Characterization of Microbial Primary and Secondary Metabolism in the Marine Realm

By

David Edward Geller-McGrath
B.A., Clark University, 2015

Submitted to the Department of Earth, Atmospheric, and Planetary Sciences in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

and the

WOODS HOLE OCEANOGRAPHIC INSTITUTION

September 2024

Author

Joint Program in Biological Oceanography
Massachusetts Institute of Technology
and Woods Hole Oceanographic Institution
July 19, 2024

Certified by

Dr. Virginia P. Edgcomb
Senior Scientist
Woods Hole Oceanographic Institution
Thesis Supervisor

Accepted by

Dr. Jesús Pineda
Chair, Joint Committee for Biological Oceanography
Massachusetts Institute of Technology/
Woods Hole Oceanographic Institution

Characterization of Microbial Primary and Secondary Metabolism in the Marine Realm

By

David Edward Geller-McGrath

Submitted to the Department of Earth, Atmospheric, and Planetary Sciences
on July 19, 2024, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

This thesis applies meta-omics data analysis to elucidate the ecological roles of marine microorganisms in diverse habitats and includes the development of new bioinformatics tools to enhance these analyses. In my second chapter, I applied genome mining tools to analyze the gene content and expression of biosynthetic gene clusters (BGCs). The analysis of BGCs through large-scale genome mining efforts has identified diverse natural products with potential applications in medicine and biotechnology. Many marine environments, particularly oxygen-depleted water columns and sediments, however, remain under-represented in these studies. Analysis of BGCs in free-living and particle-associated microbial communities along the oxycline water column of the Cariaco Basin, Venezuela, revealed that differences in water column redox potential were associated with microbial lifestyle and the predicted composition and production of secondary metabolites. This experience set the stage for my third chapter, in which I developed MetaPathPredict, a machine learning-based tool for predicting the metabolic potential of bacterial genomes. This tool addresses the lack of computational pipelines for pathway reconstruction that predict the presence of KEGG modules in highly incomplete prokaryotic genomes. MetaPathPredict made robust predictions in highly incomplete bacterial genomes, enabling more accurate reconstruction of their metabolic potential. In my fourth chapter, I performed metagenomic analysis of microbial communities in the hydrothermally-influenced sediments of Guaymas Basin (Gulf of California, Mexico). Previous studies indicated a decline in microbial abundance and diversity with increasing sediment depth. Analysis revealed a distribution of MAGs dominated by Chloroflexota and Thermoproteota, with diversity decreasing as temperature increased, consistent with a downcore reduction in subsurface biosphere diversity. Specific archaeal MAGs within the Thermoproteota and Hadarchaeota increased in abundance and recruitment of metatranscriptome reads towards deeper, hotter sediments, marking a transition to a specialized deep biosphere. In my fifth chapter, I developed MetaPathPredict-E, a deep learning-powered extension of MetaPathPredict for eukaryotic metabolism predictions. Eukaryotic metabolism is diverse, reflecting varied lifestyles across eukaryotic kingdoms, but the complexity of eukaryotic genomes presents challenges for assembly and annotation. MetaPathPredict-E was trained on diverse eukaryotic genomes and transcriptomes, demonstrating a robust performance on test datasets, thus advancing the study of eukaryotic metabolic potential from environmental samples.

Thesis Supervisor: Virginia Edgcomb
Title: Senior Scientist
Woods Hole Oceanographic Institution

Acknowledgements

I am deeply grateful to the many individuals and institutions that made it possible for me to explore my research interests in this dissertation. I would like to extend my sincere thanks to the Massachusetts Institute of Technology and the Woods Hole Oceanographic Institution (WHOI) for providing a supportive learning environment and funding for conference presentations. My research was supported by the 2018 MIT Presidential Fellowship, the 2020 Department of Energy SCGSR Fellowship Solicitation #2 in Computational Biology and Bioinformatics, the National Science Foundation grants OCE-19924492, OCE-2046799, and OCE-1829903, and the WHOI Academic Programs Office Fellowship. I also wish to express my gratitude to all the collaborators who contributed their expertise to examine and interpret the extensive data in this thesis.

I am grateful to my advisor Virginia Edgcomb for her unwavering support and encouragement during my time in her research group. Thank you for welcoming me into your lab and fostering my growth as I explored my research interests. You gave me the freedom to go in my own direction while holding me to a high standard that shaped me into a better scientist and writer. I greatly appreciate how you always made time for me, even on your busiest days. Your attention to detail and thoughtful advice helped me navigate the complexities of my research and the challenges of the PhD process.

Thank you to Kishori Konwar, whose fundamental knowledge of machine learning methods and bioinformatics tools inspired this dissertation. Thank you for all your thoughtful advice and support, and for always taking the time to help me when I encountered challenges and needed guidance. I am thankful to Harriet Alexander for her guidance in the analysis of eukaryotic datasets and machine learning model development. Thank you, Greg Fournier, for your enthusiasm for my research, your expertise in computational biology, and for your thoughtful ideas and feedback. Thank you to Maria Pachiadaki for your mentorship and for guiding me through the complexities of computational biology research. Thank you, Travis Wheeler, for sharing your expertise in computer science and for your invaluable support in my pursuit of developing new bioinformatics methods. Thank you to Jason McDermott, whose expertise in computational biology supported my pursuit to develop new bioinformatics tools. I appreciate the special opportunity I had to work with you at Pacific Northwest National Laboratory and for all the computational biology techniques and insights you shared with me, which were essential for completing this dissertation. Thank you, Vivian Mara, for your immense support while navigating my thesis projects, and for always being there to offer encouragement and assistance whenever I needed it. Thank you to Andreas Teske for your guidance and invaluable insights. Your profound understanding of deep subsurface habitats was inspirational. Thank you to Andy Solow, for all of your support and for lending me your statistical expertise. You were always more than happy to meet whenever I needed to ask you anything, and always knew how to make me laugh, even in the most stressful moments. Thank you to Ann Tarrant and Neel Aluru for your mentorship and guidance. Thank you, David Beaudoin, for all of your guidance.

Thank you to my mom, Julia, and Erin, who have been steadfast in their support and have always brightened my days with encouragement and positivity. You have always been there to put a smile on my face and have always believed in me, inspiring me to persevere through the toughest times. I am deeply grateful for all the support from my extended family and friends as well. Thank you for all the love, optimism, and camaraderie that got me here.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

General overview of microbial primary metabolism

Metabolism is the essential cellular process of capturing energy through the oxidation of organic matter, inorganic matter, or through photosynthesis coupled with the biosynthesis of small molecules and polymerization of those into macromolecules (DeBerardinis and Thompson 2012; Hohmann-Marriott and Blankenship 2011). Unveiling the primary metabolic potential encoded in the genomes of microorganisms from diverse environments is one of the fundamental foci of research in the field of environmental microbiology. Core metabolic pathways encompass all the chemical reactions that occur within a cell and can be categorized into three groups: catabolism, anabolism, and waste removal (DeBerardinis and Thompson 2012). Catabolic metabolism consists of the metabolic pathways that cells use to enzymatically oxidize nutrients from their environment. This process releases energy that can be harnessed by the cell and stored primarily in the form of high-energy phosphorus bonds in the adenosine triphosphate (ATP) molecule. ATP acts as an energetic "currency" for all cells; its phosphorus bonds are hydrolyzed to release the energy required to synthesize small molecules such as amino acids, sugars, and nucleic acids, and to polymerize macromolecules including DNA, RNA, and proteins (DeBerardinis and Thompson 2012). Anabolic metabolism includes biosynthetic pathways that expend stored energy to synthesize new molecules to facilitate cellular growth, repair, maintenance, and reproduction.

Some similarities are observed between the metabolic pathways encoded in the genomes of prokaryotes (single-celled organisms from the domains Bacteria and Archaea) and eukaryotes (single-celled and multicellular organisms with a nucleus and other eukaryote-specific, membrane-bound organelles). These include several highly conserved metabolic pathways that are a testament to their shared evolutionary history and fundamental importance. Catabolic pathways that are present in most organisms include glycolysis (the breakdown of the sugar glucose that yields acetyl-CoA and ATP), the tricarboxylic acid cycle (TCA cycle; consumes acetyl-CoA and produces more ATP), and oxidative phosphorylation (induces chemiosmosis via an electron

transport chain (ETC) that generates ATP). It is notable that within these essential pathways there is some variation in the enzymes involved (Kwong et al., 2017), leading to pathway variants that retain the same function (Evans et al., 2024). Anabolic pathways that are largely conserved include the biosynthesis of nucleotides, as well as some amino acids, sugars, and transporters.

Metabolisms present within both prokaryotes and eukaryotes include oxygenic photosynthesis, organoheterotrophy, and diverse fermentations. Oxygenic photosynthesis is the process in which energy from sunlight is harnessed to oxidize water, generate ATP, and fix inorganic carbon. It originated in an ancestor of Cyanobacteria (Sánchez and Cardona 2020) and has been acquired in Eukaryotes through primary, secondary, and tertiary endosymbiosis (Hohmann-Marriott and Blankenship 2011). Bacterial phyla that can perform anoxygenic photosynthesis (no oxygen released) use terminal electron donors that include elemental sulfur, sulfide, thiosulfate, hydrogen and ferrous iron (Ehrenreich and Widdel 1994; Frigaard and Dahl 2008). Organoheterotrophic prokaryotes and eukaryotes acquire energy and carbon from organic sources, while fermenting organisms gain energy through the anaerobic oxidation of sugars. Mixotrophy is a trophic mode observed in some protists and bacteria that are capable of both photosynthesis and organoheterotrophy (Porter 1988; Eiler 2006). The genomes of some mixotrophs contain genes encoding photosynthetic machinery, while some protists steal chloroplasts from photoautotrophic or mixotrophic Eukaryotes through phagocytosis (Lewitus 1999; Johnson 2011).

All life forms require substrates for catabolism and to fuel cellular activities. Prokaryotes, however, exhibit a wider range of energy- and carbon-obtaining strategies than eukaryotes despite their simpler cellular architecture (Heider et al., 1998; Schäfer et al., 1999; Offre et al., 2013; Gupta and Gupta 2021). In the absence of oxygen, prokaryotes (and some eukaryotes) utilize alternative terminal electron acceptors in their electron transport chains (ETCs; Fewson and Nicholas 1961; Strohm et al., 2007; Kraft et al., 2011; Kamp et al., 2015). Electron acceptors for anaerobic microbes include nitrate (for denitrification and dissimilatory nitrate reduction to ammonium), sulfate (dissimilatory sulfate reduction), carbon dioxide (acetogenesis), and ferric iron (dissimilatory iron reduction; Sørensen 1982; Lever et al., 2012; Averill and Tiedje 1982; Mohan and Cole 2007). Chemolithotrophs include autotrophic and heterotrophic microbes that oxidize inorganic compounds for energy, including hydrogen, carbon monoxide, sulfur, methane, alkanes, ferrous iron, ammonia (for nitrification and anammox), and manganese (Frigaard and Dahl 2008;

Kim and Hegeman 1983; Chen et al., 2009; Wang et al., 2019; Ehrenreich and Freidrich 1994; Yu and Leadbetter 2020). Anaerobes can also oxidize additional inorganic molecules, including ions of selenium, arsenic, chromium, and uranium, as well as fumarate, trimethylamine *N*-oxide, and dimethyl sulfoxide (Guest 1979; Fredrickson et al., 2000; Stolz et al., 2006; Yu and Leadbetter 2020; Oren and Trüper 1990). Autotrophic chemolithotrophs acquire carbon from inorganic sources, while heterotrophic chemolithotrophs require organic carbon as their carbon source. Some bacteria and archaea are also capable of diazotrophy (nitrogen fixation; Delmont et al., 2022; Bombar et al., 2016).

General overview of microbial secondary metabolism

Secondary metabolites, encoded by biosynthetic gene clusters (BGCs), are a diverse set of compounds that play essential roles in ecological interactions within prokaryotic communities (Hibbing et al., 2010), between prokaryotes and eukaryotes (Contreras-Cornejo et al., 2016; Bi et al., 2021), and between eukaryotes (Padder et al., 2018; Jagtap et al., 2020; Li et al., 2023). These metabolites are small compounds that do not contribute directly to cellular growth, maintenance, repair or reproduction, yet can have profound effects on microorganisms by shaping their structure as well as facilitating beneficial and antagonistic interactions (Patin et al., 2017; Chevrette et al., 2022; Musilova et al., 2016). One of the key roles of these compounds in prokaryotes is to function as intercellular signaling molecules (Hibbing et al., 2010). The secretion and uptake of signaling molecules facilitates the coordination of group behaviors and responses to environmental stimuli through processes that regulate gene expression in prokaryotic microbial populations. *N*-acyl homoserine lactones are well-studied example of secondary metabolite autoinducers (signaling molecules) that facilitate quorum sensing (population-wide communication) in bacteria (Fuqua et al., 1994), and possibly in archaea (Zhang et al., 2012).

Secondary metabolites with antimicrobial properties also mediate antagonistic interactions between microorganisms. Microorganisms produce antibiotics as a defense mechanism to inhibit the growth of competitors, or to evade grazers in their environment (Matz and Kjelleberg 2005; Andrić et al., 2023; Teasdale et al., 2009; Wietz et al., 2013). Antibiotic compounds can target specific structures in target organisms, including cell walls and DNA gyrase enzymes, that can lead to growth inhibition, and ultimately cell death (Epand et al., 2016; Phillips et al., 2011). The producer of an antibiotic typically encodes a mechanism (commonly an efflux pump, a

modifying/inactivating enzyme, or a modification to the antibiotic's target structure) to resist the inhibitory effects of the compound that it is exuding into the environment to prevent self-toxicity (Cundliffe 1989). Many classes of secondary metabolites, including polyketides, non-ribosomal peptides (NRPs), terpenes, lactones, and ribosomally-synthesized and post-translationally modified peptides (RiPPs) can possess antibiotic properties (Walsh 2004; Cragg and Newman 2013; Letzel et al., 2014; Mazur et al., 2022; Yamaguchi 2022).

Biosynthetic gene clusters (BGCs) encode instructions to synthesize enzymes that facilitate the biosynthesis of secondary metabolites (Walsh 2004; Wenzel and Müller 2005). In addition to core enzymes involved in biosynthesis, BGCs often encode genes for regulation, export, resistance, and tailoring of the final product (Blin et al., 2017). While the resulting structures of secondary metabolites can vary, the core biosynthetic genes encoded in BGCs are typically highly conserved, which facilitates genome mining efforts to unveil the biosynthetic potential of recovered environmental genomes (Blin et al., 2017). The emergence of computational tools in recent years to detect BGCs in silico has been a major achievement in computational biology. Annotation of BGCs in environmental genomes has revealed a hidden world of potential for ecological interactions within prokaryotic and eukaryotic microbial communities as well as between prokaryotes and eukaryotes (Paoli et al., 2022; Li et al., 2018; Malit et al., 2021; Yan and Matsuda 2024). Large-scale genome mining efforts and localized environmental studies have shown that the potential of prokaryotes and eukaryotes to produce secondary metabolite is significantly more widespread than what has been observed in laboratory settings. This is most likely due to the lack of specific but critically important environmental stimuli in typical culturing experiments (Zazapoulos et al., 2003; Seyedsayamdost 2019). Understanding the biosynthesis, regulation, and ecological functions of secondary metabolites is essential for unraveling the complex networks of interactions that govern microbial community dynamics and ecosystem functioning.

Advances in genome sequencing technology, bioinformatics tools, and machine learning applications for prediction of microbial metabolism

Predicting metabolism in prokaryotes and eukaryotes requires a roadmap of the complex biochemical networks that govern cellular functions, energy production, and the synthesis of organic molecules. Our knowledge of microbial biochemistry has expanded since the integration

of next-generation sequencing (massively parallel, high-throughput sequencing technology; Slatko et al., 2018), third-generation sequencing (single molecule real-time sequencing of longer reads, and detection of DNA modification; Athanasopoulou et al., 2021), computational biology, systems biology, and machine learning (McElhinney et al., 2022; Jiang et al., 2022; Faure et al., 2021).

Metagenomics currently utilizes next- and third-generation sequencing strategies to study the collective metabolic capabilities encoded in the DNA of whole microbial communities from environmental samples (Akaçin et al., 2022). This culture-independent approach yields insights into how different microbes may contribute to overall community metabolism and on their potential to interact with their surroundings. There are several challenges associated with analyzing metagenomics datasets. One major challenge is distinguishing genetic data from different closely-related species and sub-species within microbial communities due to the difficulty of clustering DNA sequences from closely related taxa separately into distinct bins (Quince et al., 2017). This can lead to the presence of chimeric sequences in metagenome-assembled genomes (MAGs). Additionally, there is composition-based bias in the coverage of metagenomic assemblies and their resulting MAGs, often in the form of under-coverage of GC-rich and GC-poor regions (Browne et al., 2020). Adjustments to protocols for sequencing library preparation can reduce coverage biases (Browne et al., 2020). Fluorescence-activated cell sorting coupled with single-cell amplified genome (SAG) sequencing provides an alternative approach (Rinke et al., 2014; Bowers et al., 2017). This method facilitates the strain-resolved analysis of microbial population structure in resulting genomic assemblies (Chen et al., 2020). Single-cell genomics approaches, however, often yield fragmented assemblies because they typically do not amplify a cell's genome completely due to the low quantity of DNA within a single cell (Chen et al., 2020).

The incomplete, fragmented recovery of MAGs and SAGs from environmental 'omics datasets can bias inferences into their functional potential due to missing, contaminated, or fragmented gene content (Eisenhofer et al., 2023; Chen et al., 2020). The inherent complexity of eukaryotic genomes (including expansive repetitive regions) further contributes to the difficulty in recovering their genomes from metagenomes, whole-genome sequencing, and single-cell sequencing (Tørresen et al., 2019; Saraiva et al., 2023; Biscotti et al., 2015). Long-read sequencing technologies, however, can help reduce the fragmented nature of assemblies produced by short-read sequencing (Koren and Phillippy 2015; Chakraborty et al., 2016). Taken together, the advent

of metagenomics and single-cell genomics have produced significant insights into the coding potential of uncultured microbial populations.

Bioinformatic tools are essential for all steps in the analyses of 'omics datasets. Algorithms and pipelines have been developed to run quality control on sequencing reads (Martin 2011; Bolger et al., 2014), assemble DNA sequences (Li et al., 2015; Bankevich et al., 2012) to extract genomes coming from individual taxa from assemblies (Kang et al., 2019; Sieber et al., 2018), and assemble RNA (Haas et al., 2013) for metatranscriptome analyses to examine expressed genes within communities in environmental sequence read datasets. Computational methods have also been created to map reads to assemblies for quantification purposes (Patro et al., 2017; Li 2013), as well as to predict gene sequences and annotate them to specific protein families (Hyatt et al., 2010; Aramaki et al., 2020; Altschul et al., 1990). Gene annotations can then be mapped to the reaction steps of metabolic pathways (Neely et al., 2020; Karp et al., 2021). This core computational framework provides valuable insights into the functional potential and ecological roles of microorganisms and has become increasingly available to the scientific community through the creation of open-source tools and pipelines.

One of the first steps in predicting metabolism is annotating genomes and transcriptomes to identify genes and their functions. Tools such as Prokka (Seemann 2014) and RAST (Rapid Annotation using Subsystem Technology; Aziz et al., 2008) or pipelines such as MetaSanity (Neely et al., 2020) or EukMetaSanity for eukaryotes (Neely et al., 2021) are used to predict the function of genes. For gene annotations of either prokaryotes or eukaryotes, tools such as KofamScan (Aramaki et al., 2020), DRAM (Shaffer et al., 2020), InterProScan (Jones et al., 2014), eggNOG-mapper (Huerta-Cepas et al., 2017), and BLAST (Altschul et al., 1990) are utilized. Once genes are annotated, they can be mapped to metabolic pathways using databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa 2002). KEGG is a resource that contains biochemical networks classified into function-based hierarchical structures that consist of metabolic pathways, modules, and reactions composed of orthologous groups of genes. MetaCyc (Karp et al., 2002) is another metabolic pathway database which contains experimentally verified pathways and reactions from prokaryotic and eukaryotic organisms. Cross-referencing gene annotations to metabolic pathway databases facilitates analysis of the metabolic capabilities of organisms.

Mapping gene annotations to metabolic pathway databases can underestimate the full biochemical potential of a genome or transcriptome due to several factors. Aside from possible incomplete representation of metabolic pathways in databases, genes may be missing from a query dataset due to incomplete sequence recovery from an assembly, or they may not be labelled by homology-based annotation tools if they contain highly divergent nucleotide/amino acid sequences (Sinha et al., 2020). Genes can be missing from the genomes of taxa (or be present in a fragmented or nonfunctional form) because they have lost or are losing the capacity for a particular function (Douglas et al.,2024). For metagenomes produced from environmental samples, which often recover only incomplete genomes for individual taxa (Li et al., 2022), it is thus challenging to interpret the significance of missing genes. To address gaps in metabolic pathways, genome-scale metabolic models (GEMs) are computational models that represent the complete set of metabolic reactions in an organism (Durot et al., 2008). They are optimal only for complete genomes or nearly-complete genomes that contain few gaps or poorly annotated gene content. Tools like the COBRA Toolbox (Constraint-Based Reconstruction and Analysis Toolbox, Heirendt et al., 2019) and Gapseq (Zimmerman et al., 2021) are used to build and analyze GEMs. A common approach to studying GEMs is flux balance analysis (FBA), which uses mathematical constraints to predict the flow of metabolites through a metabolic network (Fell and Small 1986; Watson 2000; Orth et al., 2010). It can predict growth rates, optimal metabolic pathways, and responses to environmental changes. Thermodynamics-based modeling considers the energy changes associated with metabolic reactions to predict feasible metabolic pathways, and Thermodynamics-based Flux Balance Analysis (TFA) is one example (Henry et al., 2007). Additionally, methods that combine genomic, transcriptomic, proteomic, and/or metabolomic data can facilitate more comprehensive analysis of prokaryotic and eukaryotic metabolism (e.g., Chong et al., 2018; Sun et al., 2014; Kamburov et al., 2011). This systems biology approach provides insights into how different layers of cellular information interact to influence metabolism.

Emerging machine-learning techniques are now being applied to metabolism predictions in microbial genomes and transcriptomes. Machine-learning algorithms, such as random forests (Lambert et al., 2022; Alexander et al., 2023), support vector machines (Weimann et al., 2016), and deep learning algorithms (Guo et al., 2017; Shah et al., 2022) can be trained on genomic and transcriptomic data to predict the metabolic capabilities or phenotype of organisms. These models

leverage genomic features, such as gene content, synteny, and sequence similarity, to predict the presence of specific metabolic pathways and to infer the phenotype of microbes.

Microorganisms are essential drivers of all of Earth's biogeochemical cycles. Bioinformatics tools and machine-learning techniques provide powerful tools for mapping gene annotations to metabolic processes in prokaryotes and eukaryotes, offering valuable insights into the metabolic potential of microbial genomes and facilitating the study of microbial ecology. Accurately identifying metabolic pathways is crucial to assessing their role in the environment and in interactions with other organisms. However, environmental genomes (MAGs, SAGs) are often incomplete due to the limitations of sequencing platforms, bioinformatics algorithms, and gene annotation tools and databases. This makes it difficult to predict complete metabolic networks. Existing methods for creating GEMs are less effective for highly incomplete genomes and can lead to inaccurate predictions (Palù et al., 2022; Bernstein et al., 2021). As computational methods for moderately to highly incomplete ‘omics datasets continue to advance, so will our understanding of the networks of metabolic pathways and biochemical transformations that underpin the diversity and functionality of microbial life.

Oxygen-deficient and anaerobic environments in the world’s oceans

Marine oxygen-deficient water columns and sediments are prevalent globally and play critical roles in biogeochemical cycles, serving as habitats for unique microbial communities. Examples of ocean regions that have oxygen-depleted and anoxic water columns and sediments include the Eastern Tropical North and South Pacific Oxygen Minimum Zones, the Arabian Sea, and the Black Sea, the Cariaco Basin off the coast of Venezuela and Saanich Inlet in British Columbia. These ecosystems are characterized by significantly reduced or undetectable oxygen levels, and they serve as analogs of ancient ecosystems that existed before the Great Oxidation Event, often providing insights into early-diverging taxa and their biogeochemical processes. They also are analogs for some potential extraterrestrial ecosystems with anoxic conditions. Climate change and eutrophication are expanding the extent and intensity of many oxygen-deficient environments, impacting marine biodiversity, fisheries, and biogeochemical cycling (Wright et al., 2012; Breitburg et al., 2018; Bhuiyan et al., 2024), highlighting the importance of monitoring and understanding the dynamics of these habitats in a changing ocean. Some of these ecosystems harbor microorganisms that can degrade pollutants and synthesize novel bioactive compounds,

underscoring their ecological and biotechnological significance (Varjani et al., 2017; Oren et al., 1992; Zhang et al., 2022; Zhang et al., 2005; Singh et al., 2017). Research of these habitats is not only fundamental for evolutionary biology and microbial ecology but is also essential for addressing contemporary environmental challenges and for harnessing their potential biotechnological applications.

Oxygen-deficient water columns (ODWCs) are marine environments operationally defined based on oxygen concentrations ranging from dysoxic (20-90 μM), suboxic (1-20 μM), anoxic (> 1 μM) or euxinic (anoxic and sulfidic; Wright et al., 2012). These ecosystems include permanently-stratified basins, oxygen minimum zones (OMZs), and anoxic marine zones (AMZs) and cover an area that accounts for 8% of the world's oceans' surface (Paulmier and Ruiz Pino 2009). AMZs are regions in which oxygen depletion is more severe than in OMZs (Ulloa et al., 2012; Garcia-Robledo et al., 2017). At oxic/anoxic interfaces when oxygen becomes undetectable in AMZs there is typically a peak in nitrite, and below this depth sulfide can accumulate (Ulloa et al., 2012; Garcia-Robledo et al., 2017). The eastern tropical North and South Pacific as well as the Arabian Sea are examples of AMZs; OMZs include the Bay of Bengal and northeast Pacific (Ulloa et al., 2012; Garcia-Robledo et al., 2017). The Cariaco Basin and the Black Sea are examples of anoxic basins (Murray et al., 1989; Rodriguez-Mora et al., 2013). Limited mixing of ODWCs with surrounding waters due to basin geometry (i.e., presence of shallow sills) or stratification restricts the replenishment of oxygenated water in these zones (Wright et al., 2012; Friedrich et al., 2014). Remaining oxygen is consumed as aerobic microbes remineralize organic matter that descends from the surface ocean (Paulmier and Ruiz-Pino 2009; Wright et al., 2012). Climate change exacerbates oxygen deficiency in these water columns by increasing thermal stratification that restricts water mass circulation (Gilly et al., 2013; Altieri et al., 2015). Eutrophication caused by anthropogenic inputs is an additional factor that contributes to the intensification of oxygen deficiency in coastal ODWCs.

The microbial inhabitants of ODWCs are adapted to survive in oxygen-deficient conditions. They are primarily composed of strict and facultative anaerobes that perform chemoorganotrophic and chemolithoautotrophic metabolisms using various terminal electron accepters to obtain energy. The use of alternative electron accepters to oxygen in ODWCs can result in the production of climate active gases including nitrous oxide and methane. High rates of dark carbon fixation have been previously recorded in and are characteristic of ODWCs (Juniper

and Brinkhurst 1986; Taylor et al., 2001; Lengger et al., 2019). Knowledge of the microbial contributions to biogeochemical cycles in ODWCs are essential for evaluating the long-term impacts of the expansion and intensification of these environments.

Sediment ecosystems constitute at least two-thirds of the Earth's surface and are situated on the ocean floor. Marine sediments host a heterogenous microbial biosphere with activity that can act as a control on the global climate and on the carbon cycle. Organic carbon burial in sediments leads to a net removal of carbon dioxide from the atmosphere (Burdige 2007). Approximately 80% of organic matter burial in marine sediments occurs in coastal shelf and slope regions, where primary productivity in surface waters is fueled by several factors including wind-driven nutrient upwelling as well as terrestrial and anthropogenic inputs of nutrients (Gattuso et al., 1998; Rabalais et al., 2009; Grantham et al., 2004). Marine sediments consist of particulate organic matter originating from deceased and sinking organisms, fecal pellets, cellular exudates, black carbon, and other organic carbon sources that descend and accumulate on the ocean floor, resulting in Earth's most expansive pool of organic carbon. Much of the particulate and dissolved organic matter that accumulates in sediments is recalcitrant and difficult for microorganisms to catabolize (Burdige 2007). Microbial communities occupying seafloor and subseafloor ecosystems make up approximately 30% of all the Earth's living biomass (Whitman et al., 1998) and process both organic and inorganic carbon while contributing to the biogeochemical cycling of nitrogen, sulfur, and iron compounds. Electron donors for catabolic metabolism in the subseafloor include organic matter, reduced minerals, and hydrogen (Blair et al., 2007). Electron acceptors include oxygen, nitrate, and sulfate (D'Hondt et al., 2019). Despite the global significance of these microbial communities, relatively little is known about the taxonomic diversity of the microorganisms occupying marine sediments. This is due in part to the heterogeneity of sediment-occupying microbial communities, and the complicated nature of sampling these habitats.

The catabolic metabolism rate of microbes inhabiting marine sediments is several orders of magnitude slower than microorganisms grown on nutrient-rich media (Hoehler and Jørgensen, 2013). Growth is at most, very slow for microorganisms in deeper marine sediments, and evidence indicates that much if not all the limited energy of active cells in these ecosystems is used for maintenance activities (Bradley et al., 2018 Arndt et al., 2006; D'Hondt et al., 2009; Mara et al., 2023). Dormancy is a transient and reversible state characterized by low metabolic activity. Dormant microbes are widespread in marine sediments, which enables them to endure unfavorable

conditions for extended periods of time without dividing (Lever et al., 2015; Jørgensen 2011). Further studies of marine sediment microbial communities are essential for understanding the fate of buried carbon, including microorganisms, and whether they go dormant or remain active. Studies of marine sediments can also elucidate the role of biogeochemical cycles in carbon sequestration, which has important implications in the context of climate change and methods of carbon sequestration being actively considered. Finally, these studies can uncover novel microbial processes and interactions, enhancing our knowledge of one of the planet's most extensive ecosystems, and providing potential biotechnological applications.

Thesis summary

Chapter 2

ODWCs are ocean regions with very little or undetectable levels of oxygen that include anoxic basins and oxygen minimum zones. These regions have expanded globally due to climate change and pollution. The microbial communities in ODWCs are well-studied for their role in carbon, nitrogen, sulfur, and trace metal cycling. The potential for secondary metabolite production and expression in these environments, however, has been largely uncharacterized in genome mining efforts despite the major advances made in charting the secondary metabolic potential of prokaryotes globally.

To address this knowledge gap in Chapter 2, I analyzed metagenomic and metatranscriptomic samples collected along the Cariaco Basin redoxcline and mined them for BGCs, with a focus on differences among samples from two different size fractions: the particle-associated and free-living fractions. I recovered, annotated, and taxonomically labelled prokaryotic genomes. I then identified BGCs encoded within their genomes using antiSMASH (Blin et al., 2021), a bioinformatics pipeline used to detect BGCs in genomes. The identified gene clusters encoded diverse and bioactive compounds that facilitate intercellular communication as well as antagonist interactions. These included antibiotics (NRPS, polyketides, RiPPs, terpenes, lactones) as well as aryl polyenes, which protect against oxidative stress and can facilitate biofilm formation (Johnston et al., 2021). These findings provide a snapshot of the potential to express secondary metabolites in microbial communities inhabiting ODWCs.

Chapter 3

Microorganisms play a crucial role in Earth's biogeochemical cycles, and understanding their metabolic pathways through studies of their genomic data is essential for gaining insights into their ecological interactions and environmental impact given the fact that only a fraction of microorganisms in nature have been brought into culture. Advances in genomic sequencing and bioinformatics algorithms facilitate the recovery of genomes from environmental samples, however most recovered genomes are moderately to highly incomplete. Challenges remain in identifying the metabolic potential of incomplete genomes due to limitations of protein annotation methods and metabolic pathway gapfilling tools. Current methods for metabolic network gapfilling, such as network topology-based and parsimony-based approaches, are not designed to predict the presence of metabolic pathways encoded within highly incomplete genomes.

To address these limitations in Chapter 3, I developed MetaPathPredict, an open-source tool using deep learning to predict the presence of KEGG metabolic modules in bacterial genomic datasets, including isolate genomes, MAGs, and SAGs. MetaPathPredict integrates manually curated metabolic modules from the KEGG database with machine learning models trained on gene features from high-quality genomes. MetaPathPredict's deep learning models, trained on diverse bacterial isolate genomes and MAGs, demonstrated robust macro precision and recall, even on genomes of very low completeness. MetaPathPredict enhances the study of metabolic potential in environmental microbiomes, providing a valuable resource to gain further insight into microbial metabolism in the environment.

Chapter 4

The Guaymas Basin is a hydrothermally-active ocean spreading center in the Gulf of California with pronounced geothermal and geochemical gradients. The deep sediments deposited on its seafloor are host to microbial communities that perform anaerobic metabolic transformations. Strong geothermal heatflow from magmatic sill intrusions and hydrothermal fluids drives the pyrolysis of buried organic carbon to form a complex mixture of petroleum hydrocarbons, carboxylic acids, and ammonia, which are transported by hydrothermal fluids, fostering diverse and active microbial communities. These microbes perform chemosynthetic carbon fixation, heterotrophic organic matter remineralization, and assimilate fossil carbon into the benthic biosphere. Despite its potential, the deep biosphere of Guaymas Basin below the most surficial

sediments that have been studied intensively, remains underexplored, with limited studies to date. These studies included methanogen enrichments from deep biosphere sediments and bacterial and archaeal diversity surveys using 16S rRNA amplicon sequencing. The spatial extent, diversity, and metabolic activity of its deep biosphere has remained largely unknown.

I bridged the gap in knowledge of the metabolic potential of the Guaymas Basin deep biosphere in Chapter 4, by analyzing deep drill core metagenomic and metatranscriptomic data sets produced from samples collected during the International Ocean Discovery Program Expedition 385. Bacterial and archaeal taxonomic distributions, potential metabolisms, and transcriptional activity were analyzed along geothermal and geochemical gradients. Results indicated that while moderate temperatures correlated with biogeochemical parameters influencing microbial community composition, temperatures above 45ºC significantly reduced microbial diversity. However, specific archaeal lineages, including orders from the Thermoproteota and Hadarchaeota, thrived under these more extreme conditions, marking the transition to a specialized deep, hot biosphere. This study underscored the influence of temperature and energy availability on microbial survival in the deep subsurface and explored the genomic potential of microbes in Guaymas Basin's varying geothermal and geochemical environments, contributing to an understanding of the hydrothermally-influenced deep biosphere's diversity and activity.

Chapter 5

Eukaryotic metabolism displays remarkable diversity across its kingdoms, reflecting varied lifestyles. The complexities of eukaryotic genome architecture pose significant challenges for genome assembly and annotation algorithms, leading to challenges in predicting the metabolic pathways they encode. The improvements of next-generation sequencing technologies and advancements in bioinformatics methods have made possible the extraction of eukaryotic genomes from environmental datasets, as well as the prediction and functional annotation of their genes, allowing for insight into their metabolic potential and trophic modes. Due to large genome sizes and the complexities of eukaryotic genomes, most eukaryotic genomes recovered from environmental samples, however, are highly incomplete and KEGG metabolic module prediction tools for eukaryotes are lacking.

To address these limitations in Chapter 5, I designed MetaPathPredict-E, a deep learning-powered extension of MetaPathPredict for eukaryotic metabolic module prediction. This tool reconstructs and predicts KEGG metabolic modules from eukaryotic datasets including isolate genomes and transcriptomes, as well as MAGs and SAGs. MetaPathPredict-E's models are trained on data from taxonomically diverse eukaryotic genomes and transcriptomes, leveraging metabolic module information from the KEGG database. MetaPathPredict-E's deep learning models demonstrated robust macro precision and recall on test datasets, even when those data were highly incomplete. By facilitating the study of eukaryotic genomes and transcriptomes from environmental samples, MetaPathPredict-E enhances the ability to decipher eukaryotic metabolic potential from environmental samples.

Chapter 6

In Chapter 6, I summarize the implications of my thesis research, and include an exploration of some exciting avenues for future research in the fields of marine microbiology, computational biology, and machine learning.

References

Akacin, Ilayda, et al.,"Comparing the significance of the utilization of next generation and third generation sequencing technologies in microbial metagenomics." Microbiological Research 264 (2022): 127154.

Alexander, H., Hu, S. K., Krinos, A. I., Pachiadaki, M., Tully, B. J., Neely, C. J., & Reiter, T. (2023). Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *mBio*, *14*(6), e01676-23.

Altieri, A. H., & Gedan, K. B. (2015). Climate change and dead zones. *Global change biology*, *21*(4), 1395-1406.

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

Andrić, S., Rigolet, A., Argüelles Arias, A., Steels, S., Hoff, G., Balleux, G., ... & Ongena, M. (2023). Plant-associated Bacillus mobilizes its secondary metabolites upon perception of the siderophore pyochelin produced by a Pseudomonas competitor. *The ISME Journal*, *17*(2), 263-275.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), 2251-2252.

Arndt, S., Brumsack, H. J., & Wirtz, K. W. (2006). Cretaceous black shales as active bioreactors: a biogeochemical model for the deep biosphere encountered during ODP Leg 207 (Demerara Rise). *Geochimica et Cosmochimica Acta*, *70*(2), 408-425.

Athanasopoulou, Konstantina, et al.,"Third-generation sequencing: the spearhead towards the radical transformation of modern genomics." Life 12.1 (2021): 30.

Averill, B. A., & Tiedje, J. M. (1982). The chemical mechanism of microbial denitrification. *Febs Lett*, *138*(1), 8-12.

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

Bankevich, Anton, et al.,"SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." Journal of computational biology 19.5 (2012): 455-477.

Bernstein, D. B., Sulheim, S., Almaas, E., & Segrè, D. (2021). Addressing uncertainty in genome-scale metabolic model reconstruction and analysis. *Genome Biology*, *22*, 1-22.

Bhuiyan, M. M. U., Rahman, M., Naher, S., Shahed, Z. H., Ali, M. M., & Islam, A. R. M. T. (2024). Oxygen declination in the coastal ocean over the twenty-first century: Driving forces, trends, and impacts. *Case Studies in Chemical and Environmental Engineering*, *9*, 100621.

Bi, B., Wang, K., Zhang, H., Wang, Y., Fei, H., Pan, R., & Han, F. (2021). Plants use rhizosphere metabolites to regulate soil microbial diversity. *Land Degradation & Development*, *32*(18), 5267-5280.

Blair, C. C., D'Hondt, S., Spivack, A. J., & Kingsley, R. H. (2007). Radiolytic hydrogen and microbial respiration in subsurface sediments. *Astrobiology*, *7*(6), 951-970.

Blin, K., Kim, H. U., Medema, M. H., & Weber, T. (2019). Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics*, *20*(4), 1103-1113.

Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., Van Wezel, G. P., Medema, M. H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic acids research*, *49*(W1), W29-W35.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data." Bioinformatics 30.15 (2014): 2114-2120.

Bombar, Deniz, Ryan W. Paerl, and Lasse Riemann. "Marine non-cyanobacterial diazotrophs: moving beyond molecular detection." Trends in microbiology 24.11 (2016): 916-927.

Bowers, Robert M., et al.,"Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea." Nature biotechnology 35.8 (2017): 725-731.

Bradley, J. A., Amend, J. P., & LaRowe, D. E. (2018). Bioenergetic controls on microbial ecophysiology in marine sediments. *Frontiers in microbiology*, *9*, 324724.

Breitburg, D., Levin, L. A., Oschlies, A., Grégoire, M., Chavez, F. P., Conley, D. J., ... & Zhang, J. (2018). Declining oxygen in the global ocean and coastal waters. *Science*, *359*(6371), eaam7240.

Browne, Patrick Denis, et al.,"GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms." GigaScience 9.2 (2020): giaa008.

Burdige, D. J. (2007). Preservation of organic matter in marine sediments: controls, mechanisms, and an imbalance in sediment organic carbon budgets? *Chemical reviews*, *107*(2), 467-485.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*, *38*(12), 5825-5829.

Chakraborty, Mahul, et al.,"Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage." Nucleic acids research 44.19 (2016): e147-e147.

Chen, Lin-Xing, et al.,"Accurate and complete genomes from metagenomes." Genome research 30.3 (2020): 315-333.

Chen, Yin, et al.,"Life without light: microbial diversity and evidence of sulfur-and ammonium-based chemolithotrophy in Movile Cave." The ISME journal 3.9 (2009): 1093-1104.

Chevrette, M. G., Thomas, C. S., Hurley, A., Rosario-Meléndez, N., Sankaran, K., Tu, Y., ... & Handelsman, J. (2022). Microbiome composition modulates secondary metabolism in a multispecies bacterial community. *Proceedings of the National Academy of Sciences*, *119*(42), e2212930119.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S., Bourque, G., ... & Xia, J. (2018). MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic acids research*, *46*(W1), W486-W494.

Contreras-Cornejo, H. A., Macías-Rodríguez, L., Del-Val, E. K., & Larsen, J. (2016). Ecological functions of Trichoderma spp. and their secondary metabolites in the rhizosphere: interactions with plants. *FEMS microbiology ecology*, *92*(4), fiw036.

Cragg, G. M., & Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1830*(6), 3670-3695.

Cundliffe, E. (1989). How antibiotic-producing organisms avoid suicide. *Annual review of microbiology*, *43*(1), 207-233.

Delmont, Tom O., et al.,"Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean." The ISME Journal 16.4 (2022): 927-936.
Douglas, Gavin M., and B. Jesse Shapiro. "Pseudogenes act as a neutral reference for detecting selection in prokaryotic pangenomes." Nature Ecology & Evolution 8.2 (2024): 304-314.
D'Hondt, S., Pockalny, R., Fulfer, V. M., & Spivack, A. J. (2019). Subseafloor life and its biogeochemical impacts, Nat. Commun., 10, 3519.
D'Hondt, S., Spivack, A. J., Pockalny, R., Ferdelman, T. G., Fischer, J. P., Kallmeyer, J., ... & Stancin, A. M. (2009). Subseafloor sedimentary life in the South Pacific Gyre. *Proceedings of the National Academy of Sciences*, *106*(28), 11651-11656.
DeBerardinis, R. J., & Thompson, C. B. (2012). Cellular metabolism and disease: what do metabolic outliers teach us? *Cell*, *148*(6), 1132-1144.
Durot, Maxime, Pierre-Yves Bourguignon, and Vincent Schachter. "Genome-scale models of bacterial metabolism: reconstruction and applications." FEMS microbiology reviews 33.1 (2008): 164-190.
Ehrenreich, Armin, and Friedrich Widdel. "Anaerobic oxidation of ferrous iron by purple bacteria, a new type of phototrophic metabolism." Applied and environmental microbiology 60.12 (1994): 4517-4526.
Eiler, A. (2006). Evidence for the ubiquity of mixotrophic bacteria in the upper ocean: implications and consequences. *Applied and Environmental Microbiology*, *72*(12), 7431-7437.
Eisenhofer, Raphael, Iñaki Odriozola, and Antton Alberdi. "Impact of microbial genome completeness on metagenomic functional inference." ISME communications 3.1 (2023): 12.
Epand, R. M., Walker, C., Epand, R. F., & Magarvey, N. A. (2016). Molecular mechanisms of membrane targeting antibiotics. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, *1858*(5), 980-987.
Evans, S. E., Franks, A. E., Bergman, M. E., Sethna, N. S., Currie, M. A., & Phillips, M. A. (2024). Plastid ancestors lacked a complete Entner-Doudoroff pathway, limiting plants to glycolysis and the pentose phosphate pathway. *Nature Communications*, *15*(1), 1102.
Faure, Emile, Sakina-Dorothée Ayata, and Lucie Bittner. "Towards omics-based predictions of planktonic functional composition from environmental data." Nature Communications 12.1 (2021): 4361.
Fell, D. A., & Small, J. R. (1986). Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochemical journal*, *238*(3), 781-786.
Fewson, C. A., & Nicholas, D. J. D. (1961). Utilization of nitrate by micro-organisms. *Nature*, *190*(4770), 2-7.
Fredrickson, J. K., Kostandarithes, H. M., Li, S. W., Plymale, A. E., & Daly, M. (2000). Reduction of fe (III), cr (VI), U (VI), and tc (VII) by Deinococcus radiodurans R1. *Applied and Environmental Microbiology*, *66*(5), 2006-2011.
Friedrich, J., Janssen, F., Aleynik, D., Bange, H. W., Boltacheva, N., Çagatay, M. N., ... & Wenzhöfer, F. (2014). Investigating hypoxia in aquatic environments: diverse approaches to addressing a complex phenomenon. *Biogeosciences*, *11*(4), 1215-1259.
Frigaard, Niels-Ulrik, and Christiane Dahl. "Sulfur metabolism in phototrophic sulfur bacteria." Advances in microbial physiology 54 (2008): 103-200.

Fuqua, W. C., Winans, S. C., & Greenberg, E. P. (1994). Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. *Journal of bacteriology*, *176*(2), 269-275.
Garcia-Robledo, Emilio, et al.,"Cryptic oxygen cycling in anoxic marine zones." Proceedings of the National Academy of Sciences 114.31 (2017): 8319-8324.
Gattuso, J. P., Frankignoulle, M., & Wollast, R. (1998). Carbon and carbonate metabolism in coastal aquatic ecosystems. *Annual Review of Ecology and Systematics*, *29*(1), 405-434.
Gilly, W. F., Beman, J. M., Litvin, S. Y., & Robison, B. H. (2013). Oceanographic and biological effects of shoaling of the oxygen minimum zone. *Annual review of marine science*, *5*, 393-420.
Grantham, B. A., Chan, F., Nielsen, K. J., Fox, D. S., Barth, J. A., Huyer, A., ... & Menge, B. A. (2004). Upwelling-driven nearshore hypoxia signals ecosystem and oceanographic changes in the northeast Pacific. *Nature*, *429*(6993), 749-754.
Guest, J. (1979). Anaerobic growth of Escherichia coli K12 with fumarate as terminal electron acceptor. Genetic studies with menaquinone and fluoroacetate-resistant mutants. *Microbiology*, *115*(2), 259-271.
Guo, W., Xu, Y., & Feng, X. (2017). DeepMetabolism: a deep learning system to predict phenotype from genome sequencing. *arXiv preprint arXiv:1705.03094*.
Gupta, R., & Gupta, N. (2021). *Fundamentals of bacterial physiology and metabolism* (pp. 267-287). Singapore: Springer.
Haas, Brian J., et al.,"De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." Nature protocols 8.8 (2013): 1494-1512.
Heider, J., Spormann, A. M., Beller, H. R., & Widdel, F. (1998). Anaerobic bacterial metabolism of hydrocarbons. *FEMS microbiology reviews*, *22*(5), 459-473.
Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., ... & Fleming, R. M. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v. 3.0. *Nature protocols*, *14*(3), 639-702.
Henry, C. S., Broadbelt, L. J., & Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophysical journal*, *92*(5), 1792-1805.
Hibbing, M. E., Fuqua, C., Parsek, M. R., & Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature reviews microbiology*, *8*(1), 15-25.
Hoehler, T. M., & Jørgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nature Reviews Microbiology*, *11*(2), 83-94.
Hohmann-Marriott, Martin F., and Robert E. Blankenship. "Evolution of photosynthesis." *Annual review of plant biology* 62 (2011): 515-548.
Hyatt, Doug, et al.,"Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC bioinformatics 11 (2010): 1-11.
Jagtap, S. S., Bedekar, A. A., & Rao, C. V. (2020). Quorum sensing in yeast. In *Quorum Sensing: Microbial Rules of Life* (pp. 235-250). American Chemical Society.
Jiang, Yiru, et al.,"Machine learning advances in microbiology: A review of methods and applications." Frontiers in Microbiology 13 (2022): 925454.
Johnson, M. D. (2011). The acquisition of phototrophy: adaptive strategies of hosting endosymbionts and organelles. *Photosynthesis research*, *107*, 117-132.

