GENE- AND GENOME-CENTRIC ANALYSES OF BACTERIAL NICHE DIMENSIONS IN THE COASTAL OCEAN

by

BRENT NOWINSKI

(Under the Direction of Mary Ann Moran)

ABSTRACT

Heterotrophic bacteria are central players in the ocean's carbon and nutrient cycles, shaping marine ecosystems through their diverse metabolic and ecological functions. This dissertation identifies abiotic and biotic factors that shape bacterial activities and form the dimensions of their niche, a foundational ecological concept to explain the distribution of species in natural environments. Here, the niche dimensions of marine bacteria associated with phytoplankton were addressed through analysis of microbial genes, transcripts, and genomes in dynamic coastal waters. In the first study, diversity and temporal dynamics of genes encoding bacterial transformation of an important resource dimension, the abundant phytoplankton-derived osmolyte dimethylsulfoniopropionate (DMSP), were measured in surface waters of Monterey Bay, CA. Shifts in abundance of paralogous genes that encode production of the volatile sulfur gas dimethylsulfide from DMSP occurred as bacterial communities responded to environmental conditions. Positive relationships between abundance of DMSP-producing dinoflagellates and specific bacterial taxa emerged. In the second study, a time-series dataset of metagenomes, metatranscriptomes, and 16S and 18S rRNA gene libraries over 52 days of a massive dinoflagellate bloom in Monterey Bay is reported. This comprehensive sequence dataset

and accompanying measures of chemical and biological conditions will facilitate studies of the metabolic responses of heterotrophic bacteria during episodic phytoplankton blooms. In the third study, niche dimensions of a well-characterized heterotrophic marine bacterium, *Ruegeria pomeroyi*, were explored by conducting serial invasions of this model bacterium into Monterey Bay bloom seawater and assessing transcriptome composition and apparent growth rates. Differential gene expression patterns indicated relevant substrate, vitamin, nutrient, metal, stress, and biotic interaction factors serving as key niche dimensions in this environment. In the fourth study, genomes were assembled from bloom seawater communities to provide insights into the ecological capabilities of dominant taxa in the natural bacterioplankton community. This revealed two highly related, sequence-discrete species from the roseobacter group that dominated the bacterial community during the bloom. From metapangenomic analysis of 31 genomes from these species, genes involving substrate transformation (polyamines, urea, sugars, sulfur, and carboxylic acids), metal dynamics, vitamin synthesis, and phototrophy provided insights into the dimensions of niche overlap and differentiation in these sympatric species.

INDEX WORDS: Marine Bacteria, Roseobacter, Niche Dimensions, Metagenomics, Transcriptomics, Microbial Invasion Studies, DMSP, Monterey Bay, Metapangenomics

GENE- AND GENOME-CENTRIC ANALYSES OF BACTERIAL NICHE DIMENSIONS IN THE COASTAL OCEAN

by

BRENT NOWINSKI

B.S., Indiana University, 2012

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

GENE- AND GENOME-CENTRIC ANALYSES OF BACTERIAL NICHE DIMENSIONS IN THE COASTAL OCEAN

by

BRENT NOWINSKI

Major Professor:	Mary Ann Moran
Committee:	Brian M. Hopkinson
	Charles S. Hopkinson
	Elizabeth A. Ottesen
	William B. Whitman

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
August 2020

ACKNOWLEDGEMENTS

My time at UGA has been filled with amazing opportunities and memories, and so many people have been a part of making this journey special. First, I'd like to thank my advisor, Mary Ann Moran. Her positivity, work ethic, and creativity are one-of-a-kind, and I am deeply grateful for her support. Thank you to my committee and peers for shaping my scientific interests and always providing great feedback. I am grateful to members of the ESP team at MBARI and the DMSP team at DISL for the opportunity to sample at sea with such a helpful and accommodating group. I would also like to thank all past and present members of the Moran Lab for assistance in the lab and at sea, and for the always jovial office environment. Finally, I'd like to thank my friends and family for their encouragement along the way.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Marine microbes occupy niches in virtually every oceanic environment, from the surface microlayer down through the water column, deep within sediments, and within and attached to organisms and particles (Whitman *et al.*, 1998; Bar-On *et al.*, 2018; Gasol and Kirchman, 2018). In the surface ocean environment, microbes play roles that are integral to the cycling of carbon, nutrients, and energy. Here, phytoplankton are responsible for half the photosynthetic fixation of carbon on Earth (Field *et al.*, 1998), much of which fluxes into the dissolved organic matter (DOM) pool and is rapidly degraded by bacterioplankton (Pomeroy, 1974; Azam *et al.*, 1983). The DOM transformed by these microbes is diverse, with tens to hundreds of thousands of unique organic compounds represented in seawater (Hertkorn *et al.*, 2006). This vast pool of DOM is dynamic, with different groups of phytoplankton producing unique suites of metabolites (Becker *et al.*, 2014). The ability of bacterioplankton to degrade and compete for these compounds, encoded in their genomes, shapes the niche space that each taxon occupies.

In addition to the diverse resources available in the DOM pool, other abiotic and biotic factors determine the ecological success of marine bacteria. Physical factors, such as light, temperature, and salinity, drive distributions of marine microbes. For example, ecotypes of the dominant marine cyanobacterium *Prochlorococcus* exhibit physiological adaptations to high or low light conditions that drive partitioning in the water column (Moore *et al.*, 1998; Rocap *et al.*, 2003), and the abundant heterotrophic bacteria of the SAR11 clade show patterns of ecotype distribution associated with seasonal mixing and stratification of the water column (Carlson *et*

al., 2009). Chemical factors, in addition to organic substrates, drive success of bacterioplankton, including availability of nutrients and metals (Church *et al.*, 2000; Pinhassi *et al.*, 2006). Biotic interactions also greatly influence bacterial viability, including both mutualism and competition between species (Persson *et al.*, 2009; Yeung *et al.*, 2012).

The totality of these external factors form the dimensions of each bacterial species' niche and determine where it can persist in the ocean (Hutchinson, 1957). Improved identification of these dimensions will facilitate understanding and modeling of the processes, molecules, and interactions that drive biogeochemical cycling. This dissertation details aspects of a major niche dimension driving bacterial success in the surface ocean (DMSP degradation), and then identifies key niche dimensions from a genome-centric view of members of the roseobacter group that are abundant and active in the coastal ocean (Buchan *et al.*, 2005).

DMSP as an important niche dimension of marine bacterioplankton

Dimethylsulfoniopropionate (DMSP) is an abundant and ubiquitous osmolyte biosynthesized by phytoplankton at rates high enough to account for 10% of fixed carbon (Archer *et al.*, 2001). A large fraction of intracellular DMSP is released to the marine DOM pool where it is degraded rapidly by bacteria, typically within hours to days (Kiene and Linn, 2000; Zubkov *et al.*, 2002), and can meet up to 15% and 100% of bacterial carbon and sulfur demands, respectively, in marine surface waters. (Kiene *et al.*, 2000). Bacteria utilize two major pathways to degrade DMSP. In the demethylation pathway, cells use DMSP as a sulfur source through incorporation into sulfur-containing amino acids. In the cleavage pathway, the volatile gas dimethylsulfide (DMS) that is generated fluxes from the ocean surface to the atmosphere where it is implicated in regulation of climate through cloud-condensation nuclei generation and albedo

effects (Charlson *et al.*, 1987). The gene *dmdA* catalyzes the first step in demethylation, producing methylmercaptopropionate, while seven non-homologous *ddd* genes act as the first step in the cleavage pathway (Howard *et al.*, 2006; Curson *et al.*, 2008; Todd *et al.*, 2009; Todd *et al.*, 2011; Peng *et al.*, 2012; Todd *et al.*, 2012; Sun *et al.*, 2016). **Chapter 2** describes analysis of the DMSP-degrading community in Monterey Bay, CA, USA during Fall 2014. Metagenomic sampling of surface seawater was used to measure the abundance and diversity of bacterial genes encoding DMSP degradation pathways in the context of shifting phytoplankton communities sampled over a three-week period.

Methods for analyzing genes and gene expression of marine microbes

Over forty years ago, the small subunit of ribosomal RNA was recognized as a key tool for describing evolutionary relationships between organisms (Woese and Fox, 1977). Methods for analyzing gene sequences of this molecule from the environment allowed some of the first insights into marine organisms that had not been cultivated (Olsen *et al.*, 1986). Studies of 16S rRNA genes in bacteria and archaea and 18S rRNA genes in eukaryotes have shed light on the diversity and relationships of microbes in the ocean (Giovannoni and Rappé, 2000). Metagenomics, the study of genetic material recovered from the environment, has been employed more recently to gain insights into key taxa and functions present in microbial communities (Tringe *et al.*, 2005; DeLong *et al.*, 2006). These approaches have helped characterize major microbial distributions, dynamics, and associations across the oceans (Lima-Mendez *et al.*, 2015; Sunagawa *et al.*, 2015). Metatranscriptomics, the study of community-wide gene expression (Poretsky *et al.*, 2005; Frias-Lopez *et al.*, 2008), complements metagenomics by pinpointing when and where a gene is expressed, resolving timing of key processes by specific

members of the microbial community (Ottesen *et al.*, 2014; Aylward *et al.*, 2015). **Chapter 3** describes the generation of a high-resolution sampling of a massive dinoflagellate bloom in Fall of 2016 in Monterey Bay. Over 41 dates, 84 metagenomes, 82 metatranscriptomes, 88 16S rRNA gene libraries, and 88 18S rRNA gene libraries were generated, providing a sequence inventory of unprecedented coverage and temporal resolution of microbial communities in episodic blooms.

Using invasion experiments to identify niche dimensions

Hutchinson (1957) defined the ecological niche in the context of two key concepts. The fundamental niche is the set of resources and conditions which allow a species to survive in the absence of interactions with other organisms, while the realized niche incorporates these interactions. Much work in ecology has focused on negative interactions, such as competition and predation, which leads to the conclusion that the realized niche is smaller than the fundamental niche. Indeed, the ocean is a highly competitive environment, yet we also observe levels of species richness that indicate coexistence among many marine microbes with similar ecological roles (Hutchinson, 1961). This paradox is reconciled when considering the massive molecular diversity of substrates from the DOM pool (Zark *et al.*, 2017), spatial heterogeneity and resource gradients in phytoplankton phycospheres (Stocker, 2012), and predation and viral dynamics (Kirchman, 2010), offering a myriad of multi-dimensional niches in which species can exist. Recently, facilitation has been inferred to play a major role in shaping marine microbial communities, and these positive interactions, such as cross-feeding and release of public goods (Morris *et al.*, 2012; Pacheco *et al.*, 2019), could extend the conditions under which a microbe

can survive, resulting in realized niches that are greater than the corresponding fundamental niche (Bruno *et al.*, 2003).

The analysis of the genes and transcripts of bacterioplankton can help elucidate key aspects of niche theory, as the genes of an organism represent coding potential for responding to niche dimensions, while transcripts signify when these dimensions affect the ability of a bacterium to survive in a specific environment (Muller, 2019). Metatranscriptomics can identify metabolic capabilities and lifestyle strategies that allow coexistence in microbial assemblages (Gifford *et al.*, 2013). Additionally, model organism studies can address niche dimensions by testing growth responses under defined conditions (Martens-Habbena *et al.*, 2009). **Chapter 4** uses a combination of these two approaches for a serial invasion study of a well-characterized heterotrophic marine bacterium from the roseobacter group, introduced into seawater from different phases of a natural phytoplankton bloom. Resulting transcriptomic profiles of the bacterium highlighted genes that responded to factors (dimensions) in the bloom environment, and inferred viability of the bacterium through time.

Genome-centric approaches to study niche partitioning in abundant marine microbes

Sequencing of genomes of microbial isolates provides ecological and evolutionary insights. However, a continuing challenge in microbial ecology is the isolation and culturing of environmental bacteria and archaea (Steen *et al.*, 2019). Improvements in nucleic acid sequencing technology and bioinformatics methods provide the ability to access the genomic potential of uncultured microbes in their environment, without the need for culturing (McMahon, 2015). Deep sequencing of microbial communities now allows assembly of metagenome-assembled genomes (MAGs) representing the dominant populations in marine habitats (Tully *et*

al., 2018). Sorting and sequencing of microbial cells yields individual cell genomes (single amplified genomes; SAGs) for members of microbial communities (Berube *et al.*, 2018). Genomes obtained from these methods have revealed novel biological processes (Brown *et al.*, 2015) and phylogenomic diversity (Hug *et al.*, 2016). Analysis of pangenomes, the full set of genes represented in the members of a taxon, can be applied to collections of MAGs and SAGs to reveal taxon boundaries and gain insights into microbial adaptation, specialization, and evolution (Delmont and Eren, 2018; Jarett *et al.*, 2018; Moulana *et al.*, 2020). **Chapter 5** describes the use of genome-centric approaches to discover and characterize two highly-related bacterial species from the roseobacter group that were both dominant members of the Fall 2016 Monterey Bay bloom community, identifying niche dimensions that differentiate these globally abundant sympatric species.

References

Archer, S.D., Widdicombe, C.E., Tarran, G.A., Rees, A.P., and Burkill, P.H. (2001) Production and turnover of particulate dimethylsulphoniopropionate during a coccolithophore bloom in the northern North Sea. *Aquatic Microbial Ecology* **24**: 225-241.

Aylward, F.O., Eppley, J.M., Smith, J.M., Chavez, F.P., Scholin, C.A., and DeLong, E.F. (2015) Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proceedings of the National Academy of Sciences* **112**: 5443-5448.

Azam, F., Fenchel, T., Field, J.G., Gray, J., Meyer-Reil, L., and Thingstad, F. (1983) The ecological role of water-column microbes in the sea. *Marine ecology progress series*: 257-263.

Bar-On, Y.M., Phillips, R., and Milo, R. (2018) The biomass distribution on Earth. *Proceedings of the National Academy of Sciences* **115**: 6506-6511.

Becker, J.W., Berube, P.M., Follett, C.L., Waterbury, J.B., Chisholm, S.W., DeLong, E.F., and Repeta, D.J. (2014) Closely related phytoplankton species produce similar suites of dissolved organic matter. *Frontiers in microbiology* **5**: 111.

Berube, P.M., Biller, S.J., Hackl, T., Hogle, S.L., Satinsky, B.M., Becker, J.W. et al. (2018) Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. *Scientific data* **5**: 180154.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A. et al. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**: 208-211.

Bruno, J.F., Stachowicz, J.J., and Bertness, M.D. (2003) Inclusion of facilitation into ecological theory. *Trends in ecology & evolution* **18**: 119-125.

Buchan, A., González, J.M., and Moran, M.A. (2005) Overview of the marine Roseobacter lineage. *Applied and environmental microbiology* **71**: 5665-5677.

Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., and Vergin, K. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *The ISME journal* **3**: 283-295.

Charlson, R.J., Lovelock, J.E., Andreae, M.O., and Warren, S.G. (1987) Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**: 655-661.

Church, M.J., Hutchins, D.A., and Ducklow, H.W. (2000) Limitation of bacterial growth by dissolved organic matter and iron in the Southern Ocean. *Applied and Environmental Microbiology* **66**: 455-466.

Curson, A., Rogers, R., Todd, J., Brearley, C., and Johnston, A. (2008) Molecular genetic analysis of a dimethylsulfoniopropionate lyase that liberates the climate-changing gas dimethylsulfide in

several marine α-proteobacteria and Rhodobacter sphaeroides. *Environmental microbiology* **10**: 757-767.

Delmont, T.O., and Eren, A.M. (2018) Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. *PeerJ* **6**: e4320.

DeLong, E.F., Preston, C.M., Mincer, T., Rich, V., Hallam, S.J., Frigaard, N.-U. et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**: 496-503.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**: 237-240.

Frias-Lopez, J., Shi, Y., Tyson, G.W., Coleman, M.L., Schuster, S.C., Chisholm, S.W., and DeLong, E.F. (2008) Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences* **105**: 3805-3810.

Gasol, J.M., and Kirchman, D.L. (2018) *Microbial ecology of the oceans*: John Wiley & Sons.

Gifford, S.M., Sharma, S., Booth, M., and Moran, M.A. (2013) Expression patterns reveal niche diversification in a marine microbial assemblage. *The ISME journal* **7**: 281-298.

Giovannoni, S., and Rappé, M. (2000) Evolution, diversity and molecular ecology of marine prokaryotes. *Microbial ecology of the oceans*: 47-84.

Hertkorn, N., Benner, R., Frommberger, M., Schmitt-Kopplin, P., Witt, M., Kaiser, K. et al. (2006) Characterization of a major refractory component of marine dissolved organic matter. *Geochimica et Cosmochimica Acta* **70**: 2990-3010.

Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Bürgmann, H., Welsh, R. et al. (2006) Bacterial taxa that limit sulfur flux from the ocean. *Science* **314**: 649-652.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J. et al. (2016) A new view of the tree of life. *Nature microbiology* **1**: 1-6.

Hutchinson, G.E. (1957) Cold spring harbor symposium on quantitative biology. *Concluding remarks* **22**: 415-427.

Hutchinson, G.E. (1961) The paradox of the plankton. *The American Naturalist* **95**: 137-145.

Jarett, J.K., Nayfach, S., Podar, M., Inskeep, W., Ivanova, N.N., Munson-McGee, J. et al. (2018) Single-cell genomics of co-sorted Nanoarchaeota suggests novel putative host associations and diversification of proteins involved in symbiosis. *Microbiome* **6**: 1-14.

Kiene, R.P., and Linn, L.J. (2000) Distribution and turnover of dissolved DMSP and its relationship with bacterial production and dimethylsulfide in the Gulf of Mexico. *Limnology and Oceanography* **45**: 849-861.

Kiene, R.P., Linn, L.J., and Bruton, J.A. (2000) New and important roles for DMSP in marine microbial communities. *Journal of Sea Research* **43**: 209-224.

Kirchman, D.L. (2010) *Microbial ecology of the oceans*: John Wiley & Sons.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F. et al. (2015) Determinants of community structure in the global plankton interactome. *Science* **348**.

Martens-Habbena, W., Berube, P.M., Urakawa, H., José, R., and Stahl, D.A. (2009) Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* **461**: 976-979.

McMahon, K. (2015) Metagenomics 2.0. *Environmental Microbiology Reports* **7**: 38-39.

Moore, L.R., Rocap, G., and Chisholm, S.W. (1998) Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes. *Nature* **393**: 464-467.

Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio* **3**.

Moulana, A., Anderson, R.E., Fortunato, C.S., and Huber, J.A. (2020) Selection Is a Significant Driver of Gene Gain and Loss in the Pangenome of the Bacterial Genus Sulfurovum in Geographically Distinct Deep-Sea Hydrothermal Vents. *Msystems* **5**.

Muller, E.E. (2019) Determining microbial niche breadth in the environment for better ecosystem fate predictions. *Msystems* **4**.

Olsen, G.J., Lane, D.J., Giovannoni, S.J., Pace, N.R., and Stahl, D.A. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annual reviews in microbiology* **40**: 337-365.

Ottesen, E.A., Young, C.R., Gifford, S.M., Eppley, J.M., Marin, R., Schuster, S.C. et al. (2014) Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* **345**: 207-212.

Pacheco, A.R., Moel, M., and Segrè, D. (2019) Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature communications* **10**: 1-12.

Peng, M., Xie, Q., Hu, H., Hong, K., Todd, J.D., Johnston, A.W., and Li, Y. (2012) Phylogenetic diversity of the dddP gene for dimethylsulfoniopropionate-dependent dimethyl sulfide synthesis in mangrove soils. *Canadian journal of microbiology* **58**: 523-530.

Persson, O.P., Pinhassi, J., Riemann, L., Marklund, B.I., Rhen, M., Normark, S. et al. (2009) High abundance of virulence gene homologues in marine bacteria. *Environmental microbiology* **11**: 1348-1357.

Pinhassi, J., Gómez-Consarnau, L., Alonso-Sáez, L., Sala, M.M., Vidal, M., Pedrós-Alió, C., and Gasol, J.M. (2006) Seasonal changes in bacterioplankton nutrient limitation and their effects on bacterial community composition in the NW Mediterranean Sea. *Aquatic Microbial Ecology* **44**: 241-252.

Pomeroy, L.R. (1974) The ocean's food web, a changing paradigm. *Bioscience* **24**: 499-504.

Poretsky, R.S., Bano, N., Buchan, A., LeCleir, G., Kleikemper, J., Pickering, M. et al. (2005) Analysis of microbial gene transcripts in environmental samples. *Applied and Environmental Microbiology* **71**: 4121-4126.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A. et al. (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042-1047.

Steen, A.D., Crits-Christoph, A., Carini, P., DeAngelis, K.M., Fierer, N., Lloyd, K.G., and Thrash, J.C. (2019) High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME journal* **13**: 3126-3130.

Stocker, R. (2012) Marine microbes see a sea of gradients. *science* **338**: 628-633.

Sun, J., Todd, J.D., Thrash, J.C., Qian, Y., Qian, M.C., Temperton, B. et al. (2016) The abundant marine bacterium Pelagibacter simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. *Nature microbiology* **1**: 1-5.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G. et al. (2015) Structure and function of the global ocean microbiome. *Science* **348**.

Todd, J., Curson, A., Dupont, C., Nicholson, P., and Johnston, A. (2009) The dddP gene, encoding a novel enzyme that converts dimethylsulfoniopropionate into dimethyl sulfide, is widespread in ocean metagenomes and marine bacteria and also occurs in some Ascomycete fungi. *Environmental microbiology* **11**: 1376-1385.

Todd, J.D., Kirkwood, M., Newton-Payne, S., and Johnston, A.W. (2012) DddW, a third DMSP lyase in a model Roseobacter marine bacterium, Ruegeria pomeroyi DSS-3. *The ISME journal* **6**: 223-226.

Todd, J.D., Curson, A.R., Kirkwood, M., Sullivan, M.J., Green, R.T., and Johnston, A.W. (2011) DddQ, a novel, cupin-containing, dimethylsulfoniopropionate lyase in marine roseobacters and in uncultured marine bacteria. *Environmental microbiology* **13**: 427-438.

Tringe, S.G., Von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W. et al. (2005) Comparative metagenomics of microbial communities. *Science* **308**: 554-557.

Tully, B.J., Graham, E.D., and Heidelberg, J.F. (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific data* **5**: 170203.

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

Woese, C.R., and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* **74**: 5088-5090.

Yeung, L.Y., Berelson, W.M., Young, E.D., Prokopenko, M.G., Rollins, N., Coles, V.J. et al. (2012) Impact of diatom-diazotroph associations on carbon export in the Amazon River plume. *Geophysical Research Letters* **39**.

Zark, M., Christoffers, J., and Dittmar, T. (2017) Molecular properties of deep-sea dissolved organic matter are predictable by the central limit theorem: evidence from tandem FT-ICR-MS. *Marine Chemistry* **191**: 9-15.

Zubkov, M.V., Fuchs, B.M., Archer, S.D., Kiene, R.P., Amann, R., and Burkill, P.H. (2002) Rapid turnover of dissolved DMS and DMSP by defined bacterioplankton communities in the stratified euphotic zone of the North Sea. *Deep Sea Research Part II: Topical Studies in Oceanography* **49**: 3017-3038.

CHAPTER 2

MICRODIVERSITY AND TEMPORAL DYNAMICS OF MARINE BACTERIAL DIMETHYLSULFONIOPROPIONATE GENES [1]

[1] Nowinski, B, Motard-Côté, J, Landa, M, Preston, CM, Scholin, CA, Birch, JM, Kiene, RP, Moran, MA. 2019. *Environmental Microbiology* 21(5): 1687-1701.

Summary

Dimethylsulfoniopropionate (DMSP) is an abundant organic sulfur metabolite produced by many phytoplankton species and degraded by bacteria via two distinct pathways with climate-relevant implications. We assessed the diversity and abundance of bacteria possessing these pathways in the context of phytoplankton community composition over a three-week time period spanning September – October, 2014 in Monterey Bay, CA. The *dmdA* gene from the DMSP demethylation pathway dominated the DMSP gene pool and was harbored mostly by members of the alphaproteobacterial SAR11 clade and secondarily by the Roseobacter group, particularly during the second half of the study. Novel members of the DMSP-degrading community emerged from *dmdA* sequences recovered from metagenome assemblies and single-cell sequencing, including largely uncharacterized Gammaproteobacteria and Alphaproteobacteria taxa. In the DMSP cleavage pathway, the SAR11 gene *dddK* was the most abundant early in the study, but was supplanted by *dddP* over time. SAR11 members, especially those harboring genes for both DMSP degradation pathways, had a strong positive relationship with the abundance of dinoflagellates, and DMSP- degrading Gammaproteobacteria co-occurred with haptophytes. This *in situ* study of the drivers of DMSP fate in a coastal ecosystem demonstrates for the first time correlations between specific groups of bacterial DMSP degraders and phytoplankton taxa.

Introduction

The phytoplankton-produced organic sulfur compound dimethylsulfoniopropionate (DMSP) plays a major role in marine microbial food webs as a source of reduced sulfur and carbon (Kiene *et al*., 2000), and its degradation has impacts on ocean-atmosphere sulfur flux (Andreae, 1990). Marine bacteria can catabolize DMSP using the demethylation pathway,

wherein sulfur is routed to methanethiol and potentially incorporated into biomass; or using the cleavage pathway, in which case sulfur is routed to the gas dimethylsulfide (DMS) with implications for atmospheric aerosol dynamics and cloud formation (Charlson *et al*., 1987; Quinn and Bates, 2011). The gene *dmdA* catalyzes the first step in the demethylation pathway, removing a methyl group and yielding methylmercaptopropionate, while seven different non-homologous *ddd* genes can catalyze the first step in cleavage, generating DMS and acrylate (Curson *et al*., 2008; Howard *et al*., 2006; Peng *et al*., 2012; Sun *et al*., 2016; Todd *et al*., 2009; Todd *et al*., 2011; Todd *et al*., 2012). *dmdA* and the *ddd* genes can be considered gatekeeper genes for the entry of DMSP into these two competing pathways.

The main taxa implicated in most DMSP degradation studies are marine members of the class Alphaproteobacteria, with genes discovered and characterized in the Roseobacter group, the SAR11 clade, and the SAR116 cluster. The first bacterial *dmdA* gene described was in a model coastal heterotrophic bacterium from the Roseobacter group, *Ruegeria pomeroyi* (Howard *et al.*, 2006), and was subsequently found in many other roseobacters. Many members of this group also harbor DMSP cleavage genes *dddD*, *dddL*, *dddP*, *dddQ*, and *dddW*, either alone or in various combinations (Moran *et al.*, 2012). Many SAR11 bacteria also possess *dmdA* (Howard *et al.*, 2006), and recent research has discovered *dddK*, a new DMSP cleavage gene specific to SAR11 cells (Sun *et al*., 2016). SAR116 members can possess *dmdA* and *dddP* (Oh *et al.*, 2010).

The coastal upwelling system of Monterey Bay, CA, harbors diverse and rapidly-changing phytoplankton communities and high numbers of DMSP-degrading bacteria (Varaljay *et al*., 2015). How these shifting phytoplankton communities, and thus DMSP availability, may affect the DMSP- degrading bacterial community and the metabolic routing of organic sulfur to the divergent biogeochemical fates of the two degradation pathways were the questions that

drove our study. We utilized the Environmental Sample Processor (ESP), an autonomous robotic instrument than can remain deployed in an environment for an extended period to repeatedly collect and preserve microbial community samples for later nucleic acid sequencing (Ottesen *et al.*, 2011). The ESP sampled surface waters in Monterey Bay for three weeks in the fall of 2014. Six sample dates from this collection varying in DMSP and chlorophyll *a* levels were selected for metagenomic analysis of DMSP-related genes. We examined the diversity and abundance of the DMSP-degrading bacterial community using metagenomic and single cell sequencing in the context of phytoplankton dynamics and environmental parameters.

Results

Overview of Microbial Dynamics

Concentrations of chlorophyll *a* (as a proxy for phytoplankton biomass) and total dissolved plus particulate DMSP concentration (DMSPt) each varied about ten-fold during the course of the study, from 0.8 to 7.2 ug L^{-1} chlorophyll *a* and from 32 to 311 nM DMSPt (Fig. 2.1a). Based on metagenomic data, the phytoplankton community was dominated by reads mapping to diatoms (Coscinodiscophyceae, Bacillariophyceae, Mediophyceae, and Fragilariophyceae), haptophytes (Haptophyceae), picoeukaryotes (Mamiellophyceae), and dinoflagellates (Dinophyceae) (Fig. 2.1b,c), with largest changes in relative abundance occurring toward the end of the deployment. Centric diatoms from the Coscinodiscophyceae and dinoflagellates peaked in relative abundance early in the study (September 29 sample date, corresponding to the peak of DMSPt concentration). Pennate diatoms from the Bacillariophyceae exhibited a marked increase in relative abundance in the second week in October (October 8 and 13 sample dates). Haptophytes showed highest relative abundances during the first two weeks of

the deployment (September 22, 29, October 3, 6 sample dates). There was a significant negative linear regression of relative abundance over time for Haptophyceae ($R^2 = 0.35$; $p<0.05$) and Dinophyceae ($R^2 = 0.51$; $p<0.01$). At the October 13 sample date, the phytoplankton community had shifted substantially, with picoeukaryotes from the Mamiellophyceae, diatoms from the Bacillariophyceae, and pelagophytes (Pelagophyceae) increasing in relative abundance (Fig. 2.1b,c), coincident with the lowest chlorophyll *a* and DMSPt concentrations measured during the study period (Fig. 2.1a). The bacterial community was dominated by Alphaproteobacteria, including members of the known DMSP-catabolizing groups SAR11, Roseobacter, and SAR116. Metagenomic reads mapping to Gammaproteobacteria and Bacteroidetes were also highly represented (Fig. 2.1d). There was a significant negative linear regression of the relative abundance over time for SAR11 taxa ($R^2 = 0.65$; $p<0.01$).

Demethylation Gene Diversity

To characterize the phylogeny of predicted *dmdA* genes in the metagenomic data, a base tree for read placement was constructed with full-length DmdA sequences obtained by three methods: 1) assembled from the metagenomes using two different approaches (see Methods), 2) obtained from genomes of single cells collected at the ESP deployment site, and 3) acquired from the NCBI RefSeq database based on best hits to the metagenomic reads (Fig. 2.2) (See Methods). For approach 2, six single cell bacterial genomes from seawater collected at the ESP mooring were selected for sequencing following screening for 16S rRNA genes indicative of likely DMSP-degrading taxa; these included three from the SAR11 clade and three from the Roseobacter group (see Methods). After building the *dmdA* base tree, metagenomic reads identified as *dmdA* based on BLAST analysis against a custom reference database were placed

on the tree and relative abundances were calculated by normalizing to the universal single copy bacterial gene *recA* (Howard *et al.*, 2006).

From this approach, it emerged that SAR11 was the most abundant DMSP-demethylating taxon, with one in five bacterial cells in Monterey Bay possessing a SAR11-like *dmdA* (Fig. 2.2). Most SAR11 *dmdA* reads aligned with the SAR11 subclade containing strain HTCC9022, and this taxon recruited the most *dmdA* reads during the study overall. Also abundant were *dmdA* reads mapping to a SAR11 subclade represented by Monterey Bay single cell SCGC-AG-145-C19 obtained in this study. *dmdA* reads mapping to reference genomes from SAR11 subclade 1a.1, a group typical of temperate coastal regions (Brown *et al.*, 2012; Giovannoni, 2017), including HTCC1062 and HTCC1002, represented the third most abundant SAR11 DmdA group, while the subclade including HTCC7211, which more typically dominates in warm, stratified open oceans, was not abundant (Fig. 2.2).

DmdA sequences mapping to roseobacters made up the next most frequent group. The sequence from single cell genome SCGC AG-145-N17 obtained in this study recruited the highest number of Roseobacter-like reads. Reads mapping to Rhodobacterales bacterium HTCC2255, a bacterium found previously in Monterey Bay (Varaljay *et al.,* 2015), and those mapping to Rhodobacteraceae bacterium SB2, a member of the globally abundant CHAB-I-5 lineage (Billerbeck *et al.*, 2016), also accounted for a substantial fraction of *dmdA* reads (Fig. 2.2). These three Roseobacter taxa share similar characteristics, represented by cells with small genomes and low G+C content, in contrast to the readily cultured members of this group that dominate genomic databases.

The *dmdA* genes in the two SAR116 reference genomes belong to divergent clusters (Fig. 2.2), suggesting two different evolutionary origins of DMSP demethylation capability in this

taxon. The Monterey Bay reads mapped to both clusters. The majority of gammaproteobacterial *dmdA* reads recruited to a DmdA assembled in this study (MEGAHIT_03; Fig. 2.2) that has low similarity (<62% identity) to the DmdA in the only recognized DMSP-demethylating gammaproteobacterium prior to this study (strain HTCC2080; Cho and Giovannoni, 2004).

Reference Gene Recruitment Comparison

The three different approaches used to obtain full-length Monterey Bay-specific DmdA sequences were compared to determine which provided the most relevant reference sequences, i.e., those that recruited the most *dmdA*-like reads from the metagenomes. The 27 full-length sequences obtained using the approach that assembled only those reads first identified as *dmdA* by BLAST analysis of the metagenomic data (hereafter referred to as SPAdes assemblies) recruited the most metagenomic reads (48%) (Fig. 2.S1a). The nine full-length *dmdA* genes obtained from contigs assembled from the full metagenomic dataset (referred to as MEGAHIT assemblies) recruited 6% of reads. Four identical sequences were assembled independently by each of these strategies, and these recruited 3% of reads. The two *dmdA* genes identified in the screened SAGs represented novel subclusters (Pelagibacter-like SCGC AG-145-C19 and Rhodobacterales bacterium SCGC AG- 145-N17) and recruited 2% of reads. Finally, the 65 *dmdA* genes pulled from reference genomes in preexisting databases recruited 39% of reads. The only *dmdA* sequence obtained both by this reference genome method and a metagenomic assembly method was that of Rhodobacteraceae bacterium SB2 (Fig. 2.2). Overall, the approach of identifying reads from the metagenomes with similarity to known *dmdA*s and then assembling this reduced dataset provided the highest number of relevant reference genes for taxonomic placement of metagenomic reads. This method added substantial new diversity in a number of

clades, such as the abundant HTCC9022 cluster (Fig. 2.2).

Demethylation Gene Temporal Dynamics

Temporal trends in abundance for cells harboring these major DmdA clusters were analyzed using edgePCA (Matsen IV and Evans, 2013) (Fig. 2.3), a method that uses read placement variation on either side of each internal edge of a phylogenetic tree as input data for a principal components analysis. The first principal component axis (accounting for 77% of the variation) separated SAR11- dominated September samples from Roseobacter-enriched October samples (Fig. 2.3a). The SAR11 taxa that significantly influenced this axis fell into the HTCC9022 clade (Pearson's *r* of PC1 loadings versus gene abundance, $p<0.01$), 1a.1 subclade ($p<0.01$), and single cell SCGC-AG-145-C19 clade ($p<0.05$) (Fig. 2.3b, Table 2.S1). Roseobacter taxa that influenced PC1 were October-dominant members whose DmdA sequences fell into Rhodobacterales SCGC AG-145-N17 ($p<0.05$), Rhodobacteraceae bacterium SB2 ($p<0.05$), and Roseobacter sp. LE17 clades ($p<0.01$) (Fig. 2.3b). The second principal component axis (accounting for 10% of the variation) was significantly influenced by DmdA clades Gamma HTCC2080 (Pearson's *r* of PC2 loadings versus gene abundance, $p<0.01$) and 'other Gamma' ($p<0.05$) (Fig. 2.3c, Table 2.S1). These plotted with more negative loading values relative to the other groups (Fig. 2.3c) and achieved their highest abundances in the middle of the deployment (Fig. 2.2).

Cleavage Gene Overview

Genes *dddP* and *dddK* dominated the Monterey Bay DMSP cleavage gene pool, being present in 9.8% and 7.2% of cells. Genes identified as *dddQ* and *dddD* were also present but

found in fewer than 1% of cells, and orthologs to the algal DMSP cleavage gene Alma1 were also in low abundance. *dddL*, *dddY*, and *dddW* were not observed. Cleavage gene *dddK* is present in only a subset of SAR11 genomes. Of the fifteen complete SAR11 reference genomes obtained for this study from preexisting databases, all have *dmdA* but only seven have *dddK*. Fourteen *dddK* genes were assembled from reads identified by BLAST analysis of the metagenomic data using SPAdes, and nine genes were obtained from contigs assembled from all metagenome reads using MEGAHIT (Fig. 2.4a). A *dddK* gene was also retrieved from Pelagibacter-like SCGC AG-145-C19 obtained in this study. Greater than 90% of *dddK*s placed to two main clades, one containing all sequences from preexisting databases (including the three with verified function) plus the single-cell genome sequence, and the other with only sequences assembled from this study (and presumed to also represent SAR11 *dmdA*s). Overall, reference sequences assembled from the metagenomes recruited the largest proportion of *dddK* reads (Fig. 2.S1b).

Reads predicted to be DMSP cleavage gene *dddP* placed to both Alphaproteobacteria and Gammaproteobacteria sequences (Fig. 2.4b). Roseobacters recruited the most reads, primarily to the sequences from SAG SCGC AG-151-C16, *Planktomarina temperata*, SB2, HIMB11, and HTCC2255. Metagenomic reads from the SAR116 clade also recruited to *dddP*. Whereas most *dddP* reads that mapped to reference sequences did so to groups of bacteria already known to cleave DMSP, one exception is the reads recruiting to the Gammaproteobacteria member *Candidatus* Thioglobus singularis (SUP05 group) (Fig. 2.4b). The SUP05 group is already known to transform inorganic sulfur (Callbeck et al., 2018; Marshall and Morris, 2015; Shah et al., 2017) and emerges here as a potential contributor to DMSP cleavage. A second exception is the *dddP* reads placing to the "other Alphaproteobacteria" cluster containing reference genomes from the Rhizobiales. Together, these suggest that the full diversity of *dddP* has not yet been

characterized. DddP sequences obtained by the two metagenomic assembly methods were not as successful at recruiting reads as they were for DmdA and DddK (only 19% of reads compared to 59% and 92% for DmdA and DddK; Fig. 2.S1c), possibly because the longer average length of DddP (451 residues compared to 371 and 125 residues for DmdA and DddK) made assembly more difficult.

DMSP Gene Frequency in Bacterioplankton Cells

We calculated the proportion of bacterial cells genetically capable of degrading DMSP on each sample date (although whether they were expressing this capability at the time of sampling is not known). Some bacterial genomes contain two copies of *dmdA*, and therefore reads with phylogenetic placements identifying them as a likely second copy (some members of SAR11 and Roseobacter; Table 2.S2) were not used in the calculation of demethylation-capable cells. Following normalization to the universal single copy bacterial gene *recA* (Howard *et al.*, 2006), an average of 30% of Monterey Bay bacterial cells were estimated to be capable of DMSP demethylation, varying from a maximum of 36% on September 22 to a minimum of 25% on October 8. In the case of cleavage genes, *dddP* and *dddK* do not have overlapping taxonomic ranges, since *dddK* is exclusively found in SAR11 genomes while *dddP* is not in SAR11. Together, they indicate that an average of 17% of Monterey Bay bacterial cells were capable of DMSP cleavage, varying from 19% on October 6 to 16% on October 13 (Fig. 2.S2). Overall, *dmdA* genes were 1.9-fold more abundant in Monterey Bay than cleavage genes (*dddK, dddP, dddQ*, or *dddD*). The relative abundances of the two dominant cleavage genes shifted over the course of the study, with a *dddP:dddK* ratio averaging 0.9:1 in the September samples and 1.8:1 in the October samples (Fig. 2.S2). This change in relative abundance fits with the observed

pattern for *dmdA* showing a shift from SAR11 to Roseobacter genes over time.

Based on an analysis of reference genomes, most bacteria containing *dddP* or *dddK* also possess a *dmdA* copy. There are some exceptions, with a few reference genomes in the Roseobacter, Rhizobiales, and SUP05 Gammaproteobacteria groups having *dddP* without an accompanying *dmdA* (Table 2.S3). Adding the percent of cells from this study that recruit to reference genomes that possess only *dddP* (3%) to the estimated percent of demethylation-capable cells (30%), we calculate that one out of every three bacteria in the Monterey Bay community can use DMSP.

For cases in which a *dmdA* clade and *dddP*/*dddK* clade mapped to the same reference genomes (Figs. 2.2,2.4), both demethylation and cleavage capabilities are anticipated to be present in the same genome and therefore gene dynamics should be correlated through time. We checked this for four sets of demethylation and cleavage genes that mapped to the same references: SAR11 1a.1, SAR116 Clade 1, Roseobacter strain SB2, and HTCC2255. Three of the four had significant linear relationships between relative abundances of *dmdA* and *dddP/K* during the course of the study (Pearson's r, $p<0.05$; Fig. 2.S3)

Relationships to Phytoplankton Community Composition and DMSP Availability

DMSP is an important component of the metabolites released by phytoplankton, although taxa differ in the amount of DMSP they produce. In this study, groups reported to produce high amounts of DMSP under laboratory conditions, including haptophytes (up to 12 pg DMSP/cell) and dinoflagellates (up to 384 pg DMSP/cell) (Keller, 1989), accounted for 31% of the phytoplankton reads identified in the metagenomes (Table 2.1; Fig. 2.1b,c). Diatoms, which range from having undetectable intracellular DMSP to 35 pg DMSP/cell, made up 28% of the

phytoplankton reads. Members of the Mamiellophyceae phylum of picophytoplankton have < 1 pg DMSP/cell in laboratory cultures (Keller, 1989) (Table 2.1). Mamiellophyceae peaked in relative abundance on the final sampling date at 31% of metagenomic reads mapping to phytoplankton, coincident with lower DMSP and chlorophyll concentrations. Because these phytoplankton groups have widely varying genome sizes, from about 20 Mbp for Mamiellophyceae (Worden *et al.*, 2009; Moreau *et al.*, 2012) to over 200 Mbp for some dinoflagellates, read percentages are not a good proxy for cell numbers. Nonetheless, the relative contributions of these high and low DMSP-producing groups shifted during the three-week study.

The maximal information-based nonparametric exploration program (MINE; Reshef *et al.*, 2011) was used to test for associations between specific clades of DMSP genes, the phytoplankton community composition, and environmental parameters (Table 2.S4). MINE analyzes scatterplots of pair-wise datasets to find the grid with the most mutual information (i.e., when knowing one variable provides the most information about the other). Bacteria harboring *dmdA* genes from SAR11 subclades generally had positive relationships with dinoflagellate abundance, and these relationships were significant for SAR11 subclades 1a.1 and "other SAR11" (Fig. 2.5). Some members of the SAR11 subclade 1a.1 also carry *dddK* in their genomes, consistent with the positive correlation also found between relative abundance of cells with *dddK* and dinoflagellate abundance (Fig. 2.5, Table 2.S5). Further, both total DMSP (DMSPt) concentration and the rate of consumption of dissolved DMSP (DMSPd) were positively related to the dinoflagellate inventory (Table 2.S5). Possessing both DMSP degradation pathways may provide an advantage for SAR11 cells when DMSP concentrations are high. Because DMSP demethylation (*dmdA*-mediated) provides reduced sulfur for

biomolecule synthesis whereas DMSP cleavage (*dddK*-mediated) does not, differential routing of DMSP through these two pathways may be kinetically controlled to increase DMSP cleavage once cellular sulfur demand has been met (Kiene *et al*., 2000; Pinhassi *et al*., 2005; Sun, *et al*., 2016). MINE analysis also indicated that the two SAR11 subclades had a significant positive relationship with DMSPd consumption (Table 2.S5), suggesting they may be among the more active DMSP degraders.

Cells carrying SAR11 *dmdA* also varied positively with abundance of haptophytes, as did cells with *dmdA* genes mapping to Gammaproteobacteria genomes (Fig. 2.5). Haptophytes represent another high DMSP-producing phytoplankton group (Keller *et al*., 1989) and had a positive relationship with DMSPd concentration. Going against this pattern were cells with DMSP genes in the Roseobacter group, which generally had negative relationships to dinoflagellate and haptophyte abundances, but positive relationships to pennate diatoms in the Bacillariophyceae and to Mamiellophyceae. Abundance of these roseobacters had no relationship to DMSP concentration or consumption rate (Table 2.S5).

Discussion

Bacterial cells capable of DMSP degradation accounted for 1 out of every 3 cells in the productive coastal system of Monterey Bay, similar to the high frequencies found in other marine environments (Moran *et al*., 2012; Varaljay *et al*., 2012). This implicates DMSP as a highly reliable phytoplankton-derived metabolite available to heterotrophic marine bacterioplankton. Previous studies have estimated that this single compound supports up to 10% of bacterial carbon demand (Archer *et al*., 2001; Howard *et al*., 2006), and plays a central role in carbon transfer between microbial autotrophs and heterotrophs in the surface ocean (Landa *et al*.,

2017).

dmdA genes have consistently been found to be more abundant than cleavage genes in metagenomic analysis [3.8-fold more abundant in the Global Ocean Sampling metagenomic dataset (Moran *et al.*, 2012) and 4.7-fold in Station ALOHA metagenomes (Varaljay *et al*., 2012)]. In this Monterey Bay bacterial community, *dmdA* genes were 1.9-fold more abundant, lower than previous studies as a result of inclusion of *dddK* in the analysis. The abundant DMSP degraders in Monterey Bay recruit to genomes with different suites of DMSP degradation genes. For example, SAR11 member HTCC9022 has *dmdA* and *dddK*, Roseobacter strains LE17, SB2, and HTCC2255 have *dmdA* and *dddP*, and SUP05 clade member *Candidatus* Thioglobus singularis has only *dddP*. Such a diversity of gene repertoires may result from adaptations to environments with DMSP concentrations that are low and stable [e.g., SAR11 subclades more common in the oligotrophic open ocean (such as HTCC711) have only *dmdA*] versus dynamic [e.g., subclades more numerous in dynamic coastal environments (such as SAR11 strain HTCC9022 and subclade 1a.1) have both *dmdA* and *dddK*]. Bacterial taxa lacking *dmdA* may fulfill reduced sulfur quotas through other mechanisms of acquisition yet still take up DMSP for use as an osmolyte (Motard-Côté & Kiene, 2015) or as a predator deterrent (Strom *et al.*, 2003).

The DMSP genes found using reference-free methods (metagenomic assembly and SAGs) allowed more resolution of the DMSP-degrading bacteria by capturing local gene sequences not represented in existing databases. Of the metagenomic reads that were identified as *dmdA, dddK,* and *dddP* from the full 12-sample metagenomic dataset, 59%, 92%, and 32%, respectively, mapped with highest identity to a reference gene from a metagenomic or single-cell assembly from this study. The MEGAHIT pipeline found a predicted novel gammaproteobacterial *dmdA* highly divergent from the only previously known

gammaproteobacterial *dmdA* known. The two metagenome assembly pipelines yielded different genes, with overlap between the SPAdes and MEGAHIT pipelines accounting for only 3%, 10%, and 0% of the *dmdA, dddK,* and *dddP* reads. While the SPAdes assembly method yielded *dmdA* sequences that were nearly identical to the two SAGs from this study (Fig. 2.2), the performance of single-cell sequencing for untargeted recovery of specific functional genes is likely to improve as gene screening methodologies advance (Yu *et al.*, 2017).

SAR11 cells dominated the DMSP degrading community throughout the study despite variations over time in phytoplankton composition and DMSP concentration, although subgroup abundance varied (Fig. 2.2). The positive relationship of the SAR11 groups harboring both *dmdA* and *dddK* with dinoflagellate abundance and DMSP concentration suggest the possibility that possessing both pathways is advantageous when DMSP concentrations are high and dynamic. Roseobacters, on the other hand, were correlated with diatom abundance and lower DMSP concentrations. Roseobacters typically reach their highest percent abundance in marine bacterial communities during bloom conditions (Buchan *et al*., 2014), but the fall bloom typical of this system did not occur during the study and chlorophyll *a* levels only reached 7 ug L^{-1} compared to 27 ug L^{-1} in a previous fall season (Varaljay *et al.*, 2015). Gammaproteobacteria capable of DMSP degradation had some of the strongest relationships with haptophytes. Based on available genomic data, DMSP-degrading Gammaproteobacteria can either demethylate (HTCC2080-like taxa) or cleave (Thioglobus-like taxa) but not both. Thus, unlike the Roseobacter and SAR11 cells that harbor genetic capabilities for both pathways, regulatory control of DMSP routing at the cellular level does not appear to occur in these Gammaproteobacteria.

