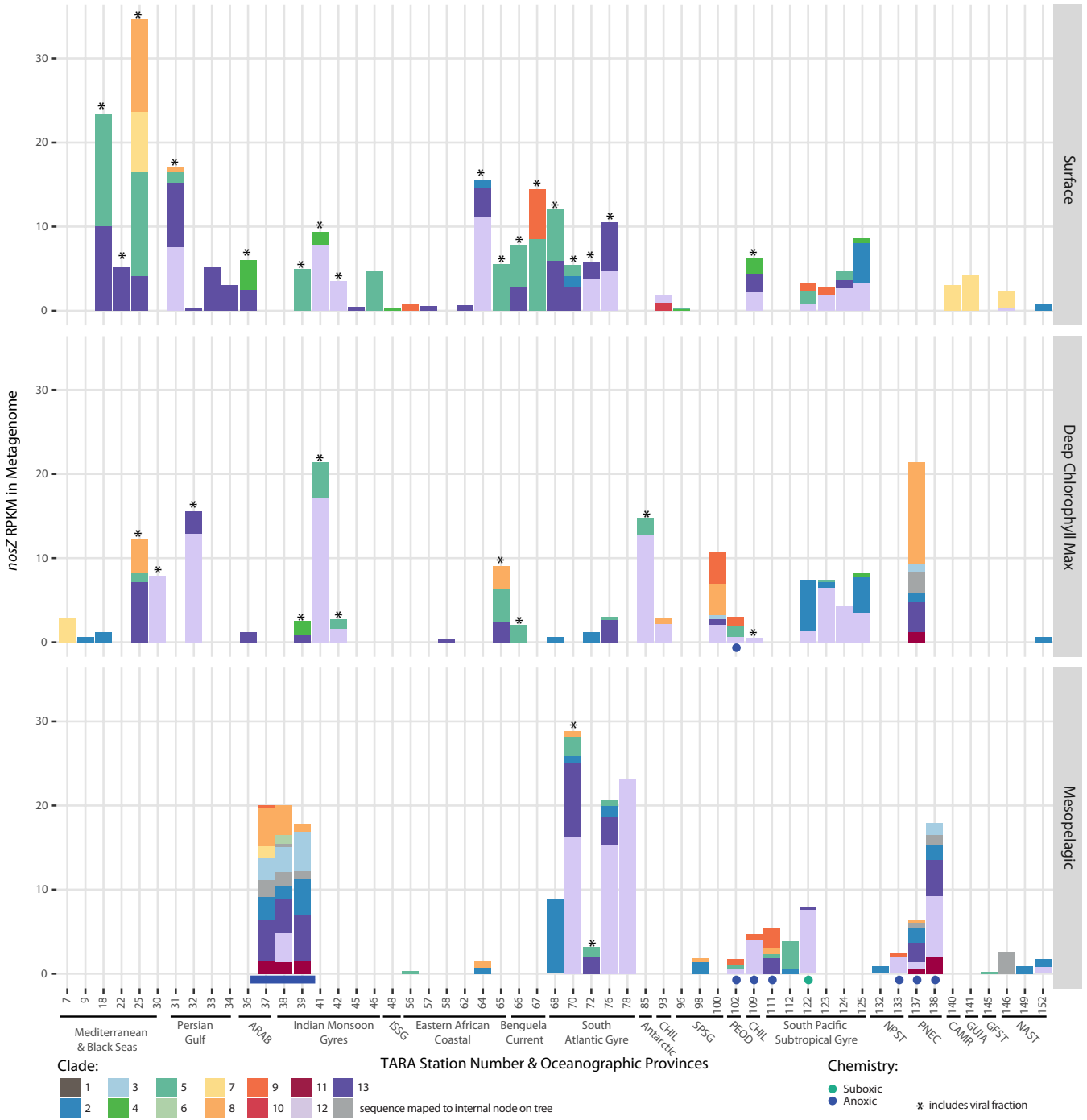

nosZ RPKM in Metagenome
Surface
Deep Chlorophyll Max
Mesopelagic
0
10
20
30
7 9 18 22 25 30 31 32 33 34 36 37 38 39 41 42 45 46 48 56 57 58 62 64 65 66 67 68 70 72 76 78 85 93 96 98 100 102 109 111 112 122 123 124 125 132 133 137 138 140 141 145 146 149 152
Mediterranean & Black Seas
Persian Gulf
ARAB
Indian Monsoon Gyres
ISSG
Eastern African Coastal
Benguela Current
South Atlantic Gyre
Antarctic
CHIL
SPSG
PEOD
CHIL
South Pacific Subtropical Gyre
NPST
PNEC
CAMR
GUIA
GFST
NAST
TARA Station Number & Oceanographic Provinces
Clade:
1 2 3 4 5 6 7 8 9 10 11 12 13
sequence maped to internal node on tree
Chemistry:
Suboxic
Anoxic
* includes viral fraction

Figure D.5: Abundance of *nosZ* clades along TARA Oceans cruise track (previous page). Abundance of *nosZ* in the metagenome from the Tara Global Oceans cruise [264] for Surface, Deep Chlorophyll Maximum and Mesopelagic depths. Oceanographic Provinces abbreviations: Northwest Arabian Sea Upwelling Province (ARAB); Indian South Subtropical Gyre Province(ISSG); Chile-Peru Current Coastal Province (CHIL); South Pacific Subtropical Gyre Province (SPSG); Pacific Equatorial Divergence Province (PEOD); North Pacific Subtropical and Polar Front Provinces (NPST); North Pacific Equatorial Countercurrent Province (PNEC); Central American Coastal Province (CAMR); Guianas Coastal Province(GUIA); Gulf Stream Province (GFST); North Atlantic Subtropical Gyral Province(NAST). For OMZ samples chemical status of the water during sampling is indicated by coloured dot at the base of the stacked bar.