Johnston, I., Osborn, L. J., Markley, R. L., McManus, E. A., Kadam, A., Schultz, K. B., ... & Claesen, J. (2021). Identification of essential genes for Escherichia coli aryl polyene biosynthesis and function in biofilm formation. *npj Biofilms and Microbiomes*, *7*(1), 56.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., ... & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236-1240.

Jørgensen, B. B. (2011). Deep subseafloor microbial cells on physiological standby. *Proceedings of the National Academy of Sciences*, *108*(45), 18193-18194.

Juniper, S. K., & Brinkhurst, R. O. (1986). Water-column dark CO_2 fixation and bacterial-mat growth in intermittently anoxic Saanich Inlet, British Columbia. *Marine Ecology Progress Series*, 41-50.

Kamburov, A., Cavill, R., Ebbels, T. M., Herwig, R., & Keun, H. C. (2011). Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, *27*(20), 2917-2918.

Kamp, A., Høgslund, S., Risgaard-Petersen, N., & Stief, P. (2015). Nitrate storage and dissimilatory nitrate reduction by eukaryotic microbes. *Frontiers in microbiology*, *6*, 174121.

Kanehisa, M. (2002, November). The KEGG database. In *'In silico' simulation of biological processes: Novartis Foundation Symposium 247* (Vol. 247, pp. 91-103). Chichester, UK: John Wiley & Sons, Ltd.

Kang, Dongwan D., et al.,"MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies." PeerJ 7 (2019): e7359.

Karp, P. D., Riley, M., Paley, S. M., & Pellegrini-Toole, A. (2002). The metacyc database. *Nucleic acids research*, *30*(1), 59-61.

Karp, Peter D., et al.,"Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology." Briefings in bioinformatics 22.1 (2021): 109-126.

Kim, Young M., and George D. Hegeman. "Oxidation of carbon monoxide by bacteria." International Review of Cytology 81 (1983): 1-32.

Koren, Sergey, and Adam M. Phillippy. "One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly." Current opinion in microbiology 23 (2015): 110-120.

Kraft, B., Strous, M., & Tegetmeyer, H. E. (2011). Microbial nitrate respiration–genes, enzymes and environmental distribution. *Journal of biotechnology*, *155*(1), 104-117.

Kwong, W. K., Zheng, H., & Moran, N. A. (2017). Convergent evolution of a modified, acetate-driven TCA cycle in bacteria. *Nature microbiology*, *2*(7), 1-3.

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., ... & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, *119*(7), e2100916119.

Lengger, S. K., Rush, D., Mayser, J. P., Blewett, J., Schwartz-Narbonne, R., Talbot, H. M., ... & Pancost, R. D. (2019). Dark carbon fixation in the Arabian Sea oxygen minimum zone contributes to sedimentary organic carbon (SOM). *Global Biogeochemical Cycles*, *33*(12), 1715-1732.

Letzel, A. C., Pidot, S. J., & Hertweck, C. (2014). Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. *BMC genomics*, *15*, 1-21.
Lever, M. A. (2012). Acetogenesis in the energy-starved deep biosphere–a paradox? *Frontiers in microbiology*, *2*, 284.
Lewitus, A. J., Glasgow Jr, H. B., & Burkholder, J. M. (1999). Kleptoplastidy in the toxic dinoflagellate Pfiesteria piscicida (Dinophyceae). *Journal of Phycology*, *35*(2), 303-312.
Li, Dinghua, et al.,"MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph." Bioinformatics 31.10 (2015): 1674-1676.
Li, Heng. "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM." arXiv preprint arXiv:1303.3997 (2013).
Li, L., Pan, Y., Zhang, S., Yang, T., Li, Z., Wang, B., ... & Li, X. (2023). Quorum sensing: cell-to-cell communication in Saccharomyces cerevisiae. *Frontiers in Microbiology*, *14*, 1250151.
Li, Tang, and Yanbin Yin. "Critical assessment of pan-genomic analysis of metagenome-assembled genomes." Briefings in Bioinformatics 23.6 (2022): bbac413.
Li, Y. X., Zhong, Z., Zhang, W. P., & Qian, P. Y. (2018). Discovery of cationic nonribosomal peptides as Gram-negative antibiotics through global genome mining. *Nature communications*, *9*(1), 3273.
Malit, J. J. L., Liu, W., Cheng, A., Saha, S., Liu, L. L., & Qian, P. Y. (2021). Global genome mining reveals a cytochrome P450-catalyzed cyclization of crownlike cyclodipeptides with neuroprotective activity. *Organic Letters*, *23*(17), 6601-6605.
Mara, P., Zhou, Y. L., Teske, A., Morono, Y., Beaudoin, D., & Edgcomb, V. (2023). Microbial gene expression in Guaymas Basin subsurface sediments responds to hydrothermal stress and energy limitation. *The ISME journal*, *17*(11), 1907-1919.
Martin, Marcel. "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet. journal 17.1 (2011): 10-12.
Matz, C., & Kjelleberg, S. (2005). Off the hook–how bacteria survive protozoan grazing. *Trends in microbiology*, *13*(7), 302-307.
Mazur, M., & Masłowiec, D. (2022). Antimicrobial activity of lactones. *Antibiotics*, *11*(10), 1327.
McElhinney, James MWR, et al.,"Interfacing machine learning and microbial omics: a promising means to address environmental challenges." Frontiers in Microbiology 13 (2022): 851450.
Mohan, S. B., & Cole, J. A. (2007). The dissimilatory reduction of nitrate to ammonia by anaerobic bacteria. In *Biology of the nitrogen cycle* (pp. 93-106). Elsevier.
Murray, J. W., et al.,"Unexpected changes in the oxic/anoxic interface in the Black Sea." Nature 338.6214 (1989): 411-413.
Musilova, L., Ridl, J., Polivkova, M., Macek, T., & Uhlik, O. (2016). Effects of secondary plant metabolites on microbial populations: changes in community structure and metabolic activity in contaminated environments. *International journal of molecular sciences*, *17*(8), 1205.
Neely, C. J., Graham, E. D., & Tully, B. J. (2020). MetaSanity: an integrated microbial genome evaluation and annotation pipeline. *Bioinformatics*, *36*(15), 4341-4344.

Neely, C. J., Hu, S. K., Alexander, H., & Tully, B. J. (2021). The high-throughput gene prediction of more than 1,700 eukaryote genomes using the software package EukMetaSanity. *bioRxiv*, 2021-07.
Offre, P., Spang, A., & Schleper, C. (2013). Archaea in biogeochemical cycles. *Annual review of microbiology*, *67*, 437-457.
Oren, A., & Trüper, H. G. (1990). Anaerobic growth of halophilic archaeobacteria by reduction of dimethylsulfoxide and trimethylamine N-oxide. *FEMS Microbiology Letters*, *70*(1), 33-36.
Oren, Aharon, et al.,"Microbial degradation of pollutants at high salt concentrations." Biodegradation 3 (1992): 387-398.
Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, *28*(3), 245-248.
Padder, S. A., Prasad, R., & Shah, A. H. (2018). Quorum sensing: A less known mode of communication among fungi. *Microbiological research*, *210*, 51-58.
Palù, M., Basile, A., Zampieri, G., Treu, L., Rossi, A., Morlino, M. S., & Campanaro, S. (2022). KEMET–A python tool for KEGG Module evaluation and microbial genome annotation expansion. *Computational and Structural Biotechnology Journal*, *20*, 1481-1486.
Paoli, L., Ruscheweyh, H. J., Forneris, C. C., Hubrich, F., Kautsar, S., Bhushan, A., ... & Sunagawa, S. (2022). Biosynthetic potential of the global ocean microbiome. *Nature*, *607*(7917), 111-118.
Papoutsakis, E. T. (2000). Equations and calculations for fermentations of butyric acid bacteria. *Biotechnology and bioengineering*, *67*(6), 813-826.
Patin, N. V., Schorn, M., Aguinaldo, K., Lincecum, T., Moore, B. S., & Jensen, P. R. (2017). Effects of actinomycete secondary metabolites on sediment microbial communities. *Applied and environmental microbiology*, *83*(4), e02676-16.
Patro, Rob, et al.,"Salmon provides fast and bias-aware quantification of transcript expression." Nature methods 14.4 (2017): 417-419.
Paulmier, A., & Ruiz-Pino, D. (2009). Oxygen minimum zones (OMZs) in the modern ocean. *Progress in Oceanography*, *80*(3-4), 113-128.
Phillips, J. W., Goetz, M. A., Smith, S. K., Zink, D. L., Polishook, J., Onishi, R., ... & Singh, S. B. (2011). Discovery of kibdelomycin, a potent new class of bacterial type II topoisomerase inhibitor by chemical-genetic profiling in Staphylococcus aureus. *Chemistry & biology*, *18*(8), 955-965.
Porter, K. G. (1988). Phagotrophic phytoflagellates in microbial food webs. *Hydrobiologia*, *159*, 89-97.
Quince, Christopher, et al.,"DESMAN: a new tool for de novo extraction of strains from metagenomes." Genome biology 18 (2017): 1-22.
Rabalais, N. N., Turner, R. E., Díaz, R. J., & Justić, D. (2009). Global change and eutrophication of coastal waters. *ICES Journal of Marine Science*, *66*(7), 1528-1537.
Rinke, Christian, et al.,"Obtaining genomes from uncultivated environmental microorganisms using FACS–based single-cell genomics." Nature protocols 9.5 (2014): 1038-1048.
Rodriguez-Mora, Maria J., et al.,"Bacterial community composition in a large marine anoxic basin: a Cariaco Basin time-series survey." FEMS microbiology ecology 84.3 (2013): 625-639.
Sánchez-Baracaldo, P., & Cardona, T. (2020). On the origin of oxygenic photosynthesis and Cyanobacteria. *New Phytologist*, *225*(4), 1440-1446.

Saraiva, Joao Pedro, et al.,"Recovery of 197 eukaryotic bins reveals major challenges for eukaryote genome reconstruction from terrestrial metagenomes." Molecular Ecology Resources 23.5 (2023): 1066-1076.

Schäfer, G., Engelhard, M., & Müller, V. (1999). Bioenergetics of the Archaea. *Microbiology and molecular biology reviews*, *63*(3), 570-620.

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

Seyedsayamdost, M. R. (2019). Toward a global picture of bacterial secondary metabolism. *Journal of Industrial Microbiology and Biotechnology*, *46*(3-4), 301-311.

Sieber, Christian MK, et al.,"Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy." Nature microbiology 3.7 (2018): 836-843.

Singh, Monika, et al.,"Endophytic bacteria: a new source of bioactive compounds." 3 Biotech 7 (2017): 1-14.

Sinha, Swati, Andrew M. Lynn, and Dhwani K. Desai. "Implementation of homology based and non-homology based computational methods for the identification and annotation of orphan enzymes: using Mycobacterium tuberculosis H37Rv as a case study." BMC bioinformatics 21 (2020): 1-18.

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

Shah, H. A., Liu, J., Yang, Z., Zhang, X., & Feng, J. (2022). DeepRF: A deep learning method for predicting metabolic pathways in organisms based on annotated genomes. *Computers in Biology and Medicine*, *147*, 105756.

Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel. "Overview of next-generation sequencing technologies." Current protocols in molecular biology 122.1 (2018): e59.

Sørensen, J. (1982). Reduction of ferric iron in anaerobic, marine sediment and interaction with reduction of nitrate and sulfate. *Applied and Environmental Microbiology*, *43*(2), 319-324.

Stolz, J. F., Basu, P., Santini, J. M., & Oremland, R. S. (2006). Arsenic and selenium in microbial metabolism. *Annu. Rev. Microbiol.*, *60*, 107-130.

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

Strohm, T. O., Griffin, B., Zumft, W. G., & Schink, B. (2007). Growth yields in bacterial denitrification and nitrate ammonification. *Applied and environmental microbiology*, *73*(5), 1420-1424.

Sun, H., Wang, H., Zhu, R., Tang, K., Gong, Q., Cui, J., ... & Liu, Q. (2014). iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics*, *30*(5), 737-739.

Taylor, G. T., Iabichella, M., Ho, T. Y., Scranton, M. I., Thunell, R. C., Muller-Karger, F., & Varela, R. (2001). Chemoautotrophy in the redox transition zone of the Cariaco Basin: a significant midwater source of organic carbon production. *Limnology and Oceanography*, *46*(1), 148-163.

Teasdale, M. E., Liu, J., Wallace, J., Akhlaghi, F., & Rowley, D. C. (2009). Secondary metabolites produced by the marine bacterium Halobacillus salinus that inhibit quorum sensing-controlled phenotypes in gram-negative bacteria. *Applied and environmental microbiology*, *75*(3), 567-572.

Tørresen, Ole K., et al.,"Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases." Nucleic acids research 47.21 (2019): 10994-11006.
Ulloa, Osvaldo, et al.,"Microbial oceanography of anoxic oxygen minimum zones." Proceedings of the National Academy of Sciences 109.40 (2012): 15996-16003.
Varjani, Sunita J., and Vivek N. Upasani. "A new look on factors affecting microbial degradation of petroleum hydrocarbon pollutants." International Biodeterioration & Biodegradation 120 (2017): 71-83.
Walsh, C. T. (2004). Polyketide and nonribosomal peptide antibiotics: modularity and versatility. *Science*, *303*(5665), 1805-1810.
Wang, Yinzhao, et al.,"Diverse anaerobic methane-and multi-carbon alkane-metabolizing archaea coexist and show activity in Guaymas Basin hydrothermal sediment." Environmental microbiology 21.4 (2019): 1344-1355.
Weimann, A., Mooren, K., Frank, J., Pope, P. B., Bremges, A., & McHardy, A. C. (2016). From genomes to phenotypes: Traitar, the microbial trait analyzer. *MSystems*, *1*(6), 10-1128.
Wenzel, S. C., & Müller, R. (2005). Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways. *Current opinion in biotechnology*, *16*(6), 594-606.
Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, *95*(12), 6578-6583.
Wietz, M., Duncan, K., Patin, N. V., & Jensen, P. R. (2013). Antagonistic interactions mediated by marine bacteria: the role of small molecules. *Journal of chemical ecology*, *39*, 879-891.
Wright, J. J., Konwar, K. M., & Hallam, S. J. (2012). Microbial ecology of expanding oxygen minimum zones. *Nature Reviews Microbiology*, *10*(6), 381-394.
Yamaguchi, T. (2022). Antibacterial effect of the combination of terpenoids. *Archives of Microbiology*, *204*(8), 520.
Yan, D., & Matsuda, Y. (2024). Global genome mining-driven discovery of an unusual biosynthetic logic for fungal polyketide–terpenoid hybrids. *Chemical Science*.
Yu, H., & Leadbetter, J. R. (2020). Bacterial chemolithoautotrophy via manganese oxidation. *Nature*, *583*(7816), 453-458.
Zazopoulos, E., Huang, K., Staffa, A., Liu, W., Bachmann, B. O., Nonaka, K., ... & Farnet, C. M. (2003). A genomics-guided approach for discovering and expressing cryptic metabolic pathways. *Nature biotechnology*, *21*(2), 187-190.
Zhang, G., Zhang, F., Ding, G., Li, J., Guo, X., Zhu, J., ... & Dong, X. (2012). Acyl homoserine lactone-based quorum sensing in a methanogenic archaeon. *The ISME journal*, *6*(7), 1336-1344.
Zhang, Lixin, et al.,"Exploring novel bioactive compounds from marine microbes." Current opinion in microbiology 8.3 (2005): 276-281.
Zhang, Ting, and Houjin Zhang. "Microbial consortia are needed to degrade soil pollutants." Microorganisms 10.2 (2022): 261.
Zimmermann, J., Kaleta, C., & Waschina, S. (2021). Gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome biology*, *22*, 1-35.

Chapter 2

Diverse secondary metabolites are expressed in particle-associated and free-living microorganisms of the permanently anoxic Cariaco Basin

David Geller-McGrath[1#], Paraskevi Mara[2#], Gordon T. Taylor[3], Elizabeth Suter[3,4], Virginia Edgcomb[2*], Maria Pachiadaki[1*]

[1]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

[2]Geology & Geophysics Department, Woods Hole Oceanographic Institution, Woods Hole, MA, USA

[3]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY, USA

[4]Biology, Chemistry and Environmental Studies Department, Molloy College, Rockville Centre, NY, USA

# These authors contributed equally

* These authors jointly supervised this work

Statement of contribution

M.P, V.E., and G.T. designed the original research project. M.P., V.E., D.G.M., and P.M. designed the secondary metabolite study. G.T, M.P, E.S. and V.E. sampled the ecosystem. M.P conducted DNA and RNA extractions. M.P, assembled the metagenomes and binned metagenome-assembled genomes. D.G.M. performed genome-resolved metagenomics and metatranscriptomics. P.M. manually curated the meta-omics data, and P.M and D.G.M. conducted the analysis of secondary metabolite data. P.M. and D.G.M. wrote the paper with input from M.P. and V.E. G.T. and E.S. contributed to the final manuscript.

This chapter was originally published as:

Geller-McGrath, D., Mara, P., Taylor, G. T., Suter, E., Edgcomb, V. P., & Pachiadaki, M. (2023). Diverse secondary metabolites are expressed in particle-associated and free-living microorganisms of the permanently anoxic Cariaco Basin. *Nature Communications*, *14*(1), 656.

Abstract

Secondary metabolites play essential roles in ecological interactions and nutrient acquisition, and are of interest for their potential uses in medicine and biotechnology. Genome mining for biosynthetic gene clusters (BGCs) can be used for the discovery of new compounds. Here, we use metagenomics and metatranscriptomics to analyze BGCs in free-living and particle-associated microbial communities through the stratified water column of the Cariaco Basin, Venezuela. We recovered 565 bacterial and archaeal metagenome-assembled genomes (MAGs) and identified 1154 diverse BGCs. We show that differences in water redox potential and microbial lifestyle (particle-associated vs. free-living) are associated with variations in the predicted composition and production of secondary metabolites. Our results indicate that microbes, including understudied clades such as Planctomycetota, potentially produce a wide range of secondary metabolites in these anoxic/euxinic waters.

Introduction

Secondary metabolites are low-molecular-mass compounds that are not required for the growth or reproduction of an organism. Nonetheless, they can serve a variety of functions, including the facilitation of intercellular communication, inhibition of competitors, nutrient acquisition, and interactions with the surrounding environment (Hibbing et al., 2010). Many classes of these molecules can have antibiotic properties, such as polyketides, non-ribosomal peptides, and ribosomally synthesized post-translationally modified peptides (RiPPs; Cragg and Newman, 2013). Other examples of these compounds include terpenes, aryl polyenes, and lactones with diverse roles (e.g., pigments, quorum sensing). Groups of co-located genes, referred to as biosynthetic gene clusters (BGCs), encode instructions to build these molecules. While the chemical structures of secondary metabolites vary significantly, the biosynthetic gene sequences that encode them are often highly conserved (Blin et al., 2019). The high similarity of the amino acid sequences of core biosynthetic enzymes facilitates the mining of genome data for the presence

of specific classes of BGCs. Core biosynthetic genes are frequently flanked by regulatory, export, and resistance genes, as well as genes encoding tailoring enzymes that modify the compound scaffold (Blin et al., 2019). Genome mining has revealed that the secondary metabolic potential of both prokaryotes and eukaryotes is much broader than what is observed under laboratory conditions (Gavriilidou et al., 2022; Paoli et al., 2022). This could be due to the absence of specific stimuli in laboratory settings that are requisite to upregulate or activate compound production in cultures (Scherlach and Hertweck, 2009).

Large genome mining efforts have revealed widespread and diverse biosynthetic capability among prokaryotes; yet "extreme" environments such as oxygen-depleted water columns (ODWCs) are underrepresented in these studies (Gavriilidou et al., 2022; Paoli et al., 2022). ODWCs are oceanic realms with low (< 20 μM) to undetectable oxygen concentrations (Gilly et al., 2013), and include permanently-stratified basins as well as oxygen minimum zones. ODWCs have expanded and intensified globally over the past 50 years (Schmidtko et al., 2017) due to global climate change and anthropogenic pollution. This expansion causes changes in water column stratification and upper water column primary production, and results in shifts in the cycling of trace gases that produce feedbacks on climate (e.g., methane, nitrous oxide, carbon dioxide; Naqvi et al., 2010). The Cariaco Basin is a permanently stratified marine system off the north coast of Venezuela. The Basin's water column is fully oxic at the surface but stratification below the mixed layer (<80 m; Scranton et al., 1987; G. T. Taylor et al., 2001) causes a sharp oxygen decline. A strong vertical redox gradient (redoxcline) extends from ~200 m to ~250–350 m depth, where oxygen becomes undetectable. Below 350 m the water becomes euxinic with sulfide concentrations approaching 80 μM near the basin floor (Scranton et al., 2001, 2014). This relatively stable redoxcline makes the Cariaco Basin an ideal natural laboratory for studying how microbes organize and function in specific redox conditions. ODWCs are relatively well-studied regarding the microbially mediated biogeochemical transformations of carbon, nitrogen, sulfur, and redox-sensitive trace metals (Canfield et al., 2010; Dalsgaard et al., 2012; Rapp et al., 2019; Schlosser et al., 2018). However, secondary metabolite genomic potential and expression in ODWCs hasn't yet been studied. Further, analyses of size-fractionated water samples are required in order to assess the role of particles in the production of secondary metabolites in the environment. Particles provide colonizable, nutrient-rich substrates where metabolites can be

concentrated and exchanged and can provide protection for oxygen- or sulfide-sensitive microbiota.

In order to address this critical gap in secondary metabolite knowledge and assess the role of particles in the production of secondary metabolites in ODWC environments, we analyzed size-fractionated water samples along various oxygen and sulfide regimes in the water column of Cariaco Basin. We reconstructed 565 metagenome-assembled genomes (MAGs) and we estimated their relative abundance and fraction partitioning along Cariaco's redoxcline using metagenomic read recruitment and DESeq2 (Love et al., 2014), and we identify the encoded BGCs using antiSMASH (Blin et al., 2021). For this environmental survey of secondary metabolites, we use metatranscriptomes constructed from *in situ* filtration and preservation of water samples to compare the biosynthetic transcript expression profiles of particle-associated (PA > 2.7 μm) and free-living (FL; 0.2-2.7 μm) fractions. *In situ* filtration and fixation minimizes artifacts that can be introduced into RNA pools due to sample handling and physico-chemical changes (Edgcomb et al., 2016). The detected biosynthetic clusters encode for production of auxiliary compounds with chemical diversity and bioactivity that can provide competitive advantages via antimicrobial compounds (e.g., non-ribosomal peptide synthetases [NRPS], polyketides, RiPPs), or can have a broader impact on microbial survival via the synthesis of pigments and toxins (e.g., aryl polyenes, terpenes) or via their possible role in biofilm formation (e.g., RiPPs and phenazines).

Results

We recovered 565 metagenome-assembled genomes (MAGs) with ≥75% bin completeness and ≤5% bin contamination from sulfidic layers of Cariaco Basin using a PA and FL size fraction co-assembly. Recovered MAGs belonged to 44 bacterial and 8 archaeal phyla (Supplementary Figs. 1-2; Supplementary Data 1-2). The overall taxonomic profile resembled patterns previously observed in MAGs recovered from the Black Sea (Cabello-Yeves et al., 2021). Nonetheless, while identified MAGs from the Black Sea affiliated with the Bdellovibrionota and Nitrospirota phyla were not recovered in our Cariaco samples, we did recover genomes from 26 phyla not reported thus far from the Black Sea (Supplementary Data 3). This was likely due to the analysis of two different size fractions in the present dataset.

Differential abundance (fraction partitioning) of recovered genomes

Differential abundance analysis revealed size fraction partitioning of the recovered MAGs from a taxonomic perspective, and the results are largely consistent with marker gene profiles from the same samples (Suter et al., 2017). The majority of MAGs from the Planctomycetota, Myxococcota, Verrumicrobiota, and the candidate phyla Krumholzibacteriota were differentially abundant in PA metagenomes (Supplementary Figure 3). Planctomycetota and Verrumicrobiota were previously reported to be more abundant in the PA fraction of 16S rRNA gene amplicon samples in various marine environments (Duret et al., 2019; J. Li et al., 2021; Mestre, Borrull, et al., 2017; Mestre, Ferrera, et al., 2017; Pelve et al., 2017). However, the two Planctomycetota MAGs belonging to the genus Scalindua, a group known to perform anaerobic ammonia oxidation (anammox; Sinninghe Damsté et al., 2005), were more abundant in the FL metagenomes as shown previously (Fuchsman et al., 2012). Proteobacteria (primarily *Alphaproteobacteria* and *Gammaproteobacteria*), Nanoarchaeota, Crenarchaeota, and Iainarchaeota, as well as the candidate phyla Omnitrophota, Marinisomatota, Margulisbacteria, SAR324, and Patescibacteria were more abundant in the FL fraction. MAGs from the Desulfobacterota and Thermoplasmatota did not exhibit a preferred association with either fraction at oxycline and shallow anoxic depths, while the majority of MAGs from these phyla were more abundant in the FL fraction at the euxinic depth.

Identification of secondary metabolite biosynthetic gene clusters

Anaerobic/microaerophilic bacteria and archaea have been overlooked as a potential source of bioactive secondary metabolites (Scherlach & Hertweck, 2021). Yet, genomic studies now show that these organisms can contain enormous biosynthetic potential, much of which remains unknown (Letzel et al., 2013). The antiSMASH 6 pipeline identified and annotated 1,154 BGCs longer than 10kb (1,369 total clusters identified), which contained 23,845 genes, in 68% of the recovered MAGs in this study (Supplementary Figure 4, Supplementary Data 4). The majority of BGCs detected in our study encoded for RiPP (332 BGCs), terpene (191 BGCs), non-ribosomal peptide (NRPS: 113 BGCs), and polyketide synthases (112 BGCs; types I, II, and III). There were additionally 130 hybrid clusters composed of overlapping BGCs (e.g., non-ribosomal peptide-polyketide combinations), as well as four ectoine clusters (Supplementary Data 4). Sixty-five percent of the BGCs had a boundary on a contig edge, indicating potentially incomplete recovery

of the whole sequence for nearly two-thirds of our predicted BGCs. BiG-SCAPE (Navarro-Muñoz et al., 2020) analysis revealed that the majority of detected BGCs longer than 10kb did not cluster together with the other 1,154 biosynthetic clusters. BiG-SCAPE created 15 gene cluster families (GCFs) of size 2, while 1,139 clusters were placed into singleton GCFs. Antibiotic Resistance Target Seeker (Alanjary et al., 2017; Mungan et al., 2020) identified putative antibiotic resistance genes within some BGCs.

Distribution and expression of secondary metabolite biosynthetic gene clusters (BGCs) across recovered MAGs

We detected BGCs in MAGs recovered from most of the recovered phyla; exceptions in BGC detection were the bacterial phyla Spirochaetota, Ratteibacteria, Dadabacteria, Dependentiae and Aerophobota that were underrepresented (1-2 MAGs per phylum), and the archaeal phylum Iainarchaeota where 10 genomes were reconstructed. Lack of detection of BGCs in the aforementioned phyla can be attributed to the small sample size analyzed.

MAGs from phyla Myxococcota, Verrucomicrobiota, and Acidobacteriota were more prevalent within the PA fraction and contained some of the largest diversity of biosynthetic gene clusters among phyla with an apparent PA preference (Figure 1a). This likely reflects adaptations to a particle-associated lifestyle where intercellular communication and competition are relatively intense compared to the free-living state, as revealed in culture studies (Long & Azam, 2001; Waters et al., 2010). Genomes from these phyla contained several classes of BGCs including RiPPs, NRPS, as well as polyketide and terpene synthases (Figure 1a).

The recovered Cariaco MAGs expressed BGCs in all the PA and FL metatranscriptomes (Figure 2). BGCs in the PA fraction were expressed predominantly from MAGs affiliated with the Omnitrophota, Desulfobacterota, Planctomycetota, Myxococcota, and Gammaproteobacteria. Planctomycetota genomes have recently been reported to contain diverse biosynthetic potential (Graça et al., 2016), while the Gammaproteobacteria and Myxococcota are prolific producers of secondary metabolites (Murphy et al., 2021; Scherlach & Hertweck, 2021). The majority of expressed BGCs in the PA fraction encoded terpenes, RiPPs, and non-ribosomal peptides. Notably, 2,980 biosynthetic genes were only expressed in PA metatranscriptomes, while 1,901 were exclusively expressed in FL. Additionally, the PA fraction showed higher differential expression of 21 ($P < 0.05$; False Discovery Rate (FDR) = 5%) BGC genes across all sampled depths, while

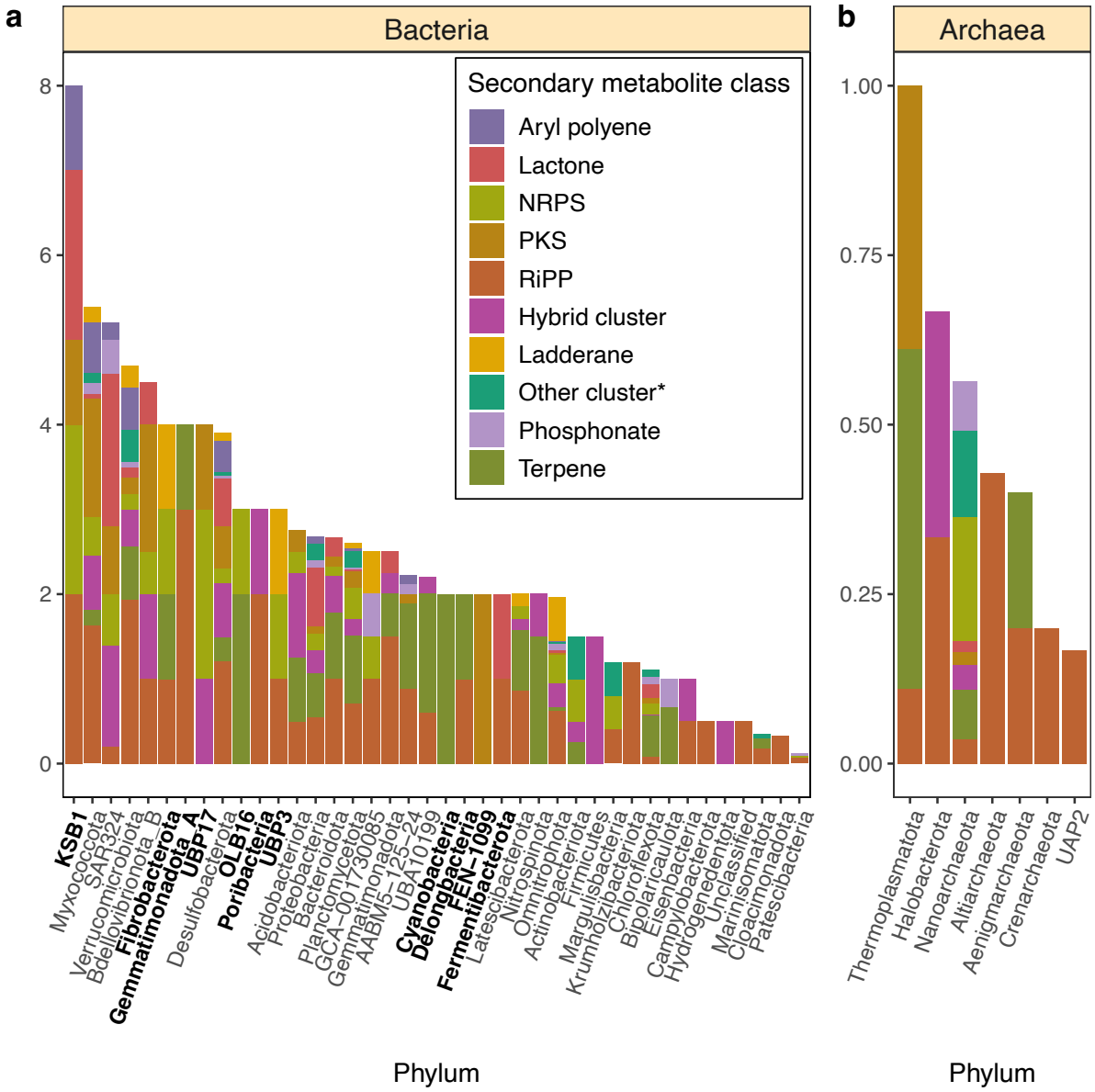


Figure 2-1. Normalized biosynthetic gene cluster count per phylum. a Normalized bacterial biosynthetic gene cluster count by phylum. b Normalized archaeal biosynthetic gene cluster count by phylum. Bold labels denote underrepresented phyla (phyla with only one representative MAG). BGC counts were normalized by dividing the

138 transcripts were significantly more abundant in the FL samples. MAGs more abundant in PA metagenomes ($P < 0.05$; FDR = 5%) exhibited expression of BGC genes almost exclusively from PA metatranscriptomes with little evidence of expression in FL samples (Supplementary Figure 5a-c). This suggests MAGs with an apparent PA preference primarily expressed biosynthetic gene clusters while associated with particles.

Despite the generally small size of free-living marine prokaryote genomes, diverse sets of

Figure 2-2. Expression of secondary metabolite biosynthetic transcripts from individual MAGs in metatranscriptomic samples. Each row represents a biosynthetic transcript, each column represents a sample from the PA fraction (left) and FL fraction (right), and the color represents the log-normalized transcripts per million.

BGCs have been previously reported from free-living marine prokaryotes (Pachiadaki et al., 2019).

The production and potential release of secondary metabolites by free-living prokaryotes has received little consideration so far. This may be primarily due to the perception that bioactive secondary metabolites would be ineffective in dilute planktonic environments, and thus, not confer the selective advantages experienced by microbes associated with highly structured microhabitats (e.g., particles, sediments, soils). Yet, analysis of the FL metatranscriptomes in this study revealed expression of BGC transcripts associated primarily with *Alphaproteobacteria*, Omnitrophota, SAR324, and Desulfobacterota MAGs. The transcripts were predominantly from terpene, non-ribosomal peptide, and lactone clusters with inferred antibiotic activity, as well as roles in oxidative stress and cell-to-cell signaling. Notably, MAGs with an apparent FL preference ($P < 0.05$; FDR = 5%) expressed BGC genes in both the FL and PA fractions with a high degree of overlap (Supplementary Figure 5a-c). We postulate some of the free-living cells, and thus their transcripts, could have been captured by the 2.7 μm PA filters during sampling. It is also possible some MAGs more abundant in the FL fraction dissociated from particles during sample processing. Microorganisms likely also attach to, and disassociate from, particles as they sink through the water column. Some cells that associate with particles in the surface ocean may remain trapped within particles as they sink into realms that no longer favor their survival in the FL state. These hypotheses require further investigation, as do other possible roles these secondary metabolites might play in the free-living state, including grazer avoidance. Bacterial MAGs with high levels of BGC expression in both sample types were affiliated to Desulfobacterota, Omnitrophota, and Planctomycetota. The presence of MAGs from these phyla in both FL and PA, and the robust expression of BGC transcripts in both fractions may indicate these taxa interact intermittently with particles.

Archaea have only recently gained attention for their potential to produce secondary metabolites (Charlesworth & Burns, 2015; Scherlach & Hertweck, 2021; S. Wang & Lu, 2017). Fifty-eight BGCs affiliated to Nanoarchaeota, Thermoplasmatota, Aenigmarchaeota, Altiarchaeota, Halobacterota, Crenarchaetoa, and Undinarchaeota (previously UAP2) MAGs were detected and encoded primarily terpene and polyketide synthases, as well as RiPPs and NRPSs (Figure 1b). Thirty-one clusters were identified within Nanoarchaota MAGs, a group reported previously to possess biosynthetic genes for molecules with putative antibiotic properties (Castelle et al., 2015). Metatrancriptomics revealed NRPS and terpene cluster expression from FL oxycline samples from 6 FL-abundant Nanoarchaeota MAGs. Terpene and polyketide synthase transcripts

from 4 Thermoplasmatota MAGs with no fraction preference were also expressed at oxycline, shallow anoxic, and euxinic depths in PA and FL samples. Further studies of these clades may provide a better ecological understanding of these archaea and new natural product discoveries.

Distribution and expression of BGCs across gradients

We observed differences in BGC abundance and expression across size fraction in both the metagenomic and metatranscriptomic samples (Figure 3a, 3b). Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) analysis of metagenomic and metatranscriptomic read recruitment to biosynthetic clusters primarily separated PA from FL sample types in most datasets (Figure 3a, 3b). For the same size fraction and redox regime, UMAP analysis further separated most datasets between the two sampling points (May vs. November), particularly for BGC expression in oxycline and euxinic water features.

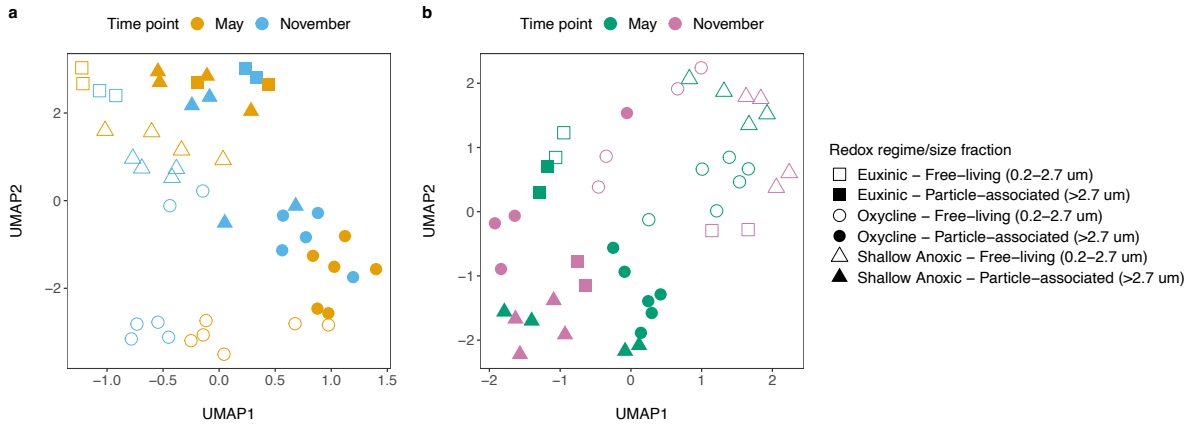


Figure 2-3. Uniform manifold approximation and projection (UMAP) analysis of metagenomic and metatranscriptomic reads recruited to BGCs. UMAP analysis for read mapping data of particle-associated and free-living metagenomes (**a**) and metatranscriptomes (**b**) to all BGCs longer than 10 kb total length detected in Cariaco MAGs. Each point represents the BGC expression profile in a sample, with redox regimes denoted by different shapes. The two size fractions are represented by filled-in and hollow shapes, and sampling time points are colored differently.

Ladderane biosynthetic cluster detection

We detected ladderane BGCs in bacterial MAGs affiliated with Desulfobacterota, Fibrobacterota, Myxococcota, Verrucomicrobiota, Planctomycetota, Latescibacterota, Omnitrophota, GCA-001730085, and UBP3. Ladderane lipids are strictly associated with bacterial genera within the Planctomycetota phylum that perform anammox, but the pathway of ladderane biosynthesis and associated enzymes is unknown. BGCs that resemble ladderane clusters have been reported for non-Planctomycetota genomes (Rattray et al., 2009), but an association between those and the

presence of ladderane lipids was not made. Assessment of contigs containing ladderane BGCs by GUNC (Orakov et al., 2021) could not identify any contaminated or chimeric contigs. Clusters annotated as ladderanes were expressed by all the phyla to which they were attributed (Supplementary Figure 6). The two Planctomycetota MAGs that expressed ladderane clusters were differentially abundant in the FL fraction and were from the anammoxer genus *Scalindua*. The highest expression was observed at shallow anoxic depths (Supplementary Figure 6). We conclude that the *Scalindua* ladderane clusters were accurately annotated, based on prior knowledge of anammoxers lipids and our expression profiles. Clusters of remaining MAGs encoding ladderanes may serve unknown functions in Cariaco Basin. Plausible *in silico* explanations for ladderanes in non-anammox taxa include possible involvement in fatty acid biosynthesis (Rattray et al., 2009) and in lineage divergence of closely related taxa via acquisition of ladderane genes (Choudoir et al., 2018). These could apply to the Cariaco Basin but needs to be validated experimentally.

Oxidative stress genes in biosynthetic gene clusters

We annotated genes within 118 BGCs (primarily RiPPs, terpenes, NRPSs and lactones) encoding for proteins that detoxify, promote biofilm formation (Y. Li & Rebuffat, 2020), or counter damage from free radicals. These BGCs were primarily associated with *Alphaproteobacteria*, Desulfobacterota, Omnitrophota, Planctomycetota, and Myxococcota MAGs. Oxidative stress-related genes from these clusters were functionally annotated mostly as alkyl hydroperoxide reductase subunit C (*ahpC*), glutathione S-transferase (*gst*), and nickel superoxide dismutase. It is possible these enzymes assist in the intracellular regulation of the free radicals' concentrations, albeit previous studies found AhpC and GST to contribute directly to secondary metabolite biosynthesis (Davis et al., 2011; Ma & Payne, 2012). In FL metatranscriptomes, transcripts associated with oxidative stress from lactone, phosphonate, and terpene clusters were primarily expressed by FL-abundant *Scalindua* Planctomycetota, Chloroflexota, and SAR324 MAGs from oxycline and shallow anoxic depths, habitats that exhibit oxygen fluctuations.

To further identify redox-related compounds in the Cariaco BGCs, we compared them to the MIBiG (Kautsar et al., 2020) database which contains community-curated clusters with known functions. Cellular level redox-cycling antibiotics can infiltrate and impose oxidative stress on target cells (Orakov et al., 2021). As an example, four of the BGCs contained genes encoding phenazine or phenazine-like biosynthesis proteins. Phenazines are redox-active compounds known

to contribute to formation of bacterial biofilms and to cause debilitating oxidative stress in targeted cells by forming intracellular free radicals of both reactive oxygen and nitrogen species (Laursen & Nielsen, 2004; Y. Wang et al., 2011). Nevertheless, expression of phenazines could increase microbial fitness in Cariaco Basin by enhancing phosphorus cycling. Within the redoxcline of the Cariaco Basin exists a challenging variability in phosphate concentrations whose fate (precipitation vs. remobilization) is controlled by the delivery of iron and manganese in the water column (McParland et al., 2015). Phenazines are phosphorus/iron-regulated antibiotics suggested to promote microbial growth under phosphorus starvation via solubilization of phosphates through reduction of iron oxides (McRose & Newman, 2021). Expression of genes associated with redox-cycling antibiotics was found primarily in FL metatranscriptomes at all water layers.

Genomes encoding clusters with antibiotic properties often contain genes coding for proteins within the same cluster that prevents self-toxicity (Cundliffe, 1989). We applied Antibiotic Resistant Target Seeker (ARTS; Alanjary et al., 2017; Mungan et al., 2020) to detect antibiotic resistance genes in our BGCs. We detected only 4 types of proteins/protein domains involved in resistance (Supplementary Data 5). These include 13 MAGs that had ABC and RND efflux pumps, RND-type membrane proteins of the efflux complex MexW/MexI/MexH (Webber & Piddock, 2003), and 8 MAGs that contained pentapeptide repeats (Vetting et al., 2006). These were all associated with BGCs that coded for terpenes, bacteriocins, T1PKS/T3PKS, homoserine lactone (hserlactone), NRPS/NRPS-like, betalactones, arylpolyenes and hgIE-KS.

Core biosynthetic genes and tailoring enzymes in Cariaco biosynthetic gene clusters

Analysis of antiSMASH results revealed the presence of core biosynthetic genes that are highly conserved and essential to secondary metabolite biosynthesis. The most frequently detected core biosynthetic gene encoded β-ketoacyl synthase, an essential enzyme in fatty acid biosynthesis (Kauppinen et al., 1988; Figure 4a). We also detected pyrroloquinoline quinone subunit D-like (PqqD-like) synthase, which is essential to the biosynthesis of RiPP-recognition element-dependent RiPPs (Kloosterman et al., 2020; Figure 4a). In the PA fraction, there were 235 differentially expressed or uniquely expressed PqqD-like and beta-ketoacyl synthase transcripts (Figure 4b), while only 94 were detected in the FL fraction (Figure 4b).

Various tailoring enzymes identified in the recovered Cariaco BGCs suggest the implementation of diverse chemical transformations and post-translational modifications. Oxygen

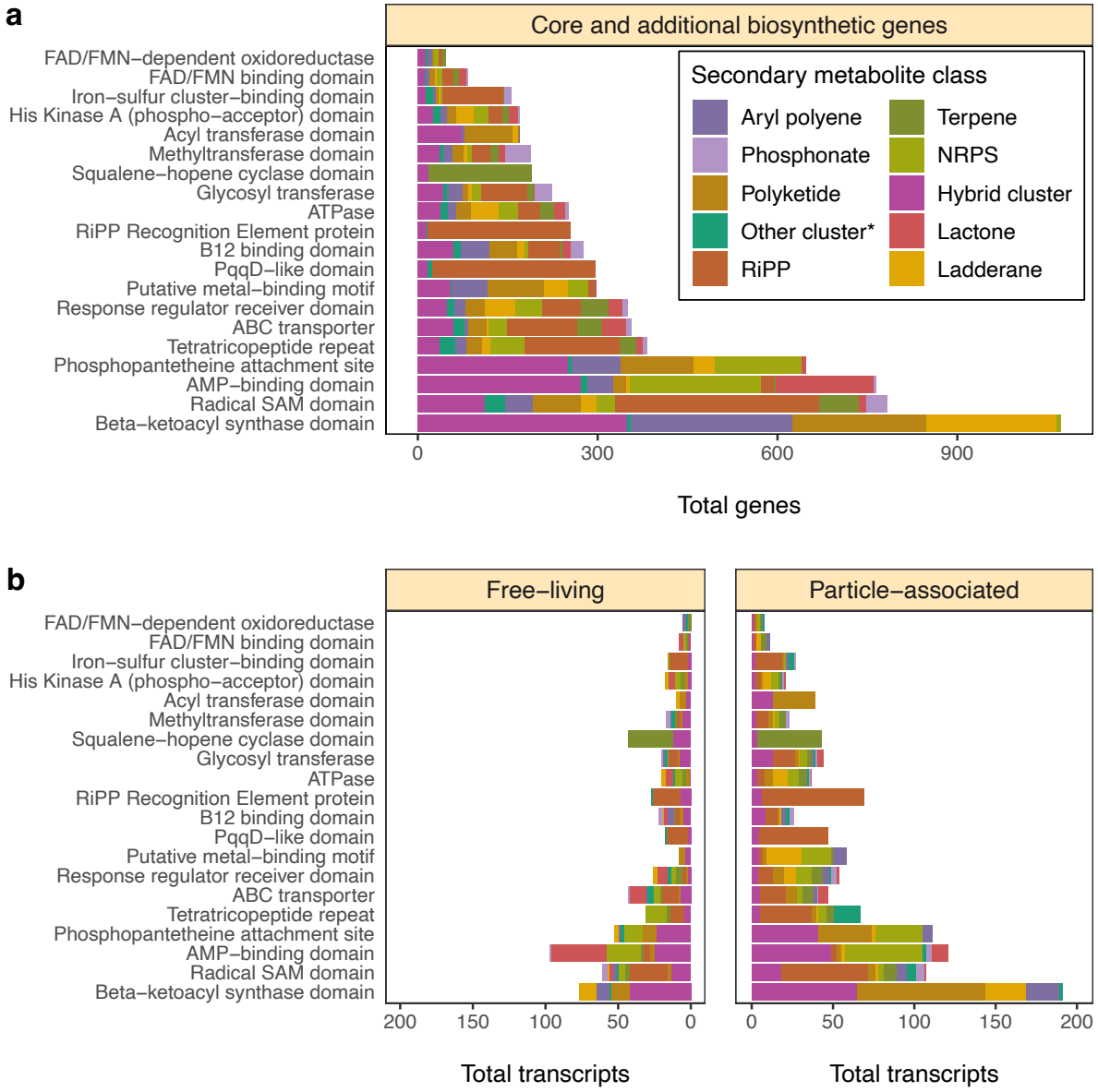


Figure 2-4. Distribution of core and additional biosynthetic genes or domains and transcripts from biosynthetic gene clusters. a Distribution of the most frequently detected core/additional biosynthetic genes, genes encoding tailoring enzymes, and biosynthetically important protein domains in clusters ≥10 kb in length. The "Core and additional biosynthetic genes" strip title of a refers to genes or domains. **b** Core/additional biosynthetic transcripts, as well as transcripts encoding tailoring enzymes and biosynthetically important protein domains differentially expressed or solely expressed in the free-living (left-hand panel) and particle-associated (right-hand panel) metatranscriptomes; colored by BGC product class.

availability along the oxycline in Cariaco Basin could impact the distribution/number of BGCs and tailoring enzymes utilizing molecular oxygen. We searched the BGCs for tailoring enzymes, including Rieske non-heme iron oxygenases (ROs). These enzymes contain oxygen-sensitive [2Fe-2S] clusters and are involved in synthesis of bioactive natural products (Barry & Challis, 2013). Overall, we detected 8 types of ROs in 32 MAGs encoding BGCs for terpenes, betalactones,

T1PKS/T3PKS, phosphonates, RiPPs (lasso/thio/ranthipeptides, linear azole/azoline-containing peptides), NRPS (cyclodipeptides) and RiPP- and NRPS-like clusters. These ROs/ROs-domains were annotated to dioxygenases associated with degradation of aromatic amino acids (tyrosine/tryptophan), phosphonate and sulfur (taurine) cycling, pigment biosynthesis (carotenoids/betalain), and glyoxalase/bleomycin/validamycin dioxygenase superfamilies. This suggests that the identified ROs/ROs can be directly (e.g., synthesis of pigments, antibiotics) or indirectly (via nutrient/amino acid cycling) involved in the synthesis of these secondary metabolites.