This fine-scale dissection of predicted genes mediating the fate of DMSP in Monterey Bay provides a nuanced view of the ecology of DMSP metabolism in marine surface waters,

including relationships between phytoplankton sources and bacterial sinks, as well as correlations with environmental patterns. As many as 15 distinguishable clades of *dmdA* appeared during the three- week time period, as the community shifted from peak numbers of centric diatoms(Coscinodiscophyceae) and dinoflagellates to peak numbers of pennate diatoms (Bacillariophyceae) and Mamiellophyceae. Some of these suggest new *dmdA* clades not known previously and found only through assembling metagenomic data and SAG sequencing. Similarly, a new clade of predicted *dddK* sequences was assembled that was equal in abundance to the clade containing all previously identified *dddK* sequences (those from NCBI RefSeq; Fig. 2.4a). For *dddP*, nine major clades appeared during the deployment. This diversity of bacterial groups having the genetic capability of transforming DMSP in a coastal ecosystem is striking, as is the finding that their population dynamics track with various DMSP-producing phytoplankton. Taken together, these results suggest that taxon-specific bacteria-phytoplankton interactions could play important roles in the fate of DMSP-derived carbon and sulfur in the coastal ocean.

Methods

Sample Collection

The ESP was deployed near Monterey Bay Station M0 (36.835 N, 121.901 W; water depth ~76 m) at a depth of ~6 m from September 18 to October 15, 2014. The instrument filtered up to 1 L of seawater through a 25 mm, 0.2 μm pore-size polyethersulfone filter to capture microbial plankton during 118 sampling events, with four sequential replicate samples taken each day. During filtration, pressure was maintained across the membrane between 25 and 28 psi. Nucleic acids were preserved onboard the ESP. First, seawater was evacuated from filters followed twice with a 2 min incubation with 1 ml of RNA*later*™. RNAlater was evacuated, and

filters were stored in the ESP until they were transferred to -80°C upon instrument recovery. Duplicate filters from six dates were chosen for metagenomic sequencing to coincide with the dates when environmental parameter sampling showed variability in chlorophyll *a* and total seawater DMSP concentrations.

For the environmental sampling, 10 L grab samples of seawater were taken at the ESP mooring at the same depth (6 m) using an SBE 19plus SEACAT CTD (Sea-Bird) and 5.0 L Niskin bottles, timed to coincide with the ESP sampling. Water was transferred to low-density polyethylene cubitainers and stored at *in situ* temperature until returned to shore. Seawater was processed within 80 min of collection for triplicate measurements of chlorophyll *a*, total DMSP concentration (DMSPt), and dissolved DMSP (DMSPd) consumption rate. Samples for chlorophyll *a* were collected by vacuum filtration of 200 mL of seawater onto Whatman GF/F filters and extracted in 90% acetone at -20°C prior to fluorometric quantification (Pennington and Chavez, 2000).

DMSPt concentrations

Upon return to the laboratory, the cubitainer of water was gently mixed by inversion and three replicate 10 ml sub-samples were removed by pipette into individual 15 ml centrifuge tubes (Corning, polypropylene). The samples were immediately acidified with 0.3 ml of 50% concentrated HCl (1.5% final concentration of concentrated HCl) to preserve total DMSP (dissolved plus particulate). These DMSPt samples were closed tightly and stored until analysis (described below) which took place within three months of collection.

DMSPd consumption

To measure the consumption rate of dissolved DMSP, we used the glycine betaine (GBT) inhibition technique (Kiene & Gerard, 1995; Li et al., 2016). Immediately upon return to the laboratory, six 250 ml black Teflon bottles were filled with seawater from the gently-mixed cubitainer. Three of the bottles were treated with 25 µl of a 100 mM GBT anhydrous reagent (Sigma) solution (10 µM final GBT concentration), and three were left untreated as controls. Bottles were incubated in seawater maintained within 1°C of the *in situ* temperature. Immediately after GBT addition, the first time point was collected by simultaneously filtering ~50 ml sub-samples from each bottle through 47 mm Whatman GF/F filters using the small volume gravity drip filtration protocol of Kiene and Slezak (2006). The first 3.5 ml of filtrate from each sample was collected into 15 ml centrifuge tubes (Corning, polypropylene) that contained 100 µl of 50% HCl to immediately preserve any DMSP passing through the GF/F filter, which is defined as dissolved DMSP (DMSPd). Additional time points from each bottle were collected at 3 and 6 h. The rate of change of DMSPd in no-treatment bottles was subtracted from the rate of change in the +GBT bottles to obtain an estimate of DMSPd consumption rate (Kiene and Gerard, 1995).

DMSP Analysis

DMSP was quantified by proxy as the amount of DMS released from samples after alkaline cleavage (White, 1982). For DMSPt, 0.05 to 0.5 ml of each preserved sample was pipetted into a 14 ml glass serum vial, with the volume being adjusted based on the concentration of DMSPt in the sample. For DMSPd, the volume pipetted was 1.0 to 3.0 ml. Each serum vial was treated with 1 ml of 5 M NaOH and capped with a Teflon-faced serum stopper (Wheaton).

After 1 h, the amount of DMS in each vial was quantified by purge and trap gas chromatography with flame photometric detection. Briefly, each vial was attached to the purge system and a flow of helium (~95 ml min^{-1}) allowed bubbling of the solution. An excurrent needle led to a Nafion dryer and six-port valve (Valco). The DMS in the samples was cryotrapped in a Teflon tubing loop immersed in liquid nitrogen. After a 4 min sparge, during which >99% of the DMS in the samples was removed, hot water replaced the liquid nitrogen to introduce the DMS into the Shimadzu GC-2014 gas chromatograph. Separation of the sulfur gases was achieved with a Chromosil 330 column (Supelco; Sigma) maintained at 60ºC with a helium carrier flow of 25 ml min^{-1}. The flame photometric detector was operated in sulfur mode and maintained at 175ºC. Minimum detection limits during this study were 0.5 to 1 pmol DMS per sample with minimum detectable concentrations ranging from 0.17 to 10 nM, depending on the volume analyzed. The GC-FPD system was calibrated with a gas stream containing known amounts of DMS from a permeation system.

DNA Extraction and Sequencing

DNA was extracted using the phenol-chloroform extraction method of Crump et al. (2003) after placing the filters into 1 ml of DNA extraction buffer. Extracted DNA was sheared ultrasonically to ~350 bp fragments, and library preparation was performed at the Georgia Genomics and Bioinformatics Core (GGBC) facility. Single-end 250 bp sequencing was performed using an Illumina HiSeq Rapid Run at HudsonAlpha Genomic Services Laboratory (Huntsville, AL, USA). Metagenomic data are deposited in the NCBI SRA under project number PRJNA505827.

Single Amplified Genomes (SAGs)

On three dates corresponding to the metagenomic sequencing (Sept. 22, Sept. 29, Oct. 8), a 1 ml aliquot of seawater from the grab sample was preserved with 100 µl glyTE stock (20 ml 100 x TE pH 8.0; 60 ml deionized water; 100 ml molecular-grade glycerol), flash frozen in liquid nitrogen upon return to shore, and stored at -80°C. Samples were later processed and sequenced at the Bigelow Single Cell Genomics Center (East Boothbay, ME, USA) (Stepanauskas and Sieracki, 2007). DNA amplified from single cells was screened by 16S rRNA sequencing, and six cells with strong DNA amplification and representing taxa whose members are known to degrade DMSP were selected for sequencing; these included three Roseobacter and three SAR11 cells. Single cell genome data are deposited in NCBI SRA under project number PRJNA505827.

Bioinformatic and Statistical Analyses

Metagenomic sequencing reads with a quality score of $\geq$ 20 over at least 80% of the read length were retained for analysis. For taxonomic characterization of eukaryotes, reads were annotated using RAPSearch2 (Zhao *et al*., 2012) against a custom sequence database containing representative genomes and transcriptomes of major marine eukaryotes and encompassing much of the known diversity of surface layer marine microbes (MarineRefII, http://roseobase.org/data/). Relative abundances of major phytoplankton groups were calculated. For taxonomic characterization of bacteria and archaea, all quality-filtered reads were classified using Kaiju (Menzel *et al*., 2016) against the NCBI Reference Sequence Database (RefSeq), and relative abundances of bacterial and archaeal groups were calculated.

To identify DMSP genes, metagenomic reads were queried using BLASTx against a custom database containing reference sequences for each DMSP degradation gene, comprised of

genes with experimentally verified function whenever possible (Table 2.S6). The gene databases also included sequences of the closest paralogs to the genes of interest in order to reduce false positives. Reads hitting orthologs with a bit score ≥ 40 were kept. These were analyzed further by a BLASTx search against all bacterial and archaeal reference proteins in RefSeq followed by manual annotation to determine a final cut-off above which mostly orthologs were present. RefSeq bit- score cutoffs used were 71 for *dmdA*, 85 for *dddP*, and 79 for *dddK*, and reads above these cutoffs were retained. The percentage of cells containing a particular DMSP gene was calculated as (# homologs x 100)/# of single-copy gene *recA* hits following normalization to gene size (Howard *et al.*, 2006). To quantify the abundance of *recA* hits, each metagenome was queried using BLASTx against the *Escherichia coli* K-12 *recA* gene, and reads hitting with a bit score of ≥ 40 were retained. Kept reads were queried using Diamond (Buchfink *et al*., 2015) against RefSeq to verify that top hits were to RecA.

To find Monterey Bay-specific DMSP degradation genes without close reference sequences in existing databases, three different assembly approaches were used: 1) all reads from the 12 metagenomic libraries were co-assembled using MEGAHIT (Li *et al*., 2015), and the assembled contigs were searched for *dmdA*-, *dddK*, and *dddP*-like sequences; 2) the subset of reads in the metagenomic libraries previously annotated as *dmdA, dddP,* or *dddK* using the pipeline described above were co-assembled using SPAdes (Bankevich *et al.*, 2012); and 3) six genomes assembled from the single cell analysis were searched for DMSP gene sequences. Prodigal (Hyatt *et al.*, 2010) was used to find open reading frames (ORFs) in the contigs from all three approaches, and predicted DMSP genes were identified using the custom gene database approach as described above. MEGAHIT was used for the co-assembly of the 12 metagenomes due to lower memory requirements than SPAdes.

For phylogenetic analysis of DMSP demethylating taxa from this study, the translated DMSP sequences were aligned using Clustal Omega (Sievers and Higgins, 2014) along with the translated reference sequences representing the best hits to the metagenomic reads. Gblocks (Castresana, 2000) was used to trim the alignment, ProtTest (Darriba *et al*., 2011) selected the best-fit model of protein evolution, and raxmlGUI (Silvestro and Michalak, 2012) was used to construct the phylogenetic trees. The DMSP reads identified in this study were mapped onto trees using pplacer (Matsen *et al*., 2010). In the case of the more abundant *dmdA* reads, edge principal component analysis (ePCA; Matsen IV and Evans, 2013) was used to assess the importance of reads from different DmdA lineages in driving the compositional differences between communities. Briefly, this program applies principal component analysis to a data matrix generated based on read placement on a phylogenetic tree. The resulting tree has edges whose thickness reflects the number of placed reads, and variation in read placement between samples is captured on the principal component axes (Fig. 2.3). Read placement to the right of orange edges in Fig. 2.3 drives samples in a positive direction along the axes, while read placement to the right of green edges moves samples in a negative direction along the axes.

To examine relationships between DMSP gene abundances, phytoplankton abundances, and environmental parameters, the Maximal Information-based Non-parametric Exploration program (MINE; Reshef *et al*., 2011) was used with the '-equitability' parameter. Input data consisted of abundance of DMSP genes by phylogenetic clade assignment, phytoplankton relative abundance, and environmental parameters obtained as described above (Table 2.S4). MINE finds both linear and non-linear associations and assigns a normalized maximum information coefficient (MIC) value from 0 (no relationship) to 1 (strongest relationship). In our analysis, we kept pairwise relationships with MIC > 0.7, along with any statistically significant

linear relationships (Pearson's r, $p < 0.05$) (Table 2.S5).

Acknowledgements

We thank R. Marin III, A. Burns, C. Smith, J. Figurski, C. Wahl, B. Kieft, and T. Pennington for sampling assistance, advice, and technical expertise; the Captain and crew of the R/V Carson and R/V Paragon; the Moss Landing Marine Laboratories Small Boat Facility; and S. Sharma for bioinformatic assistance. This work was funded by NSF grants OCE-1342694 and OCE-1342699. This publication is dedicated to the memory of our friend and colleague, Ronald P. Kiene.

References

Andreae, M.O. (1990) Ocean-atmosphere interactions in the global biogeochemical sulfur cycle. *Mar. Chem.* **30**: 1–29.

Archer, S.D., Widdicombe, C.E., Tarran, G.A., Rees, A.P., and Burkill, P.H. (2001) Production and turnover of particulate dimethylsulphoniopropionate during a coccolithophore bloom in the northern North Sea. *Aquat. Microb. Ecol.* **24**: 225–241.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**: 455–77.

Billerbeck, S., Wemheuer, B., Voget, S., Poehlein, A., Giebel, H.-A., Brinkhoff, T., et al. (2016) Biogeography and environmental genomics of CHAB-I-5, a pelagic lineage of the marine Roseobacter clade. *Nat. Microbiol.* 1–8.

Brown, M. V, Lauro, F.M., Demaere, M.Z., Muir, L., Wilkins, D., Thomas, T., et al. (2012) Global biogeography of SAR11 marine bacteria. *Mol. Syst. Biol.* **8**: 595.

Buchan, A., LeCleir, G.R., Gulvik, C.A., and González, J.M. (2014) Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat. Rev. Microbiol.* **12**: 686–698.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**: 59–60.

Callbeck, C.M., Lavik, G., Ferdelman, T.G., Fuchs, B., Gruber-Vodicka, H.R., Hach, P.F., et al. (2018) Oxygen minimum zone cryptic sulfur cycling sustained by offshore transport of key sulfur oxidizing bacteria. *Nat. Commun.* **9**: 1729.

Castresana, J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**: 540–552.

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

Cho, J.C. and Giovannoni, S.J. (2004) Cultivation and growth characteristics of a diverse group of oligotrophic marine gammaproteobacteria. *Appl. Environ. Microbiol.* **70**: 432–440.

Crump, B.C., Kling, G.W., Bahr, M., and Hobbie, J.E. (2003) Bacterioplankton community shifts in an arctic lake correlate with seasonal changes in organic matter source. *Appl. Environ. Microbiol.* **69**: 2253–68.

Curson, A.R.J., Rogers, R., Todd, J.D., Brearley, C.A., and Johnston, A.W.B. (2008) Molecular genetic analysis of a dimethylsulfoniopropionate lyase that liberates the climate-changing gas dimethylsulfide in several marine α-proteobacteria and *Rhodobacter sphaeroides*. *Environ. Microbiol.* **10**: 757–767.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2011) ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**: 1164–1165.

Giovannoni, S.J. (2017) SAR11 bacteria: the most abundant plankton in the oceans. *Ann. Rev. Mar. Sci.* **9**: 231–255.

Howard, E.C., Henriksen, J.R., Buchan, A., Reisch, C.R., Bürgmann, H., Welsh, R., et al. (2006) Bacterial taxa that limit sulfur flux from the ocean. *Science* **314**: 649–652.

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

Keller, M.D. (1989) Dimethyl sulfide production and marine phytoplankton: the importance of species composition and cell size. *Biol. Oceanogr.* **6**: 375–382.

Keller, M.D., Bellows, W.K., and Gulliard, R.L. (1989) Dimethyl sulfide production in marine phytoplankton. *Am. Chem. Soc.* **81**: 168–182.

Kiene, R.P. and Gerard, G. (1995) Evaluation of glycine betaine as an inhibitor of dissolved dimethylsulfoniopropionate degradation in coastal waters. *Mar. Ecol. Prog. Ser.* **128**: 121– 131.

Kiene, R.P., Linn, L.J., and Bruton, J.A. (2000) New and important roles for DMSP in marine microbial communities. *J. Sea Res.* **43**: 209–224.

Kiene, R.P. and Slezak, D. (2006) Low dissolved DMSP concentrations in seawater revealed by small volume gravity filtration and dialysis sampling. *Limnol. Oceanogr. Methods* **4**: 80–95.

Landa, M., Burns, A.S., Roth, S.J., and Moran, M.A. (2017) Bacterial transcriptome remodeling during sequential co-culture with a marine dinoflagellate and diatom. *ISME J.* **11**: 2677– 2690.

Li, C., Yang, G.P., Kieber, D.J., Motard-Côté, J., and Kiene, R.P. (2016) Assessment of DMSP turnover reveals a non-bioavailable pool of dissolved DMSP in coastal waters of the Gulf of Mexico. *Environ. Chem.* **13**: 266–279.

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

Marshall, K.T. and Morris, R.M. (2015) Genome sequence of "*Candidatus Thioglobus singularis*" strain PS1, a mixotroph from the SUP05 clade of marine Gammaproteobacteria. *Genome Announc.* **3**: e01155-15.

Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**: 538.

Matsen IV, F.A. and Evans, S.N. (2013) Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS One* **8**: e56859.

Menzel, P., Ng, K.L., and Krogh, A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* **7**: 11257.

Moran, M.A., Reisch, C.R., Kiene, R.P., and Whitman, W.B. (2012) Genomic insights into bacterial DMSP transformations. *Ann. Rev. Mar. Sci.* **4**: 523–542.

Moreau, H., Verhelst, B., Couloux, A., Derelle, E., Rombauts, S., Grimsley, N., et al. (2012) Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol.* **13**: R74.

Motard-Côté, J. and Kiene, R.P. (2015) Osmoprotective role of dimethylsulfoniopropionate (DMSP) for estuarine bacterioplankton. *Aquat. Microb. Ecol.* **76**: 133–147.

Oh, H.M., Kwon, K.K., Kang, I., Kang, S.G., Lee, J.H., Kim, S.J., and Cho, J.C. (2010) Complete genome sequence of "*Candidatus* Puniceispirillum marinum"; IMCC1322, a representative of the SAR116 clade in the Alphaproteobacteria. *J. Bacteriol.* **192**: 3240–1.

Ottesen, E.A., Marin, R., Preston, C.M., Young, C.R., Ryan, J.P., Scholin, C.A., and Delong, E.F. (2011) Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J.* **5**: 1881–1895.

Peng, M., Xie, Q., Hu, H., Hong, K., Todd, J.D., Johnston, A.W.B., and Li, Y. (2012) Phylogenetic diversity of the *dddP* gene for dimethylsulfoniopropionate-dependent dimethyl sulfide synthesis in mangrove soils. *Can. J. Microbiol.* **58**: 523–530.

Pennington, T.J. and Chavez, F.P. (2000) Seasonal fluctuations of temperature, salinity, nitrate, chlorophyll and primary production at station H3/M1 over 1989–1996 in Monterey Bay, California. *Deep Sea Res. Part II Top. Stud. Oceanogr.* **47**: 947–973.

Pinhassi, J., Simó, R., González, J.M., Vila, M., Alonso-Sáez, L., Kiene, R.P., et al. (2005) Dimethylsulfoniopropionate turnover is linked to the composition and dynamics of the bacterioplankton assemblage during a microcosm phytoplankton bloom. *Appl. Environ. Microbiol.* **71**: 7650–7660.

Quinn, P.K. and Bates, T.S. (2011) The case against climate regulation via oceanic phytoplankton sulphur emissions. *Nature* **480**: 51–56.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., et al. (2011) Detecting novel associations in large data sets. *Science* **334**: 1518–24.

Shah, V., Chang, B.X., and Morris, R.M. (2017) Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. *ISME J.* **11**: 263– 271.

Sievers, F. and Higgins, D.G. (2014) Clustal Omega. *Curr. Protoc. Bioinformatics* **48**: 3.13.1-3.13.16.

Silvestro, D. and Michalak, I. (2012) raxmlGUI: a graphical front-end for RAxML. *Org. Divers. Evol.* **12**: 335–337.

Stepanauskas, R. and Sieracki, M.E. (2007) Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci.* **104**: 9052–9057.

Strom, S., Wolfe, G., Slajer, A., Lambert, S., and Clough, J. (2003) Chemical defense in the microplankton II: Inhibition of protist feeding by ß-dimethylsulfoniopropionate (DMSP). *Limnol. Oceanogr.* **48**: 230–237.

Sun, J., Todd, J.D., Thrash, J.C., Qian, Y., Qian, M.C., Guo, J., et al. (2016) The abundant marine bacterium *Pelagibacter* simultaneously catabolizes dimethylsulfoniopropionate to the gases dimethyl sulfide and methanethiol. *Nat. Microbiol.* **11**: 1–24.

Todd, J.D., Curson, A.R.J., Dupont, C.L., Nicholson, P., and Johnston, A.W.B. (2009) The *dddP* gene, encoding a novel enzyme that converts dimethylsulfoniopropionate into dimethyl sulfide,

is widespread in ocean metagenomes and marine bacteria and also occurs in some Ascomycete fungi. *Environ. Microbiol.* **11**: 1376–1385.

Todd, J.D., Curson, A.R.J., Kirkwood, M., Sullivan, M.J., Green, R.T., and Johnston, A.W.B. (2011) DddQ, a novel, cupin-containing, dimethylsulfoniopropionate lyase in marine roseobacters and in uncultured marine bacteria. *Environ. Microbiol.* **13**: 427–438.

Todd, J.D., Kirkwood, M., Newton-Payne, S., and Johnston, A.W.B. (2012) DddW, a third DMSP lyase in a model Roseobacter marine bacterium, *Ruegeria pomeroyi* DSS-3. *ISME J.* **6**: 223– 226.

Varaljay, V.A., Gifford, S.M., Wilson, S.T., Sharma, S., Karl, D.M., and Moran, M.A. (2012) Bacterial dimethylsulfoniopropionate degradation genes in the oligotrophic North Pacific subtropical gyre. *Appl. Environ. Microbiol.* **78**: 2775–2782.

Varaljay, V.A., Robidart, J., Preston, C.M., Gifford, S.M., Durham, B.P., Burns, A.S., et al. (2015)
Single-taxon field measurements of bacterial gene regulation controlling DMSP fate. *ISME J.* doi: 10.1038/ismej.2015.23.

White, R. (1982) Analysis of dimethyl sulfonium compounds in marine algae. *J Mar Res* **40**: 529– 536.

Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., et al. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science* **324**: 268–72.

Yu, F.B., Blainey, P.C., Schulz, F., Woyke, T., Horowitz, M.A., and Quake, S.R. (2017) Microfluidic- based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. *eLife* **6**: e26580.

Zhao, Y., Tang, H., and Ye, Y. (2012) RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**: 125–126.

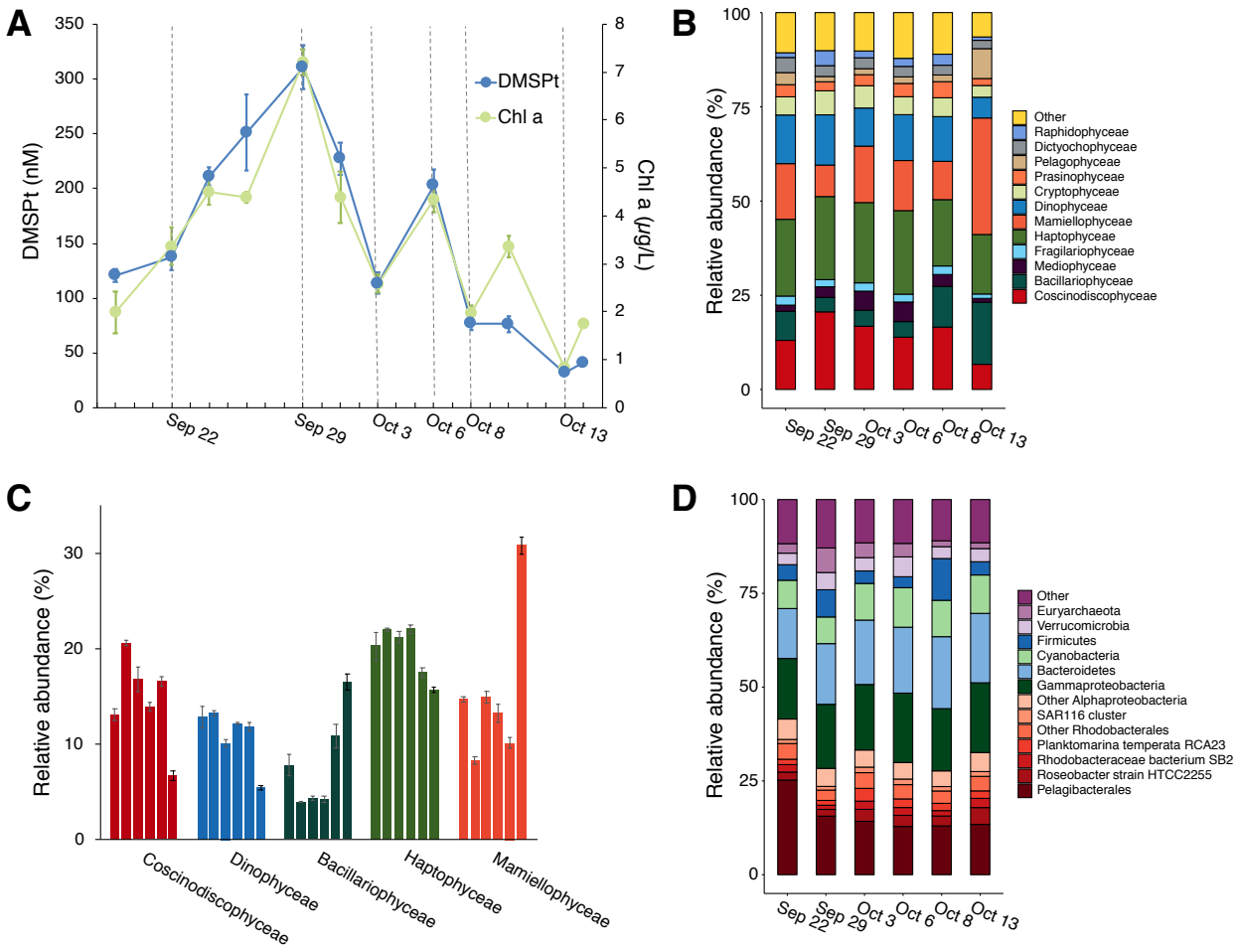


Figure 2.1. a) DMSP and chlorophyll *a* dynamics during the ESP deployment. Dashed vertical lines indicate dates selected for metagenomic sequencing. b) Overview of phytoplankton community composition. Metagenomic reads were aligned to a custom database of marine microbes and relative abundance of total phytoplankton reads by taxonomic class was calculated. c) Relative abundance of total phytoplankton reads of the five most abundant classes over the six sample dates. d) Bacterial and archaeal community composition.

Figure 2.2. Diversity and abundance of genes predicted to be DMSP demethylation gene *dmdA*. Bar graphs display the percent of bacterial cells (*recA*-normalized) in Monterey Bay possessing the highlighted gene clade. Colors of taxon labels indicate NCBI RefSeq sequences (black), sequences obtained from MEGAHIT co-assembly of all metagenomes (green), sequences obtained from SPAdes assembly of all metagenomic reads identified as *dmdA* (blue), and sequences from SAGs collected at the ESP (orange). Black circles indicate nodes with bootstrap values $\geq$50%. Red asterisks indicate functionally verified proteins.

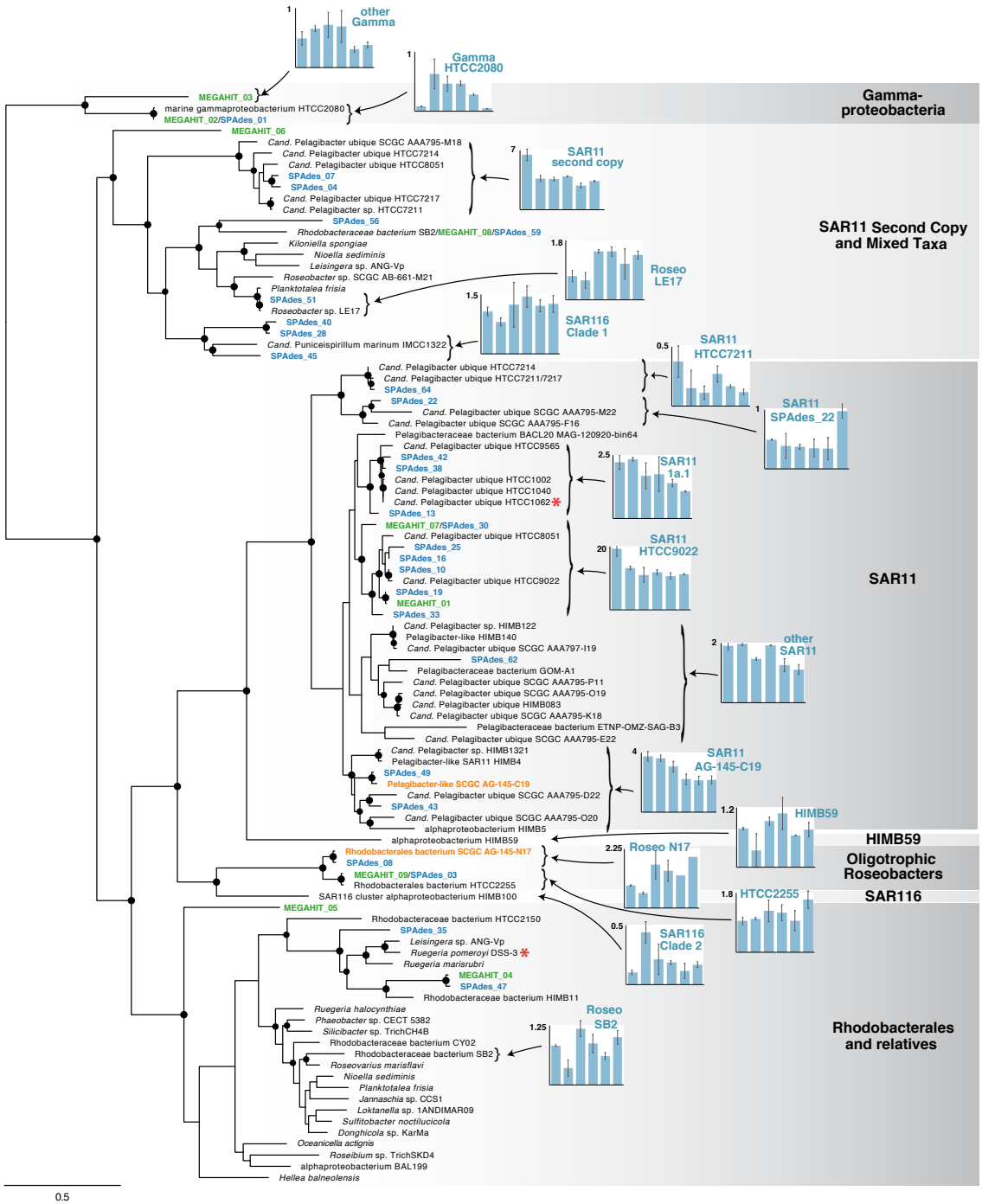

other Gamma
Gamma HTCC2080
MEGAHIT_03
marine gammaproteobacterium HTCC2080
MEGAHIT_02/SPAdes_01
MEGAHIT_06
Gamma-proteobacteria
Cand. Pelagibacter ubique SCGC AAA795-M18
Cand. Pelagibacter ubique HTCC7214
Cand. Pelagibacter ubique HTCC8051
SPAdes_07
SPAdes_04
Cand. Pelagibacter ubique HTCC7217
Cand. Pelagibacter sp. HTCC7211
SPAdes_56
SAR11 second copy
Rhodobacteraceae bacterium SB2/MEGAHIT_08/SPAdes_59
Kiloniella spongiae
Nioella sediminis
Leisingera sp. ANG-Vp
Roseobacter sp. SCGC AB-661-M21
Planktotalea frisia
SPAdes_51
Roseobacter sp. LE17
Roseo LE17
SPAdes_40
SPAdes_28
Cand. Puniceispirillum marinum IMCC1322
SPAdes_45
SAR116 Clade 1
SAR11 Second Copy and Mixed Taxa
Cand. Pelagibacter ubique HTCC7214
Cand. Pelagibacter ubique HTCC7211/7217
SPAdes_64
SPAdes_22
Cand. Pelagibacter ubique SCGC AAA795-M22
Cand. Pelagibacter ubique SCGC AAA795-F16
SAR11 HTCC7211
SAR11 SPAdes_22
Pelagibacteraceae bacterium BACL20 MAG-120920-bin64
Cand. Pelagibacter ubique HTCC9565
SPAdes_42
SPAdes_38
Cand. Pelagibacter ubique HTCC1002
Cand. Pelagibacter ubique HTCC1040
Cand. Pelagibacter ubique HTCC1062
SPAdes_13
SAR11 1a.1
MEGAHIT_07/SPAdes_30
Cand. Pelagibacter ubique HTCC8051
SPAdes_25
SPAdes_16
SPAdes_10
Cand. Pelagibacter ubique HTCC9022
SPAdes_19
MEGAHIT_01
SPAdes_33
SAR11 HTCC9022
SAR11
Cand. Pelagibacter sp. HIMB122
Pelagibacter-like HIMB140
Cand. Pelagibacter ubique SCGC AAA797-I19
SPAdes_62
Pelagibacteraceae bacterium GOM-A1
Cand. Pelagibacter ubique SCGC AAA795-P11
Cand. Pelagibacter ubique SCGC AAA795-O19
Cand. Pelagibacter ubique HIMB083
Cand. Pelagibacter ubique SCGC AAA795-K18
Pelagibacteraceae bacterium ETNP-OMZ-SAG-B3
Cand. Pelagibacter ubique SCGC AAA795-E22
other SAR11
Cand. Pelagibacter sp. HIMB1321
Pelagibacter-like SAR11 HIMB4
SPAdes_49
Pelagibacter-like SCGC AG-145-C19
Cand. Pelagibacter ubique SCGC AAA795-D22
SPAdes_43
Cand. Pelagibacter ubique SCGC AAA795-O20
SAR11 AG-145-C19
alphaproteobacterium HIMB5
alphaproteobacterium HIMB59
HIMB59
HIMB59
Rhodobacterales bacterium SCGC AG-145-N17
SPAdes_08
MEGAHIT_09/SPAdes_03
Rhodobacterales bacterium HTCC2255
SAR116 cluster alphaproteobacterium HIMB100
Roseo N17
Oligotrophic Roseobacters
HTCC2255
SAR116
MEGAHIT_05
Rhodobacteraceae bacterium HTCC2150
SPAdes_35
Leisingera sp. ANG-Vp
Ruegeria pomeroyi DSS-3
Ruegeria marisrubri
MEGAHIT_04
SPAdes_47
Rhodobacteraceae bacterium HIMB11
SAR116 Clade 2
Ruegeria halocynthiae
Phaeobacter sp. CECT 5382
Silicibacter sp. TrichCH4B
Rhodobacteraceae bacterium CY02
Rhodobacteraceae bacterium SB2
Roseovarius marisflavi
Roseo SB2
Nioella sediminis
Planktotalea frisia
Jannaschia sp. CCS1
Loktanella sp. 1ANDIMAR09
Sulfitobacter noctilucicola
Donghicola sp. KarMa
Oceanicella actignis
Roseibium sp. TrichSKD4
alphaproteobacterium BAL199
Hellea balneolensis
Rhodobacterales and relatives
0.5

Figure 2.3. Edge principal components analysis (edge PCA) of *dmdA* read recruitment to a DmdA phylogenetic tree. (a) Principal components analysis clustering by sample. (b) The first principal component, representing 77% of the variance in *dmdA* read placement to the tree. Orange edges contribute to positive location of the sample on the principal component axis, while green edges contribute to the negative direction. The thickness of the edge is proportional to the magnitude of read placement below the edge driving between-sample heterogeneity. (c) The second principal component, representing 10% of the variance in *dmdA* read placement to the tree.

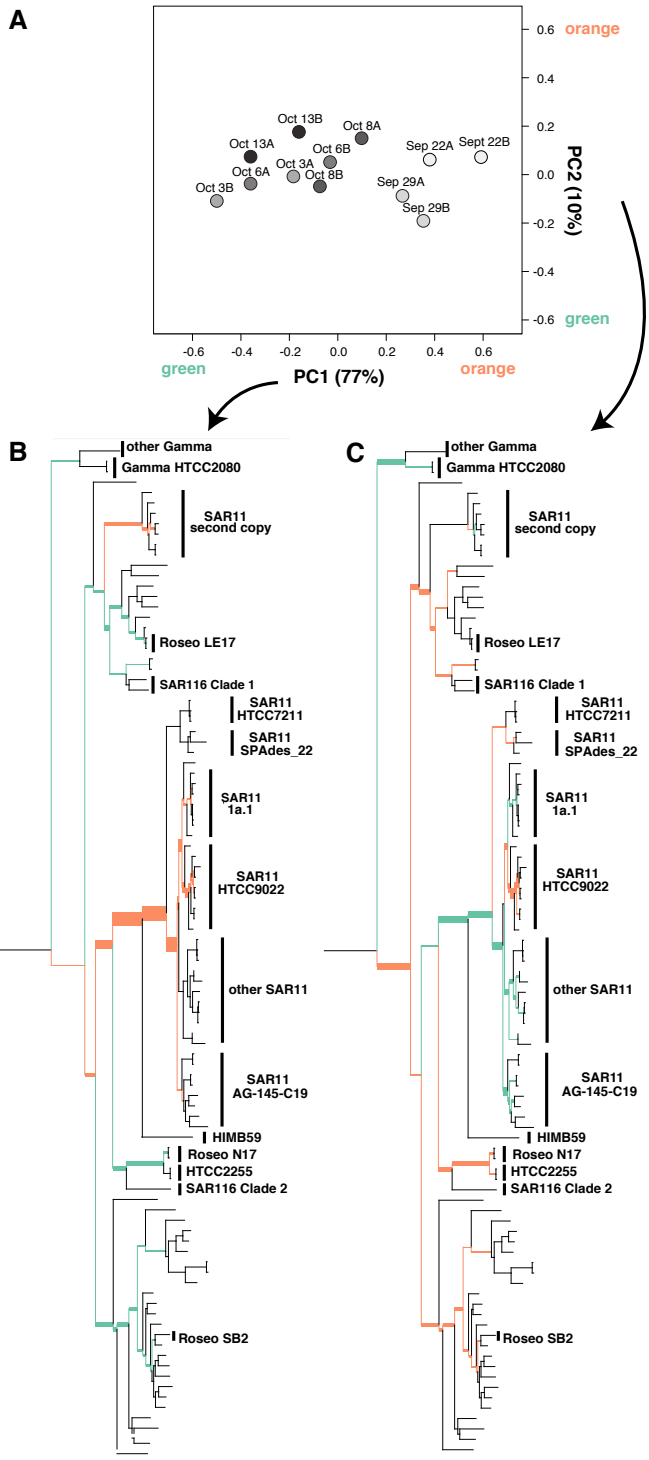

A
Oct 13B
Oct 8A
Oct 13A
Oct 6B
Sep 22A
Sept 22B
Oct 3A
Oct 6A
Oct 8B
Oct 3B
Sep 29A
Sep 29B
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
orange
green
PC2 (10%)
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
green
PC1 (77%)
orange
B
other Gamma
Gamma HTCC2080
SAR11 second copy
Roseo LE17
SAR116 Clade 1
SAR11 HTCC7211
SAR11 SPAdes_22
SAR11 1a.1
SAR11 HTCC9022
other SAR11
SAR11 AG-145-C19
HIMB59
Roseo N17
HTCC2255
SAR116 Clade 2
Roseo SB2
C
other Gamma
Gamma HTCC2080
SAR11 second copy
Roseo LE17
SAR116 Clade 1
SAR11 HTCC7211
SAR11 SPAdes_22
SAR11 1a.1
SAR11 HTCC9022
other SAR11
SAR11 AG-145-C19
HIMB59
Roseo N17
HTCC2255
SAR116 Clade 2
Roseo SB2

Figure 2.4. Diversity and abundance of DMSP cleavage genes *dddK* (a) and *dddP* (b). Bar graphs display percent of bacterial cells (*recA*-normalized) in Monterey Bay possessing the highlighted gene clade. Colors of taxon labels indicate NCBI RefSeq sequences (black), sequences obtained from MEGAHIT co-assembly of all metagenomes (green), sequences obtained from SPAdes assembly of all metagenomic reads identified as *dmdA* (blue), and sequences from SAGs collected at the ESP (orange). Black circles indicate nodes with bootstrap values ≥50%. Red asterisks indicate functionally verified proteins.

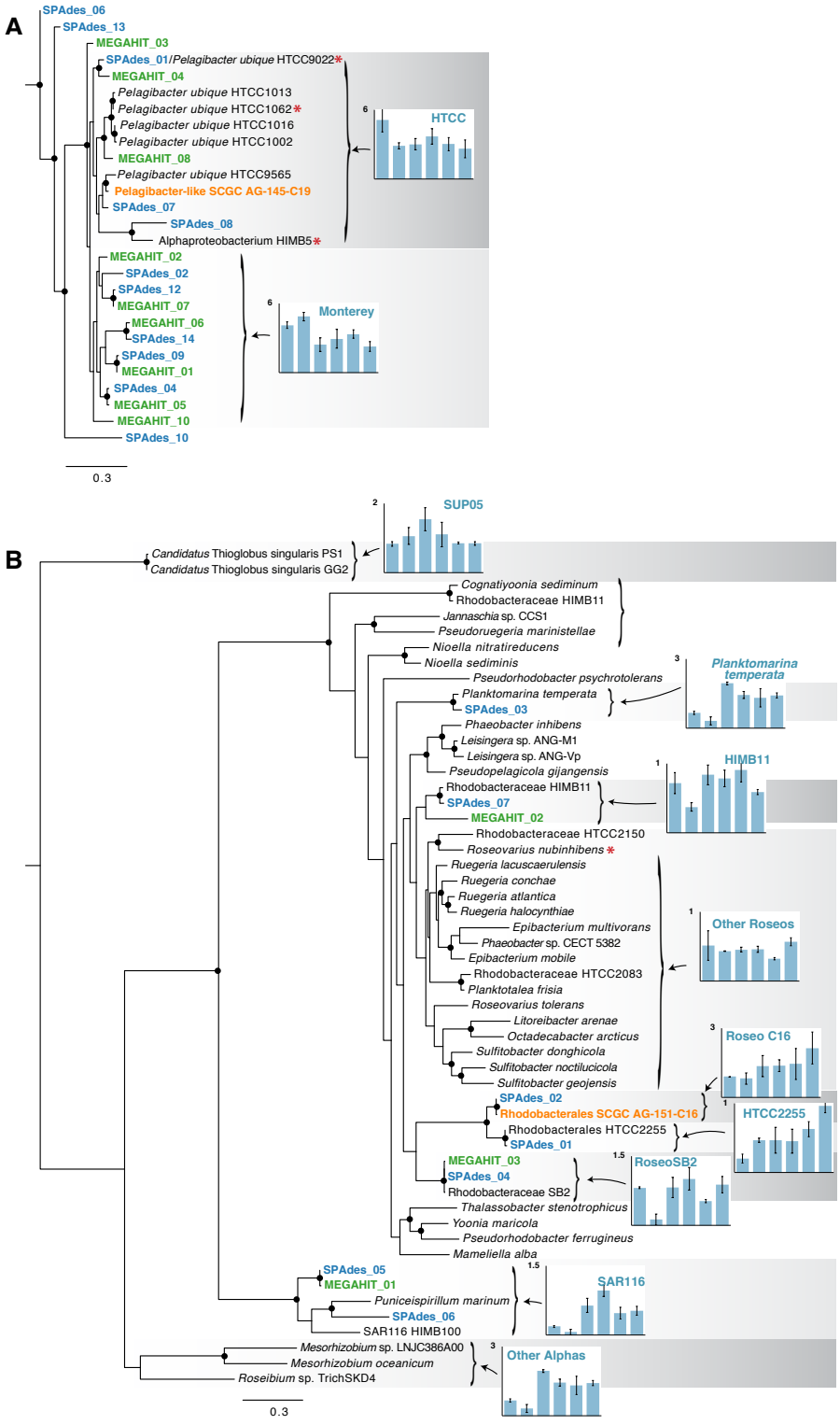
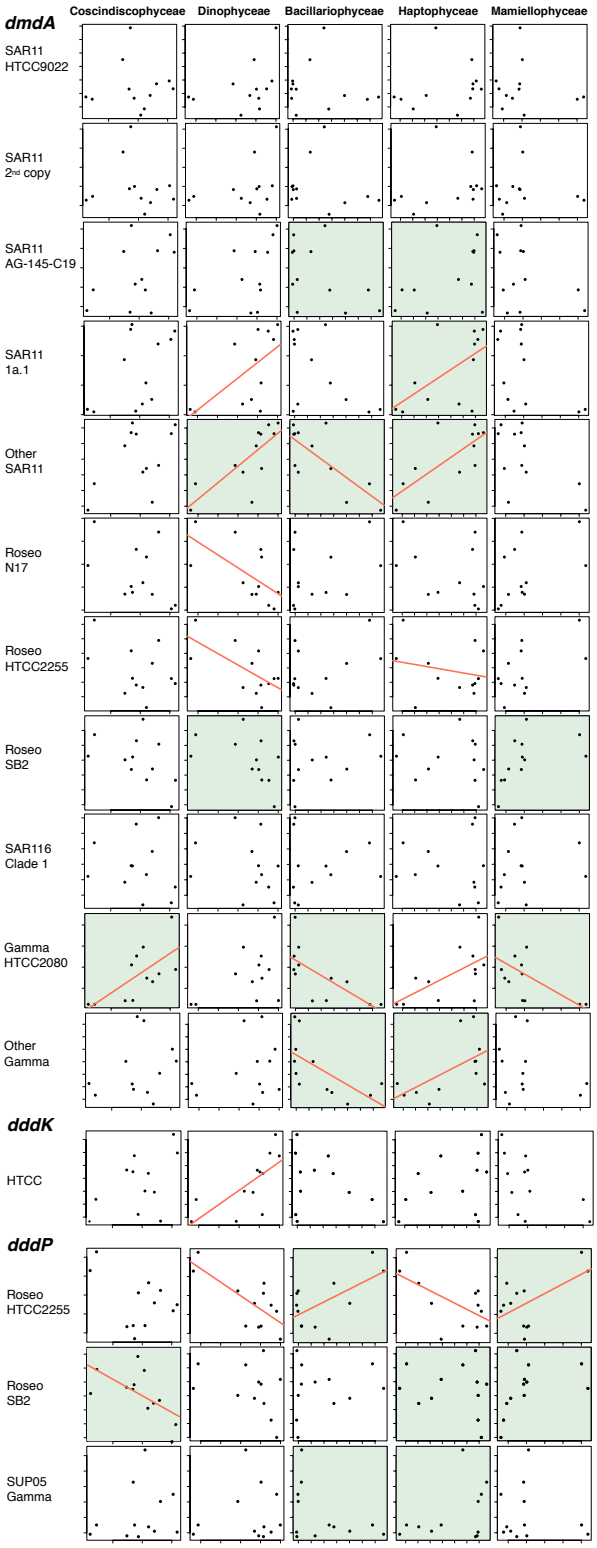

A
SPAdes_06
SPAdes_13
MEGAHIT_03
SPAdes_01/Pelagibacter ubique HTCC9022*
MEGAHIT_04
Pelagibacter ubique HTCC1013
Pelagibacter ubique HTCC1062*
Pelagibacter ubique HTCC1016
Pelagibacter ubique HTCC1002
MEGAHIT_08
Pelagibacter ubique HTCC9565
Pelagibacter-like SCGC AG-145-C19
SPAdes_07
SPAdes_08
Alphaproteobacterium HIMB5*
MEGAHIT_02
SPAdes_02
SPAdes_12
MEGAHIT_07
MEGAHIT_06
SPAdes_14
SPAdes_09
MEGAHIT_01
SPAdes_04
MEGAHIT_05
MEGAHIT_10
SPAdes_10
0.3
6
HTCC
6
Monterey
B
2
SUP05
Candidatus Thioglobus singularis PS1
Candidatus Thioglobus singularis GG2
Cognatiyoonia sediminum
Rhodobacteraceae HIMB11
Jannaschia sp. CCS1
Pseudoruegeria marinistellae
Nioella nitratireducens
Nioella sediminis
Pseudorhodobacter psychrotolerans
Planktomarina temperata
SPAdes_03
3
Planktomarina temperata
Phaeobacter inhibens
Leisingera sp. ANG-M1
Leisingera sp. ANG-Vp
Pseudopelagicola gijangensis
Rhodobacteraceae HIMB11
SPAdes_07
MEGAHIT_02
1
HIMB11
Rhodobacteraceae HTCC2150
Roseovarius nubinhibens*
Ruegeria lacuscaerulensis
Ruegeria conchae
Ruegeria atlantica
Ruegeria halocynthiae
Epibacterium multivorans
Phaeobacter sp. CECT 5382
Epibacterium mobile
Rhodobacteraceae HTCC2083
Planktotalea frisia
Roseovarius tolerans
Litoreibacter arenae
Octadecabacter arcticus
Sulfitobacter donghicola
Sulfitobacter noctilucicola
Sulfitobacter geojensis
1
Other Roseos
3
Roseo C16
SPAdes_02
Rhodobacterales SCGC AG-151-C16
Rhodobacterales HTCC2255
SPAdes_01
1
HTCC2255
MEGAHIT_03
SPAdes_04
Rhodobacteraceae SB2
1.5
RoseoSB2
Thalassobacter stenotrophicus
Yoonia maricola
Pseudorhodobacter ferrugineus
Mameliella alba
SPAdes_05
MEGAHIT_01
Puniceispirillum marinum
SPAdes_06
SAR116 HIMB100
1.5
SAR116
Mesorhizobium sp. LNJC386A00
Mesorhizobium oceanicum
Roseibium sp. TrichSKD4
3
Other Alphas
0.3

Table 2.1. Relative abundance and per-cell DMSP concentrations for phytoplankton taxa represented in the metagenomes.

Phylum/Class	Relative abundance (%)	Order/Suborder	Relative abundance (%)	Species/Clone from Keller (1989)	DMSP (pg cell^{-1})	DMSP (μmol cm^{-3})
Haptophytes:						
Haptophyceae	19.9	Prymnesiales	10.3	*Prymnesium parvum*	1.7	112
		Isochrysidales	3.8	*Emiliania huxleyi*	0.8	166
		Phaeocystales	3.2	*Phaeocystis* sp. 677–3;	2.3	260
				Phaeocystis sp. 1209	1.0	113
		Coccolithales	1.3	*Pleurochrysis carterae*	12	170
Dinoflagellates:						
Dinophyceae	11.0	Gonyaulacales	3.4	*Gonyaulacales spinifera*	145	17
		Gymnodiniales	2.5	*Amphidinium carterae*	19	377
		Peridiniales	2.6	*Scrippsiella trochoidea*	384	350
		Suessiales	1.9	*Symbiodinium microadraticum*	24	345
		Prorocentrales	0.4	*Prorocentrum minimum*	21	111
Diatoms:						
Coscinodiscophyceae	14.6	Chaetocerotophycidae	7.4	*Chaetoceros affinis*	N.D.	
		Thalassiosirophycidae	6.0	*Skeletonema menellii*	0.1	30
Bacillariophyceae	7.9	Bacillariales	7.4	*Cylindortheca closterium*	1.5	41
Mediophyceae	3.2	Cymatosirophycidae	2.5	none		
Fragilariophyceae	2.0	Fragilariales	1.0	*Asterionella glacialis*	N.D.	
Chlorophyta:						
Mamiellophyceae	15.4	Mamiellales	15.2	*Micromonas pusilla*	0.03	162
Prasinophyceae	3.0	Pyramimonadales	1.0	*Pyramimonas sp.*	0.02	0.5
Chlorophyceae	2.0	Chlamydomonadales	2.0	*Chlamydomonas sp.*	N.D.	
Cryptophyta	5.0	Pyrenomonadales	3.8	*Rhodomonas lens*	N.D.	
Pelagophyceae	2.9	Pelagomonadales	2.3	none		
Dictyochophyceae	2.9	Pedinellales	1.2	none		
		Florenciellales	1.2	none		
Raphidophyceae	2.2	Chattonellales	2.2	*Chattonella harima*	N.D.	

Major groups of phytoplankton are divided into phylum/class and order/suborder levels. Intracellular DMSP concentrations measured by Keller (1989) are shown. N.D. = not detectable.

Figure 2.5. Data pairs of DMSP gene abundances (*recA*-normalized % of bacterial cells) and phytoplankton class abundance (% of phytoplankton reads). Y-axes represent percent of bacterial cells harboring the indicated gene, scaled individually for each taxon (see Figure 2 and 4 for scales); X-axes represent abundance of reads assigned to the indicated phytoplankton taxon as percent of total phytoplankton reads. Green shading indicates data pairs with a maximum information coefficient (MIC) value > 0.7. Statistically significant Pearson correlation coefficients ($p < 0.05$) are indicated with an orange line.


Coscindiscophyceae
Dinophyceae
Bacillariophyceae
Haptophyceae
Mamiellophyceae
dmdA
SAR11 HTCC9022
SAR11 2nd copy
SAR11 AG-145-C19
SAR11 1a.1
Other SAR11
Roseo N17
Roseo HTCC2255
Roseo SB2
SAR116 Clade 1
Gamma HTCC2080
Other Gamma
dddK
HTCC
dddP
Roseo HTCC2255
Roseo SB2
SUP05 Gamma

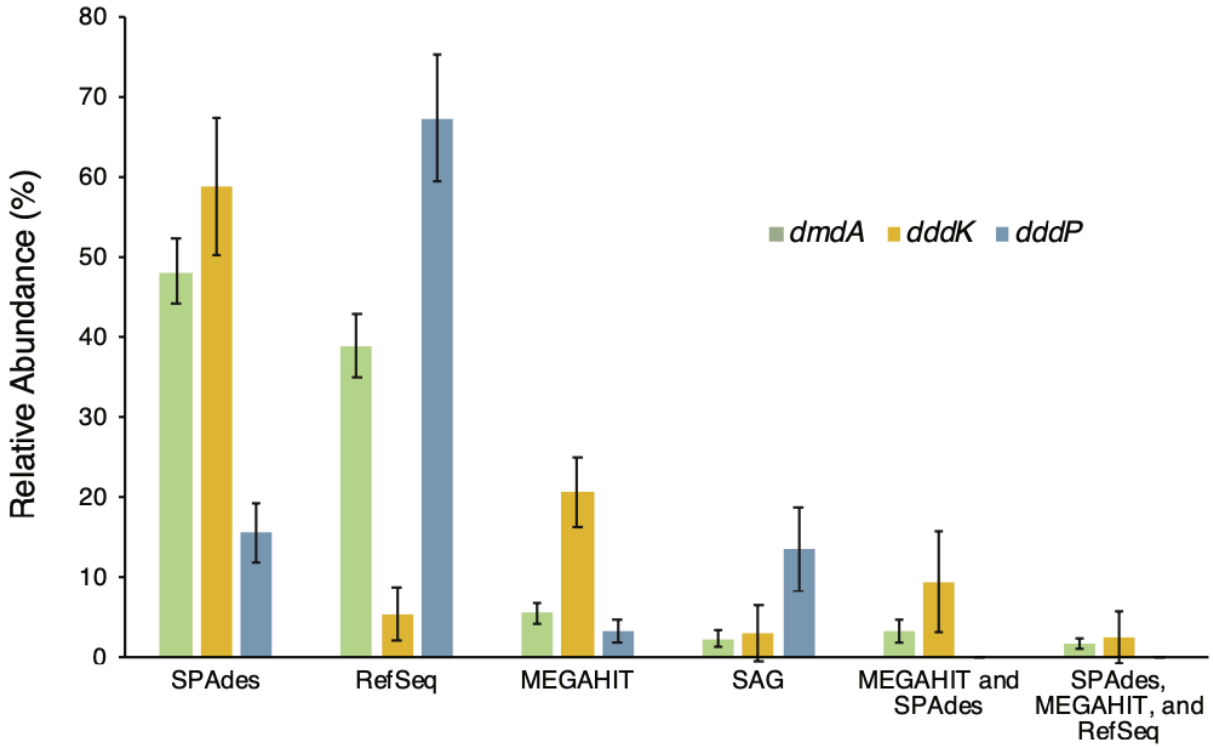


Figure 2.S1. Read placement distribution by sequence assembly method. The bars represent the average percent of reads in the 12 metagenomic samples recruited as the top hit. Error bars represent 1 SD

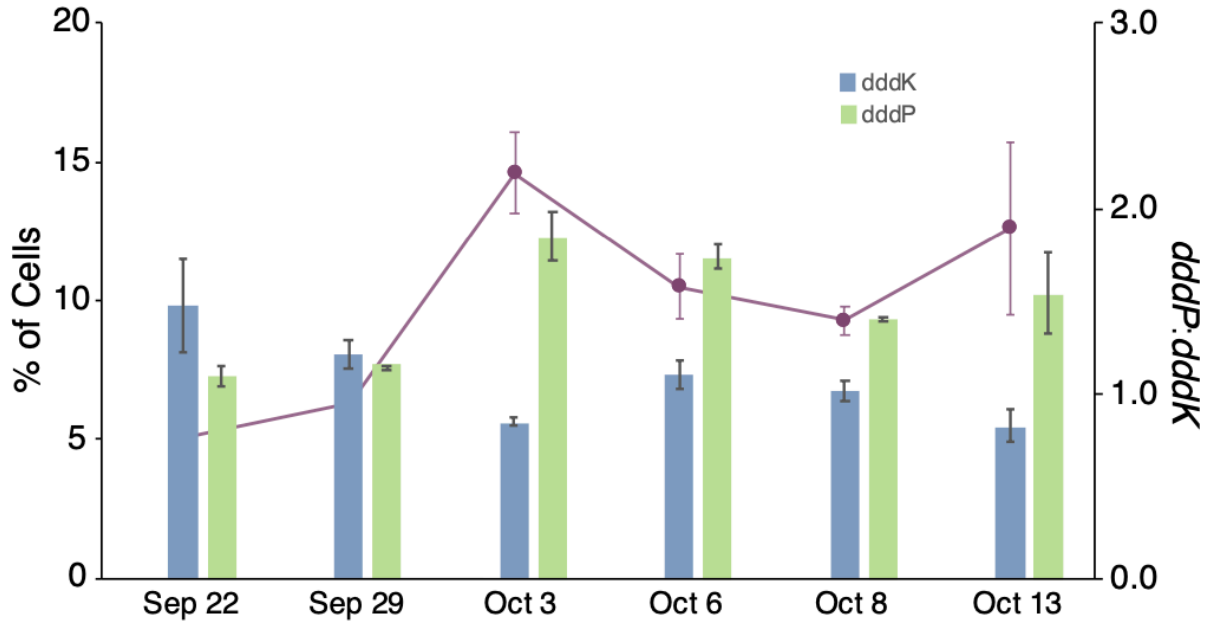


Figure 2.S2. Percent of bacterial cells in possessing DMSP cleavage genes dddP and dddK and the dddP:dddK ratio in fall 2016 in Monterey Bay.

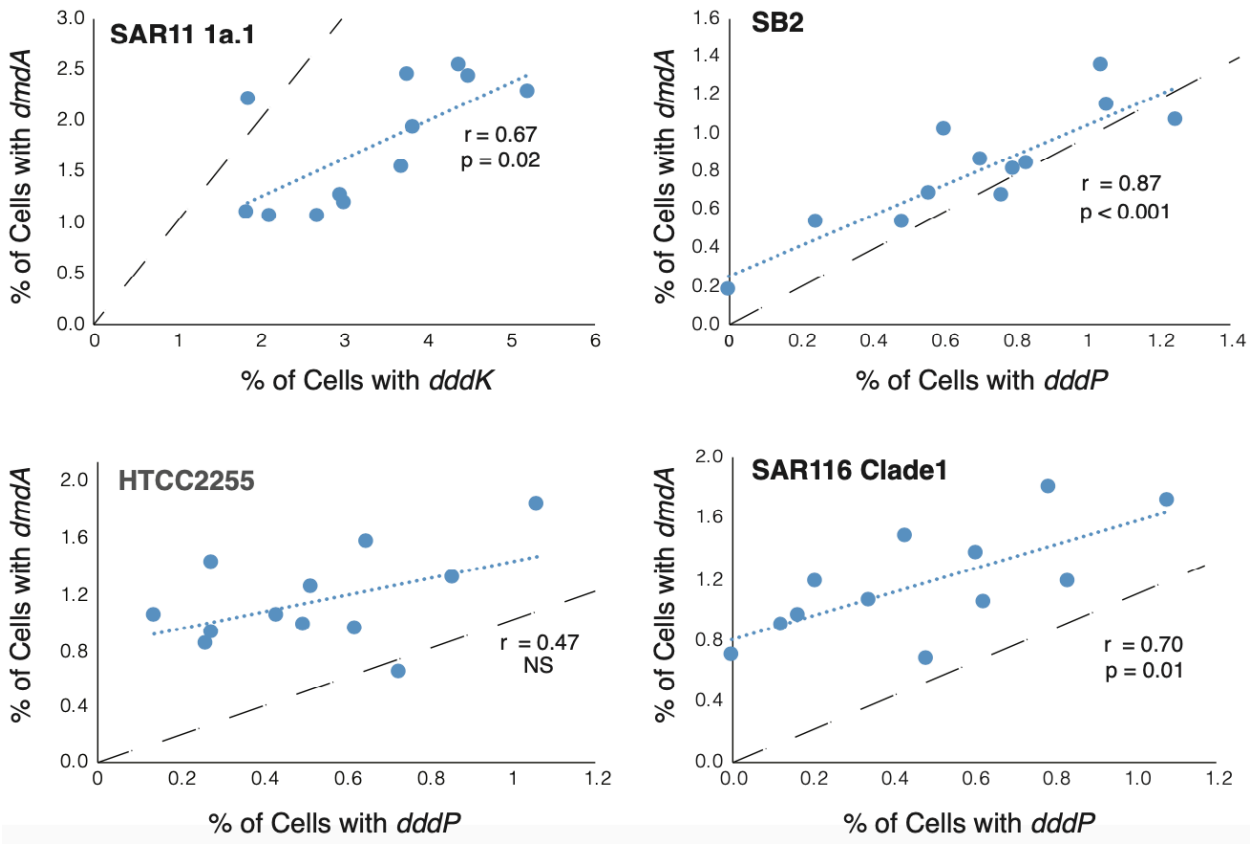


Figure 2.S3. Relationship between percent of cells with a demethylation gene (*dmdA*) and percent of cells with a cleavage gene (*dddK* or *dddP*) for the four cases in which a clade defined by the same reference sequences was present in both gene trees. Pearson correlation and significance level shown, along with the correlation line (dotted blue) and 1:1 line (dashed black). For the SAR11 1a.1 clade, all reference genomes harbour a *dmdA* gene while only half harbour a *dddK*, removing the expectation for a 1:1 relationship.

Figure 2.S4. Additional data pairs of DMSP gene abundances (*recA*-normalized percent of bacterial cells) and phytoplankton class abundance (percent of phytoplankton reads). Green shading indicates relation- ships with a maximum information coefficient (MIC) value > 0.7. Statistically significant Pearson correlation coefficients ($p < 0.05$) are indicated with an orange line.

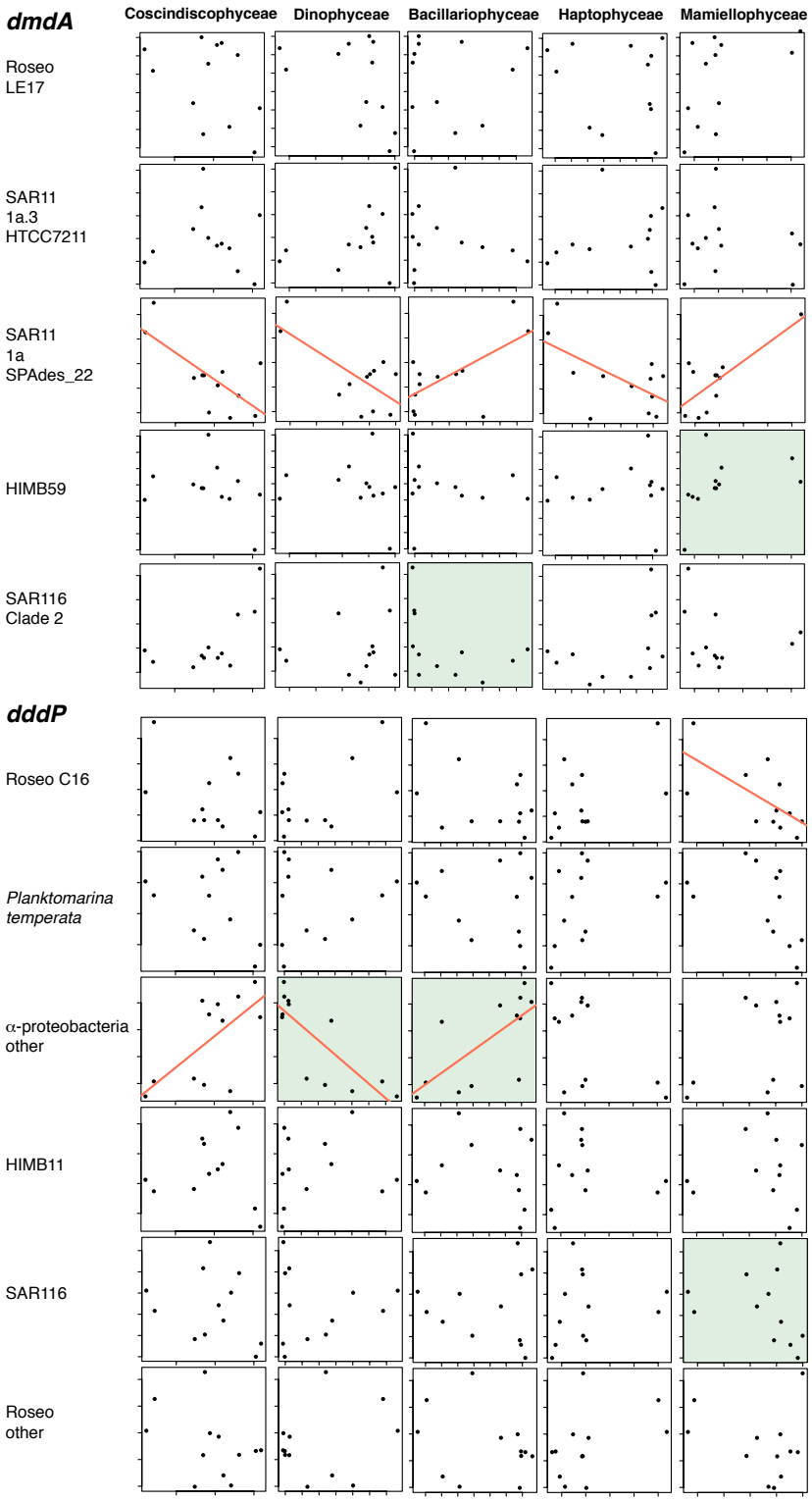

dmdA
Coscindiscophyceae
Dinophyceae
Bacillariophyceae
Haptophyceae
Mamiellophyceae
Roseo LE17
SAR11 1a.3 HTCC7211
SAR11 1a SPAdes_22
HIMB59
SAR116 Clade 2
dddP
Roseo C16
Planktomarina temperata
α-proteobacteria other
HIMB11
SAR116
Roseo other

Table 2.S1. Pearson correlation coefficients for principal component loadings (PC1 and PC2, Figure 3) versus relative abundance of DmdA gene clades in 12 metagenomic samples. * = $p<0.05$; **=$p<0.01$.

DmdA Clade	**PC1**	**PC2**
Roseo N17	0.69*	-0.23
other Gamma	0.29	0.66*
HIMB59	0.56	-0.20
Roseo SB2	0.67*	-0.26
SAR116 Clade 1	-0.23	-0.38
SAR116 Clade 2	-0.17	0.54
SAR11 HTCC9022	-0.81**	-0.18
SAR11 AG-145-C19	-0.61*	0.51
other SAR11	-0.50	0.55
SAR11 1a.1	-0.72**	0.29
SAR11 1a.3 HTCC7211	-0.44	-0.23
Gamma HTCC2080	0.10	0.84**
Roseo LE17	0.78**	-0.14
SAR11 SPAdes_22	0.21	-0.39
SAR11 second copy	-0.64*	0.13

Table 2.S2. Reference genomes containing two copies of *dmdA*.

Taxon	Group
Pelagibacter sp. HTCC7211	SAR11
Pelagibacter ubique HTCC7214	SAR11
Pelagibacter ubique HTCC7217	SAR11
Pelagibacter ubique HTCC8051	SAR11
Pelagibacter sp. RS40 DmdA	SAR11
Pelagibacter ubique SCGC AAA795-M18 *	SAR11
Leisingera sp. ANG-Vp	Roseobacter
Nioella sediminis	Roseobacter
Planktotalea frisia	Roseobacter
Rhodobacteraceae bacterium SB2	Roseobacter

* Although the *Pelagibacter ubique* SCGC AAA795-M18 genome bin does not contain a second copy, it is an incomplete single-cell genome. The DmdA that is encoded is 99% similar at the amino acid level to one in *Pelagibacter* sp. RS40, a SAR11 cell which has two copies.

Table 2.S3. Reference genomes recruiting *dddP* reads and lacking *dmdA.*

Taxon	Group
Candidatus Thioglobus singularis	SUP05 Gammaproteobacteria
Mesorhizobium oceanicum	Rhizobiales
Mesorhizobium sp. LNJC386A00	Rhizobiales
Mameliella alba	Roseobacter
Planktomarina temperata	Roseobacter
Pseudorhodobacter psychrotolerans	Roseobacter
Sulfitobacter donghicola	Roseobacter

Table 2.S6. Reference sequences for BLASTx analysis of Monterey Bay reads consisting of orthologs and close paralogs of DMSP genes *dmdA, dddK*, and *dddP*. The sequence header includes the NCBI accession number and/or GenInfo Identifier. Genes with experimental verification of function are noted in the sequence header.

>gi|294083644|ref|YP_003550401.1| aminomethyl transferase [Candidatus Puniceispirillum marinum IMCC1322] /dmdA/
MAQSGLNMSRRIRRSPFTDKVEEYGVRGFSVVNHMLLPKAFETSVEDDYWHLREHVQI
WDVGVQRQVQITGLDAARLVQMMTPRDVRQAKIGQCLYVPMIDEDAGMLNDPVLIKL
ADDKFWLSIADSDILLWVKGLALGLKLNVDVEEPDVSPLAIQGPKAIALMADLFGEAIR
DLGYFQYGIFDVLGTRQLIARSGYSKQGGFEIYLHGGHLGSDLWDMIYQAGKQYNIMP
GCPNLIERIEGGLMSYGNEFTRDNNPLECGFEELCYFGDDIDYIGKIALRRIAEEGPQKLIR
GIKFGGGKAPPCGKPFLVTTRDNIHIGQITSGIYSPRLKCNVGMSMMAKGHWDFGTVVF
VHTPDGIVREGTVSPLPF

>gi|114771227|ref|ZP_01448647.1| aminomethyl transferase family protein [alpha proteobacterium HTCC2255] /dmdA/
MASAMIFPSRRLRATPFTSRVSKLGVSGFTVYNHMLLPTVFESLQEDYKHLKEYVQMW
DVSVERQVQLLGKDAHKLACMISARDLTNAQTGRCYYAPICDQSGAIINDPIALRLADD
KYWFSIADSDLLLWVQGIALGLDLNVEICEPDVSPLAIQGPMAEDLMVDVFGAEIRNIKF
FHFKEFPFNGRMLNIARSGWSKQGGFEIYLNDSQLGPELWDTIWEKGEKYNIRPGCPNLI
ERIEAGLLSYGNDMNREDSPLEIGLEKYISLDSNVDFIGKKALLKQRKDGIKKRLLGIEID
GSEMPPLSMPEEVFKDGKKIGIVTSAVFSPDYNGNIGFAMIEASNATAGTEVSVDSKAGI
RKGKLCEIGDFWSQVQSKN

>gi|254456019|ref|ZP_05069448.1| aminomethyltransferase [Candidatus Pelagibacter sp. HTCC7211] /dmdA/
MKKFPIAKSRRLRSTPYTDRIERQGVSSYTVYNHMLLPASFVSVEADYHHLKEFVQVW
DVAAERQVEISGKDSAQLVQLMTCRDLSKSKVGKCYYAPIIDGQGNLVNDPIINKLAEN
RWWLSIADSDVIFFAKGLASGNKFDVDIKEPDVNILAVQGPLSDKLMSKVFGEKISQLKF
FNFDYFEFKGMKHFIARSGWSKQGGFEIYVENAEAGKELYDYLFEAGLEFNVKPGCPNL
IERIEGALLSYGNDFDNRDNPLEANFEKYTNLDSEVEFLGKDRLKKIRDKGVKRKLMGV
KIDHDQIDMYCEKTLFDDNNNIIGFVRSATYSPTFKKVIGIAMINKPYWNSKNSFKIEINE
KIHLGNICDLPFI

>gi|254456670|ref|ZP_05070099.1| probable aminomethyltransferase, putative [Candidatus Pelagibacter sp. HTCC7211] /dmdA/
MSKNIKLNMSRRIRRTPYTNRVEQHGVSDFTVVNHMLLPKGFKNTVEEDYLHLSKEVQ
MWDVSCQRQVQICGPDAAKLIQKLTPRSIKDMTIGKCFYIPMLNENAGMINDPVLLKLD
DDMFWISIADSDILLWAKGLALGLNLNVVIEEPDVYPLAIQGPKSEELMVSIFGDEIKKIK
FFNFRVIDFEGTKQIIARSGYSKQDGFEIYFKVHENYFDKVEMGEKLWDTIWEAGKKFNI
SPGCPNLIDRIEAGLMSYGNDFTGENNPLECNLEKYCKADASHDFVGKQALTKIQSEGII
QKMRGIIFDGAPCAATGQPLKIFSKDNKRIGQITSGIFSPRIKKNIGLSMILKDYWNVGNE
VIIETLDGEKRNGTITSLPFPD

>gi|119503015|ref|ZP_01625100.1| aminomethyl transferase family protein [marine gamma proteobacterium HTCC2080] /dmdA/
MNAPVISFSRRLRVSPFELRSLEGSKSASVYNHLVLPTCYESLEADYWHLREHVQLWDV
ACQVQVEVQGPDAAEFVEYLTPRDVSRCQVGQCIYTPLIDEAAGIINDPLVLRLAEDRF
WISLSDSDVLLWAKGLALGKGFDVRVFDPDVFPMSIQGPKSADLLSRVLGDSIRELKFFR
FVETEIAGTPVVVARTGWSGQGGYEIYLQEPDAGVTLWDTLAAAGEDLQVRVGCPNLI
ERIESGLLSFGNDMTLANNPLEAGLDRFFKLGKSADYLGRAALEAIAEEGVKNRLVKLV
IEGEPIANPRTVYTVQGESGENIGTVTSAVYSPRLCCNIGLGYLPVSYCDEGKAAIVLTPQ
GPRELRIANNDWSS

>gi|86139921|ref|ZP_01058486.1| aminomethyl transferase family protein [Roseobacter sp. MED193] /dmdA/
MAFISPSRRLRRTPFSEGVEAAQVKAYTVYNHMLLPTVFESPEADYHHLKKHVQIWDVS
CERQVELRGPDAGRLMQMLTPRDLRGMLPGQCYYVPIVDETGGMLNDPVAVKLAEDR
WWISIADSDLLYWVKGIANGWRLDVLVDEPDVSPLAIQGPKSEELLVRVFGESIRSIRFF
RFGTFQFQGRDLVIARSGYSKQGGFEIYVEDSEIGMPLWNKLFEAGQDLEVRAGCPNLIE
RIEGGLLSYGNDMTDDNTPHECGLGRFCDTHTAIGCIGRDALLRVAKEGPVQQIRAISIA
GEPVPACTEFWPLYAGGKRVGRVSSAAWSPDFRTNVAIGMVRMTHWDAGAKLEVETP
DGMRQATVREKFWI

>gi|56696787|ref|YP_167148.1| dmdA gene product [Ruegeria pomeroyi DSS-3] /functionally ratified dmdA/
MASIFPSRRVRRTPFSAGVEAAGVKGYTVYNHMLLPTVFDSLQADCAHLKEHVQVWD
VACERQVSIQGPDALRLMKLISPRDMDRMADDQCYYVPTVDHRGGMLNDPVAVKLAA
DHYWLSLADGDLLQFGLGIAIARGFDVEIVEPDVSPLAVQGPRADDLMARVFGEAVRDI
RFFRYKRLAFQGVELVVARSGWSKQGGFEIYVEGSELGMPLWNALFAAGADLNVRAG
CPNNIERVESGLLSYGNDMTRENTPYECGLGKFCNSPEDYIGKAALAEQAKNGPARQIR
ALVIGGEIPPCQDAWPLLADGRQVGQVGSAIHSPEFGVNVAIGMVDRSHWAPGTGMEV
ETPDGMRPVTVREGFWR

>gi|71082952|ref|YP_265671.1| dmdA gene product [Candidatus Pelagibacter ubique HTCC1062] /functionally ratified dmdA/
MKNFSIAKSRRLRSTPYTSRIEKQGVTAYTIYNHMLLPAAFGSIEDSYKHLKEHVQIWDV
AAERQVEISGKDSAELVQLMTCRDLSKSKIGRCYYCPIIDENGNLVNDPVVLKLDENKW
WISIADSDVIFFAKGLASGHKFDVKIVEPVVDIMAIQGPKSFALMEKVFGKKITELKFFGF
DYFDFEGTKHLIARSGWSKQGGYEVYVENTQSGQKLYDHLFEVGKEFNVGPGCPNLIE
RIESALLSYGNDFDNNDNPFECGFDQYVSLDSDINFLGKEKLKEIKLKGPQKKLRGVKID
IKEISLTGSKNIYDENNNVIGELRSACYSPHFQKVIGIAMIKKSHWEASQGFKIQINDNTIN
GNVCDLPFI

>gi|399156270|ref|ZP_10756337.1| dimethyl sulfoniopropionate demethylase [SAR324 cluster bacterium SCGC AAA001-C10] /dmdA/
MRPELLISSRTRSTPFSSRVEACGVKAYSVYNHMLLPLIFRSLEEDYWHLCESVQVWDV
SCQRQVEITGPDTQKLVQLMTPRDLSQAELGQCFYAPLCDETGGMINDPILIKHSNNHW
WLSIADSDVMLWAKGLATGFGLDALVTEPDIWPLAVQGPKAEELLSRVFGKEISKILFFR
SSTFNYQGTKMLVARSGWSKQGGFEIYVNDAELGGQLWDELFAKGEDLNVGPGCPNLI

ERIESGLLSFGGDMGYDTTPFECGLEKYVSLDADIESLSLDELRTRKSKTKLIGIVIDRKV
NLINKRVFIGSNVVGKITSDTWSPRYSAFLAFARCELKHLEDAQNXGIDXGTXX

>gi|167753016|ref|ZP_02425143.1| PARALOG hypothetical protein ALIPUT_01280 [Alistipes putredinis DSM 17216] /dmdA_Paralog/
MKTTVFTKHHIANGAKMAEFAGYNMPIEFTGINEEHLAVRNNAGVFDVSHMGEVWVK
GPKAEAFLQHITTNDVAALYDGKVQYTTMPNGKGGIVDDLLVYRIDAETYLLVINAANI
DKDWNHIVEEGKKFGLEAGHGKQLYNASDEICQLAIQGPNAMKIVQKLCTEPVEDMEY
YTFKKLKVAGVDAILSITGYTGAGGCEIYVANEDGEKLWKALWQEGSKEGLKNIGLGA
RDTLRLEMGFCLYGNDIDDTTSPIEAGLNWLTKFVDGKEFIDRKLMEEQKAGGLTRKLV
GFKMIDRGIPRHGYQIASPEGDIIGQVTSGTMSPCLKQGIGMGYVKKEFAKVGTQIAIVIR
EKLMKAEVVKLPFIQQK

>gi|94968297|ref|YP_590345.1| PARALOG glycine cleavage T protein (aminomethyl transferase) [Candidatus Koribacter versatilis Ellin345] /dmdA_Paralog/
MPIGTAFHERTFGLCQSLSYREWSGYYTVSSYETHHEHEYNAIRNACALIDISPLFKYLIT
GDDATQFVNRVITRDIKKVAINQVIYCCWCDQDGKVIDDGTITRLGENTYRWTAADPSL
RWFRQNSIAMKVQIEDISESVSALALQGPTSAALLASVAEADIANLKYFRMTKGRINGID
VDISRTGYTGDLGYEIWIPWEHSLRVWDALATAGNAFDLHPVGMLALDVARIEAGLLLI
EVDYFSSKKALIDSQKYSPFELGFDKMVHLDKETFVGREALLKEKGSRTGRKLVGLEFD
WTAVEKLYDRVGLPPQVPSAASRVPVPVYRGNVQAGKATSTTWSPILKKMIALASVDA
AHSAIGTELQAEITIEAVRYKTAVKVVQLPFFNPARKSAVPPRL

>gi|119718058|ref|YP_925023.1| PARALOG FAD dependent oxidoreductase [Nocardioides sp. JS614] /dmdA_Paralog/
MTNLPDRARVVVIGGGVIGCSVAYHLAHAGWSDVVLLERDRLTSGTTWHAAGLMTCF
GSTSETSTAIRLYSRDLYARLEAETGQATGFRPVGLIEAAADEARLEEYRRVAAFQRHLG
LEVHEISPREMADLFPWARTDDLLAGFHVPGDGRVNPVDLTLALAKGARRLGVRIVEG
VSVSDVQVSPGPAGGTDRVTGVTTTAGDIECEYVVNCAGMWARELGARNGLVIPNQA
AEHYYLITDTIEGLDPDAPVFEDPASYGYYREEGGGMMVGLFEPVAAPWRVDGVPADF
SFGTIPPDWDRMGPFLEKAMARVPVTLDAGVRTFFCGPESFTPDLAPAVGEAPGLRNYF
VAAGMNSVGVLSAGGLGRVLAEWITTGRPDVDVTGFDVHRFRPWQADDAYRAARTTE
ILGTVYAAHTPGTQLRSARGTLLSPVHDRLVEQGGYLREVSGWEGADWFAGPGTTPVA
EPSWGRAPWFREWAAEHRAVREGVGLMDMSFMAKLAVRGAGAAALLDRVSAGDVT
ASVETITYTQWLDERGRIEADLTVTKLADDDFLVVASDTAHGHTLAWLRGAVADGTDV
RIEDVTADYAQLNVQGPRSRDLLAALTDADLSTAAFGFRTARWIEVAGVRVLCARITYL
GELGYELYVPAGSGLKVYDALQDAGPAYGLRPVGLKALASLRMEKGYRDFGHDIDNT
DCPLEVGLGFALSLDKPGGFVGRDAVLERKAANAAAGGMGQRLVQVRLLDPDPLLHH
AEVVHRDGVPVGYVRAASYGWTLGGAVGLAMVSGQGAPVTPDWLSGGTWEVDVAG
TRHRAEVSLRPMYDPASARVRA

>gi|284801733|ref|YP_003413598.1| PARALOG glycine cleavage system aminomethyltransferase T [Listeria monocytogenes 08-5578] /dmdA_Paralog/
MCYDRAYELYKPFCYVKKEDIIMTELLKTPIHPLYAKYGAKTIDFGGWDLPVQFAGIKA
EHEAVRTDAGLFDVSHMGEILVKGPDSTSYLQYLLSNDIEKIKIGKAQYNIMCYETGGT
VDDLVVYKKSETEYILVVNAANTDKDFEWMVKNIRGDVSVTNVSSEYGQLALQGPNA

EKILSKLTDVDLSSISFFGFVEDADVAGVKTIISRSGYTGEDGFEIYMPSADAGKVFEAIL
AEGVAPIGLGARDTLRLEAVLALYGQELSKDITPLEAGLNFAVKLKKEADFIGKEALIKQ
KEAGLNRKLVGIELIERGIPRHDYPVFLNEEQIGVVTSGTQSPTLGINIGLALIDTAYTELG
QEVEIGIRNKKVKAKIVPTPFYKRAK

>gi|254281873|ref|ZP_04956841.1| PARALOG aminomethyltransferase [gamma proteobacterium NOR51-B] /dmdA_Paralog/
MNIVAPADPELSSDDRFAGERLKLSPFHPRQAELNIRDAWSAWNGYKFADYYYEATYE
YFCIRNTCGTYDICPMQKFLVEGEDALAMLDRMVTRDLTKLRVNRITYCCWCDDTGR
MIDDGTIFRLDDNRYMLTCGSPCLAWLAKSALGFDKVTITEHTEQLAGLSLQGPTSFSTL
KNMGVGDAVAELKPFGFTRVPFVGTELMISRTGFTGDLGYELWIDAEYALPLWDALYE
AGEDYGIQPYGEAATNMARLEAGFIMPYMEFNEAPKTINFEHDQTPLELNLGWLVDFK
KPHFNGRRALLEQKQKGTKQLLVKLDIEGNKPAEEAILYDSKGCRNQIGYVTSAMWSPS
VKANIALAMIDTKALTGEIWAEIYHYKELRPYRKVAKCKVQDKPFWMPPRARQTPPGE
F

>gi|56751803|ref|YP_172504.1| PARALOG gcvT gene product [Synechococcus elongatus PCC 6301] /dmdA_Paralog/
MTLTVTVSLLSSPLHSVCTSAGARFTGFAGWELPLQFQGLMQEHLAVRERAGLFDISHM
GKFQLRGSGLRAALQRLLPSDLTTLLPGQAQYSVLLNEAGGCLDDLIVYWQGIVDGVE
QAFLIVNAATTDSDRLWLTEHLPPAIALLDLSQDLALVAIQGPQAIAFLQPLVSCDLAELP
RFSHTVTSIAGQPAFVARTGYTGEDGCEVMLPPAAAITLWQQLTAAGVVPCGLGARDT
LRLEAAMPLYGHELDTDTNPLEAGLGWVVHLDRNPDFLGRDRLVQAKTNGLERRLVG
LELPGRNIARHGYPVAIADTTVGIVTSGSWSPTLSKAIALAYVPPALANLGQELWVEIRG
KQVPATVVKRPFYRGSQFR

>gi|148655664|ref|YP_001275869.1| PARALOG glycine cleavage system T protein [Roseiflexus sp. RS-1] /dmdA_Paralog/
MSEAVGLRRTPLYERHLALGARMVAFGGWEMPVQYSGIIEEHRAVREAAGLFDISHMG
EVEVRGPDALPFLQYLVTYDVAAIPPGRANYALMCRPDGGIIDDTFIYNLGDYYLIVVN
AANTAKDVAWMHECAKGFNVTVSDVSDQTGMLALQGPLAEALLAQVADADLAALPF
HGVMQGRVVHTPAIVARTGYTGEDGFEIFVAAGDVTRVWDELLDAGRTIGLKPCGLGA
RDSLRFEACLALYGHEITEETNPYEARLGWVVKLDKGDFIGREALQRIKQEGVARRLTG
FEMAGRGIARSEYEIRDLEGAPIGRVTSGMPSPTLGKNLGMGYVPVAFSTEGSEFDVVV
RDRPVRARAVKMPFYRPRYKKG

>gi|148655664|ref|YP_001275869.1| PARALOG glycine cleavage system T protein [Roseiflexus sp. RS-1] /dmdA_Paralog/
MSEAVGLRRTPLYERHLALGARMVAFGGWEMPVQYSGIIEEHRAVREAAGLFDISHMG
EVEVRGPDALPFLQYLVTYDVAAIPPGRANYALMCRPDGGIIDDTFIYNLGDYYLIVVN
AANTAKDVAWMHECAKGFNVTVSDVSDQTGMLALQGPLAEALLAQVADADLAALPF
HGVMQGRVVHTPAIVARTGYTGEDGFEIFVAAGDVTRVWDELLDAGRTIGLKPCGLGA
RDSLRFEACLALYGHEITEETNPYEARLGWVVKLDKGDFIGREALQRIKQEGVARRLTG
FEMAGRGIARSEYEIRDLEGAPIGRVTSGMPSPTLGKNLGMGYVPVAFSTEGSEFDVVV
RDRPVRARAVKMPFYRPRYKKG

>gi|300778775|ref|ZP_07088633.1| PARALOG aminomethyltransferase [Chryseobacterium gleum ATCC 35910] /dmdA_Paralog/
MKKTALYDKHVSLGAKIVPFAGFEMPVQYSGVTEEHFAVREKAGLFDVSHMGQFFIEG
PGSKDLLQFVTTNNVDTLENGKAQYSCLPNENGGIVDDLIVYKMEDDKYFVVVNASNI
DKDWNHISKYNTFGAKMTNASDEMSLLAVQGPKATEILQKLTDVNLSEIPYYHFTVGS
VAGENDVIISNTGYTGSGGFEIYFKNESAEKLWDAVMEAGQEEGIIPCGLAARDTLRLEK
GFCLYGNDIDDTTSPIEAGLGWITKFDKDFVSKDVFAKQKEEGVSRKLVGFELTDKGVP
RHDYPVVDAEGNVIGKVTSGTQSPMKKVGLGLAYVDKPHFKLGSEIFIQVRNKNIPAKV
VKAPFV

>gi|188587345|ref|YP_001918890.1| PARALOG aminomethyltransferase [Natranaerobius thermophilus JW/NM-WN-LF] /dmdA_Paralog/
MTHPQKTPLYDIHKERGGKIIDFGGWYLPVQFTGIIDEVMTTRKEAGLFDVSHMGEIIVE
GPKALEYLQKMVPNDVARLKPGKILYTPMCYENGGTVDDFLIYKMDENKFLLIVNAAN
TDKDFEWLQENNTEGVELKNLSDEYGQIAIQGPKAEKILQRLTDTPLKEIKFFNFKEDVD
LDGVKALISRTGYTGENGFEIYIKAEETAKLWEKIEDAGENDGLKPIGLGARDVLRFEVC
LPLYGNELSPEITPLEARLNPFVKLNKTEDFLGKDVLVNQKEQGLERVLVGFEMIDRGIP
RTNYILMKDGQEIGFVSSGSQSPTLDKALGLGFIKPEHDQEGNEIEVKIRKKTAKAKIVKT
PFYRRG

>gi|124003958|ref|ZP_01688805.1| PARALOG glycine cleavage system T protein [Microscilla marina ATCC 23134] /dmdA_Paralog/
MDLKTTDLKQTALNDIHVALGGKMVEFAGYSMPVRYTSDKEEHFAVRENVGVFDVSH
MGEFLLKGEGALDLIQKVSSNDASKLYPGRVQYSCLPNDQGGIVDDLVIYMIAENEYYL
VVNASNVQKDWDWISKHNTYGVEMTNLSDQTSMLAIQGPKATQALQSLTDVKLDDM
KFYTFEKATFAGVPDVIISATGYTGLGGVELYVPNEHAETIWNKIFEAGKDYHIQAIGLG
ARDTLRLEKGYCLYGNDIDDTTSPLEAGLGWVTKFTKDFVNSEALKKQKEEGVKRKLV
AFKMVDKGIPRHGYELLDTDGKNIGKVTSGSMSPSLNIGIGLGYVTKELSKPGNEIMVQ
VRNKQLKAEVIKLPFI

>gi|297559405|ref|YP_003678379.1| PARALOG glycine cleavage system protein T [Nocardiopsis dassonvillei subsp. dassonvillei DSM 43111] /dmdA_Paralog/
MPMSDAASAPRATALREVHEKAGATLVDFAGWLMPLRYGSETAEHRAVREAAGLFDL
SHMGEIRLTGPQAAQALDHALVGHLSQVKVGRARYSMITAEDGGVLDDLIVYRLREDE
YLVVANAANTAVVAPALAERAAGFDVEVRDESAEYALIAVQGPRAVDVLAPLTDADL
DGIRYYAGYEHTVAGEPVLLARTGYTGEDGFEIFVSPADRAPKVWDALMAEGERHGLV
PAGLSARDTLRMEAGMPLYGQELTADLTPFDAGLGRVVKFDKGDFVGRAALEEASRSS
RPRRLIGLVARGRRPLRQGQEVLRDGTPVGTITSGAPSPTLGRPIAMAYVDGDLDTSTGA
FTVDVRGRGEDVDVVELPFYKRQS

>gi|295134431|ref|YP_003585107.1| PARALOG glycine cleavage system aminomethyltransferase T [Zunongwangia profunda SM-A87] /dmdA_Paralog/
MKEVALANKHKELGAKMVPFAGYNMPVSYEGVNAEHHNVREKLGVFDVSHMGEFLV
TGENALALIQLISSNDASKLVDGQAQYTCMPNEKGGIVDDMIIYRMNAEKYLLVVNAA
NIEKDWNWISKHNTMDANLTDLSEELSLLAIQGPKAAEAMQSLTDVDLSAMKFYTFEIG
TFAGMEKVIISATGYTGSGGFEIYFKNECAQEIWDKVMEAGKDYGIQPIGLAARDTLRLE

MGFCLYGNDIDDTTSPIEAKLGWITKFTKDFINAEALKQEKEEGPKRKLVAFELDERGIP
RQGYDIVNDEGEVIGNVTSGTMSPSLEKGIGLGYVKSEYTGFGKKINIQIRKKAVSATQV
KLPFYKG

>gi|91762978|ref|ZP_01264943.1| PARALOG putative aminomethyltransferase protein [Candidatus Pelagibacter ubique HTCC1002] /dmdA_Paralog/
MSNEFDYTKLNHVTSVDQSDREVPYNLRQSGPTKVEMLISTRVRKSPYWHLSMEAGC
WRATVYNRIYHPRGYVKPEDGGAMVEYEAIKNHVTMWNVAVERQIRVKGPDAEKFTD
YVITRDATKISPMRARYVILCNAYGGVLNDPILLRISKDEFWFSLSDSDIGMYLQGVNAD
GRFDCTIEEIDVCPVQIQGPKSKALMKDLIGDQVDLDNMPFYGLAEAKVGGRSCVISQS
GFSGEAGYEIYLREATKYADDMWNAVLEAGKKHSLMVIAPAHHRRIQAGILSWGQDM
DHQHNPFQCNLGYQVSLSGKGEWNKKADYVGKAALEKMGADLKAGQKPYKLQLVGL
ELGGKPIEEYAPDFWLVSPESGGDPVGFITSPWYHPEKGQNIAMGYVPFDGTLNANGFP
KGKVGTKYKVHLPAKYSDTPGTPVDAVVVDIPFTESFNANTREVVKG

>gi|254479868|ref|ZP_05093116.1| PARALOG Glycine cleavage T-protein (aminomethyl transferase) [marine gamma proteobacterium HTCC2148] /dmdA_Paralog/
MNENQFISALTIGPRVRKSPFFDATLAAGVKSFTVYNHMYLPTSYGDPLQEYWAMVEG
VTLWDVSCERQIEVSGSDAIEFTQLLTPRDVASCPVGRCRYVVFTDHDGGVINDAIMLR
LEESRFWLSPGDGDVLLWAQGVAARSGMDVKLTEPDVSPLQLQGPLAPKVARKLFGD
VAVEMGYYHLHELELNGIPLVLSRTGWSGELGYEIYLRDGSRGTELWDLVMAAGEEFG
IKPACPSAMRTIEGGILSYASDITREDTPFTIGMERLLDLDKSQDYIGKAALQQIAKEGTP
RRLVGIEIDGDPIGGNDRFWDVFENQDKVGHLTRCAWSPRLERNIGLVNLPTELAEPGT
ALKVQTLDDLRDGIVVALPWFKSITKIPDDL