Other tailoring enzymes in BGCs included radical SAM proteins, glycosyl transferases, and flavoenzymes (Figure 4a). The most abundant of these were radical SAM proteins, which are known for imparting diverse post-translational modifications on RiPPs (Benjdia et al., 2017). Post-translational glycosylation of secondary metabolites by glycosyl transferases can have a variety of effects, such as toxicity reduction for the producer of the metabolite (Pandey et al., 2018). Flavoenzymes help tailor structurally diverse secondary metabolites through various redox reactions, including single-electron transfers (Argueta et al., 2015). A total of 169 transcripts encoding the above tailoring enzymes were differentially expressed or uniquely expressed in the PA fraction, compared to 94 transcripts in the FL samples (Figure 4b).

Genes encoding specialized domains involved in peptide biosynthesis were also detected in the Cariaco biosynthetic clusters (Figure 4a). A prevalence of B_{12}-binding domains identified in biosynthetic genes raises the possibility of B_{12}-dependent methylation during synthesis or post-translational modification of biosynthesized peptides (Benjdia et al., 2017; Jarrett, 2019). Phosphopantetheine attachment site domains were also ubiquitous in Cariaco BGCs. Phosphopantetheine prosthetic groups from acyl carrier proteins are transferred by acyl transferases (Byers & Gong, 2007), both of which were numerous in Cariaco clusters. Tetratricopeptide repeats were prevalent as well, which can mediate protein-protein interactions in diverse cell processes (D'Andrea & Regan, 2003) by binding many distinct types of ligands (Ganesh et al., 2015). The presence of various biosynthetically important protein domains present in the recovered BGCs suggests a variety of diverse chemical transformations and post-translational modifications that could shape the secondary metabolites synthesized by the particle-associated and free-living microbes identified in the water column of Cariaco Basin.

Discussion

We performed genome mining to detect and classify BGCs across a diverse set of bacterial and archaea phyla recovered from the anoxic depths of the Cariaco Basin. Although the increasing number of available genomes and bioinformatic approaches have revolutionized the discovery of secondary metabolites (Gavriilidou et al., 2022), a key issue that remains is linking the detected clusters to biological activity. Previous studies of marine Planctomycetota showed that aqueous and organic extracts of isolates that contain bioinformatically-predicted BGCs exhibited antimicrobial and antifungal activity (Graça et al., 2016). However, the vast majority of microorganisms, particularly those from challenging environments like oxygen-depleted systems, escape cultivation, thus hindering our ability to use similar approaches to explore the bioactive potential of predicted BGCs. Recently, an analysis of >1000 publicly available marine metagenomes revealed ~40,000 putative BGCs (Paoli et al., 2022). Nonetheless, this analysis did not include samples from sulfidic waters. For this environmental survey of BGCs, we used mapped metatranscriptomes collected and preserved *in situ* to our BGCs to unveil the expression profiles of detected biosynthetic clusters, and to investigate the potential role of redox conditions and particles in the observed patterns.

Particles provide colonizable, nutrient-rich substrates where metabolites can be concentrated and exchanged and can provide protection for oxygen- or sulfide-sensitive microbiota. Previous work shows particle-associated microbial assemblages from the Eastern Tropical North Pacific oxygen minimum zone possess genes coding for antibiotic resistance, motility, cell-to-cell transfer, and signal recognition (Ganesh et al., 2014, 2015), and microorganisms are able to proliferate in particles in suboxic to anoxic zones where reducing conditions can persist for extended periods of time (Alldredge & Cohen, 1987; Fuchsman et al., 2011). Consistent with previous culture-based studies, BGCs may allow PA taxa to compete for precious resources, prevent growth of other potential particle colonizers, and aid survival in oxygen-depleted conditions (Long & Azam, 2001). Our study revealed enhanced expression of BGCs by members of Myxococcota, Desulfobacterota, Omnitrophota, Planctomycetota and *Gammaproteobacteria* within the PA fractions.

Analysis with UMAP of biosynthetic cluster abundances and expression profiles revealed a marked separation between the PA and FL size fractions in both the metagenomic and metatranscriptomic data. The niche preferences of taxa behind the MAGs we recovered, as well

as the two different sampling times likely play a role in the observed differences in expression profiles of biosynthetic clusters in our PA vs. FL samples. We detected differences between sampling season and redox regime within the metagenomes. In the metagenomic samples, the PA euxinic and deep anoxic samples, as well as the FL euxinic samples clustered together (Figure 3a). The abundance of metagenome reads mapped to MAGs across size fractions was similar at depths where oxygen is very limited or absent, and contributed to the clustering of BGC read abundances within these samples. The FL shallow anoxic and oxycline, and the PA oxycline samples formed three distinct clusters, suggesting differing redox conditions shaped BGC composition and abundance in these samples. Within oxycline samples, the influence of oxygen and the separation between PA and FL size fractions is evident.

UMAP analysis of the metatranscriptomic data revealed BGC expression profiles that differentiated primarily by size fraction as well as season of sampling (Figure 3b). Some overlap is observed between BGCs expressed in the PA and FL fractions, consistent with the idea that some taxa may transiently associate with particles as they sink through the water column. Paoli et al., (Paoli et al., 2022) examined MAGs recovered from PA and FL fractions in global datasets (that did not include sulfidic end-members) and they found genes for terpenes and RiPPs enriched in the FL fraction, and NRPS and PKS genes enriched in PA samples. This supports the idea that taxa and the genes they carry are shaped by their FL vs. PA lifestyle (niche requirements). Seasonal differences in primary productivity can also shape microbial communities and the genes they express. In Cariaco Basin, upwelling of nutrients occurs between January and March, fueling increased primary productivity (Scranton et al., 2006). This may be a contributing factor to the observed separation of most PA vs. FL BGC profiles (Figure 3a, b) because in the FL state microorganisms will experience environmental shifts more directly than those protected within particles.

While little is known about secondary metabolite expression in free-living marine prokaryotes, biosynthetic potential is known to be widespread in their genomes (Pachiadaki et al., 2019). We detected expression of BGCs in the FL metatranscriptomes, predominantly from *Alphaproteobacteria*, Omnitrophota, SAR324, and Desulfobacterota MAGs. The overlap in BGC expression detected in the PA and FL transcriptomes mapping onto preferentially FL-associated MAGs was unique, as we did not observe the same phenomenon in expression pattern of preferentially PA-associated MAGs. This supports our hypothesis that the PA metatranscriptomes

captured some of the BGC expression signal from the FL samples. While it is less clear how free-living microbes could benefit from release of secondary metabolites, we conclude that interactions with particles alone cannot account for all the expression of biosynthetic transcripts in the FL samples. Some of these compounds (such as the ladderanes), likely only serve intracellular roles within free-living prokaryotes. It is also plausible that higher expression of BGCs in PA samples reflects more commonplace release of secondary metabolites within particles than within free-living ODWC microbial populations. The role of secondary metabolites in microbial fitness is an open debate because possession of secondary metabolism can enhance the overall fitness, but not all products of secondary metabolism will necessarily have an effect on the producer (Firn & Jones, 2000). Nonetheless, secondary metabolites are reported to affect niche utilization, shape microbial community assembly, and act as a functional trait driving ecological diversification among closely-related bacteria inhabiting the same microenvironments (Junkins et al., 2022; Penn et al., 2009). Likewise, the example of phenazines and phosphorus acquisition can be a paradigm of dual/pleiotropic functions of secondary metabolites where they can serve as potential antibiotics and regulators of nutrient cycling. Co-localization of antibiotic resistance genes within biosynthetic clusters has been previously observed (Thaker et al., 2013). Bacteria may have evolved pleiotropic switching capabilities that allow simultaneous expression of secondary metabolites with other co-localized genes in a cluster (e.g., antibiotic regulation and resistance) as a survival strategy under unfavorable conditions, and as a self-protection mechanism (Cundliffe, 1989). Co-localized genes encoding antibiotic resistance were present in the BGCs we identified. In addition to antibiotic resistance genes, clusters contained a diverse array of core biosynthetic synthases, tailoring enzymes, and significant protein domains involved in secondary metabolite synthesis and post-translational modification.

In summary, our investigation of BGCs in metagenomic and metatranscriptomic datasets from an oxygen-depleted marine water column provides considerable evidence for secondary metabolite synthesis over a wide taxonomic distribution from 44 bacterial and 8 archaeal phyla. More BGCs were expressed (particularly coding for non-ribosomal peptides, polyketides, RiPPs, and terpenes) by taxa whose MAGs were particle-associated than in the free-living fraction. The BGCs identified here hint at a complex network of ecological interactions coupled to a competitive, yet communicative lifestyle mediated by chemical and toxin production not only within PA microbes, but to a lesser extent, within the FL communities. These findings open the

door for future laboratory characterization of genes for novel bioactive metabolites with potential ecological and pharmaceutical importance.

Methods

Sample collection

Water samples for metagenomic analyses were collected from 6 depths during two cruises in May 2014 and in November 2014 using Niskin bottles, as described in detail in Suter et al., 2017 (Supplementary Data 6). Specifically, 8-10 L water samples for metagenomic analysis were gravity-filtered sequentially through EMD Millipore 2.7 µm glass fiber membranes 47mm diameter (PA fraction), and then through 0.2 µm Sterivex filters (FL fraction) and stored frozen at -20 °C in the field and then -80 °C in the laboratory until extraction. Water samples were also collected and preserved *in situ* for isolation of RNA and construction of metatranscriptome libraries from depths selected to capture anoxic and sulfidic water layers. RNA sample collections were conducted with a "Microbial Sampler – Submersible Incubation Device" (MS-SID; Pachiadaki et al., 2016; C. D. Taylor & Doherty, 1990). Water (2 L) was sequentially filtered through EMD Millipore 2.7 µm glass fiber filters and then through 0.2 µm Millipore Express polysulfone membranes at depth. The filters were preserved immediately *in situ* with RNAlater®. Upon MS-SID retrieval, preserved filters were transferred to cryovials with additional RNAlater and stored frozen at -20 °C in the field and then -80 °C in the laboratory until extraction. Biogeochemical data collected in support of molecular samples are available at https://www.bco-dmo.org/dataset/652313/data.

DNA extractions and sequencing

DNA was extracted from all samples according to Frias-Lopez et al., (Frias-Lopez et al., 2008) and Ganesh et al., Ganesh et al., 2014) and described in detail in Suter et al., 2017 (Suter et al., 2017). DNA was extracted from all samples according to Frias-Lopez et al., (Frias-Lopez et al., 2008) and Ganesh et al., (Ganesh et al., 2014) and described in detail in Suter et al., (Suter et al., 2017). Briefly, lysozyme solution (2 mg in 40 µL of lysis buffer) was added directly to the tube containing the 2.7 µm membrane filter or to the Sterivex cartridge, and was incubated for 45 min at 37 °C. Subsequently, Proteinase K solution (1 mg in 100 µl lysis buffer, with 100 µl 20% SDS) was added, and then incubated for 2 h at 55 °C. The lysate was transferred to a clean tube and nucleic acids were extracted once with phenol:chloroform:isoamyl alcohol (25:24:1) and once with

chloroform:isoamyl alcohol (24:1). The aqueous phase was concentrated using Amicon Ultra-4w/100 kDa MWCO centrifugal filters (Millipore). After extraction, DNA was purified with the Genomic DNA Clean and Concentrator-25 kit (Zymo Research), eluted into 10 mM Tris-HCl, and frozen until downstream analysis. Aliquots of the extracted DNA were sent to the Georgia Genomics for library preparation and paired-end 2x150 bp Illumina NextSeq sequencing. The R1 and R2 reads were filtered using Trimmomatic 0.39 (Bolger et al., 2014). Trimmomatic performs a "sliding window" trimming removing sequence data when the average quality within the window (eight nucleotides used here) drops below a threshold (set to 12). The length of the trimmed sequences was set to a minimum of 50 nucleotides.

RNA extraction and sequencing

RNA was extracted using a modification of the mirVana miRNA Isolation kit (Ambion, Life Technologies, Carlsbad, CA, USA) as in Stewart et al., (Stewart et al., 2012). Briefly, filters were thawed on ice and the RNA stabilizing buffer (RNAlater) was removed from the cryovials. Cells on filters were lysed by adding lysis buffer and miRNA homogenate additive (Ambion) into the cryovial or cartridge. After vortexing and incubation on ice, lysates were transferred to RNAase-free tubes and processed using an acid–phenol/chloroform extraction following the manufacturer's suggestions. We used TURBO DNA-free kit (Ambion, Foster City, CA, USA) to remove carryover DNA and we purified the extracts using the RNeasy MinElute Cleanup Kit (Qiagen, Hilden, Germany). Removal of DNA was confirmed by PCR using the forward primer (5′-AYTGGGYDTAAAGNG-3′) and a mix of reverse primers (5′-GCCTTGCCAGCCCGCTCAG, TACCRGGGTHTCTAATCC, TACCAGAGTATCTAATTC, CTACDSRGGTMTCTAATC and TACNVGGGTATCTAATCC-3′ in a 6:1:2:12 ratio, respectively), designed to cover most hypervariable regions of bacterial 16S rRNA (Cole et al., 2009; Conroy et al., 2009). cDNA libraries were prepared using the ScriptSeq RNA-Seq Library Preparation Kit (Illumina). Excess nucleotides and PCR primers were removed from the library using the Agencourt AMPure™ XP (Beckman-Coulter) kit. The Illumina NextSeq platform was used for paired-end 2x150 sequencing at the Georgia Genomic Facility. The R1 and R2 reads were filtered using Trimmomatic.

MAG co-assembly, binning, and taxonomic assignment

Metagenomes originating from adjacent regions (such as geographic regions or in this case, adjacent depths targeted in this study) are likely to overlap in the sequence space, increasing the mean coverage and extent of reconstruction of MAGs when using a co-assembly approach. In order to reconstruct MAGs, the trimmed reads of metagenomic datasets from the anoxic to sulfidic depths (314m and 900m in May, 267m and 900m in November) and both PA and FL fractions were co-assembled into contigs using SPAdes 3.11.1 (Bankevich et al., 2012) with default values and flag "--meta". Assembled contigs were binned using MetaBAT 2.12.1 (Kang et al., 2015) with default values. CheckM 1.0.1161 (Parks et al., 2015) was used to estimate the completeness and contamination of the reconstructed genomes. Only MAGs with ≥75% complete and ≤5% contamination were used for the downstream analysis (Supplementary Data 1). The taxonomic placement of the MAGs was performed with GTDB-Tk 2.1.1 (Chaumeil et al., 2020). The taxonomic identification of the recovered MAGs revealed the presence of three known lab contaminants, including *Burkholderia* contaminans (Giovannoni et al., 1990) that was reconstructed in the first marine genomic study.

Removal of redundant MAGs

To collapse any redundancy, a workflow using Anvi'o 4 (Eren et al., 2015) was implemented as described in Delmont et al., (Delmont et al., 2018). Scaffolds from all MAGs were concatenated into a single FASTA file for mapping and processing. Anvi'o commands 'anvi-gen-contigs-database' and 'anvi-run-hmms' were run with default parameters to create a MAG contig database, and scan MAGs with HMMs, respectively. Contigs were then used to recruit short reads from all metagenomic samples using the Bowtie2 2.5.0 (Langmead & Salzberg, 2012) commands 'bowtie2-build' and 'bowtie2' with default parameters, and SAMtools 1.16.1 (H. Li et al., 2009) commands 'view', 'sort', and 'index' to convert resulting Sequence Alignment/Map format (SAM) files into Binary Alignment/Map (BAM) format, as well as sort and index the converted BAM files. The Anvi'o command 'anvi-merge' was implemented to create a merged profile database of all MAGs, describing the distribution and detection statistics of scaffolds in MAGs across all the metagenomic samples. The Average Nucleotide Identity (ANI) and Pearson correlation values were then calculated to identify MAGs with high sequence similarity and those that were distributed similarly across metagenomes. Pairwise Pearson correlations of the MAGs were

calculated using the 'cor' function in R (Team, 2013), and ANI values were calculated for MAGs using NUCmer (from the MUMmer [Marçais et al., 2018] 3.23 package) with default settings, grouped by their taxonomy, such that ANI was not computed for pairs of MAGs which did not belong to the same phylum. Anvi'o scripts were then used to classify MAGs as redundant if (1) their ANI was at least a 98% match (with a minimum alignment of 75% of the smaller genome in each comparison), and (2) the Pearson coefficient for their distribution across datasets was > 0.9.

Calculation of MAG relative abundances

Reads from 48 metagenomic samples (2 sample types: PA- and FL-fraction, 2 replicates per fraction from 6 depths, taken over 2 sampling time points = 48 metagenomes; Supplementary Table 5) were mapped to each MAG using the BWA (H. Li, 2013) 2.0 aligner via the CoverM 0.6.1 (Woodcroft, B.; https://github.com/wwood/CoverM) command line tool. The CoverM tool automatically concatenated all the MAGs into a single file, and metagenome reads were recruited to MAG contigs, setting the parameter --min-read-percent-identity to 95 and --min-read-aligned-percent to 50. The "Relative Abundance" CoverM method on the "genome" setting was used to calculate relative abundances of the 565 MAGs in each of the metagenomic samples. Each relative abundance was calculated as the percentage of total reads from the sample that uniquely mapped to a MAG (with consideration for the percentage of unmapped reads in the sample). The relative abundance values were log-transformed using the log1p formula: $y = \log(x + 1)$ and were used for heatmap plotting.

Assessing BGC-containing contigs for chimerism and contamination using GUNC

The GUNC (Orakov et al., 2021) command line tool was used with default settings to assess contigs from Cariaco MAGs for contamination that contained ladderane BGCs predicted by AntiSMASH (Blin et al., 2021). The flat file outputs from running GUNC 1.0.5 on all contigs containing BGCs were then manually assessed for potential chimerism and/or contamination.

Differential abundance analysis

Differential abundance (PA- or FL-abundant) was determined for individual MAGs exhibiting differences in abundance between sample type (PA, FL) and water layer (oxycline, shallow anoxic, euxinic). Reads were recruited to a concatenated FASTA file containing whole genome contigs using CoverM (Ben Woodcroft, n.d.). A counts matrix was created with rows containing individual

MAG read mapping counts and with metagenomic samples as columns. Count data for all MAGs was analyzed to calculate DESeq2 1.34.0 size factors for cross-sample count normalization. The differential abundance of MAGs between fraction sizes and water layers was modeled using the DESeq2 negative binomial model with the metadata variables of fraction size and water layer in which "count" was the dependent variable and "fraction" as well as "water layer" were independent variables. The significant differential abundances of MAGs (with an FDR-corrected $P < 0.05$) identified by comparing the PA and FL samples were grouped by water layer, and direct comparisons were made between normalized counts of significantly differentially abundant MAGs from the oxycline, shallow anoxic, and euxinic depths.

Similarity clustering of BGCs using BiG-SCAPE

The redundancy of the predicted biosynthetic cluster sequences recovered from the Cariaco MAGs was assessed using the BiG-SCAPE 1.1.4 (Navarro-Muñoz et al., 2020) command line tool with default parameters. The resulting Gene Cluster Families (GCFs) from this sequence similarity network analysis were visually assessed using BiG-SCAPE's default index.html output file.

Scanning of MAGs for BGCs, functionally annotating genes within clusters, and comparing mined clusters to the MiBIG database BGCs

Genes were predicted using Prodigal (Hyatt et al., 2010) 2.6.3 for all 565 MAGs. The resulting genes of each MAG were individually scanned for BGCs using antiSMASH 6 (Blin et al., 2021) with default parameters. Gene clusters with a total length less than 10kb were discarded from downstream analysis to minimize the inclusion of fragmented BGCs in our data. The genes predicted using Prodigal were scanned using the InterProScan 581(Jones et al., 2014) and Prokka 1.14.6 (Seemann, 2014) command line tools with default parameters for functional annotations, as well as during the implementation of the antiSMASH 6 pipeline with the antiSMASH HMM databases. We manually searched the resulting annotations for genes and domains that encoded a variety of functions, such protein domains involved in post-translational modifications. The results of comparing the mined Cariaco BGCs to the MiBIG database BGCs was scraped using R from the output HTML files from scanning each MAG with antiSMASH.

Detection of antibiotic resistance genes in BGCs using ARTS

The presence of putative antibiotic resistance genes was examined with ARTS version 2 (Alanjary et al., 2017; Mungan et al., 2020) that implements antiSMASH 5.0. The web interface was used. The results in the "Proximity: BGC table with localized hits" were manually inspected. The location and the annotation of potential antibiotic resistance genes that show colocalization with BGS is recorded in Supplementary Table 5.

Metatranscriptomic read mapping of RNA-seq data to BGCs

Metatranscriptomic samples were individually mapped to the concatenated gene FASTA file using the minimap2 2.24-r1122 (H. Li, 2018) sequence alignment algorithm with default parameters. The resulting output files in PAF format were manually filtered of supplementary alignments using a custom R script, and only alignments incorporating at least 50% of the length of a read pair with at least 95% percent identity were retained. The same custom script was utilized to concatenate all individual alignment counts into a single file in a matrix format, with each sample representing a column and each row representing RNA-Seq alignment counts to a gene. The metatranscriptomic counts matrix was normalized to Transcripts Per Million (TPM) and the values were log-transformed using the log1p formula: $y = \log(x + 1)$ and were used for heatmap plotting.

Differential gene expression analysis

To detect differential expression of individual genes within differently expressed biosynthetic clusters between sampling depths, the read counts matrix was modeled in the context of the metadata variable size fraction using a negative binomial model implemented with DESeq2 1.34.0 in R. Count data for all genes from all MAGs was analyzed independently so that the DESeq2 size factors for cross-sample count normalization would reflect the total transcriptomic activity of MAGs in each sample. This approach is robust to biases in total transcriptomic activity per organism between samples, and it is used to identify differences in gene expression independent of changes in taxonomic composition, similar to previously reported methods (Bray et al., 2016; Love et al., 2014). After size factor normalization, read counts were fit to a negative binomial model in which "count" was the dependent variable and "size fraction" was an independent variable. To test whether any genes exhibited differential expression associated with different size fractions, the differential expression results were saved and analyzed. The significant genes (with an FDR-corrected $P < 0.05$) identified by comparing the PA and FL samples and direct

comparisons were made between normalized counts of genes that differed significantly in expression profiles. This method confirmed differential expression of individual genes within each differentially expressed biosynthetic cluster.

UMAP analysis on BGC abundances in metatranscriptomic and metagenomic datasets

The normalized abundances of BGC read mapping data from both metagenomic and metatranscriptomic read recruitment using minimap2 were used as input for UMAP analysis in R using the umap 0.2.9.0 (Konopka & Konopka, 2018) package (https://github.com/tkonopka/umap). The results were plotted using ggplot2 (Wickham et al., 2016) 3.3.6. Clustering of the UMAP embedding was done using the hierarchical density-based spatial clustering (HDBSCAN) function from the dbscan 1.1-11 (Hahsler et al., 2019) package.

Data Availability

The metatranscriptome and metagenome data generated in this study have been deposited in the NCBI database under accession code PRJNA326482 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA326482). The processed metagenome-assembled genomes (in FASTA format) and biosynthetic gene cluster files (in ZIP format) are available at OSF (https://osf.io/usm8r/). The biogeochemistry data from the CARIACO Basin Time Series Station for May to November 2014 are available through the Biological and Chemical Oceanography Data Management Office (BCO-DMO) at the Woods Hole Oceanographic Institution (https://www.bco-dmo.org/dataset/652313/data). The MIBiG 2.0 database is publicly available (https://mibig.secondarymetabolites.org). Source data are provided with this paper.

Code Availability

The scripts used for all bioinformatic pipelines, data processing, and plotting used in this study are available in the following GitHub repository: https://github.com/d-mcgrath/cariaco_basin (Geller-McGrath et al., 2023).

Source Data

Links to all datasets used in this analysis, including source data used to generate the figures in this chapter can be found here: https://www.nature.com/articles/s41467-023-36026-w.

References

Alanjary, M., Kronmiller, B., Adamek, M., Blin, K., Weber, T., Huson, D., Philmus, B., & Ziemert, N. (2017). The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Research*, *45*(W1), W42–W48.

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

Argueta, E. A., Amoh, A. N., Kafle, P., & Schneider, T. L. (2015). Unusual non-enzymatic flavin catalysis enhances understanding of flavoenzymes. *FEBS Letters*, *589*(8), 880–884.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., & Prjibelski, A. D. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477.

Barry, S. M., & Challis, G. L. (2013). Mechanism and catalytic diversity of Rieske non-heme iron-dependent oxygenases. *ACS Catalysis*, *3*(10), 2362–2370.

Ben Woodcroft. (n.d.). *CoverM* (0.6.1).

Benjdia, A., Balty, C., & Berteau, O. (2017). Radical SAM enzymes in the biosynthesis of ribosomally synthesized and post-translationally modified peptides (RiPPs). *Frontiers in Chemistry*, *5*, 87.

Blin, K., Kim, H. U., Medema, M. H., & Weber, T. (2019). Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics*, *20*(4), 1103–1113.

Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*, 1.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527.

Byers, D. M., & Gong, H. (2007). Acyl carrier protein: structure–function relationships in a conserved multifunctional protein family. *Biochemistry and Cell Biology*, *85*(6), 649–662.

Cabello-Yeves, P. J., Callieri, C., Picazo, A., Mehrshad, M., Haro-Moreno, J. M., Roda-Garcia, J. J., Dzhembekova, N., Slabakova, V., Slabakova, N., & Moncheva, S. (2021). The microbiome of the Black Sea water column analyzed by shotgun and genome centric metagenomics. *Environmental Microbiome*, *16*(1), 1–15.

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

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., Frischkorn, K. R., Tringe, S. G., Singh, A., & Markillie, L. M. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current Biology*, *25*(6), 690–701.

Charlesworth, J. C., & Burns, B. P. (2015). Untapped resources: biotechnological potential of peptides and secondary metabolites in archaea. *Archaea*, *2015*.

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

Choudoir, M. J., Pepe-Ranney, C., & Buckley, D. H. (2018). Diversification of secondary metabolite biosynthetic gene clusters coincides with lineage divergence in Streptomyces. *Antibiotics*, *7*(1), 12.

Conroy, J. L., Restrepo, A., Overpeck, J. T., Steinitz-Kannan, M., Cole, J. E., Bush, M. B., & Colinvaux, P. A. (2009). Unprecedented recent warming of surface temperatures in the eastern tropical Pacific Ocean. *Nature Geoscience*, *2*(1), 46–50.

Cragg, G. M., & Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. *Biochimica et Biophysica Acta (BBA)-General Subjects*, *1830*(6), 3670–3695.

Cundliffe, E. (1989). How antibiotic-producing organisms avoid suicide. *Annual Review of Microbiology*, *43*(1), 207–233.

Dalsgaard, T., Thamdrup, B., Farías, L., & Revsbech, N. P. (2012). Anammox and denitrification in the oxygen minimum zone of the eastern South Pacific. *Limnology and Oceanography*, *57*(5), 1331–1346.

D'Andrea, L. D., & Regan, L. (2003). TPR proteins: the versatile helix. *Trends in Biochemical Sciences*, *28*(12), 655–662.

Davis, C., Carberry, S., Schrettl, M., Singh, I., Stephens, J. C., Barry, S. M., Kavanagh, K., Challis, G. L., Brougham, D., & Doyle, S. (2011). The role of glutathione S-transferase GliG in gliotoxin biosynthesis in Aspergillus fumigatus. *Chemistry & Biology*, *18*(4), 542–552.

Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T. M., Rappé, M. S., McLellan, S. L., Lücker, S., & Eren, A. M. (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, *3*(7), 804–813.

Duret, M. T., Lampitt, R. S., & Lam, P. (2019). Prokaryotic niche partitioning between suspended and sinking marine particles. *Environmental Microbiology Reports*, *11*(3), 386–400.

Edgcomb, V. P., Taylor, C., Pachiadaki, M. G., Honjo, S., Engstrom, I., & Yakimov, M. (2016). Comparison of Niskin vs. in situ approaches for analysis of gene expression in deep Mediterranean Sea water samples. *Deep Sea Research Part II: Topical Studies in Oceanography*, *129*, 213–222.

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

Firn, R. D., & Jones, C. G. (2000). The evolution of secondary metabolism–a unifying model. *Molecular Microbiology*, *37*(5), 989–994.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., & DeLong, E. F. (2008). Microbial community gene expression in ocean surface waters. *Proceedings of the National Academy of Sciences*, *105*(10), 3805–3810.

Fuchsman, C. A., Kirkpatrick, J. B., Brazelton, W. J., Murray, J. W., & Staley, J. T. (2011). Metabolic strategies of free-living and aggregate-associated bacterial communities inferred from biologic and chemical profiles in the Black Sea suboxic zone. *FEMS Microbiology Ecology*, *78*(3), 586–603.

Fuchsman, C. A., Staley, J. T., Oakley, B. B., Kirkpatrick, J. B., & Murray, J. W. (2012). Free-living and aggregate-associated Planctomycetes in the Black Sea. *FEMS Microbiology Ecology*, *80*(2), 402–416.

Ganesh, S., Bristow, L. A., Larsen, M., Sarode, N., Thamdrup, B., & Stewart, F. J. (2015). Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. *The ISME Journal*, *9*(12), 2682–2696.

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

Gavriilidou, A., Kautsar, S. A., Zaburannyi, N., Krug, D., Müller, R., Medema, M. H., & Ziemert, N. (2022). Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes. *Nature Microbiology*, *7*(5), 726–735.

Gilly, W. F., Beman, J. M., Litvin, S. Y., & Robison, B. H. (2013). Oceanographic and biological effects of shoaling of the oxygen minimum zone. *Annual Review of Marine Science*, *5*, 393–420.

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, *345*(6270), 60–63.

Graça, A. P., Calisto, R., & Lage, O. M. (2016). Planctomycetes as novel source of bioactive molecules. *Frontiers in Microbiology*, *7*, 1241.

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*, 1–30.

Hibbing, M. E., Fuqua, C., Parsek, M. R., & Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology*, *8*(1), 15–25.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 1–11.

Jarrett, J. T. (2019). Surprise! A hidden B12 cofactor catalyzes a radical methylation. *Journal of Biological Chemistry*, *294*(31), 11726–11727.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., & Nuka, G. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240.

Junkins, E. N., McWhirter, J. B., McCall, L.-I., & Stevenson, B. S. (2022). Environmental structure impacts microbial composition and secondary metabolism. *ISME Communications*, *2*(1), 1–10.

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

Kauppinen, S., Siggaard-Andersen, M., & von Wettstein-Knowles, P. (1988). β-ketoacyl-ACP synthase I of Escherichia coli: Nucleotide sequence of thefabB gene and identification of the cerulenin binding residue. *Carlsberg Research Communications*, *53*(6), 357–370.

Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., Van Der Hooft, J. J. J., Van Santen, J. A., Tracanna, V., Suarez Duran, H. G., & Pascal Andreu, V. (2020). MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research*, *48*(D1), D454–D458.

Kloosterman, A. M., Shelton, K. E., van Wezel, G. P., Medema, M. H., & Mitchell, D. A. (2020). RRE-Finder: a genome-mining tool for class-independent RiPP discovery. *Msystems*, *5*(5), e00267-20.

Konopka, T., & Konopka, M. T. (2018). R-package: umap. *Uniform Manifold Approximation and Projection*.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.
Laursen, J. B., & Nielsen, J. (2004). Phenazine natural products: biosynthesis, synthetic analogues, and biological activity. *Chemical Reviews*, *104*(3), 1663–1686.
Letzel, A.-C., Pidot, S. J., & Hertweck, C. (2013). A genomic approach to the cryptic secondary metabolome of the anaerobic world. *Natural Product Reports*, *30*(3), 392–428.
Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv Preprint ArXiv:1303.3997*.
Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.
Li, J., Gu, L., Bai, S., Wang, J., Su, L., Wei, B., Zhang, L., & Fang, J. (2021). Characterization of particle-associated and free-living bacterial and archaeal communities along the water columns of the South China Sea. *Biogeosciences*, *18*(1), 113–133.
Li, Y., & Rebuffat, S. (2020). The manifold roles of microbial ribosomal peptide–based natural products in physiology and ecology. *Journal of Biological Chemistry*, *295*(1), 34–54.
Long, R. A., & Azam, F. (2001). Antagonistic interactions among marine pelagic bacteria. *Applied and Environmental Microbiology*, *67*(11), 4975–4983.
Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21.
Ma, L., & Payne, S. M. (2012). AhpC is required for optimal production of enterobactin by Escherichia coli. *Journal of Bacteriology*, *194*(24), 6748–6757.
Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, *14*(1), e1005944.
McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint ArXiv:1802.03426*.
McParland, E., Benitez-Nelson, C. R., Taylor, G. T., Thunell, R., Rollings, A., & Lorenzoni, L. (2015). Cycling of suspended particulate phosphorus in the redoxcline of the Cariaco Basin. *Marine Chemistry*, *176*, 64–74.
McRose, D. L., & Newman, D. K. (2021). Redox-active antibiotics enhance phosphorus bioavailability. *Science*, *371*(6533), 1033–1037.
Mestre, M., Borrull, E., Sala, M. M., & Gasol, J. M. (2017). Patterns of bacterial diversity in the marine planktonic particulate matter continuum. *The ISME Journal*, *11*(4), 999–1010.
Mestre, M., Ferrera, I., Borrull, E., Ortega-Retuerta, E., Mbedi, S., Grossart, H., Gasol, J. M., & Sala, M. M. (2017). Spatial variability of marine bacterial and archaeal communities along the particulate matter continuum. *Molecular Ecology*, *26*(24), 6827–6840.
Mungan, M. D., Alanjary, M., Blin, K., Weber, T., Medema, M. H., & Ziemert, N. (2020). ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Research*, *48*(W1), W546–W552.
Murphy, C. L., Yang, R., Decker, T., Cavalliere, C., Andreev, V., Bircher, N., Cornell, J., Dohmen, R., Pratt, C. J., & Grinnell, A. (2021). Genomes of novel Myxococcota reveal severely curtailed machineries for predation and cellular differentiation. *Applied and Environmental Microbiology*, *87*(23), e01706-21.

Naqvi, S. W. A., Bange, H. W., Farías, L., Monteiro, P. M. S., Scranton, M. I., & Zhang, J. (2010). Marine hypoxia/anoxia as a source of CH 4 and N 2 O. *Biogeosciences*, *7*(7), 2159–2190.

Navarro-Muñoz, J. C., Selem-Mojica, N., Mullowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L. C., Yeong, M., Cruz-Morales, P., & Abubucker, S. (2020). A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology*, *16*(1), 60–68.

Orakov, A., Fullam, A., Coelho, L. P., Khedkar, S., Szklarczyk, D., Mende, D. R., Schmidt, T. S. B., & Bork, P. (2021). GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biology*, *22*(1), 1–19.

Pachiadaki, M. G., Brown, J. M., Brown, J., Bezuidt, O., Berube, P. M., Biller, S. J., Poulton, N. J., Burkart, M. D., La Clair, J. J., & Chisholm, S. W. (2019). Charting the complexity of the marine microbiome through single-cell genomics. *Cell*, *179*(7), 1623–1635.

Pachiadaki, M. G., Rédou, V., Beaudoin, D. J., Burgaud, G., & Edgcomb, V. P. (2016). Fungal and prokaryotic activities in the marine subsurface biosphere at Peru Margin and Canterbury Basin inferred from RNA-based analyses and microscopy. *Frontiers in Microbiology*, *7*, 846.

Pandey, R. P., Parajuli, P., & Sohng, J. K. (2018). Metabolic engineering of glycosylated polyketide biosynthesis. *Emerging Topics in Life Sciences*, *2*(3), 389–403.

Paoli, L., Ruscheweyh, H.-J., Forneris, C. C., Hubrich, F., Kautsar, S., Bhushan, A., Lotti, A., Clayssen, Q., Salazar, G., & Milanese, A. (2022). Biosynthetic potential of the global ocean microbiome. *Nature*, 1–8.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055.

Pelve, E. A., Fontanez, K. M., & DeLong, E. F. (2017). Bacterial succession on sinking particles in the ocean's interior. *Frontiers in Microbiology*, *8*, 2269.

Penn, K., Jenkins, C., Nett, M., Udwary, D. W., Gontang, E. A., McGlinchey, R. P., Foster, B., Lapidus, A., Podell, S., & Allen, E. E. (2009). Genomic islands link secondary metabolism to functional adaptation in marine Actinobacteria. *The ISME Journal*, *3*(10), 1193–1203.

Rapp, I., Schlosser, C., Menzel Barraqueta, J.-L., Wenzel, B., Lüdke, J., Scholten, J., Gasser, B., Reichert, P., Gledhill, M., & Dengler, M. (2019). Controls on redox-sensitive trace metals in the Mauritanian oxygen minimum zone. *Biogeosciences*, *16*(21), 4157–4182.

Rattray, J. E., Strous, M., Op den Camp, H. J. M., Schouten, S., Jetten, M. S. M., & Damsté, J. S. S. (2009). A comparative genomics study of genetic products potentially encoding ladderane lipid biosynthesis. *Biology Direct*, *4*(1), 1–16.

Scherlach, K., & Hertweck, C. (2009). Triggering cryptic natural product biosynthesis in microorganisms. *Organic & Biomolecular Chemistry*, *7*(9), 1753–1760.

Scherlach, K., & Hertweck, C. (2021). Mining and unearthing hidden biosynthetic potential. *Nature Communications*, *12*(1), 1–12.

Schlosser, C., Streu, P., Frank, M., Lavik, G., Croot, P. L., Dengler, M., & Achterberg, E. P. (2018). H 2 S events in the Peruvian oxygen minimum zone facilitate enhanced dissolved Fe concentrations. *Scientific Reports*, *8*(1), 1–8.

Schmidtko, S., Stramma, L., & Visbeck, M. (2017). Decline in global oceanic oxygen content during the past five decades. *Nature*, *542*(7641), 335–339.

Scranton, M. I., Astor, Y., Bohrer, R., Ho, T.-Y., & Muller-Karger, F. (2001). Controls on temporal variability of the geochemistry of the deep Cariaco Basin. *Deep Sea Research Part I: Oceanographic Research Papers*, *48*(7), 1605–1625.

Scranton, M. I., McIntyre, M., Astor, Y., Taylor, G. T., Müller-Karger, F., & Fanning, K. (2006). Temporal variability in the nutrient chemistry of the Cariaco Basin. In *Past and present water column anoxia* (pp. 139–160). Springer.

Scranton, M. I., Sayles, F. L., Bacon, M. P., & Brewer, P. G. (1987). Temporal changes in the hydrography and chemistry of the Cariaco Trench. *Deep Sea Research Part A. Oceanographic Research Papers*, *34*(5–6), 945–963.

Scranton, M. I., Taylor, G. T., Thunell, R., Benitez-Nelson, C. R., Muller-Karger, F., Fanning, K., Lorenzoni, L., Montes, E., Varela, R., & Astor, Y. (2014). Interannual and subdecadal variability in the nutrient geochemistry of the Cariaco Basin. *Oceanography*, *27*(1), 148–159.

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

Sinninghe Damsté, J. S., Rijpstra, W. I. C., Geenevasen, J. A. J., Strous, M., & Jetten, M. S. M. (2005). Structural identification of ladderane and other membrane lipids of planctomycetes capable of anaerobic ammonium oxidation (anammox). *The FEBS Journal*, *272*(16), 4270–4283.

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

Suter, E. A., Pachiadaki, M., Taylor, G. T., Astor, Y., & Edgcomb, V. P. (2018). Free-living chemoautotrophic and particle-attached heterotrophic prokaryotes dominate microbial assemblages along a pelagic redox gradient. *Environmental Microbiology*, *20*(2), 693–712.

Taylor, C. D., & Doherty, K. W. (1990). Submersible Incubation Device (SID), autonomous instrumentation for the in situ measurement of primary production and other microbial rate processes. *Deep Sea Research Part A. Oceanographic Research Papers*, *37*(2), 343–358.

Taylor, G. T., Iabichella, M., Ho, T.-Y., Scranton, M. I., Thunell, R. C., Muller-Karger, F., & Varela, R. (2001). Chemoautotrophy in the redox transition zone of the Cariaco Basin: a significant midwater source of organic carbon production. *Limnology and Oceanography*, *46*(1), 148–163.

Team, R. C. (2013). *R: A language and environment for statistical computing*.

Thaker, M. N., Wang, W., Spanogiannopoulos, P., Waglechner, N., King, A. M., Medina, R., & Wright, G. D. (2013). Identifying producers of antibacterial compounds by screening for antibiotic resistance. *Nature Biotechnology*, *31*(10), 922–927.

Vetting, M. W., Hegde, S. S., Fajardo, J. E., Fiser, A., Roderick, S. L., Takiff, H. E., & Blanchard, J. S. (2006). Pentapeptide repeat proteins. *Biochemistry*, *45*(1), 1–10.

Wang, S., & Lu, Z. (2017). Secondary metabolites in archaea and extreme environments. In *Biocommunication of archaea* (pp. 235–239). Springer.

Wang, Y., Wilks, J. C., Danhorn, T., Ramos, I., Croal, L., & Newman, D. K. (2011). Phenazine-1-carboxylic acid promotes bacterial biofilm development via ferrous iron acquisition. *Journal of Bacteriology*, *193*(14), 3606–3617.

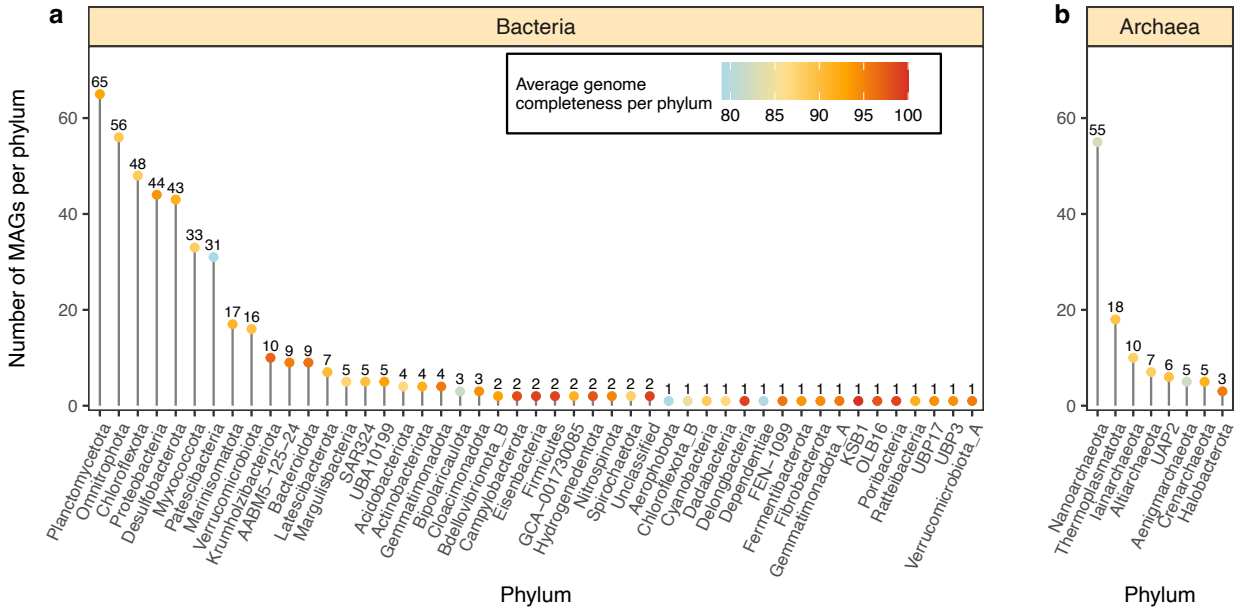
Waters, A. L., Hill, R. T., Place, A. R., & Hamann, M. T. (2010). The expanding role of marine microbes in pharmaceutical development. *Current Opinion in Biotechnology*, *21*(6), 780–786.

Webber, M. A., & Piddock, L. J. V. (2003). The importance of efflux pumps in bacterial antibiotic resistance. *Journal of Antimicrobial Chemotherapy*, *51*(1), 9–11.
Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2.' *Create Elegant Data Visualisations Using the Grammar of Graphics. Version*, *2*(1), 1–189.