>gi|304394587|ref|ZP_07376506.1| PARALOG aminomethyl transferase family protein [Ahrensia sp. R2A130] /dmdA_Paralog/
MFSIFPTARLRPSPFYDATVAEGMNSAMVYNGMILPASYGDREAEYWRLINGVSQWDV
AVERQVQLKGPDAAELAQILSPRDLTKCKVGQGKYVAMCNHDGAIVNDPILLKLDEDL
FWFSIADSDVWLWASAIAAERKLDVEITEPDVSPMALQGPMAEDVVAHVCGDWVRDL
KYFWFRESHIDGIPVAVQRSGWSKQGGFEIYLRDGTKGTQLWNIFKEAGQPWGIGPGAP
TSAERTESGLVSVGGDTSDNTNPYEVRLGRYVDLHVPDHVVGIQALRKIAEEGPKRHQL
GVILEGDVPAPLGLNWEPIILNGEHLGDMTNCVWSPRMNANIGYALISVKAQIGDDVTI
QRPAGEVSAKLVDLPFI

>gi|56696532|ref|YP_166889.1| PARALOG glycine cleavage system protein T [Ruegeria pomeroyi DSS-3] /dmdA_Paralog/
MFSISPSTRLRPSPFYEATLADGVCAMTTYNQMLMPTSYGHPEEEYWRIINGVSMWDV
AVERQVQLMGPDAGRLAQILAPRDLSKCKIGQGKYVPLCNHNGVLINDPILLKLDEDRY
WLSIADSNIWFWAEAIARERGLKVEVSEPDVSPLAVQGPKAETVVASIFGDWVRDLKYF
WFRETEIDGIPVAVARSGWSKQGGFEIYLMDGTKGTALWNIVKEAGQPQGIGPGNPNW
CERVESGLVSYGGDSDGQTNPFEVRMGKYVDLDLPDDTIGIEALRRIAAEGPKRHQLGV
VLDNSEPVKAEFTWNDIDMDGMRIGDMTTCVWSYRMNKNIGFALVATSARPGDRVVV
RRAAGAVEGTLCDLPFL

>gi|294085739|ref|YP_003552499.1| PARALOG aminomethyl transferase [Candidatus Puniceispirillum marinum IMCC1322] /dmdA_Paralog/
MSLALSVGPRVRKSPFFSSARKAGLAAASVYNHMYMPTSYGDPMAEYDRLINGVAMW
DVAVERQVALKGPDAIALAKYLTPRNLDNLKVGVGKYVPLCDFNGMLINDPVLLQISE
DEVWLSIADSDVKLWAAGIAGARGMDVRVYEPDVSPLAIQGPKASDVVRDLFGDWVN
EIKYFGFRATELKGIPLVLARSGWSKQGGFELYLQDGSKGDALWDIVAEAGKPYGIGPG
TPNYIERVESGLISYGADTDEMSNPFELGMDRLIDLDQPQDFVGKAALSDIKARGATRRF
MGLIIDGEKFTSTNESRWPVEWNGANAGYVSASAYSPRLDANIAMAMVSVAAIESGDK
VHVLNETGRLTAKIVSLPMV

>gi|254454805|ref|ZP_05068242.1| PARALOG glycine cleavage T protein [Octadecabacter arcticus 238] /dmdA_Paralog/
MTQANDFGFGTQIRKSPYFDATVRWGAKGFSVYNHMYIPRDFGDPEQNFWNLVDKAIL
CDVAVERQVEITGPDAAKFVQILTPRDLSKMAVGQCKYILITNADGGLLNDPILLRLAEN
HFWISLADSDILLWAQGVAVHSGMDVQIVEPDVSPLQLQGPNSGLIMQELFGESIMDLK
YYWLREVELDGIPLVVSRTGWSSELGYELYLRDGSRGDLLWERIMAAGMEYGLKPGHT
SSIRRIEGGMLSYHADADIHTNPYELGFDRLVNLDMDADFIGKAALRRIKDEGPKRKQV
GLVIDCEPLTGPNTMFWTINQGGADIGKVTSAVYSPRLEKNIALAMVAADAAVIGAEVE
VVTKSGPTKATVVERPFYDPKKQIAAA

>gi|149374589|ref|ZP_01892363.1| PARALOG putative aminomethyltransferase protein [Marinobacter algicola DG893] /dmdA_Paralog/
MAVKFEQALLDYPQQRAGAARQPDSVDQSDRRVPINLRQSGPTPVEMLISTRVRKSPY
WHLAYEAGCWRATVYNRMYHPRGYVRPEDGGAMVEYESLIHDVTMWNVAVERQIQV
KGPDAERFVNYVITRDATKIKPMRGKYVILCNEDGGILNDPVLLRVAEDEFWFSLSDSD
LEFWLRGVNIGMGFNVTIAEIDVAPVQIQGPKSEALMADLFGERVKEIPYYGLMEGQVA
GHDVIISQTGFTGEKGYEIYLKEATKYAEDLWYTVLAAGEAHNLRVIAPAHHRRIAAGIL
SWGQDVDQETLPFQCNLAYQVPRNKDADYIGKQKLEKVRDQLDAGRPPFSHIMVGIRF
GGGQVTDYANDFWLVSGPDGGEPEGYVTSPWYSPELETNIGLAYVPFDLRAVGTRLMV
HLPVEYAATDGSTAVEAEVVEVPFRPSVNPNARERARAKGIDFAD

>gi|294085739|ref|YP_003552499.1| PARALOG aminomethyl transferase [Candidatus Puniceispirillum marinum IMCC1322] /dmdA_Paralog/
MSLALSVGPRVRKSPFFSSARKAGLAAASVYNHMYMPTSYGDPMAEYDRLINGVAMW
DVAVERQVALKGPDAIALAKYLTPRNLDNLKVGVGKYVPLCDFNGMLINDPVLLQISE
DEVWLSIADSDVKLWAAGIAGARGMDVRVYEPDVSPLAIQGPKASDVVRDLFGDWVN
EIKYFGFRATELKGIPLVLARSGWSKQGGFELYLQDGSKGDALWDIVAEAGKPYGIGPG
TPNYIERVESGLISYGADTDEMSNPFELGMDRLIDLDQPQDFVGKAALSDIKARGATRRF
MGLIIDGEKFTSTNESRWPVEWNGANAGYVSASAYSPRLDANIAMAMVSVAAIESGDK
VHVLNETGRLTAKIVSLPMV

>gi|84516541|ref|ZP_01003900.1| PARALOG aminomethyl transferase family protein [Loktanella vestfoldensis SKA53] /dmdA_Paralog/
MQADDFGFGTQIRKSPYFDATLRWGAKGFSVYNHMYIPRDFGDAEQNFWNLVNDAILC
DVAVERQVEITGPDAAQFTQMLTCRDLSKMAVGQCKYILITNADGGILNDPILLRLAEN
HFWISLADSDILLWAQGVAIHSGLNVTIREPDVSPLQLQGPKSGEIMKALFGEDILDLRY

YWLREVELNGIPLVVSRTGWSSELGYEIYLRDGAKGDLLWETIMAAGMEFGLKPGHTS
SIRRIEGGMLSYHADADMTTNPFELGFDRLVNLDMEADFIGKAALRRIKDEGVSRKQIG
LIIDGDPLAGPNTTFWAINLGGDTIGKVTSAVYSPRLKQNIALAMVSAEHANIGAVVEVV
THSGPTIATICERPFYDPKKQIAAA

>gi|110667811|ref|YP_657622.1| PARALOG aminomethyltransferase, glycin cleavage system T protein [Haloquadratum walsbyi DSM 16790] /dmdA_Paralog/
MMNTIIYCSIYLEYISMANSEDHPNYPSIDQSDRTLPRNLRQTGDPGIEMLVSTRVRKSPF
FDKSFNEEGAWRCTVYNRIYHPRGLVEPEDGGAMAEYDALTEAVTLWDVAVERQIRV
KGPDAEALTNYVITRDATEIDPMHGKYVILCNEDGGILNDPILLRVAEDEFWFSISDSTL
MQWIEGVNVGMDFDVEVDEIDVAPMQIQGPRSEDVMVDVVGEEVSEIPYYGLMEAEIG
GAEVLISQTGFSGEKGFEIYVRDAMETAERVWDPVLDSVKDHGGMQIAPGHHRRIAAGI
LSWGQDMDHETSPFQVNLGYQVPDNKQADYIGKEELERQQALIDDGEYPFNLKLVGLK
MSGEPIRDYAPDFWLVSDPDTGEECGYMTSPWWNPDLETNIGLGFVPADKLEAETDAL
LNDEIYENDLDLEFQVHLPEEYAESGGPAYATVAEVPFKESVNPSAREQAKLGARQQAE
ANDD

>gi|134102067|ref|YP_001107728.1| PARALOG aminomethyltransferase [Saccharopolyspora erythraea NRRL 2338] /dmdA_Paralog/
MTINQNPGVLQYPRLRKSPFYYASRRHGVALYSVYNHTYHPRHYGDPVAEYWHLLEG
VTLWDVGVERQVEITGPDAFEFTNMLVPRDLNKCKVGQCKYVFVTAEDGGIINDPVLL
RLGENHFWLSLADSDVLLWAKGLAHSLGMDVQIHEPDVGPVQIQGPKSREVMADLFGE
SILDVPYYYAVDRELDGMQVVVSRTGYTAELGYEVYLHNASRDGVRLWDAIWQAGEP
HDLRVIGPCHIRRIEAGILSWGCDLTYDTNPFEVGYGFETTWMVDLEQEADFIGKQALTR
IRDEGVSRKLVGVEIGGPGVGSFNDGNMIDVFDVHDPRGLRIGEVTSACYSPRLERNIGY
AMVPVAYQEYGTELVVHTQHGPQEAVVVQKPFLDPTKSIPKRLVRASA

>gi|257069117|ref|YP_003155372.1| PARALOG glycine cleavage system T protein (aminomethyltransferase) [Brachybacterium faecium DSM 4810] /dmdA_Paralog/
MTVNPNPHVLLYPRIRKSPFFYASRRHGVAMYSVYNHTYHPRHYGDPVAEYWALLEG
VTLWDVGVERQIEISGPDAFDFTNLLVTRDLSKCAVGQCKYVFLTDQHGGILNDPILLRL
EENRFWLSLADSDILLWARGVATHAGMDVSIEEIDVGPVQVQGPKSYAVMRDLLGEAV
ADLRYYYLHDFTLDGIDVTVSRTGYTGEIGYEIYVHDASQNAEKLWQLVLEAGEPHGL
RVIGPCHIRRIEGGMLAHGADITVQTTPFEVGMGYDWMVDLEQEADFVGKDALRRLKA
EGPRCKLVGLEIGGEPLGSYNDGSMIDAFPVHHDGAVVGQVTSACHSPRLEKNIGLALV
PAALSEIGTRFQIDTGPRPGAQLPSGEELVEAVVVPKPFIDPTKEQPKGDVTALGRGEDT
RSTDAAAGSRDAARA

>gi|71083370|ref|YP_266089.1| PARALOG gcvT gene product [Candidatus Pelagibacter ubique HTCC1062] /dmdA_Paralog/
MDILKTALYSLHQKHGAKFVPFAGYQMPIQYSKGIIEEHKSTRENAGIFDVSHMGQLFIK
GDDKLAKDLEKIFPAELSKAKLNQSKYSFLMNDEAGIYDDLIITKVEGGFNIVLNAACKN
TDFKLLTKLLEDKYEMILSEELSLIAIQGPKAVQILEKIINGVSDLKFMNGDTFNYLKEDI
YITRSGYTGEDGFEISIKNENAEVFVQKLIDEGANLIGLGARDTLRLEAGLCLYGHDMDI
NKSPVEANLKWAISKNRILEGGFIGCEKIKSQIEKGVSKIRVGIKPEGRIIAREKTSIFSEDD

KNIGEITSGTFGPSVQAPVAMGYVENSFSKIDTKVFLEVRGKKYPAIISNLPFYKKSYVK
GASK

>gi|254456244|ref|ZP_05069673.1| PARALOG glycine cleavage T protein [Candidatus Pelagibacter sp. HTCC7211] /dmdA_Paralog/
MSNKNFGFGTQIRKSPYFDSTVKWGATGFSVYNHMYIPRDFGNPEQNFWNLIQTAILCD
VAVERQVEITGPDAYKFIQLLTPRDLSKLAIGQCKYVLITNNDGGILNDPVLLRLAENHF
WLSLADSDVLLWAQGVAVNSGLNVQIKEPDVSPLQLQGPNSGEIMVKLFGEGIRELKY
YWLREYDLDGIPLIVSRTGWSSELGYEIYLRDGSKGNELYEKIMEAGKTHGLQPGHTSSI
RRIEGGMLSYHADADINTNPFELGLDRLVNLDADINFVGKDALKKIKQDGIKRKQIGIEI
DCEPLKGPNTTFWELQKDNKIIGKVTSAVYSPRLKKNIALAMVEIQQTEIGNKFEVISNE
GKFNCTVVEKPFYDPKKKIASSS

>gi|254455529|ref|ZP_05068958.1| PARALOG aminomethyltransferase, putative [Candidatus Pelagibacter sp. HTCC7211] /dmdA_Paralog/
MGFQINDITGLKFRFNLLKNYMHKSLRNIRFSIKPQQEESGRPVELARTMSIHPLTYQELP
YDPEYSHYAGRLTTEKLSNATPDEQYWKTKREIILRHTGEHPYEISGPDALKLLQRIFPR
DISKVKKGRCSYQFACYHDGGIITDGLLLRIDENCYWFAQADGDMLSWYKANSEGLDV
EIKEPNVFVSQIQGPKSMELLDQLIDEPIANTWKYFDWVEITMANEKVIISRTGFTNELG
WEIYFRPENDAEKLGNLILENGKKMGMIITATPSFRGRRIEAGLLSAGQDFSNETNPFSV
GLGRFVDLKKDNFIGKKALLNADKECRSWGIRVVDGIAKKGRYIKINNQSIGKITSSTWS
PYQVCGVGIVLLDKSDIRPGTVVDVECTDEKIHKAELCKLPMYDPKGEIVRGINKKIPTK
AEPWSGIKN

>gi|119503920|ref|ZP_01626002.1| PARALOG aminomethyl transferase family protein [marine gamma proteobacterium HTCC2080] /dmdA_Paralog/
MVMTNQPSQEFGFGTQIRKSPYFDATVRWGAQSFSVYNHMYIPRDFGDPEQNFWNLVN
TAILCDVAVERQVQITGPDAARFVQLLTPRDLSKLAVGQCKYVMITNNDGCILNDPVLL
RLAEDKFWLSLADSDILLWAQGVAVNAGMDVHICEPDVSPLQLQGPNSGEIAKVLFGD
DIADLRYYWLREYTLDGIPLIVSRTGWSSELGYEIYLLDGSRGDDLWEAIMAAGEPFGL
KPGHTSTIRRIEGGMLSYHADMDNQTNPFEVGLGHWAAIDTDLEFVGKAALTAIRDAG
VTRQQVGLEIDGEALPAPNTRFWELSVDEAPVGKVTSAVYSPRLKKNIALAMVDCAAA
ALGTEIAVAMPDGVRLATVVEKPFYDPKKQITAASLSAATAPSAV

>gi|499600944|ref|WP_011281678.1| cupin [Candidatus Pelagibacter ubique HTCC1062] /functionally ratified dddK/
MIFVKNLASVLSQEWSSTEKYPGVRWKFLIDADFDGSSGLSLGFAEIAPGGDLTLHYHSP
AEIYVVTNGKGILNKSGKLETIKKGDVVYIAGNAEHALKNNGKETLEFYWIFPTDRFSEV
EYFPAKQKSG

>gi|654569998|ref|WP_028037226.1| cupin [Candidatus Pelagibacter ubique HTCC9022] /functionally ratified dddK/
MIFVNNLKSVSDQEWSSTEKYPGVRWKFLIDADYTKSSGLSLGFAEIAPGGDLTLHYHS
PAEIYVVTNGTGILNKSGQLEEIKKGDVVYIAGNAKHALKNNGKETLEFYWIFPTDRFSE
VKYLS

>gi|504765971|ref|WP_014953073.1| cupin [alpha proteobacterium HIMB5] /functionally ratified dddK/
MIFIKNMNSVSDQDWTTSEKYPGVRWKFLIDEDYNGSKGLSCGFAEIEPGGNLTLHHHA
PDEIYVVTNGSGTLNKSGELEEIKKGDVVYIAGNAKHALQNNGKEVLGFYWVFPTNKF
KDVEYISDE

>gi|496746257|ref|WP_009359929.1| cupin domain-containing protein [alpha proteobacterium HIMB114] /dddK_Paralog/
MKQIKEKIFKEQKIKPIKRFGSVKTKIFINKKQGSKKMISGITIIPQNKSINLHYHNCEEAV
MILEGTAIAEINKKKYILKKGEVSWIPAKIPHRFMNKKKEKLKIYWTYANANATRTDVL
TDKTNKILNEHK

>gi|503460787|ref|WP_013695448.1| mannose-6-phosphate isomerase [Candidatus Pelagibacter sp. IMCC9063] /dddK_Paralog/
MLNNKPAIFKEKKIKSIKRFGTVVTKIFVNKNSGSKSMISGTTLIPKDKSINLHYHNCEEA
VLVLKGTALAEINKKKYTLKEGEACWIPAKVPHRFINNNKSNLKIYWTYANVNATRTD
VLTKKTYKILDEHKKKL

>WP_009813101.1 peptidase M24 [Roseovarius nubinhibens] /functionally ratified dddP/
MNQHYSETRKIDPSRGATLGDNTPNDNNRIEIGPTQLAFGEWATAGLALPDLQRMREFR
WNRLTQAVVDRDYGGVLMFDPLNIRYATDSTNMQLWNAHNPFRALLVCADGYMVIW
DYKNSPFLSTFNPLVREQRFGADLFYFDRGDKVDVAADAFSNEVRTLIAEHGGGNMRL
AVDKIMLHGLRALEAQGFEIMEGEELTEKTRAIKGPDEILAMRCAVHACETSVAAMEHF
AREAVPQGNTSEDDVWAVLHAENIKRGGEWIETRLLASGPRTNPWFQECGPRIIQNNEII
SFDTDLIGSYGICVDISRSWWVGDAAPPADMVYAMQHAHEHIMTNMEMLKPGVTIPEL
SERSHRLDEQFQAQKYGCLMHGVGLCDEWPLVAYPDQAVPGSYDYPLEPGMVLCVEA
AVGAVGGNFTIKLEDQVLITETGYENLTSYPFDPALMGR

>WP_015494462.1 peptidase M24 [Octadecabacter arcticus] /dddP/
MNTHYRDTRKIDPSKGSVLGDGSPNDNDRVEIGPTQLAFGEWDTAGLVLPNLQNMREY
RWQRLTQHIVDRGYGGLLMFDPLNIRYATDSTNMQLWNTHNPFRAVLLCADGYMVIW
DYKISPFLSAFNPLVRERRSGASMFYFSNGNKGLQAASTFADEVKDIMGEHAGTNTRLA
VDKIMVDGLRALEARGFEVMEGEEVTEHARSIKGVDEILAMRCANHACETAVKVMEDF
ARNRSGNGITSEDDIWSVLHGENIKRGGEWIETRLLASGPRTNPWFQECGPRITQPNEILA
FDTDLIGSYGICIDISRTWWIGDEKPRPDMVEAMKHGVEHIETNMQMLKPGVNIQDLSR
NTHVLDAKYQKQKYGCLMHGVGLCDEWPLVAYPDSMVDGAFDHELKAGMVLCVEA
LLGEEGGDFSIKLEDQVLITEDGFENLTTYPHDDALMGR

>WP_014880246.1 peptidase M24 [Phaeobacter inhibens DSM 17395] /dddP/
MSSDTFETMSNTEPEMNEHYRDTRKIDPTRGATLGDNTPNDQDRVEIGPTQLAFGEWA
AAGLQLPDLQAMRRYRWERLTRFINDRDYAGLLVFDPMNIRYATDSTNMQLWNTHNP
FRALLICADGYMVMWDYKQAPFLSEFNPLVREQRAGADLFYFDRGDKVDVAADAFAN
EVRTLLAEHSGGNTRLAVDKIMLHGLRALEAQGLEVFPGEELTEKCRAVKGPDEILAMR
CANHACETTVAEMERYARSAIPGGQISEDDVWAVLHAENIRRGGEWIETRLLTSGPRTN
PWFQECGGRIIQNNEIISFDTDLVGSYGICIDISRSWWIGDRAPPADMVYAMQHGVEHIQ

SNMEMLKPGVNLQELSRNCHLLDAQFQKQKYGCMMHGVGLCDEWPLVAYPDAMVE
GAFDYDLEPGMVLCVEALVSPEGGDFSIKLEDQVLITETGYENLTTYPFDPALMGTTR

>WP_008033539.1 peptidase M24 [Rhodobacterales bacterium HTCC2255] /dddP/
MINVKYDTFAVFNKAKFTINSNETVFDGINKMQNKYSEIRKIDPSQGQFLVDGTPNNSN
RVEIGPTLLAINEWKKAGLVQPNLTKMREYRWQRLTQHIVDRDWGGLLMFDPLNIRYA
TDSTNMQLWNTHNPFRAVLLCADGYMVIWDYKNSPFLSSFNSLVKEQRSGADLFYFDR
GDKIDVAADLFSSEITELITEHGGRNMNLGLDKGMIHGIRALEAQGFEIMDGEECTEKCR
SIKGPDEILAMKCASHSCELSIHEMQNKISIGMSEDAIWAELHKSNIARGGEWIETRLLTT
GPRTNPWFQECGPRQLQNNEILAFDTDLIGCYGFCIDVSRTWWIGNEKPRADMVYAMQ
HAHEHIMTNMEMLKPDTPFRDLTFNGHQLDSQYDKGKYSCRFHGVGLCDEWPLISYSD
NFIDGAFDYKLKAGMVLCVEALVSPEKGDFSIKLEDQVLVTEDGFENITKFPFCPHLMG
ET

>WP_011454901.1 peptidase M24 [Jannaschia sp. CCS1] /dddP/
MNQAYRRNVRKIDPTKGVMLPDGTLNDNDRIEIGPTALAYAEWAAAGLTLPNLQTMRE
YRLGRLVGQLQERDLAGVLMTDPLNIRYATDATNMQLWNTHNPFRACLVCADGHMV
LWEYKNAPFLAEHNPLVREIRSGASMFYFTAGDRGDAVAETFSGEVADLLREHAGTNT
RLAVDKIMLHGARALEARGLTVSDGELVTEHARKIKCADEILAMRCAVDACEKSLKAM
EDAIEPGKSEDEIWAVLHAENIKRGGEWIETRLFSSGPRTNPWFQECGPRTLLPNEISALD
TDLIGCYGLCVDISRTWWTGPEKPRPDMIEAMQHAHEHIMVNMDRLKPGRSINDLVHN
GHRLADKYWARKYSCQMHGVGLCDEWPHVGYPDHHHDDAFDYVLEPGMMLCVEAL
VGEEGGDFCIKLEDQVLITEDGYENLTTYPFDAALMGAS

>WP_013046297.1 peptidase M24 [Candidatus Puniceispirillum marinum] /dddP/
MNQLIVGSNRKIDPTRRLHLKPDNTPDDNDRVEIGPTALAFEEWKQLGLTAPDMPALRA
YRLERLQQQIRIHDCAGLLLFDPLNIRYATDATNMQLWTSHNMARACFVPPEGKMILW
DFHNCEHLSAHLPLVGELRGGASFFYFETGDRTAEAAKAFADQMLDIMHHYAPGNKRL
AVDKMENLGYAALVGLGVEVLEGQVLTEHGRSIKNENELNAMRCAIATCELAVEEMR
DEMRAGISENELWATLHAGNIKRGGEWIETRILSSGPRTNPWFQECGPRIMQDGDLMAF
DTDLVGTYGYCCDISRTWLVGDGSPSDAQKHLYQVAYDHVMTNIGLIEPGMRFADMT
RIAHRLPEEYRALRYGVLAHGVGLCDEYPSVRYPEDVEHHGYGGCFEVGMTLCVEAYV
GAVGGRDGVKLEEQVVVTDQGAIPLSTYRYEDAFLS

>WP_053819505.1 peptidase M24 [Candidatus Thioglobus singularis] /dddP/
MSFTSAKRHHAKIGSHLKGEEDIYSLNKHALGPGELAESEWLEAGLANPDMTKIREYRL
KRVREKLVEFDCAGILLYDPLNIRYATDSTNMSLWTSHNAARYALVMTDGPVIIFEFDA
HDFLSNHNPLITEVRHAVTYLYFTAGDKSKERAKIWASEIVDIVREYGKGNKKLALDHC
APEGIHELQSNGLELANGEEVMELARLIKSDDEMKAMRRSIFSCEKSMELMRNHFKPGI
TEQELWSRFQMEAVSRGAEWIETRLLASGPRANPWYQECSSRPILSGELMGFDTDLVGS
YGYCTDMSRTWLCGDEKATDEQKEIYTMGYEQIQNNMKLLKPGVTFKELTLNAKEYS
KQEFRHYSVLFHGVGLCDEFPAIPFSWELNENSFDGVLQPGMVLCVETYVGRFSGGPGV
KLEEQVLVTEGGHELLTNYPFETELLI

>gi|261250753|ref|ZP_05943327.1| PARALOG hypothetical protein VIA_000771 [Vibrio orientalis CIP 102891] /dddP_Paralog/

MKRIVDQLQARDLAGVLLFDPLNIRYATDSTNMQLWIAHNHARACFVSAEGYMILWDF
HNCEHLSAHLPLVKEVRNGASFFYFETGNRTNEHAHHFAKEIADIVKQYGGGSNRIAVD
KIEIVGLRELDKQGLELFDGQEVMELARAVKNIDEINAMRCSIASTEIAMKKMQEATVP
GVTENDIWSVLHAENIKRGGEWIECRILSSGPRTNPWFQECGPRVVKEGELLAFDTDLIG
PYGFCADLSRTWLIGDVEATEEQRHLYRVAYEHIQHNMEILKPGMTFEEVTRSGLLLPE
KYRPQRYGVMMHGVGLCDEYPSIRYPEDLEGHGYDGVLEPGMALCVEAYVGAVGGN
EGVKLEDQVIITEDGFENLTNYPFEKELLK

>gi|226311600|ref|YP_002771494.1| PARALOG Xaa-Pro dipeptidase/Xaa-Pro aminopeptidase [Brevibacillus brevis NBRC 100599] /dddP_Paralog/
MFQERISKLHTFLTEQELNAVLITSPKHVYYLTGFFTDPHERFMGLVIPAEGKPSLIVPAL
DREAAAEASFVQDIHTHTDIQNPYEILKQVLPANLAKLGIEKSHMTVERYEALGQVVLA
SSYVDVEEPLREMRLIKSADEVDRLKHAVQLVEDSLRETLKKVKTGMTETEIVAELEFQ
MKRLGAEGPSFTSMVLAGEKSALPHGKPGTRQVQEGDLLLFDIGVAANGYVSDITRTFA
VGKISAQLQEIYETVLAANEAAIAEIRPGVTFAHLDKTARDVITAKGYGEYFMHRLGHG
LGMDVHEYPSVHSQNQEVLRPGMVFTIEPGIYLPGVGGVRIEDDVLVTETGVEILTQFPK
KLTSINQ

>gi|262275658|ref|ZP_06053467.1| PARALOG aminopeptidase YpdF (MP- MA- MS- AP- NP-specific) [Grimontia hollisae CIP 101886] /dddP_Paralog/
MSIVKPTEPEIIRESDIEPGWDWSKRIPAPGRMSVDFEQRVDFNRLHRYRVGRARDALKN
SGLGAVLCFDNNNIRYLTSTVIGEWARDKIARYTLFTGNSDPYLWDFGSAAKHHQLYQ
GLIQPEHFKAGMLGLRGSVAKEAGLFKNAAKDIKALLVEEGVADMPLGIDVCEKPMLE
ALEAEGIEVRDCQQVMLEARQIKSMDEVVLLNMAATMVDGAYHQLAENLKPGKRENE
SVADANKFLYDNGSDDVEAINAVSGERGSPHPHNFTDRMYRPGDQAFFDIIHSFMGYRT
CYYRTLNVGSASQAQQDAYKQAREWIDAAIDLIRPGMTTDKIAAVWPKAEQFGFASEM
EAFGLQFGHGLGLALHERPIISRLVSMENPFELQEGMVFALETYCPSADGNGAARIEEEV
VVTADGCEIITLFPAQELFIANKY

>gi|399993476|ref|YP_006573716.1| PARALOG metallopeptidase, family M24 [Phaeobacter gallaeciensis DSM 17395] /dddP_Paralog/
MSERPGNALMPNLLTPMDLEPNWEWRDKLPAHGHMSVDFERRIDHDRLRRYRLARTR
QSLKNSNAGTLLLFDVNNIRYVSATKIGEWERDKMCRFCLLTGDDSPYVWDFGSAAEH
HKRHSDWLEPSHCLAGVVGMRGTIPPEFGLMKKYAKQIAGLIRDAGMADMPVGVDYA
ETAMFHALQEEGLNVVDGQQIMLAAREIKNTDEIQLLTQAAAMVDGVYHMIYEELKPG
VRENDIVALSNKMLYEMGSDDVEAINAISGERCNPHPHNFTDRLIRPGDQAFFDILQSYQ
GYRTCYYRTFNVGRATPSQNDAYTKAREWIDASIAMIKPGVSTDKVAEVWPTAQELGF
ASEDQAFGLQFGHGLGLALHERPIISRAVSMDHPMEIQTGMVFALETYCPATDGYSAAR
IEEEVVVTETGCEVISLFPAEELPIANRY

CHAPTER 3

MICROBIAL METAGENOMES AND METATRANSCRIPTOMES DURING A COASTAL PHYTOPLANKTON BLOOM [2]

[2] Nowinski, B, Smith, CB, Thomas, CM, Esson, K, Marin, R, Preston, CM, *et al*. 2019. *Scientific Data* 6(1): 129.

Reprinted here with permission of the publisher.

Abstract

Metagenomic and metatranscriptomic time-series data covering a 52-day period in the fall of 2016 provide an inventory of bacterial and archaeal community genes, transcripts, and taxonomy during an intense dinoflagellate bloom in Monterey Bay, CA, USA. The dataset comprises 84 metagenomes (0.8 terabases), 82 metatranscriptomes (1.1 terabases), and 88 16S rRNA amplicon libraries from samples collected on 41 dates. The dataset also includes 88 18S rRNA amplicon libraries, characterizing the taxonomy of the eukaryotic community during the bloom. Accompanying the sequence data are chemical and biological measurements associated with each sample. These datasets will facilitate studies of the structure and function of marine bacterial communities during episodic phytoplankton blooms.

Background & Summary

In pelagic marine ecosystems, a major proportion of primary production is transformed by heterotrophic microbes on the scale of hours to days(Williams, 1981; Azam *et al.*, 1983; Moran, 2015). Much of this rapidly-processed primary production is made available in the form of dissolved organic carbon (DOC), released from phytoplankton by direct excretion or through trophic interactions. Bacterial uptake of DOC produces living biomass and regenerates inorganic nutrients(Azam *et al.*, 1983).

Monterey Bay is a coastal ecosystem with high primary production driven by frequent upwelling of nutrient-rich waters(Pennington and Chavez, 2000; Ryan *et al.*, 2009). Intense phytoplankton blooms can develop(Schulien *et al.*, 2017), and these vary dynamically in terms of taxonomic composition. In 2016, the fall phytoplankton bloom (Fig. 3.1) was dominated by an unusually intense bloom of the dinoflagellate *Akashiwo sanguinea*(Wells *et al.*, 2017). *A.*

sanguinea cell abundances reached 4.9 x 10^6 cells L^{-1}, and chlorophyll *a* concentrations reached 57 µg L^{-1} (at ~6 m depth) over the period spanning mid-September to mid-November. Here we present metagenomic, metatranscriptomic, and iTag data on the bacterial and archaeal communities during a 52-day period spanning this unusual plankton bloom in Monterey Bay (Table 3.1). iTag data on the eukaryotic microbial communities provides contextual information on community dynamics of the bloom-forming phytoplankton and grazer communities.

Methods

Sampling Protocol

From September 26 through November 16, 2016, microbial cells were collected at Monterey Bay station M0 for sequence analysis. A moored autonomous robotic instrument, the Environmental Sample Processor (ESP)(Scholin *et al.*, 2006), filtered up to 1 L of seawater sequentially through a 5.0 µm pore-size polyvinylidene fluoride filter to capture primarily eukaryotic microbes, which was stacked on top of a 0.22 µm pore-size polyvinylidene fluoride filter to capture primarily bacteria and archaea (Table 3.1). The samples were collected between 5 and 7 m depth at approximately 10 a.m. PST. Samples were collected daily except during October 7 – November 1 when the ESP was offline for repair. ESP filters were preserved with RNAlater at the completion of sample collection and stored in the instrument until retrieval. While the ESP was offline, grab samples were collected by Niskin bottle at the M0 mooring site 2-3 times per week, with time of sampling, depth of sampling, and filters the same as for the ESP samples except that filters were flash frozen in liquid nitrogen.

Environmental data (temperature, salinity, chlorophyll *a* fluorescence, light transmission, and dissolved O_2 concentrations) were collected by a CTD instrument mounted with the

ESP(Moran, 2019). Additional environmental data were obtained from grab samples collected at the M0 mooring 2-3 times per week [total dimethylsulfoniopropionate concentration (DMSPt), dissolved DMSP concentration (DMSPd), DMSPd consumption rate, chlorophyll *a,* and cell counts by flow cytometry and microscopy](Moran and Kiene, 2019; Nowinski *et al.*, 2019) (Online-only Table 1).

DNA/RNA Extraction

Total community nucleic acids for metagenome, metatranscriptome, and 16S iTag sequencing were obtained from the same 0.22 µm filter (0.22 - 5.0 µm size fraction) using the ZymoBIOMICS DNA/RNA Miniprep Kit (Zymo Research, Irvine CA). At extraction start, internal standards were added to the lysis buffer tube (see Usage Notes), and the filter was cut into small pieces under sterile conditions to facilitate extraction. RNA was treated according to the manufacturer's instructions with in-column DNase I treatment. After elution, RNA was treated with Turbo DNase (Invitrogen, Carlsbad CA) and concentrated using Zymo RNA Clean and Concentrator (Zymo Research). Except for a few cases of low nucleic acid yields, duplicate filters were sequenced for each sample date.

DNA for 18S rRNA gene sequencing was extracted from the 5.0 µm filters using the DNeasy Plant Mini Kit (Qiagen, Venlo NL) with modifications. Filters were cut into pieces and added into a prepared lysis tube containing ~200 µl of 1:1 mixed 0.1 and 0.5 mm zirconia/silica beads (Biospec Products, Bartlesville, OK) and 400 µl Buffer AP1. Internal standards (see Usage Notes) were added just prior to extraction. Three freeze-thaw cycles were performed using liquid nitrogen and a 65 °C water bath. Following freeze-thaw, bead beating was performed for 10 min, followed by centrifugation at 8,000 rpm for 10 min to remove foam. Following centrifugation,

45 µl of proteinase K (>600 mAU/ml, solution, Qiagen) was added to each tube and incubated at 55 °C for 90 min with gentle rotation. Filters were then removed and the tubes incubated at 55 °C for 1 h. The DNeasy kit protocol was resumed at the RNase A addition step. Final DNA was eluted in 75 µl of diluted (1:10) TE buffer.

Metagenome Sequencing and Analysis

Sequence data were generated at the Department of Energy (DOE) Joint Genome Institute (JGI) using Illumina technology. Libraries were constructed and sequenced using the HiSeq-2000 1TB platform (2x151 bp). For assembly, reads were trimmed and screened, and those with no mate pair were removed using BFC (v r181)(Li, 2015). Remaining reads were assembled using SPAdes (v 3.11.1)(Bankevich *et al.*, 2012). The read set was mapped to the final assembly and coverage information generated using BBMap (v 37.78)(Bushnell, 2014) with default parameters. Assembled metagenomes were processed through the DOE JGI Metagenome Annotation Pipeline (MAP) and loaded into the *Integrated Microbial Genomes and Microbiomes* (IMG/M) platform(Huntemann *et al.*, 2016; Chen *et al.*, 2018).

Metatranscriptome Sequencing and Analysis

Sequence data were generated at the DOE JGI using Illumina technology. Libraries were constructed and sequenced using the HiSeq-2500 1TB platform (2x151 bp). Metatranscriptome reads were assembled using MEGAHIT (v 1.1.2)(Li *et al.*, 2016). Cleaned reads were mapped to the assembly using BBMap.

16S and 18S iTag Sequencing and Analysis

Sequence data were generated at the DOE JGI using Illumina technology. Primers 515FB(Parada *et al.*, 2016) (5'-GTGYCAGCMGCCGCGTAA) and 806RB(Apprill *et al.*, 2015) (5'-GGACTACNVGGGTWTCTAAT) were used for 16S rRNA gene amplification, and primers 565F (5'-CCAGCASCYGCGGTAATTCC) and 948R (5'-ACTTTCGTTCTTGATYRA) were used for 18S rRNA gene amplification(Stoeck *et al.*, 2010). Libraries were constructed and sequenced using the Illumina MiSeq platform (2x301 bp). Contaminant reads were removed using the kmer filter in BBDuk, and filtered reads were processed by the JGI iTagger (v 2.2) pipeline (https://bitbucket.org/berkeleylab/jgi_itagger).

To generate an overview of microbial community composition during the bloom (Fig. 3.2 and 3.3), the 16S and 18S rRNA amplicon libraries (raw reads) were primer-trimmed using Cutadapt (v 1.18)(Martin, 2011) and analyzed using QIIME2 (v 2018.6)(Bolyen *et al.*, 2018). The DADA2(Callahan *et al.*, 2016) plugin in QIIME2 was used to generate exact sequence variants (ESVs), which were classified using the QIIME2 naive Bayes classifier trained on 99% Operational Taxonomic Units (OTUs) from the SILVA rRNA database (v 132)(Quast *et al.*, 2012) after trimming to the primer region. Taxonomic bar plots were generated using QIIME2.

Code Availability

Software versions and parameters used are as follows:

BFC v r181

MEGAHIT v 1.1.2: --k-list 23,43,63,83,103,123

SPAdes v 3.11.1: -m 2000, -k 33, 55, 77, 99, 127 –meta

BBDuk v 38.08 for 16S, v 38.06 for 18S

BBMap v 37.78

iTagger v 2.2

For 16S iTags:

Cutadapt v 1.18: --interleaved -g GTGYCAGCMGCCGCGGTAA -G

GGACTACNVGGGTWTCTAAT -m 275 --discard-untrimmed

QIIME2 v 2018.6:

qiime dada2 denoise-paired \

--p-trunc-len-f 210 \

--p-trunc-len-r 181

For 18S itags:

Cutadapt v 1.18: --interleaved -g CCAGCASCYGCGGTAATTCC -G

ACTTTCGTTCTTGATYRA -m 275 --discard-untrimmed

QIIME2 v2018.6:

qiime dada2 denoise-paired \

--p-trunc-len-f 259 \

--p-trunc-len-r 200

Data Records

The raw Illumina sequencing reads for metagenomes, metatranscriptomes, and 16S rRNA and 18S rRNA iTags are available from the NCBI Sequence Read Archive(Institute, 2018). Contigs assembled within each individual metagenome and metatranscriptome are available from the JGI Integrated Microbial Genomes portal (Online-only Table 2). Chemical and biological data associated with each sample are available at the Biological and Chemical Oceanography Data Management Office (BCO-DMO)(Moran, 2019; Moran and Kiene, 2019). Measured

parameters include temperature, salinity, depth, light transmission, concentrations of dissolved oxygen and chlorophyll, concentration and consumption rates of DMSP, and cell counts for heterotrophic bacteria, *Synechococcus*, *Akashiwo*, and photosynthetic eukaryotes.

Technical Validation

For metagenomic and metatranscriptomic Illumina data, BBDuk (version 37.95; https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) was used to remove contaminants, trim reads that contained adapter sequence, and trim reads where quality dropped to zero. BBDuk was used to remove reads that contained four or more 'N' bases, had an average quality score across the read <3, or had a minimum length ≤51 bp or 33% of the full read length. Reads mapped with BBMap to masked human, cat, dog and mouse references at ≥93% identity were separated into a chaff file. Reads aligned to common microbial contaminants were also separated into a chaff file. For metatranscriptomic data, reads containing ribosomal RNA and known JGI spike-in sequences were removed and placed into separate fastq files. The internal DNA and mRNA standards added for quantification purposes at the nucleic acid extraction step (see Usage Notes) were recovered at 0.5-5.0% of sequences as expected.

For 16S rRNA and 18S rRNA, BBDuk was used to remove contaminants and trim reads that contained adapter sequence. This program was also used to remove reads that contained one or more 'N' bases, had an average quality score across the read of <10, or had a minimum length ≤51 bp or 33% of the full read length. Reads mapped with BBMap to masked human, cat, dog and mouse references at ≥93% identity or aligned to common microbial contaminants were separated into a chaff file. The 16S and 18S rRNA reads amplified from the internal DNA

standards added for quantification purposes (see Usage Notes) were recovered at their expected level (0.5-5.0% of sequences).

Sequence datasets were checked for consistency with the expected composition of coastal marine microbial communities. Taxonomic assignments of 16S and 18S rRNA ESVs matched those of marine microbes common in coastal areas in general(Gifford *et al.*, 2013; Satinsky *et al.*, 2014) and in Monterey Bay seawater in particular(Nowinski *et al.*, 2019) (Fig. 3.2 and 3.3). Taxonomic assignments of protein-encoding genes from metagenomic datasets were likewise representative of coastal and Monterey Bay microbial communities, and had taxonomic assignments consistent with the iTag datasets.

Usage Notes

Sample processing included the addition of internal standards to allow for calculation of volume-based absolute copy numbers for each gene or transcript type (i.e., counts L^{-1} rather than % of sequence library)(Moran *et al.*, 2013; Satinsky *et al.*, 2013). The DNA standards consisted of genomic DNA from *Thermus thermophilus* DSM7039 HB8(Satinsky *et al.*, 2013) and *Blautia producta* strain VPI 4299 (American Type Culture Collection, Manassas, VA). mRNA standards consisted of custom-designed 1006 nt artificial transcripts(Satinsky *et al.*, 2013). Artificial transcript sequences are available at Addgene Plasmid Repository (https://www.addgene.org; products MTST5 and MTST6). All four standards (two DNA and two mRNA) were added to the 0.22 μm pore size samples at the initiation of nucleic acid extraction. In the case of 18S iTag samples, genomic DNA from *Arabidopsis* (BioChain Institute, Inc., Newark, CA) and *Mus musculus* (Millipore Sigma, Burlington MA) was similarly added to the 5.0 μm pore size samples at initiation of extraction. Added amounts of internal standards were estimated at ~1%

of final yields of DNA or mRNA based on prior recoveries from similar filters. Actual yields averaged ~2% of reads. The internal standards should be removed from the raw data prior to analysis. Information on how internal standards can be used for volume-based quantification is available elsewhere(Satinsky *et al.*, 2013; Lin *et al.*, 2019).

Environmental data collected in association with the nucleic acid samples are given in Online-only Table 1. Available data differ between sampling dates depending on whether sampling was done by the ESP, from Niskin grab samples, or both.

Acknowledgements

We thank B. Roman, B. Ussler, J. Figurski, C. Wahl, B. Kieft, S. Gifford, T. Pennington for sampling and protocol assistance, advice, and technical expertise; K. Selph for flow cytometric analysis, L. Ziccarelli for *A. sanguinea* microscopy analysis, J. Christmann and the crew of the R/V Shana Rae; the Moss Landing Marine Laboratories Small Boat Facility; J. Ryan for processing MODIS satellite images; and S. Sharma and C. Edwardson for bioinformatic assistance. This work was funded by NSF grants OCE-1342694, OCE-1342699, OCE-1342734, the DOE Community Science Program, and partial support from the David and Lucile Packard Foundation through funds allocated to the Monterey Bay Aquarium Research Institute (MBARI). The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

Apprill, A., McNally, S., Parsons, R., and Weber, L. (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology* **75**: 129-137.

Azam, F., Fenchel, T., Field, J.G., Gray, J., Meyer-Reil, L., and Thingstad, F. (1983) The ecological role of water-column microbes in the sea. *Mar Ecol Prog Ser* **10**: 257-263.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S. et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comp Biol* **19**: 455-477.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C., Al-Ghalith, G.A. et al. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. In: PeerJ Preprints.

Bushnell, B. (2014) BBMap: a fast, accurate, splice-aware aligner. In: Lawrence Berkeley National Laboratory, DOE Joint Genome Institute.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Meth* **13**: 581.

Chen, I.-M.A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J. et al. (2018) IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res*.

Gifford, S.M., Sharma, S., Booth, M., and Moran, M.A. (2013) Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J* **7**: 281.

Huntemann, M., Ivanova, N.N., Mavromatis, K., Tripp, H.J., Paez-Espino, D., Tennessen, K. et al. (2016) The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v. 4). *Stand Genomic Sci* **11**: 17.

Institute, D.J.G. (2018) Seawater microbial communities from Monterey Bay, California, United States. In. NCBI Sequence Read Archive

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K. et al. (2016) MEGAHIT v1. 0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**: 3-11.

Li, H. (2015) BFC: correcting Illumina sequencing errors. *Bioinformatics* **31**: 2885-2887.
Lin, Y., Gifford, S., Ducklow, H., Schofield , O., and Cassar, N. (2019) Towards quantitative microbiome community profiling using internal standards. *Appl Environ Microbiol* **85**: e02634-02618.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**: pp. 10-12.

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

Moran, M.A. (2019) Environmental data from CTD during the Fall 2016 ESP deployment in Monterey Bay, CA. In. Biological and Chemical Oceanography Data Management Office (BCO-DMO).

Moran, M.A., and Kiene, R.P. (2019) Environmental data from Niskin bottle sampling during the Fall 2016 ESP deployment in Monterey Bay, CA. In. Biological and Chemical Oceanography Data Management Office (BCO-DMO).

Moran, M.A., Satinsky, B., Gifford, S.M., Luo, H., Rivers, A., Chan, L.-K. et al. (2013) Sizing up metatranscriptomics. *ISME J* **7**: 237.

Nowinski, B., Motard-Côté, J., Landa, M., Preston, C.M., Scholin, C.A., Birch, J.M. et al. (2019) Microdiversity and temporal dynamics of marine bacterial dimethylsulfoniopropionate genes. *Environ Microbiol* **12**: 1687-1701.

Parada, A.E., Needham, D.M., and Fuhrman, J.A. (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18**: 1403-1414.

Pennington, J.T., and Chavez, F.P. (2000) Seasonal fluctuations of temperature, salinity, nitrate, chlorophyll and primary production at station H3/M1 over 1989–1996 in Monterey Bay, California. *Deep Sea Res Part II: Top Stud Oceanogr* **47**: 947-973.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-D596.

Ryan, J.P., Fischer, A.M., Kudela, R.M., Gower, J.F., King, S.A., Marin III, R., and Chavez, F.P. (2009) Influences of upwelling and downwelling winds on red tide bloom dynamics in Monterey Bay, California. *Cont Shelf Res* **29**: 785-795.

Satinsky, B.M., Gifford, S.M., Crump, B.C., and Moran, M.A. (2013) Use of internal standards for quantitative metatranscriptome and metagenome analysis. In *Meth Enzymol*: Elsevier, pp. 237-250.

Satinsky, B.M., Crump, B.C., Smith, C.B., Sharma, S., Zielinski, B.L., Doherty, M. et al. (2014) Microspatial gene expression patterns in the Amazon River Plume. *Proc Nat Acad Sci USA* **111**: 11085-11090.

Scholin, C., Jensen, S., Roman, B., Massion, E., Marin, R., Preston, C. et al. (2006) The Environmental Sample Processor (ESP)-an autonomous robotic device for detecting microorganisms remotely using molecular probe technology. *OCEANS 2006*: 1-4.