Acknowledgments

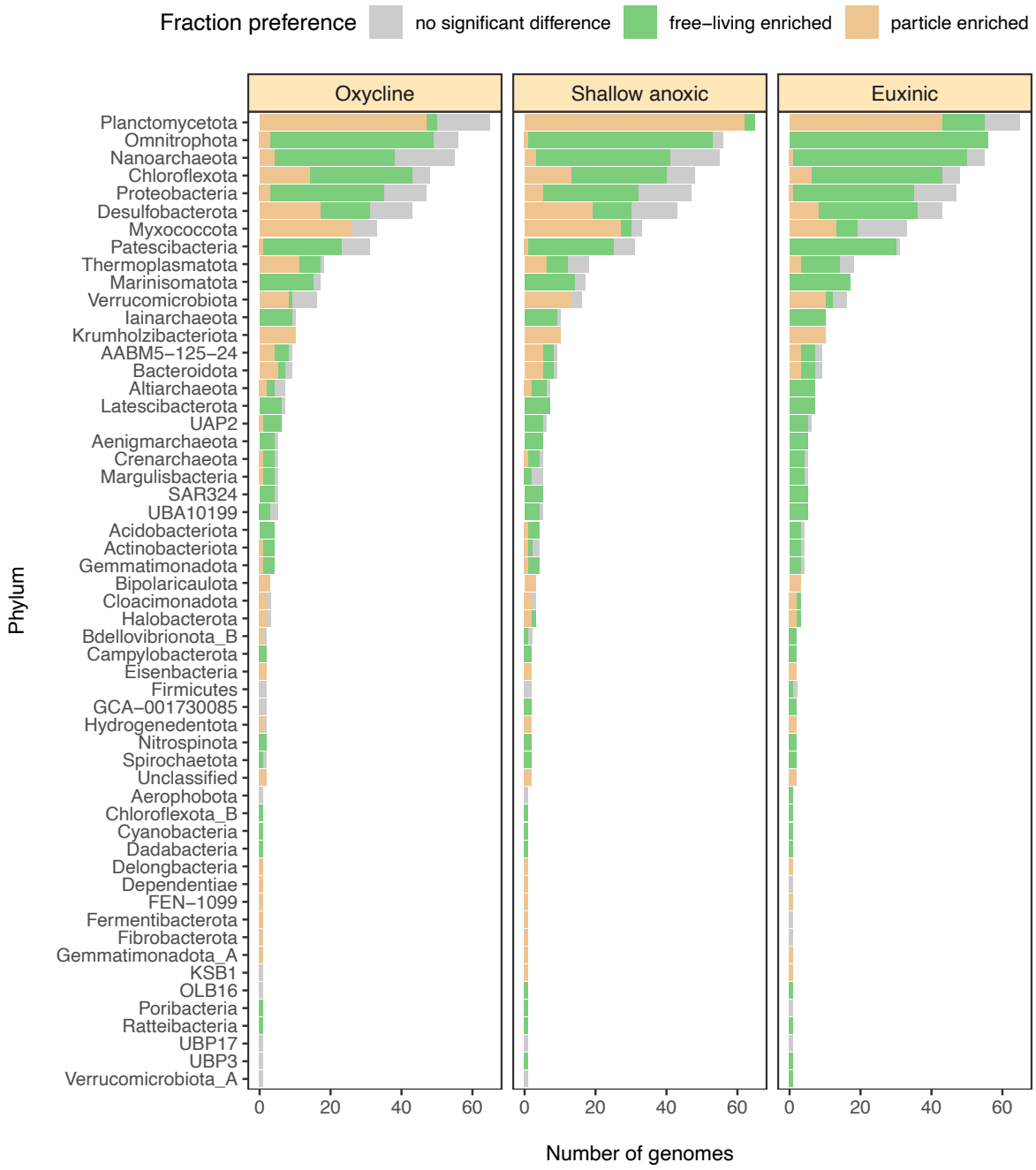
We thank the staff of Fundación La Salle de Ciencias Naturales (FLASA), EDIMAR, Porlamar, Edo Nueva Esparta, Venezuela and the crew of the R/V *Hermano Ginés* for their support during field work for this study, especially Y. Astor and R. Varela. The field work that provided samples and data for this study was supported by National Science Foundation (NSF) grants (OCE-1336082 to VE and OCE-1335436 and OCE-1259110 to Gordon Taylor, Stony Brook University). Analysis of the data was partially supported by NSF grant OCE-19924492 to MP and VE and Simons Foundation award 929985 to MP.

Supplementary Figures

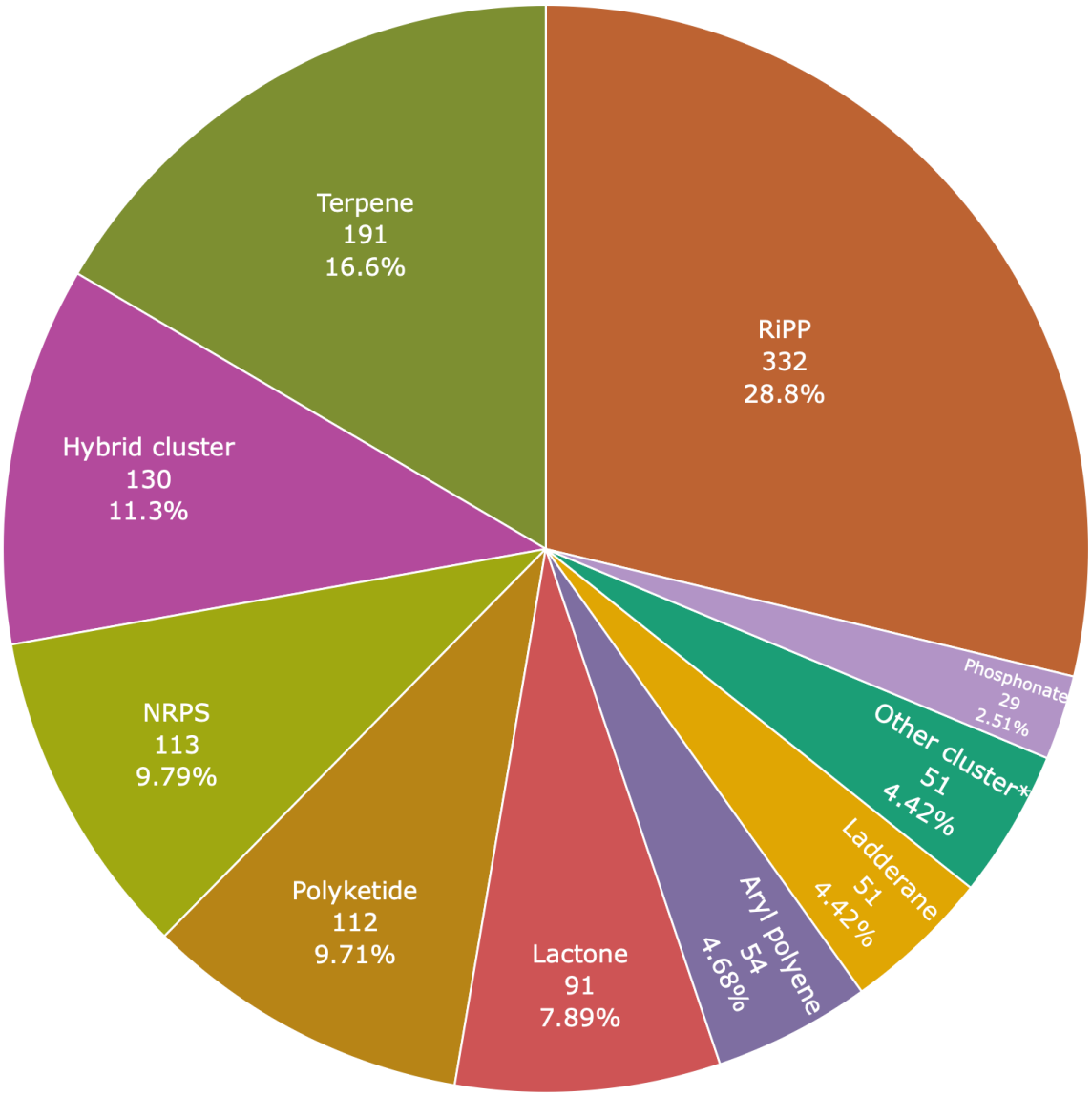


Supplementary Figure 2-1. Frequency of Cariaco prokaryotic MAGs (≥75% completeness, ≤5% contamination) by bacterial (a) and archaeal phylum (b). Colored dots at the end of each line segment correspond to the mean genome completeness of the phylum; the number above the dot quantifies the number of genomes recovered from the phylum.

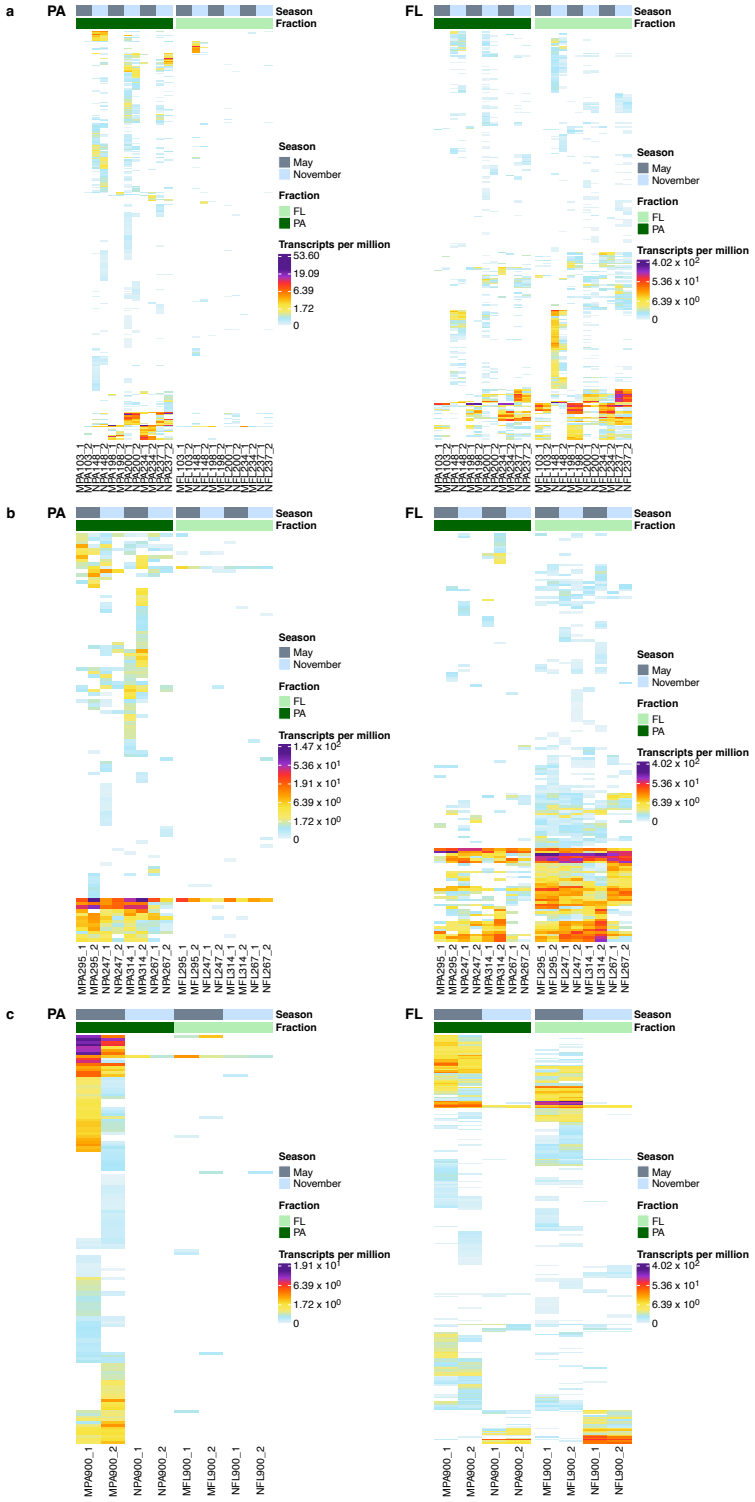
Supplementary Figure 2-2. MAG relative abundances. The heatmap shows for each column the percentage of total pre-processed reads from a metagenomic sample that mapped to the 100 most abundant MAGs (across all samples) using a log1p scale ranging from 0-32.12% relative abundance. Each row represents an individual MAG (additional information about the MAGs can be found in Supplementary Tables 1-2, 4 and 5).



Supplementary Figure 2-3. MAG differential abundance (fraction preference). Bar charts of the differential abundances of MAGs, grouped by phylum (DESeq2; $P < 0.05$; False Discovery Rate (FDR) = 5%) for each layer of the water column (oxycline, shallow anoxic, euxinic). The light brown color indicates MAGs that were more abundant in the PA metagenomes, the green color represents MAGs that were more abundant in the FL metagenomes, and the grey color indicates there was not statistically significant difference between the FL and PA fractions. Phyla represented in all three panels are from groups for which at least 5 genomes were recovered.



Supplementary Figure 2-4. Distribution of the biosynthetic gene clusters identified using antiSMASH 6.0. Each label corresponds to the total amount of biosynthetic gene clusters >= 10kb recovered from a given class, with the total number listed below the class followed by the percentage (%) out of the total BGC count (1,154). Other cluster* includes resorcinol, nucleoside, linear azol(in)e-containing peptide, acyl amino acid, cyclodipeptide, ectoine, redox-cofactor, non-alpha poly-amino group acid, siderophore, polybrominated diphenyl ether, and indole biosynthetic gene clusters.



Supplementary Figure 2-5. Biosynthetic transcript expression of MAGs with strict PA and FL fraction preferences. Heatmaps of the expression (transcripts per million) of biosynthetic gene cluster transcripts from MAGs with an apparent PA fraction preference (DESeq2; $P < 0.05$; False Discovery Rate (FDR) = 5%) in left-hand panels labelled "PA", and from MAGs more abundant in the FL fraction in the right-hand panels labelled "FL" from a oxycline, b shallow anoxic and c euxinic water layers.

Supplementary Figure 2-6. Expression of biosynthetic transcripts from gene clusters annotated as ladderanes. Heat map of the expression (transcripts per million) of biosynthetic gene cluster transcripts associated with gene clusters annotated by antiSMASH 6 as ladderanes. The column color annotations from top to bottom correspond to water layer, sampling season, and sampling fraction. The row color annotations are color-coded to differentiate MAG fraction preference (DESeq2; $P < 0.05$; False Discovery Rate (FDR) = 5%) as well as the phylum the MAG expressing the transcript is affiliated with. Each individual row is the aggregated TPM of a ladderane biosynthetic cluster.

Chapter 3

Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict

This chapter was originally published as:

Geller-McGrath, D., Konwar, K. M., Edgcomb, V. P., Pachiadaki, M., Roddy, J. W., Wheeler, T. J., & McDermott, J. E. (2024). Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict. *eLife*, *13*, e85749.

Statement of contribution

D.G.M. designed the research project with input from all co-authors. D.G.M. downloaded and extracted features from genomic data and trained machine learning models. D.G.M. developed the MetaPathPredict software with advice from K.M.K, J.W.R., T.J.W, and J.E.M. D.G.M. wrote the paper with input from all co-authors.

Abstract

The reconstruction of complete microbial metabolic pathways using 'omics data from environmental samples remains challenging. Computational pipelines for pathway reconstruction that utilize machine learning methods to predict the presence or absence of KEGG modules in incomplete genomes are lacking. Here, we present MetaPathPredict, a software tool that incorporates machine learning models to predict the presence of complete KEGG modules within bacterial genomic datasets. Using gene annotation data and information from the KEGG module database, MetaPathPredict employs deep learning models to predict the presence of KEGG modules in a genome. MetaPathPredict can be used as a command line tool or as a Python module, and both options are designed to be run locally or on a compute cluster. Benchmarks show that MetaPathPredict makes robust predictions of KEGG module presence within highly incomplete genomes.

Introduction

Microorganisms play a key role in all major biogeochemical cycles on Earth. Accurate and more complete identification of microbial metabolic pathways within genomic data is crucial to understanding their potential activities. This identification of pathways within genomic data, and assessment of their expression, provides important insight into their influence on the chemistry of their environment and their mediation of interactions with other organisms.

In recent decades, the scientific community has significantly advanced its capability to gather and sequence genomes from microorganisms. Key steps in the process of working with isolated genomes, single-amplified genomes (SAGs), or metagenome assembled genomes (MAGs), are identifying genes coding for enzymes that catalyze metabolic reactions and inferring the metabolic potential of the associated organism from these data. These analyses involve comparing protein-coding sequences with homologous sequences from reference metabolic pathway databases including KEGG (Kanehisa et al., 2000) and MetaCyc (Caspi et al., 2018). Environmental genomes that are recovered from high-throughput sequencing samples vary in their degree of completeness due to numerous factors including limited coverage of low-abundance microbes, composition-based coverage biases, insertion-deletion errors, and substitution errors (Browne et al., 2020). Enzymes encoded in genomes are also missed due to limitations in protein annotation methods, that is, undiscovered protein families may be undetected by traditional homology-based methods. This can limit the ability to determine the extent to which these organisms (or communities) can catalyze metabolic reactions and form pathways.

Sequencing biases, novel protein families, and incomplete gene and protein annotation databases lead to missing, ambiguous, or inaccurate gene annotations that create incomplete metabolic networks in recovered environmental genomes. This leads to a challenge in genome analysis: given a set of annotated genes that incompletely covers some known metabolic network, predict whether the network is, in fact, present in that organism (i.e. to predict whether one or more unobserved network components is likely present but unobserved for some reason). Existing algorithms for this metabolic network "gapfilling" largely fall into two categories of approaches: those based on network topology, such as the method utilized by Gapseq (Zimmerman et al., 2021), and those that utilize pre-defined KEGG module cutoffs, such as those used by METABOLIC (Zhou et al., 2022). Network topology and pathway gene presence/absence cutoffs, however, can lead to underestimation of pathways that are present, particularly in highly incomplete genomes.

Parsimony-based algorithms such as MinPath detect gaps in a metabolic network and identify the minimum number of modifications to the network that can be made to activate those reactions (Ye et al., 2009); its conservative approach, however, can lead to underestimation of the metabolic pathways present in a sample. KEMET (Palù et al.,, 2022) can detect gaps in metabolic pathways by searching unannotated genes in a genome with custom Hidden Markov Models (HMMs) created based on the genome's taxonomy. This approach, however, is limited by the genome taxonomies available in the KEGG GENES database. Other modern tools, such as DRAM (Shaffer et al., 2020), provide annotations for metagenomic sequences but do not closely tie these to metabolic pathways. Flux-balance analysis (e.g. Escher-FBA; Rowe et al., 2018) utilizes genome-scale metabolic models of organisms and requires experimental growth data for model parameterization; it is not easily applied to incomplete genome data, and the additional required experimental measurements may prohibit application in many use cases.

An emerging set of methods utilize machine learning models to a related problem of classifying microbial organisms' niches based on their genomic features. One such example is a tool called Traitar, which utilizes Support Vector Machines (SVMs) to predict lifestyle and pathogenic traits in prokaryotes based on gene family abundance profiles (Weiman et al., 2016). Other recent approaches have used machine learning approaches to train models using eukaryote MAG and transcriptome data to classify trophic mode (autotroph, mixotroph, or heterotroph) based on gene family abundance profiles (Lambert et al., 2021, Alexander et al., 2021). To our knowledge, there are no existing tools that predict the presence/absence of KEGG metabolic modules via machine learning models trained on gene features of high-quality genomes.

Here, we present "MetaPathPredict", an open-source tool for metabolic pathway prediction based on a deep learning classification framework. MetaPathPredict addresses critical deficiencies in existing metabolic pathway reconstruction tools that limit the utility and predictive power of 'omics data: it connects manually curated, current knowledge of metabolic pathways from the KEGG database with machine learning methods to reconstruct and predict the presence or absence of KEGG metabolic modules within genomic datasets including bacterial isolate genomes, MAGs, and SAGs.

The models underlying MetaPathPredict contain metabolic reaction and pathway information from taxonomically diverse bacterial isolate genomes and MAGs found in the NCBI RefSeq (O'Leary et al., 2016) and Genome Taxonomy (GTDB, Parks et al., 2021) databases. The

set of metabolic modules from the KEGG database is the basis of the tool's metabolic module reconstruction and prediction. The KEGG database contains metabolic pathway information for thousands of prokaryotic species and strains. KEGG modules are functional units of metabolic pathways composed of sets of ordered reaction steps. Examples include carbon fixation pathways, nitrification, biosynthesis of vitamins, and transporters or two-component systems (see Supplementary File 1a for a description of the distribution of modules covered by MetaPathPredict). MetaPathPredict is designed to run on the command-line locally or on a computing cluster and is available as a Python module on GitHub (https://github.com/d-mcgrath/MetaPathPredict).

A detailed overview of the MetaPathPredict pipeline is provided in Figure 1. The tool accepts as input gene annotations of one or more (possibly-incomplete) genomes, with associated KEGG ortholog (KO) gene identifiers. Because the genomes may be incomplete, it is possible that a KEGG module that is truly present in the organism will not be fully represented in the available data. MetaPathPredict first reconstructs both complete and incomplete KEGG metabolic modules, then predicts whether incomplete modules are in fact present. Input annotations can come from tools such as KofamScan (Aramaki et al., 2020), DRAM, blastKOALA (Kanehisa et al., 2016), ghostKOALA (Kanehisa et al., 2016), or a custom list of KO identifier gene annotations. MetaPathPredict classification models produced accurate results on held-out test genome annotation datasets even when the data were highly incomplete. A set of two deep learning models (5 hidden layers each) made predictions with a high degree of precision on all test datasets and with high recall on genomes with an estimated completeness as low as 30%. One model was trained to classify the presence or absence of 96 KEGG modules that were present in ≥10% and ≤90% of training genomes. The second model classifies 94 modules with an imbalanced profile of presence/absence (i.e. were present in <10% or >90% of training genomes). False positive predictions were rare in all tests, while false negatives increased when predictions were made with highly incomplete gene annotation information, as would be expected. We believe that MetaPathPredict is a valuable tool to further enhance studies of metabolic potential in environmental microbiome studies as well as synthetic biology efforts.

Results and Discussion

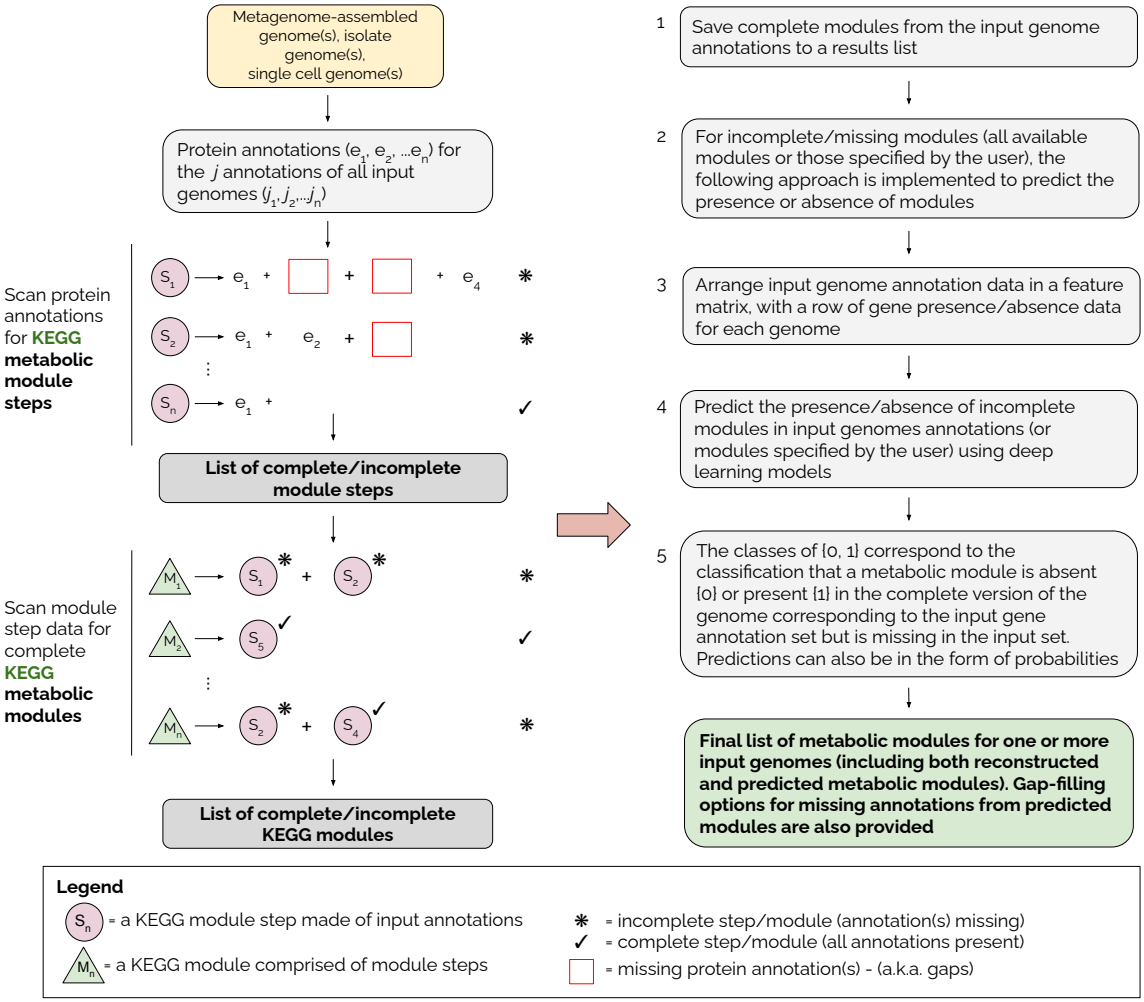


Figure 3-1. Overview of the MetaPathPredict pipeline. Input genome annotations are read into MetaPathPredict as a data object. The data are scanned for present KEGG modules and are formatted into a feature matrix. The feature matrix is then used to make predictions for all incomplete modules (or modules specified by the user). A summary and detailed reconstruction and prediction objects, along with gapfilling options are returned in a list as the final output.

MetaPathPredict is designed to predict the presence of a metabolic module even when annotation support for that module is incomplete, for example due to incomplete sequencing/annotation of the constituent proteins. It was trained on both complete and down-sampled genomes for this task. Complete genomes containing the genes required to non-redundantly complete a KEGG module were labelled as containing the module, otherwise the module was labelled as absent. To create down-sampled genomes for training, protein annotations were randomly removed form complete genomes in increasing increments while still retaining KEGG module class labels (from those

complete genomes). To assess MetaPathPredict's efficacy in this "gapfilling" task, we performed a variety of benchmarking experiments in which the complete genomes/proteomes were down-sampled to artificially produce incomplete modules.

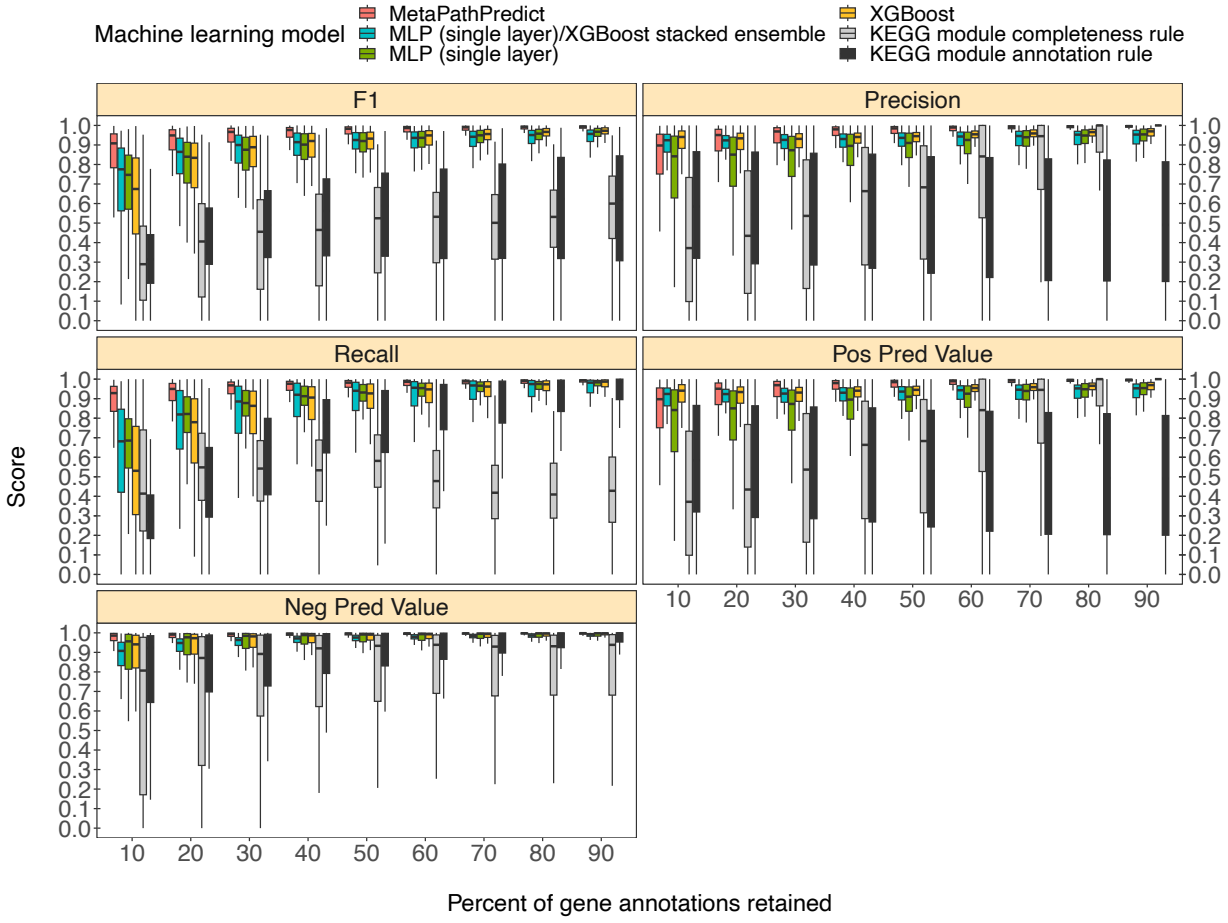


Figure 3-2. Comparison of performance metrics of MetaPathPredict's pair of deep learning multi-label classification models to next-best performing XGBoost, single-layer neural network, and XGBoost/single-layer neural network stacked ensemble machine learning models as well as two naïve classification rules. Down-sampled gene annotations of high-quality genomes used in held-out test sets are from NCBI RefSeq and GTDB. Each boxplot displays the distribution of model performance metrics for predictions on randomly sampled versions of the gene annotation test sets in downsampling increments of 10% (90% down to 10%, from right to left). The binary classifier performances are based on the classification of the presence or absence of KEGG modules in the complete versions of the gene annotations that were down-sampled for model testing.

MetaPathPredict's exhibited superior performance to other recently developed metabolic pathway reconstruction and prediction approaches. Its performance metrics on held-out test datasets suggest its models predict with high fidelity when at least 30% of gene annotations are recovered from a reconstructed genome (Figure 2). The efficacy of MetaPathPredict models was assessed using incomplete gene annotation data simulated from whole genomes, as well as from genomes reconstructed from reads that had been randomly down-sampled. We further

benchmarked MetaPathPredict against custom presence/absence classification rules, and existing gapfilling tools METABOLIC and Gapseq.

Benchmarking MetaPathPredict on down-sampled NCBI RefSeq and GTDB data

We compared the performance of MetaPathPredict's deep learning models to two classes of competitor classifiers: naive rule-based classifiers and various other machine learning model architectures. The evaluation was performed on test datasets comprised of isolate and metagenome-assembled genomes from GTDB and NCBI (30,596 total genomes; see Methods). When evaluating with the same sets of randomly down-sampled gene annotations, we found that each competing method showed poorer performance than MetaPathPredict (Figure 2). We assessed two naïve classification methods. First, we devised a classification rule based on the completeness of a KEGG module relative to the number of genes retained after downsampling: if, in a down-sampled genome, the number of genes involved in a KEGG module are present in at least the same proportion of all genes retained, the KEGG module is classified as 'present,' otherwise it is labeled 'absent'. For example, if 50% of gene annotations were removed from a genome during downsampling, then any KEGG module for which 50% of its associated genes are retained would be reported as 'present'. The results of this naive approach (Figure 2) show that the relative completeness of a KEGG module alone is not a robust classification strategy. The second classification rule that we tested was: for all gene annotation sets in the dataset, if any genes were present in an annotation set that were unique to a KEGG module (relative to other modules) then the module was classified as 'present', otherwise it was labelled 'absent'. The results of this naive approach (Figure 2) suggest that the presence of rare protein annotations or genes unique to a certain KEGG module is not always a strong indicator of the presence of a module in a genome. Ultimately, the performance of these naive classifiers indicate that MetaPathPredict's models have the advantage of incorporating information from genes outside of KEGG modules. We additionally compared the performance of various machine learning model architectures. Of these, the XGBoost, neural network (single hidden layer) and XGBoost/neural network (single hidden layer) stacked ensemble architectures were the next-best performing models and are included in Figure 2.

MetaPathPredict's deep learning strategy produced the best observed performance. Mean F1 score (a summary metric of the predictive performance) of the models was 0.96 when predicting

on test datasets in which 30-90% of gene annotations had been retained. MetaPathPredict rarely made false positive predictions based on data from highly incomplete gene annotation sets; the average precision of the models was consistently above 0.94 for all held-out test sets. MetaPathPredict also did not misclassify most negative class observations. The recall of MetaPathPredict's models was greater than 0.96 on average for test datasets containing at least 30% of the complete gene annotation data. The mean recall decreased to 0.89 and the mean precision decreased to 0.86 on genomes containing only 20% or less of the complete gene annotation data. The models' ability to achieve notably high recall even with significantly reduced sampling rates implies that it compensates for limited sequence availability by becoming more assertive in labeling a module as 'present' at the cost of decreased precision.

Benchmarking MetaPathPredict against Genomes from Earth's Microbiomes repository MAGs

MetaPathPredict was further tested on gene annotations from a set of 40 high-quality metagenome-assembled genomes from the Genomes from Earth's Microbiomes (GEM; Nayfach et al., 2021) genome repository. This repository contains a set of MAGs recovered from a diverse array of environments that make it ideal for benchmarking MetaPathPredict's performance (Figure 3). The MAGs selected from the repository had an estimated completeness of 100 and estimated contamination of 0, MIMAG quality score of "High Quality". The genomes belonged to 7 taxonomic phyla and were recovered from 9 different environments, primarily from human-associated and built environment metagenomes (see Appendix–Figure 1 for GEM genome taxonomic distributions and environmental sources). We created a set of 9 GEM datasets by randomly downsampling the data to retain 10% to 90% of gene annotations (in 10% increments) as in the previous section. MetaPathPredict classified the presence/absence of KEGG modules in each MAG. Overall, results were comparable to MetaPathPredict's performance on the GTDB/NCBI benchmark. The models excelled at predicting the presence or absence of KEGG modules in genomes when at least 40% of gene annotations were randomly retained. Predictions were less reliable though still accurate when 30% or less gene annotation data was retained.

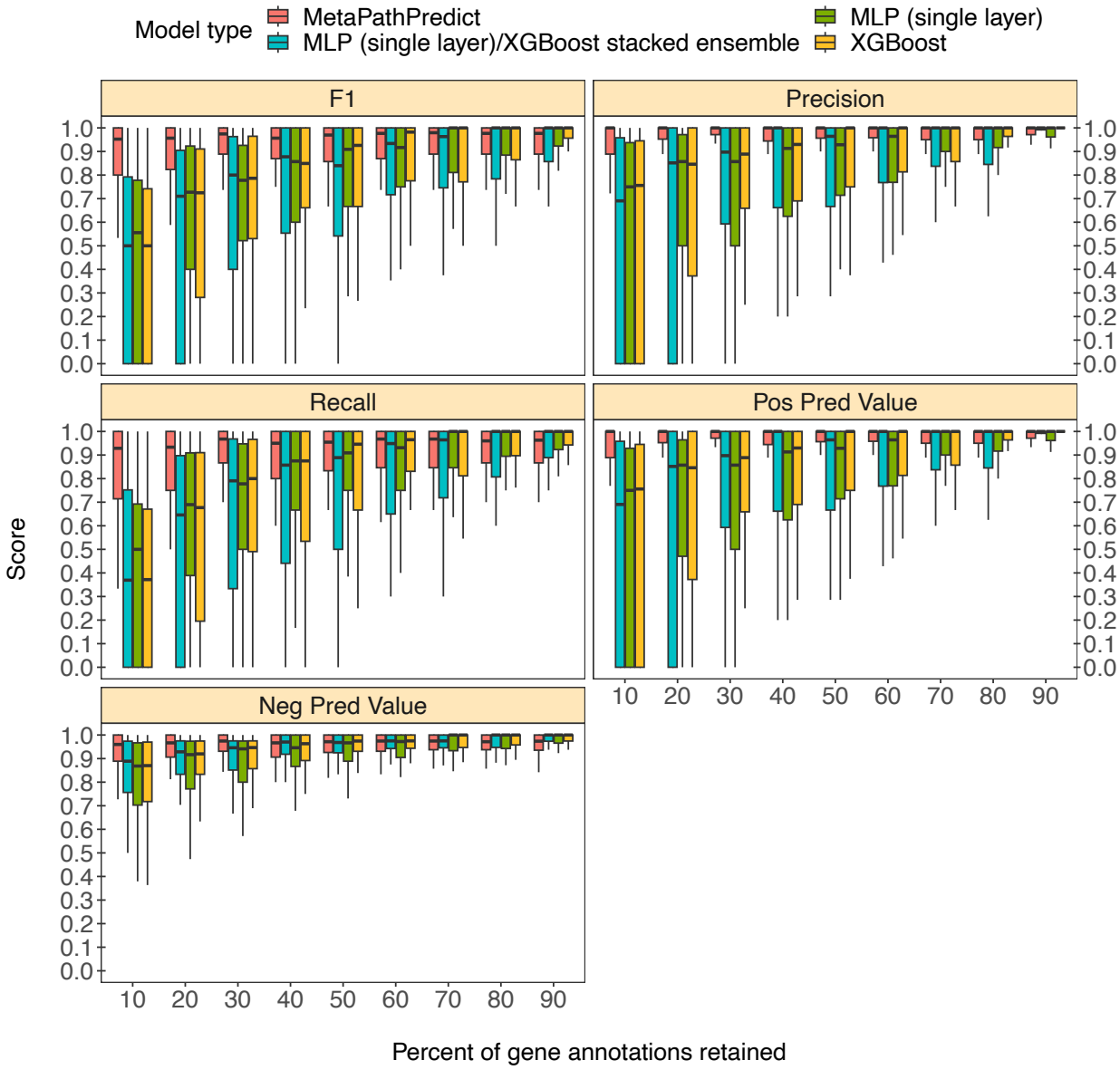


Figure 3-3. Boxplots of performance metrics of MetaPathPredict models on high-quality bacterial GEM MAGs (n = 40). Model performance metrics are for predictions on down-sampled versions of GEM genome gene annotations in decreasing increments of 10% (retaining 10-90% of the annotations in each test set). MetaPathPredict's deep learning models were benchmarked against XGBoost and neural network model architectures.

Benchmarking MetaPathPredict against existing tools on a dataset with down-sampled reads

In addition to model assessments made through down-sampling protein annotations, we evaluated a second set of held-out test set genomes from the GTDB/NCBI dataset (n = 50). In this analysis, the sequence reads for each genome were randomly down-sampled to simulate genomes incompletely recovered from an environmental sample. This analysis replicates situations with

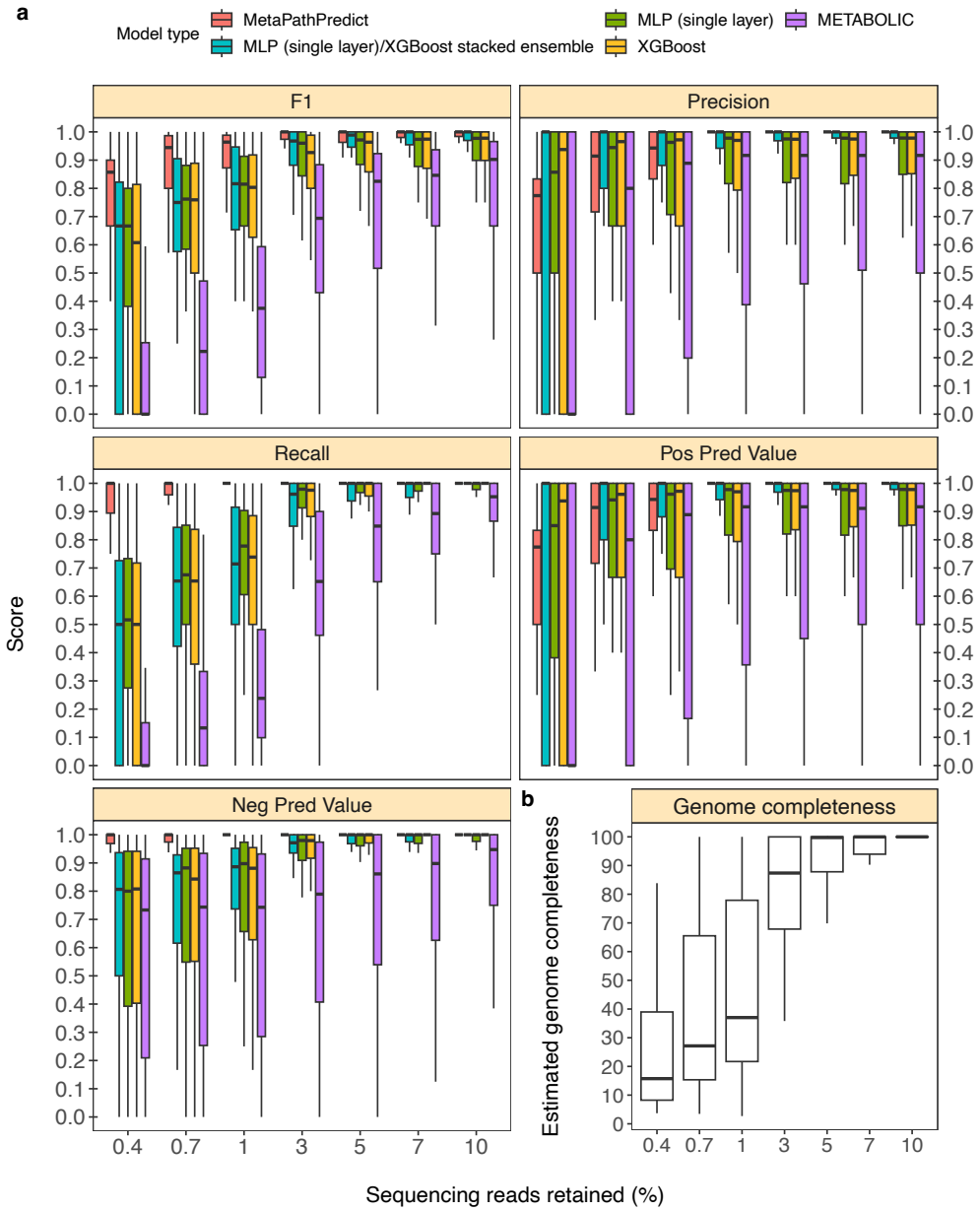


Figure 3-4. Performance metrics boxplots of 2 deep learning classification models. Down-sampled sequence reads of high-quality genomes used as a second held-out test set are from NCBI RefSeq and GTDB databases. **Panel a:** Boxplots display the distribution of model performance metrics for predictions of KEGG module presence/absence on simulated incomplete genomes down-sampled at the sequence read level by MetaPathPredict models, various next-best performing machine learning architectures, and METABOLIC. Downsampling increments were chosen based on average estimated completeness of the test set genomes at each increment to reflect a range of estimated completeness thresholds. **Panel b:** Average estimated genome completeness distributions of test set genomes that were down-sampled at the sequence read level using SeqTK and then assembled with SPAdes.

lower sequencing coverage, which can cause proteins to be unobserved due to incomplete or error-filled assemblies. As an example, using only 3% of reads (equivalent to an average of 1.5x

coverage of the genomes), roughly 86% of the genomes' proteins were assembled; meanwhile a reduction to 1% of reads caused assembly to recover ~40% of proteins. MetaPathPredict's performance on this test set resembled protein annotation random sampling results (Figure 4a, Figure 4b), though with greater loss in precision for down-sampling <3%. MetaPathPredict had an average F1 score for all 190 modules of 0.96 on the second held-out test set observations that had an average estimated genome completeness of at least 30%. The similarity of these results to the gene annotation downsampling approach validates the latter approach that was used more broadly to assess MetaPathPredict.

In addition to evaluating MetaPathPredict against our custom competitor models, we tested the software METABOLIC, which is a command line tool that performs gene annotations and estimates the completeness of individual KEGG modules in genomes and prokaryotic microbial communities (Figure 4a). METABOLIC showed much poorer recall at all levels of read sampling.

MetaPathPredict was also compared to another gapfilling tool, Gapseq (Figure 5a). Gapseq makes predictions of the presence or absence of KEGG pathways, and thus indirectly makes predictions of all modules and reactions they contain (instead of predictions for individual modules or reactions). We facilitated the direct comparison of MetaPathPredict to Gapseq by classifying a single KEGG pathway to be present if all modules it contained were predicted "present" by MetaPathPredict. MetaPathPredict outperformed predictions made by Gapseq, particularly on genomes with low read sampling prior to assembly.

Analysis of model feature importance using SHAP

Though the neural networks of MetaPathPredict produce accurate predictions of module presence/absence, it is not immediately clear what input features contribute to its decision-making process. To gain some insight into this, we calculated the importance of the various features of MetaPathPredict's models using the SHAP method (Lundberg and Lee, 2017), a mathematical method to explain the predictions of machine learning models. Features with large absolute SHAP values play an important role in calculating a model's predictions. SHAP values in the first model (trained to classify modules present in ≥10% and ≤90% of training genomes) indicated that 30 of the top 100 most important features (genes) influencing predictions were direct components of KEGG modules predicted by this model (Supplementary File 1b). In the second model (classifies

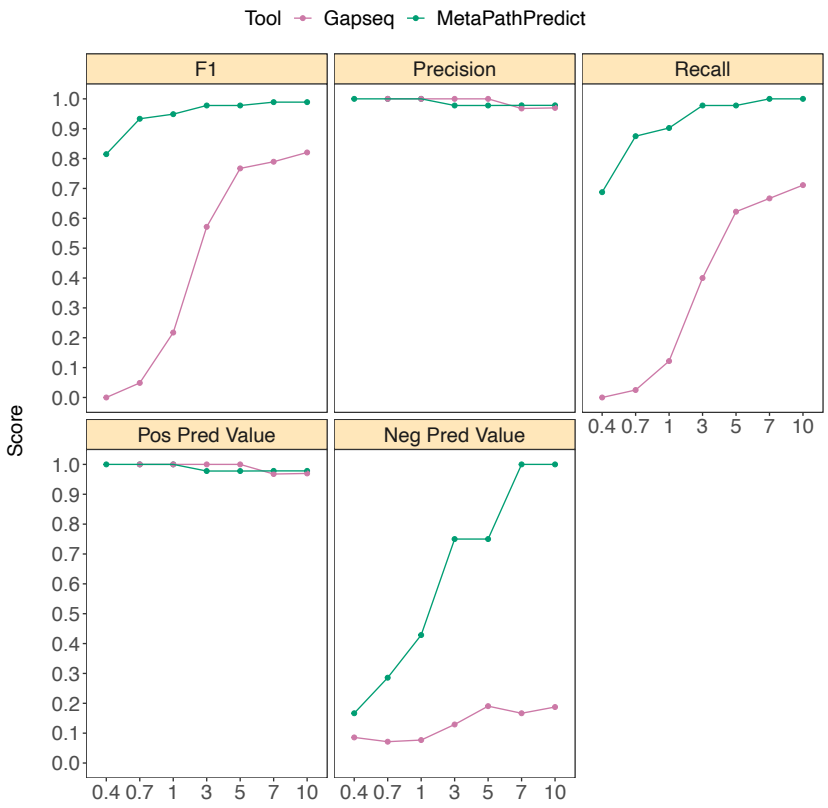


Figure 3-5. Performance metrics boxplots of MetaPathPredict and Gapseq predictions for KEGG pathway map00290 (Valine, leucine, and isoleucine biosynthesis) which contains KEGG modules M00019, M00432, M00535, and M00570. For MetaPathPredict predictions, the whole KEGG pathway was considered present if the aforementioned KEGG modules were all present. Down-sampled sequence reads of high-quality genomes used as a second held-out test set are from NCBI RefSeq and GTDB databases. Line segments display model performance metrics for MetaPathPredict and Gapseq predictions of KEGG pathway map00290 presence/absence on simulated incomplete genomes down-sampled at the sequence read level. Downsampling increments were chosen based on average estimated completeness of the test set genomes at each increment to reflect a range of estimated completeness thresholds.

modules present in <10% or >90% of training genomes), 37 of the 100 most influential features were part of KEGG modules this model was trained to predict.

The two models share 14 most important features out of the top 50 that are not part of KEGG modules. We examined these top features and found that there were a number of proteins annotated as sensors or transcriptional regulators (Supplementary File 1b). Also, we noted a number of transporters annotated as top features in both models and, interestingly, factors related to pathogenesis like toxins and mobile elements. Given the multi-label architecture of our models, it is difficult to draw direct conclusions from SHAP analysis. However, it is clear that MetaPathPredict's predictions are in part influenced by select genes present in KEGG modules, and also to a larger extent by genes not directly participating in KEGG module reactions.

Conclusion

MetaPathPredict is an open-source tool that can be used to characterize the functional potential of one or more sample genomes by detecting complete KEGG modules and predicting the presence or absence of those that are incomplete or missing. The tool accepts sets of gene annotations of individual genomes in KO gene identifier format as input. This type of annotation format can be acquired by annotating a genome of interest using KEGG-based annotation tools such as KofamScan (Aramaki et al., 2020), DRAM, blastKOALA (Kanehisa et al., 2016), or ghostKOALA (Kanehisa et al., 2016). MetaPathPredict also provides gene gapfilling options by listing putative KO gene annotations that could fill in missing gaps in predicted modules.

MetaPathPredict further validates the use of gene family presence or absence within a genome as a feature for bacterial metabolic function predictions. The performance metrics of MetaPathPredict on NCBI/GTDB and GEM test datasets validated the use of deep learning models to predict the presence/absence of KEGG metabolic modules with high fidelity on sparse to near-complete bacterial genomes. MetaPathPredict's multi-label classification models consistently made predictions with high precision and recall on simulated and real genomes using gene annotation and sequence read downsampling methods. The predictive power of the deep learning models was most limited when predicting on 10%-30% of protein annotations, and when the mean estimated completeness of reconstructed genomes from down-sampled reads was below 30%. We suggest that optimal performance with MetaPathPredict can be achieved when at least 40% of a genome has been recovered in an input bacterial gene annotation dataset.

Based on our performance tests of MetaPathPredict, the recall of its models was robust (mean > 0.9) even when protein sets were down-sampled to 10%. However, MetaPathPredict also surprisingly shows a decrease in precision (i.e. an increase in false calls of module presence). This, combined with surprisingly high recall rates at such low sampling rates, suggests that the model directly compensates for low general sequence availability by increasing aggressiveness in calling a module "present". This over-exuberant positive class prediction problem arose in our analyses only when <30% of gene annotation data was retained. Though such low sampling rates are not expected to be typical, it suggests an opportunity for method improvement.

Due to the multi-label architecture of MetaPathPredict's models, it is difficult to draw connections between the important features identified for the models and individual KEGG modules. However, the presence of sensing proteins (e.g. iron sensing and chemotaxis), pathogen

proteins (e.g. toxins and lysins), and transporters in these lists may indicate the contribution of lifestyle and environmental factors in predicting presence or absence of individual modules. Additionally, transcriptional regulators may be important due to their outsized influence on the expression of many genes (and thus modules) in each organism. Perhaps the most intriguing finding was that components of mobile elements (transposases) were found to be important features of both models. This could indicate that insertional elements are being used by the model to indicate, e.g., evolutionary lineage, which could be used to inform predictions of KEGG module composition.

MetaPathPredict facilitates more complete and accurate reconstruction of the metabolic potential encoded within bacterial genomes from a diverse array of environments and will enhance the ability to infer what metabolisms they are capable of, and/or how they may respond to perturbations. MetaPathPredict connects the field of machine learning with the growing community of environmental microbiologists using genomic sequencing techniques and will help transform and improve the way they work with environmental genomic datasets.

Materials and Methods

Filtering genome database metadata, downloading high-quality genomes, and gene annotations

The NCBI RefSeq (Release 205) database metadata file was downloaded and filtered to retain only the information for all bacterial genomes classified as "Complete genome". These are defined on the NCBI assembly help webpage: "all chromosomes are gapless and have no runs of 10 or more ambiguous bases (Ns), there are no unplaced or unlocalized scaffolds, and all the expected chromosomes are present (i.e. the assembly is not noted as having partial genome representation). Plasmids and organelles may or may not be included in the assembly but if present, then the sequences are gapless." This resulted in 17,491 complete NCBI genomes.

The GTDB bacterial metadata file for release 95 was downloaded and filtered to keep the information for all genomes with an estimated completeness greater than 99, an estimated contamination of 0, and a MIMAG (Bowers et al., 2017) quality score of "High Quality". A total of 30,760 non-redundant bacterial genomes from the two database metadata files were downloaded using the ncbi-genome-download command line tool (Blin, K.). The RefSeq genomes (17,491

total) were downloaded from the RefSeq FTP server, and the GTDB genomes (13,105 total) were downloaded from the GenBank FTP server (Appendix–Figure 2). Genes were called using Prodigal (Hyatt et al., 2010), and the KofamScan command line tool (Aramaki et al., 2020) was used to generate gene annotations (in KO gene identifier format) for all of the genomes using the KOfam set of HMMs available for download from the KEGG database (Kanehisa et al., 2002). KofamScan-derived annotations had to surpass their HMM's adaptive scoring threshold to be included in the training dataset. This approach provides resilience to using specific e-value cutoffs by preventing inflation of our training and assessment datasets with less-confident gene annotations.

Formatting gene annotation data, fitting KEGG module classification models

The full dataset of simulated incomplete genomes (n = 305,960) was split so that 75% of genomes were used for training and the remaining 25% as a test dataset. The training dataset was further split into 80% training/20% validation test sets. Each observation in the train/test/validation datasets contained a vector of length 8,853 that consisted of KO gene identifier (protein family) presence/absence indicated by ones and zeroes, respectively.

Training and test sets contained both complete and incomplete gene annotations of bacterial genomes from a diverse array of phyla (Appendix–Figure 2). The incomplete annotations used in training and testing of MetaPathPredict's models were constructed from complete genome annotation observations that were randomly down-sampled to retain 10-90% of the total gene content while the presence/absence class labels were kept unchanged for all down-sampled data. All complete and down-sampled versions of genomes were retained. The training datasets had a size of 305,960 observations, and the test datasets each contained 76,490 observations. The percent of observations with a positive class (a complete KEGG module 'present' in the gene annotations) in the training and test datasets varied, with a mean of 26.2% (Appendix–Figure 3).

The gene copy number data of the downloaded genomes was formatted in a matrix containing KO gene identifier presence/absence (1 or 0, respectively) in columns and genomes in rows. The label of each model was the presence/absence (1 or 0) of a KEGG module, as was determined using the KEGG modules downloaded from the KEGG database and the Anvi'o Python module (Eren et al., 2015). The "unroll_module_definition" function from the Anvi'o module was utilized with downloaded KEGG module data to create a list of all possible KEGG

Ortholog combinations to complete each module. For the module to be categorized as present, at least one possible combination of every step of the module had to be present in a genome, otherwise it was designated as absent. Two models were constructed for 190 KEGG modules for which at least 306 (0.1%) of the complete genomes (n = 30,596) contained the module genes, due to an improvement in performance when two models were trained (one model for the more balanced labels, one for highly imbalanced labels) instead of one. The models were trained using the gene annotation data of the genomes consolidated from the NCBI and GTDB databases. The first model was trained to classify modules within ≥10% and ≤90% of training genomes, while the second model classified modules within <10% or >90% of training genomes. The constructed models classify the presence or absence of complete KEGG modules based on the gene annotations of a genome.

A deep learning classification approach was chosen to model the relationship between whole genome KO gene identifier annotation data and the presence of metabolic modules. The same training data was used to train both of the models. MetaPathPredict is built on the Keras deep learning library (Chollet et al., 2015). L2-regularization was utilized to adjust hidden unit weights during training, with a learning rate of 0.001. Features used in each training dataset for classification were the presence or absence of protein families that were assigned KO gene identifiers. A deep learning architecture consisting of one input layer, five hidden layers, and one output layer were used as the machine learning architectures in MetaPathPredict's models. The input layer consisted of the presence/absence vector of KO gene identifiers (n = 8,853), and the hidden layers each contained 2,048 hidden units and were fully connected. The output layer of the first and second models contained 94 and 96 nodes for a total of 190 module presence/absence predictions when prediction outputs from both are combined.

Stratified sampling is a sampling method that ensures that all groups within the training and test data are represented in the same proportion as they are in the population as a whole. A multi-label stratified sampling method (Sechidis et al., 2011) was used to generate 75% train/25% test dataset splits that each contained data observations with preserved proportions of positive ('KEGG module present') and negative ('KEGG module absent') classes that were present in the genome dataset (see boxplot of the distribution of module presence/absence classes in Appendix–Figure 2, and an example of a held-out test dataset in Appendix–Figure 4). The

training dataset was further separated into 80% train/20% validation dataset splits to fit the deep learning models.

The binary cross entropy loss function was used in tandem with the Adaptive Moment Estimation (Adam) optimizer. The input and hidden layers utilized the rectified linear unit (ReLu) activation function; the output layer contained a sigmoid activation function. Dropout (Srivastava et al., 2014) was applied to 10% of edges at all layers except the final layer to avoid overfitting the training data. The input and hidden layers utilized the "he_uniform" layer weight initializer, and each of these layers contained 2,048 hidden units.

We assessed and benchmarked MetaPathPredict's models against two naïve classification methods. First, we devised a simple model that predicted the presence of a KEGG module if, after downsampling test sets of gene annotations, the proportion of module genes present in the dataset was greater than or equal to the percentage of annotations retained in the dataset. If the proportion of genes involved in a KEGG module were present in a dataset observation at least equivalently to the proportion of gene annotations retained after downsampling, the module was classified as 'present', otherwise it was classified as 'absent'. The second naïve classification rule was: for all gene annotation sets in the dataset, if any genes were present in an annotation set that were unique to a KEGG module (relative to all other KEGG modules) then the module was classified as 'present', otherwise it was labelled 'absent'. We additionally benchmarked MetaPathPredict's deep learning models against several other machine learning model types including single-layer neural network, XGBoost, and neural network/XGBoost stacked ensemble models, each trained on the same input data.

Evaluating models on test data, including test data randomly down-sampled to simulate varying degrees of genome incompleteness

Each of MetaPathPredict's models was validated on a held-out test set consisting of a combination of 76,490 complete and simulated incomplete genomes, and the performance metrics were extracted using the Scikit-learn (Pedregosa et al., 2011) Python module. The genome annotations in each test set were created by randomly downsampling complete genomes to simulate recovered gene annotations from incomplete genomes. 10% to 90% of genes from each annotation set were randomly retained (in increments of 10%) and used as input for MetaPathPredict predictions of

KEGG module presence/absence. The performance metrics used in evaluating the models were precision, recall, F1 score, positive predictive value, and negative predictive value (Table 1).

Metric	Definition
Precision	true positive/(true positive + false positive)
Recall	true positive/(true positive + false negative)
Specificity*	true negative/(true negative + false positive)
F1 score	2 × ((precision × recall)/(precision + recall))
Positive predictive value	recall × prevalence / ((recall × prevalence) + ((1 – specificity) × (1 – prevalence))
Negative predictive value	specificity × (1 – prevalence) / ((1 – recall) × prevalence) + (specificity × (1 – prevalence))
Prevalence*	(true positive + false negative) / (true positive + false positive + true negative + false negative)

Table 3-1. Definitions of machine learning model performance metrics used to assess MetaPathPredict models.
*Specificity and prevalence are defined due to their use in the definitions of negative and positive predictive value.

Testing models with a set of high-quality metagenome-assembled genomes from the Genomes from Earth's Microbiomes online repository

MetaPathPredict was further validated on another test set of genome annotations extracted from the GEM repository of MAGs. The GEM metadata file was downloaded from the repository and filtered to retain a random sample of 40 MAGs with a CheckM2 (Chklovski et al., 2023) estimated completeness of 100, an estimated contamination of 0, and a MIMAG quality score of "High Quality". The method for this assessment was the same as was described above for testing MetaPathPredict model performances on the held-out test data.

Evaluating models on test data down-sampled at the read level

A second held-out set of complete genomes (n = 50), independent of the training dataset, was downloaded from NCBI/GTDB databases using the SRA Toolkit (SRA Toolkit Development Team) and SRA explorer (Phil Ewels). The raw sequencing reads were filtered using fastp (Chen et al., 2018), and the quality-filtered reads were randomly down-sampled using seqtk (Li, H. 2012). Down-sampled reads were assembled into genomes using the SPAdes assembler (Bankevich et al., 2012), genes were called with Prodigal and then annotated using KofamScan. MetaPathPredict's deep learning models were then used to predict the presence or absence of all 190 KEGG modules in each genome and predictions were then cross-referenced with their known presence/absence based on the unmodified test dataset. In addition to simple approaches described above, the METABOLIC (Zhou et al., 2022) and Gapseq (Zimmerman et al., 2021) tools were evaluated on the same benchmark dataset. Both tools were used with default settings. Gapseq makes predictions of the presence or absence of entire KEGG pathways, and therefore it was benchmarked against MetaPathPredict by evaluating predictions for the presence or absence of the KEGG pathway map00290 (Valine, leucine, and isoleucine biosynthesis). This pathway consists of KEGG modules M00019, M00432, M00535, and M00570. In order to facilitate a direct comparison to Gapseq's predictions, the whole KEGG pathway was considered present if the aforementioned KEGG modules were all predicted as present by MetaPathPredict, otherwise it was classified as absent.

Gapfilling for incomplete modules predicted as present

MetaPathPredict provides enzyme gapfilling options for KEGG modules predicted as present by suggesting putative KO gene annotations missing from an input genome's gene annotations that could fill in missing gaps in predicted modules.

Data Availability

Genomic data used for creation of MetaPathPredict models is available from the NCBI Bacterial RefSeq Genomes database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/, version 209) and the Genome Taxonomy Database (https://data.gtdb.ecogenomic.org/releases/latest/, version r95). The GEM genomes used for model benchmarking are available at the GEM repository

(https://portal.nersc.gov/GEM/genomes/). The sequencing reads used for model benchmarking are available at the NCBI Sequence Read Archive website (https://www.ncbi.nlm.nih.gov/sra).

Code Availability

The scripts used for all data processing, model training, model benchmarking, and figure creation used in this study are available in the following GitHub repository: https://github.com/Microbiaki-Lab/MetaPathPredict_workflow. The MetaPathPredict Python module is available from the following GitHub repository: https://github.com/d-mcgrath/MetaPathPredict and XetHub repository: https://xethub.com/dgellermcgrath/MetaPathPredict.

Source Data

Links to all datasets used in this analysis, including source data used to generate the figures in this chapter can be found here: https://elifesciences.org/articles/85749.

References

1. Kanehisa, M. *et al.,* KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 28, 27–30 (2000). https://www.genome.jp/tools/kofamkoala/, Kanehisa et al., 2002
2. Caspi, R. *et al.,* The MetaCyc database of metabolic pathways and enzymes. *Nucleic acids research* 46, D633–D639 (2018).
3. Browne, Patrick Denis, et al., "GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms." *GigaScience* 9.2 (2020): giaa008.
4. Zimmermann, Johannes, Christoph Kaleta, and Silvio Waschina. "Gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models." *Genome biology* 22.1 (2021): 1-35.
5. Zhou, Zhichao, et al., "METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks." *Microbiome* 10.1 (2022): 1-22.
6. Ye, Y. *et al.,* A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput Biol* 5, e1000465 (2009).
7. Palù, Matteo, et al., "KEMET–A python tool for KEGG Module evaluation and microbial genome annotation expansion." *Computational and Structural Biotechnology Journal* 20 (2022): 1481-1486.
8. Shaffer, M. *et al.,* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Research* 48, 8883 8900 (2020).
9. Rowe, Elliot, Bernhard O. Palsson, and Zachary A. King. "Escher-FBA: a web application for interactive flux balance analysis." BMC systems biology 12.1 (2018): 1-7.

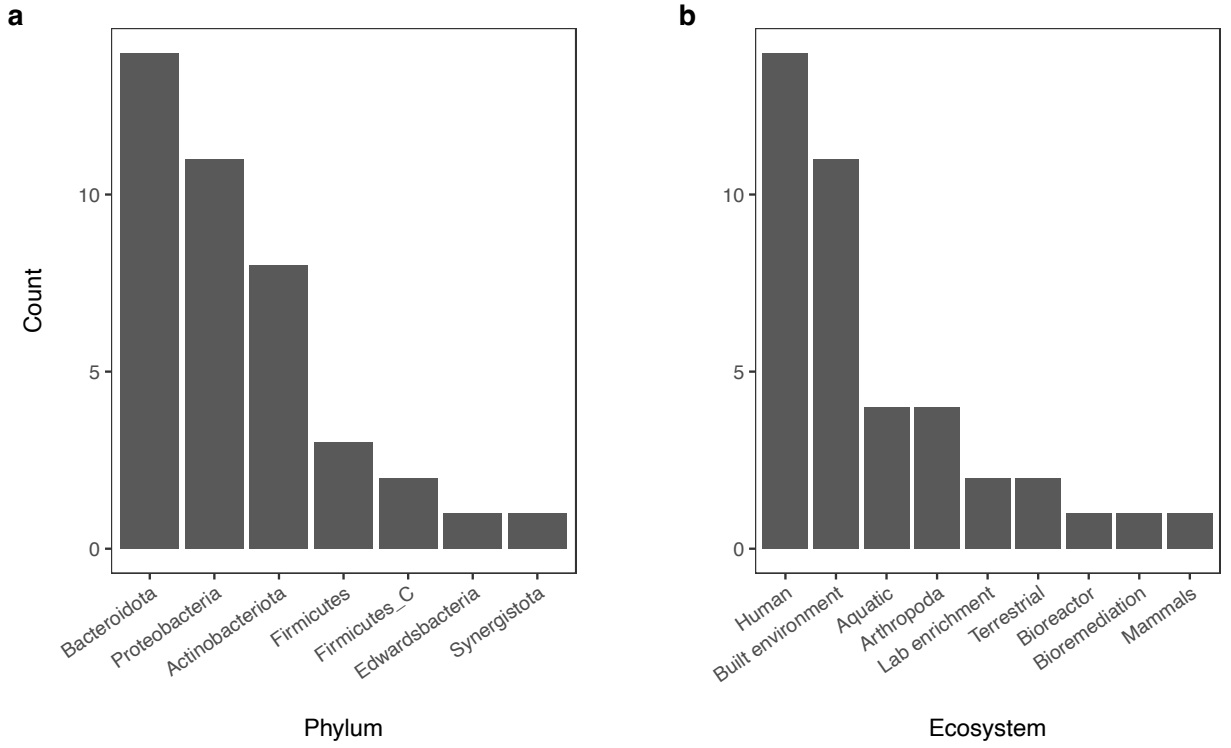
10. Weimann, Aaron, et al., "From genomes to phenotypes: Traitar, the microbial trait analyzer." *MSystems* 1.6 (2016): e00101-16.
11. Lambert, Bennett S., et al., "The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics." *Proceedings of the National Academy of Sciences* 119.7 (2022): e2100916119.
12. Alexander, Harriet, et al., "Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton." *bioRxiv* (2021).
13. O'Leary, Nuala A., et al., "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." *Nucleic acids research* 44.D1 (2016): D733-D745.
14. Parks, Donovan H., et al., "GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy." *Nucleic Acids Research* (2021).
15. Aramaki, Takuya, et al., "KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold." *Bioinformatics* 36.7 (2020): 2251-2252.
16. Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. "BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences." *Journal of molecular biology* 428.4 (2016): 726-731.
17. Bowers, Robert M., et al., "Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea." *Nature biotechnology* 35.8 (2017): 725-731.
18. Blin, K., https://github.com/kblin/ncbi-genome-download, version 0.2.10
19. Hyatt, Doug, et al., "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC bioinformatics* 11.1 (2010): 1-11.
20. Kanehisa, Minoru. "The KEGG database." *Novartis found symp*. Vol. 247. 2002.
21. Eren, A. Murat, et al., "Anvi'o: an advanced analysis and visualization platform for 'omics data." *PeerJ* 3 (2015): e1319.
22. Chollet, F., et al., Keras (2015). Retrieved from https://keras.io.
23. Pedregosa, Fabian, et al., "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12 (2011): 2825-2830.
24. Sechidis, Konstantinos, Grigorios Tsoumakas, and Ioannis Vlahavas. "On the stratification of multi-label data." Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22. Springer Berlin Heidelberg, 2011.
25. Srivastava, Nitish, et al., "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15.1 (2014): 1929-1958.
26. Nayfach, Stephen, et al., "A genomic catalog of Earth's microbiomes." *Nature biotechnology* 39.4 (2021): 499-509. https://genome.jgi.doe.gov/portal/GEMs/GEMs.home.html
27. Chklovski, Alex, et al., "CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning." *Nature Methods* (2023): 1-10.
28. SRA Toolkit Development Team, https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software)
29. Phil Ewels, https://sra-explorer.info/
30. Chen, Shifu, et al., "fastp: an ultra-fast all-in-one FASTQ preprocessor." *Bioinformatics* 34.17 (2018): i884-i890.

31. Li, Heng. "seqtk Toolkit for processing sequences in FASTA/Q formats." *GitHub* 767 (2012): 69.
32. Bankevich, Anton, et al., "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *Journal of computational biology* 19.5 (2012): 455-477.

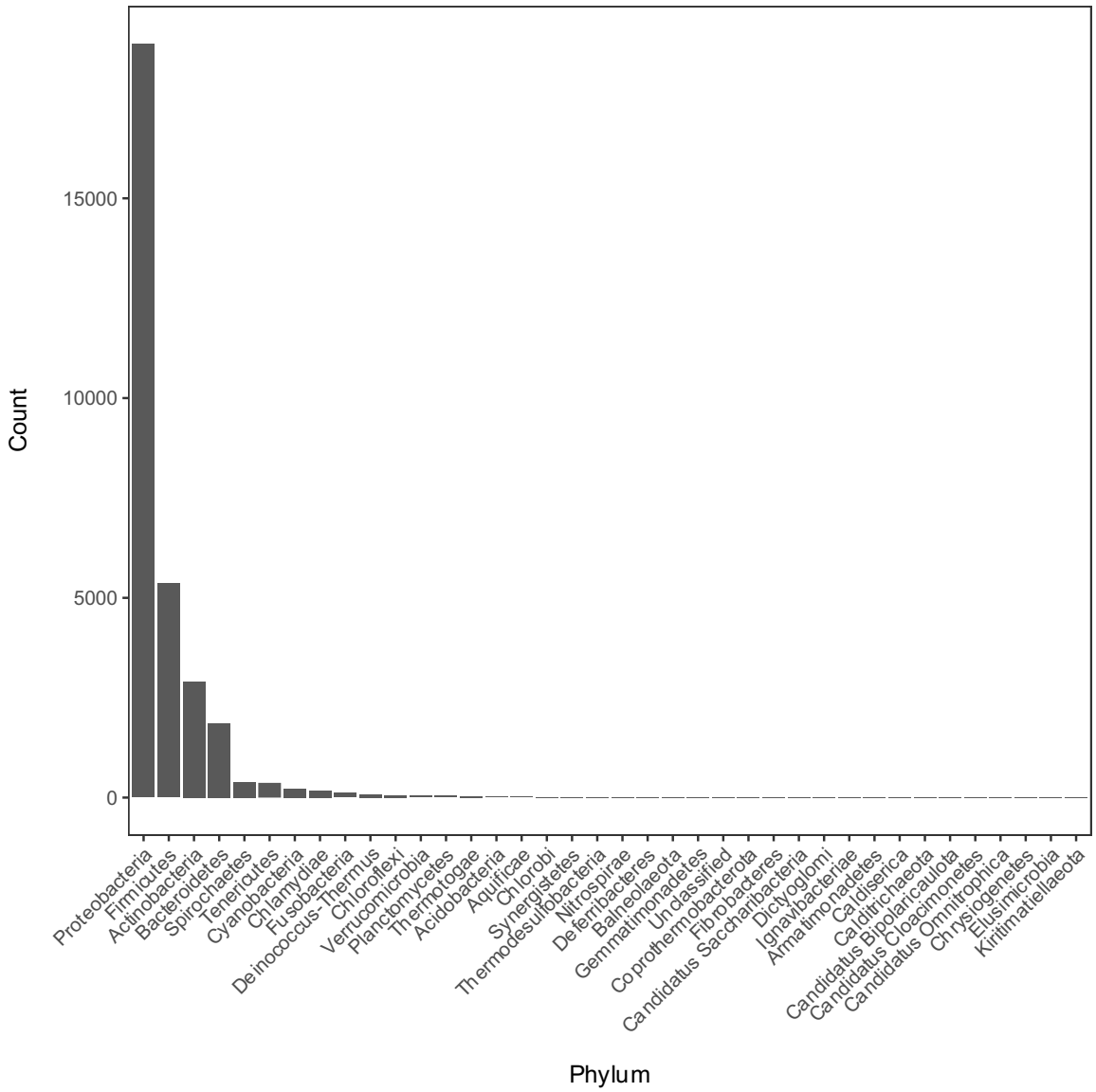
Acknowledgements

D. Geller-McGrath acknowledges funding from the Department of Energy (DOE) SCGSR Fellowship for the 2020 Solicitation 2 in Computational Biology and Bioinformatics. We would like to thank A. Solow (WHOI) for helpful initial discussion about statistical approaches. JWR and TJW were supported by NIH NIGMS R01GM132600. JEM, JWR, and TJW were supported by the DOE Office of Biological and Environmental Research (BER) through the "Machine-Learning Approaches for Integrating Multi-Omics Data to Expand Microbiome Annotation" project. PNNL is operated for the DOE by Battelle Memorial Institute under Contract DE-AC05-76RL01830.

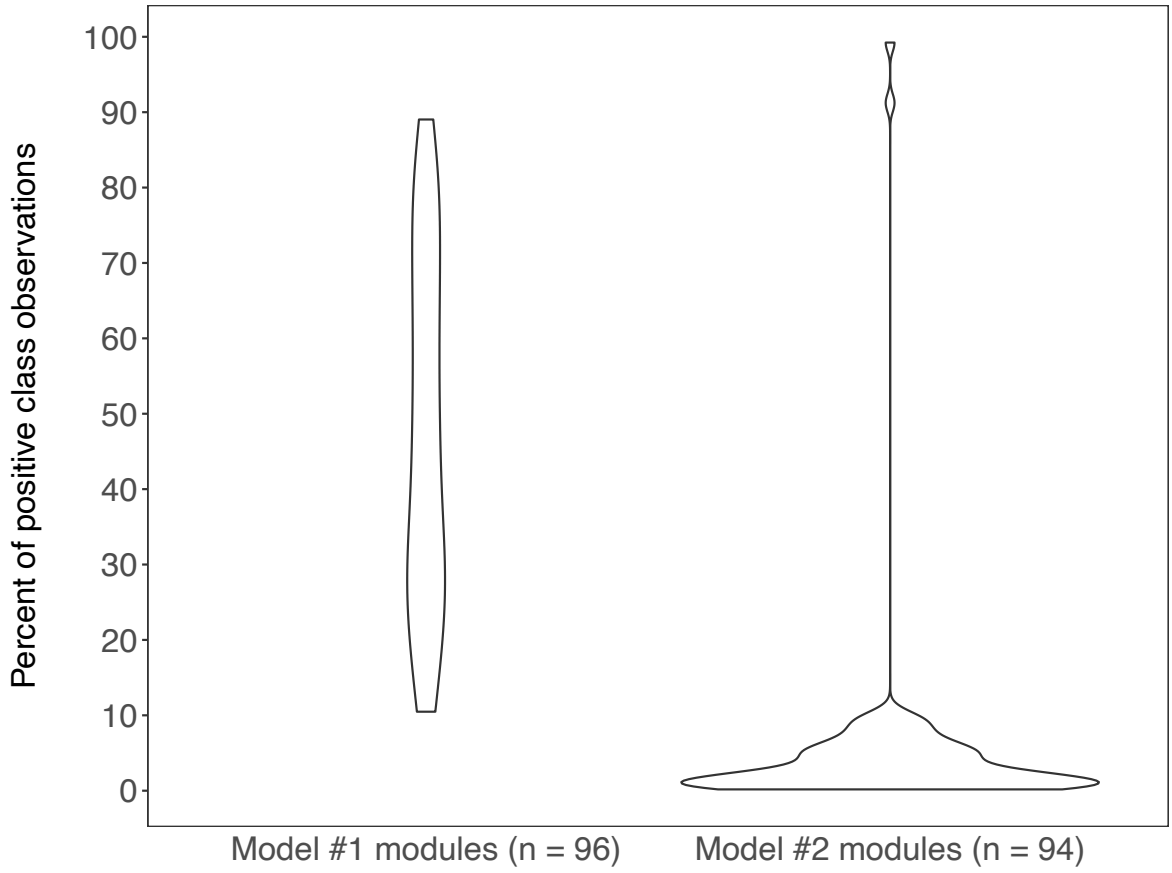
Appendix–Figures



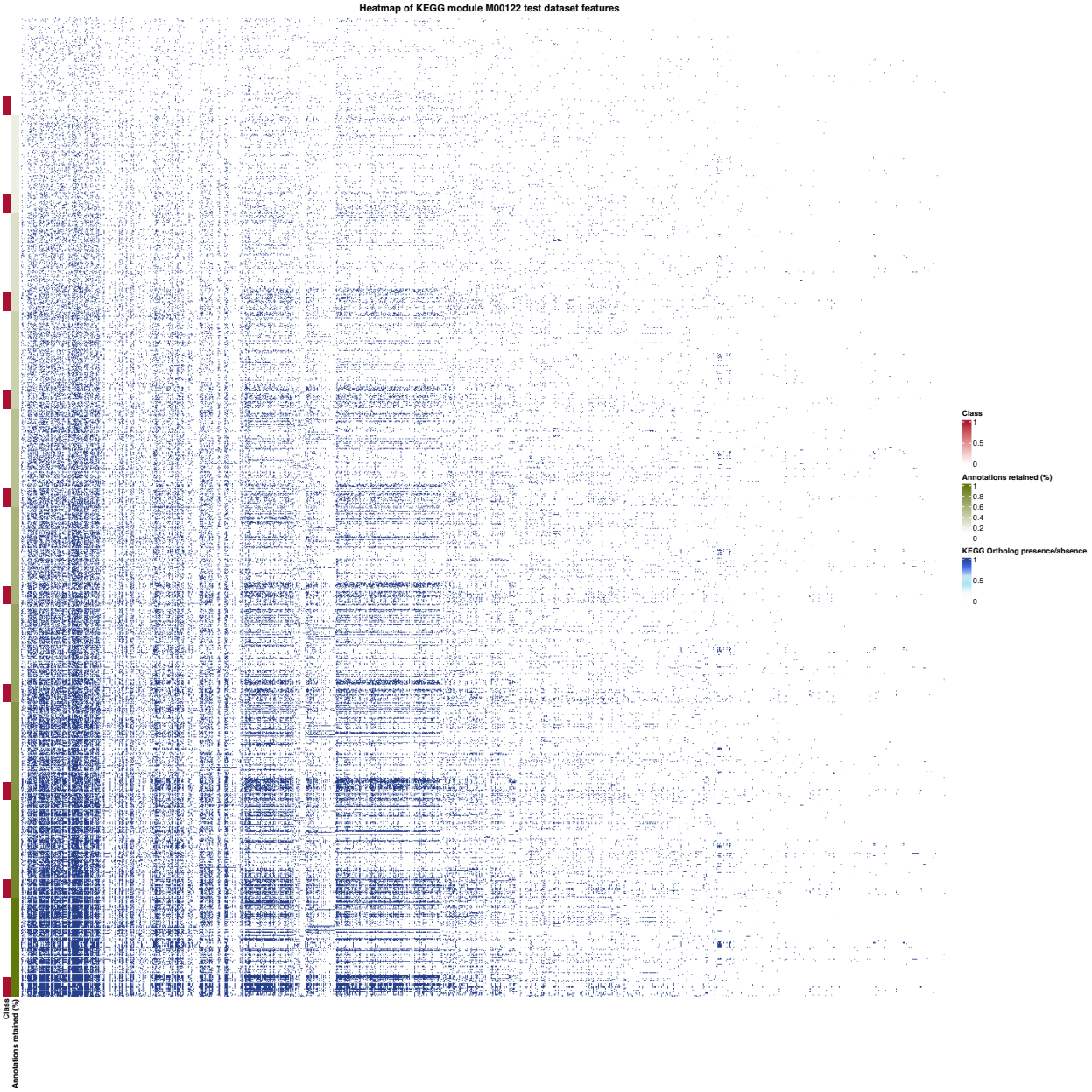
Supplementary Figure 3-1. Panel a: Bar chart of the taxonomic distribution of genomes (n = 40) from the GEM repository used during model validation. Panel b: Bar chart of the environmental sources of metagenomes the MAGs from this test set were recovered from.



Supplementary Figure 3-2. Distribution of phyla of bacterial genomes from which annotation data was used during model training and testing. See Supplementary File 1c for the full metadata table.



Supplementary Figure 3-3. Violin plots of the percent of positive "KEGG module present" classes for genomes from MetaPathPredict's deep learning training and test datasets for both of its models (model #1 on the left-hand side; model #2 on the right-hand side). Each train/test split contains the same distribution of positive and negative classes.



Supplementary Figure 3-4. Heatmap of held-out test data for the set of features (KEGG Ortholog presence/absence) used by MetaPathPredict's deep learning models. The annotation row on the left-hand side of the plot is annotated with classes and predictions for KEGG module M00122 (cobalamin biosynthesis), and is sorted by the percentage of protein annotations retained in each observation (increasing in protein annotations retained from top to bottom).

Chapter 4

Metagenomic profiles of archaea and bacteria within thermal and geochemical gradients of the Guaymas Basin deep subsurface

Paraskevi Mara[1*], David Geller-McGrath[2*], Virginia Edgcomb[1], David Beaudoin[1], Yuki Morono[3], Andreas Teske[4#]

[1]Geology and Geophysics Department, Woods Hole Oceanographic Institution, Woods Hole, MA, 02543, USA

[2]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA, 02543, USA

[3]Kochi Institute for Core Sample Research, Institute for Extra-cutting-edge Science and Technology Avantgarde Research (X-STAR), Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Monobe, Nankoku, Kochi, Japan

[4]Department of Earth, Marine and Environmental Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

*These authors contributed equally

[#]Corresponding author

Statement of contribution

VE, AT, and PM designed the study. VE and YM took primary responsibility for collecting the samples during IODP Expedition 385 together with shipboard scientists. AT served as Co-Chief Scientist of IODP Expedition 385. DB and PM extracted DNA and RNA for metagenomes and metatranscriptomes, respectively. DB prepared metagenome libraries for sequencing, and handled data deposition into GenBank. DGM took primary responsibility for bioinformatic processing of metagenome data and mapping of transcripts to MAGs. PM and DGM analyzed the MAG data and VE and AT contributed to data interpretation. YM provided cell count data. PM, DGM, VE,

and AT co-wrote the first draft of the manuscript. AT and PM led writing of all subsequent drafts, and all authors contributed to its final form.

This chapter was originally published as:

Mara, P., Geller-McGrath, D., Edgcomb, V., Beaudoin, D., Morono, Y., & Teske, A. (2023). Metagenomic profiles of archaea and bacteria within thermal and geochemical gradients of the Guaymas Basin deep subsurface. *Nature Communications*, *14*(1), 7768.

Abstract

Previous studies of microbial communities in subseafloor sediments reported that microbial abundance and diversity decrease with sediment depth and age, and microbes dominating at depth tend to be a subset of the local seafloor community. However, the existence of geographically widespread, subsurface-adapted specialists is also possible. Here, we use metagenomic and metatranscriptomic analyses of the hydrothermally heated, sediment layers of Guaymas Basin (Gulf of California, Mexico) to examine the distribution and activity patterns of bacteria and archaea along thermal, geochemical and cell count gradients. We find that the composition and distribution of metagenome-assembled genomes (MAGs), dominated by numerous lineages of Chloroflexota and Thermoproteota, correlate with biogeochemical parameters as long as temperatures remain moderate, but downcore increasing temperatures beyond ca. 45ºC override other factors. Consistently, MAG size and diversity decrease with increasing temperature, indicating a downcore winnowing of the subsurface biosphere. By contrast, specific archaeal MAGs within the Thermoproteota and Hadarchaeota increase in relative abundance and in recruitment of transcriptome reads towards deeper, hotter sediments, marking the transition towards a specialized deep, hot biosphere.

Introduction

The interplay between temperature stress and energy availability determines microbial survival in the subsurface biosphere, and delineates the extent and limits of life in the deep subsurface biosphere (Hoehler et al., 2013; Heuer et al., 2019). As microbial communities in cool, relatively shallow subsurface sediments transition into more deeply buried and increasingly warm and finally hot sediments, it should be possible to track how subsurface bacteria and archaea react to these

gradually harsher regimes downcore on the levels of cellular activity and community change. While microbial abundance and diversity are generally expected to decline downcore (Starnawski et al., 2017; Kirkpatrick et al., 2019), it is also possible that particular subsurface-adapted microbial populations benefit from conditions that would eliminate others, and constitute a specialized deep, hot biosphere. Recent studies indicated active microbial populations in extremely deep and hot sediments, yet without sequence-based identification (Heuer et al., 2020; Beulig et al., 2022). To learn more about bacterial and archaeal communities of the deep, hot biosphere from a genomic perspective, downcore trends of diversity and activity in increasingly hot sediments need to be examined, and microbial communities and their genomes have to be tracked downcore, as far as microbial biomass and DNA yield allow. Yet, investigating downcore changes in microbial abundance, community composition and activity in well-characterized geochemical and thermal gradients requires a suitable field site where extensive physical, chemical and microbial gradients can be sampled in adequate resolution by sediment coring and drilling.

An ideal natural laboratory for such a research task is Guaymas Basin, a hydrothermally-active ocean spreading center in the Gulf of California, covered by several hundred meters of sediment that host basaltic sill intrusions (Lizarralde et al., 2023) and strong geothermal heat flow (Neumann et al., 2023). Pyrolysis of buried organic carbon in these organic-rich sediments produces a complex milieu of petroleum hydrocarbons, including light hydrocarbons and methane, alkanes, and aromatic compounds, as well as carboxylic acids, and ammonia (Von Damm et al., 1985; Simoneit et al., 1995). These compounds are transported via hydrothermal fluids through Guaymas Basin's thick sediments, supporting diverse and active microbial communities (Teske et al., 2014). Collectively, these communities not only perform chemosynthetic carbon fixation and heterotrophic organic matter remineralization, but they also assimilate fossil carbon into the benthic biosphere (Pearson et al., 2005). Yet, few studies to date have explored the microbiology of deep subsurface sediments in Guaymas Basin. Methanogens were enriched from sediments collected during Deep Sea Drilling Project Leg 64 to Guaymas Basin (Oremland et al., 1982), and bacterial and archaeal communities in piston cores were surveyed using 16S rRNA amplicon sequencing (Ramírez et al., 2020; Teske et al., 2019; Vigneron et al., 2014). Aside from these studies, the spatial extent, diversity and activity of the deep biosphere in Guaymas Basin have remained largely unknown.

International Ocean Discovery Program (IODP) Expedition 385 drilled into Guaymas Basin at eight locations that differ in their degree of hydrothermal influence and heatflow (Neumann et al., 2023), and to survey their resulting characteristics (Teske et al., 2021a). Drilling

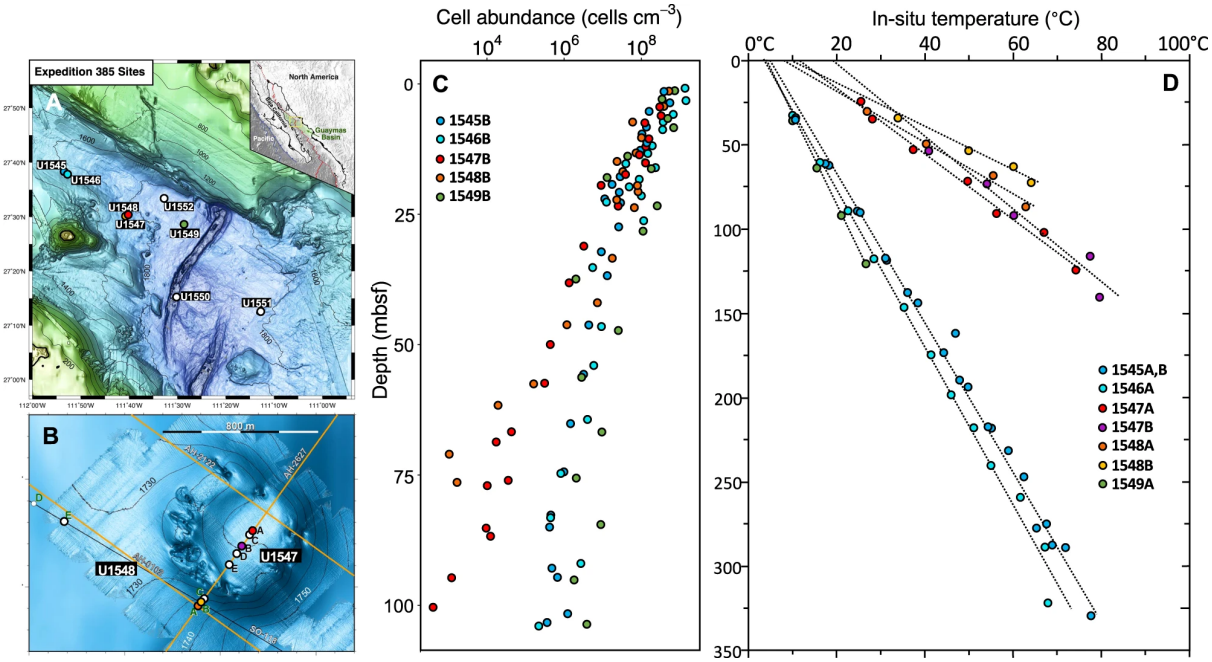


Figure 4-1. Locations, cell count profiles and temperature profiles for IODP Expedition 385 drilling sites. **A** Guaymas Basin bathymetry with drill sites. **B** Bathymetry of Ringvent with drill sites within and on the periphery of the Ringvent site. **C** Cell counts for drill sites (U1545, U1546, U1547, U1548, and U1549) where metagenomic and metatranscriptomic samples were collected. **D** Temperature profiles for drill sites where metagenomic and metatranscriptomic samples were collected. The lines indicated linear functions that were fitted to in-situ temperature measurements. Bathymetric maps, courtesy of D. Lizarralde (WHOI).

sites followed broadly a northwest-to-southeast transect across the northern Guaymas axial trough (Figure 1). Two neighboring sites (U1545 and U1546) on the northwestern end of Guaymas Basin (Teske et al., 2021b, Teske et al., 2021c) essentially differ by the presence of a massive, thermally equilibrated sill between 350 to 430 meters below seafloor (mbsf) at Site U1546 (Lizarralde et al., 2023). Two drilling sites (U1547, U1548) targeted the hydrothermally active Ringvent area, approximately 28 km northwest of the spreading center (Teske et al., 2019), where a shallow, recently emplaced and hot sill creates steep thermal gradients and drives hydrothermal circulation (Teske et al., 2021d). Drilling Site U1549 (Teske et al., 2021e) explores the periphery of an off-axis methane cold seep, Octopus Mound, located ~9.5 km northwest of the northern axial graben (Teske et al., 2021f).

These contrasting sites provide an opportunity for a comprehensive analysis of subsurface microbiota at different temperatures and depths. To assess the environmental distribution and genomic potential of microbes living in the deep biosphere of Guaymas Basin, we analyzed

reconstructed metagenome-assembled genomes (MAGs) from depths ranging from 0.8 to 219.5 mbsf at these thermally and geochemically contrasting sites. We also provide evidence for the activity of specific bacterial and archaeal lineages by mRNA transcript mapping on bacterial and archaeal MAGs.

Results and Discussion

Sampling sites and depths

Metagenomes were produced from sediment samples at drilling sites U1545B to U1549B that follow a northwest-to-southeast transect across the northwestern flanking region of Guaymas Basin (Figure 1A) and include an off-axis hydrothermal system, the Ringvent site (Figure 1B). The samples were selected to coordinate with depths used for separate ongoing analyses, and ranged from 1.7 m to 219.5 mbsf at Site U1545B, 0.8-16.3 mbsf at U1546B, 2.1-75.7 mbsf at U1547B, 9.1-69.4 mbsf at U1548B, and 16.5 mbsf at U1549B (Figure 1; Table 1). For all samples, a wide range of geochemical parameters was analyzed shipboard (Supplementary Dataset 1). The sites represent distinctly different thermal gradients and cell densities; generally, sites with steeper downcore temperature gradients are characterized by more rapidly decreasing cell counts (Figure 1C, D). U1545B is the reference site for IODP Expedition 385 because of the absence of seepage, hydrothermal influence, and massive sill intrusions (Teske et al., 2021b). Here, metagenome libraries extended down to 219.5 mbsf, at *in-situ* temperatures of 54.3°C. Cell count trends for sites U1545, U1546 and U1549 were similar, and showed a decrease over three orders of magnitude within 100 meters (Figure 1C). At the hot Ringvent sites U1547B and U1548B (Neumann et al., 2023; Teske et al., 2021d), comparable temperatures of 50-55°C were already reached near 70 mbsf (Figure 1D), and cell counts decreased by four to five orders of magnitude within this depth range (Figure 1C). To describe temperature-related trends in MAG recovery and diversity, we categorized our samples into three groups according to temperature; cool (2-20°C), warm (20-45°C) and hot (>45°C).

Subsurface Biogeochemical zonation

Most samples collected for metagenomes are from sediments within the sulfate-reducing zone where sulfate is still available at near-seawater concentrations (~28 mM) or becomes gradually

depleted with depth (Table 1). At those same sediments hydrogen sulfide concentrations are gradually increasing towards multiple millimolar concentrations. Metagenome samples from site U1545B also include depths spanning the sulfate-methane transition zone (SMTZ) at ~ 64 mbsf where sulfate is consumed by microbial sulfate reduction, and methane begins to accumulate. At the SMTZ sulfate concentrations drop from 21.1 mM to 0.7 mM, sulfide reaches peak concentrations of 8.9 mM, and methane concentrations increase from picomolar to 1.5 mM (Table 1). High methane concentrations persist also in deeper samples from U1545B, and decrease only in the very deepest samples (> 200 msbf). The deep subsurface methane reservoir at this and other sites results from long-term thermogenic and biological methane accumulation (Bojanova et al., 2023). In contrast to site U1545B, samples from Ringvent sites U1547B and U1548B show gradual downcore sulfate consumption (from 27.9 to 18.8 mM) but not depletion, combined with hydrogen sulfide accumulation (max. 7.1 mM at 75.7 mbsf at U1547B); methane does not accumulate in these samples. Ammonia concentrations increase from < 1 mM towards 3 to 5 mM downcore at most sites, and reach 9 to 25 mM below the SMTZ in U1545B. Dissolved inorganic carbon (DIC) and alkalinity concentrations are generally highest at Site U1545B where they peak in the SMTZ (~28 and 60 mM, respectively). Ammonia, DIC and alkalinity remain elevated not only in the upper sediment column but also in the deeper samples of Site U1545B, presumably due to cumulative bioremineralization of buried organic matter over time at this undisturbed site. In contrast, the Ringvent samples (sites U1547B and U1548B) generally have lower ammonia, alkalinity and DIC porewater concentrations, suggesting reduced remineralization of organic matter at these sites, most likely a consequence of hydrothermal activity due to recent volcanic sill emplacement (Teske et al., 2019). Dissolved organic carbon (DOC) remained ~10 to 20 mg/L in most samples but increased towards 70 mg/L in the sulfate-methane transition zone of U1545B and remained between 20 and 50 mg/L in the deeper sediments of U1545B. This suggests DOC enrichment and decreased heterotrophic DOC consumption in deep methanogenic sediments of U1545B where energy-rich electron acceptors for heterotrophic carbon remineralization are not available. While total nitrogen and total organic carbon generally decrease with depth at all sites, the Ringvent sites have moderately elevated TOC/TN ratios (Table 1), likely reflecting the influence of nitrogen-depleted hydrothermal carbon sources (Ramírez et al., 2020). Total petroleum hydrocarbon, saturated and polyaromatic hydrocarbon content remain each quite similar across a wide range of sediments and temperatures, before increasing considerably in hot

sediments (>80°C) near deep sill intrusions (Supplementary Dataset 1, and Supplementary Figures 1A, B).

MAG diversity, distribution, and evidence of activity

A total of 142 metagenome-assembled genomes (MAGs) were recovered from a co-assembly of all metagenomic samples (Supplementary Dataset 2). MAGs that matched those from negative controls were excluded from further analysis (Supplementary Dataset 3). For downstream analysis, we retained 89 bacterial and archaeal MAGs that had at least ≥ 50% bin completeness and ≤ 10% bin contamination (Bowers et al., 2017). Genome completeness ranged from ~50 to 97% (Supplementary Figure 2; Supplementary Dataset 4).