Schulien, J.A., Peacock, M.B., Hayashi, K., Raimondi, P., and Kudela, R.M. (2017) Phytoplankton and microbial abundance and bloom dynamics in the upwelling shadow of Monterey Bay, California, from 2006 to 2013. *Mar Ecol Prog Ser* **572**: 43-56.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D., BREINER, H.W., and Richards, T.A. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19**: 21-31.

Wells, B.K., Schroeder, I.D., Bograd, S.J., Hazen, E.L., Jacox, M.G., Leising, A. et al. (2017) State of the California Current 2016–17: Still anything but normal in the north. *CalCOFI Rep* **58**: 1-55.

Williams, P. (1981) Microbial contribution to overall marine plankton metabolism-direct measurements of respiration. *Ocean Acta* **4**: 359-364.

Table 3.1. Sequence datasets from the fall bloom in Monterey Bay, CA, 2016.

Source Name	Sampling Dates	Geographical Location	Sampling Method	Sequence Type	Sample Identifiers from GOLD (Gaxxx) or the JGI Portal (Project ID xxx)	BioProject Accession IDs from the NCBI SRA
Monterey Bay Station M0	September 26 - November 16, 2016	Monterey Bay, CA, USA, 36.835 N, 121.901 W	Autonomous collection by the Environmental Sample Processor and Niskin bottle sampling	All		Umbrella project PRJNA533622
Monterey Bay Station M0	September 26 - November 16, 2016	Monterey Bay, CA, USA, 36.835 N, 121.901 W	Autonomous collection by the Environmental Sample Processor and Niskin bottle sampling	Metagenomics	Ga0228601 - Ga0228678; GA0233393 - Ga0233402	PRJNA467720 - PRJNA467773, PRJNA468208 - PRJNA468214, PRJNA502407 - PRJNA502427, PRJNA502440 - PRJNA502442
Monterey Bay Station M0	September 26 - November 16, 2016	Monterey Bay, CA, USA, 36.835 N, 121.901 W	Autonomous collection by the Environmental Sample Processor and Niskin bottle sampling	Metatranscriptomics	Ga0228679 - Ga0232167; Ga0247556 - Ga0247607; Ga0256411 - Ga0256417	PRJNA467774 - PRJNA467774, PRJNA468143 - PRJNA468143, PRJNA468299 - PRJNA468332, PRJNA502451 - PRJNA502468, PRJNA502608 - PRJNA502612
Monterey Bay Station M0	September 26 - November 16, 2016	Monterey Bay, CA, USA, 36.835 N, 121.901 W	Autonomous collection by the Environmental Sample Processor and Niskin bottle sampling	16S rRNA iTags	JGI Project ID 1190879	PRJNA511156 - PRJNA511206, PRJNA511216 - PRJNA511252
Monterey Bay Station M0	September 26 - November 16, 2016	Monterey Bay, CA, USA, 36.835 N, 121.901 W	Autonomous collection by the Environmental Sample Processor and Niskin bottle sampling	18S rRNA iTags	JGI Project ID 1190880	PRJNA511207 - PRJNA511215, PRJNA511253 - PRJNA511331

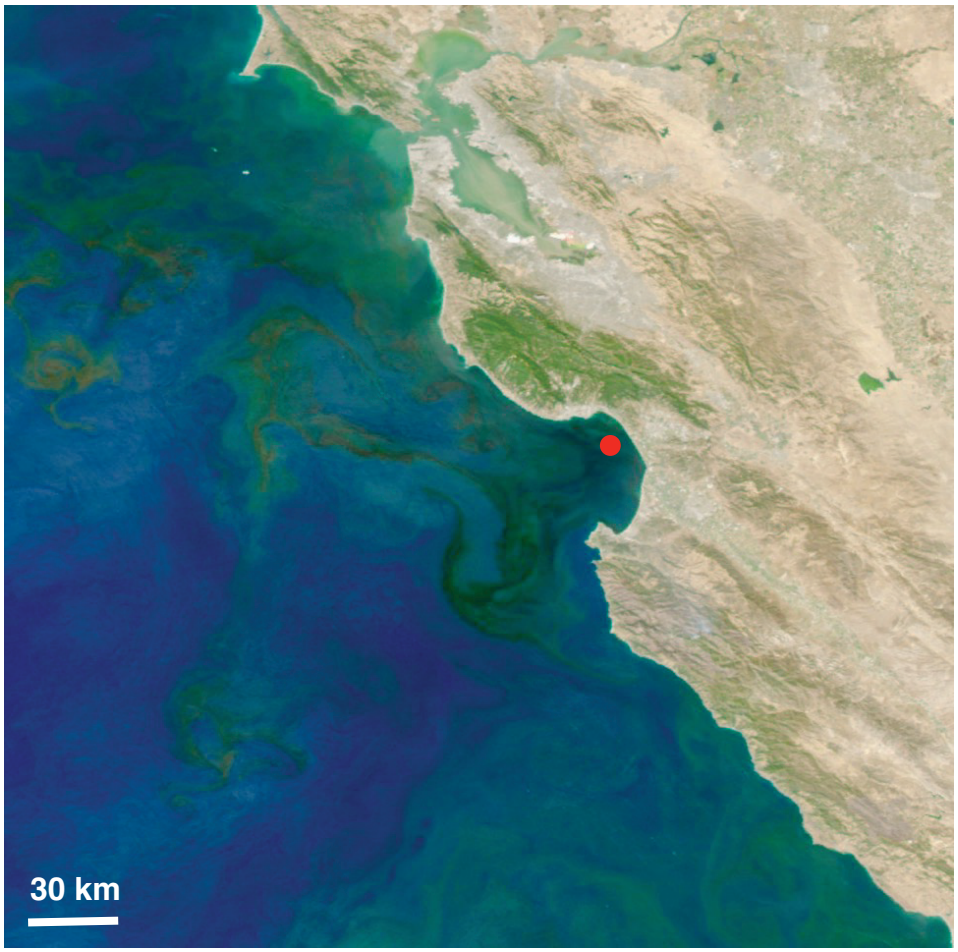


Figure 3.1 MODIS satellite image on September 26, 2016 of the phytoplankton bloom occurring in Monterey Bay and extending into the Pacifc. The red dot represents the sampling station M0, located at 36.835 N, 121.901 W.

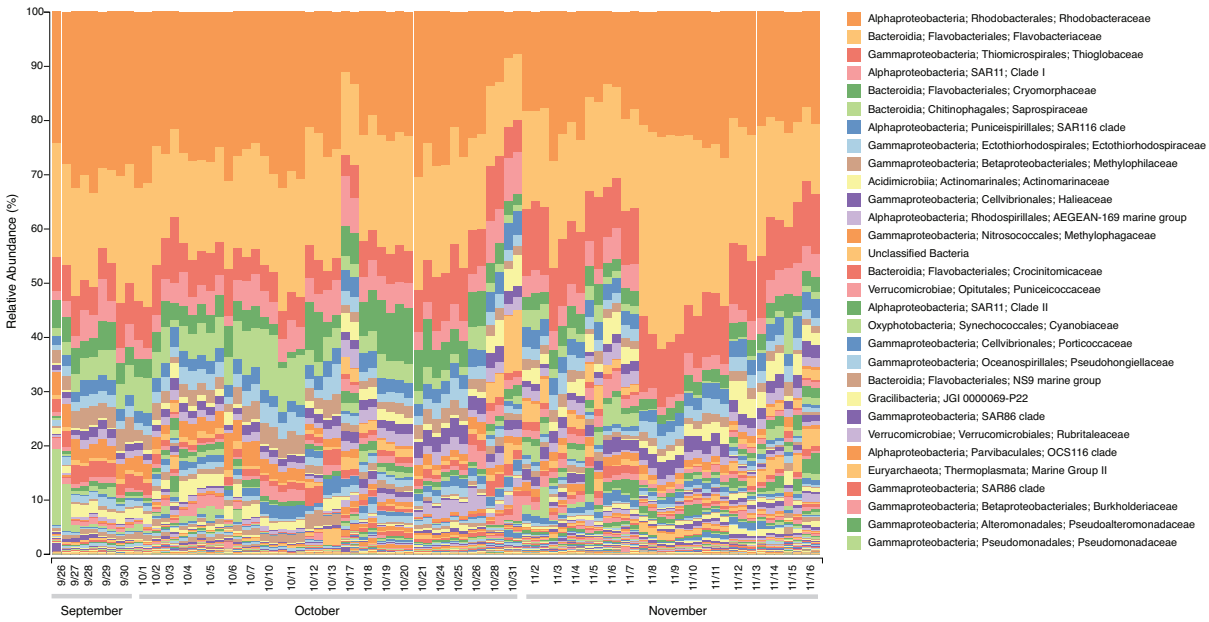


Figure 3.2. Relative abundance of bacterial and archaeal taxa at Monterey Bay station M0 during the fall of 2016. Samples were collected at ~6m, and 16S rRNA genes were amplified from community DNA in the 0.22 to 5.0 µm size range. Taxonomic groups were defined based on exact sequence variants using DADA2 in QIIME 2 (https://qiime2.org) and assigned taxonomy with the naive Bayes q2-feature-classifer trained using the 515F/806R region from 99% operational taxonomic units from the SILVA 132 16S rRNA database. Assignments of the 30 most abundant taxa are given at the family level.

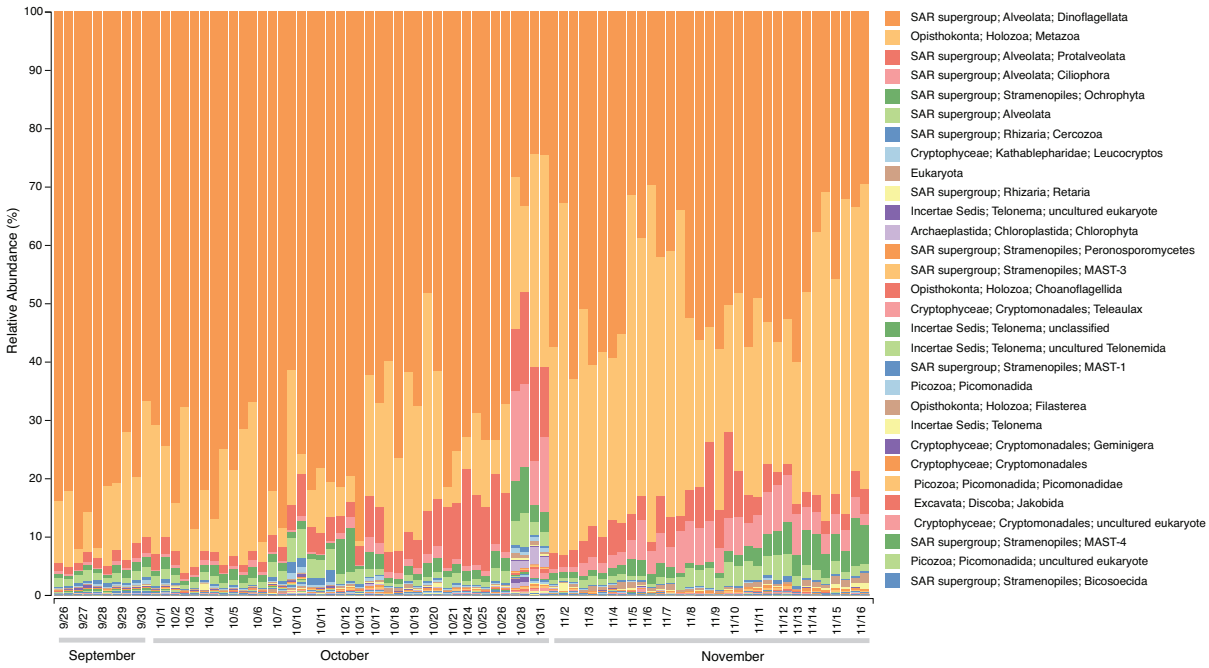


Figure 3.3. Relative abundance of eukaryotic taxa at Monterey Bay station M0 during the fall of 2016. Samples were collected at ~6m, and 18S rRNA genes were amplified from community DNA in the >5.0 µm size range. Taxonomic groups were defined based on exact sequence variants using DADA2 in QIIME 2 (https://qiime2. org) and assigned taxonomy with the naive Bayes q2-feature-classifer trained using the 565F/948R region from 99% operational taxonomic units from the SILVA 132 18S rRNA database.

CHAPTER 4

IDENTIFYING MARINE BACTERIAL NICHE DIMENSIONS BY AN EXPERIMENTAL INVASION [3]

[3] Nowinski, B and Moran, MA. Submitted to *Nature Microbiology*.

Abstract

Niche theory is a foundational ecological concept to explain the distribution of species in natural environments. However, identifying the key dimensions of an organism's niche is challenging because of the many factors in an environment that can affect viability, and particularly so for bacteria, whose potential metabolic functions and ecological adaptations are extraordinarily diverse. Here we used serial invasion experiments that introduced a well-characterized heterotrophic marine bacterium into a natural coastal phytoplankton bloom for 90 minutes on 14 dates during bloom progression. Changes in transcriptome composition were used to identify the bacterium's responses to its surroundings, indicative of its abiotic and biotic niche dimensions in this dynamic ecosystem. Over 43 substrate dimensions, 9 vitamin dimensions, 5 nutrient dimensions, and 4 metal dimensions affected the bacterium's viability. Biotic interaction dimensions were represented by 24 mechanisms including both antagonism and resistance. Although the peak bloom was characterized by favorable substrate dimensions, which typically control bacterial viability in coastal phytoplankton blooms, low apparent growth rate of the invading bacterium indicated that other dimensions had narrowed the bacterium's realized niche at the height of the bloom. Among the possible negative biotic dimensions were interactions with the dominant red tide dinoflagellate. These serial invasion studies with a marine bacterium detected a diversity of bloom conditions affecting survival, highlighting factors that determine where bacterial species survive and function in ocean environments.

Introduction

Heterotrophic bacteria are central players in Earth's carbon and nutrient cycles, shaping natural and human ecosystems through their metabolic and ecological functions. In the ocean,

bacteria mediate flux between major carbon reservoirs, control nutrient availability, and engineer the base of marine food webs. Many environmental factors govern these activities, including physical factors such as seawater mixing and sunlight (Carlson *et al.*, 2009; Palovaara *et al.*, 2014), chemical factors such as substrates and nutrients (Church *et al.*, 2000; Poretsky *et al.*, 2010), and biological factors such as mutualistic and antagonistic interactions with other species (Persson *et al.*, 2009; Yeung *et al.*, 2012). Together these factors form the dimensions of each bacterial species' niche, a foundational ecological concept in which the abiotic and biotic variables that influence birth and death rates of a species determine where it can exist in nature (Hutchinson, 1957; Colwell and Fuentes, 1975).

Identifying the dimensions of marine bacterial niches promises to improve understanding and modeling of the processes that drive ocean biogeochemistry. Hutchinson defined the 'fundamental' niche as the full range of external conditions in which an organism is viable (i.e., has an intrinsic growth rate, r_0, ≥ 0) in the absence of biotic interactions; and the 'realized' niche as the fundamental niche narrowed to the dimensions invoked in the presence of other species (Hutchinson, 1957). Yet while the niche concept has long been a useful ecological framework (Cohan, 2002; Erguder *et al.*, 2009), the myriad of possible environmental influences controlling ecological success of a species makes comprehensive analyses of niche dimensions difficult in practice. For marine bacteria, selected niche dimensions have been addressed by correlations to measured environmental factors (Meier *et al.*, 2017) and growth responses under defined laboratory conditions (Martens-Habbena *et al.*, 2009). Since basic rules of marine bacterial ecology are still being discovered, however, untargeted methodologies – those not limited to preselected factors – provide for a more inclusive accounting of features determining the balance of a bacterium's birth and death rates in natural environments.

Shifts in bacterial messenger RNA (mRNA) pools *in situ* represent untargeted proxies for the factors, or realized niche dimensions, invoking bacterial sensing and phenotypic response (Gifford *et al.*, 2013; Ottesen *et al.*, 2013; Landa *et al.*, 2017). In complex natural ecosystems, transcripts drawn randomly from multiple community members make it challenging to characterize niche dimensions of the many co-existing taxa. Mapping transcripts to genomes assembled from single cell sequencing or metagenomic data parses the aggregated community response (Galambos *et al.*, 2019; Nuccio *et al.*, 2020), although these reference genomes are typically incomplete and, when derived from metagenomic data, can blur responses through aggregation of multiple populations (Shaiber and Eren, 2019). The extent to which environmental stimuli are manifested in transcriptomes also differs among taxa, with bacteria that harbor few regulatory elements exhibiting only minor transcriptional shifts compared to those with well-regulated genomes experiencing identical perturbations (Cottrell and Kirchman, 2016; Landa *et al.*, 2017).

Here, we undertake the identification of bacterial niche dimensions using a novel variation of an invasion study, an approach typically employed in ecology to address principles governing the outcome of the invasion of an existing community by a foreign species (Mallon *et al.*, 2015; Bell, 2019). Our variation uses transcriptional responses by a heterotrophic marine bacterium to the experimental invasion of a natural phytoplankton bloom to identify factors that make up a bacterium's niche dimensions. On 14 dates over a 1-month period, the well-characterized, metabolically responsive bacterium *Ruegeria pomeroyi* DSS-3 was pre-grown using a standard protocol and added into surface seawater samples from Monterey Bay, CA, USA (the 'introduction' step of the invasion) (Mallon *et al.*, 2015). The additions occurred during the natural progression of a massive bloom of the dinoflagellate *Akashiwo sanguinea*

(Kiene *et al.*, 2019), an ecologically relevant environmental regime for this bacterium (González *et al.*, 2003; Anderson *et al.*, 2018; Anderson and Harvey, 2019). Transcriptome analysis of gene expression patterns of invading *R. pomeroyi* cells was conducted after 90 min and used to identify the abiotic and biotic factors of the shifting bloom environment that triggered phenotypic responses (the 'establishment' stage of the invasion) (Mallon *et al.*, 2015). From this, we present an untargeted window into the dynamic factors affecting the viability of a heterotrophic bacterium whose life history strategy is tied to the ecology of marine microphytoplankton (Luo and Moran, 2014).

Results and Discussion

Invasion manipulation

In the fall of 2016, surface waters of Monterey Bay, CA, USA hosted a large bloom of the dinoflagellate *A. sanguinea* (>50 mg chl *a* L-1) (Kiene *et al.*, 2019). Analysis of 18S rRNA gene sequencing and phytoplankton biomass calculated from cell counts show the protist community shifting over the course of the study to a mixture of diatoms and dinoflagellates, including dinoflagellate parasites. *A. sanguinea* accounted for >99% of phytoplankton biomass in late September but <20% in late October (Figs. 4.1a, 4.S1). A rapid decrease of *A. sanguinea* biomass on October 12 followed by recovery highlighted a distinct water mass moving through the sampling station.

On each of 14 dates between September 28 and October 31, the heterotrophic marine Rhodobacterales bacterium *Ruegeria pomeroyi* was pre-grown in the laboratory under a standard protocol (Fig. 4.1b; n = 3), washed free of culture medium, introduced into the Monterey Bay bloom community at 1 x 10^{10} cells L^{-1} seawater (2:1 ratio of *R. pomeroyi* to native heterotrophic

bacteria), and incubated in the laboratory under identical conditions of temperature (24ºC) and light (ambient laboratory). Gene expression patterns of *R. pomeroyi* were characterized by extracting and sequencing mRNA from the seawater communities 90 min after introduction. The sensitivity of the invasion protocol was assessed under controlled conditions by preparing *R. pomeroyi* cells as for the field samples but inoculating into defined minimal medium with and without 10 mM glucose (Fig. 4.1b). Transcriptomes from these controls verified robust *R. pomeroyi* expression responses 90 min post-introduction, and confirmed that the responses were consistent with conditions in the defined media (Fig. 4.1c). PCA analysis of genome-wide expression patterns indicated that the 90 min *R. pomeroyi* bloom profiles diverged from the laboratory controls and were distinctly non-random. Transcriptomes followed a trajectory that, with the exception of the Oct. 12 samples, tracked iteratively with date of sample collection (Fig. 4.2a) and showed a relationship with protist community composition (Fig. 4.2b). A growth rate index for *R. pomeroyi* (ribosomal proteins as % of transcriptome) was low in field samples (0.4 – 1.2%) and the no-glucose controls (1.9%), but elevated in glucose controls (6.1%).

Gene Expression Patterns

The 4,278 protein-encoding genes in the *R. pomeroyi* genome were categorized into 17 expression modules based on transcriptional patterns in the 14 independent introductions to bloom seawater (Fig. 4.2c). The largest module contained 1,417 genes that had maximum relative expression in the initial samples, corresponding to the peak of the bloom, and a decrease as the bloom aged (turquoise module); four additional modules were positively correlated with this module (Pearson's $R = 0.70 – 0.98$, $p<0.01$, d.f. = 12), and all were merged to create a super-module of 2,087 genes with highest expression in the peak bloom stages. The second largest

module contained 915 genes that had the opposite pattern – minimum relative expression under peak bloom conditions and increases over time (orange module); two additional modules were significantly positively correlated with this module ($R = 0.75$ and 0.76; $p < 0.01$) and all were merged to create a super-module of 1,423 genes with expression maxima in late bloom stages (Fig. 4.2c). Together, these two super-modules accounted for 82% of the *R. pomeroyi* genome (49% were peak bloom genes, 33% were late bloom genes) and were negatively correlated with each other (Pearson's $R = -0.91$, $p < 0.001$, d.f. = 12). The genes with maxima occurring during the peak bloom correlated positively with *A. sanguinea* biomass and chlorophyll *a* concentration; the late-bloom genes correlated positively with diatom biomass and dinoflagellate parasites in the Syndiniales clade (Fig. 4.S2).

Niche Dimension Analysis

We considered the aspects of niche theory that can be informed by sequential microbial invasion experiments. Hutchinson (1957) defined niche as the existing conditions in a specific geographic space that allow an organism to "survive and reproduce" (Colwell and Rangel, 2009), a criterion formalized as an intrinsic growth rate ≥ 0 (Holt, 2009). Although niches are genetically determined (Baltar *et al.*, 2019; Alneberg *et al.*, 2020), neither a genome nor transcriptome can delineate the Hutchinsonian niche because they do not themselves indicate whether growth is possible under existing conditions. For example, genomic data might indicate a microbe's capability for metabolizing a particular substrate, but not whether the supply of this substrate in the environment is sufficient to support growth. Genomes and transcriptomes do, however, indicate niche dimensions – the factors that have the potential to influence a species' growth in a given environmental space; genomes provide insights into the functions an organism

can invoke in reaction to a dimension, and transcriptomes indicate whether these functions are currently invoked (Muller, 2019). Niche theory also distinguishes between the 'fundamental' and 'realized' niche (Hutchinson, 1957), the former referring to the environmental conditions in which a species grows without consideration of other organisms, and the latter to the (typically narrower) set of conditions in which a species grows in the presence of other organisms, reflecting the outcome of competition for substrates or space, for example. Invasion experiments in which introductions are made into intact natural systems, as in this study, include dimensions imposed by competition, predation, and other biotic interactions and therefore represent realized niche dimensions of the invading microbe.

The realized niche dimensions influencing the viability of heterotrophic marine bacterium *R. pomeroyi* in the 2016 fall Monterey Bay phytoplankton bloom were operationally defined from the functional annotation of genes with significant relative expression changes through time. In other words, the subset of the bacterium's total niche dimensions whose variation elicited a response during the course of the phytoplankton bloom were identified from shifts in the bacterium's transcriptome. We focused first on the 3,510 genes of the peak-bloom and late-bloom super-modules because they had clear temporal patterns and accounted for the majority of the *R. pomeroyi* genome. For this group, a significant difference in relative expression between the two initial and two final time points was considered a measurable transcriptional response by the bacterium to an influential environmental signal (DeSeq2, adjusted $p<0.05$). This criterion narrowed the group to 1,382 genes. Eighteen percent of *R. pomeroyi* genes were not included in the super-modules. For these, a significant difference in relative expression between the two highest and two lowest time points was considered a measurable transcriptional response to an influential environment signal, which added 423 genes.

Chemical niche dimensions

Invasions into the phytoplankton bloom community identified at least 43 substrate-based niche dimensions for *R. pomeroyi*, recognized from transcriptional responses of genes transporting organic molecules that support bacterial growth. Another 13 potential substrate dimensions were suggested from transcriptional response of genes catabolizing organic molecules, although these may already be counted among the transporters that lack definitive substrate annotations. Nitrogen-containing compounds made up a surprising 67% of the substrate-based niche dimensions (29 compounds), and all but two of these invoked the highest expression levels when the bacterium invaded samples from the *A. sanguinea*-dominated peak bloom (Fig. 4.1a). These peak-bloom nitrogen-containing substrate dimensions included *N*-methyl compounds (TMAO, carnitine, choline, glycine betaine), amino acids and related molecules (polyamines, peptides), and sulfonated organic nitrogen compounds (taurine, *N*-acetyltaurine, cysteate). The two N-containing substrate dimensions that elicited highest expression later in the bloom were ectoine and a putative branched chain amino acid (Fig. 4.3). *R. pomeroyi* also reacted to five sulfur-containing substrates, including the three sulfonated organic nitrogen compounds given above plus dimethylsulfoniopropionate (DMSP) and isethionate, and these also elicited strongest responses when the bacterium invaded the peak bloom samples. Niche dimensions based on organic compounds with carbon-only backbones included the C1 molecules carbon monoxide and formate, six carboxylic acids including lactate, three sugars including ribose, and the aromatic compounds ferulate, catechol, and protocatechuate (Fig 4.3, Table 4.S1); again, *R. pomeroyi* reacted to all of these most strongly when invading peak bloom samples. The carbon-only compounds whose influence was the highest when the bacterium was inoculated into late bloom samples were those processed

through the ethylmalonyl CoA pathway for C2 substrate catabolism. Bacterial initiation of carbon storage as polyhydroxybutanoate in the late bloom indicated that the degree to which excess organic carbon accumulated in the bloom environment was a relevant dimension for *R. pomeroyi*'s ecological success (Fig. 4.3).

Other chemical dimensions of *R. pomeroyi*'s niche in the Monterey Bay bloom were based on nutrient and metal concentrations. Mirroring responses when cultured previously under nitrogen limitation (Chan *et al.*, 2012), the bacterium showed significant ammonium and urea uptake responses at the earliest two sample dates, suggesting nitrogen limitation during the peak bloom (Fig. 4.3). *A. sanguinea* biomass was positively correlated with the expression of these genes (Pearson's $R = 0.95$, $p < 0.01$, d.f. = 12; Fig. 4.S3), and given the dinoflagellate's high affinity for inorganic nitrogen and preference for ammonium (Kudela *et al.*, 2010), likely drew down nitrogen concentrations in the early bloom. Indeed, *R. pomeroyi* cannot take up nitrate and relies on ammonium as its inorganic nitrogen source (Moran *et al.*, 2004). *R. pomeroyi* also differentially expressed genes for acquisition of sulfate, phosphonate, and phosphate (Fig. 4.3). As was found for organic carbon, bacterial initiation of phosphorus storage as polyphosphate suggested that build-up of available phosphate in the late bloom environment was an operational niche dimension for *R. pomeroyi*. Four metals served as influential niche dimensions, with *R. pomeroyi* increasing expression of magnesium transport when introduced into the peak bloom, and manganese, iron, and zinc transport at later times (Fig. 4.3).

R. pomeroyi showed different patterns of gene expression for synthesis or utilization of six B vitamins and three co-factors depending on the stage of the bloom. Upon introduction into peak bloom conditions, it responded with increased transcription of genes linked to requirements for pyrroloquinoline quinone (coenzyme PQQ), and in late bloom conditions for thiamine,

nicotinamide, B3, riboflavin, pyridoxal phosphate (PLP), pantothenate, biotin, molybdopterin, and folate (Fig. 4.3). These responses may reflect external conditions that elicited a change in the bacterium's cellular requirements, or shifts in availability of these molecules or their precursors in bloom seawater.

Biotic interaction niche dimensions

On about half of the invasion experiment dates, a feature of the environment induced *R. pomeroyi* to synthesize indole acetic acid, a hormone that enhances growth of co-occurring phytoplankton (Amin *et al.*, 2015). On two dates in late September and early October, *R. pomeroyi* ramped up transcription of genes encoding a diffusible killing mechanism that targets diverse bacterial taxa (Sharpe *et al.*, 2019) (Fig. 4.3). The specific dimensions that drive transcription of these non-trophic biotic interaction genes after only a 90 min exposure to the bloom remain unidentified, but are likely to include the presence of specific protist or bacterial taxa in the invaded microbial community (Fig. 4.S1). Antagonistic genes encoded type I secretion and efflux systems for toxins and antibiotics, while resistance genes encoded antibiotic resistance and detoxification.

Differences in transcription of motility-related genes suggested that patchy distribution of deterrents or resources affected *R. pomeroyi* during the bloom. *R. pomeroyi* showed the greatest investment in building motility machinery when invading the bloom peak, and less as the bloom aged. The bacterium also initiated expression of pilus assembly genes, for attachment or conjugation, and quorum sensing genes, for cell-to-cell chemical signaling, and both had increased importance in the peak bloom invasions. *R. pomeroyi* harbors a gene transfer agent (GTA) system (Moran *et al.*, 2004) encoded by 16 genes that package random ~5 kb genome

fragments into virus-like particles and release them extracellularly to initiate intraspecific gene transfer (Biers *et al.*, 2008). Transcription of these genes was invoked on six consecutive invasion experiments in mid-October after a 90 min of exposure to the bloom microbial community, suggesting a persistent environmental condition triggering initiation of DNA transfer. Expression of this system highlights how the genetic basis of niche dimensions can evolve over relatively few generations (Vergin *et al.*, 2007; Gravel *et al.*, 2011), particularly for bacteria with GTA or other mechanisms enabling high horizontal gene transfer rates (McDaniel *et al.*, 2010).

Stress niche dimensions

On both peak- and late-bloom invasion experiment dates, *R. pomeroyi* encountered environmental conditions that elicited enhanced transcription of genes for repair and recombination of DNA, for heat shock proteins that refold damaged proteins (Nuss *et al.*, 2010), and for responding to oxidative stress (Fig. 4.3, Table 4.S1). The *R. pomeroyi* genome has two σ-32 genes encoding the RpoH protein that were invoked upon introduction into the late bloom environment and may have been the master regulators of stress responses (Zhao *et al.*, 2005; Nuss *et al.*, 2010; Berghoff *et al.*, 2011). Temperature and light exposure were kept at ambient laboratory conditions for all invasion experiments and salinity shifts were minor (33.35 to 33.60); these factors were therefore not likely to have differentially affected bacterial viability. Influence from exposure to UV light and formation of reactive oxygen species, either formed from organic matter in seawater or generated by the microbial community (Diaz *et al.*, 2013; Wietz *et al.*, 2013), are more likely drivers of these transcriptional responses.

Niche boundaries

The sequential invasion experiments characterized features of a natural coastal phytoplankton bloom environment eliciting transcriptional responses from a bacterium, but not whether the values of those features would allow its survival and reproduction. That is, the transcriptomic data addressed niche dimensions but not niche space. We looked for attributes of *R. pomeroyi*'s transcriptome that might indicate when the bacterium would have been successful in the "growth and spread" stage (Mallon *et al.*, 2015) had the invasion study been carried to completion. Transcription of ribosomal proteins was maximum in the late bloom invasions (Fig. 4.4a), accounting for >2-fold more of the transcript pool in late versus peak bloom experiments (2.3 vs. 1.2% for the two final vs. two initial experiments; Mann Whitney U-test, $p < 0.01$, d.f. = 9, two-tailed). Because up to 40% of a bacterium's energy is allocated to protein synthesis, cells strictly regulate ribosomal proteins transcripts to match available resources (Wei *et al.*, 2001; Wilson and Nierhaus, 2007; Maguire, 2009), and experimental studies have confirmed that *R. pomeroyi* ribosomal protein transcript inventory correlates with growth rate (Vinas, 2015; Cottrell and Kirchman, 2016). In addition, the bacterium's transcription of σ-70 (RpoD), the major regulator of housekeeping gene expression that shifts with growth (Ishihama, 2000), was also maximum in the late bloom invasions, accounting for >4-fold more of the transcript pool in late versus peak bloom experiments (0.83 vs. 0.20%; Mann Whitney U-test, $p < 0.01$, d.f. = 9, two-tailed) (Fig. 4.4a); experimental studies have similarly confirmed that *R. pomeroyi rpoD* transcript inventory is correlated with growth rate (Vinas, 2015). A third growth proxy based on abundance of the Rhodobacterales subclade to which *R. pomeroyi* belongs in the invaded community 16S rRNA gene pool (Fig. 4.4b), an index independent of transcription patterns, revealed that late bloom conditions supported >100-fold higher *Ruegeria-Sulfitobacter* clade

populations compared to peak bloom microbial communities (Fig. 4.4a) (2.22 vs. 0.02% of Rhodobacterales 16S rRNA sequences; Mann Whitney U-test, $p < 0.05$, d.f. = 7, two-tailed). Yet the agreement of these three indices that growth potential was higher for invading *R. pomeroyi* during the late bloom is counter to indications from transporter expression that substrates supporting the bacterium's growth were maximally available in the peak bloom (Fig. 4.3). One potential explanation for asynchrony between opportunity for substrate acquisition and growth, signaling a narrowing of the bacterium's niche, is that the bacterium was limited by nitrogen availability at the peak of the bloom and unable to capitalize on substrate supply. Transcription patterns suggest a variety of organic nitrogen molecules were available, however (Fig. 4.3). Alternatively, transport/catabolism transcripts may be unreliable reporters of substrate availability because they do not track closely with substrate supply. However, measures of DMSP concentration made at each invasion date (Kiene *et al.*, 2019) were strongly correlated to relative expression levels of the DMSP catabolism gene *dmdA* in invading *R. pomeroyi* cells (Pearson's $R = 0.87$, $p < 0.001$, d.f. = 12) (Fig. 4.4c). Thus for at least one key substrate of marine Rhodobacterales in the *Akashiwo* bloom (González *et al.*, 1999), gene expression patterns were synchronized with substrate supply.

A third possible explanation for the mismatch in timing of substrate transport expression and growth is that *R. pomeroyi*'s ability to capitalize on substrate availability was affected by negative biotic interactions that were strongest at the bloom peak. Secondary metabolite release by red tide species *A. sanguinea* is linked to toxicity in seabirds, fish, and invertebrates(Jessup *et al.*, 2009; Jones *et al.*, 2017; Xu *et al.*, 2017), and potentially narrowed the realized niche space of *R. pomeroyi* at the bloom peak. Measures of native bacterial community uptake of DMSP at each invasion date were unusually low on these dates, with turnover of dissolved DMSP

averaging 10% day^{-1} compared to typical bloom values 30% - 100% day^{-1} (Kiene and Linn, 2000; Motard-Côté *et al.*, 2016; Kiene *et al.*, 2019). The native bacterial community therefore processed DMSP from bloom seawater at unusually low rates despite high availability, and *R. pomeroyi* may have been similarly affected. Expression of bacterial genes for the synthesis of an RTX toxin, a secreted protein with cytotoxic and hemolytic activities toward eukaryotic cells (Lally *et al.*, 1999), and two polyketides were also highest at the peak bloom. Indications that negative biotic dimensions can be at least as important as resource dimensions in the realized niche space of phytoplankton-associated bacteria was unexpected, as resource-driven assembly of bacterial communities is common in other coastal phytoplankton blooms (Billen and Fontigny, 1987; Pinhassi *et al.*, 2004; Teeling *et al.*, 2012; Buchan *et al.*, 2014; Bunse *et al.*, 2016). Recognition of the importance of non-trophic biotic interactions in determining surface ocean microbial viability is increasing (Morris *et al.*, 2011; Stock *et al.*, 2019).

Negative biotic interactions are included in classical niche theory as the features that narrow an organism's niche from fundamental to realized, for example by competition for resources or toxicity from secondary metabolites (Hutchinson, 1957; Colwell and Rangel, 2009) (Fig. 4.4d). The inclusion in niche theory of positive biotic interactions (i.e., 'facilitation'(Bruno *et al.*, 2003)) such as public goods dimensions (e.g., vitamins; Morris *et al.*, 2012) or mutualism dimensions (e.g., nitrogen cross-feeding ; Morris *et al.*, 2012; Pacheco *et al.*, 2019) implies that, absent from early formulations, realized niche dimensions can extend beyond fundamental niche dimensions if interacting species broaden the conditions under which a microbe can survive (Bruno *et al.*, 2003; Colwell and Rangel, 2009) (Fig. 4.4d). This perspective is not necessarily counter to Hutchinson's conceptualizations (Colwell and Rangel, 2009), but is an aspect of niche

theory that may be particularly important in highly interconnected microbial communities such as coastal seawater blooms.

Ecological invasion studies, in their simplest form, add taxa to extant natural communities to uncover principles governing the ability of an invading species to successfully exploit the invaded community's resources (Bell, 2019). The niche dimensions that govern the invading species' success are notoriously difficult to identify, however, given the vast number of potentially influential environmental features, most of which have yet to be recognized (Saupe *et al.*, 2018). Further, correlated signals between inventoried features and microbial responses do not address causal relationships. In this study, a metabolically responsive species representative of marine bacteria with life history strategies linked to phytoplankton-derived metabolites (Luo and Moran, 2014; Fu *et al.*, 2020) was introduced into an environment it might reasonably invade. Each experimental invasion reported the species' *de novo* detection of environmental conditions, with transcriptional responses spotlighting 43 substrates, 9 vitamins or cofactors, 4 metals, and 24 mechanisms for non-trophic biotic interactions that influenced its ecological success during a dinoflagellate-dominated phytoplankton bloom. Invasion studies with model microorganisms can bridge the gap between ecologically-relevant field studies and mechanistically-informative model organism studies, improving understanding of the factors that determine where bacterial species survive and function in the seawater environment.

Methods

Experimental setup

The pre-incubation protocol for *Ruegeria pomeroyi* DSS-3 began two days before each invasion experiment. The bacterium was inoculated into ½ YTSS liquid medium and grown

overnight to exponential phase at 30° C. Cells were then washed twice, inoculated into marine basal medium (MBM)(González *et al.*, 1997) with 10 mM glucose, and incubated at 30° C for ~26 h. After washing 3 times and resuspending in artificial seawater (28 g L^{-1}; Sigma sea salts), the bacterium was added into triplicate 350 ml aliquots of unfiltered Monterey Bay surface seawater collected at Station M0(Nowinski *et al.*, 2019) at approximately 10 am PST, for a final concentration of ~1 x 10^{10} *R. pomeroyi* cells L^{-1}. Incubations were stirred at 120 rpm at ambient temperature (24ºC) and light for 90 min and then filtered sequentially through 2-µm polycarbonate filters to remove most non-bacterial community members and 0.2-µm polycarbonate filters to collect the bacterial size fraction. Filters were flash frozen in liquid nitrogen. Two control experiments were set up in which *R. pomeroyi* cells were prepared as described above for the field studies but inoculated into defined media consisting of MBM with no substrate or with 10 mM glucose (Fig. 4.1b).

RNA extraction and sequencing

RNA was extracted from filters using the ZymoBIOMICS RNA Miniprep Kit (Zymo Research), treated with Turbo DNase (Invitrogen, Waltham, MA, USA), and cleaned using RNA Clean & Concentrator (Zymo Research). Ribosomal rRNA was removed using the Ribo-Zero Bacteria Kit (Illumina, San Diego, CA, USA). Stranded libraries were prepared and sequenced using an Illumina Next-Seq SE75 High Output flow cell at the Georgia Genomics and Bioinformatics Core (University of Georgia).

Bioinformatic analysis

The FASTX toolkit was used to retain reads with a minimum quality score of 20 over 80% of read length. Reads were mapped to the *R. pomeroyi* genome using BWA (Li and Durbin, 2009) and counted using HTSeq (Anders *et al.*, 2015). Counts were converted to transcripts per million (TPM). Weighted transcriptomic correlation network analysis was performed on z-score transformed TPMs to cluster genes based on their expression across sample dates into modules within a correlation network using the R package WGCNA (Langfelder and Horvath, 2008). Differential expression of genes between sample dates was calculated using DESeq2 (Love *et al.*, 2014). PCA analysis was carried out on mean-normalized TPMs using the R program prcomp.

Environmental data

Bacteria and eukaryotic microbes were counted by flow cytometry and microscopy, respectively (Nowinski *et al.*, 2019) and total particulate + dissolved dimethylsulfoniopropionate concentrations (DMSPt) and bacterial uptake rates were measured at each sample date (Kiene *et al.*, 2019). 16S and 18S rRNA gene libraries were analyzed from seawater collected from Station M0 at the time of each *R. pomeroyi* addition as described previously (Nowinski *et al.*, 2019). The SILVA v132 rRNA gene database (16S) (Quast *et al.*, 2012) and the Protist Ribosomal Reference database (PR2; 18S) (Guillou *et al.*, 2013) were used to classify sequences. Heterotrophic protist ASVs were removed before community analysis. UPGMA clustering of unweighted UniFrac distances was run in Qiime2(Bolyen *et al.*, 2018).

Ruegeria-Sulfitobacter *clade abundance*

16S rRNA gene amplicon sequence variants (ASVs) from the 16S rRNA gene libraries classified in the Rhodobacteraceae were aligned using blastn to the Joint Genome Institute IMG/M All Isolates database. Top hits to a marine Rhodobacterales genome were used to construct a phylogenomic tree using GToTree v1.4.1(Lee, 2019) based on HMM profiles of 117 alphaproteobacterial single-copy genes. ASV counts were mapped onto the tree, and relative abundance of a well-supported clade that included the *R. pomeroyi* genome was calculated as a percent of all Rhodobacterales ASV hits.

Data Availability Statement

Data that support the findings of this study have been deposited in NCBI SRA with BioProject numbers PRJNA641119 (RNA-seq) and PRJNA511156 - PRJNA511331 (16S and 18S rRNA data), and the Biological and Chemical Oceanography Data Management Office under DOI:10.1575/1912/bco-dmo.756413.2 at https://www.bco-dmo.org/dataset/756413/data (environmental data).

Acknowledgements

C. Preston, J. Birch, C. Sholin and the MBARI ESP team provided field sampling infrastructure and expertise, S. Sharma provided bioinformatic assistance, C. Smith, C. Thomas, and K. Esson assisted with field and laboratory techniques, R. Michisaki provided data for microbial biomass estimates, and the University of Georgia Genomics and Bioinformatics Core (GGBC) supplied sequencing services. This work was supported by Simons Foundation (grant 542391 to MAM) within the Principles of Microbial Ecosystems (PriME) Collaborative and NSF

(IOS-1656311). rRNA amplicon sequencing was provided through the DOE Joint Genome Institute Community Sequencing Program.

Author Contributions

BN and MAM conceived of the study, BN collected the data, and BN and MAM analyzed data and wrote the paper.

References

Alneberg, J., Bennke, C., Beier, S., Bunse, C., Quince, C., Ininbergs, K. et al. (2020) Ecosystem-wide metagenomic binning enables prediction of ecological niches from genomes. *Comm Biology* **3**: 1-10.

Amin, S.A., Hmelo, L.R., Van Tol, H.M., Durham, B.P., Carlson, L.T., Heal, K.R. et al. (2015) Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature* **522**: 98-101.

Anders, S., Pyl, P.T., and Huber, W. (2015) Genome analysis HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166-169.

Anderson, S.R., and Harvey, E.L. (2019) Seasonal variability and drivers of microzooplankton grazing and phytoplankton growth in a subtropical estuary. *Front Mar Sci* **6**: 174-174.

Anderson, S.R., Diou-Cass, Q.P., and Harvey, E.L. (2018) Short-term estimates of phytoplankton growth and mortality in a tidal estuary. *Limnol Oceanogr* **63**: 2411-2422.

Baltar, F., Bayer, B., Bednarsek, N., Deppeler, S., Escribano, R., Gonzalez, C.E. et al. (2019) Towards integrating evolution, metabolism, and climate change studies of marine ecosystems. *Trends Ecol Evol* **34**: 1022-1033.

Bell, T. (2019) Next-generation experiments linking community structure and ecosystem functioning. *Environ Microbiol Rep* **11**: 20-22.

Berghoff, B.A., Glaeser, J., Nuss, A.M., Zobawa, M., Lottspeich, F., and Klug, G. (2011) Anoxygenic photosynthesis and photooxidative stress: a particular challenge for Roseobacter. *Environ Microbiol* **13**: 775-791.

Biers, E.J., Wang, K., Pennington, C., Belas, R., Chen, F., and Moran, M.A. (2008) Occurrence and expression of gene transfer agent genes in marine bacterioplankton. *Appl Environ Microbiol* **74**: 2933-2939.

Billen, G., and Fontigny, A. (1987) Dynamics of a *Phaeocystis*-dominated spring bloom in Belgian coastal waters. II. Bacterioplankton dynamics. *Mar Ecol Prog Ser* **37**: 249-257.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C., Al-Ghalith, G.A. et al. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. In: PeerJ Preprints.

Bruno, J.F., Stachowicz, J.J., and Bertness, M.D. (2003) Inclusion of facilitation into ecological theory. *Trends Ecol Evol* **18**: 119-125.

Buchan, A., LeCleir, G.R., Gulvik, C.A., and González, J.M. (2014) Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nat Rev Microbiol* **12**: 686-698.

Bunse, C., Bertos-Fortis, M., Sassenhagen, I., Sildever, S., Sjöqvist, C., Godhe, A. et al. (2016) Spatio-temporal interdependence of bacteria and phytoplankton during a Baltic Sea spring bloom. *Front Microbiol* **7**: 517-517.

Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., and Vergin, K. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* **3**: 283-295.

Chan, L.-K., Newton, R.J., Sharma, S., Smith, C.B., Rayapati, P., Limardo, A.J. et al. (2012) Transcriptional changes underlying elemental stoichiometry shifts in a marine heterotrophic bacterium. *Front Microbiol* **3**: 159.

Church, M.J., Hutchins, D.A., and Ducklow, H.W. (2000) Limitation of bacterial growth by dissolved organic matter and iron in the Southern Ocean. *Appl Environ Microbiol* **66**: 455-466.

Cohan, F.M. (2002) What are bacterial species? *Ann Rev Microbiol* **56**: 457-487.

Colwell, R.K., and Fuentes, E.R. (1975) Experimental studies of the niche. *Ann Rev Ecol Syst* **6**: 281-310.

Colwell, R.K., and Rangel, T.F. (2009) Hutchinson's duality: the once and future niche. *Proc Nat Acad Sci* **106**: 19651-19658.

Cottrell, M.T., and Kirchman, D.L. (2016) Transcriptional control in marine copiotrophic and oligotrophic bacteria with streamlined genomes. *Appl Environ Microbiol* **82**: 6010-6018.

Diaz, J.M., Hansel, C.M., Voelker, B.M., Mendes, C.M., Andeer, P.F., and Zhang, T. (2013) Widespread production of extracellular superoxide by heterotrophic bacteria. *Science* **340**: 1223-1226.

Erguder, T.H., Boon, N., Wittebolle, L., Marzorati, M., and Verstraete, W. (2009) Environmental factors shaping the ecological niches of ammonia-oxidizing archaea. *FEMS Microbiol Rev* **33**: 855-869.

Fu, H., Uchimiya, M., Gore, J., and Moran, M.A. (2020) Ecological drivers of bacterial community assembly in synthetic phycospheres. *Proc Nat Acad Sci* **117**: 3656-3662.

Galambos, D., Anderson, R.E., Reveillaud, J., and Huber, J.A. (2019) Genome-resolved metagenomics and metatranscriptomics reveal niche differentiation in functionally redundant microbial communities at deep-sea hydrothermal vents. *Environ Microbiol* **21**: 4395-4410.

Gifford, S.M., Sharma, S., Booth, M., and Moran, M.A. (2013) Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J* **7**: 281-298.

González, J.M., Kiene, R.P., and Moran, M.A. (1999) Transformation of sulfur compounds by an abundant lineage of marine bacteria in the α-subclass of the class Proteobacteria. *Appl Environ Microbiol* **65**: 3810-3819.

González, J.M., Mayer, F., Moran, M.A., Hodson, R.E., and Whitman, W.B. (1997) *Microbulbifer hydrolyticus* gen. nov., sp. nov., and *Marinobacterium georgiense* gen. nov., sp. nov., two marine bacteria from a lignin-rich pulp mill waste enrichment community. *Int J Syst Evol Microbiol* **47**: 369-376.

González, J.M., Covert, J.S., Whitman, W.B., Henriksen, J.R., Mayer, F., Scharf, B. et al. (2003) *Silicibacter pomeroyi* sp. nov. and *Roseovarius nubinhibens* sp. nov., dimethylsulfoniopropionate-demethylating bacteria from marine environments. *Int J Syst Evol Microbiol* **53**: 1261-1269.

Gravel, D., Bell, T., Barbera, C., Bouvier, T., Pommier, T., Venail, P., and Mouquet, N. (2011) Experimental niche evolution alters the strength of the diversity-productivity relationship. *Nature* **469**: 89-94.

Guillou, L., Bachar, D., phane Audic, S., Bass, D., dric Berney, C., Bittner, L. et al. (2013) The Protist Ribosomal Reference database (PR^2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**: 597-604.

Holt, R.D. (2009) Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proc Nat Acad Sci* **106**: 19659-19665.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symp Quant Biol* **22**: 415-427.

Ishihama, A. (2000) Functional modulation of *Escherichia coli* RNA polymerase. *Annu Rev Microbiol* **54**: 499-518.

Jessup, D.A., Miller, M.A., Ryan, J.P., Nevins, H.M., and Kerkering, H.A. (2009) Mass stranding of marine birds caused by a surfactant-producing red tide. *PLoS one* **4**: 4550-4550.

Jones, T., Parrish, J.K., Punt, A.E., Trainer, V.L., Kudela, R., Lang, J. et al. (2017) Mass mortality of marine birds in the Northeast Pacific caused by Akashiwo sanguinea. *Mar Ecol Prog Ser* **579**: 111-127.

Kiene, R.P., and Linn, L.J. (2000) Distribution and turnover of dissolved DMSP and its relationship with bacterial production and dimethylsulfide in the Gulf of Mexico. *Limnol Oceanogr* **45**: 849-861.

Kiene, R.P., Nowinski, B., Esson, K., Preston, C., Marin III, R., Birch, J. et al. (2019) Unprecedented DMSP concentrations in a massive dinoflagellate bloom in Monterey Bay, CA. *Geophys Res Lett* **46**: 12279-12288.

Kudela, R.M., Seeyave, S., and Cochlan, W.P. (2010) The role of nutrients in regulation and promotion of harmful algal blooms in upwelling systems. *Prog Oceanogr* **85**: 122-135.

Lally, E.T., Hill, R.B., Kieba, I.R., and Korostoff, J. (1999) The interaction between RTX toxins and target cells. *Trends Microbiol* **7**: 356-361.

Landa, M., Burns, A.S., Roth, S.J., and Moran, M.A. (2017) Bacterial transcriptome remodeling during sequential co-culture with a marine dinoflagellate and diatom. *ISME J* **11**: 2677-2690.

Langfelder, P., and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**: 559.

Lee, M.D. (2019) GToTree: A user-friendly workflow for phylogenomics. *Bioinformatics* **35**: 4162-4164.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Love, M.I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550-550.

Luo, H., and Moran, M.A. (2014) Evolutionary ecology of the marine Roseobacter clade. *Microbiol Mol Biol Rev* **78**: 573-587.

Maguire, B.A. (2009) Inhibition of bacterial ribosome assembly: a suitable drug target? *Microbiol Mol Biol Rev* **73**: 22-35.

Mallon, C.A., Van Elsas, J.D., and Salles, J.F. (2015) Microbial invasions: the process, patterns, and mechanisms. *Trends Microbiol* **23**: 719-729.

Martens-Habbena, W., Berube, P.M., Urakawa, H., José, R., and Stahl, D.A. (2009) Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* **461**: 976-979.

McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B., and Paul, J.H. (2010) High frequency of horizontal gene transfer in the oceans. *Science* **330**: 50-50.

Meier, D.V., Pjevac, P., Bach, W., Hourdez, S., Girguis, P.R., Vidoudez, C. et al. (2017) Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *ISME J* **11**: 1545-1558.

Moran, M.A., Buchan, A., González, J.M., Heidelberg, J.F., Whitman, W.B., Kiene, R.P. et al. (2004) Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. *Nature* **432**: 910-913.

Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012) The black queen hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**: e00036-00012.

Morris, J.J., Johnson, Z.I., Szul, M.J., Keller, M., and Zinser, E.R. (2011) Dependence of the cyanobacterium *Prochlorococcus* on hydrogen peroxide scavenging microbes for growth at the ocean's surface. *PloS one* **6**.

Motard-Côté, J., Kieber, D.J., Rellinger, A., and Kiene, R.P. (2016) Influence of the Mississippi River plume and non-bioavailable DMSP on dissolved DMSP turnover in the northern Gulf of Mexico. *Environ Chem* **13**: 280-280.

Muller, E.E. (2019) Determining microbial niche breadth in the environment for better ecosystem fate predictions. *mSystems* **4**: e00080-00019.

Nowinski, B., Smith, C.B., Thomas, C.M., Esson, K., Marin, R., Preston, C.M. et al. (2019) Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. *Sci Data* **6**: 1-7.

Nuccio, E.E., Starr, E., Karaoz, U., Brodie, E.L., Zhou, J., Tringe, S.G. et al. (2020) Niche differentiation is spatially and temporally regulated in the rhizosphere. *ISME J* **14**: 999-1014.

Nuss, A.M., Glaeser, J., Berghoff, B.A., and Klug, G. (2010) Overlapping alternative sigma factor regulons in the response to singlet oxygen in *Rhodobacter sphaeroides*. *J Bact* **192**: 2613-2623.

Ottesen, E.a., Young, C.R., Eppley, J.M., Ryan, J.P., Chavez, F.P., Scholin, C.a., and DeLong, E.F. (2013) Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc Nat Acad Sci* **110**: E488-497.

Pacheco, A.R., Moel, M., and Segrè, D. (2019) Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat Comm* **10**: 1-12.

Palovaara, J., Akram, N., Baltar, F., Bunse, C., Forsberg, J., Pedrós-Alió, C. et al. (2014) Stimulation of growth by proteorhodopsin phototrophy involves regulation of central metabolic pathways in marine planktonic bacteria. *Proc Nat Acad Sci* **111**: E3650-E3658.

Persson, O.P., Pinhassi, J., Riemann, L., Marklund, B.I., Rhen, M., Normark, S. et al. (2009) High abundance of virulence gene homologues in marine bacteria. *Environ Microbiol* **11**: 1348-1357.

Pinhassi, J., Sala, M.M., Havskum, H., Peters, F., Guadayol, Ò., Malits, A., and Marrasé, C. (2004) Changes in bacterioplankton composition under different phytoplankton regimens. *Appl Environ Microbiol* **70**: 6753-6766.

Poretsky, R.S., Sun, S., Mou, X., and Moran, M.A. (2010) Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* **12**: 616-627.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**: D590-D596.

Saupe, E.E., Barve, N., Owens, H.L., Cooper, J.C., Hosner, P.A., and Peterson, A.T. (2018) Reconstructing ecological niche evolution when niches are incompletely characterized. *System Biol* **67**: 428-438.

Shaiber, A., and Eren, A.M. (2019) Composite metagenome-assembled genomes reduce the quality of public genome repositories. *mBio* **10**: e00725-00719.

Sharpe, G.C., Gifford, S.M., and Septer, A.N. (2019) A model Roseobacter employs a diffusible killing mechanism to eliminate competitors. *bioRxiv*: 766410-766410.

Stock, F., Syrpas, M., Graff van Creveld, S., Backx, S., Blommaert, L., Dow, L. et al. (2019) N-acyl homoserine lactone derived tetramic acids impair photosynthesis in *Phaeodactylum tricornutum*. *ACS Chem Biol* **14**: 198-203.

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

Vergin, K.L., Tripp, H.J., Wilhelm, L.J., Denver, D.R., Rappé, M.S., and Giovannoni, S.J. (2007) High intraspecific recombination rate in a native population of *Candidatus* Pelagibacter ubique (SAR11). *Environ Microbiol* **9**: 2430-2440.

Vinas, N. (2015) Relationships between growth rate and gene expression in Ruegeria pomeroyi DSS-3, a model marine alphaproteobacterium. In *Department of Microbiology*. Clemson, SC: M.S. Thesis, Clemson University.

Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A., and Larossa, R.A. (2001) High-density microarray-mediated gene expression profiling of *Escherichia coli*. *J Bact* **183**: 545-556.

Wietz, M., Duncan, K., Patin, N.V., and Jensen, P.R. (2013) Antagonistic interactions mediated by marine bacteria: the role of small molecules. *J Chem Ecol* **39**: 879-891.

Wilson, D.N., and Nierhaus, K.H. (2007) The weird and wonderful world of bacterial ribosome regulation. *Crit Rev Biochem Mol Biol* **42**: 187-219.

Xu, N., Wang, M., Tang, Y., Zhang, Q., Duan, S., and Gobler, C.J. (2017) Acute toxicity of the cosmopolitan bloom-forming dinoflagellate A*kashiwo sanguinea* to finfish, shellfish, and zooplankton. *Aquat Microb Ecol* **80**: 209-222.

Yeung, L.Y., Berelson, W.M., Young, E.D., Prokopenko, M.G., Rollins, N., Coles, V.J. et al. (2012) Impact of diatom-diazotroph associations on carbon export in the Amazon River plume. *Geophys Res Lett* **39**.

Zhao, K., Liu, M., and Burgess, R.R. (2005) The global transcriptional response of *Escherichia coli* to induced σ32 protein involves σ32 regulon activation followed by inactivation and degradation of σ32 in vivo. *J Biol Chem* **280**: 17758-17768.

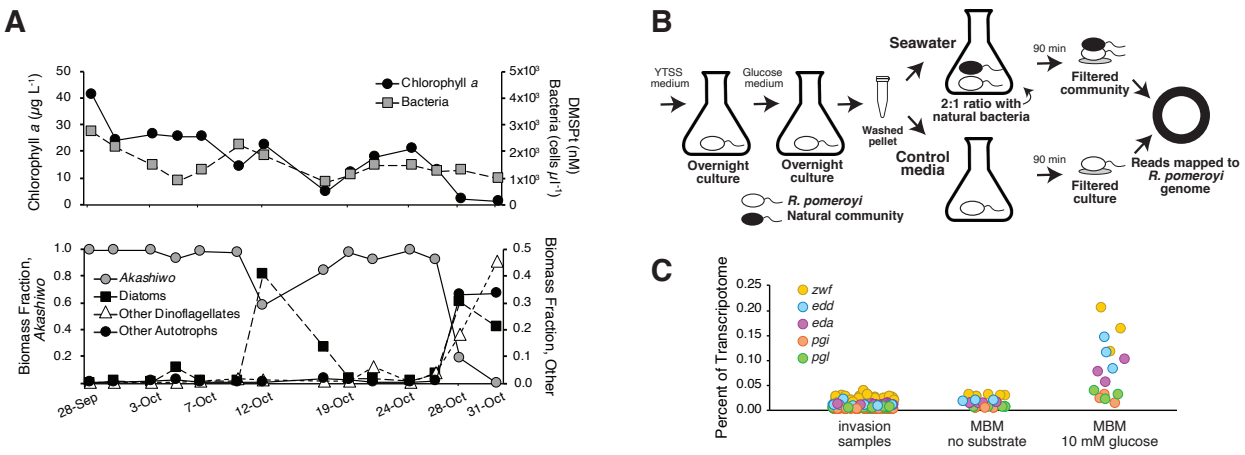


Figure 4.1. A) Chemical and biological features of the 2016 Monterey Bay fall bloom. Top panel, chlorophyll a, n=3; bacterial cell counts, n=2. Error bars (+ 1 s.d.) are within the symbols. Bottom panel, n = 1. Note: different scales are used on the left and right axes of the bottom panel. B) Protocol for invasion experiments; n = 3. C) Relative expression of five R. pomeroyi glucose catabolism genes in 14 invasion studies compared to controls incubated in marine basal medium without a substrate or with glucose. MBM glucose samples are signifcantly higher than the other treatments (ANOVA $p<0.0001$, Tukey HSD $p<0.01$, d.f. = 2).

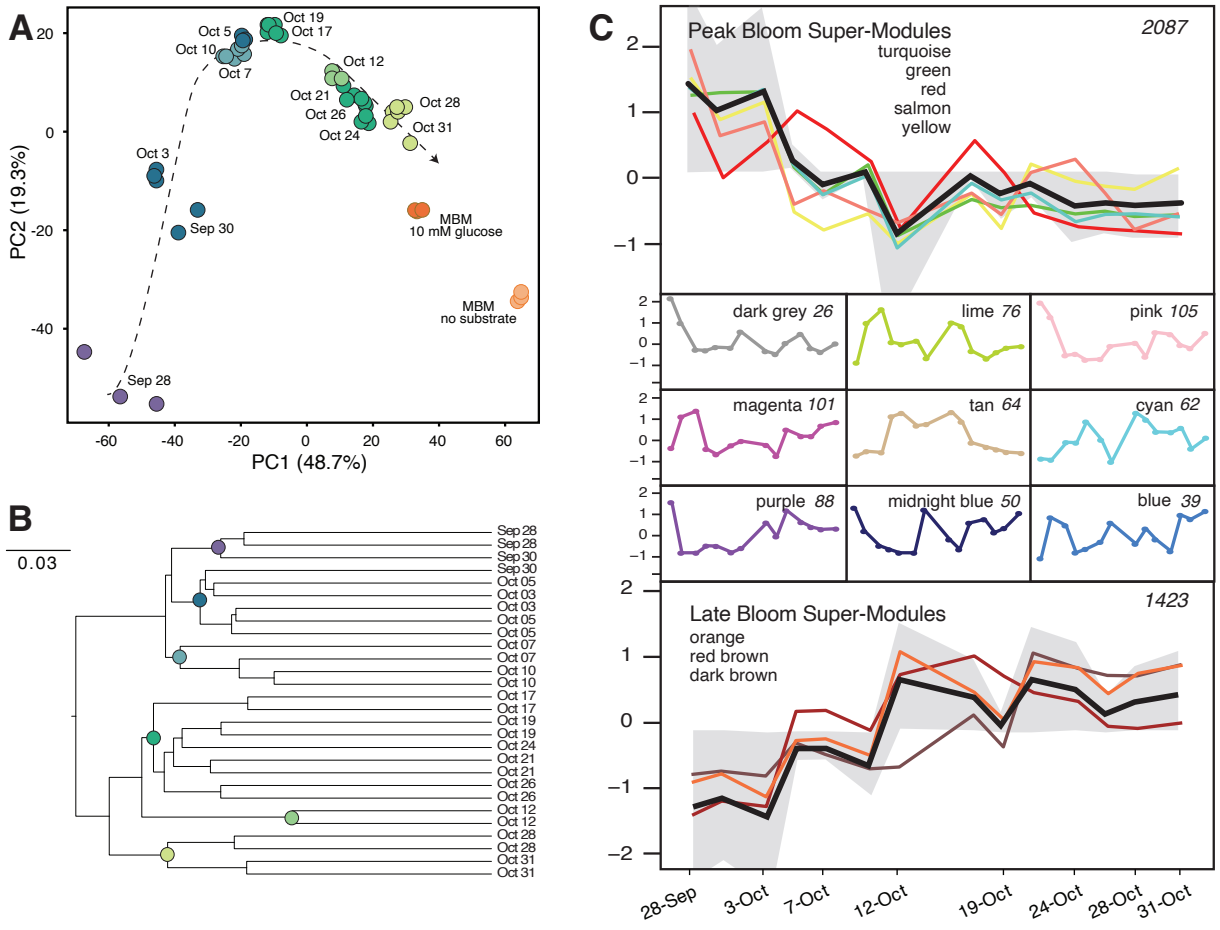
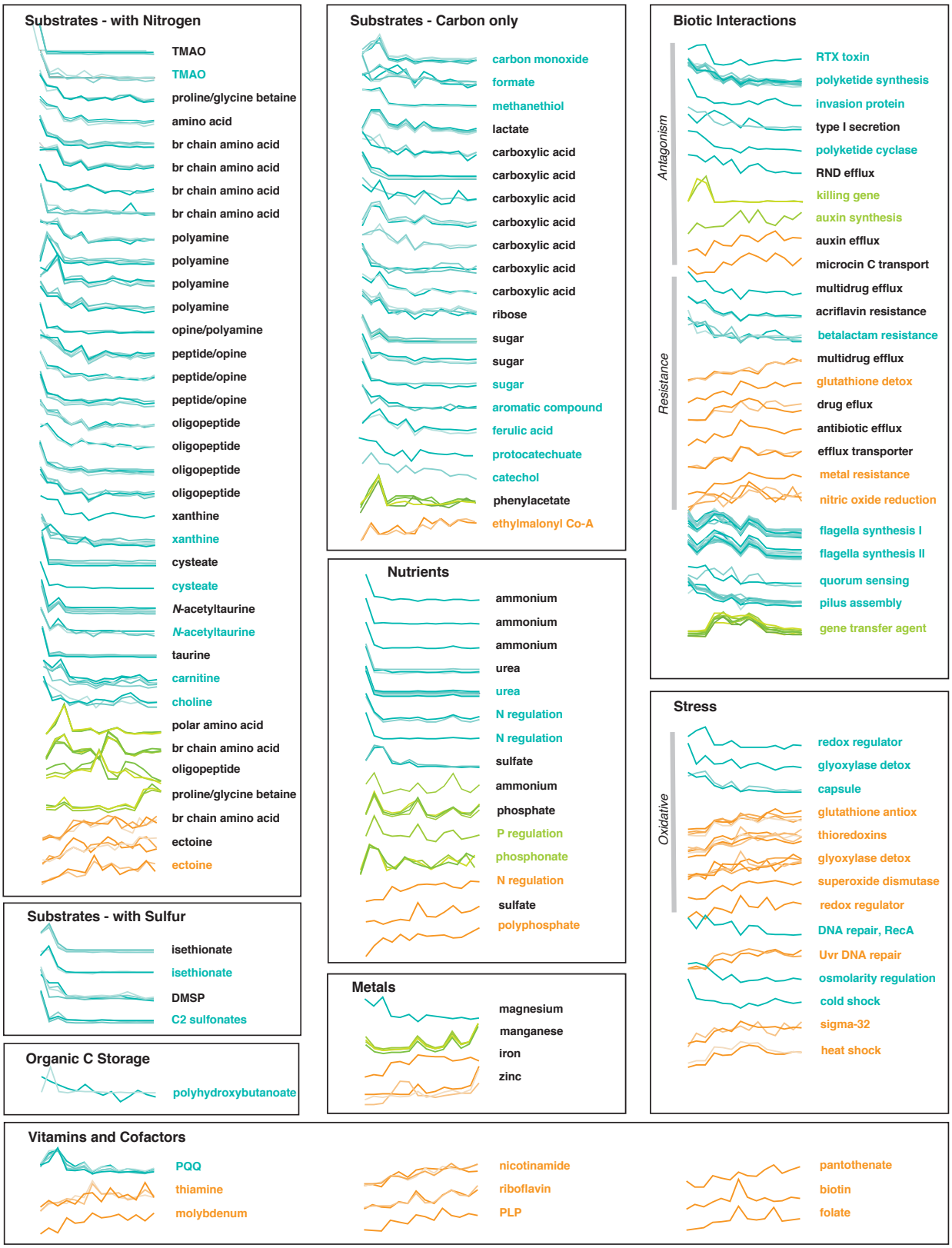


Figure 4.2. A) Principal components analysis of *R. pomeroyi* whole transcriptome expression patterns in invasion and control experiments. Invasion experiments are labeled by date and symbols are colored based on taxonomic composition of the protist community. B) Cluster analysis of protist communities. Colors at clade nodes correspond to those in panel A. 18S rRNA taxonomic details are given in Fig. 4.S1. C) *R. pomeroyi* genes classified based on expression patterns over the 14 invasion experiments. Significantly correlated modules were merged into peak bloom or late bloom super-modules. Data are Z-scores of mean expression for 3 replicate samples. Numbers in the upper right corner indicate the genes in each module/super-module. Black lines are the average of all genes in the super-modules and grey shading is the standard deviation.

Figure 4.3. Time course of relative gene expression indicating *R. pomeroyi* responses to niche dimensions in the invaded community. Genes are organized by dimension type and colored by expression module; peak bloom super-module, turquoise lines; late bloom super-module, orange lines; other modules, green lines. Multiple lines represent protein subunits or functionally related genes. Functional assignments in black font denote transporters. All genes shown are significantly different at two highest and lowest time points based on DeSeq2, $p < 0.05$. Data are mean Z scores; $n = 3$ for all dates except date 2 (Sept. 30), where $n = 2$.


Substrates - with Nitrogen
TMAO
TMAO
proline/glycine betaine
amino acid
br chain amino acid
br chain amino acid
br chain amino acid
br chain amino acid
polyamine
polyamine
polyamine
polyamine
opine/polyamine
peptide/opine
peptide/opine
peptide/opine
oligopeptide
oligopeptide
oligopeptide
oligopeptide
xanthine
xanthine
cysteate
cysteate
N-acetyltaurine
N-acetyltaurine
taurine
carnitine
choline
polar amino acid
br chain amino acid
oligopeptide
proline/glycine betaine
br chain amino acid
ectoine
ectoine
Substrates - with Sulfur
isethionate
isethionate
DMSP
C2 sulfonates
Organic C Storage
polyhydroxybutanoate
Substrates - Carbon only
carbon monoxide
formate
methanethiol
lactate
carboxylic acid
carboxylic acid
carboxylic acid
carboxylic acid
carboxylic acid
carboxylic acid
carboxylic acid
ribose
sugar
sugar
sugar
aromatic compound
ferulic acid
protocatechuate
catechol
phenylacetate
ethylmalonyl Co-A
Nutrients
ammonium
ammonium
ammonium
urea
urea
N regulation
N regulation
sulfate
ammonium
phosphate
P regulation
phosphonate
N regulation
sulfate
polyphosphate
Metals
magnesium
manganese
iron
zinc
Biotic Interactions
Antagonism
Resistance
RTX toxin
polyketide synthesis
invasion protein
type I secretion
polyketide cyclase
RND efflux
killing gene
auxin synthesis
auxin efflux
microcin C transport
multidrug efflux
acriflavin resistance
betalactam resistance
multidrug efflux
glutathione detox
drug eflux
antibiotic efflux
efflux transporter
metal resistance
nitric oxide reduction
flagella synthesis I
flagella synthesis II
quorum sensing
pilus assembly
gene transfer agent
Stress
Oxidative
redox regulator
glyoxylase detox
capsule
glutathione antiox
thioredoxins
glyoxylase detox
superoxide dismutase
redox regulator
DNA repair, RecA
Uvr DNA repair
osmolarity regulation
cold shock
sigma-32
heat shock
Vitamins and Cofactors
PQQ
thiamine
molybdenum
nicotinamide
riboflavin
PLP
pantothenate
biotin
folate

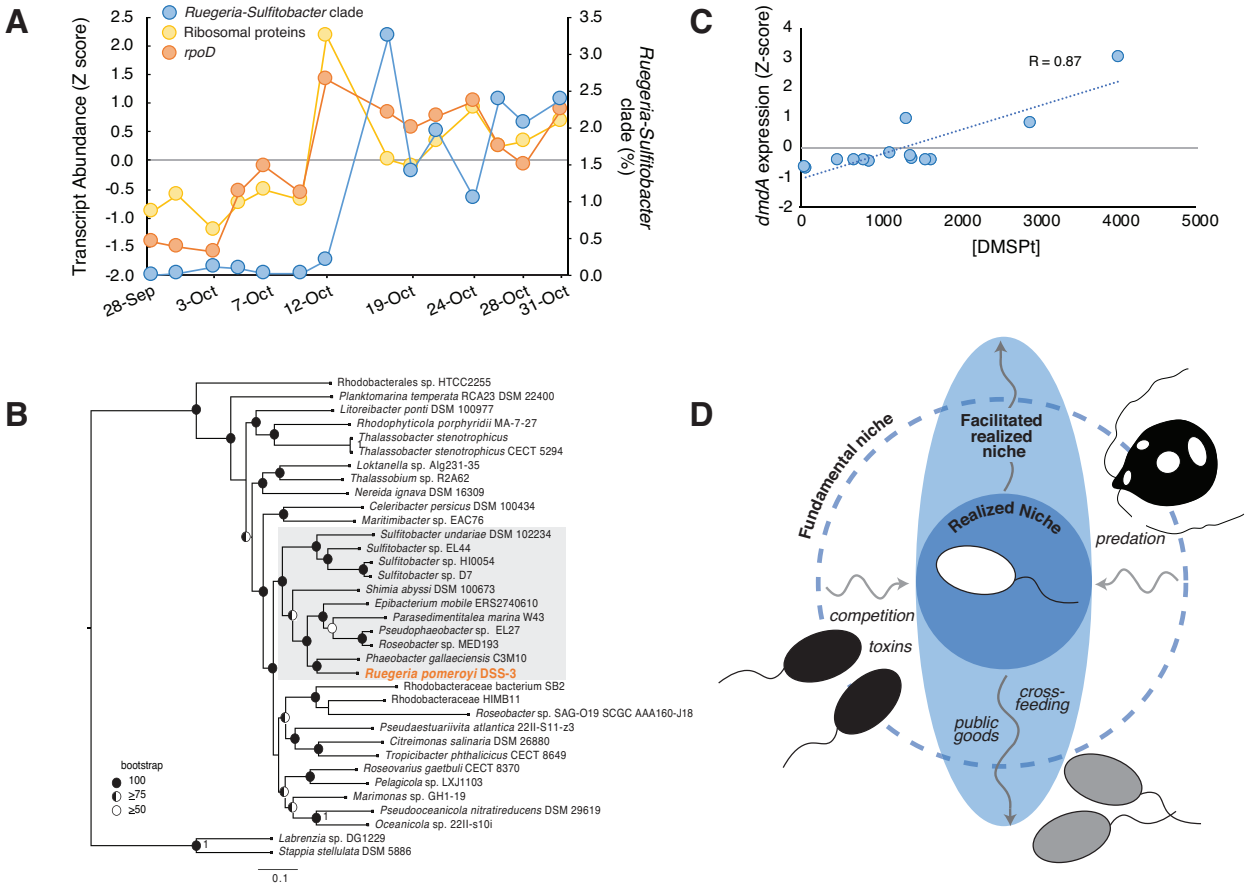


Figure 4.4. A) Growth metrices of *R. pomeroyi* on 14 dates during the 2016 fall Monterey Bay bloom. Expression of ribosomal protein genes (sum of 54 genes) and *rpoD* (σ-70 factor) (left axis) is plotted as a Z-score normalization of mean values. n = 3. *Ruegeria-Sufitobacter* clade abundance in the bloom communities without added *R. pomeroyi* (right axis) is shown as percent of Rhodobacterales sequences. n = 2 or 3. B) Phylogenomic tree of genomes most closely related to Rhodobacterales ASVs in 16S rRNA libraries (Fig. S1); grey shading indicates the *Rugeria-Sulfitobacter* clade shown in panel A. C) Correlation of DMSP concentrations with relative expression of DMSP catabolism gene *dmdA* in invading *R. pomeroyi* cells. Pearson's R, $p < 0.01$, d.f. = 12. D) Conceptualization of a microbial fundamental niche (dashed line) reflecting the environmental conditions determining where net growth rate is > 0 in the absence of biotic interactions. Negative biotic interactions such as competition and toxins narrow niche space to the realized niche (dark blue circle), while positive biotic interactions, such as the provision of resources, can expand realized niche space beyond the fundamental niche along certain dimensions. Modified from Bruno *et al.* 2003.

Figure 4.S1. A,B) Protist and C) bacterial community composition during the 2016 Monterey Bay fall bloom based on rRNA gene amplicon sequencing. Each bar represents 1 replicate sample.

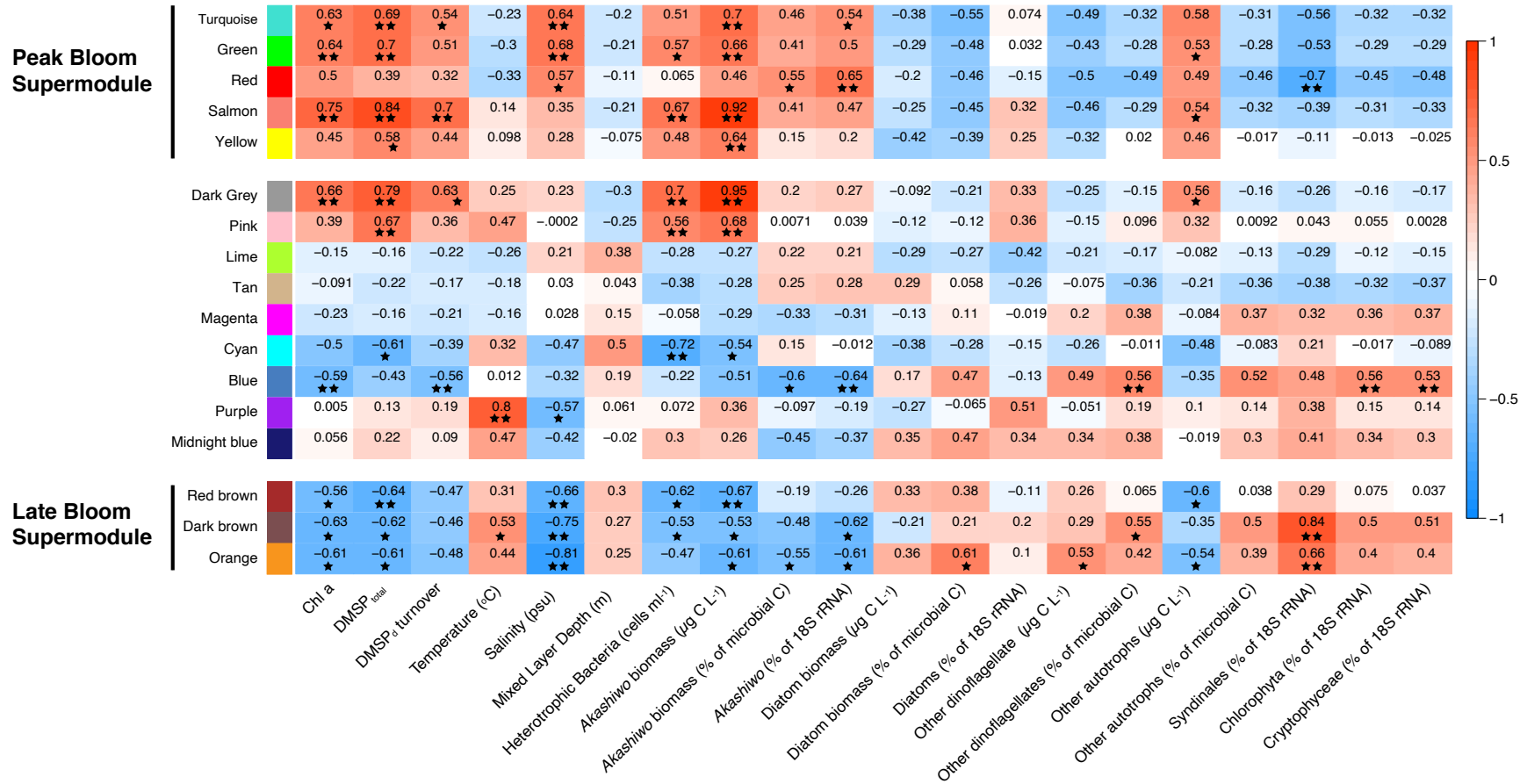


Figure 4.S2. Correlations of Z-score normalized *R. pomeroy*i gene expression modules with environmental data measured on the 14 sample dates at Monterey Bay Station M0 in Fall, 2016. Cells are colored by Pearson's R parameter. Two stars indicate correlations at $p < 0.01$; one star indicates correlations at $p < 0.05$. d.f. = 12.

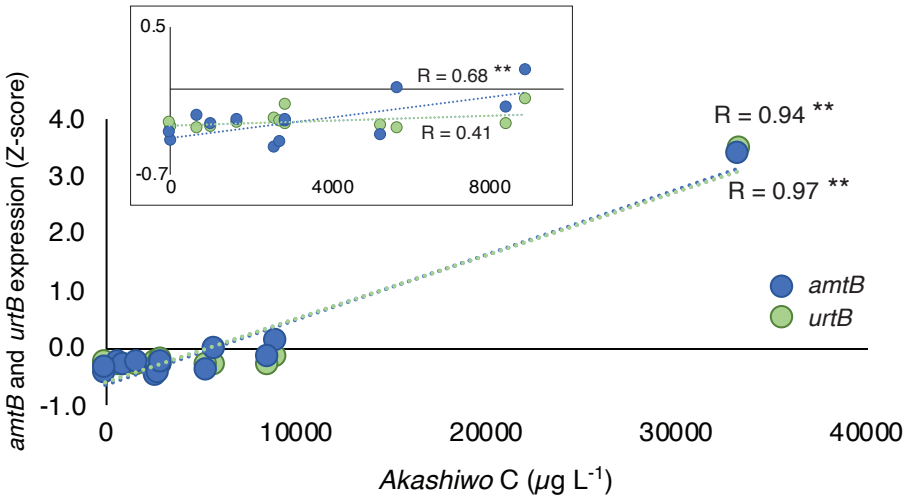


Figure 4.S3. Correlations of representative *R. pomeroyi* genes for transport of ammonium (*amtB*) and urea (*urtB*) with *Akashiwo* biomass. The extreme value of *Akashiw*o biomass is removed in the inset. **, Pearson's R, $p < 0.01$. d.f. = 12, main figure, d.f. = 11, inset.

Table 4.S1. *R. pomeroyi* genes with significant changes in relative expression and informative of realized niche dimensions in the 2016 fall bloom in Monterey Bay, CA, USA. Gene expression patterns are shown in Fig. 4.3

Gene locus tags	Gene name	Gene annotation	Super-Module
Chemicals			
Substrate, organic N			
SPO1548-1550		TMAO ABC transporter	P
SPO1562, 1579, 1580-582, 2750, 3400		TMAO degradation	P
SPO1131-1133		proline/glycine betaine ABC transporter-1	P
SPO0520-0522		amino acid ABC transporter	P
SPO0822-0825		branched-chain amino acid ABC transporter-1	P
SPO1846, 1848-1850		branched-chain amino acid ABC transporter-3	P
SPO1936, 1937		branched-chain amino acid ABC transporter-4	P
SPOA0296, A0298-A0300	*livF-2*	branched-chain amino acid ABC transporter-5	P
SPO1607-1609	*potA,B,C*	spermidine/putrescine ABCtransporter-1	P
SPO2006-2009		spermidine/putrescine ABC transporter-2	P
SPO3466-3469,	*potF,G,H,I*	putrescine ABC transporter	P
SPO3472-3475		polyamine ABC transporter	P
SPO2700, 2702		opine/polyamine ABC transporter	P
SPO1543-1547		peptide/opine/nickel ABC transporter-1	P
SPO2814-2816		peptide/opine/nickel ABC transporter-2	P
SPO2995-2998,		peptide/nickel/opine ABC transporter-3	P
SPO0558-0561		oligopeptide ABC transporter-1	P
SPO0703, 0705		oligopeptide ABC transporter-2	P
SPO1210-1213		oligopeptide ABC transporter-3	P
SPO1656-1659		oligopeptide/dipeptide ABC transporter	P
SPO0874		xanthine permease	P
SPO0652-0654, 0830, 0831	*xdhA,B,C*	xanthine degradation	P
SPO2658-2661		cysteate ABC transporter	P
SPO2657	*cuyA2*	cysteate degradation	P
SPO0660-0664	*naaA,B,B',C,C'*	*N*-acetyltaurine ABC transporter	P
SPO0657-0659	*naaR,S,T*	*N*-acetyltaurine degradation	P
SPO0674-0676	*tauA,B,C*	taurine ABC transporter	P
SPO2704-2706, 2708, 2711	*caiD-1*	carnitine degradation	P
SPO0800, 1082, 1083	*betC,I*	choline sulfate degradation	P
SPOA0068-0071		polar amino acid	
SPO1829-SPO1833		branched-chain amino acid ABC transporter-6	
SPO2441-2443		proline/glycine betaine ABC transporter-2	
SPO3774-3778		oligopeptide ABC transporter-4	
SPO1491-1493		branched-chain amino acid ABC transporter-2	L
SPO1146, 1147	*uehA,B*	ectoine TRAP transporter	L
SPO1143, 1144	*eutA*	ectoine degradation	L
Substrate, organic S			
SPO2356-2358	*iseM,L,K*	isethionate transporter	P
SPO2355-2359	*iseR,J*	isethionate degradation	P
SPO0453, 1912-1914	*dddW; dmdA,R; acuI*	DMSP degradation	P
SPO3560-3562	*pta, xsc, tauR*	C2 sulfonate degradation	P

Substrate, C only			
SPO1517-1520, 2396	*coxG,S-1,L-1*	carbon monoxide oxidation	P
1555-1557, 1794, 1796	*fhs-1*	formate degradation	P
SPOA0268, A0269	*mtoX*	methanethiol degradation	P
SPO1017-1021		lactate ABC transporter	P
SPO1814-1816		dicarboxylate TRAP transporter 1	P
SPO2431-2433		dicarboxylate TRAP transporter 2	P
SPO2545, 2546		dicarboxylate TRAP transporter 3	P
SPO2626-2628		dicarboxylate TRAP transporter 4	P
SPO3693, 3695		dicarboxylate TRAP transporter 5	P
SPOA0249-A0251		dicarboxylate TRAP transporter 6	P
SPOA0372, A0373		dicarboxylate TRAP transporter 7	P
SPOA0255-0258	*rbsA,C-2*	ribose ABC transporter	P
SPO0608-0612		sugar ABC transporter 1	P
SPO0649-0651		sugar ABC transporter 2	P
SPO2427, 2436		sugar degradation	P
SPO1450, 1451		aromatic compound degradation	P
SPO3696, 3698	*fcs*	ferulic acid degradation	P
SPOA0044	*pcaH*	protocatechuate degradation	P
SPOA0434	*catD*	catechol degradation	P
SPO0753-0757	*paaG,H,J,K*	phenylacetate ABC transporter	
SPO0370, 0693	*ccrA*	ethylmalonyl-CoA pathway	L
C storage			
SPO1015		PHA degradation	P
SPO0112	*phaC*	PHA synthesis	L
Nutrients			
SPO1554		ammonium transporter	P
SPO1578	*amtB*	ammonium transporter	P
SPO3723	*amt-2*	ammonium transporter 2	P
SPO1709, 1710	*urtA,B*	urea ABC transporter	P
SPO1712-1718	*ureA,B,C,D,E, F,G*	urea degradation	P
SPO2087,2088	*ntrB,C*	nitrogen regulation protein NtrC	P
SPO3724	*glnB-2*	nitrogen regulatory protein P-II	P
SPO1789, 1790		sulfate/tungstate ABC transporter	P
SPO2093	*amt-1*	ammonium transporter 1	-
SPO1948-1951	*pstA,B,C,S*	phosphate ABC transporter	-
SPO1953	*phoB*	phosphate regulatory protein	-
SPO0468,0371,0473	*phnG,J,L*	alkylphosphonate utlitzation protein	-
SPO0077	*ptsN*	PTS IIA-like nitrogen-regulatory protein PtsN	L
SPO3058		sulfate transporter	L
SPO1256	*ppk2*	polyphosphate synthesis	L
Metals			
SPOA0218		magnesium transporter	P
SPO3663-3666		manganese ABC transporter	
SPO0382		iron transporter	L
SPO0985-0987	*zur; znuA,C*	zinc ABC transporter	L
Vitamins/coenzymes			
SPO1500-1504	*pqqA,B,C,D,E*	coenzyme PQQ synthesis	P
SPO0045- 0047	*thiD,G*	thiamine biosynthesis; vitamin B1	L

SPO0410	*moeB*	molybdopterin biosynthesis	L
SPO1374, 3243, 3245	*nadA,C*	nicotinamide biosynthesis; vitamin B3	L
SPO1761, 1762	*ribB,A,H-2*	riboflavin biosynthesis, vitamin B2	L
SPO1921		PLP biosyntheis; vitamin B6	L
SPO3200	*acpS*	pantothenate biosynthesis; vitamin B5	L
SPO3338	*bioB,Y*	biotin synthesis, vitamin B8	L
SPO3903		folate biosynthesis	L
Biotic Interactions			
Antagonism			
SPO0227	*paxA*	RTX toxin	P
SPO0838, 0841, 0842, 0846-0849,0852-0854		polyketide synthesis	P
SPO1649		invasion protein IbeA	P
SPO2352, 2586, 2828		type I secretion	P
SPO2652a		Polyketide cyclase	P
SPO2718		RND efflux	P
SPOA0342, 0343		afsA-like killing gene	
SPO1489		auxin biosynthesis	L
SPO2744		auxin efflux	L
SPO3534-3537	*yejA,B,E,F*	microcin C ABC transporter	L
Antibiotic/toxin resistance			
SPO0645		multidrug efflux	P
SPO1397, 1398		acriflavin resistance	P
SPO1641, 2502, 2696, 3036		beta-lactam resistance	P
SPO0027, 0028		multidrug efflux	L
SPO1191	*gst*	glutathione S-transferase detoxification	L
SPO1093, 2071		drug resistance transporter	L
SPO1430		antibiotic efflux	L
SPO2331, 2332		efflux transporter	L
SPO2852	*czcN*	cobalt-zinc-cadmium resistance	L
SPOA0049, A0054, A0217	*norC; nosL,R*	nitric oxide reduction	L
Motility			
SPO0170-0183	*flhA,B; flgA,B,C,F, G,H; fliE,I,Q,R flgE,I,K,L; fliF,L,N,P;*	flagellar biosynthesis	P
SPO0193-0203	*motA,B*	flagellar biosynthesis	P
Quorum sensing			
SPO0371, 2287	*luxR-1; luxI*	quorum sensing	P
Biomolecule transfer			
SPO3086-3092	*cpaB,C,E,F; ompA; gspF*	pilus assembly	P
Horizontal gene transfer			
SPO2255, 2256, 2259-2260, 2263-2264, 2266-2270	*gta*	gene transfer agent	
Stress			
Oxidative			
SPO0314	*soxR*	redox-sensitive transcriptional activator SoxR	P
SPO0917		glyoxalase proteins	P

SPO1755, 1757	*kpsC,S*	capsular polysaccharide export	P
SPO0401, 1328, 3208, 3494	*gshB; gor*	glutathione synthesis, reductase	L
SPO0442, 0903, 3383, 3423, 3874	*trxB*	thioredoxin proteins	L
SPO2127, 2466, 2566, 2570		glyoxalase proteins	P
SPO2340	*sodB*	superoxide dismutase, Fe	L
SPO3866	*senC*	redox regulator	L
Light			
SPO2034	*recA*	RecA	P
SPO2218, 3637	*uvrA,C*	UvrABC system	L
Osmolarity			
SPO3495	*mscL*	mechanosensitive channel protein	P
General			
SPO1275		cold shock family protein	P
SPO0406, 1409	*rpoH-1, rphH-2*	RNA polymerase sigma-32 factors 1,2	L
SPO0895, 3484		heat shock proteins, Hsp20 family 1,2	L

CHAPTER 5

NICHE DIFFERENTIATION OF TWO HIGHLY RELATED, ABUNDANT SPECIES OF STREAMLINED BLOOM-ASSOCIATED ROSEOBACTERS [1]

[1] Nowinski B, Preston CM, Scholin CA, Birch JM, Whitman WB, Moran, MA. To be submitted to *Environmental Microbiology*.

Abstract

Time-series sequencing of bacterial genes and transcripts over a 7-week coastal phytoplankton bloom revealed two highly-related Rhodobacteraceae taxa from the deeply-branching NAC11-7 lineage that have identical 16S rRNA amplicon sequences but sufficient genomic divergence to distinguish them at the species level. Although both were abundant throughout the bloom, each species maintained higher cell numbers than the other at different bloom stages. Metapangenomic analysis resolved the abundance, temporal patterns, and genomic variation of 31 genomes from these two streamlined species that co-existed over the Fall 2016 Monterey Bay phytoplankton bloom. A total of 2,296 genes were detected across the two species, of which 215 were unique to one or the other. Unique genes indicated key niche dimensions underlying their ecological differentiation that included substrate acquisition (of polyamines, sugars, and carboxylic acids) and vitamin synthesis (of riboflavin). Metatranscriptomic data allowed comparisons of gene expression levels for shared genes as another indication of niche differentiation. Genes with expression differences averaging >2-fold included those involved in trace element binding, phosphate transport, and transcriptional regulators. Our analyses revealed niche dimensions that differentiate these sympatric bacterial species that are globally distributed and frequently abundant in coastal and open ocean phytoplankton blooms.

Introduction

Members of the diverse roseobacter group of marine bacterioplankton are ubiquitous and abundant in surface seawater environments (Moran *et al.*, 2007; Newton *et al.*, 2010). Species within the group can vary dramatically in life history characteristics, with most cultured isolates

possessing large genomes enriched with characteristics of copiotrophs and r-strategists, while the more abundant free-living roseobacters have smaller genomes and features of K-strategists and are often difficult to cultivate (Luo *et al.*, 2012). Of the latter, four main lineages have been studied (CHAB-I-5, DC5-80-3, SAG-O19, and NAC11-7) that together represent >60% of the global distribution of pelagic roseobacters (Zhang *et al.*, 2016). Key life history strategies for these lineages include photoheterotrophy and organic sulfur metabolism, and many of the genomes from these lineages lack genes facilitating cell-cell interactions that are typical of copiotrophic relatives, including quorum sensing, type IV and type VI secretion, and a gene transfer agent system (Zhang *et al.*, 2016).

Members within these streamlined lineages often exhibit low divergence in their 16S rRNA gene sequences (Buchan *et al.*, 2005; Giebel *et al.*, 2011; Zhang *et al.*, 2016), such that marker gene studies can conflate closely related species and blur evidence of niche differentiation (Rodriguez-R *et al.*, 2018). Metagenomic studies offer a work-around to this issue by delineating sequence-discrete groups, defined as environmental populations with >95% average nucleotide identity (ANI) within a cluster and 70-85% ANI with the nearest relative (Konstantinidis and DeLong, 2008; Caro-Quintero and Konstantinidis, 2012). These sequence-discrete populations provide a foundation for understanding the mechanisms that define and maintain bacterial species, and from which insights into ecological niches can be drawn (Rodriguez-R and Konstantinidis, 2014).

In the fall of 2016, a massive dinoflagellate bloom occurred in Monterey Bay, CA, USA that was dominated by the red tide dinoflagellate species *Akashiwo sanguine*a (Kiene *et al.*, 2019). An autonomous instrument that collected and filtered seawater samples for microbial community nucleic acid analysis (the Environmental Sample Processor (Scholin *et al.*, 2017))

was deployed in the bay over a 52-day period from bloom peak in late September through bloom demise in mid-November. Metagenomic sequencing and genomes recovered from these community samples indicated two sequence-discrete, closely-related roseobacter species from the NAC11-7 lineage. As these represented the two most abundant bacterial species in the bloom microbial community, we undertook an examination of the genome characteristics and gene expression patterns underlying niche differentiation in these sympatric taxa.

Results/Discussion

Abundant sequence-discrete, highly related roseobacters

We used 16S rRNA gene sequencing, metagenome-assembled genomes (MAGs), single amplified genomes (SAGs), and quantitative metagenomics and metatranscriptomics to characterize members of the surface seawater (6 m depth) bacterioplankton community on 41 days during the Fall 2016 Monterey Bay bloom. Sequencing of the V4 region of the 16S rRNA gene revealed a dominant amplicon sequence variant (ASV) that averaged 15% of the sequences recovered from the community, twice that of the next most abundant ASV (Fig. 5.1). This ASV aligned with 100% identity to the sequence of the roseobacter *Rhodobacterales* bacterium HTCC2255, an isolate genome from the NAC11-7 roseobacter lineage found previously to dominate sequence libraries from nearshore surface seawater in Monterey Bay and the North Sea (Riedel *et al.*, 2010). This bacterium is also an important member of open ocean bacterioplankton communities (Yooseph *et al.*, 2010) and typically associates with phytoplankton blooms (Buchan *et al.*, 2005; Wagner-Döbler and Biebl, 2006; West *et al.*, 2008; Rich *et al.*, 2011). HTCC2255 represents the deepest branching member of the roseobacter group (Luo and Moran, 2014; Simon *et al.*, 2017) and has characteristics distinct from most roseobacter

isolates. It possesses one of the smallest genomes and lowest %G+C content among roseobacters (Luo and Moran, 2014) and is one of only two roseobacters with a proteorhodopsin gene for capturing energy from sunlight (Newton *et al.*, 2010; Sun *et al.*, 2017).