Of these 89 MAGs, 26 MAGs were assigned to 6 archaeal phyla, and 63 MAGs were assigned to 13 bacterial phyla (Figure 2); the phylogenetic spectrum includes lineages documented previously in 16S rRNA gene amplicon sequencing of shallow subsurface sediments (Ramírez et al., 2020; Teske et al., 2019), and in metagenomic surveys of shallow hydrothermal sediments of Guaymas Basin (Dombrowski et al., 2018). In parallel to downcore decreasing cell numbers (Figure 1), MAG diversity decreased downcore at all sites as temperatures increased (Figure 2). In samples from cool (3-20°C) sediments from all sites, reads mapped to diverse bacterial and archaeal phyla, including the bacterial phyla Chloroflexota, Acidobacteriota, Desulfobacterota, WOR-3, Aerophobota, and Bipolaricaulota, and the archaeal phyla Thermoproteota, Thermoplasmatota, and Aenigmatarchaeota (Figure 2). In samples with warm temperatures (20-45°C), reads were predominantly assigned to bacterial phyla Chloroflexota (mostly order-level group G1F9), Acidobacteriota, WOR-3 (order-level group UBA3073), Aerophobota and Bipolaricaulota, and to archaeal phyla Thermoproteota, Hadarchaeota, and Aenigmatarchaeota. At hot temperatures (45-60°C), bacterial reads mapped primarily to a single Chloroflexota MAG (class Dehalococcoidia), a single WOR-3 MAG and two Aerophobota MAGs (class Aerophobia). In contrast, several Archaeal MAGs show a marked preference for hot sediments, and mapped to the Thermoproteota (class Bathyarchaeia), and Hadarchaeota (class Hadarchaeia). Our recovered MAGs reflected metabolisms predicted for the deep biosphere including sulfur, nitrogen and methane cycling, hydrocarbon degradation, and carbon fixation (Chklovski et al., 2023; Supplementary Note, and Supplementary Figures 3 and 4). Desulfobacterota MAGs linked to sulfate reduction contained the *dsr* operon (e.g., *dsrB/J/K/D*) that is essential for dissimilatory

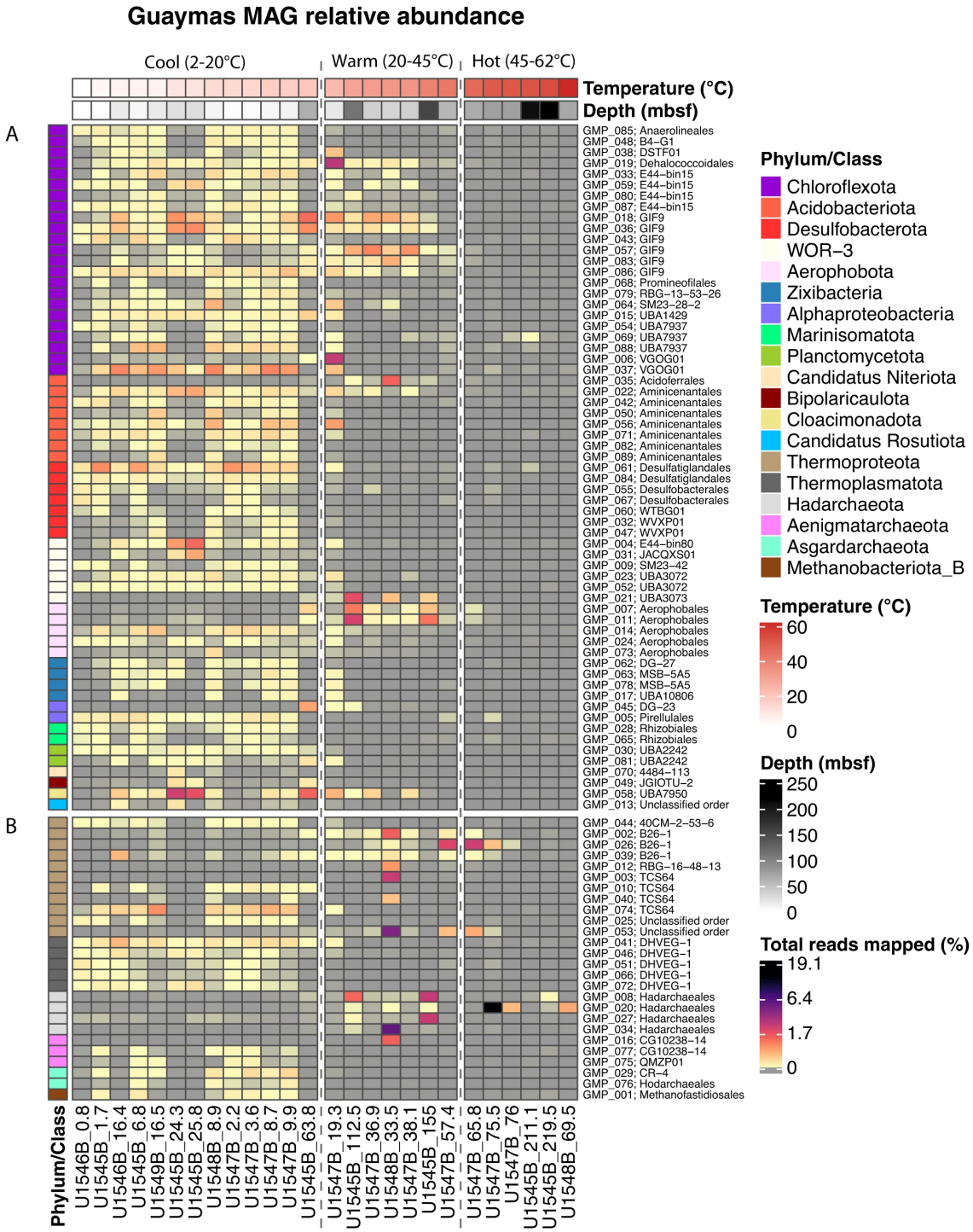


Figure 4-2. Heatmap of MAG relative abundance. Each column shows the percentage of total pre-processed metagenomic reads (relative abundance) that mapped to all 89 MAGs, for samples ordered by increasing temperature from left to right on the x-axis (annotated by site numbers and depths in mbsf). Temperature regimes (Cool, Warm, and Hot) are separated by vertical dashed lines. Each row shows the abundance profile of an individual MAG across all samples. MAGs are color-coded by phylum on the left, and annotated by GMP (Guaymas MAG Prokaryote) numbers 001 to 089 and order-level affinity to the right. RKPM, reads mapped per kilobase of genome, per million mapped reads. Panel section **A** denotes bacterial MAGs and panel section **B** denotes archaeal MAGs.

sulfate reduction (Anantharaman et al., 2018), and were recovered from shallow sediments with

available sulfate. These MAGs also shared potential for iron reduction using extracellular electron transfer mechanisms, such as *mtrA, mtoA,* and *eetB* genes (Garber et al., 2020; Chatterjee et al., 2021). Marker genes of dissimilatory iron reduction (Garber et al., 2020; e.g., *dmkA*, *dmkB*, *eetA*, *eetB*, *fmnA*, *fmnB*, *pplA*, *ndh2*) with the potential for extracellular electron transfer (EET) were identified in 86/89 MAGs within all recovered phyla (Supplementary Note, and Supplementary Datasets 5, 6).

To determine any intra-phylum differences in metabolic activity, we mapped reads of the Guaymas Basin subsurface metatranscriptome (Mara et al., 2023) to our recovered MAGs, for samples collected at the same sites (Figure 3). Since the metagenome and metatranscriptome of the Guaymas Basin subsurface remain incompletely covered by sequence data, the absence of transcript read mapping to particular MAGs cannot be taken as evidence of microbial inactivity. Microbial activity of the deep biosphere is certainly constrained but not eliminated by substrate and energy limitation (Hoehler et al., 2013). To avoid these ambiguities that are inherent in negative transcript mapping results, we focus on positive transcript mapping results that support the activity of specific MAGs in the subsurface. Actively transcribed genes are present for MAGs within all phyla discussed here, albeit at variable levels; some MAGs within individual phyla show no or much lower apparent activity than others (Figure 3).

Most transcriptionally active bacterial and archaeal MAGs from warm and hot sediments belong to uncultured lineages previously detected in hydrothermal chimneys, sulfidic springs and seeps, and in Guaymas Basin surficial hydrothermal sediments (Figure 3). Bacterial transcripts from warm sediments were affiliated with four MAGs (GMP_018, GMP_083, GMP_036, and GMP_057) of the Chloroflexota GIF19 lineage, a dominant group in carbonate hydrothermal chimneys (Frouin et al., 2018). Other transcripts from warm sediments mapped to MAG GMP_019 within the dehalogenating Dehalococcoides lineage, to MAGs GMP_007 and GMP_011 within the subsurface Aerophobota, and MAG GMP_58 within the Bipolaricaudota lineage UBA7950, found at the Lost City hydrothermal vents (Brazelton et al., 2022). Archaeal transcripts in warm and hot sediment samples were mapped to MAG GMP_008 within the Hadarchaeota, MAG GMP_075 of the Aenigmatarchaeota QMZP01 lineage from a terrestrial sulfur spring (Hahn et al., 2022), MAG GMP_040l within the thermoproteotal brine pool lineage TCS64 (Zhang et al., 2016), and to three Thermoproteota MAGs GMP_002, GMP_026, and GMP_039 within the B26-1 lineage from Guaymas Basin hydrothermal sediments (He et al., 2016). Transcriptional activity of

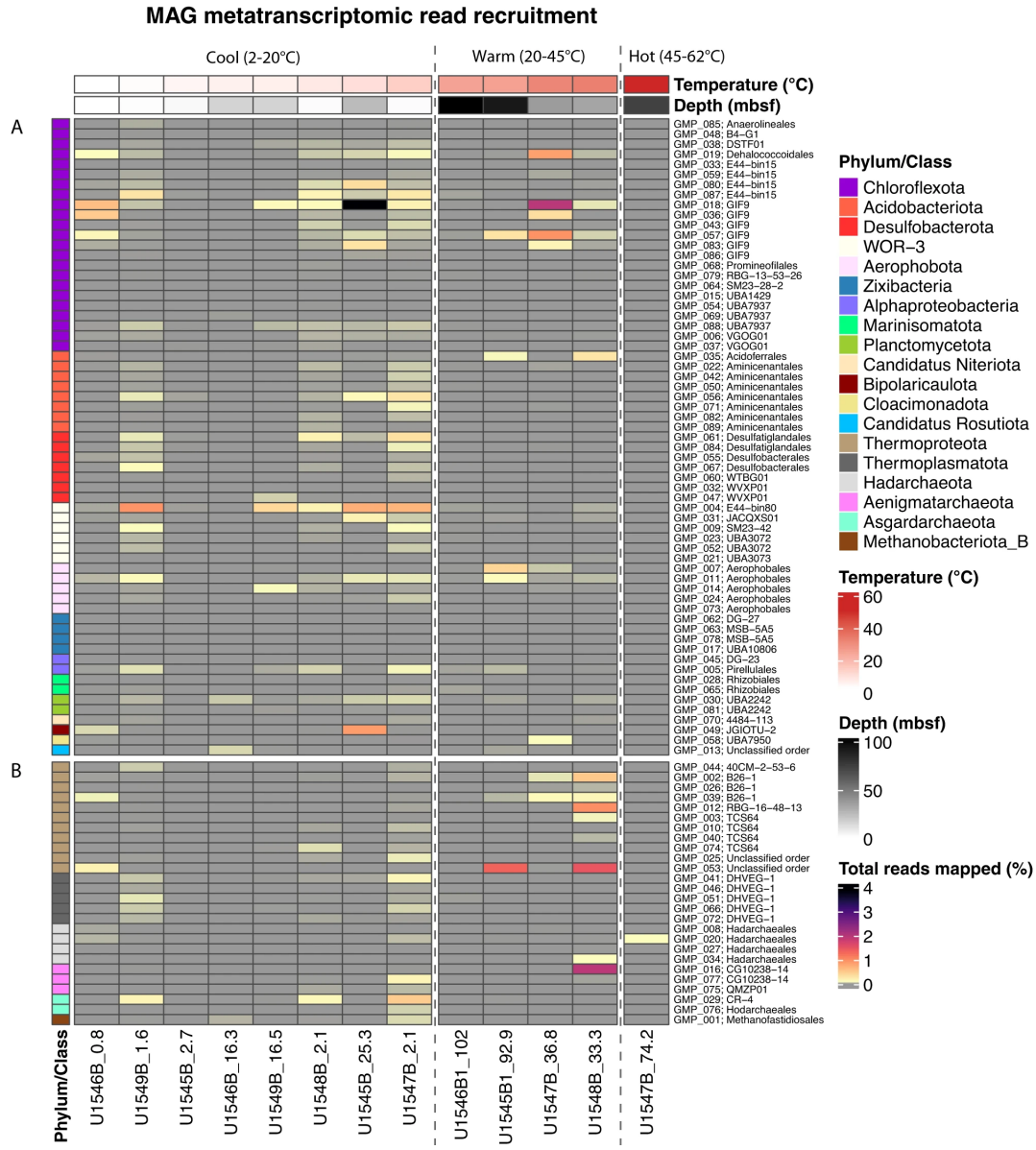


Figure 4-3. Heatmap of MAG Metatranscriptomic read recruitment. Each column shows the percentage of total pre-processed metatranscriptome reads (relative abundance) that mapped to all 89 MAGs, for samples ordered by increasing temperature from left to right on the x-axis (annotated by site numbers and depths in mbsf). Temperature regimes (Cool, Warm, and Hot) are separated by vertical dashed lines. Each row shows the abundance profile of an individual MAG across all samples. MAGs are color-coded by phylum on the left, and annotated by GMP (Guaymas MAG Prokaryote) numbers 001 to 089 and order-level affinity to the right. RKPM, reads mapped per kilobase of genome, per million mapped reads. Panel section **A** denotes bacterial MAGs and panel section **B** denotes archaeal MAGs.

these MAGs suggests their inherent physiological adaptations to warm and reducing habitats are advantageous in the Guaymas Basin subsurface as well.

The influence of environmental factors on MAG composition

The relationship between environmental parameters (Supplementary Dataset 1) and the taxonomic composition of MAGs from cool (2-20°C), warm (20-45°C), and hot temperatures (45-62°C) was investigated using non-metric multidimensional scaling (nMDS) (Figure 4) and Canonical Correspondence Analysis (CCA) (Supplementary Figure 5). In both analyses, bacterial and archaeal MAGs clustered consistently by temperature, comparable to previous analyses of temperature-dependent microbial community composition in surficial sediments in Guaymas Basin (Teske et al., 2021f). In particular, MAGs from hot sediments aligned with temperature as the strongest influencing factor (Figure 4). Total sulfide concentration (H_2S) was aligned with temperature in the CCA plot (Supplementary Figure 5). For samples from warm sediments, nMDS analyses revealed that methane, alkalinity, and dissolved inorganic carbon (DIC) concentrations exerted a significant effect ($p < 0.05$) on the MAG community (Figure 4). For cool sediments, nMDS and CCA showed consistently that MAGs clustered in the direction of total organic carbon (TOC) and total nitrogen (TN) content, CO concentration, pH and salinity (Figure 4, Supplementary Figure 5). The influence of TN and TOC on MAG diversity in cool samples may reflect increased availability of labile sources of dissolved and particulate organic matter in near-surface sediments. The consistent impact of pH and salinity on MAG diversity in cool samples, in both CCA and nMDS analyses, reflects persistent downcore trends towards lower pH and slightly reduced porewater salinity (Supplementary Dataset 1).

To summarize, the environmental parameters that impact MAG composition change downcore, from surface-linked factors such as TN, TOC, pH and salinity that impact MAGs in cool sediments, to biogeochemical parameters reflecting terminal organic matter degradation, such as increasing DIC, alkalinity and methane concentrations, in deeper and warmer sediments. An organic substrate-depleted, DIC- and methane-enriched deep subsurface environment may select for specific phyla or taxa with autotrophic capabilities (e.g., Hadarchaeota). For MAGs from deep and hot samples, carbon or nitrogen substrates, or other chemical factors, become secondary to the impact of temperature itself.

Metagenomic features with wide subsurface distribution

In addition to genes for core metabolic processes (e.g., glycolysis, biosynthesis of nucleotides and amino acids), Guaymas Basin MAGs contain widespread genomic features that extend across

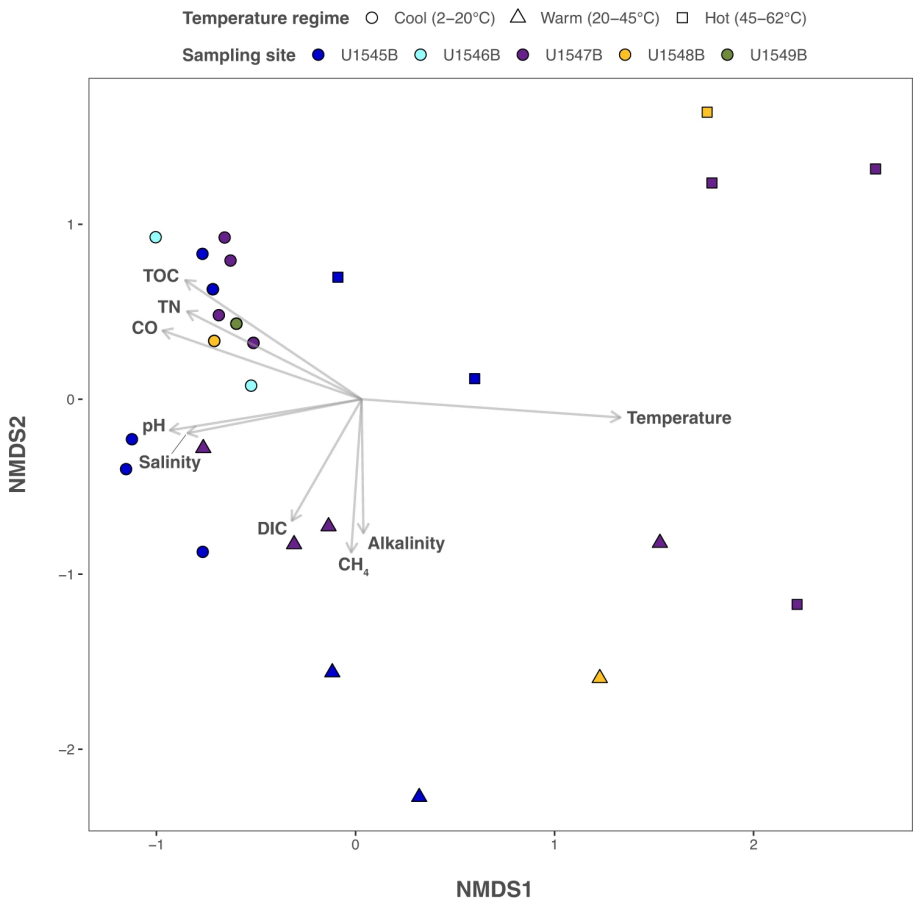


Figure 4-4. Non-metric multidimensional scaling (nMDS) ordination plot of MAGs and environmental parameters. The nMDS plot depicts the correlation of Guaymas Basin MAG occurrence with in-situ environmental parameters (plot stress: 0.106). On the basis of Fisher's method for combining p-values, we show environmental variables with p-values < 0.05 resulting from a two-sided permutation test. The directions of the arrows indicate a positive or negative correlation among the environmental parameters with the ordination axes (temperature, p = 0.0001; pH, p = 0.0041; salinity, p = 0.0119; alkalinity, p = 0.0445; dissolved inorganic carbon (DIC), p = 0.0457; methane (CH4), p = 0.0215; carbon monoxide (CO), p = 0.0011; total organic carbon (TOC), p = 0.0003; total nitrogen (TN), p = 0.0024). Arrow length reflects correlation strength between environmental parameter and MAG occurrence. The samples are color-coded by site, and their temperature regimes are indicated by shape (circles for 2–20 °C; triangles for 20–45 °C and squares for 45–62 °C).

multiple bacterial and archaeal phyla. Some of these widespread genomic features have obvious adaptive value and are thus retained for survival, while others challenge assumptions on subsurface adaptations and evolutionary constraints under subsurface conditions.

Among genes that confer survival advantages, two-component systems (TCSs) can induce metabolic shifts, and are used extensively by bacteria and some archaea to respond and adapt to environmental changes (Beier et al., 2006). Generally, archaea acquire TCS genes through horizontal gene transfer from bacteria (Schaller et al., 2011). In Guaymas Basin, TCS genes occur in the majority of bacterial MAGs but not in archaeal MAGs (Supplementary Datasets 5, 6), and they may help cells to adapt to long term burial. For example, the KinABCDE-Spo0FA system is

present in almost all our bacterial MAGs and plays a role in sporulation by shifting cellular metabolism from active growth to dormancy/sporulation (Quisel et al., 2001). Likewise, the RegB/RegA redox-signaling mechanism involved in carbon fixation, hydrogen oxidation and anaerobic respiration (Elsen et al., 2004) is present in the majority of the bacterial MAGs.

Among widely distributed genes, we found a variety of transporters and efflux pumps associated with microbial defense, and biosynthetic gene clusters involved in synthesis of diverse secondary metabolites (Supplementary Datasets 5, 6). While all bacterial and archaeal MAGs encoded transporters (Chklovski et al., 2023), efflux pumps (found in 67% of all MAGs) included a large proportion of multidrug resistance pumps, detected in, 44% of all MAGs. In 25% of all MAGs, biosynthetic gene clusters were involved in the biosynthesis of diverse secondary metabolites (Supplementary Dataset 7). Archaeal biosynthetic gene clusters were primarily annotated as polyketide synthases, ribosomally synthesized and post-translationally modified peptides. Additionally, archaeal genes encoded the synthesis of terpenes (e.g., geranylgeranyl diphosphate synthase; Supplementary Dataset 6) that can be part of their lipid membranes, or function as pigments, antimicrobial agents and (in plants) as thermoprotectants (Yang et al., 2012).

Genes involved in chemotaxis (*cheA/B/R/W/Y*) and motility (*flgB/C/E/G/H/I* and *fliE/F/G/*) were present in 56% of all bacterial and archaeal MAGs (Supplementary Datasets 5, 6). These findings suggest potential for cell-cell interaction, cell movement and competition for resources in the Guaymas Basin subsurface microbial community – a surprising result given deep biosphere microorganisms are trapped in tight pore spaces that limit movement and interaction (Morono et al., 2020). While cell motility genes are gradually depleted downcore in the marine subsurface (Biddle et al., 2008), they do not disappear. Cell motility and secondary metabolite biosynthesis genes were present and expressed in marine subsurface MAGs from Peru Margin and Canterbury Basin sediments at depths down to 345 mbsf (Pachiadaki et al., 2016).

Our results can be interpreted as evidence that evolution in the deep biosphere proceeds at extremely slow rates. Cells deep below the sediment surface must use available energy to maintain their cellular integrity over possibly geological timeframes while greatly attenuating cell division and genome replication (Hoehler et al., 2013; Morono et al., 2020), unless some physical disturbance or fluid flow returns them to the sediment surface. Under subsurface conditions, attenuated gene loss slows down the impact of selection that gradually shapes the subsurface biosphere (Starnawski et al., 2017). However, the adaptive value for genes of motility and

competition may be reduced with depth but is unlikely to expire entirely, since pore space constraints do not preclude slow microbial movement (millimeters over months), as demonstrated experimentally by gradual recolonization of deep subsurface sediments (Parkes et al., 2000).

Characteristics and distribution of dominant bacterial and archaeal groups

The Guaymas Basin subsurface yields predominantly MAGs affiliated with specific phylum-and order-level lineages that show distinct mesophilic and thermophilic preferences. This suggests lineages appearing at specific depths and temperature ranges respond to environmental factors, which in turn shape their occurrence patterns. Our central working hypothesis is that the Guaymas Basin subsurface community is not a random assemblage, but reveals phylogenetic and functional structure that can be tracked downcore. Our account of this structured community focuses on dominant bacterial and archaeal phyla (Chloroflexota, Thermoproteota, Hadarchaeota); an extended overview on further bacterial and archaeal MAGs is provided in the Supplementary Note.

Dominant subsurface bacteria

Of 63 bacterial MAGs found in the Guaymas Basin subsurface, 23 are members of the phylum Chloroflexota, one of the dominant phyla in marine sediments with metabolically diverse fermentative and dehalogenating lineages (Fincker et al., 2020; Supplementary Note). Within the Guaymas subsurface, Chloroflexota MAGs comprise 12 order-level lineages, and account for a significant fraction of recruited metagenomic reads per sample (up to 8.3%) (Figure 2, Supplementary Figure 6). At site U1545B, Chloroflexota MAGs were widespread within cool samples (2-20°C) and persisted occasionally into deep and warm sediments; at Ringvent site U1547B they were ubiquitous in cool samples but also widely found in warm sediments (20-45°C) (Figure 2, Supplementary Figure 6). MAGs that occur in warm sediments are affiliated with the subsurface and hydrothermal GIF9 group (Frouin et al., 2018; Hug et al., 2013), the VGOG01 lineage from the sulfidic, warm water column of tropical Lake Tanganyika (Tran et al., 2021), and the dehalogenating Dehalococcoidales lineage. In hot sediments above 45°C, Chloroflexota MAGs appear only in traces (Supplementary Figure 6). Thus, the Guaymas Basin subsurface Chloroflexota generally prefer cool or moderately warm habitats, and avoid temperatures above ca. 40°C.

Metagenomes were assembled and annotated for all Chloroflexota in our data sets to gather additional information about their metabolic potential (Supplementary Figure 3, and Supplementary Dataset 8). Using the KEGG framework for functional annotation, within the general category "central carbohydrate metabolism" we find core genes that can participate in the TCA cycle, glycolysis, gluconeogenesis, and the pentose phosphate pathway (Supplementary Figure 3). This category includes one specific module (K0378) that encodes an aldehyde ferrodoxin oxidoreductase (AOR), a tungsten-containing enzyme identified in mesophilic bacteria that can reduce aromatic compounds (Arndt et al., 2019). Within the category of "other carbohydrate metabolism" we find genes affiliated with galactonate/galactose degradation that can be linked to biosynthesis of alkaloids (e.g., terpenoid alkaloids). We also detect modules (assigned as "photorespiration") that are involved in the glycine cleavage system and shared between different amino acid biosynthetic pathways (Supplementary Note). In the general category "metabolic capacity" we detected genes assigned to oxygenic and anoxygenic photosynthesis, nonetheless, these are genes (e.g., pyruvate phosphate dikinase and citryl-CoA lyase) involved in carbon fixation. An expanded KEGG module analysis of the whole community metagenome (Supplementary Dataset 9) reveals many of the same genes, including those within the categories "central carbohydrate metabolism" and "other carbohydrate metabolism", reflecting the dominance of Chloroflexota among the recovered MAGs (Supplementary Figure 4).

Dominant subsurface archaea

Although the archaea contributed only 26 MAGs compared to 63 bacterial MAGs to our total, and represent fewer phylum-level lineages, they exhibit greater thermal range (Figure 2). MAGS of two dominant archaeal phyla – the Thermoproteota (11 MAGs) and the Hadarchaeota (4 MAGs) – prefer warm and hot subsurface sediments, and are introduced here in greater detail; additional archaeal lineages are discussed in the Supplementary Note.

Archaeal MAGs were dominated by the Thermoproteota, an archaeal phylum consisting of four major lineages, the Thaumarchaeota, Aigarchaeota, Korarchaeota and Bathyarchaeia (Oren et al., 2021). All 11 Thermoproteota MAGs belonged to the uncultured class Bathyarchaeia; these were detected at all examined sites but primarily at the Ringvent sites U1547B and U1548B (Figure 2 and Supplementary Figure 6). Order-level identification of bathyarchaeial MAGs reveals linkages to subsurface, seep and hydrothermal sediment habitats. Five MAGs assigned to the

order-level lineages TCS64 and 40CM−2−53−6 were recovered primarily between 0.8-15 mbsf sediments at sites U1545B and U1547B with cool temperatures ranging from 2.8-17.4°C. These bathyarchaeial orders have been reported also from deep sea brine pool samples (Zhang et al., 2016) and from soil samples (Butterfield et al., 2016). The order-level lineage B26-1, previously found in Guaymas Basin hydrothermal sediments (Dombrowski et al., 2018; He et al., 2016), included three MAGs from warmer sediments (19-40°C) below 63.8 mbsf at U1545B, and from warm to hot sediments (24-47°C) between 19.3 and 65.8 mbsf at Ringvent site U1547B. The order-level lineage RBG-16-48-13, recovered previously from terrestrial subsurface cores (Anantharaman et al., 2016), was represented by a MAG detected at site U1548 at 20-45°C (Figure 2). Two bathyarchaeial MAGs could not be classified at the order level, but one of these MAGs was abundant at temperatures between 39.5-47°C at U1547B (Supplementary Figure 6). The detection of bathyarchaeial MAGs over a wide temperature spectrum, and the link of bathyarchaeial orders to specific temperature regimes, suggests distinct thermal preferences among different lineages of Bathyarchaeia (Qi et al., 2021). The ubiquitous presence of Bathyarchaeia in anaerobic sediments (Lloyd et al., 2013), including hydrothermal sediments (He et al., 2016), can be attributed to their capacity to metabolize multiple organic substrates, e.g., polysaccharides, urea, acetate, detrital proteins, and aromatics compounds such as benzoate and lignin (Feng et al., 2019), potential substrates in the hydrocarbon-rich Guaymas Basin subsurface. Based on MAG gene content, Bathyarchaiea can potentially utilize formaldehyde and shuttle it into carbon fixation via the Wood-Ljungdahl pathway (Supplementary Note). Lineage-specific thermophilic adaptations among the Bathyarchaeia include reverse DNA gyrase that facilitates DNA supercoiling under extreme temperatures (Feng et al., 2019).

Hadarchaeota thrive in subsurface sediments by a combination of heterotrophic traits (fermentation of carbohydrates) with autotrophic energy generation, specifically the oxidation of carbon monoxide and hydrogen (Baker et al., 2016). Hadarchaeota were previously recovered from surficial hydrothermal sediments in Guaymas Basin (Dombrowski et al., 2018). Consistently, the 4 hadarchaeotal MAGs (GMP_008, GMP_020, GMP_027, GMP_034) did not recruit any reads from cool samples but only from warm and hot samples, indicating a preference for elevated temperatures (Figure 2). In contrast to changing thermal preferences for MAGs from different bathyarchaeial orders, the Hadarchaeota, originally detected in hot and deep terrestrial aquifers

(Takai et al., 2001), consistently prefer elevated temperatures in deep sediments of the Guaymas Basin subsurface (Figure 2).

Hadarchaeotal genomic features

Abundant and highly expressed hadarchaeotal MAGs were examined for characteristic features in their genomes and transcriptomes. One Hadarchaeota MAG, Guaymas_P_008, recruited ~19% of all metagenomic reads at 74.3 msbf (*in-situ* temperature 51°C) at Ringvent site U1547B (Figures 2, 3). This MAG contained genes for carbohydrate hydrolysis (α-RHA, β-galactosidase) and nucleoside uptake and degradation (nucleoside transporters, purine nucleosidases) that suggest purine/pyrimidine synthesis from nucleosides. This MAG also contained carbon monoxide oxidation genes (*coxM*, *coxS*) that were absent in the other Hadarchaeota MAGs that encoded genes for fermentation (*porA*, *ack*, *acdA*) and aromatics degradation (*ubiX*) (Supplementary Datasets 6, 7). The ability to utilize a wider range of carbohydrates may support higher temperature tolerance, as reported for thermally-adapted Bathyarchaeia genomes (Qi et al., 2021). The potential for hydrocarbon utilization in Hadarachaeota and other phyla (Supplementary Note) might contribute to reduced hydrocarbon concentrations at intermediate sediment depths and temperatures (Supplementary Figure 1). One of our Hadaracheota MAGs (P_034) contained homologs to *mcrC* and *mcrG* that regulate the expression and assembly of the alkyl/methyl coenzyme M reductase operon (Shao et al., 2022), the essential methane and alkane-activating genes in archaeal methanogens, methane oxidizers and short-chain alkane oxidizers (Wang et al., 2021). Finally, we note the presence of KaiC histidine in Hadarchaeota, a circadian clock protein that regulates cell division and allows prokaryotes to adapt to changes in environmental conditions (Jabbur et al., 2022), and the gene for programmed cell death (protein 5) that is linked to anti-virus defense and triggers dormancy under hostile conditions (Koonin et al., 2017).

Genome size trends in the deep biosphere

Comparisons of estimated genome sizes for all MAGs that recruited at least 0.1% of metagenomic reads from cool, warm, and hot sediments revealed a difference in average genome size. The most abundant genomes in cool sediments were on average significantly (two-sided partially overlapping samples *t*-test, adjusted $p < 0.05$) larger (~32%) than those recovered from hot sediments (Figure 5A, B). The estimated genome size of MAGs recovered from our shallow (2-15

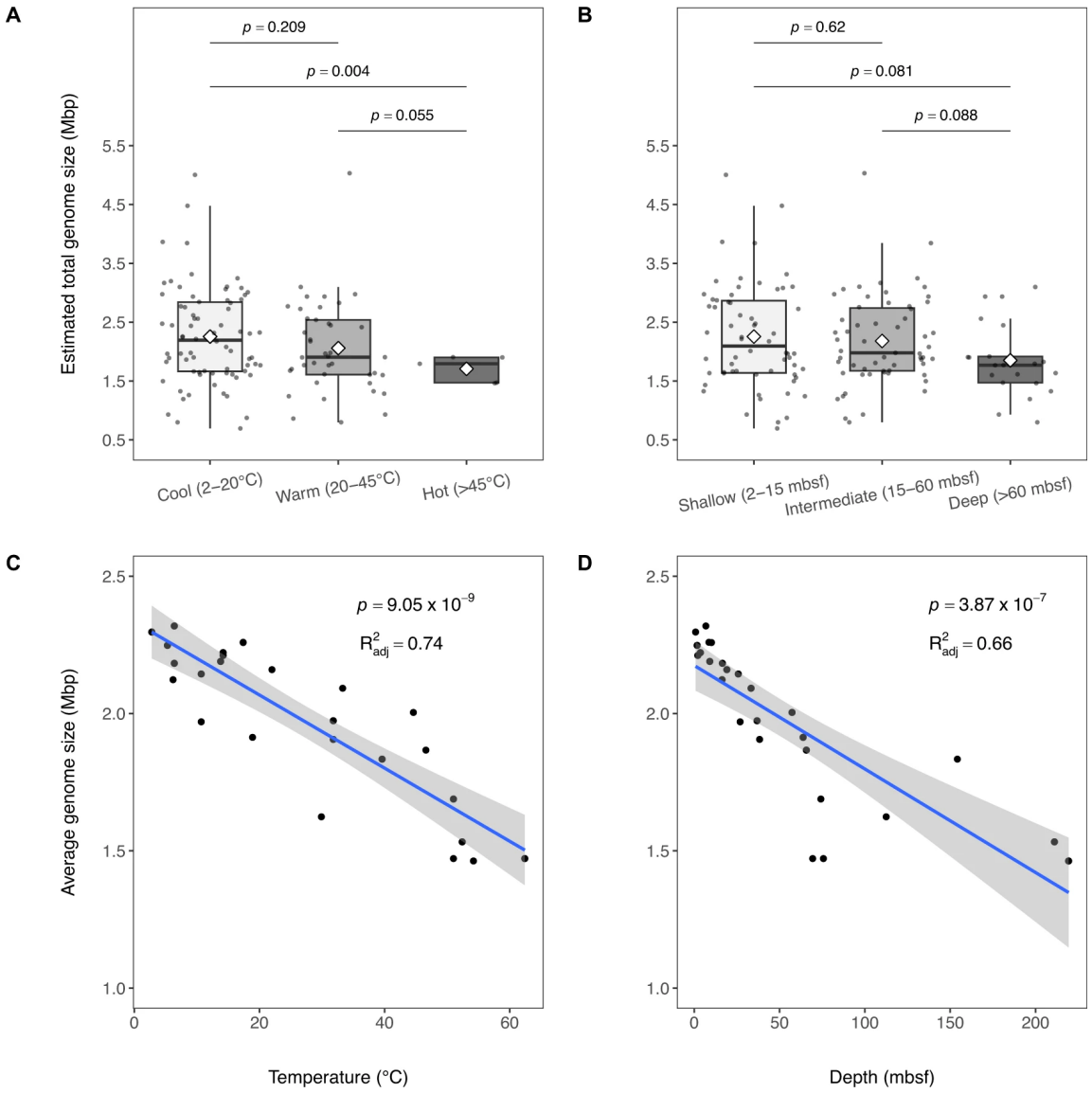


Figure 4-5. Estimated and average genome size vs. temperature and depth. Boxplots show estimated genome size of MAGs that recruited at least 0.1% of metagenomic reads from samples collected in cool (2–20 °C), warm (20–45 °C), and hot sediments (45–60 °C) in panel **A**, and at shallow (2–15 mbsf), intermediate (15-60 mbsf), and deep (>60 mbsf) depths in panel **B**. The Median is shown as the middle horizontal lines, the mean as the white diamonds, and interquartile ranges are shown as boxes (whiskers extend to 1.5 times the interquartile range). Each data point is overlaid on the boxplots and values at the top denote adjusted p-values from two-sided partially overlapping samples t-tests comparing estimated genome sizes by temperature (panel **A**) and depth (panel **B**) regimes. Panels **C** and **D** display the relationship between average estimated genome size in each metagenomic sample plotted against temperature (**C**) and depth (**D**) using linear regression. The blue lines in panels **C** and **D** denote the regression lines, with the fitted values +/− 1.96 standard error indicated by the grey bands. The values at the top of panels **C** and **D** denote the p-value and adjusted R-squared value of the fit.

mbsf) samples was also ~22% larger on average than those detected in deeper (> 60 mbsf) and warmer (>30-40°C) sediments. Linear regression analysis demonstrated a general reduction in average genome size in our samples as both temperature and depth increased (Figure 5C, D). Elevated *in situ* temperatures are thought to select for smaller genome sizes via genome

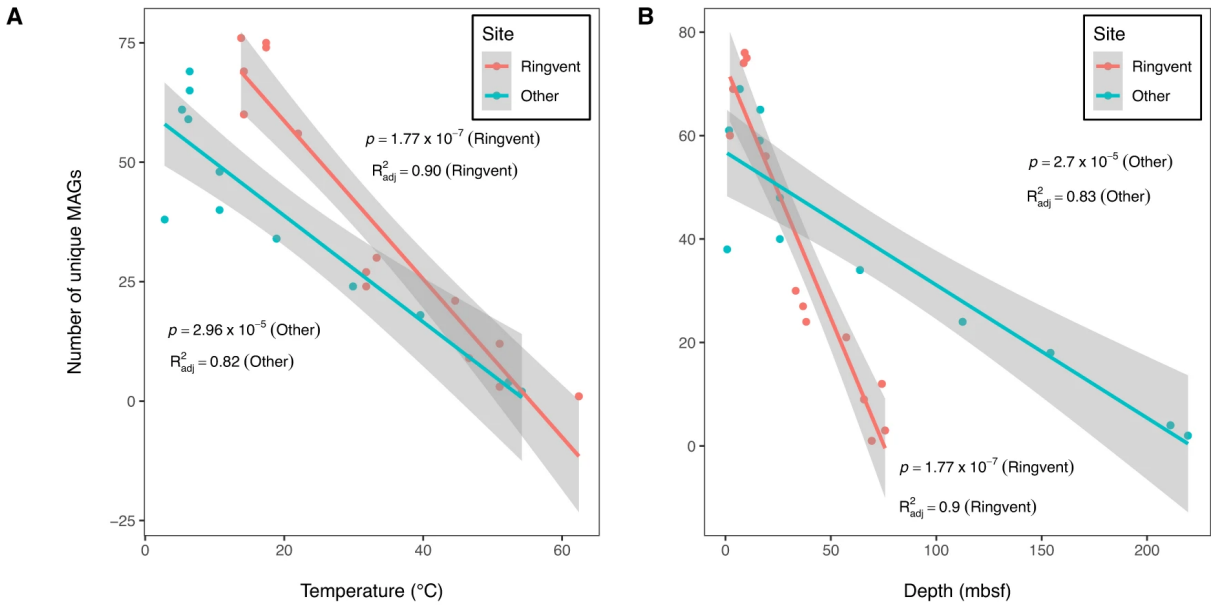


Figure 4-6. MAG recovery at sampled sites vs. temperature and depth. Panels **A** and **B** show the MAGs that recruited metagenomic reads from samples t sites U1547B and U1548B (in red) and sites U1545B, U1546B, and U1549B (in green) plotted against temperature (**A**) and depth (**B**) using best-fit linear regression. The solid lines in panels **A** and **B** denote the regression lines, with the fitted values +/− 1.96 standard error indicated by the grey bands. For both panels **A** and **B**, the p-values and adjusted R-squared values of the fits of each regression line are shown. MAG numbers for all samples (n = 26), and their depths and temperatures are provided in the Source Data file.

streamlining (Sabath et al., 2013), for example increased gene loss after duplication; the effects of genomic streamlining are pervasive and result in the elimination of hundreds of genes all over the genome (Simonsen et al., 2022). Reduced genome size lowers the metabolic cost required for microbial DNA replication, as suggested for hadal microorganisms in the Challenger Deep at Mariana Trench (Zhou et al., 2022). Microbes with smaller genomes would gain a relative survival advantage and gradually dominate the microbial community in the subsurface, as metabolically more demanding microbial community members with large genomes die off. Such a mechanism would contribute to the selection of subsurface-adapted microbial communities that has been documented already within the top few meters below seafloor (Starnawski et al., 2017), and it would explain the small size of microbial cells in deep subsurface sediments, near 0.5 micrometer (Biddle et al., 2006).

Temperature impact on MAG recovery

The environmental stresses that increasingly exclude microbial lineages, reduce genome size and reduce overall microbial population size (and thus, quantity of recovered DNA) are reflected in decreased recovery of MAGs in warmer and deeper samples from all sites (Figure 6A, B). Plotted

against depth, MAG recovery declines more quickly for the two hotter Ringvent sites U1547B and U1548B than for the cooler sites U1545B, U1546B and U1549B (Figure 6B). When plotted against temperature, declining MAG recovery for the hot Ringvent sites and the cooler sites converge towards a shared minimum between ca. 50 and 60°C (Figure 6A). These comparisons show that the decline of MAG recovery with depth is locally modified, behaves differently at different sites, and does not follow a uniform depth-related decay rate. In contrast, the influence of increasing temperature is pervasive, reduces microbial diversity at all sites, and occludes the emergence of MAGs representing new microbial lineages beyond approximately 50-60°C.

Conclusions and outlook

While improved DNA and RNA recovery could potentially compensate for declining downcore cell density, and extend the recovery of new bacterial and archaeal MAGs towards deeper and hotter sediments, the observed trend towards increasingly limited microbial diversity in the subsurface stands in marked contrast to the numerous bacterial and archaeal lineages that thrive in surficial hydrothermal sediments of Guaymas Basin, where fluidized sediments are permeated by pulsating, extremely hot (> 80°C) and highly reducing fluids (Dombrowski et al., 2018). We ascribe the difference to contrasting energy supply (Lagostina et al., 2021), and suggest that relatively moderate temperatures in IODP boreholes have a disproportionally greater impact on the energy-limited microbial deep biosphere, whereas surficial microbial communities that are well-supplied with energy-rich circulating hydrothermal fluids can tolerate high temperatures. The latter conditions select for thermophilic and hyperthermophilic, frequently chemolithoautotrophic bacteria and archaea (Anderson et al., 2015; Reysenbach et al., 2020). We suggest that this difference ultimately results in distinct microbial communities in surficial hydrothermal sites, and in subsurface sediments where dominant bacteria and archaea (Chloroflexota, Thermoproteota, Acidobacteriota, Desulfobacterota) resemble the largely heterotrophic and mesophilic microbiota of non-hydrothermal benthic sediments (Baker et al., 2021; Parkes et al., 2014).

Yet, we note that specific archaea, in particular the Hadarchaeota, show a preference for deep, hot sediments of Guaymas Basin. These archaea extend consistently into the deep sediment column, not only by MAG detection but also in 16S rRNA gene surveys (Mara et al., 2023), and appear to represent deep subsurface thermophiles that are sustained by substrates and energy sources of deep, hot sediments. Observations of microbial cells and activity in extremely deep, hot

subsurface environments (Heuer et al., 2020; Inagaki et al., 2015) could indicate such thermophile communities that have adapted to deep subsurface conditions. Since candidate archaea for deep, hot biosphere communities were consistently detected in the hydrothermally influenced Ringvent sites where a hot volcanic sill is emplaced into organic-rich marine sediments, we extrapolate further that the mineralogically and morphologically complex basalt interface (Teske et al., 2021d) could provide microbial substrates and energy sources (Thiel et al., 2019), calling for further studies.

Methods

Sample collection

Sediment cores were collected during IODP Expedition 385 using the drilling vessel *JOIDES Resolution*. Holes at each site were first advanced using advanced piston coring (APC), then half-length APC, and then extended core barrel (XCB) coring as necessary. Temperature measurements used the advanced piston corer temperature (APCT-3) and Sediment Temperature 2 (SET2) tools (Neumann et al., 2023). Downhole logging conducted after coring used the triple combination and Formation MicroScanner sonic logging tool strings. After bringing core sections onto the core receiving platform of the D/V *JOIDES Resolution*, whole round samples for microbiology were retrieved within ~30 minutes using ethanol-cleaned spatulas. Samples for biogeochemical measurements were obtained and processed shipboard (Teske et al., 2021a). Whole round samples for DNA-based studies were capped with ethanol-sterilized endcaps, transferred to the microbiology laboratory, and stored briefly at 4°C in heat-sealed tri-foil gas-tight laminated bags flushed with nitrogen until processing. Masks, gloves and laboratory coats were worn during sample handling in the laboratory where core samples were transferred from their gas-tight bags onto sterilized foil on the bench surface inside a Table KOACH T 500-F system, which creates an ISO Class I clean air environment (Koken Ltd., Japan). In addition, the bench surface was targeted with a fanless ionizer (Winstat BF2MA, Shishido Electrostatic Co., Ltd., Japan). Within this clean space, the exterior 2 cm of the extruded core section were removed using a sterilized ceramic knife. The core interior was transferred to sterile 50-mL Falcon tubes, labeled, and immediately frozen at -80°C for post cruise analyses. For RNA-based studies, sampling occurred immediately after core retrieval on the core receiving platform by sub-coring with a sterile, cutoff 50cc syringe into

the center of each freshly cut core section targeted. These sub-cores were immediately frozen in liquid nitrogen and stored at -80°C.

DNA extraction and sequencing

DNA was extracted from selected core samples using a FastDNA SPIN Kit for Soil (MP Biomedicals). Up to 5 grams of sediment were processed following a modified manufacturer's protocol (Ramírez et al., 2018). Briefly, each sediment sample was homogenized twice (vs. once that the manufacturer suggests) in Lysing Matrix E tubes for 40 seconds at speed 5.5 m/s, using the MP biomedicals bench top homogenizer equipped with 2 ml tube adaptors. Between the two homogenization rounds the samples were placed on ice for 2 minutes. After the second homogenization the samples were centrifuged at 14,000 x *g* for 5 minutes. For each sample, the supernatant and the top layer of the pellet was transferred to a clean 2 ml tube where proteins were precipitated by the addition of the protein precipitation solution (PPS) provided in the extraction kit. The rest of the extraction protocol followed the manufacturer's recommendations. When parallel extractions were performed, the extracts were pooled and concentrated using EMD 3kDa Amicon Ultra-0.5 ml Centrifugal Filters (Millipore Sigma). A control extraction, in which no sediment was added, was included to account for any laboratory contaminants (Supplementary Materials). All libraries for metagenome sequencing (n =29; 26 samples and 3 controls; Supplementary Data 2) were prepared from genomic DNA extracts that were submitted at the University of Delaware DNA Sequencing & Genotyping Center. Thirteen libraries were sequenced with NovaSeq S4 PE150 (Illumina) at the University of California, Davis Genome Center, and thirteen libraries were sequenced with NextSeq550 (Illumina) at the University of Delaware DNA Sequencing & Genotyping Center. Metagenome sequence reads were deposited to the National Center for Biotechnology Information Sequence Read Archive under access numbers SRR23614663-23614677 and SRR22580794-SRR22580807 (Bioproject PRJNA909197).

Metagenomic co-assembly, binning, dereplication and taxonomic assignment

Metagenomic reads originating from adjacent regions (such as adjacent depths targeted in this study) are likely to overlap in the sequence space, increasing the mean coverage and extent of reconstruction of MAGs when using a co-assembly approach. Before assembly, reads were trimmed for quality and adapters removed using Trimmomatic v0.39 (Bolger et al., 2014;

parameters: leading:20; trailing:20; sliding window: 0-24; min length 50). The quality of reads was verified with FastQC v0.11.9 (Andrews et al., 2012). For MAG reconstructions, we used the trimmed reads of metagenomic datasets from all 29 Guaymas samples sequenced in this study (Supplementary Data 2). The 26 metagenomes were co-assembled into contigs using MEGAHIT 1.2.9 (Li et al., 2016) with default parameters. For determining non-redundant MAGs, assembled contigs were binned using three different binners, MetaBAT2 2.12.183 (Kang et al., 2015), MaxBin2 2.2.7 (Wu et al., 2016), as well as CONCOCT 1.1.0 (Alneberg et al., 2014). Output bins from all three binning algorithms were refined and dereplicated using DAS Tool 1.1.6 (Sieber et al., 2018). DasTool determines a unique MAG through a single-copy gene (SCG) scoring strategy coupled to an iterative bin de-replication procedure that produces the highest-scoring set of non-redundant bins (in terms of SGC completeness/contamination) from input bins generated by different binners. Completeness, size, and contamination levels of the reconstructed genomes were estimated using CheckM2 1.0.0 (Chklovski et al., 2023). Only MAGs that were at least 50% complete and contained less than 10% contamination were used for downstream analyses (Supplementary Data 4). The taxonomic placement of the MAGs was performed with GTDB-Tk 2.1.0 (Chaumeil et al., 2020).

To account for seawater and laboratory contamination (Supplementary Note), control samples (Supplementary Data 2) identified MAGs of lab/control contaminants, including Patescibacteria (Paceibacteria, Microgenomatia), Actinobacteriota (Actinomycetia, Humimicrobia), Gammaproteobacteria (Pseudomonadales, Burkholderiales), and Firmicutes (Staphylococcales); these were removed from downstream analyses (Supplementary Data 3).

Calculation of MAG relative abundances

Metagenomic reads from 26 samples were mapped to each MAG using the CoverM 0.6.1 (https://github.com/wwood/CoverM) command line tool with the BWA 2.0 aligner (Vasimuddin et al., 2019). The CoverM tool automatically concatenated all the MAGs into a single file, and metagenomic reads were recruited to MAG contigs, setting the parameter --min-read-percent-identity to 95 and --min-read-aligned-percent to 50. The “Relative Abundance” CoverM method on the “genome” setting was used to calculate the percent of total metagenomic reads per sample that mapped to each of the 89 MAGs. A custom R script was utilized to concatenate all CoverM output files into a single file in a matrix format (with each sample representing a column and each

row representing total percent of DNA-Seq reads per sample that mapped to a MAG) and was used for heatmap plotting.

Gene annotation, and prediction of KEGG metabolic module presence/absence using MetaPathPredict

Genes were called for all MAGs using Prodigal 2.6.3 (Hyatt et al., 2010) and then annotated using Prokka 1.14.6 (Seeman 2014), KofamScan 1.3.0 (Aramaki et al., 2020), and METABOLIC 4.0 (Zhou et al., 2022) using default settings. The KofamScan annotations were used to assign KEGG annotations to KEGG modules to give broad overview of the metabolisms present in the genomes recovered from Guaymas Basin. The associated script was used to generate Supplementary Figures 3 and 4, and Supplementary Data 5. To present additional data on Chloroflexota, Thermoproteota, Acidobacteriota, Desulfobacterota, Aerophobota, and WOR-3 in a user-friendly format, METABOLIC was used to annotate MAGs to identify putatative metabolisms that we predicted would be present in Guaymas Basin. The associated script was used to generate Supplementary Data 6b. It is recognized that different databases used by the different tools can provide slightly different information. KEGG modules for bacterial MAGs were reconstructed using gene annotations from the KofamScan 1.3.0 command line tool, and the presence or absence of incomplete modules in the genomes was predicted using MetaPathPredict 1.0.0 (Geller-McGrath et al., 2022) with default settings. MetaPathPredict cannot yet be applied to archaeal MAGs. Briefly, Prodigal was used to call genes, and KofamScan was used to annotate them. Gene annotations were generated for predicted genes from bacterial MAGs, and were used as input to MetaPathPredict, which generated predictions for the presence or absence of KEGG modules based on the gene annotations of all bacterial MAGs.

CCA and nMDS analyses of metagenomic abundance datasets and associated environmental parameters

The abundances of metagenomic reads mapped to MAGs were normalized using the “transcripts per million” normalization method (Wagner et al., 2012) with the read mapping "counts" output from coverM (https://github.com/wwood/CoverM). The abundance data were analyzed using canonical correlation analysis (CCA) as well as non-metric multidimensional scaling (nMDS) and were fitted with the environmental parameters in Supplementary Data 1 using R (R Core Team

2018). The cca and metaMDS functions were used for CCA and nMDS analyses, respectively, as well as the envfit function from the vegan 2.6-4 package (Dixon 2003). The results were plotted using ggplot2 3.3.6 (Wickham et al., 2016) with sample shapes corresponding to temperature regime.

Estimated genome size analysis

The estimated genome size of all 89 MAGs was calculated by dividing the MAG assembly size (total base pair length of the MAG) by the fractional CheckM2 completeness of the MAG (the default CheckM2 completeness output divided by 100; a number between 0 and 1). Difference in genome size distributions for MAGs that recruited at least 0.1% of metagenomic reads from samples across temperature (cool [2-20°C], warm [20-45°C], hot [45-62°C]) and depth (shallow [2-15 mbsf], intermediate [15-60 mbsf], deep [>60 mbsf]) regimes was assessed using the two-sided partially overlapping samples *t*-test (Derrick et al., 2017), and resulting *p*-values were adjusted for multiple comparisons via Benjamini-Hochberg correction. The average estimated genome size of MAGs that recruited at least 0.1% of reads from metagenomic samples (n = 26) was fitted using linear regression against temperature and depth measurements affiliated with the samples.

MAG recovery at sampled sites versus temperature and depth

The number of non-redundant MAGs that recruited at least 0.1% of reads from metagenomic samples (n = 26) was fitted using linear regression against temperature and depth measurements affiliated with the samples. Temperature values were interpolated for each sample using linear regression of the local thermal gradient (°C/m) multiplied by depth (mbsf), plus the y-axis intercept: U1545B, T = 0.225 x depth + 4.899; U1546B, T = 0.221 x depth + 2.627; U1547B, T = 0.511 x depth + 13.01; U1548B, T = 0.804 x depth + 6.499; U1549A/B, T = 0.194 x depth + 3.532.

Scanning of MAGs for secondary metabolite biosynthetic gene clusters

All 89 MAGs were individually scanned for secondary metabolic biosynthetic gene clusters using antiSMASH 6.0 (Blin et al., 2021) with default parameters. Resulting gene cluster prediction results (in GenBank format) were parsed and their gene content was analyzed. Clusters with a total

length less than 5kb were discarded from downstream analysis to minimize the inclusion of fragmented biosynthetic clusters in the analysis.

RNA extraction, library preparation, sequencing, and mapping of RNA reads to the MAGs

RNA was extracted from 19 sediment samples from sites U1545B-U1552B and a blank sample (control) using the RNeasy PowerSoil Total RNA Kit (Qiagen) following the manufacturer's protocol with modifications which are discussed below. RNA samples were prepared from samples spanning the depths 0.8 to 101.9 mbsf. All samples, including a blank control, were first washed twice with absolute ethanol (200 proof; purity ≥ 99.5%), and sterile DEPC water (once) to reduce hydrocarbons and other inhibitory elements that otherwise resulted in low RNA yield. In brief, 13-15 grams of frozen sediments were transferred into UV-sterilized 50 ml Falcon tubes (RNAase/DNase free) using clean, autoclaved and ethanol-washed metallic spatulas. Each sample transferred into the 50 ml Falcon tube received an equal volume of absolute ethanol and was shaken manually for 2 min followed by 30 seconds of vortexing at full speed to create a slurry. Samples were spun in an Eppendorf centrifuge (5810R) for 2 minutes at 2000 x *g*. The supernatant was decanted and after the second wash with absolute ethanol, an equal volume of DEPC water was added into each sample and samples were spun for 2 minutes at 2000 x *g*. The supernatant was decanted, and each sediment sample was immediately divided into three 15 mL Falcon tubes containing beads provided in the PowerSoil Total RNA Isolation Kit (Qiagen). The RNA extraction protocol was followed as suggested by the manufacturer, with the modification that the RNA extracted from the three aliquots was pooled into one RNA collection column. All steps were performed in a UV-sterilized clean hood equipped with HEPA filters. Surfaces inside the hood and pipettes were thoroughly cleaned with RNase AWAY™ (Thermo Scientific™) before every RNA extraction and in between extraction steps.

Trace DNA contaminants were removed from RNA extracts using TURBO DNase (Thermo Fisher Scientific) and the manufacturer's protocol. Removal of DNA was confirmed by negative PCR reactions using the bacterial primers BACT1369F/PROK1541R (F: 5'CGGTGAATACGTTCYCGG 3', R: 5'AAGGAGGTGATCCRGCCGCA 3') targeting the 16S rRNA gene (Suzuki et al., 2000). Each 25 µl PCR reaction was prepared using 0.5 U μl^{-1} GoTaq® G2 Flexi DNA Polymerase (Promega), 1X Colorless GoTaq® Flexi Buffer, 2.5 mM $MgCl_2$, (Promega) 0.4 mM dNTP Mix (Promega), 4 µM of each primer (final concentrations), and DEPC

water. PCR reactions used an Eppendorf Mastercycler Pro S Vapoprotect (Model 6321) thermocycler with following conditions: 94°C for 5 min, followed by 35 cycles of 94°C (30 s), 55°C (30 s), and 72°C (45 s). The PCR products were run in 2% agarose gels (Low-EEO/Multi-Purpose/Molecular Biology Grade Fisher BioReagents™) to confirm absence of DNA amplification. Amplified cDNAs from the DNA-free RNA extracts were prepared using the Ovation RNA-Seq System V2 (Tecan) following manufacturer's suggestions. All steps through cDNA preparation were completed the same day to avoid freeze/thaw cycles. cDNAs were submitted to the Georgia Genomics and Bioinformatics Core for sequencing using NextSeq 500 PE 150 High Output (Illumina). The cDNA library generated from our control did not contain detectable DNA. It was nonetheless submitted for sequencing, but it failed to generate any sequences that met the minimum length criterion of 300-400 base pairs.

Reads from the 13 metatranscriptome samples collected from sites that metagenomic samples were taken from were mapped to each MAG using the CoverM 0.6.1 (https://github.com/wwood/CoverM) command line tool with the BWA 2.0 aligner (Vasimuddin et al., 2019). The CoverM tool automatically concatenated all the MAGs into a single file, and metatranscriptome reads were recruited to MAG contigs, setting the parameter --min-read-percent-identity to 95 and --min-read-aligned-percent to 50. A custom R script was utilized to concatenate all coverM output files into a single file in a matrix format, with each sample representing a column and each row representing total percent of RNA-Seq reads per sample that mapped to a MAG. The output was used in this study for heatmap plotting to examine evidence for activity of the taxa for which we recovered MAGs. Metatranscriptome reads were deposited to the National Center for Biotechnology Information Sequence Read Archive under accession numbers SRR22580929-SRR22580947 (Bioproject PRJNA909197).

Cell counts

The sediment sampling for cell counts occurred immediately after core retrieval on the core receiving platform by sub-coring with a sterile, tip-cut 2.5 cc syringe from the center of each freshly cut core section. Approximately 2 cm^3 sub-cores were immediately put into tubes containing fixation solution consisting of 8 mL of 3xPBS (Gibco™ PBS, pH 7.4, Fischer) and 5% (v/v) neutralized formalin (Thermo Scientific™ Shandon™ Formal-Fixx™ Neutral Buffered Formalin). If necessary, the mixture was stored at 4°C.

Fixed cells were separated from the slurry using ultrasonication and density gradient centrifugation (Morono et al., 2013). For cell detachment, a 1 mL aliquot of the formalin-fixed sediment slurry was amended with 1.4 mL of 2.5% NaCl, 300 μL of pure methanol, and 300 μL of detergent mix (Kallmeyer et al., 2008), 100 mM ethylenediamine tetraacetic acid [EDTA], 100 mM sodium pyrophosphate, 1% [v/v] Tween-80). The mixture was thoroughly shaken for 60 min (Shake Master, Bio Medical Science, Japan), and subsequently sonicated at 160 W for 30 s for 10 cycles (Bioruptor UCD-250HSA; Cosmo Bio, Japan). The detached cells were recovered by centrifugation based on the density difference of microbial cells and sediment particles, which allows collection of microbial cells in a low-density layer. The sample was transferred onto a set of four density layers composed of 30% Nycodenz (1.15 g cm^{-3}), 50% Nycodenz (1.25 g cm^{-3}), 80% Nycodenz (1.42 g cm^{-3}), and 67% sodium polytungstate (2.08 g cm^{-3}). Cells and sediment particles were separated by centrifugation at 10,000 × *g* for 1 h at 25°C. The light density layer was collected using a 20G needle syringe. The heavy fraction, including precipitated sediment particles, was resuspended with 5 mL of 2.5% NaCl, and centrifuged at 5000 × *g* for 15 min at 25°C. The supernatant was combined with the previously recovered light density fraction. With the remaining sediment pellet, the density separation was repeated. The sediment was resuspended using 2.1 mL of 2.5% NaCl, 300 μL of methanol, and 300 μL of detergent mix and shaken at 500 rpm for 60 min at 25°C, before the slurry sample was transferred into a fresh centrifugation tube where it was layered onto another density gradient and separated by centrifugation just as before. The light density layer was collected using a 20G needle syringe, and combined with the previously collected light density fraction and supernatant to form a single suspension for cell counting.

For cell enumeration, a 50%-aliquot of the collected cell suspension was passed through a 0.22-μm polycarbonate membrane filter. Cells on the membrane filter were treated with SYBR Green I nucleic acid staining solution (1/40 of the stock concentration of SYBR Green I diluted in Tris-EDTA [TE] buffer). The number of SYBR Green I– stained cells were enumerated either by direct microscopic counts (Inagaki et al., 2015) or image-based discriminative counts (Morono et al., 2009). For image-based discriminative counting, the Count Nuclei function of the MetaMorph software (Molecular Devices) was used to detect and enumerate microbial cells.

Data availability

The raw metagenome and metatranscriptome sequence data generated in this study have been deposited in the NCBI GenBank database under the Bioproject accession number PRJNA909197 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA909197). Metatranscriptome reads are deposited under the accession numbers SRR22580929-SRR22580947. Metagenome reads are deposited under the accession numbers SRR22580794-SRR2258807 and SRR23614663-SRR236114677. Biogeochemical and thermal shipboard data for all IODP385 sites discussed in this study (U1545-U1552) are publicly available on the IODP Expedition 385 online report (http://publications.iodp.org/proceedings/385/385title.html). Shipboard data can be downloaded for each drilling site individually, as numbered excel tables. Post-cruise geochemical data sets (DIC, TOC, TN, hydrocarbons) have been submitted to the Biological and Chemical Oceanography database (BCO-DMO) and are publicly available under project number 833856 (https://www.bco-dmo.org/project/833856). Publicly available datasets used in this study include the CheckM2 database (https://zenodo.org/record/4626519), the GTDB-Tk database release R214 (https://ecogenomics.github.io/GTDBTk/installing/index.html), the KOfam database (ftp://ftp.genome.jp/pub/db/kofam/), the METABOLIC database (https://github.com/AnantharamanLab/METABOLIC), the MEROPS database (https://www.ebi.ac.uk/merops/download_list.shtml), the dbCAN2 database (http://bcb.unl.edu/dbCAN2/download/Databases/dbCAN-old@UGA/dbCAN-fam-HMMs.txt), ISfinder database (https://isfinder.biotoul.fr/), NCBI Bacterial Antimicrobial Resistance database (https://www.ncbi.nlm.nih.gov/bioproject/313047), UniProtKB (SwissProt) database (https://www.uniprot.org/uniprot/?query=reviewed:yes), Prokka databases (https://github.com/tseemann/prokka) and the antiSMASH 6.0 databases (https://dl.secondarymetabolites.org/releases/).

Code availability

All custom scripts used for data analysis and figure creation are available in the GitHub repository at https://github.com/d-mcgrath/guaymas_basin (Geller-McGrath et al., 2023).

Source Data

Links to all datasets used in this analysis, including source data used to generate the figures in this chapter can be found here: https://www.nature.com/articles/s41467-023-43296-x.

References

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11, 1144-1146.

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., Thomas, B. C., Singh, A., Wilkins, M. J., & Karaoz, U. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*(1), 13219.

Anantharaman, K., Hausmann, B., Jungbluth, S.P., Kantor, R.S., Lavy, A., Warren, L.A., Rappe, M.S., Pester, M., Loy, A., Thomas, B.C., and Banfield, L.F. (2018) Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *ISME J* 12, 1715–1728.

Anderson, R. E., Sogin, M. L., & Baross, J. A. (2015). Biogeography and ecology of the rare and abundant microbial lineages in deep-sea hydrothermal vents. *FEMS Microbiology Ecology*, *91*(1), 1–11.

Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., & Wingett, S. (2012). FastQC: a quality control tool for high throughout sequence data. Babraham Institute, UK. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), 2251–2252.

Arndt, F., Schmitt, G., Winiarska, A., Saft., M., Seubert, A., Kahnt, J., Heider, J. 2019. Characterization of an Aldehyde Oxidoreductase from the mesophilic bacterium *Aromatoleum aromaticum* EbN1, a member of a new subfamily of tungsten-containing enzymes. *Frontiers in Physiology and Metabolism* 10, 71.doi:10.3389/fmicb.2019.00071

Baker, B. J., Appler, K. E., & Gong, X. (2021). New microbial biodiversity in marine sediments. *Annual Review of Marine Science*, *13*, 161–175.

Baker, B. J., Saw, J. H., Lind, A. E., Lazar, C. S., Hinrichs, K.-U., Teske, A. P., & Ettema, T. J. G. (2016). Genomic inference of the metabolism of cosmopolitan subsurface Archaea, Hadesarchaea. *Nature Microbiology*, *1*(3), 1–9.

Beier, D., & Gross, R. (2006). Regulation of bacterial virulence by two-component systems. *Current Opinion in Microbiology*, *9*(2), 143–152.

Beulig, F., Schubert, F., Adhikari, R.R. *et al.,* (2022). Rapid metabolism fosters microbial survival in the deep, hot subseafloor biosphere. *Nat Commun* **13**, 312. https://doi.org/10.1038/s41467-021-27802-7

Biddle, J. F., Fitz-Gibbon, S., Schuster, S. C., Brenchley, J. E., & House, C. H. (2008). Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *Proc. Nat. Acad. Sci. USA*, 105:10583–10588.

Biddle, J. F., Lipp, J. S., Lever, M. A., Lloyd, K. G., Sørensen, K. B., Anderson, R., Fredricks, H. F., Elvert, M., Kelly, T. J., Schrag, D. P., Sogin, M. L., Brenchley, J. E., Teske, A.,

House, C.H., &. Hinrichs, K.-U. (2006). Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proc. Nat. Acad. Sci. USA 103*(10), 3846–3851.
Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*, 49, W29-W35.
Bojanova, D.P., De Anda, V.Y., Haghnegahdar, M.A., Teske, A.P., Ash, J.L., Young, E.D., Baker, B.J., LaRowe, D.E., Amend, J.P. (2023). Well-hidden Methanogenesis in deep, organic-rich sediments of Guaymas Basin, Gulf of California. *ISME J.,* doi:10.1038/s41396-023-01485-y.
Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15), 2114–2120.
Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., & Eloe-Fadrosh, E. A. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725–731.
Brazelton W.J., McGonigle, J.M., Motamedi, S. Pendleton, H. L., Twing, K.I., Miller, B.C., Lowe, W.J., Hoffman, A.M., Prator, C.A. Chadwick, G.L., Anderson, R.E., Thomas, E., Butterfield, D.A., Aquino, K.A., Früh-Green, G.L. Schrenk, M.O., Lang, S.Q. (2022). Metabolic strategies shared by basement residents of the Lost City hydrothermal field. *Appl. Environ. Microbiol.* 88:17, doi: 10.1128/aem.00929-22
Butterfield, C. N., Li, Z., Andeer, P. F., Spaulding, S., Thomas, B. C., Singh, A., Hettich, R. L., Suttle, K. B., Probst, A. J., & Tringe, S. G. (2016). Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ*, *4*, e2687.
Chatterjee, M., Fan, Y., Cao, F., Jones, A. A., Pilloni, G., & Zhang, X. (2021). Proteomic study of *Desulfovibrio ferrophilus* IS5 reveals overexpressed extracellular multi-heme cytochrome associated with severe microbiologically influenced corrosion. *Scientific Reports*, *11*(1), 1–11.
Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Oxford University Press.
Chklovski, A., Parks, D. H., Woodcroft, B. J., & Tyson, G. W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nature* Methods 20(8):1203-1212
Derrick, B., Toher, D., & White, P. (2017). How to compare the means of two samples that include paired observations and independent observations: A companion to Derrick, Russ, Toher and White (2017). *The Quantitative Methods for Psychology*, 13(2), 120–126.
Dixon, P. (2003). VEGAN, a package of R functions for community ecology. *Journal of Vegetation Science*, *14*(6), 927–930.
Dombrowski, N., Teske, A. P., & Baker, B. J. (2018). Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nature Communications*, *9*(1), 4999.
Elsen, S., Swem, L. R., Swem, D. L., & Bauer, C. E. (2004). RegB/RegA, a highly conserved redox-responding global two-component regulatory system. *Microbiology and Molecular Biology Reviews*, *68*(2), 263–279.

Feng, X., Wang, Y., Zubin, R., & Wang, F. (2019). Core metabolic features and hot origin of Bathyarchaeota. *Engineering*, *5*(3), 498–504.
Fincker, M., Huber, L.A., Orphan, V.J., Rappe, M.S., Teske, A., and Spormann, A.M. (2020) Metabolic strategies of marine subseafloor Chloroflexi inferred from genome reconstructions. *Environ. Microbiol*. 22(8), 3188-3203.
Frouin, E., Bes, M., Ollivier, B., Quéméneur, M., Postec, A., Debroas, D., Armougom, F., & Erauso, G. (2018). Diversity of rare and abundant prokaryotic phylotypes in the Prony Hydrothermal Field and comparison with other serpentinite-hosted ecosystems. *Frontiers in Microbiology*, *9*, 102.
Garber, A. I., Nealson, K. H., Okamoto, A., McAllister, S. M., Chan, C. S., Barco, R. A., & Merino, N. (2020). FeGenie: a comprehensive tool for the identification of iron genes and iron gene neighborhoods in genome and metagenome assemblies. *Frontiers in Microbiology*, 11:37. DOI=10.3389/fmicb.2020.00037
Geller-McGrath, D. Metagenomic profiles of Archaea and Bacteria within thermal and geochemical gradients of the Guaymas Basin deep subsurface. Zenodo https://zenodo.org/record/8422630 (2023).
Geller-McGrath, D., Konwar, K., Edgcomb, V. P., Pachiadaki, M., Roddy, J., Wheeler, T., & McDermott, J. E. (2022). MetaPathPredict: A machine learning-based tool for predicting metabolic modules in incomplete bacterial genomes. *BioRxiv*, 2022.
Hahn, C.R., Farag, I.F., Murphy, C.L. Podar, M., Elshahad, M.S., Youssef, N.H. 2022. Microbial diversity and sulfur cycling in an early Earth analogue: From novelty to modern commonality. mBio 13:2, doi: 10.1128/mbio.00016-22
He, Y., Li, M., Perumal, V., Feng, X., Fang, J., Xie, J., Sievert, S. M., & Wang, F. (2016). Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nature Microbiology*, *1*(6), 1–9.
Heuer, V., Lever, M.A., Morono, Y., and Teske A. 2019. The limits of life and the biosphere in Earth's interior. In: *Special Issue on Scientific Ocean Drilling: looking to the future.* Oceanography 32:208-211
Heuer, V.B. et al., (2020). Temperature limits to deep subseafloor life in the Nankai Trough subduction zone. *Science* 370, 1230–1234.
Hoehler, T. M., & Jørgensen, B. B. (2013). Microbial life under extreme energy limitation. *Nature Reviews Microbiology*, *11*(2), 83–94.
Hug, L. A., Castelle, C. J., Wrighton, K. C., Thomas, B. C., Sharon, I., Frischkorn, K. R., Williams, K. H., Tringe, S. G., & Banfield, J. F. (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome*, *1*(1), 1–17.
Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 1–11.
Inagaki, F. et al., (2015). Exploring deep microbial life in coal-bearing sediment down to ~2.5 km below the ocean floor. *Science* 349, 420-424.
Jabbur, M.L. and Johnson, C.H. (2022). Spectres of clock evolutions: Past, Present and Yet to come. *Frontiers in Physiology*, 12:815847; doi: 10.3389/fphys.2021.815847
Kallmeyer, J., Smith, D.C., Spivack, A.J., & D'Hondt, S. (2008). New cell extraction procedure applied to deep subsurface sediments. *Limnology & Oceanography Meth.* 6, 236-245.

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

Kirkpatrick, J.B., Walsh, E.A., and D'Hondt, S. (2019). Microbial selection and survival in subseafloor sediment. *Frontiers in Microbiology* 10:956, doi: 10.3389/fmicb.2019.00956

Koonin, E.V., Makarova, K.S., Wolf, Y.I. (2017). Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu Rev Microbiol*. 8, 71:233-261. doi: 10.1146/annurev-micro-090816-093830.

Lagostina, L., Frandsen, S., MacGregor, B., Glombitza, C., Deng, L., Fiskal, A., Li, J., Doll, M., Geilert, S., Schmidt, M., Scholz, F., Bernasconi, S.M., Jørgensen, B.B., Hensen, C., Teske, A., and Lever, M.A. (2021). Interactions between temperature and energy supply drive microbial communities in hydrothermal sediment**.** *Communications Biology* 4:1006.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., & Lam, T.-W. (2016). MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, *102*, 3–11.

Lizarralde, D., Teske, A., Höfig, T. W., González-Fernández, A., & Scientists, I. E. 385. (2023). Carbon released by sill intrusion into young sediments measured through scientific drilling. *Geology*, *51*(4), 329–333.

Lloyd, K. G., Schreiber, L., Petersen, D. G., Kjeldsen, K. U., Lever, M. A., Steen, A. D., Stepanauskas, R., Richter, M., Kleindienst, S., Lenk, S., Schramm, A., & Jørgensen, B.B. (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature*, *496*(7444), 215–218.

Mara, P., Zhou, Y., Teske, A., Morono, Y., Beaudoin, D., Edgcomb, V.P. (2023). Microbial gene expression in Guaymas Basin subsurface sediments responds to hydrothermal stress and energy limitation. *ISME J*., doi: /10.1038/s41396-023-01492-z.

Morono, Y., Ito, M., Hoshino, T. Terada, T., Hori, T., Ikehara, M., D'Hondt, S. and Inagaki, F. (2020) Aerobic microbial life persists in oxic marine sediment as old as 101.5 million years. *Nat Commun* 11, 3626.

Morono, Y., Terada, T., Kallmeyer, J., *&* Inagaki, F. (2013). An improved cell separation technique for marine subsurface sediments: applications for high-throughput analysis using flow cytometry and cell sorting. *Environ. Microbiol*. 15, 2841-2849.

Morono, Y., Terada, T., Masui, N., & Inagaki, F. (2009). Discriminative detection and enumeration of microbial life in marine subsurface sediments. *ISME J* 3, 503–511.

Neumann, F., Negrete-Aranda, R., Harris, R. N., Contreras, J., Galerne, C. Y., Peña-Salinas, M. S., Spelz, R. M., Teske, A., Lizarralde, D., & Höfig, T. W. (2023). Heat flow and thermal regime in the Guaymas Basin, Gulf of California: Estimates of conductive and advective heat transport. *Basin Research,* 35:1308-1328.

Oremland, R.S., Culbertson, C., and Simoneit, B.R.T., (1982). Methanogenic activity in sediment from Leg 64, Gulf of California. In Curray, J.R., Moore, D.G., et al., Initial Reports of the Deep-Sea Drilling Project, 64: Washington, DC (U.S. Government Printing Office), 759–762. https://doi.org/10.2973/dsdp.proc.64.122.1982

Oren, A., & Garrity, G. M. (2021). Valid publication of the names of forty-two phyla of prokaryotes. *Int. J. System. Evol. Microbiol*. *71*(10), 005056.

Pachiadaki, M. G., Rédou, V., Beaudoin, D. J., Burgaud, G., & Edgcomb, V. P. (2016). Fungal and prokaryotic activities in the marine subsurface biosphere at Peru Margin and

Canterbury Basin inferred from RNA-based analyses and microscopy. *Frontiers in Microbiology*, *7*, 846.
Parkes, R. J., Cragg, B., Roussel, E., Webster, G., Weightman, A., & Sass, H. (2014). A review of prokaryotic populations and processes in sub-seafloor sediments, including biosphere: geosphere interactions. *Marine Geology*, *352*, 409–425.
Parkes, R.J., B.A. Cragg, and Wellsbury, P. (2000). Recent studies on bacterial populations and processes in subseafloor sediments: A review. *Hydrogeology Journal* 8:11-28.
Pearson, A., Seewald, J. S., & Eglinton, T. I. (2005). Bacterial incorporation of relict carbon in the hydrothermal environment of Guaymas Basin. *Geochimica et Cosmochimica Acta*, *69*(23), 5477–5486.
Qi, Y.-L., Evans, P. N., Li, Y.-X., Rao, Y.-Z., Qu, Y.-N., Tan, S., Jiao, J.-Y., Chen, Y.-T., Hedlund, B. P., & Shu, W.-S. (2021). Comparative genomics reveals thermal adaptation and a high metabolic diversity in "Candidatus Bathyarchaeia." *Msystems*, *6*(4), e00252-21.
Quisel, J. D., Burkholder, W. F., & Grossman, A. D. (2001). In vivo effects of sporulation kinases on mutant Spo0A proteins in Bacillus subtilis. *J. Bacteriol*. *183*(22), 6573–6578.
R Core Team (2018). *R: A language and environment for statistical computing*. R Core Team R Foundation for Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/.
Ramírez, G. A., Graham, D., & D'Hondt, S. (2018). Influence of commercial DNA extraction kit choice on prokaryotic community metrics in marine sediment. *Limnology and Oceanography: Methods*, *16*(9), 525–536.
Ramírez, G. A., McKay, L. J., Fields, M. W., Buckley, A., Mortera, C., Hensen, C., Ravelo, A. C., & Teske, A. P. (2020). The Guaymas Basin subseafloor sedimentary archaeome reflects complex environmental histories. *IScience*, *23*(9), 101459.
Reysenbach, A.-L., St. John, E., Meneghin, J., Flores, G. E., Podar, M., Dombrowski, N., Spang, A., L'Haridon, S., Humphris, S. E., & de Ronde, C. E. J. (2020). Complex subsurface hydrothermal fluid mixing at a submarine arc volcano supports distinct and highly diverse microbial communities. *Proc. Nat. Acad. Sci. USA*, *117*(51), 32627–32638.
Sabath, N., Ferrada, E., Barve, A., & Wagner, A. (2013). Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biology and Evolution*, *5*(5), 966–977.
Schaller, G. E., Shiu, S.-H., & Armitage, J. P. (2011). Two-component systems and their co-option for eukaryotic signal transduction. *Current Biology*, *21*(9), R320–R330.
Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.
Shao, N., Fan, Y., Chou, C.W., Yavari, S., Williams, R.V., Amster, I.J., Brown, S.M., Drake, I.J., Duin, E.C., Whitman, W.B., Liu, Y. (2022) Expression of divergent methyl/alkyl coenzyme M reductases from uncultured archaea. *Commun. Biol*. 5:1113. doi: 10.1038/s42003-022-04057-6.
Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., & Banfield, J. F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*(7), 836–843.
Simoneit, B. R. T., Oros, D.R., Leif, R.N., & Medeiros, P.M. (1995). Weathering and biodegradation of hydrothermal petroleum in the north rift of Guaymas Basin, Gulf of California. *Revista Mexicana de Ciencias Geologicas 36(2), 159-169*.

Simonsen, A.K., 2022. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *ISME J.* 16:423-434.

Starnawski, P., Bataillon, T., Ettema, T. J. G., Jochum, L. M., Schreiber, L., Chen, X., Lever, M. A., Polz, M. F., Jørgensen, B. B., Schramm, A., & Kjeldsen, K.U. (2017). Microbial community assembly and evolution in subseafloor sediment. *Proc. Nat. Acad. Sci. USA*, *114*(11), 2940–2945.

Suzuki, M. T., Taylor, L. T., & DeLong, E. F. (2000). Quantitative analysis of small-subunit rRNA genes in mixed microbial populations via 5′-nuclease assays. *Appl. Environ. Microbiol*. *66*, 4605–4614.

Takai, K. E. N., Moser, D. P., DeFlaun, M., Onstott, T. C., & Fredrickson, J. K. (2001). Archaeal diversity in waters from deep South African gold mines. *Appl. Environ. Microbiol*. *67*:5750–5760.

Teske, A. D. Lizarralde, T.W. Höfig, I.W. Aiello, J.L. Ash, D.P. Bojanova, M.D. Buatier, V.P. Edgcomb, C.Y. Galerne, S. Gontharet, V.B. Heuer, S. Jiang, M.A.C. Kars, S. Khogenkumar Singh, J.-H. Kim, L.M.T. Koornneef, K.M. Marsaglia, N.R. Meyer, Y. Morono, R. Negrete-Aranda, F. Neumann, L.C. Pastor, M.E. Peña-Salinas, L.L. Pérez Cruz, L. Ran, A. Riboulleau, J.A. Sarao, F. Schubert, J.M. Stock, L.M.A.A. Toffin, W. Xie, T. Yamanaka, Zhuang, G. (2021b). Site U1545. In Teske, A., Lizarralde, D., Höfig, T.W., and the Expedition 385 Scientists, *Guaymas Basin Tectonics and Biosphere*. Proceedings of the International Ocean Discovery Program, 385, College Station, TX. https://doi.org/10.14379/iodp.proc.385.103.2021

Teske, A. D. Lizarralde, T.W. Höfig, I.W. Aiello, J.L. Ash, D.P. Bojanova, M.D. Buatier, V.P. Edgcomb, C.Y. Galerne, S. Gontharet, V.B. Heuer, S. Jiang, M.A.C. Kars, S. Khogenkumar Singh, J.-H. Kim, L.M.T. Koornneef, K.M. Marsaglia, N.R. Meyer, Y. Morono, R. Negrete-Aranda, F. Neumann, L.C. Pastor, M.E. Peña-Salinas, L.L. Pérez Cruz, L. Ran, A. Riboulleau, J.A. Sarao, F. Schubert, J.M. Stock, L.M.A.A. Toffin, W. Xie, T. Yamanaka, Zhuang, G. (2021d). Sites U1547 and U1548. In Teske, A., Lizarralde, D., Höfig, T.W., and the Expedition 385 Scientists, *Guaymas Basin Tectonics and Biosphere*. 385, Proceedings of the International Ocean Discovery Program, 385, College Station, TX. https://doi.org/10.14379/iodp.proc.385.105.2021

Teske, A., Callaghan, A.V., and LaRowe, D.E. (2014). Biosphere frontiers of subsurface life in the sedimented hydrothermal system of Guaymas Basin. *Frontiers in Microbiology* 5:362; doi:10.3389/fmicb.2014.00362.

Teske, A., D. Lizarralde, T.W. Höfig, I.W. Aiello, J.L. Ash, D.P. Bojanova, M.D. Buatier, V.P. Edgcomb, C.Y. Galerne, S. Gontharet, V.B. Heuer, S. Jiang, M.A.C. Kars, S. Khogenkumar Singh, J.-H. Kim, L.M.T. Koornneef, K.M. Marsaglia, N.R. Meyer, Y. Morono, R. Negrete-Aranda, F. Neumann, L.C. Pastor, M.E. Peña-Salinas, L.L. Pérez Cruz, L. Ran, A. Riboulleau, J.A. Sarao, F. Schubert, J.M. Stock, L.M.A.A. Toffin, W. Xie, T. Yamanaka, Zhuang, G. (2021c). Site U1546. In Teske, A., Lizarralde, D., Höfig, T.W., and the Expedition 385 Scientists, *Guaymas Basin Tectonics and Biosphere*. Proceedings of the International Ocean Discovery Program, 385, College Station, TX. https://doi.org/10.14379/iodp.proc.385.104.2021

Teske, A., D. Lizarralde, T.W. Höfig, I.W. Aiello, J.L. Ash, D.P. Bojanova, M.D. Buatier, V.P. Edgcomb, C.Y. Galerne, S. Gontharet, V.B. Heuer, S. Jiang, M.A.C. Kars, S. Khogenkumar Singh, J.-H. Kim, L.M.T. Koornneef, K.M. Marsaglia, N.R. Meyer, Y. Morono, R. Negrete-Aranda, F. Neumann, L.C. Pastor, M.E. Peña-Salinas, L.L. Pérez

Cruz, L. Ran, A. Riboulleau, J.A. Sarao, F. Schubert, J.M. Stock, L.M.A.A. Toffin, W. Xie, T. Yamanaka, Zhuang, G. (2021e). Site U1549. In Teske, A., Lizarralde, D., Höfig, T.W., and the Expedition 385 Scientists, *Guaymas Basin Tectonics and Biosphere*. Proceedings of the International Ocean Discovery Program, 385, College Station, TX https://doi.org/10.14379/iodp.proc.385.106.2021

Teske, A., Lizarralde, D., Höfig, T.W., Aiello, I.W., Ash, J.L., Bojanova, D.P., Buatier, M.D., Edgcomb, V.P., Galerne, C.Y., Gontharet, S., Heuer, V.B., Jiang, S., Kars, M.A.C., Khogenkumar Singh, S., Kim, J.-H., Koornneef, L.M.T., Marsaglia, K.M., Meyer, N.R., Morono, Y., Negrete-Aranda, R., Neumann, F., Pastor, L.C., Peña-Salinas, M.E., Pérez Cruz, L.L., Ran, L., Riboulleau, A., Sarao, J.A., Schubert, F., Stock, J.M., Toffin, L.M.A.A., Xie, W., Yamanaka, T., and Zhuang, G. (2021a). Expedition 385 Methods. In Teske, A., Lizarralde, D., Höfig, T.W., and the Expedition 385 Scientists, *Guaymas Basin Tectonics and Biosphere*. Proceedings of the International Ocean Discovery Program, 385, College Station, TX. https://doi.org/10.14379/iodp.proc.385.102.2021

Teske, A., McKay, L. J., Ravelo, A. C., Aiello, I., Mortera, C., Núñez-Useche, F., Canet, C., Chanton, J. P., Brunner, B., & Hensen, C. (2019). Characteristics and evolution of sill-driven off-axis hydrothermalism in Guaymas Basin–the Ringvent site. *Scientific Reports*, *9*(1), 13847.

Teske, A., Wegener, G., Chanton, J. P., White, D., MacGregor, B., Hoer, D., de Beer, D., Zhuang, G., Saxton, M. A., & Joye, S. B. (2021f). Microbial communities under distinct thermal and geochemical regimes in axial and off-axis sediments of Guaymas Basin. *Frontiers in Microbiology*, *12*, 633649.

Thiel, J., Byrne, J.M., Kappler, A., Schink, B., and Pester, M. (2019). Pyrite formation from FeS and H_2S is mediated through microbial redox activity. *Proc. Nat. Acad. Sci. USA* 116, 6897-6902.

Tran, P.Q., Bachand, S.C., McIntyre, P.B., Kraemer B.M., Vadeboncoeur, Y., Kimirei, I.A., Tamatamah, R., McMahon, K.D., and Anantharaman, K. (2021) Depth-discrete metagenomics reveals the roles of microbes in biogeochemical cycling in the tropical freshwater Lake Tanganyika. *ISME J* 15, 1971–1986.

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 314–324.

Vigneron, A., Cruaud, P., Roussel, E. G., Pignet, P., Caprais, J.-C., Callac, N., Ciobanu, M.-C., Godfroy, A., Cragg, B. A., & Parkes, J. R. (2014). Phylogenetic and functional diversity of microbial communities associated with subsurface sediments of the Sonora Margin, Guaymas Basin. *PloS One*, *9*(8), e104427.

Von Damm, K. L., Edmond, J. M. t, Measures, C. I., & Grant, B. (1985). Chemistry of submarine hydrothermal solutions at Guaymas Basin, Gulf of California. *Geochimica et Cosmochimica Acta*, *49*(11), 2221–2237.

Wagner, G. P., Kin, K., and Lynch, V.J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences* 131, 281-285.

Wang, Y., Wegener, G., Ruff, S.E., Wang, F. 2021. Methyl/alkyl-coenzyme M reductase-based anaerobic alkane oxidation in archaea. *Environ. Microbiol*. 23:530-541.

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2.' Create Elegant Data Visualizations Using the Grammar of Graphics. Version, *2*(1), 1–189.

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*, 605–607.

Yang, J., Xian, M., Su, S., Zhao, G., Nie, Q., Jiang, X., Zheng, Y., Liu, W. (2012). Enhancing production of bio-isoprene using hybrid MVA pathway and isoprene synthase in *E. coli*. *PLoS One* **7**(4):e33509, doi: 10.1371/journal.pone.0033509.

Zhang, W., Ding, W., Yang, B., Tian, R., Gu, S., Luo, H., & Qian, P.-Y. (2016). Genomic and transcriptomic evidence for carbohydrate consumption among microorganisms in a cold seep brine pool. *Frontiers in Microbiology*, *7*, 1825.

Zhou, Y.-L., Mara, P., Cui, G.-J., Edgcomb, V. P., & Wang, Y. (2022). Microbiomes in the Challenger Deep slope and bottom-axis sediments. *Nature Communications*, *13*(1), 1515.

Zhou, Z., Tran, P.T., Breister, A.M., Liu, Y., Kieft, K., Cowley, E.S., Karaoz, U., & Anantharaman, K. (2022). METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* 10.1: 33.

Acknowledgments

The authors would like to acknowledge the crew and entire science party for IODP Expedition 385 for their assistance with sample collection. Without their assistance this study would be impossible. The authors would also like to thank Gustavo Ramírez for his assistance with DNA extractions using his method. We thank Mark Shaw and Bruce Kingham in the University of Delaware DNA Sequencing & Genotyping Center for assistance with sample preparation and Illumina sequencing. This study was supported by NSF Grant OCE-2046799 to VE, PM, AT, and R. Hatzenpichler, by NSF grant OCE-1829903 to VE, PM, and AT, and by JSPS KAKENHI Grants JP19H00730 and JP23H00154 to YM.

Tables

	Sample ID	T°C	Depth (mbsf)	Alkalinity (mM)	SO_4^{2-} (mM)	PO_4^{3-} (µM)	H_2S (µM)	NH_4^+ (mM)	CH_4 (mM)	CO (nM)	DOC (mg/L)	DIC (mM)	TOC (wt %)	TN (wt %)	TOC /TN
	U1545B_1H2	5.3	1.7	6	26.9	33.9	44.1	0.5	0	360	24.2	2.4	4.87	0.61	9.3
	U1545B_2H3	6.4	6.8	6.9	26.3	36.4	1220.5	0.5	0	367	24.2	2.4	4.03	0.63	7.45
	U1545B_4H2	10.4	24.3	21	21.1	51	6068	5.3	0	189	25.7	7.6	2.43	0.38	7.4
U1545B	U1545B_4H3	10.7	25.8	21	21.1	51	6068	5.3	0	254	25.7	7.6	2.43	0.38	7.4
	U1545B_8H3	19.3	63.8	59.5	0.7	78.3	8947	9.2	1.5	174	73.2	27.9	2.73	0.34	9.4
	U1545B_13H4	30.2	112.5	40.4	0.4	77.1	1891	15.1	1.2	161	52.3	14.1	1.97	0.29	8.0
	U1545B_19F3	39.8	155.0	35.3	0.3	46.9	3.2	15.9	1.9	116	48.2	10.7	2.17	0.34	7.5
	U1545B_32F3	52.4	211.1	28.5	0.3	16.2	0	23.9	0.4	160	50.7	4.8	1.74	0.28	7.3
	U1545B_34F3	54.3	219.5	26.7	0.2	12.2	0	25.6	0.6	168	40.9	3	2.42	0.4	7
U1546B	U1546B_1H2	2.8	0.8	5.1	27.7	12.4	428	0.7	0	665	21.8	2.2	4.36	0.52	9.7
6B	U1546B_3H2	6.2	16.4	7.3	26.7	28.7	1226.6	0.3	0	260	21	2.5	2.15	0.34	7.4
	U1547B_1H2	14.2	2.2	3	27.9	16.5	68	0.1	0	350	13.9	1.2	3.76	0.45	9.7
	U1547B_1H3	15.0	3.6	3	27.9	16.5	68	0.1	0	350	13.9	1.2	3.76	0.45	9.7
	U1547B_2H2	17.5	8.7	5.3	26.8	22.8	759	0.3	0	427	14.7	1.2	3.53	0.29	14
	U1547B_2H3	17.8	9.9	5.3	26.8	22.8	759	0.3	0	427	14.7	1.2	3.53	0.29	14
	U1547B_3H3	23.7	19.3	7.8	26.1	33.3	1538	0.6	0	252	14.6	1.5	2.13	0.26	9.4
U1547B	U1547B_5H2	31.9	36.9	14.5	23.3	24.7	3725.5	1.5	0	197	14.3	1.8	2.63	0.29	10.5
7B	U1547B_5H3	34.0	38.1	14.5	23.3	24.7	3725.5	1.5	0	197	14.3	1.8	2.63	0.29	10.5
	U1547B_7H3	42.3	57.4	13.3	21.1	27	5110	2.6	0	83	23.3	1.5	2.03	0.26	9
	U1547B_8H2	46.6	65.8	13	20.2	25.4	6606	2.9	0	91	23.3	1.5	1.25	0.15	9.5
	U1547B_9H2	51.0	74.3	13.7	18.8	28	7151	3.7	0	54	15	1.9	2.2	0.26	10
	U1547B_9H3	51.8	76.0	13.7	18.8	28	7151	3.7	0	54	15	1.9	2.2	0.26	10
U1548B	U1548B_2H3	13.7	8.9	4.8	26.8	24.5	732.5	0.3	0	554	13.1	1.4	2.4	0.22	12.9
8B	U1548B_4H7	33.5	33.5	9.5	25.4	18.6	1913	0.6	0	303	9.4	1.2	1.9	0.22	10.3
	U1548B_8H5	62.4	69.5	13	20.2	25.4	6606	2.9	0	51	1.5	0.2	1.57	0.2	9
U1549B **9B**	U1549B_3H2	6.4	16.5	18.7	17.5	98.1	2567	3.3	0	188	23.6	5.7	3.12	0.59	6.1

Table 4-1. Geochemical, depth and temperature data for metagenomic samples. The samples are sorted by increasing temperature. "Sample ID" is composed by site and core section IDs. Geochemical data are compiled from Expedition 385 site chapters using the best available sample matches, and *in situ* temperatures represent linear interpolation based on published temperature gradients in the site chapters (Teske et al., 2021b-e).