Metagenomic and qPCR surveys have found abundant gene sequences recruiting strongly to the HTCC2255 genome but with variable sequence identities (Ottesen *et al.*, 2011; Rich *et al.*, 2011; Yao *et al.*, 2011; Varaljay *et al.*, 2015). It is unclear if the sequence variation represents diverse populations within the HTCC2255 species, or if multiple coherent species with high identity to HTCC2255 exist. The lack of other isolates from the HTCC2255 group and the loss of the original isolate from culture (Luo *et al.*, 2014) has made studies of the phylogenetic and functional attributes of this novel branch of the roseobacter group challenging. However, MAGs and SAGs captured during the progression of the Fall 2016 Monterey Bay bloom yielded two genome clusters that clearly separated into sequence-discrete groups, one representing species HTCC2255, and the other representing a highly-related sympatric species. This provided a unique opportunity for detailed ecological genomics analysis of this taxon.

MAGs were assembled independently from 84 individual metagenomic libraries using nucleotide composition and read recruitment patterns across all samples. Bins from all samples were then dereplicated into clusters of near-identical genomes with >75% completeness from which the best genome was selected. Two of these MAGs represented internal standard genomes from genomic DNA of *Thermus thermophilus* and *Blautia producta* added to each sample filter at the start of nucleic acid extraction; these MAGs were nearly identical in length and identity to their respective reference genomes (*T. thermophilus*: 99.58% complete, 99.97% ANI; *B. producta*: 99.36% complete, 99.98% ANI), and provided confidence in the assembly and binning of Monterey Bay MAGs.

The dereplication process found 50 high quality MAGs closely related to the HTCC2255 isolate genome. Forty-two MAGs had near-identical similarity with each other (>99.5% ANI), of which the best (1.8 Mbp; MB-MAG-HTCC2255; Table 5.1) had 99.1% ANI to the HTCC2255 isolate genome. The other 8 MAGs had >99% ANI with each other, of which the best MAG (1.7 Mbp) was only 83.7% ANI to the HTCC2255 isolate genome. Thus two closely related species with identical 16S rRNA sequences in the V4 region were present. The representative MAG from the second species (hereafter the MB-C16 clade) had high similarity (>98.6% ANI) to 3 SAGs sampled at the same location in Fall 2014 (Nowinski *et al.*, 2019a) (Table 5.1). SAGs were also generated from the Fall 2016 bloom, and 14 of these fell within the HTCC2255 clade (>95% ANI with HTCC2255) while 11 fell within the MB-16 clade (>95% ANI to SAG SCGC-AG-151-C16).

Pairwise ANI-comparison of the 2 dereplicated MAGs, 28 SAGs, and HTCC2255 isolate genome confirmed that the HTCC2255 and MB-C16 clades represent sequence-discrete species (Fig. 5.2A). Full-length 16S rRNA gene sequences were 99.9% identical between the two ANI-delineated species, and the v4 region shares 100% identity, rendering it impossible to track these species separately in the ocean by routine methods for small subunit rRNA amplicon sequencing. These combined analyses suggested that two highly related, deeply-branching roseobacter species co-occurred through the full set of samples spanning the 7 weeks in which the 2016 Monterey Bay bloom community was inventoried.

An ANI of 83% is at the upper limit of what typically delineates bacterial and archaeal species. In a comparison of pairwise ANI values between all bacterial and archaeal genomes in the NCBI Genome database, only 2.5% of pairwise ANI values that were in the species range (76-100% ANI) were higher than 83% (Jain *et al.*, 2018; Cohan, 2019). Further, this level of

interspecies sequence similarity has been found only rarely in the same environmental sample (Caro-Quintero and Konstantinidis, 2012). Recruitment mapping of the Monterey Bay metagenomic data to the HTCC2255 isolate genome displays peaks at 100% and ~85% nucleotide identity (Fig. 5.2B), highlighting the co-occurrence of populations of HTCC2255 and MB-C16.

Genome abundance

Abundances of the HTC2255 and MB-16 clades were tracked though the bloom using 84 metagenomic datasets (n=2 or 3 for 35 of the 41 sample dates, n=1 for 6). The set of 31 genomes contained 19,520 open reading frames indicating protein-encoding genes (HTCC2255 = 9,937; MB-C16 = 9,583), and metagenomic and metatranscriptomic reads were mapped to this gene set. To calculate genes and transcripts per liter of bloom seawater for each species at each date, the recovery ratio of the internal standards (either standard genomes or artificial mRNAs) in the metagenomic libraries was applied to the coverage of the 19,250 protein-encoding genes.

Overall, abundances of genomes from the two species tracked each other closely across the sample dates, averaging 9.4×10^7 and 9.6×10^7 genomes L seawater $^{-1}$ for HTCC2255 and MB-C16 genomes respectively (Fig. 5.3). Yet from Sept. 26 through Oct. 12, environmental conditions favored the growth of HTCC2255, as it averaged 30% more genomes L seawater $^{-1}$ than MB-C16. From Oct. 13 to Nov. 16, conditions instead favored the growth of MB-C16, which averaged 37% more genomes L seawater $^{-1}$ than HTCC2255 (Fig. 5.3). The niche overlap between the species is likely high given the nearly equal numbers and highly similar dynamics through the bloom regime, yet a mid-study shift in some aspect of the seawater environment differentially affected their ecological success. Environmental data collected through the

declining bloom were analyzed against the ratio of the species' abundances (HTCC2255/MB-C16). Significant positive correlations were found with parameters that decreased with bloom decline, including chlorophyll *a* concentrations, the biomass of *A. sanguinea*, and the concentration of dimethylsulfoniopropionate (DMSP), a dinoflagellate metabolite that both bacteria can catabolize (Fig. 5.S1). To look for insights into which physical situations, chemical conditions, and/or biotic interactions might underlie this differential success, we examined divergences in species genome content and in situ gene expression.

Niche overlap and partitioning – genome content

Pangenomic clustering (Delmont and Eren, 2018) of the 19,250 protein-encoding genes from the 31 genomes yielded 2,296 shared gene clusters (detected in 2 or more genomes) and 576 singleton genes (detected in only 1 genome) (Fig. 5.4). Within the shared gene clusters, 2,081 (91%) were detected in both species (referred to as "core" hereafter), while 124 (5%) were in HTCC2255 genomes only ("HTCC2255 unique"; Table 5.S1) and 91 (4%) were in MB-C16 genomes only ("MB-C16 unique"; Table 5.S2). As expected, core genes encoded proteins involved in basic cellular processes, including ribosomal proteins, chaperonins, translation elongation factors, and ATP synthase (Table 5.2). The core gene set included many transporters and the ability to use a broad range of nitrogen-containing compounds, including polyamines and taurine. Genes encoding urease were also found in the core genes, somewhat unexpected given that the 95% complete HTCC2255 isolate genome is missing these genes. Genomes from both species contain genes degrading the organosulfur metabolite DMSP, including both the demethylation pathway which routes the reduced sulfur to assimilation, and the cleavage pathway which produces the volatile sulfur gas dimethylsulfide. Both species also possess

proteorhodopsin genes that work as sunlight-powered proton pumps, likely playing a role in the global success of the these species in the upper ocean.

The gene clusters identified as unique to either the HTCC2255 or MB-C16 species were analyzed for insights into genomic features that distinguish the two species, since these represent resources or environmental factors that contribute to niche partitioning (Table 5.S1, Table 5.S2). We identified 40 transporter-related gene clusters that were unique to one or the other species. In HTCC2255 genomes, a unique spermidine/putrescine ABC transporter in the same transporter neighborhood as a unique gene involved in polyamine deamination (PuuC) was observed. The HTCC2255 genomes also have two unique sugar transporters whose annotations and gene neighborhoods suggest fucose and an aldopentose (five carbon sugar with an aldehyde functional group such as ribose, xylose, or arabinose) as possible substrates. Likewise, the MB-C16 genomes have two unique sugar transporters, in this case with annotations suggesting sugar alcohol substrates; ribulose and mannitol are candidates based on unique catabolic genes in the transporter neighborhoods. Lastly, three unique TRAP transporters were found, all of which were unique to HTCC2255 genomes. The target substrates are unknown, but typically these transporters take up small mono- or di-carboxylic acids.

Other unique genes included a manganese transporter in MB-C16 genomes, and a second copy proteorhodopsin gene in two HTCC2255 SAGs that were distinct from the core proteorhodopsin present in genomes of both species (~35% similar); the second-copy proteorhodopsins had similarity to proteorhodopsins in the order Rhizobiales. The core pathway for taurine degradation takes taurine to sulfoacetaldehyde via the *xsc* gene, which is typical of roseobacters. However, four HTCC2255 genomes have an alternate pathway that uses *tauD* instead to degrade taurine to 2-aminoacetaldehyde. Additionally, HTCC2255 genomes have two

copies of *ribH,* the gene encoding the penultimate step in riboflavin biosynthesis, while MB-C16 genomes only possess one *ribH* copy; this gene copy disparity might manifest physiologically as differences in the rate of riboflavin biosynthesis.

We also looked for transcriptional regulators among the unique genes, since differential gene regulation could facilitate niche differentiation between HTCC2255 and MB-C16 without loss or gain of functional genes. Thirteen regulators appeared among the unique genes, 10 in HTCC2255 genomes (Table 5.S1) and 3 in MB-C16 genomes (Table 5.S2). For most there is little information available on what genes they control, although one unique MB-C16 regulator (*cueR*) is annotated to control copper homeostasis in bacterial cells (Grass and Rensing, 2001). This regulator is located upstream and trans to a multi-copper oxidase unique to MB-C16, suggesting a novel function mediated by a copper-containing oxidase. These unique genes suggest that niche differentiation between these sympatric species may involve utilization of compounds in the diverse sugar and polyamine metabolite pools, as well as potential differences in interacting with metals and generating energy from sunlight.

Niche overlap and partitioning – gene expression

Gene abundance (genes L^{-1}) and expression (transcripts L^{-1}) were computed for each gene cluster and used to calculate gene expression ratios (transcripts L^{-1} / genes L^{-1}). Among the genes exhibiting the highest expression ratios were core genes important for central cellular processes such as ribosomal protein synthesis, protein folding, and ATP generation, as is typical (Table 5.2). Other highly-expressed core gene classes enabled acquisition of carbon and nitrogen. Both species also had high expression of proteorhodopsin genes.

Some core gene clusters were differentially expressed. Those with higher relative expression in HTCC2255 genomes included a lipoprotein and glycosyltransferase implicated in lipopolysaccharide biosynthesis , an inorganic phosphate transporter, and choline dehydrogenase (Table 5.S3). Genes with higher expression in MB-C16 genomes included copper binding and transport, two repair-related genes (peptide methionine sulfoxide reductase *msrB* and a selenium-binding protein), along with sulfur oxidation gene soxS.

Conclusions – The globally distributed NAC11-7 clade is regularly associated with phytoplankton-rich coastal waters. This was the case for member species HTCC2255 and MB-C16 in the Monterey Bay bloom in Fall 2016, where they not only dominated the Roseobacter 16S rRNA amplicons, but accounted for as much as 25% of the total amplicon pool. Although these streamlined bacteria are not amenable to isolation, the time-series metagenomic and metatranscriptomic sequencing provided an extended window into their genome content, gene expression, and abundance during this large and dynamic bloom event. Differences involving both acquisition of unique genes and differential regulation of shared genes emerged as important determinants of ecological niche partitioning between these highly related sympatric species.

Methods

Internal standard addition, DNA and RNA sequencing, and assembly of libraries representing the microbial community were carried out as described previously(Nowinski *et al.*, 2019b). For each metagenomic assembly, Bowtie2 (2.3.4.1) (Langmead and Salzberg, 2012) was used to map reads from all metagenomic samples to contigs to generate coverage patterns of the

contigs across the time-series. The contigs were binned using MetaBAT 2.12.1 (Kang *et al.*, 2015), which incorporated the coverage patterns across all metagenomic samples to generate genomic bins for each sample. All bins were dereplicated using dRep (Olm *et al.*, 2017). Single-cell genomes (SAGs) were sampled from surface seawater as described previously (Nowinski *et al.*, 2019a). Estimates of genome completeness were generated using CheckM v1.0.12 (Parks *et al.*, 2015). Average Nucleotide Identity (ANI) between all genomes was calculated using the ANI/AAI-Matrix program as part of the enveomics collection toolbox (Rodriguez-R and Konstantinidis, 2016). BBmap v38.73 (Bushnell, 2014) was used to calculate the percent identity distribution of reads mapping to the HTCC2255 isolate genome with a minimum alignment identity cutoff of 0.60.

The metagenomic and metatranscriptomic reads for each sample were mapped to an index consisting of the 31 genomes using Bowtie2. Anvi'o v6.1 (Eren *et al.*, 2015) was used to create a database of the 31 genomes consisting of DNA and amino acid sequences. Genes were annotated using 1) the program 'anvi-run-ncbi-cogs' with the December 2014 release of the Clusters of Orthologous Groups (COGs) database (Tatusov *et al.*, 2000), 2) eggNOG-mapper v1.0.3 with the eggnog v4.5.1 database (Huerta-Cepas *et al.*, 2017), 3) KofamKOALA v. 2019-03-20 (Aramaki *et al.*, 2020), and 4) Reciprocal Best Hits using blastp with the well-characterized roseobacter *Ruegeria pomeroyi*, with E value cutoffs of 1e-5, and an identity > 30%. Protein-encoding genes were clustered based on sequence homology to form the pangenome using the program 'anvi-pan-genome' with parameters '--use-ncbi-blast', '--minbit 0.5', and '--mcl-inflation 10'. The pangenome was visualized using the program 'anvi-display-pan'.

Acknowledgements

We thank Shalabh Sharma, A. Murat Eren, and the Georgia Advanced Computing Resource Center at the University of Georgia for bioinformatic assistance.

References

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020) KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**: 2251-2252.

Buchan, A., González, J.M., and Moran, M.A. (2005) Overview of the marine Roseobacter lineage. *Appl Environ Microbiol* **71**: 5665-5677.

Bushnell, B. (2014) BBMap: a fast, accurate, splice-aware aligner. No. LBNL-7065E. In: Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Caro-Quintero, A., and Konstantinidis, K.T. (2012) Bacterial species may exist, metagenomics reveal. *Environmental Microbiology* **14**: 347-355.

Cohan, F.M. (2019) Systematics: The Cohesive Nature of Bacterial Species Taxa. In: Cell Press, pp. R169-R172.

Delmont, T.O., and Eren, E.M. (2018) Linking pangenomes and metagenomes: The Prochlorococcus metapangenome. *PeerJ* **2018**: e4320-e4320.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and Delmont, T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**: e1319.

Giebel, H.-A., Kalhoefer, D., Lemke, A., Thole, S., Gahl-Janssen, R., Simon, M., and Brinkhoff, T. (2011) Distribution of Roseobacter RCA and SAR11 lineages in the North Sea and characteristics of an abundant RCA isolate. *The ISME journal* **5**: 8-19.

Grass, G., and Rensing, C. (2001) Genes involved in copper homeostasis in Escherichia coli. *Journal of bacteriology* **183**: 2145-2147.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C., and Bork, P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular biology and evolution* **34**: 2115-2122.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**: 1-8.

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

Kiene, R.P., Nowinski, B., Esson, K., Preston, C., Marin III, R., Birch, J. et al. (2019) Unprecedented DMSP Concentrations in a Massive Dinoflagellate Bloom in Monterey Bay, CA. *Geophysical Research Letters* **46**: 12279-12288.

Konstantinidis, K.T., and DeLong, E.F. (2008) Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *The ISME journal* **2**: 1052-1065.

Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**: 357.

Luo, H., and Moran, M.A. (2014) Evolutionary ecology of the marine Roseobacter clade. *Microbiol Mol Biol Rev* **78**: 573-587.

Luo, H., Löytynoja, A., and Moran, M.A. (2012) Genome content of uncultivated marine Roseobacters in the surface ocean. *Environmental Microbiology* **14**: 41-51.

Luo, H., Swan, B.K., Stepanauskas, R., Hughes, A.L., and Moran, M.A. (2014) Comparing effective population sizes of dominant marine alphaproteobacteria lineages. *Environmental microbiology reports* **6**: 167-172.

Moran, M.A., Belas, R., Schell, M., González, J., Sun, F., Sun, S. et al. (2007) Ecological genomics of marine roseobacters. *Appl Environ Microbiol* **73**: 4559-4569.

Newton, R.J., Griffin, L.E., Bowles, K.M., Meile, C., Gifford, S., Givens, C.E. et al. (2010) Genome characteristics of a generalist marine bacterial lineage. *The ISME journal* **4**: 784-798.

Nowinski, B., Motard-Côté, J., Landa, M., Preston, C.M., Scholin, C.A., Birch, J.M. et al. (2019a) Microdiversity and temporal dynamics of marine bacterial dimethylsulfoniopropionate genes. *Environmental microbiology* **21**: 1687-1701.

Nowinski, B., Smith, C.B., Thomas, C.M., Esson, K., Marin, R., Preston, C.M. et al. (2019b) Microbial metagenomes and metatranscriptomes during a coastal phytoplankton bloom. *Scientific data* **6**: 1-7.

Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017) dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal* **11**: 2864-2868.

Ottesen, E.A., Marin, R., Preston, C.M., Young, C.R., Ryan, J.P., Scholin, C.A., and DeLong, E.F. (2011) Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *The ISME journal* **5**: 1881-1895.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**: 1043-1055.

Rich, V.I., Pham, V.D., Eppley, J., Shi, Y., and DeLong, E.F. (2011) Time-series analyses of Monterey Bay coastal microbial picoplankton using a 'genome proxy'microarray. *Environmental Microbiology* **13**: 116-134.

Riedel, T., Tomasch, J., Buchholz, I., Jacobs, J., Kollenberg, M., Gerdts, G. et al. (2010) Constitutive expression of the proteorhodopsin gene by a flavobacterium strain representative of the proteorhodopsin-producing microbial community in the North Sea. *Appl Environ Microbiol* **76**: 3187-3197.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2014) Bypassing cultivation to identify bacterial species. *Microbe* **9**: 111-118.

Rodriguez-R, L.M., and Konstantinidis, K.T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. In: PeerJ Preprints.

Rodriguez-R, L.M., Castro, J.C., Kyrpides, N.C., Cole, J.R., Tiedje, J.M., and Konstantinidis, K.T. (2018) How much do rRNA gene surveys underestimate extant bacterial diversity? *Applied and Environmental Microbiology* **84**.

Scholin, C.A., Birch, J., Jensen, S., Marin, R., Massion, E., Pargett, D. et al. (2017) The quest to develop ecogenomic sensors a 25-year history of the environmental sample processor (ESP) as a case study. In: Oceanography Society, pp. 100-113.

Simon, M., Scheuner, C., Meier-Kolthoff, J.P., Brinkhoff, T., Wagner-Döbler, I., Ulbrich, M. et al. (2017) Phylogenomics of Rhodobacteraceae reveals evolutionary adaptation to marine and non-marine habitats. *The ISME journal* **11**: 1483-1499.

Sun, Y., Zhang, Y., Hollibaugh, J.T., and Luo, H. (2017) Ecotype diversification of an abundant Roseobacter lineage. *Environmental microbiology* **19**: 1625-1638.

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

Varaljay, V.A., Robidart, J., Preston, C.M., Gifford, S.M., Durham, B.P., Burns, A.S. et al. (2015) Single-taxon field measurements of bacterial gene regulation controlling DMSP fate. *The ISME journal* **9**: 1677-1686.

Wagner-Döbler, I., and Biebl, H. (2006) Environmental biology of the marine Roseobacter lineage. *Annu Rev Microbiol* **60**: 255-280.

West, N.J., Obernosterer, I., Zemb, O., and Lebaron, P. (2008) Major differences of bacterial diversity and activity inside and outside of a natural iron-fertilized phytoplankton bloom in the Southern Ocean. *Environmental Microbiology* **10**: 738-756.

Yao, D., Buchan, A., and Suzuki, M.T. (2011) In situ activity of NAC11-7 roseobacters in coastal waters off the Chesapeake Bay based on ftsZ expression. *Environmental microbiology* **13**: 1032-1041.

Yooseph, S., Nealson, K.H., Rusch, D.B., McCrow, J.P., Dupont, C.L., Kim, M. et al. (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**: 60-66.

Zhang, Y., Sun, Y., Jiao, N., Stepanauskas, R., and Luo, H. (2016) Ecological genomics of the uncultivated marine Roseobacter lineage CHAB-I-5. *Appl Environ Microbiol* **82**: 2100-2111.

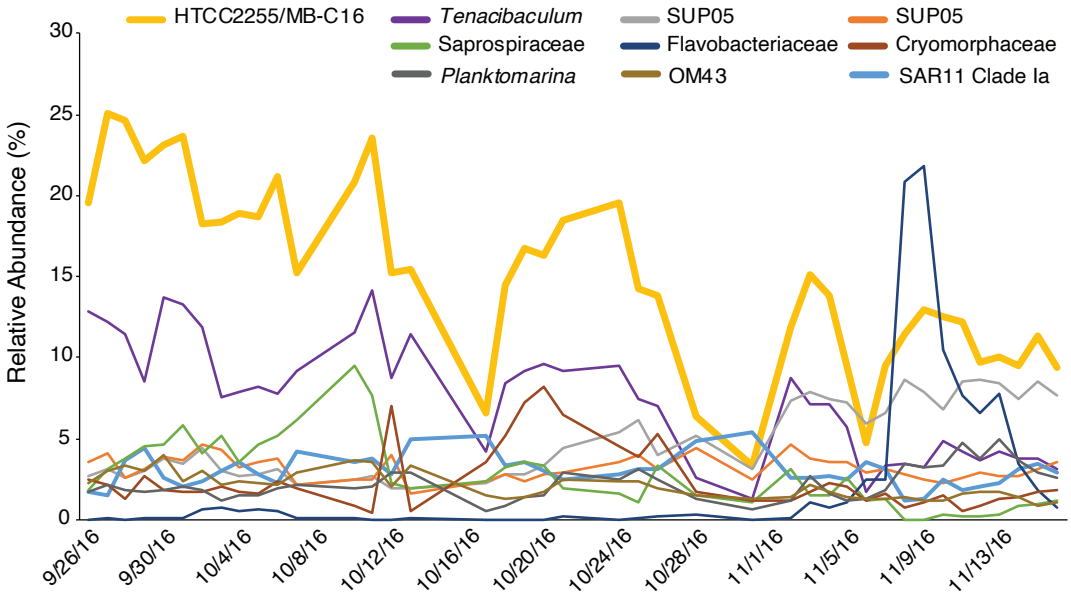


Figure 5.1. The 10 most abundant ASVs during the Fall 2016 Monterey Bay Bloom.

Table 5.1. Overview of genomes used in the analysis and CheckM reported statistics.

Genome	*Clade*	*Source*	*Genome size (bp)*	*Completeness*	*Redundancy*	*# contigs*	*GC*
HTCC2255	HTCC2255	Isolate	2224475	95.03	0.00	12	36.70
MB-MAG-HTCC2255	HTCC2255	MAG	1879461	89.00	0.20	92	36.60
Ga0315444	HTCC2255	SAG	667059	20.69	0.00	72	36.90
Ga0315366	HTCC2255	SAG	660517	31.05	0.00	91	38.10
Ga0315394	HTCC2255	SAG	591024	22.26	0.00	47	37.10
Ga0315349	HTCC2255	SAG	443698	16.73	0.00	50	37.80
Ga0315369	HTCC2255	SAG	432691	16.84	0.00	54	37.40
Ga0315399	HTCC2255	SAG	422354	19.07	0.00	54	37.60
Ga0315350	HTCC2255	SAG	419733	21.49	0.00	41	37.20
Ga0315447	HTCC2255	SAG	403395	13.79	0.00	58	37.60
Ga0315420	HTCC2255	SAG	343115	15.52	0.00	56	37.10
Ga0315456	HTCC2255	SAG	317688	15.03	0.00	42	37.00
Ga0315459	HTCC2255	SAG	250591	0.00	0.00	42	37.50
Ga0315455	HTCC2255	SAG	237086	7.37	0.00	39	36.60
Ga0315524	HTCC2255	SAG	178301	6.91	0.00	20	37.50
Ga0315414	HTCC2255	SAG	115890	10.34	0.00	22	38.90
MB-MAG-C16	MB-C16	MAG	1664462	77.23	1.37	180	36.40
SCGC-AG-151-C16	MB-C16	SAG	943511	39.66	0.00	45	36.20
SCGC-AG-145-N17	MB-C16	SAG	822380	44.01	0.00	63	36.80
SCGC-AG-145-A05	MB-C16	SAG	781374	39.66	0.86	43	36.10
Ga0315360	MB-C16	SAG	768688	34.27	0.00	86	37.00
Ga0315475	MB-C16	SAG	553517	23.27	1.98	61	36.70
Ga0315365	MB-C16	SAG	486281	20.56	1.01	73	37.40
Ga0315446	MB-C16	SAG	465132	23.27	0.00	53	37.10
Ga0315392	MB-C16	SAG	457233	9.56	0.00	58	37.20
Ga0315515	MB-C16	SAG	436111	15.15	0.15	60	37.20
Ga0315374	MB-C16	SAG	418842	14.58	0.00	54	36.80
Ga0315411	MB-C16	SAG	418778	13.79	0.00	57	37.30
Ga0315435	MB-C16	SAG	415818	17.95	0.00	60	37.00
Ga0315381	MB-C16	SAG	384573	19.18	0.00	67	37.50
Ga0315431	MB-C16	SAG	137046	8.33	4.17	28	37.20

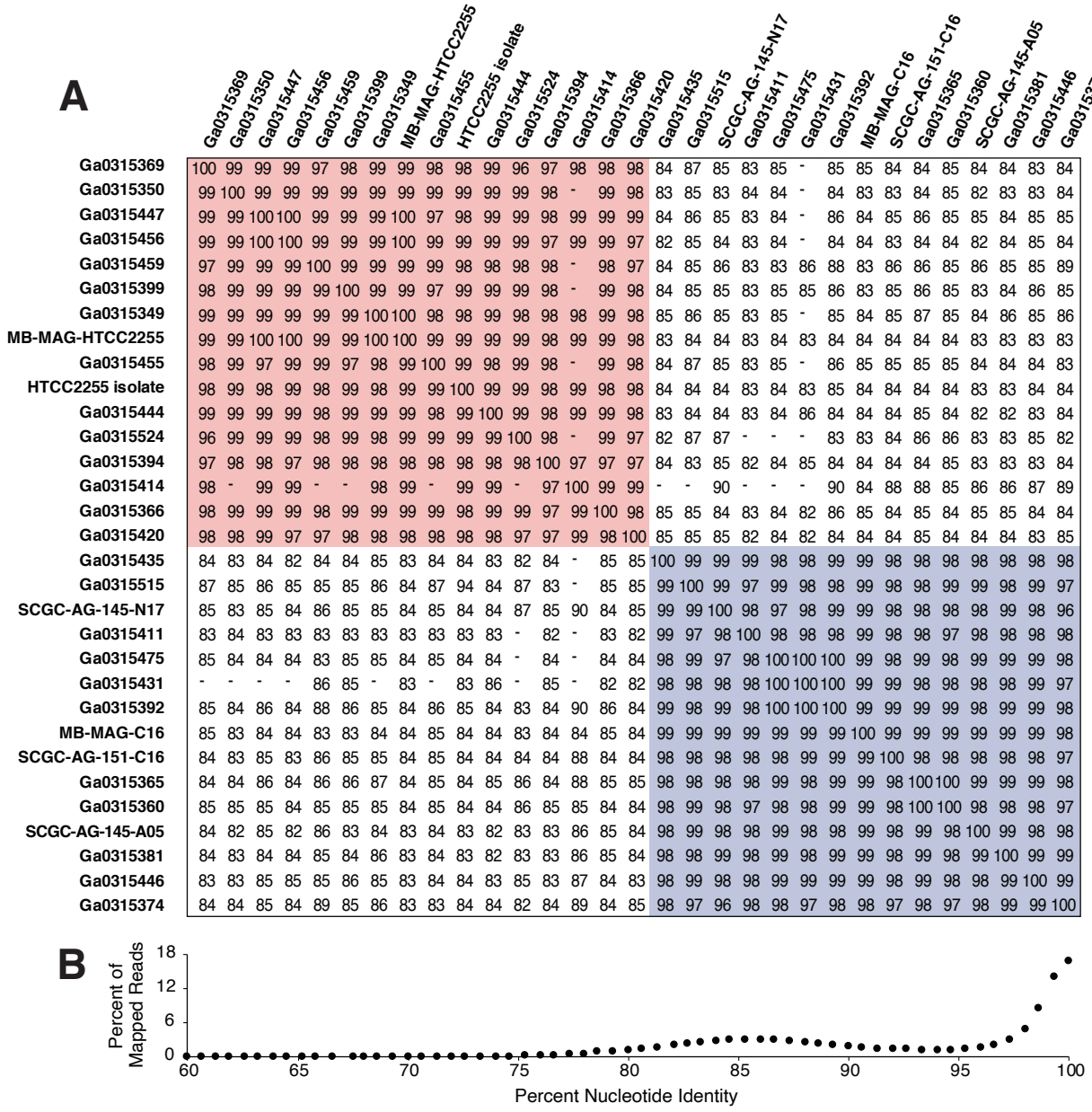


Figure 5.2. Two sequence-discrete clusters in the NAC11-7 lineage. (A) Estimated all-vs-all Average Nucleotide Identity (ANI) distances and similarity clustering in genomes used in this study. Value is missing if size of shared genes was too small to accurately calculate ANI. Red shading indicates all HTCC2255 clade genomes clustering at >95% ANI, and blue shading indicates all MB-C16 clade genomes clustering at >95% ANI. (B) The percent of mapped reads with the indicated nucleotide identity to the HTCC2255 isolate genome.

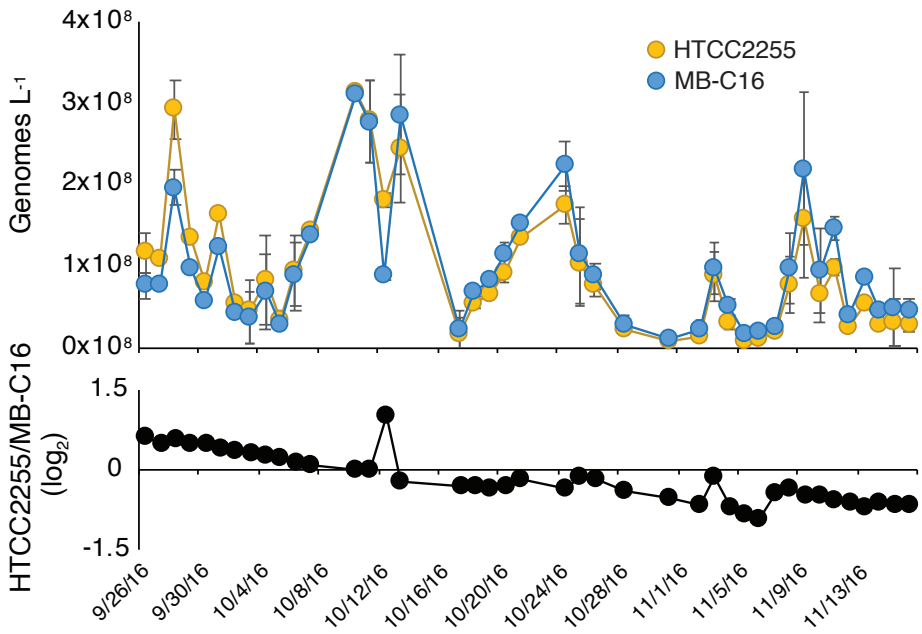


Figure 5.3. Genomes per liter seawater of the two NAC11-7 clades based on the number of DNA bases per liter seawater mapping to each clade.

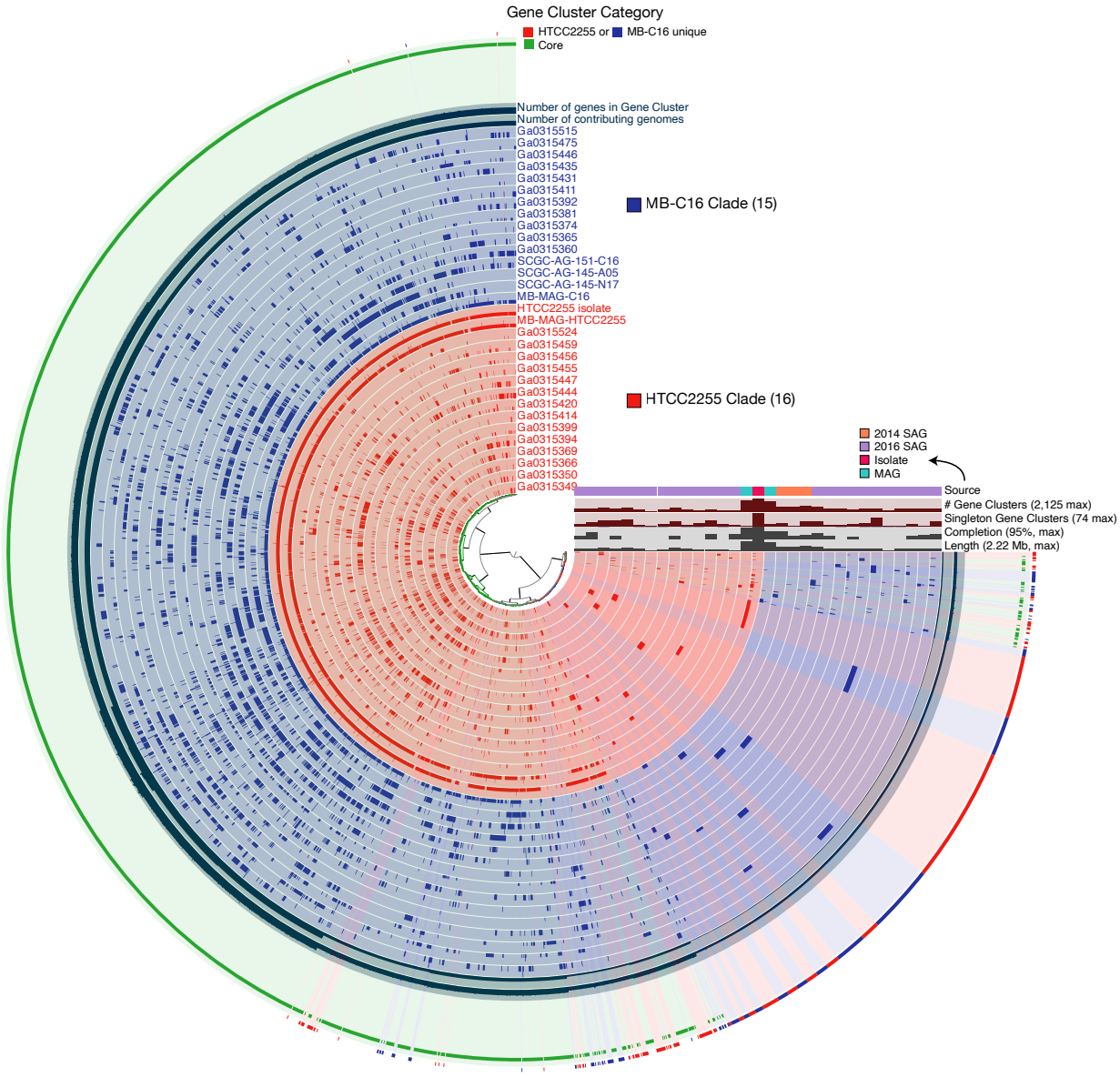


Figure 5.4. The HTCC2255 and MB-C16 pangenome, consisting of 31 genomes, 19,250 genes, and 2,872 gene clusters found in one or more genomes. Bars in the first 31 rings (starting with innermost ring) represent the occurrence of a gene cluster in a given genome. The next two layers describe the number of genomes contributing to the gene cluster and the total number of genes in the gene cluster. The outermost ring indicates if genes in each gene cluster are found in only one clade, or are core members of both clades.

Table 5.2. Top 30 core gene clusters by gene expression ratio, averaged across the two genomes.

gene_cluster_id	*mean t/g across clades*	*HTCC2255 t/g*	*MB-C16 t/g*	*genomes*	*genes*	*SPO*	*annotation*
GC_00002146	4.90	5.76	4.04	3	3	SPO3430	outer membrane porin
GC_00001306	1.77	1.92	1.63	8	8	SPO3484	heat shock protein, Hsp20 family
GC_00000804	1.70	1.15	2.25	9	9	SPO0229	ribosomal protein S21
GC_00001846	1.63	1.30	1.96	5	5	SPO1275	cold shock family protein
GC_00001223	1.56	1.62	1.49	8	8	SPO0861	xylose ABC transporter, periplasmic xylose-binding protein
GC_00001404	1.17	1.36	0.97	7	7	SPO2274	acyl carrier protein
GC_00000801	0.96	0.95	0.97	9	9	SPO0886	chaperonin, 10 kDa
GC_00000384	0.89	0.97	0.81	11	11	SPO3510	ribosomal protein L10
GC_00001002	0.88	0.84	0.92	9	9	SPO0887	chaperonin, 60 kDa
GC_00001452	0.84	0.87	0.81	7	7	NA	proteorhodopsin
GC_00000385	0.82	0.87	0.78	11	11	SPO0519	glutamate/glutamine/aspartate/asparagine ABC transporter, periplasmic substrate-binding protein
GC_00000628	0.80	0.83	0.78	10	10	SPO3509	ribosomal protein L7/L12
GC_00001304	0.67	0.64	0.71	8	8	SPO2441	glycine betaine/proline ABC transporter, periplasmic glycine betaine/proline-binding protein
GC_00000936	0.66	0.68	0.65	9	9	SPO3164	ATP synthase F1, alpha subunit
GC_00000254	0.66	0.78	0.53	11	12	SPO0043	chaperone protein DnaK
GC_00001092	0.57	0.61	0.53	8	8	SPO3235	ATP synthase F0, C subunit
GC_00000916	0.52	0.50	0.54	9	9	SPO3165	ATP synthase delta chain
GC_00000054	0.52	0.51	0.52	14	14	SPO3744	DNA-binding protein HU
GC_00001146	0.50	0.51	0.49	8	8	SPO3162	ATP synthase F1, beta subunit
GC_00001331	0.48	0.04	0.92	8	8	NA	Copper-binding protein CopC (methionine-rich)
GC_00000005	0.46	0.49	0.44	15	19	SPO3498	translation elongation factor Tu
GC_00000580	0.46	0.51	0.42	10	10	SPO3499	translation elongation factor G
GC_00001438	0.44	0.48	0.40	7	7	NA	integral membrane protein
GC_00001677	0.43	0.38	0.48	6	6	NA	extracellular solute-binding protein
GC_00000361	0.41	0.46	0.36	11	11	SPO0674	taurine ABC transporter, periplasmic taurine-binding protein
GC_00002032	0.39	0.40	0.38	4	4	NA	extracellular solute-binding protein
GC_00000834	0.39	0.41	0.37	9	9	SPO1383	cytochrome c oxidase, aa3-type, subunit I
GC_00001035	0.39	0.33	0.45	7	9	SPO3473	polyamine ABC trasnporter, periplasmic polyamine-binding protein
GC_00000703	0.34	0.32	0.37	9	10	SPO0379	sugar ABC transporter, periplasmic sugar-binding protein
GC_00001669	0.31	0.34	0.32	6	6	SPO0186	bordetella uptake gene family protein

Table 5.S1. Unique gene clusters in HTCC2255 genomes. t/g = transcripts L^{-1} / genes L^{-1}; g/L = genes L^{-1}; t/L = transcripts L^{-1}.

id	*t/g*	*g/L*	*t/L*	*genomes*	*genes*	*eggNOG*	*KofamKoala*	*COG*	*R. pomeroyi*
GC_00002190	0.47	5.74e+07	1.97e+07	2	2	NA	NA	Tripartite-type tricarboxylate transporter, receptor component TctC	NA
GC_00001844	0.39	1.50e+08	4.62e+07	5	5	Protein of unknown function (FYDLN_acid)	NA	Uncharacterized protein	SPO3623; hypothetical protein
GC_00002269	0.26	9.54e+06	1.94e+06	2	2	Ribose binding protein of ABC transporter	NA	ABC-type sugar transport system, periplasmic component	NA
GC_00001869	0.19	6.26e+07	1.14e+07	5	5	Spermidine putrescine ABC transporter, SBP	NA	Spermidine/putrescine-binding periplasmic protein	SPOA0381; spermidine/putrescine ABC transporter, periplasmic substrate-binding protein
GC_00002093	0.10	7.75e+07	6.97e+06	3	3	Aldehyde dehydrogenase	aldehyde dehydrogenase [EC:1.2.1.-]	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	SPO0084; betaine aldehyde dehydrogenase
GC_00002134	0.09	8.84e+07	5.65e+06	3	3	NA	NA	NA	NA
GC_00002098	0.07	3.97e+07	2.55e+06	3	3	Periplasmic binding protein LacI transcriptional regulator	fructose transport system substrate-binding protein	ABC-type sugar transport system, periplasmic component	NA
GC_00002230	0.06	6.86e+06	2.20e+05	2	2	NA	NA	NA	NA
GC_00002201	0.06	1.49e+07	8.42e+05	2	2	Bacterial extracellular solute-binding protein, family 7	NA	TRAP-type mannitol/chloroaromatic compound transport system, periplasmic component	SPO2606; bacterial extracellular solute-binding protein, family 7
GC_00002168	0.06	1.49e+08	7.01e+06	3	3	Dihydrolipoyl dehydrogenase	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	dihydrolipoamide dehydrogenase	SPO0340; 2-oxoglutarate dehydrogenase, E3 component, dihydrolipoamide dehydrogenase
GC_00002092	0.06	7.46e+07	3.43e+06	3	3	Protein of unknown function (DUF779)	uncharacterized protein	Uncharacterized conserved protein, DUF779 family	SPO3794; hypothetical protein
GC_00002006	0.05	8.46e+07	3.92e+06	4	4	NA	NA	NA	NA
GC_00002177	0.04	6.02e+07	1.84e+06	3	3	tricarboxylic transport membrane protein	putative tricarboxylic transport membrane protein	TctA family transporter	SPO2384; tricarboxylate transporter family protein
GC_00001962	0.04	6.25e+07	2.55e+06	5	5	spermidine/putrescine ABC transporter	nonpolar-amino-acid-transporting ATPase [EC:7.4.2.2]	ABC-type Fe3+/spermidine/putrescine transport systems, ATPase components	SPOA0382; spermidine/putrescine ABC transporter, ATP-binding protein
GC_00002288	0.04	8.73e+06	2.46e+05	2	2	ABC transporter	NA	Ribose/xylose/arabinose/galactoside ABC-type transport system, permease component	NA
GC_00002058	0.04	1.14e+08	4.26e+06	4	4	NA	NA	NA	NA
GC_00002183	0.04	1.08e+07	3.55e+05	2	2	Spfh domain band 7 family protein	flotillin	Uncharacterized membrane protein YqiK, contains Band7/PHB/SPFH domain	NA
GC_00001746	0.04	1.04e+08	3.73e+06	5	6	cell division protein; ABC transporter permease protein	cell division protein ZapA	Cell division protein ZapA, inhibits GTPase activity of FtsZ; Branched-chain amino acid ABC-type transport system, permease component	SPO1867; hypothetical protein; SPO1020; branched-chain amino acid ABC transporter, permease protein
GC_00002199	0.04	9.94e+06	3.31e+05	2	2	Bac_rhodopsin	NA	Bacteriorhodopsin	NA
GC_00002258	0.04	1.02e+07	3.05e+05	2	2	(ABC) transporter	NA	ABC-type sugar transport system, ATPase component	SPO0651; sugar ABC transporter, ATP-binding protein
GC_00002291	0.04	4.01e+07	8.48e+05	2	2	Trap dicarboxylate transporter, dctp subunit	NA	TRAP-type C4-dicarboxylate transport system, periplasmic component	SPOA0374; TRAP dicarboxylate transporter, DctP subunit
GC_00002076	0.04	6.65e+07	2.01e+06	4	4	ABC transporter permease protein	putative spermidine/putrescine transport system permease protein	ABC-type spermidine/putrescine transport system, permease component I	SPOA0383; spermidine/putrescine ABC transporter, permease protein
GC_00002181	0.03	5.98e+07	1.28e+06	2	2	methylitaconate delta2-delta3-isomerase	NA	2-Methylaconitate cis-trans-isomerase PrpF (2-methyl citrate pathway)	NA
GC_00002170	0.03	2.34e+06	5.43e+04	3	3	Polysaccharide biosynthesis protein	NA	NDP-sugar epimerase	NA

GC_00002000	0.03	6.57e+07	1.62e+06	4	4	ABC transporter permease protein	putative spermidine/putrescine transport system permease protein	ABC-type spermidine/putrescine transport system, permease component II	SPOA0384; spermidine/putrescine ABC transporter, permease protein
GC_00002268	0.03	8.51e+06	1.56e+05	2	2	Sodium:solute symporter family	NA	Na+/proline symporter	NA
GC_00001706	0.03	1.18e+08	2.76e+06	6	6	NA	NA	NA	NA
GC_00002296	0.03	3.63e+07	7.18e+05	2	2	DegT DnrJ EryC1 StrS aminotransferase	CDP-4-dehydro-6-deoxyglucose reductase, E1 [EC:1.17.1.1]	dTDP-4-amino-4,6-dideoxygalactose transaminase	NA
GC_00002213	0.03	1.15e+08	2.39e+06	2	2	Protein of unknown function (DUF2899)	NA	NA	NA
GC_00002016	0.03	6.32e+07	1.27e+06	4	4	Protein of unknown function (DUF861)	uncharacterized protein	Predicted enzyme of the cupin superfamily	SPO3326; hypothetical protein
GC_00002128	0.02	2.99e+07	5.50e+05	2	3	nad-dependent epimerase dehydratase	GDP-L-fucose synthase [EC:1.1.1.271]	Nucleoside-diphosphate-sugar epimerase	NA
GC_00002265	0.02	1.32e+08	2.67e+06	2	2	NA	NA	NA	NA
GC_00002198	0.02	6.11e+06	1.17e+05	2	2	NA	NA	NA	NA
GC_00002184	0.02	5.80e+07	9.66e+05	2	2	Uncharacterized protein conserved in bacteria (DUF2256)	NA	Uncharacterized protein	NA
GC_00002077	0.02	9.50e+07	1.64e+06	4	4	NA	NA	NA	NA
GC_00002192	0.02	7.82e+07	1.34e+06	2	2	NA	NA	NA	NA
GC_00001896	0.02	1.15e+08	1.78e+06	5	5	penultimate step in the biosynthesis of riboflavin	6,7-dimethyl-8-ribityllumazine synthase [EC:2.5.1.78]	6,7-dimethyl-8-ribityllumazine synthase (Riboflavin synthase beta chain)	SPO1762; riboflavin synthase, beta subunit
GC_00001950	0.02	1.14e+08	1.84e+06	5	5	KpsF GutQ family protein	arabinose-5-phosphate isomerase [EC:5.3.1.13]	CBS domain|D-arabinose 5-phosphate isomerase GutQ	SPO0082; arabinose 5-phosphate isomerase
GC_00002028	0.02	3.89e+07	6.45e+05	4	4	ABC transporter	fructose transport system permease protein	Ribose/xylose/arabinose/galactoside ABC-type permease	SPOA0254; ribose ABC transporter, permease protein
GC_00002042	0.02	1.04e+08	1.67e+06	4	4	NA	NA	NA	NA
GC_00002240	0.02	7.51e+07	9.44e+05	2	2	Inherit from proNOG: DoxX Family	NA	NA	NA
GC_00002010	0.02	1.30e+08	2.00e+06	4	4	NA	NA	NA	NA
GC_00002187	0.02	6.00e+07	7.11e+05	2	2	D-isomer specific 2-hydroxyacid dehydrogenase	NA	Phosphoglycerate dehydrogenase or related dehydrogenase	SPO3355; D-3-phosphoglycerate dehydrogenase
GC_00002113	0.02	1.20e+08	1.51e+06	3	3	Transcriptional regulator, gntR family	NA	DNA-binding transcriptional regulator, GntR family	SPO3621; transcriptional regulator, GntR family
GC_00002047	0.02	1.38e+08	1.75e+06	4	4	acetyltransferase	NA	Protein N-acetyltransferase, RimJ/RimL family	SPO3184; acetyltransferase, GNAT family
GC_00002232	0.02	6.70e+07	8.49e+05	2	2	BLUF	NA	Ribosome recycling factor	NA
GC_00002031	0.02	6.53e+07	9.55e+05	4	4	High affinity sulfate transporter (SulP)	sulfate permease, SulP family	Sulfate permease or related transporter, MFS superfamily	NA
GC_00001782	0.01	8.61e+07	1.16e+06	5	6	ABC transporter	NA	ABC-type lipoprotein export system, ATPase component	SPO2181; ABC transporter, ATP-binding protein
GC_00002193	0.01	5.77e+07	7.27e+05	2	2	regulator, luxr family	NA	DNA-binding response regulator, NarL/FixJ family, contains REC and HTH domains	SPO0161; DNA-binding response regulator, LuxR family
GC_00001860	0.01	1.33e+08	1.45e+06	5	5	NA	NA	NA	NA
GC_00002272	0.01	1.70e+07	2.12e+05	2	2	Domain of unknown function (DUF1989)	uncharacterized protein	Uncharacterized conserved protein YcgI, DUF1989 family	NA
GC_00002094	0.01	4.02e+07	4.62e+05	3	3	ABC transporter	fructose transport system ATP-binding protein	ABC-type sugar transport system, ATPase component	SPOA0255; sugar ABC transporter, ATP binding protein
GC_00002287	0.01	3.27e+07	3.17e+05	2	2	Trap dicarboxylate transporter, dctm subunit	NA	TRAP-type C4-dicarboxylate transport system, large permease component	SPOA0373; TRAP dicarboxylate transporter, DctM subunit
GC_00002264	0.01	2.12e+06	2.44e+04	2	2	repeat-containing protein	protein O-GlcNAc transferase [EC:2.4.1.255]	Tetratricopeptide (TPR) repeat	NA

GC_00001889	0.01	1.10e+08	1.02e+06	5	5	NA	NA	NA	NA
GC_00002215	0.01	1.13e+08	1.16e+06	2	2	Fatty acid desaturase	acyl-lipid omega-6 desaturase [EC:1.14.19.23 1.14.19.45]	Fatty acid desaturase	SPO2327; fatty acid desaturase
GC_00001853	0.01	1.32e+08	1.37e+06	5	5	16S rRNA (guanine527-N7)-methyltransferase	16S rRNA (guanine527-N7)-methyltransferase [EC:2.1.1.170]	16S rRNA G527 N7-methylase RsmG (former glucose-inhibited division protein B)	SPO0002; glucose-inhibited division protein B
GC_00002044	0.01	3.58e+07	2.52e+05	4	4	Aldehyde dehydrogenase	4-(gamma-glutamylamino)butanal dehydrogenase [EC:1.2.1.99]	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	SPOA0377; aldehyde dehydrogenase
GC_00001718	0.01	9.12e+07	8.92e+05	6	6	NA	NA	NA	NA
GC_00002084	0.01	6.43e+07	6.53e+05	4	4	Carboxymuconolactone decarboxylase family	NA	Alkylhydroperoxidase family enzyme, contains CxxC motif	NA
GC_00001972	0.01	6.78e+07	6.03e+05	5	5	NA	NA	Mannose-6-phosphate isomerase, cupin superfamily	SPO1596; DMSP lyase
GC_00002120	0.01	5.57e+07	5.61e+05	2	3	NA	NA	NA	NA
GC_00001998	0.01	9.59e+07	1.00e+06	4	4	Trap-t family transporter, periplasmic binding protein	putative tricarboxylic transport membrane protein	Tripartite-type tricarboxylate transporter, receptor component TctC	NA
GC_00002002	0.01	1.24e+08	1.04e+06	4	4	LysE type translocator	homoserine/homoserine lactone efflux protein	Threonine/homoserine/homoserine lactone efflux protein	SPO0290; transmembrane amino acid efflux protein
GC_00002008	0.01	4.26e+07	3.29e+05	4	4	Aldolase	NA	Fructose-bisphosphate aldolase class Ia, DhnA family	NA
GC_00002005	0.01	8.56e+07	6.56e+05	4	4	NA	NA	NA	NA
GC_00001703	0.01	1.27e+08	1.04e+06	6	6	shikimate dehydrogenase	shikimate dehydrogenase [EC:1.1.1.25]	Shikimate 5-dehydrogenase	SPO3891; shikimate 5-dehydrogenase
GC_00001887	0.01	1.10e+08	8.69e+05	5	5	ArsC family	NA	Arsenate reductase and related proteins, glutaredoxin family	NA
GC_00001997	0.01	7.60e+07	5.49e+05	4	4	Taurine dioxygenase	NA	Taurine dioxygenase, alpha-ketoglutarate-dependent	NA
GC_00001822	0.01	1.13e+08	7.94e+05	5	6	mandelate racemase muconate lactonizing	NA	L-alanine-DL-glutamate epimerase or related enzyme of enolase superfamily	SPO1594; mandelate racemase/muconate lactonizing enzyme, putative
GC_00002087	0.01	4.93e+07	2.78e+05	3	3	Haloacid dehalogenase type II	2-haloacid dehalogenase [EC:3.8.1.2]	FMN phosphatase YigB, HAD superfamily	NA
GC_00002125	0.01	3.59e+07	1.73e+05	3	3	epoxide hydrolase	NA	Pimeloyl-ACP methyl ester carboxylesterase	SPOA0376; epoxide hydrolase
GC_00002053	0.01	5.50e+07	3.58e+05	4	4	Protein of unknown function (DUF1498)	D-lyxose ketol-isomerase [EC:5.3.1.15]	D-lyxose ketol-isomerase	NA
GC_00002282	0.01	6.00e+07	3.21e+05	2	2	Ammonia monooxygenase	uncharacterized protein	Uncharacterized membrane protein AbrB, regulator of aidB expression	NA
GC_00002013	0.01	8.80e+07	5.48e+05	4	4	membrane	putative tricarboxylic transport membrane protein	TctA family transporter	SPO2384; tricarboxylate transporter family protein
GC_00002090	0.01	1.53e+08	9.41e+05	3	3	NA	NA	NA	NA
GC_00001995	0.01	8.17e+07	4.58e+05	4	4	NA	NA	NA	SPO3382; aldehyde dehydrogenase family protein
GC_00002021	0.01	1.07e+08	5.73e+05	4	4	Protein of unknown function (DUF3445)	NA	NA	SPO2298; hypothetical protein
GC_00002041	0.01	4.28e+07	2.04e+05	4	4	Oxidoreductase family, NAD-binding Rossmann fold	NA	Predicted dehydrogenase	NA
GC_00001739	0.01	8.64e+07	4.54e+05	5	6	ABC transporter	NA	ABC-type antimicrobial peptide transport system, permease component	SPO2182; permease, putative
GC_00001957	0.01	8.32e+07	4.59e+05	4	5	Hydroxyneurosporene synthase (CrtC)	NA	Predicted secreted hydrolase	SPO2183; hypothetical protein
GC_00002046	0.01	1.17e+08	6.12e+05	4	4	TspO and MBR like protein	translocator protein	Tryptophan-rich sensory protein (mitochondrial benzodiazepine receptor homolog)	NA

GC_00001851	0.01	1.44e+08	7.90e+05	5	5	NA	NA	NA	NA
GC_00002029	0.01	5.37e+07	2.53e+05	4	4	RbsD or FucU transport	L-fucose mutarotase [EC:5.1.3.29]	L-fucose mutarotase/ribose pyranase, RbsD/FucU family	NA
GC_00002285	0.01	1.37e+07	7.53e+04	2	2	Trap dicarboxylate transporter, dctm subunit	NA	TRAP-type mannitol/chloroaromatic compound transport system	SPO2605; TRAP dicarboxylate transporter, DctM subunit
GC_00002121	0.01	1.22e+08	5.99e+05	3	3	3-hydroxyacyl-CoA dehydrogenase	NA	Enoyl-CoA hydratase/carnithine racemase|3-hydroxyacyl-CoA dehydrogenase	SPO2920; fatty oxidation complex, alpha subunit
GC_00002024	0.01	4.66e+07	2.31e+05	4	4	Transcriptional regulator	NA	DNA-binding transcriptional regulator, LysR family	NA
GC_00002096	0.01	8.77e+07	4.03e+05	3	3	Trap-t family transporter	NA	NA	NA
GC_00001937	0.01	9.40e+07	4.46e+05	5	5	choline dehydrogenase	choline dehydrogenase [EC:1.1.99.1]	Choline dehydrogenase or related flavoprotein	SPO1088; choline dehydrogenase
GC_00002066	0.01	1.17e+08	5.77e+05	4	4	Enoyl-CoA hydratase	enoyl-CoA hydratase [EC:4.2.1.17]	Enoyl-CoA hydratase/carnithine racemase	SPO0147; enoyl-CoA hydratase
GC_00002056	0.01	3.01e+07	1.25e+05	3	4	Transcriptional regulator	GntR family transcriptional regulator / MocR family aminotransferase	DNA-binding transcriptional regulator, MocR family	SPOA0375; GntR family transcriptional regulator
GC_00002144	0.01	8.55e+07	3.63e+05	3	3	mucin-desulfating sulfatase	NA	Arylsulfatase A or related enzyme	NA
GC_00002012	0.01	5.03e+07	2.19e+05	4	4	Pantothenic acid kinase	NA	Uridine kinase	NA
GC_00002124	0.01	8.62e+07	3.79e+05	3	3	Transcriptional regulator, gntR family	NA	DNA-binding transcriptional regulator, GntR family	SPO0762; transcriptional regulator, GntR family
GC_00002009	0.01	7.74e+07	2.96e+05	4	4	alpha beta hydrolase fold-3 domain protein	epsilon-lactone hydrolase [EC:3.1.1.83]	Acetyl esterase/lipase	SPO3002; lipase, putative
GC_00002131	0.01	1.44e+08	6.72e+05	3	3	NA	NA	NA	NA
GC_00001411	0.01	1.12e+08	4.75e+05	6	7	LysR family transcriptional Regulator	NA	DNA-binding transcriptional regulator, LysR family	SPO0870; transcriptional regulator, LysR family
GC_00002050	0.01	1.05e+08	4.23e+05	4	4	Cys/Met metabolism PLP-dependent enzyme	NA	O-acetylhomoserine/O-acetylserine sulfhydrylase, pyridoxal phosphate-dependent	NA
GC_00002036	0.01	4.60e+07	1.90e+05	4	4	(ROK) family	NA	DNA-binding transcriptional regulator, MarR family	NA
GC_00001068	0.01	1.17e+08	5.17e+05	8	9	Dehydrogenase	D-amino-acid dehydrogenase [EC:1.4.5.1]	Glycine/D-amino acid oxidase (deaminating)	SPO0543; hypothetical protein
GC_00002018	0.01	1.27e+08	4.96e+05	4	4	Transcriptional regulator IclR family	NA	DNA-binding transcriptional regulator, IclR family	SPOA0143; IclR family transcriptional regulator
GC_00002167	0.01	8.91e+07	3.53e+05	3	3	Asp Glu racemase	maleate isomerase [EC:5.2.1.1]	Maleate cis-trans isomerase	SPOA0117; Asp/Glu/hydantoin racemase family protein
GC_00001914	0.01	8.45e+07	3.25e+05	5	5	LysR family (Transcriptional regulator	NA	DNA-binding transcriptional regulator, LysR family	SPO3240; transcriptional regulator, LysR family
GC_00001911	0.01	1.23e+08	4.22e+05	5	5	Ribosomal-protein-alanine acetyltransferase	[ribosomal protein S18]-alanine N-acetyltransferase [EC:2.3.1.266]	Ribosomal protein S18 acetylase RimI and related acetyltransferases	SPO0380; ribosomal-protein-alanine acetyltransferase, putative
GC_00002007	0.01	9.26e+07	3.76e+05	4	4	NA	NA	NA	NA
GC_00002266	0.00	1.09e+08	4.18e+05	2	2	Glutathione S-Transferase	glutathione S-transferase [EC:2.5.1.18]	Glutathione S-transferase	SPO1324; glutathione S-transferase family protein
GC_00002271	0.00	7.88e+06	1.90e+04	2	2	Reverse transcriptase (RNA-dependent DNA polymerase)	NA	Retron-type reverse transcriptase	NA
GC_00002033	0.00	1.24e+08	4.18e+05	4	4	Zinc metallopeptidase-like protein	uncharacterized protein	Predicted metal-dependent hydrolase	SPO3622; hypothetical protein
GC_00002162	0.00	1.41e+08	5.16e+05	3	3	NA	NA	Uncharacterized conserved protein	NA

GC_00002255	0.00	5.96e+07	1.87e+05	2	2	signal transduction histidine kinase	NA	Bacteriophytochrome (light-regulated signal transduction histidine kinase)	NA
GC_00002088	0.00	1.20e+08	2.87e+05	3	3	NA	NA	NA	NA
GC_00001311	0.00	1.37e+08	4.07e+05	8	8	Monofunctional biosynthetic peptidoglycan transglycosylase	monofunctional glycosyltransferase [EC:2.4.1.129]	Membrane carboxypeptidase (penicillin-binding protein)	SPO3766; monofunctional biosynthetic peptidoglycan transglycosylase
GC_00001892	0.00	9.26e+07	2.29e+05	4	5	Protein of unknown function (DUF1499)	NA	Uncharacterized conserved protein, DUF1499 family	SPO2463; hypothetical protein
GC_00001999	0.00	9.69e+07	2.54e+05	4	4	transcriptional Regulator, LysR family	NA	DNA-binding transcriptional regulator, LysR family	NA
GC_00002233	0.00	6.45e+07	1.70e+05	2	2	NA	NA	NA	NA
GC_00002045	0.00	1.29e+08	3.18e+05	4	4	NA	NA	NA	SPO1427; hypothetical protein
GC_00002026	0.00	1.21e+08	2.67e+05	4	4	phosphoglycerate mutase family protein	NA	Broad specificity phosphatase PhoE	NA
GC_00002059	0.00	3.69e+07	8.73e+04	4	4	pfkB family carbohydrate kinase	NA	Sugar or nucleoside kinase, ribokinase family	NA
GC_00002196	0.00	6.36e+07	1.24e+05	2	2	Two component transcriptional regulator, winged helix family	NA	DNA-binding response regulator, OmpR family	SPO0187; DNA-binding response regulator
GC_00002069	0.00	1.11e+08	2.27e+05	4	4	nuclease	NA	NA	NA
GC_00002001	0.00	8.60e+07	1.40e+05	4	4	FAD linked oxidase domain protein	NA	FAD/FMN-containing dehydrogenase	SPO2387; oxidoreductase, FAD-binding protein
GC_00002119	0.00	1.33e+08	1.62e+05	3	3	Methyltransferase	NA	O6-methylguanine-DNA--protein-cysteine methyltransferase	SPO3578; ADA regulatory protein
GC_00002295	0.00	2.45e+07	3.26e+04	2	2	NA	NA	NA	NA

Table 5.S2. Unique gene clusters in MB-C16 genomes. t/g = transcripts L^{-1} / genes L^{-1}; g/L = genes L^{-1}; t/L = transcripts L^{-1}.

id	*t/g*	*g/L*	*t/L*	*genomes*	*genes*	*eggNOG*	*KofamKoala*	*COG*	*R. pomeroyi*
GC_00002274	0.31	3.52e+07	8.34e+06	2	2	Multicopper oxidase	NA	Multicopper oxidase with three cupredoxin domains (includes cell division protein FtsP and spore coat protein CotA)	SPO1361; multicopper oxidase domain protein
GC_00002117	0.22	7.31e+07	1.68e+07	3	3	Periplasmic binding protein	D-xylose transport system substrate-binding protein	ABC-type xylose transport system, periplasmic component	SPO0861; xylose ABC transporter, periplasmic xylose-binding protein
GC_00002091	0.16	4.06e+07	5.19e+06	3	3	Glutaredoxin	NA	Glutaredoxin	NA
GC_00002236	0.08	7.19e+05	2.81e+04	2	2	Polysaccharide biosynthesis protein CapD	UDP-N-acetylglucosamine 4,6-dehydratase [EC:4.2.1.115]	NDP-sugar epimerase, includes UDP-GlcNAc-inverting 4,6-dehydratase FlaA1 and capsular polysaccharide biosynthesis protein EpsC	NA
GC_00002189	0.07	6.80e+07	4.01e+06	2	2	NA	NA	NA	SPO2778; hypothetical protein
GC_00002222	0.06	7.14e+07	3.94e+06	2	2	ABC transporter substrate binding protein	NA	ABC-type uncharacterized transport system YnjBCD, periplasmic component	NA
GC_00002022	0.05	6.67e+07	2.77e+06	4	4	NA	NA	NA	NA
GC_00002057	0.05	1.25e+08	4.55e+06	4	4	NA	NA	NA	NA
GC_00001866	0.04	1.21e+08	4.31e+06	5	5	NA	NA	NA	NA
GC_00002226	0.04	2.62e+06	1.03e+05	2	2	UDP-N-acetylglucosamine 2-epimerase	UDP-N-acetylglucosamine 2-epimerase (non-hydrolysing) [EC:5.1.3.14]	UDP-N-acetylglucosamine 2-epimerase	NA
GC_00002143	0.03	7.16e+07	2.22e+06	3	3	ABC transporter, (ATP-binding protein)	D-xylose transport system ATP-binding protein [EC:3.6.3.17]	ABC-type sugar transport system, ATPase component	SPOA0255; sugar ABC transporter, ATP binding protein
GC_00002097	0.03	6.98e+07	1.98e+06	3	3	ribitol 2-dehydrogenase	ribitol 2-dehydrogenase [EC:1.1.1.56]	NADP-dependent 3-hydroxy acid dehydrogenase YdfG	SPO3440; 20-beta-hydroxysteroid dehydrogenase, putative
GC_00002061	0.03	6.59e+07	1.43e+06	4	4	Signal-recognition-particle GTPase	NA	NA	NA
GC_00002011	0.03	1.34e+08	3.01e+06	4	4	NA	NA	NA	NA
GC_00002112	0.03	9.38e+07	2.09e+06	3	3	NA	NA	NA	NA
GC_00001664	0.03	6.05e+07	1.48e+06	6	6	solute-binding protein	putative spermidine/putrescine transport system substrate-binding protein	ABC-type Fe3+ transport system, periplasmic component	NA
GC_00002027	0.02	1.07e+08	2.02e+06	4	4	NA	NA	NA	NA
GC_00002221	0.02	1.55e+06	3.07e+04	2	2	Epimerase dehydratase	UDP-2-acetamido-2,6-beta-L-arabino-hexul-4-ose reductase [EC:1.1.1.367]	Nucleoside-diphosphate-sugar epimerase|dTDP-4-dehydrorhamnose 3,5-epimerase or related enzyme	NA
GC_00002137	0.02	2.31e+07	4.47e+05	2	3	NCS1 nucleoside transporter family	nucleobase:cation symporter-1, NCS1 family	Cytosine/uracil/thiamine/allantoin permease	NA
GC_00002030	0.02	4.88e+07	8.50e+05	3	4	extracellular solute-binding protein	sorbitol/mannitol transport system substrate-binding protein	ABC-type glycerol-3-phosphate transport system, periplasmic component	NA
GC_00001952	0.02	5.62e+07	1.03e+06	4	5	Inner-membrane translocator	D-xylose transport system permease protein	ABC-type xylose transport system, permease component	NA
GC_00001952	0.02	5.62e+07	1.03e+06	4	5	Inner-membrane translocator	NA	ABC-type xylose transport system, permease component	SPO0862; xylose ABC transporter, permease protein
GC_00001952	0.02	5.62e+07	1.03e+06	4	5	Inner-membrane translocator	NA	ABC-type xylose transport system, permease component	SPOA0254; ribose ABC transporter, permease protein
GC_00002003	0.02	1.30e+08	1.82e+06	4	4	LysE type translocator	homoserine/homoserine lactone efflux protein	Threonine/homoserine/homoserine lactone efflux protein	SPO0290; transmembrane amino acid efflux protein

GC_00002227	0.02	4.18e+07	5.03e+05	2	2	dipeptide ABC transporter, periplasmic dipeptide-binding protein	peptide/nickel transport system substrate-binding protein	ABC-type transport system, periplasmic component	SPO2835; dipeptide ABC transporter, periplasmic dipeptide-binding protein
GC_00002275	0.01	8.01e+07	1.13e+06	2	2	NA	NA	NA	NA
GC_00001864	0.01	1.09e+08	1.51e+06	5	5	Zinc manganese iron ABC transporter, periplasmic zinc manganese iron-binding protein	manganese/iron transport system substrate-binding protein	ABC-type Zn uptake system ZnuABC, Zn-binding component ZnuA	NA
GC_00001888	0.01	1.01e+08	1.10e+06	5	5	NA	NA	NA	NA
GC_00002195	0.01	7.25e+07	9.88e+05	2	2	ABC, transporter	NA	ABC-type spermidine/putrescine transport system, permease component II	NA
GC_00002259	0.01	7.44e+07	8.69e+05	2	2	ABC transporter	nonpolar-amino-acid-transporting ATPase [EC:7.4.2.2]	ABC-type Fe3+/spermidine/putrescine transport systems, ATPase components	NA
GC_00002203	0.01	4.02e+07	3.05e+05	2	2	Binding-protein-dependent transport systems inner membrane component	peptide/nickel transport system permease protein	ABC-type dipeptide/oligopeptide/nickel transport system, permease component	SPO2834; dipeptide ABC trasnporter, permease protein
GC_00002105	0.01	1.57e+08	1.46e+06	3	3	Enoyl-CoA hydratase	NA	Enoyl-CoA hydratase/carnithine racemase	SPO3646; enoyl-CoA hydratase/isomerase family protein
GC_00002114	0.01	5.19e+07	5.12e+05	3	3	Transcriptional regulator	NA	DNA-binding transcriptional regulator, LacI/PurR family	NA
GC_00002210	0.01	9.19e+07	8.86e+05	2	2	acetyL-CoA acetyltransferase	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Acetyl-CoA acetyltransferase	SPO0142; beta-ketothiolase
GC_00002219	0.01	1.38e+07	9.39e+04	2	2	NA	NA	NA	NA
GC_00002262	0.01	4.93e+07	4.16e+05	2	2	Fatty acid hydroxylase	NA	Sterol desaturase/sphingolipid hydroxylase, fatty acid hydroxylase superfamily	SPO0525; sterol desaturase-like protein
GC_00001702	0.01	1.40e+08	1.30e+06	6	6	shikimate dehydrogenase	shikimate dehydrogenase [EC:1.1.1.25]	Shikimate 5-dehydrogenase	SPO3891; shikimate 5-dehydrogenase
GC_00002086	0.01	4.48e+07	3.76e+05	3	3	transcriptional regulator, merr family	MerR family transcriptional regulator, copper efflux regulator	DNA-binding transcriptional regulator, MerR family	SPO0793; Cu(I)-responsive transcriptional regulator
GC_00001905	0.01	1.05e+08	1.00e+06	5	5	ABC transporter	manganese/iron transport system ATP-binding protein	ABC-type Mn2+/Zn2+ transport system, ATPase component	SPO3365; Manganese ABC transporter, ATP-binding protein
GC_00002142	0.01	1.31e+08	1.14e+06	3	3	Glutathione-dependent formaldehyde-activating GFA	NA	Uncharacterized conserved protein	SPO3401; hypothetical protein
GC_00001854	0.01	1.47e+08	1.25e+06	5	5	16S rRNA (guanine527-N7)-methyltransferase	16S rRNA (guanine527-N7)-methyltransferase [EC:2.1.1.170]	16S rRNA G527 N7-methylase RsmG (former glucose-inhibited division protein B)	SPO0002; glucose-inhibited division protein B
GC_00002211	0.01	1.02e+08	8.62e+05	2	2	3-hydroxyacyl-COA dehydrogenase	NA	Enoyl-CoA hydratase/carnithine racemase|3-hydroxyacyl-CoA dehydrogenase	SPO0772; enoyl-CoA hydratase/isomerase/3-hydroxyacyl-CoA dehydrogenase
GC_00002218	0.01	6.08e+07	4.06e+05	2	2	Sel1 domain protein repeat-containing protein	uncharacterized protein	TPR repeat	NA
GC_00002171	0.01	1.19e+08	9.72e+05	3	3	NA	NA	NA	NA
GC_00002151	0.01	3.53e+07	2.74e+05	3	3	NA	NA	NA	NA
GC_00002239	0.01	2.27e+07	1.31e+05	2	2	Carbohydrate kinase	NA	Sugar (pentulose or hexulose) kinase	NA
GC_00002020	0.01	1.13e+08	7.69e+05	4	4	Protein of unknown function (DUF3445)	NA	NA	SPO2298; hypothetical protein
GC_00002254	0.01	5.90e+07	4.13e+05	2	2	auxin efflux carrier	uncharacterized protein	Predicted permease	NA
GC_00002176	0.01	1.17e+08	7.85e+05	3	3	NA	NA	NA	NA

GC_00001920	0.01	1.10e+08	8.07e+05	5	5	ABC transporter	manganese/iron transport system permease protein	ABC-type Mn2+/Zn2+ transport system, permease component	NA
GC_00001881	0.01	5.38e+07	3.70e+05	5	5	Transcriptional regulator	NA	DNA-binding transcriptional regulator, LacI/PurR family	SPO0590; transcriptional regulator, LacI family
GC_00002163	0.01	1.50e+08	9.63e+05	3	3	phage Tail Protein	NA	Uncharacterized conserved protein	NA
GC_00002197	0.01	7.02e+07	5.17e+05	2	2	ABC transporter, membrane spanning protein	NA	ABC-type spermidine/putrescine transport system, permease component I	NA
GC_00001912	0.01	1.23e+08	7.17e+05	5	5	Ribosomal-protein-alanine acetyltransferase	[ribosomal protein S18]-alanine N-acetyltransferase [EC:2.3.1.266]	Ribosomal protein S18 acetylase RimI and related acetyltransferases	SPO0380; ribosomal-protein-alanine acetyltransferase, putative
GC_00001093	0.01	1.13e+08	7.39e+05	8	8	NA	NA	NA	NA
GC_00002139	0.01	1.08e+08	7.29e+05	2	3	multidrug resistance protein	small multidrug resistance pump	Multidrug transporter EmrE and related cation transporters	SPO2030; multidrug resistance efflux protein, SMR family
GC_00002284	0.01	3.60e+07	2.69e+05	2	2	NA	NA	NA	NA
GC_00002200	0.01	1.20e+08	6.87e+05	2	2	NA	NA	NA	NA
GC_00001935	0.01	1.12e+08	6.86e+05	5	5	membrane	NA	Permease of the drug/metabolite transporter (DMT) superfamily	SPO1337; hypothetical protein
GC_00002237	0.01	5.99e+07	4.38e+05	2	2	ROK family	NA	Sugar kinase of the NBD/HSP70 family, may contain an N-terminal HTH domain	NA
GC_00002283	0.01	3.10e+07	1.58e+05	2	2	EamA-like transporter family	NA	Permease of the drug/metabolite transporter (DMT) superfamily	NA
GC_00002133	0.01	2.03e+07	1.18e+05	3	3	NA	NA	NA	NA
GC_00002150	0.01	6.06e+07	4.03e+05	3	3	trap transporter, 4tm 12tm fusion protein	NA	TRAP-type uncharacterized transport system, fused permease components	SPO2186; TRAP transporter, 4TM/12TM fusion protein
GC_00002122	0.01	1.34e+08	6.50e+05	2	3	Pantothenic acid kinase	NA	Uridine kinase	NA
GC_00001913	0.01	1.09e+08	5.72e+05	5	5	ABC transporter	manganese/iron transport system permease protein	ABC-type Mn2+/Zn2+ transport system, permease component	SPO3363; Manganese ABC transporter, permease protein
GC_00002261	0.01	7.26e+07	3.26e+05	2	2	transcriptional regulator	NA	Sugar kinase of the NBD/HSP70 family, may contain an N-terminal HTH domain	NA
GC_00002109	0.01	1.55e+08	8.76e+05	3	3	NA	NA	NA	NA
GC_00002075	0.01	1.15e+08	4.99e+05	4	4	phosphoglycerate mutase	NA	Broad specificity phosphatase PhoE	SPO0523; phosphoglycerate mutase family protein
GC_00002235	0.01	2.69e+07	9.55e+04	2	2	Short chain dehydrogenase	NA	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	SPO1437; 2,5-dichloro-2,5-cyclohexadiene-1,4-diol dehydrogenase
GC_00002132	0.01	1.47e+08	6.24e+05	3	3	NA	NA	NA	NA
GC_00002212	0.01	4.69e+07	1.83e+05	2	2	Dehydrogenase	NA	Threonine dehydrogenase or related Zn-dependent dehydrogenase	SPOA0272; glutathione-dependent formaldehyde dehydrogenase
GC_00002228	0.00	2.26e+07	6.45e+04	2	2	Carbohydrate kinase	NA	Sugar (pentulose or hexulose) kinase	NA
GC_00001721	0.00	5.39e+07	1.96e+05	5	6	Binding-protein-dependent transport systems inner membrane component	putative spermidine/putrescine transport system permease protein	ABC-type spermidine/putrescine transport system, permease component II	SPO2008; spermidine/putrescine ABC transporter, permease protein
GC_00001661	0.00	5.76e+07	2.60e+05	6	6	binding-protein-dependent transport systems inner membrane component	NA	ABC-type sulfate transport system, permease component	SPO0698; molybdate ABC transporter, permease protein
GC_00002188	0.00	1.21e+08	3.82e+05	2	2	NA	NA	NA	NA
GC_00002174	0.00	5.20e+07	1.69e+05	3	3	NA	NA	Tagatose-1,6-bisphosphate aldolase non-catalytic subunit AgaZ/GatZ	NA

GC_00002102	0.00	5.27e+07	1.58e+05	3	3	binding-protein-dependent transport systems inner membrane component	sorbitol/mannitol transport system permease protein	ABC-type sugar transport system, permease component	NA
GC_00002099	0.00	5.50e+07	1.69e+05	3	3	Binding-protein-dependent transport systems inner membrane component	sorbitol/mannitol transport system permease protein	ABC-type glycerol-3-phosphate transport system, permease component	NA
GC_00002204	0.00	1.43e+08	4.53e+05	2	2	NA	NA	NA	NA
GC_00001859	0.00	1.51e+08	4.37e+05	5	5	NA	NA	NA	NA
GC_00002089	0.00	8.14e+07	2.30e+05	3	3	NA	NA	NA	NA
GC_00002205	0.00	1.26e+08	3.91e+05	2	2	NA	NA	NA	NA
GC_00001945	0.00	8.62e+07	2.28e+05	3	5	Dehydrogenase	galactitol 2-dehydrogenase [EC:1.1.1.16]	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	SPO0128; oxidoreductase, short chain dehydrogenase/reductase family
GC_00002279	0.00	1.03e+08	2.87e+05	2	2	NA	NA	NA	NA
GC_00001848	0.00	1.08e+08	3.28e+05	5	5	NA	NA	NA	NA
GC_00002155	0.00	4.99e+07	1.46e+05	2	3	Mannitol dehydrogenase	mannitol 2-dehydrogenase [EC:1.1.1.67]	Mannitol-1-phosphate/altronate dehydrogenases	SPO1724; D-mannonate oxidoreductase
GC_00002165	0.00	5.02e+07	1.07e+05	3	3	Carbohydrate kinase	NA	Sugar or nucleoside kinase, ribokinase family	NA
GC_00002281	0.00	2.96e+07	8.44e+04	2	2	NA	NA	Superfamily II DNA or RNA helicase, SNF2 family|Ubiquinone/menaquinone biosynthesis C-methylase UbiE	NA
GC_00002118	0.00	1.29e+08	2.57e+05	3	3	Methyltransferase	NA	O6-methylguanine-DNA--protein-cysteine methyltransferase	SPO3578; ADA regulatory protein
GC_00002194	0.00	3.83e+07	7.20e+04	2	2	Short-chain dehydrogenase reductase SDR	3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]	NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family	SPO2417; gluconate 5-dehydrogenase
GC_00002252	0.00	3.99e+07	5.84e+04	2	2	PfkB domain protein	NA	Sugar or nucleoside kinase, ribokinase family	NA
GC_00001874	0.00	6.29e+07	7.01e+04	5	5	NA	NA	NA	NA
GC_00002257	0.00	4.08e+07	5.94e+04	2	2	NA	NA	NA	NA

Table 5.S3. Core gene clusters with >2-fold gene expression ratios between species (HTCC2255 transcripts per gene copy / MB-C16 transcripts per gene copy). t/g = transcripts L^{-1} / genes L^{-1}.

id	*htcc2255/C16 t/g*	*upregulated clade*	*eggNOG*	*KofamKoala*	*COG*	*R. pomeroyi*
GC_00001797	4.03	HTCC2255	Aminotransferase	NA	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	SPO1401; aminotransferase, class III
GC_00001843	3.52	HTCC2255	LysR family transcriptional Regulator	NA	DNA-binding transcriptional regulator, LysR family	SPO0870; transcriptional regulator, LysR family
GC_00002206	3.21	HTCC2255	Cytochrome B561	cytochrome b561	Polyisoprenoid-binding periplasmic protein YceI|Cytochrome b561	NA
GC_00002104	3.08	HTCC2255	transporter, RhaT family, DMT superfamily	S-adenosylmethionine uptake transporter	Permease of the drug/metabolite transporter (DMT) superfamily	SPO0261; hypothetical protein
GC_00001635	2.79	HTCC2255	Oxidoreductase	NA	Glycine/D-amino acid oxidase (deaminating)	SPOA0380; hypothetical protein
GC_00002130	2.64	HTCC2255	LamB YcsF family protein	5-oxoprolinase (ATP-hydrolysing) subunit A [EC:3.5.2.9]	Lactam utilization protein B (function unknown)	SPO3659; LamB/YcsF family protein
GC_00002141	2.52	HTCC2255	chaperone	fimbrial chaperone protein	P pilus assembly protein, chaperone PapD	NA
GC_00000600	2.40	HTCC2255	response regulator	NA	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain	SPO1023; response regulator
GC_00002277	2.32	HTCC2255	glycosyl transferase, family 25	NA	Glycosyltransferase involved in LPS biosynthesis, GR25 family	SPO3385; glycosyl transferase, family 25
GC_00000821	2.22	HTCC2255	Lipoprotein	NA	Uncharacterized protein	SPO3414; lipoprotein, putative
GC_00001582	2.17	HTCC2255	ribosome-binding factor A	ribosome-binding factor A	Ribosome-binding factor A	SPO3835; ribosome-binding factor A
GC_00001518	2.10	HTCC2255	Acetyltransferase GNAT family	NA	Predicted N-acyltransferase, GNAT family	NA
GC_00000814	2.08	HTCC2255	Cupin 2 Conserved Barrel Domain Protein	NA	Cupin domain protein related to quercetin dioxygenase	SPOA0273; DNA-binding protein
GC_00002107	2.07	HTCC2255	Choline dehydrogenase or related	NA	Choline dehydrogenase or related flavoprotein	SPO2359; Isethionate dehydrogenase
GC_00001461	2.06	HTCC2255	sarcosine oxidase alpha subunit	NA	Predicted molibdopterin-dependent oxidoreductase YjgC	NA
GC_00002263	2.05	HTCC2255	alcohol dehydrogenase	NA	NADPH:quinone reductase or related Zn-dependent oxidoreductase	SPO1593; alcohol dehydrogenase, zinc-containing
GC_00001120	2.02	HTCC2255	Phosphate transporter	inorganic phosphate transporter, PiT family	Phosphate/sulfate permease|Phosphate/sulfate permease	SPO0967; phosphate transporter family protein
GC_00000705	2.02	HTCC2255	2og-fe(ii) oxygenase	NA	Isopenicillin N synthase and related dioxygenases	SPO2669; oxidoreductase, 2OG-Fe(II) oxygenase family
GC_00000895	2.00	HTCC2255	thioesterase	NA	Acyl-coenzyme A thioesterase PaaI, contains HGG motif	SPO1688; thioesterase family protein
GC_00001331	0.05	MB-C16	CopC domain	NA	Copper-binding protein CopC (methionine-rich)	NA
GC_00001317	0.06	MB-C16	Copper resistance D	copper resistance protein D	Putative copper export protein	NA
GC_00000902	0.15	MB-C16	cytochrome C family protein	NA	Cytochrome c, mono- and diheme variants	SPOA0359; cytochrome c family protein

GC_00001625	0.18	MB-C16	decarboxylase	NA	Glutamate or tyrosine decarboxylase or a related PLP-dependent protein	NA
GC_00002052	0.19	MB-C16	Spore Coat Protein U domain	NA	Spore coat protein U (SCPU) domain, function unknown	NA
GC_00002223	0.25	MB-C16	Spore Coat Protein	NA	Spore coat protein U (SCPU) domain, function unknown	NA
GC_00001233	0.32	MB-C16	agmatinase	agmatinase [EC:3.5.3.11]	Arginase family enzyme	SPOA0234; agmatinase
GC_00001072	0.33	MB-C16	Nudix hydrolase	peroxisomal coenzyme A diphosphatase NUDT7 [EC:3.6.1.-]	8-oxo-dGTP pyrophosphatase MutT and related house-cleaning NTP pyrophosphohydrolases, NUDIX family	SPO0025; hydrolase, NUDIX family
GC_00001837	0.33	MB-C16	Cytochrome	cytochrome c	Cytochrome c553|Cytochrome c2	SPO1000; diheme cytochrome c SoxE
GC_00001790	0.34	MB-C16	Ferredoxin	NA	Ferredoxin-NADP reductase	SPO2377; ferredoxin
GC_00002178	0.38	MB-C16	TRAP transporter solute receptor TAXI family	uncharacterized protein	TRAP-type uncharacterized transport system, periplasmic component	SPO2187; TRAP transporter solute receptor, TAXI family
GC_00002180	0.38	MB-C16	Aldehyde dehydrogenase	NA	Acyl-CoA reductase or other NAD-dependent aldehyde dehydrogenase	NA
GC_00001628	0.41	MB-C16	polyA polymerase	NA	tRNA nucleotidyltransferase/poly(A) polymerase	SPO0026; polyA polymerase family protein
GC_00001926	0.42	MB-C16	transcriptional Regulator, LysR family	LysR family transcriptional regulator, hypochlorite-specific transcription factor HypT	DNA-binding transcriptional regulator, LysR family	SPO2656; transcriptional regulator, LysR family
GC_00000824	0.42	MB-C16	reductase	peptide-methionine (R)-S-oxide reductase [EC:1.8.4.12]	Peptide methionine sulfoxide reductase MsrB	SPO3741; methionine-R-sulfoxide reductase
GC_00001895	0.42	MB-C16	CoA-binding domain protein	uncharacterized protein	Predicted CoA-binding protein	SPO2376; CoA-binding domain protein
GC_00000115	0.43	MB-C16	Maf-like protein	septum formation protein	Predicted house-cleaning NTP pyrophosphatase, Maf/HAM1 superfamily	SPO3892; Maf
GC_00001275	0.44	MB-C16	Sulfite exporter TauE/SafE	uncharacterized protein	Uncharacterized membrane protein YfcA	SPO2319; hypothetical protein
GC_00000585	0.44	MB-C16	SmpA OmlA	NA	Outer membrane protein assembly factor BamE, lipoprotein component of the BamABCDE complex	SPO2490; lipoprotein, SmpA/OmlA family
GC_00000837	0.44	MB-C16	Regulatory protein SoxS	NA	Thiol-disulfide isomerase or thioredoxin	SPO0990; regulatory protein SoxS
GC_00000868	0.45	MB-C16	transcriptional Regulator, LysR family	NA	DNA-binding transcriptional regulator, LysR family	SPO0241; transcriptional regulator, LysR family
GC_00000610	0.45	MB-C16	ABC transporter permease protein	branched-chain amino acid transport system permease protein	Branched-chain amino acid ABC-type transport system, permease component	SPO0824; branched-chain amino acid ABC transporter, permease protein
GC_00002185	0.46	MB-C16	Dihydrolipoyl dehydrogenase	NA	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoamide dehydrogenase (E3) component or related enzyme	NA

GC_00001621	0.46	MB-C16	transcriptional regulatory protein (LysR family)	NA	DNA-binding transcriptional regulator, LysR family	NA
GC_00002149	0.47	MB-C16	deiminase	NA	Agmatine/peptidylarginine deiminase	SPO2980; porphyromonas-type peptidyl-arginine deiminase family protein
GC_00000844	0.47	MB-C16	ATP-dependent protease	ATP-dependent HslUV protease, peptidase subunit HslV [EC:3.4.25.2]	ATP-dependent protease HslVU (ClpYQ), peptidase subunit	SPO3880; ATP-dependent protease hslV
GC_00002276	0.47	MB-C16	DegT DnrJ EryC1 StrS aminotransferase	CDP-4-dehydro-6-deoxyglucose reductase, E1 [EC:1.17.1.1]	dTDP-4-amino-4,6-dideoxygalactose transaminase	NA
GC_00001341	0.47	MB-C16	Selenium-binding protein	selenium-binding protein 1	DNA-binding beta-propeller fold protein YncE	SPO2378; selenium-binding protein, putative
GC_00001670	0.48	MB-C16	Inherit from proNOG: Xylose isomerase domain protein TIM barrel	NA	Sugar phosphate isomerase/epimerase	NA
GC_00001576	0.48	MB-C16	Polyprenyl synthetase	NA	Geranylgeranyl pyrophosphate synthase	SPO0319; decaprenyl diphosphate synthase
GC_00001813	0.49	MB-C16	Inherit from bactNOG: isoprenylcysteine carboxyl methyltransferase	NA	Protein-S-isoprenylcysteine O-methyltransferase Ste14	SPO2821; isoprenylcysteine carboxyl methyltransferase family protein
GC_00001567	0.49	MB-C16	DegT DnrJ EryC1 StrS	NA	dTDP-4-amino-4,6-dideoxygalactose transaminase	SPO2795; aminotransferase, DegT/DnrJ/EryC1/StrS family
GC_00000961	0.49	MB-C16	Ribokinase	ribokinase [EC:2.7.1.15]	Sugar or nucleoside kinase, ribokinase family	SPO0013; ribokinase
GC_00001832	0.49	MB-C16	Deoxyribodipyrimidine photo-lyase	NA	Deoxyribodipyrimidine photolyase	NA
GC_00001148	0.49	MB-C16	stress responsive alpha-beta barrel domain-containing protein	NA	NA	NA
GC_00001063	0.50	MB-C16	NA	heme exporter protein A [EC:7.6.2.5]	ABC-type transport system involved in cytochrome c biogenesis, ATPase component	SPO2317; heme exporter protein CcmA

CHAPTER 6

SUMMARY

Gene- and genome-centric approaches, including metagenomics, metatranscriptomics, single-cell genomics, small subunit rRNA gene sequencing, and model organism transcriptomics were used to characterize dimensions that determine the ecological niches of heterotrophic bacteria in the coastal ocean. Chapter 2 used a metagenomic dataset in Fall 2014 Monterey Bay surface seawater to characterize the type and abundance of bacterial genes that transform dimethylsulfoniopropionate (DMSP). This study found that the demethylation gene *dmdA* dominated the DMSP gene pool, with the alphaproteobacterial members SAR11 and roseobacters harboring the most genes. In the cleavage pathway, the recently discovered DMSP lyase *dddK* was the most abundant in September, but dominance shifted to the *dddP* homolog, found mostly in roseobacters, concurrent with shifts in bacterial taxonomy. Assembly of metagenomic reads and single-cell genomes generated during the study provided an untargeted window into the diversity of DMSP genes in this ecosystem, including a novel gammaproteobacterial gene and those from SAR11 and streamlined roseobacters. The phytoplankton community, and thus the sources and supply of DMSP, shifted over the 3-week study period from centric diatoms and dinoflagellates to pennate diatoms and coscinodiscophyceae. SAR11 DMSP genes were strongly correlated with the DMSP-producing dinoflagellates.

A second expedition to Monterey Bay during Fall 2016 served as the basis for genome-centric analyses of bloom-associated bacteria in Chapters 3-5. Sampling coincided with a

massive bloom of the high-DMSP producing dinoflagellate *Akashiwo sanguinea*. Chapter 3 details an inventory of metagenomes, metatranscriptomes, and 16S and 18S rRNA gene libraries generated from microbial biomass of this bloom. During a 52-day period of the bloom, 88 16S rRNA gene amplicon libraries, 88 18S rRNA gene amplicon libraries, 84 metagenomes, and 82 metatranscriptomes were generated. Additionally, chemical and biological measurements were taken to accompany the samples, revealing high phytoplankton biomass (up to 57 $\mu g\ L^{-1}$) and record-high DMSP concentrations. *A. sanguinea* dominated the microbial community throughout, with diatoms and other dinoflagellates peaking on some dates. The 18S rRNA gene libraries showed increases in grazers and a parasite of dinoflagellates as the bloom end neared. The bacterial community was dominated by Proteobacteria, including roseobacters and SAR11, with a significant presence of Gammaproteobacteria, and a peak of flavobacters at the end of the bloom.

Chapter 4 and 5 describe two analyses of niche dimensions and niche differentiation using seawater sampled during the Fall 2016 Monterey Bay bloom. In Chapter 4, *Ruegeria pomeroyi*, a model coastal heterotrophic bacterium, was introduced to whole seawater on 14 dates over the bloom. After a 90 minute exposure to the natural microbial community, the bacterium's transcriptomes were obtained to assess the response to abiotic and biotic factors affecting its viability in this dynamic coastal system. Key niche dimensions were deduced from the functional annotation of genes with significant relative expression changes through time, and included those related to substrates, vitamins, nutrients, and metals, as well as biotic interaction dimensions, including antagonism and resistance. Genes enriched during peak bloom seawater included many substrate-related dimensions, however signals of low apparent growth rate indicated that the bacterium's realized niche was narrowed during this peak of the bloom,

potentially due to toxins and competition with the natural microbial community. These invasion experiments highlighted many factors important to niche space and survival for marine bacteria.

Chapter 5 shifted focus from a well-characterized model organism's response to the bloom seawater to an uncultured and poorly resolved bacterial taxon present during the bloom. 16S rRNA sequencing of the bloom bacterial community identified a roseobacter amplicon sequence variant dominating most samples that had 100% sequence identity to a streamlined roseobacter isolate. Two metagenome-assembled genomes and 28 single amplified genomes recovered from the bloom communities showed this taxon to consist of two sequence-discrete clusters of genomes with sufficient genetic distance to be considered separate species. The 30 genomes recovered from Monterey Bay plus the original isolate genome were compared in a metapangenomic analysis to track the abundance of the co-occurring species over time and assess gene content and the expression of shared and unique genes in this lineage. A total of 2,296 gene clusters were identified, with the criterion that they were present in at least two genomes. Of these, 215 were unique to a species. The key genes representing niche differentiation included those for substrate acquisition (sugars, carboxylic acids, polyamines), vitamin synthesis, and energy generation via a second proteorhodopsin copy.

In summary, this dissertation produced insights into important abiotic and biotic factors influencing the success of marine heterotrophic bacteria in dynamic coastal seawater. Understanding was gained of genes important in utilizing a major substrate niche dimension, DMSP. Gene expression from a model organism introduced to seawater allowed assessment of the multitude of dimensions affecting a bacterium's viability. Finally, niche differentiation was observed between two highly related and abundant bloom-associated bacterial species.