Supplementary Note

In this Supplementary Note we provide an extended overview on the genomic background of bacterial and archaeal lineages that dominate the Guaymas subsurface metagenome and MAG surveys, and we describe certain genes involved in carbon and iron cycling that we recovered on most of our MAGs. We focus our discussion on Chloroflexota (carbon fixation, DMSO reduction), Thermoproteota (acetogenesis and methane cycling), Acidobacteriota (nitrogen fixation), Desulfobacterota (sulfate and iron reduction), Aerophobota, and the White Oak River group 3 (WOR-3) (diverse heterotrophic capabilities, and CRISPR genes). We introduce further metagenomic evidence for iron reduction and oxidation, and carbon monoxide oxidation. We include a section that examines the application of bioinformatic tool "MetaPathPredict" (Geller-McGrath et al., 2022) and its insights into the metabolic potential of MAGs assigned to less dominant phyla (Zixibacteria and Cloacimonadetes) in the Guaymas Basin subsurface. Finally, at end of this Supplementary Note, we devote a section that explains the processing and the rationale behind the control samples used in this study.

Overview on the genomic background of subsurface bacterial and archaeal phyla in Guaymas Basin

Chloroflexota

Chloroflexota are abundant in the hadal ocean (Liu et al., 2022a), in marine subsurface sediments (Vuillemin et al., 2020, Fincker et al., 2020,) and hydrothermal settings (Fullerton and Moyer, 2016, Dombrowski et al., 2018; Reysenbach et al., 2020). Diverse thermophilic Chloroflexota lineages have also been enriched and isolated from hot springs (Dodsworth et al., 2014; Palmer et al., 2023).

Genes detected in our Chloroflexota MAGs are associated with fatty acid degradation, ornithine biosynthesis and the methionine salvage pathway that produces methionine by recycling sulfur-bearing metabolites (Sekowska et al., 2004). Ornithine biosynthesis was also evident in MAGs affiliated with other bacterial and archaeal phyla, including the Hadarchaeota. Genes for enzymes involved in ornithine synthesis (e.g., ornithine decarboxylase) have been documented previously in deep biosphere samples (Orsi et al., 2013), indicating that this capability may be widely utilized by subsurface microbiota (Hernández et al., 2021). Ornithine synthesis can be

involved in some amino acid metabolisms as well as urea biosynthesis (Cunin et al., 1986; Therkildsen et al., 1996).

The majority of Chloroflexota genomes contained marker genes associated with the oxidation of formate (*fdhA*, *fdhB*, *fdoG*, *fdoH*; McGonigle et al., 2020) and carbon monoxide (*coxL*, *coxM*, *coxS*; Islam et al., 2019). Their occurrence coincides with increased carbon monoxide concentrations (83-665 nM) between 0.8-60 mbsf (Supplementary Data 1). Hydrogen utilization is suggested by Ni-Fe hydrogenases and key genes of acetate/acetyl-CoA production (*pta*/*ack*, *acdA*) in Chloroflexota MAGs (Supplementary Data 5, 6). Hydrogen concentrations associated with our metagenomic samples ranged between 23-84 nM (Supplementary Data 1) indicative of active hydrogen cycling (Lin et al., 2012). Genes associated with the Wood-Ljungdahl pathway (WL; *cdhD*, *cdhE*, *cooS*) were present in 17/23 Chloroflexota MAGs, while ATP-citrate lyase (*aclA*), associated with the reductive TCA cycle (rTCA), was present in one MAG from the VGOG01 order. The *mmoB* gene associated with methane oxidation (Supplementary Data 7) was identified in two Chloroflexota MAGs from the orders Promineofilales (Speirs et al., 2019) and E44-bin15, associated with petroleum seepage in marine sediments (Dong et al., 2019). Previous DNA-SIP experiments identified members of the Chloroflexota as putative methane oxidizers and detected methane monooxygenases (*mmoX* or *pmoB)* in Chloroflexota genomes (Altshuler et al., 2022). We caution that the identified *mmoB* has a regulatory role in methane oxidation, while the catalytic activity is encoded by the *mmoH* gene in the *mmo* operon (Sirajuddin and Rosenzweig, 2015).

Genes involved in sulfate and sulfur assimilation for amino acid synthesis (e.g., cysteine; *cysN*, *sat*, *cysD*) and in sulfur oxidation *(sdo*, *dsrH)* were present in all Chloroflexota MAGs, however, none of our Chloroflexota MAGs encoded complete pathways for sulfur oxidation. Evidence for dimethyl sulfoxide (DMSO) utilization was suggested by the presence of *dmsA* and/or *dmsB* (involved in DMSO reduction) in five of our MAGs (Supplementary Data 5-6). DMSO is abundant in deep sea ecosystems and was suggested to be an electron acceptor for microbes that survive in deep-sea extreme conditions (Xiong et al., 2016). The *nirB* and *nirD* genes were encoded in one Chloroflexota MAG, which can be involved in denitrification, dissimilatory nitrate reduction to ammonium, and/or assimilatory nitrate reduction (Zumft 1997; Stolz and Basu 2002). Finally, genes for transport of tungstate were identified in 4 Chloroflexota MAGs, and transporters of molybdenum/tungsten in 10 MAGs (Supplementary Data 5, 6). Tungsten is found

abundantly in hydrothermal ecosystems (Kishida et al., 2004), and serves as a redox catalyst in metalloenzymes of thermophilic archaea inhabiting hydrothermal vents (Kletzin and Adams, 1996) and hot springs (Buessecker et al., 2022). Evidence of putative arsenate biomineralization for detoxification (e.g., *arsC*, *arsM* genes) or energy gain via arsenotrophy (*arrA* gene; Saunders et al., 2019) was present in three MAGs. Sedimentary arsenic input would explain the elevated concentrations of arsenic in Guaymas Basin hydrothermal fluids (up to 1 μmol) that exceed those at other hydrothermal vent sites (Von Damm et al., 1985). Five Chloroflexota MAGs contained CRISPR/Cas genes involved in genome editing (Supplementary Data 5,6).

The Chloroflexota are a cosmopolitan phylum of Bacteria that contain a diverse array of metabolic capabilities (Dombrowski et al., 2017; Zhou et al., 2022; Rogers et al.,2023). We found that the Chloroflexota MAGs recovered from Guaymas Basin contained the potential for hydrocarbon and fatty acid utilization, as well as formate oxidation and carbon monoxide oxidation as has been reported previously (Dombrowski et al., 2017). Some lineages contained genes for carbon fixation via the Wood-Llungdahl pathway, as well as the potential for organohalide respiration. These metabolisms for the Chloroflexota have been reported previously in Guaymas Basin and other subsurface environments (Dombrowksi et al., 2017; Fincker et al., 2020).

Thermoproteota

Bathyarchaeia recycle hydrogen and CO_2 from fermentation using the WL pathway (He et al., 2016). Eight out of 11 of our Bathyarchaeia MAGs contained genes involved in the WL pathway. In addition, we detected the marker gene for the formaldehyde activating enzyme (*fae*) in 5 MAGs affiliated with 40CM-2-53-6, B26-1 and TCS64 orders. *Fae* condenses formaldehyde and tetrahydromethanopterin to form methylene-tetrahydromethanopterin that can be reduced and utilized in the WL pathway (Timmers et al., 2017; Vorholt et al., 2000). Six Bathyarchaeia MAGs encoded the *mer* gene (*ffdA* synonym, to avoid confusion with the *mer* operon for mercury reduction) for the reduction of methylene-tetrahydromethanopterin in the WL pathway. Various Bathyarchaeia sub-lineages have been reported to encode genes for anaerobic oxidation of methane/alkane compounds (*mcr*/*acr* complex) (Evans et al., 2015, Evans et al., 2019; Qi et al., 2021; Vanwonterghem et al., 2016). We detected the methane/alkane oxidation marker genes *fwd*, *ftr*, *mtd*, *mch*, and *mtr* in all Bathyarchaeia MAGs. The *acr*/*mcr* genes encoding the methyl/alkyl-coenzyme M complex were absent in the Bathyarchaeia MAGs, indicating loss of the *mcr*/*acr*

operon as described previously for Bathyarchaeia. Nine out of 11 Bathyarchaeia MAGs also encoded genes for acetate formation (*acdA*, *ack*, *pta, acs)* which could be utilized to couple methylotrophy with acetogenesis, as has been described previously in Bathyarchaeia MAGs from deep sediments (Farag et al., 2020; He et al., 2016).

Marker genes for other specific metabolic capacities in our Bathyarchaeota MAGs included genes for fermentation (*porA*), hydrogen cycling (Ni-Fe hydrogenases) and genes involved in the anaerobic degradation of benzoate (*bcrA, bcrB, bcrD*; Kung et al., 2009). The *bcr* genes were also observed in Chloroflexota (5 MAGs), Zixibacteria (1 MAG) and Desulfobacterota (4 MAGs); the latter group can be enriched in Guaymas Basin sediments on benzoate, under sulfate-reducing conditions (Edgcomb et al., 2022).

Acidobacteriota

Members of the highly diverse heterotrophic phylum Acidobacteriota occur in a wide range of freshwater and marine seafloor environments and can utilize oxygen or other electron acceptors (e.g., nitrate, nitrite, sulfate) for respiration (Flieder et al., 2021 and references therein). We detected eight Acidobacteriota MAGs annotated to the orders of Aminicenantales (7 MAGs), and Acidoferrales (1 MAG). Aminicenantales MAGs were previously recovered from surficial Guaymas Basin sediments (Dombrowski et al., 2018), and in this study they were found at all sites between 0.8-60 meters below sea floor (mbsf). The Acidoferrales MAG was detected only below 61.6 mbsf at sites U1545B and U1548B. The overall metabolic potential of Guaymas subsurface Acidobacteriota MAGs is summarized in Supplementary Data 5 and 6.

We observed that almost all Acidobacteriota MAGs encoded the NtrY-NtrX two-component regulatory system, a redox sensor system widely distributed in Proteobacteria which regulates denitrification and nitrogen fixation genes, and senses nitrogen levels under nitrogen limitation (Pawlowski et al., 1991; DelVecchio et al., 2002; Bonato et al., 2016). Six of our eight Acidobacteriota MAGs contained at least one gene of the *nif* operon (e.g., *nifU, nifB, nifH*) which suggests putative nitrogen fixation in our subsurface samples. This is similar to previous reports of deep-sea sediment Acidobacteriota that encode *nifH* in their genomes (Kapili et al., 2020). Two Acidobacteriota MAGs contained CRISPR/Cas genes involved in genome editing (Supplementary Data 5).

Desulfobacterota

Desulfobacterota include primarily heterotrophic sulfate reducers and syntrophic sulfate-reducing lineages that couple sulfate reduction with methane and short-chain alkane oxidation by anaerobic methane oxidizers (ANME archaea). These bacteria are widespread in Guaymas Basin sediments and other hydrothermal and cold seep sites (Knittel and Boetius, 2009; Murphy et al., 2021; Speth et al., 2022; Wegener et al., 2022; Zhou et al., 2022). Seven Desulfobacterota MAGs, belonging to the Desulfobacterales, Desulfatiglandales, and WTBG01 and WVXP0 orders were recovered primarily from shallow sulfate-rich cool sediments of all sites (0.8-15 mbsf, at or above the SMTZ with temperatures 2-20ºC) (Supplementary Figure 4). Two Desulfobacterales MAGs contained the *dsr* operon (e.g., *dsrB/J/K/D*) involved in dissimilatory sulfate reduction (Venceslau et al., 2014). One MAG annotated to the WTBG01 order (found in freshwater anoxic sulfidic sediments; Murphy et al., 2021) encoded the sulfate adenylyltransferase gene (*sat*) associated primarily with sulfur assimilation. Marker genes of DMSO reduction and/or sulfur assimilation were also detected in all Desulfobacterota MAGs. One Desulfobacterota MAG contained CRISPR/Cas genes involved in genome editing (Supplementary Data 5).

The potential for iron reduction was evidenced in all Desulfobacterota MAGs by the presence of *mtrA*, *mtoA* (Garber et al., 2020) and *eetB* genes, suggesting an extracellular electron transfer mechanism. In addition, *DFE* genes encoding multiheme cytochromes (e.g., *DFE_0449, DFE_0461, DFE_0451*) were found in three of our MAGs annotated to Desulfobacterota and in five MAGs annotated to Aminicenantales (Acidobacteriota). *DFE* genes are involved in iron oxidation and were originally documented in the genome of *Desulfovibrio ferrophilus* strain IS5 (Deng and Okamoto, 2018). They encode cytochromes and β-propeller proteins, which can function as electron carriers and leader peptides in extracellular electron transfer (Chatterjee et al., 2021; Deng and Okamoto, 2018). At depths where Desulfobacterota MAGs were detected, dissolved porewater iron ranged in concentration from < 1 μM to greater than 4 μM (Supplementary Data 1), indicating possibly active iron cycling with little accumulation.

Aerophobota and White Oak River group 3 (WOR-3)

The phylum Aerophobota is widely distributed in deep-sea sediments, and includes fermentative thermophiles and hyperthermophiles affiliated with hydrocarbon-rich environments such sediments from the Pescadero Basin (Speth et al., 2022), and methane hydrate-bearing sediments

(Liu et al., 2022b). We recovered five Aerophobota MAGs (order Aerophobiales) from the deep subsurface at depths below 100 mbsf, when temperatures did not exceed 40°C. This suggests that the distribution of Aerophobota appears to be constrained by thermal limits rather than by depth. Guaymas Aerophobota MAGs encoded marker genes for acetate/acetyl-CoA production (*acdA*, *ack*, *pta*) (Dong et al., 2019), fermentation (*porA*) and degradation of polysaccharides (e.g., cellulose, chitin). Two Aerophobota MAGs contained CRISPR/Cas and CRISPR/Csm genes that comprise adaptive defense systems against infectious agents in prokaryotes (Colognori et al., 2023).

Members of the bacterial WOR-3 candidate phylum were originally described from shallow estuarine sediments (Baker et al., 2015) and hydrothermal Guaymas Basin sediments (Dombrowski et al., 2017). Five of our six WOR-3 MAGs were recovered from depths at 0.8-26.9 mbsf; in contrast, a WOR-3 MAG annotated to UBA3073 (order-level) was recovered primarily from 112.5 and 154.2 mbsf at site U1545 (up to ~45°C) (Figure 2). Overall, our WOR-3 MAGs encoded various peptidases, as well as genes for H_2 cycling (Ni-Fe hydrogenase genes) and putative chitin degradation (endo-acting chitinase genes). While these results match previous findings (Baker et al., 2015), the Guaymas subsurface WOR-3 MAGs also contain genes for fermentation and acetate production (*acs, acdA*, *ack, porA*), and marker genes for endohemicellulases and amylolytic enzymes that can degrade other polysaccharides aside from chitin. One WOR-3 MAG encoded CRISPR/Cas genes.

Carbon and Iron cycling

Iron reduction and oxidation

Iron reduction is a known capability for Bacteria and Archaea associated with marine benthic sediments (Flieder et al., 2021; Jiang et al., 2019). Marker genes involved in dissimilatory iron reduction (e.g., *dmkA*, *dmkB*, *eetA*, *eetB*, *fmnA*, *fmnB*, *pplA*, *ndh2*; Garber et al., 2020) with the potential for extracellular electron transfer (EET) were identified in 86/89 MAGs from all recovered phyla (Supplementary Data 5, 6), and were only missing from one Thermoproteota, one WOR-3, and one Aenigmatarchaeota MAG. These genes participate in EET from the cell towards the surrounding environment (Light et al., 2018; Shi et al., 2016). Based on laboratory

experiments, EET is suggested to enhance iron bioavailability and iron uptake in anaerobes, and to act as a redox mechanism that may contribute to the proton motive force (Jeuken et al., 2020).

Eight out of 23 Chloroflexota, 2/11 Thermoproteota, 5/8 Acidobacteriota, 3/7 Desulfobacterota, and 1/6 WOR-3 MAGs also encoded components of the *DFE_0448-0451* and *DFE_0461-0465* operon genes (Supplementary Data 5, 6) homologous to multiheme cytochrome systems first identified in *Desulfovibrio ferrophilus* (Deng et al., 2018). According to the *D. ferrophilus* model, electrons from an external iron source move along extracellular and membrane-spanning multiheme cytochromes from the outer membrane to the periplasm of the cell and finally are passed to a terminal electron acceptor (Deng et al., 2018). Putative terminal electron acceptors for this process in MAGs containing *DFE* components include the sulfur cycle intermediates sulfite (based on the presence of sulfite reductase *asrA* and *asrB*), tetrathionate (from the detection of tetrathionate reductase gene *ttrB*), and thiosulfate (due to the annotation of thiosulfate reductase *phsA/B* genes; Supplementary Data 5, 6). The capacity for polysulfide reduction was also detected based on the presence of polysulfide reductase (*psrA*) in four Acidobacteriota MAGs containing *DFE* multiheme cytochrome components.

Electron acceptors from nitrogen cycle intermediates included nitrite (*nasD*) and nitrate (*napA*, *narB*). This process is predicted to be utilized when more energy-rich substrates such as organic compounds are limited (Deng et al., 2018). Iron (II and III) concentrations associated with metagenomic samples containing MAGs that encoded genes affiliated with iron metabolism ranged from nanomolar, up to 11.8 μM (Supplementary Data 1).

Carbon monoxide oxidation

Hydrothermal environments often contain carbon monoxide (CO), which can be produced by the breakdown of organic matter or generated by certain anaerobic microorganisms (Kochetkova et al., 2011; Sokolova et al., 2009). The potential for CO oxidation, is an energetically favorable anaerobic reaction prevalent in subsurface bacterial and archaeal MAGs (Baker et al., 2016; Magnabosco et al., 2016). Nine out of 23 Chloroflexota MAGs, 3/11 Thermoproteota, 7/8 Acidobacteriota, 3/7 Desulfobacterota, 2/5 Aerophobota, and one Hadarchaeota MAGs contained the genes *coxM* and *coxS* encoding CO dehydrogenase subunits (Supplementary Data 5, 6). Genes for catalytic nickel-containing CO dehydrogenase (*cooS*) and/or its iron-sulfur subunits (*cooF*) were further identified in 9/23 Chloroflexota, 1/11 Thermoproteota, 1/8 Acidobacteriota, 3/7

Desulfobacterota, and 2/5 Aerophobota MAGs. The presence of these genes suggests the capacity to oxidize CO and H_2O to CO_2 while reducing an electron-carrying cofactor that is accepting H_2. Bacteria and Archaea are also capable of CO oxidation coupled to the anaerobic reduction of sulfur and nitrogen compounds (King, 2006; Oelgeschläger and Rother, 2008). Potential terminal electron acceptors for our CO-oxidizing MAGs included the sulfur cycle intermediates sulfite (due to the presence of genes *asrA* and *asrB*), tetrathionate (*ttrB*), thiosulfate (*phsA/B*), and polysulfide (*psrA*; Supplementary Data 5,6). CO concentrations associated with metagenomic samples containing MAGs that encoded these genes ranged from 83 to 665 nM (Supplementary Data 1).

MetaPathPredict insights into metabolic potential of Zixibacteria and Cloacimonadetes

Some phyla for which we recovered MAGs (e.g., Zixibacteria and Cloacimonadota) remain poorly described in terms of their metabolic potential. To assess and predict the metabolic potential of some of our less-complete and poorly characterized bacterial MAGs, we applied a new tool "MetaPathPredict" (Geller-McGrath et al., 2022) to our Zixibacteria (n= 4) and Cloacimonadota (n=1) MAGs. MetaPathPredict is a software designed to predict the presence or absence of complete KEGG modules in partially complete bacterial genomes (see below). The tool utilizes machine learning models trained on bacterial gene annotations in the format of KEGG gene orthologs to predict the presence or absence of whole KEGG modules, and it has been designed to handle gene annotations from incomplete bacterial genomes. It is very common in environmental metagenomic studies that reconstructed MAGs will vary in their degree of completeness and contamination. Usually only a fraction of MAGs exceeds > 80% completeness, and it is known that less complete MAGs can result in underestimation of the functional capacity of the genome (Eisenhofer et al., 2023), which can be particularly important for uncultured bacterial taxa. Application of MetaPathPredict to such partially complete MAGs can be useful for providing predictions on whether metabolic pathways are present especially in cases where some key genes are missing. While this tool can sometimes miss predictions of pathways that should be present based on benchmarking tests with genomic data (Geller-McGrath et al., 2022), it can nonetheless be useful for gaining insights into less-complete MAGs, and also for predicting metabolic capacities of MAGs affiliated with poorly understood taxonomic groups.

Our Zixibacteria MAGs were 61-88.5% complete and our single Cloacimonadota MAG was 75% complete. The results of the MetaPathPredict analysis are described below and presented

in Supplementary Data 10 which shows: 1) KEGG modules present and complete in our Zixibacteria and Cloacimonadota MAGs that were also predicted by MetaPathPredict 2) KEGG modules absent or incomplete but predicted to be present by MetaPathPredict, and 3) incomplete KEGG modules, also predicted to be absent by MetaPathPredict.

Lineages of Zixibacteria have been documented in various marine and terrestrial subsurface ecosystems, hypersaline settings, and anoxic sediments (Anantharaman et al., 2018; Baker et al., 2015; Castelle et al., 2013; Lin et al., 2012b; Momper et al., 2017; Wong et al., 2020). This taxon is thought to be capable of dissimilatory nitrate and sulfate reduction, and it seems to lack complete carbon fixation pathways (Momper et al., 2017). The 4 Zixibacteria MAGs we recovered were detected at all sampling sites down to 25.8 mbsf and metagenome reads mapped most intensely from shallow/intermediate depths (8.6-16.2 mbsf). Zixibacteria (order MSB-5A5) have the metabolic capacity for oxidation of fatty acids, (e.g., acyl-CoA dehydrogenase) and synthesis of a suite of vitamins (e.g, B6, B1). MetaPathPredict predicted the potential for dissimilatory sulfate reduction in one of our MAGs (order DG-27), which also contained a marker gene for this process (*dsrA*). Two Zixibacteria MAGs (orders DG-27 and UBA10806) were also predicted to encode the potential for dissimilatory nitrate reduction and contained marker genes *napB* and *nrfH*. MetaPathPredict additionally predicted the synthesis of vitamin B7 in two Zixibacteria MAGs, and acetate production via the phosphate acetyltransferase-acetate kinase (Pta-Ack) pathway in all four MAGs. The Pta-Ack pathway can produce acetate/acetyl-CoA and can participate in carbon fixation by providing acetyl-CoA. Complete acetogenesis via the pta-ack pathway was also confirmed from our genomic data in two MAGs (*ack*, *pta*). MetaPathPredict predicted various transporters (e.g., phosphate, iron and ABC transporters), pathways involved in synthesis of co-factors using amino acids or tRNA reductases (e.g., from glutamate to heme; from tRNA-glutamyl to sideroheme), and the synthesis of various amino acids (e.g., threonine, serine, valine isoleucine). Many of these processes (e.g., biosynthesis of amino acids, iron reduction) were also verified with marker genes. The Zixibacteria MAGs did not contain the potential for carbon fixation.

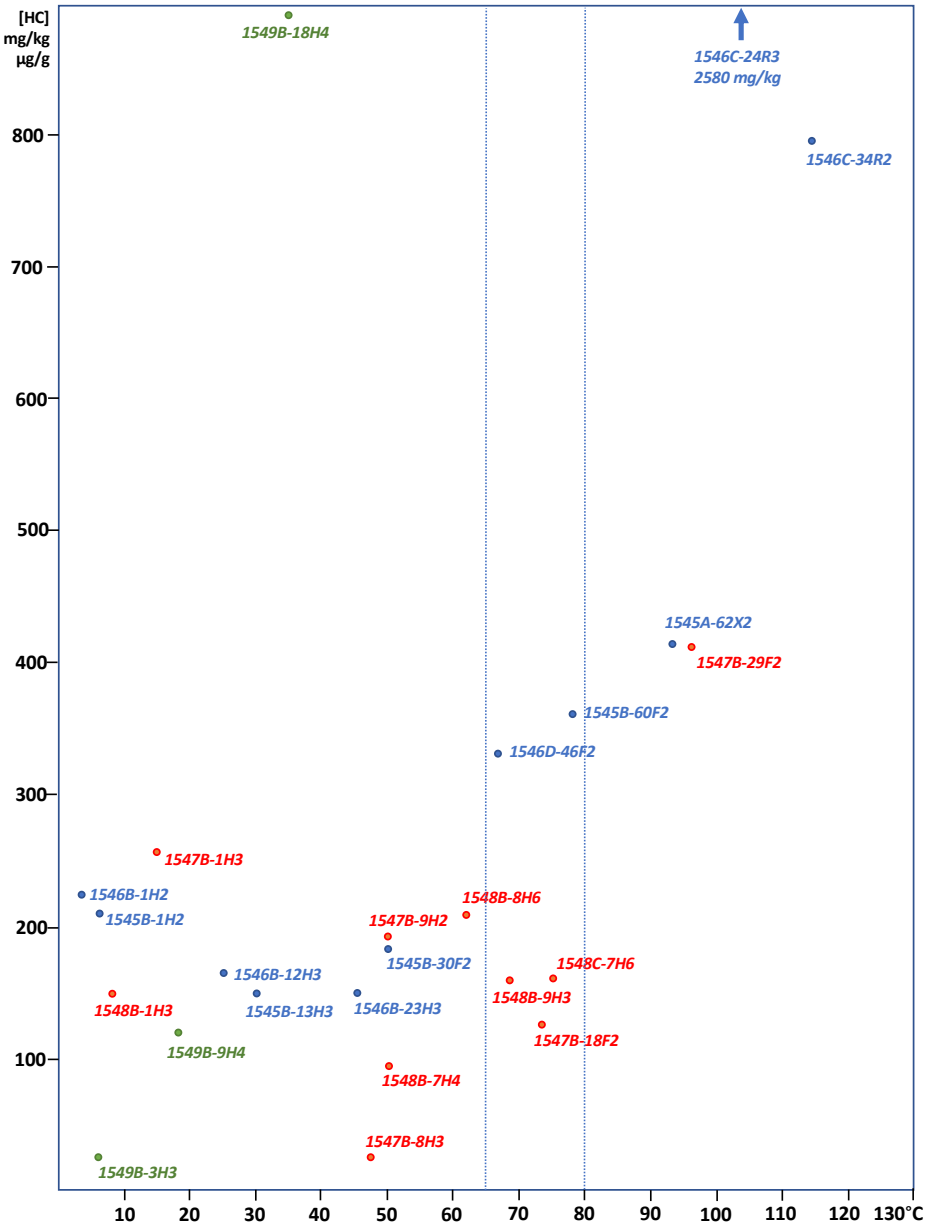
Cloacimonadota are abundant in anoxic/sulfidic water columns and cold seep brine pools (e.g., Black and Red Sea, respectively; Suominen et al., 2021; Villanueva et al., 2021; Zhang et al., 2016). They are suggested to perform diverse metabolisms including carbon fixation, fermentation, and assimilation of proteins as carbon and nitrogen sources. Our single

Cloacimonadota MAG was present in low abundance at site U1545B in metagenomes between 25.8 and 63.8 mbsf. The Cloacimonadota MAG was not predicted to contain the capacity for carbon fixation, however our data showed (and the MetaPathPredict successfully predicted) that this taxon encodes genes for fermentation via acetate production (*pta*, *ack*), and synthesis of vitamin B7 and salvage of thiamine (vitamin B1), an indispensable cofactor in amino acid and carbohydrate metabolism. Further, salvage of B1 is linked to the metabolism of pyrimidines essential for maintenance and synthesis of the DNA strands (Gonçalves and Gonçalves, 2019). Biosynthesis of purines and pyrimidines is a core metabolic process, and was detected, and predicted, in our Cloacimonadota MAG.

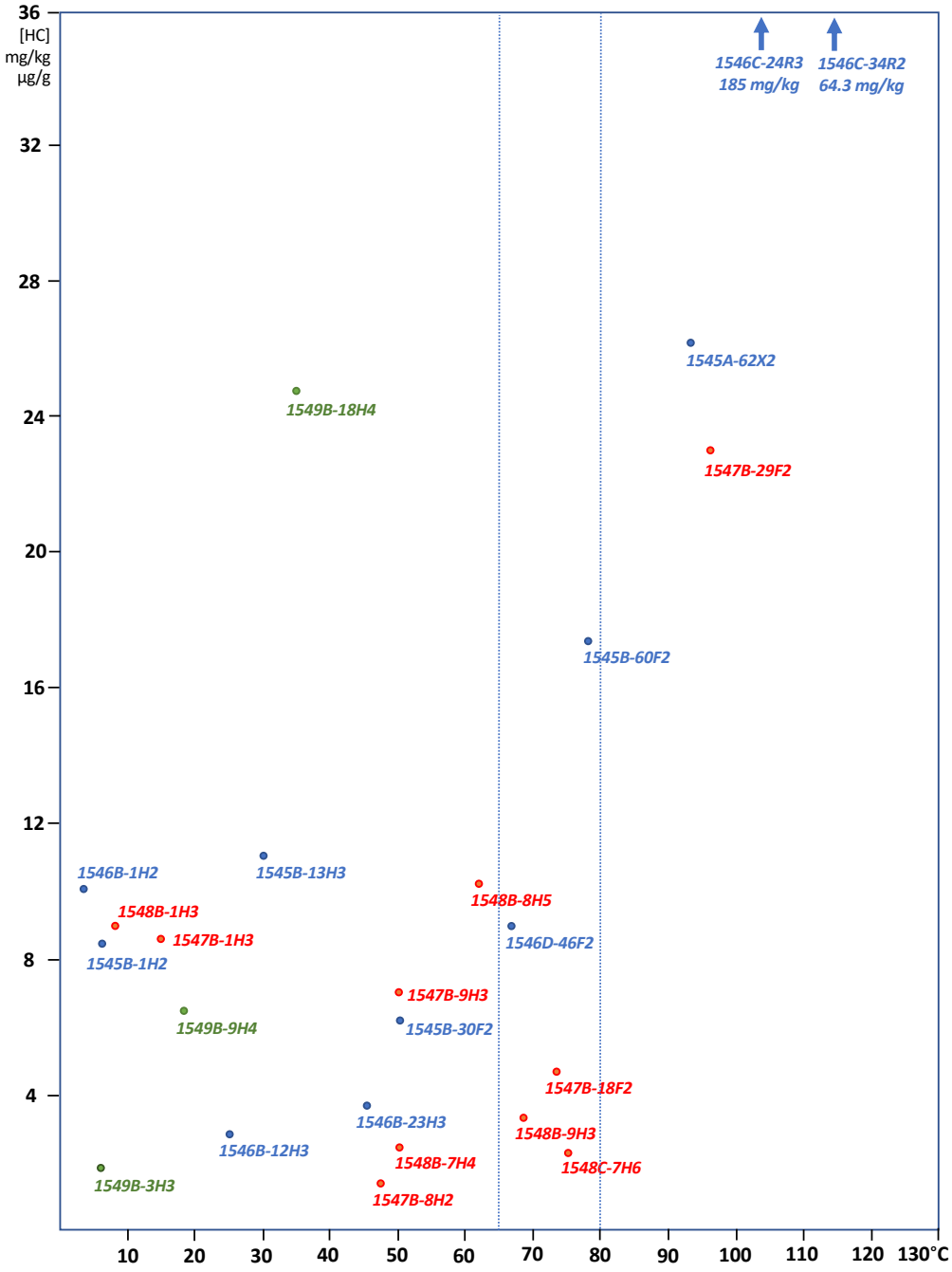
Accounting for Seawater and Laboratory Contamination

Deep-sea drilling employs a mixture of seawater and lubricant mud during drilling operations, which carries seawater-derived microbial contaminants into the sediment samples. Contamination monitoring (using polyfluorinated chemical tracers) was run throughout the drilling operations on selected samples (Lever et al., 2006), yet not every sediment sample of the thousands that are collected during an IODP expedition can be tested. It is therefore necessary to account for the potential presence of mixed microbial communities derived from seawater and drilling lubricant by sequencing a "drilling fluid control" and excluding the detected sequences from our metagenomic data. The second type of contaminants are introduced through DNA extraction kits, reagents and handling (the "kitome", e.g., Salter et al., 2014). This contaminant community needs to be accounted for based on blank extractions ("kit/method control") where no sediment sample is added. Because DNA for metagenomes was extracted in two different labs, the Amend lab (USC) and the Edgcomb lab (WHOI), our study includes two blank extraction controls. Contaminants from all three controls were identified by mapping control reads to the metagenome assembly and removing all contigs that received mapping with minimum 98% identity over minimum 75% of the read length. Contamination control samples are listed in Supplementary Data 2, and contaminant MAGs in Supplementary Data file 3.

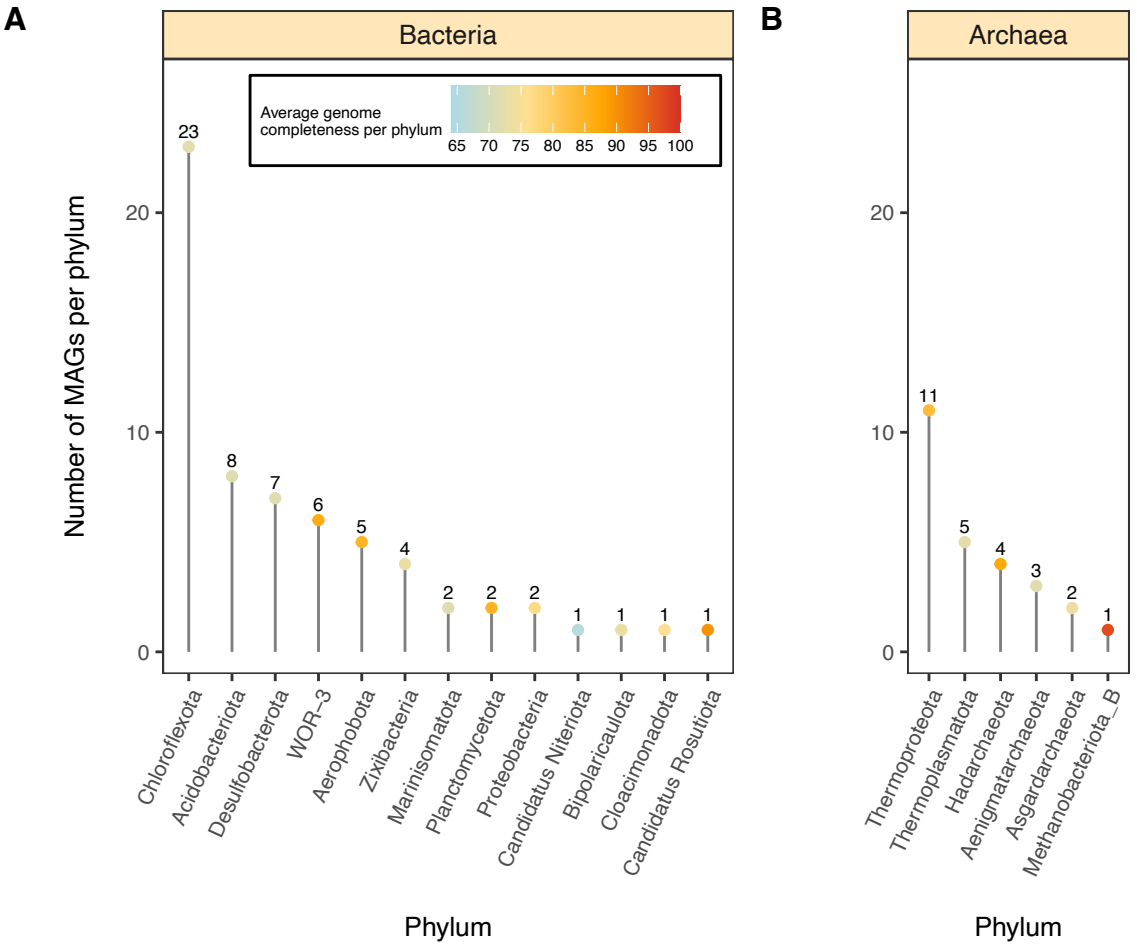
Supplementary Figures



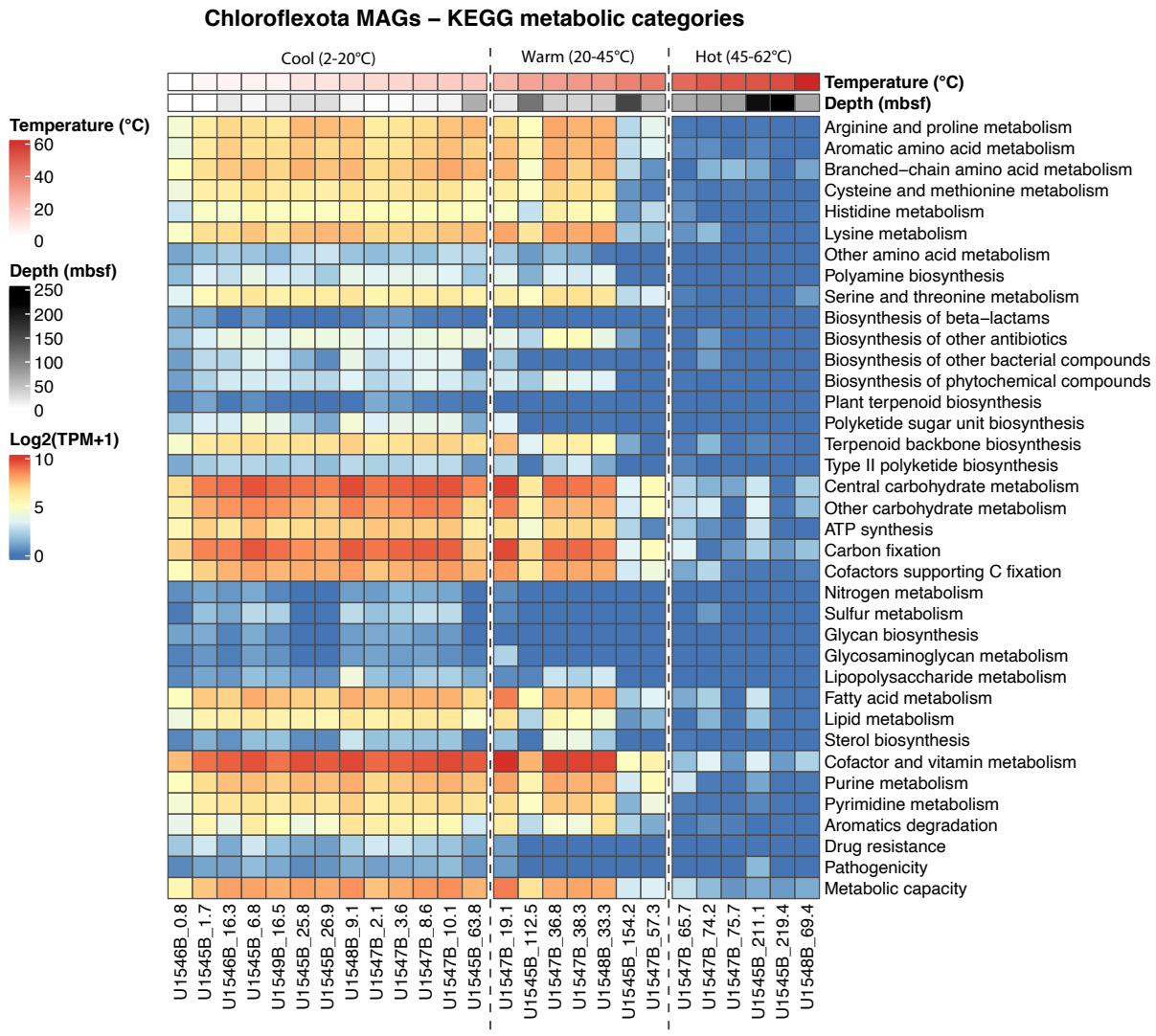
Supplementary Figure 4-1A. Total Petroleum hydrocarbon (C9-C44) content of Guaymas Basin sediments. Samples were analyzed at Alpha Analytical (Mansfield, MA, USA) for fingerprinting diagnostic compounds using EPA method 8015 (GC-FID; saturates) and a modified method 8270D (GCMS; PAHs), as detailed in Stout 2016. Hydrocarbon concentrations used in this figure are provided in the Source Data file.



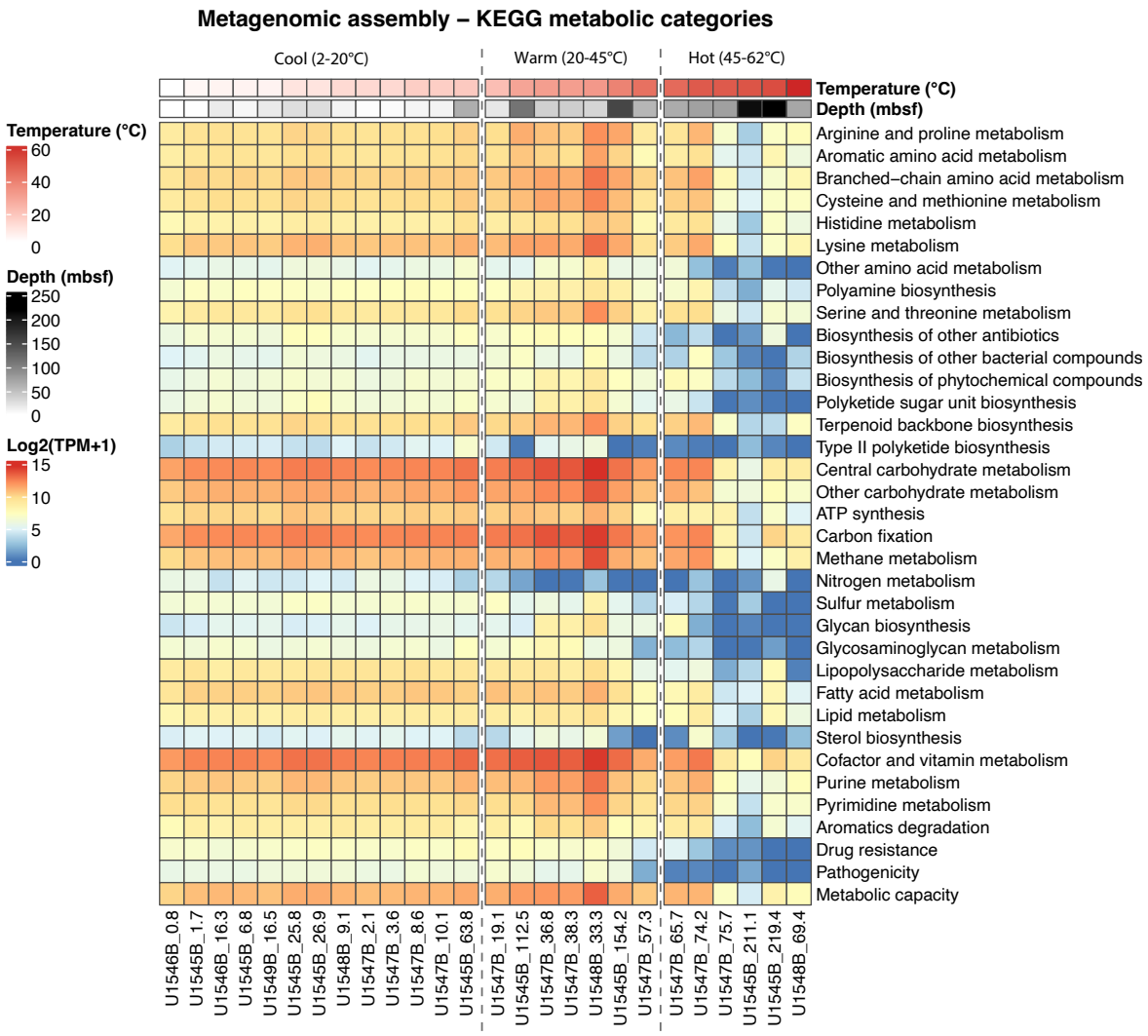
Supplementary Figure 4-1B. Total Saturated hydrocarbon content of Guaymas Basin sediments. Samples were analyzed at Alpha Analytical (Mansfield, MA, USA) for fingerprinting diagnostic compounds using EPA method 8015 (GC-FID; saturates) and a modified method 8270D (GCMS; PAHs), as detailed in Stout 2016. Hydrocarbon concentrations used in this figure are provided in the Source Data file.



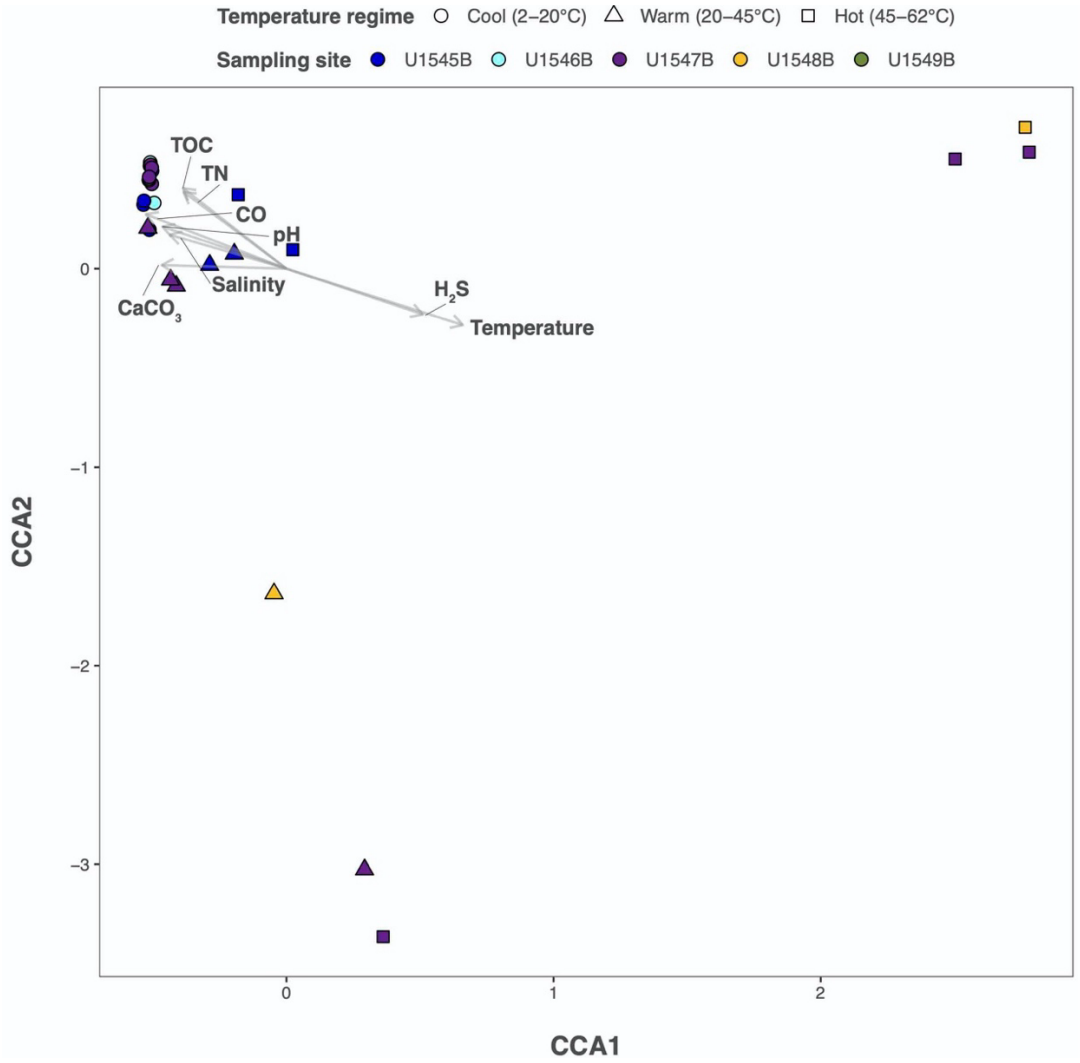
Supplementary Figure 4-2. MAG recovery frequency. Frequency of Guaymas Basin prokaryotic MAGs (≥ 50% completeness, ≤ 10% contamination) by bacterial (**A**) and archaeal phylum (**B**). Colored dots at the end of each line segment correspond to the mean genome completeness of the phylum; the number above the dot quantifies the number of genomes recovered from the phylum.



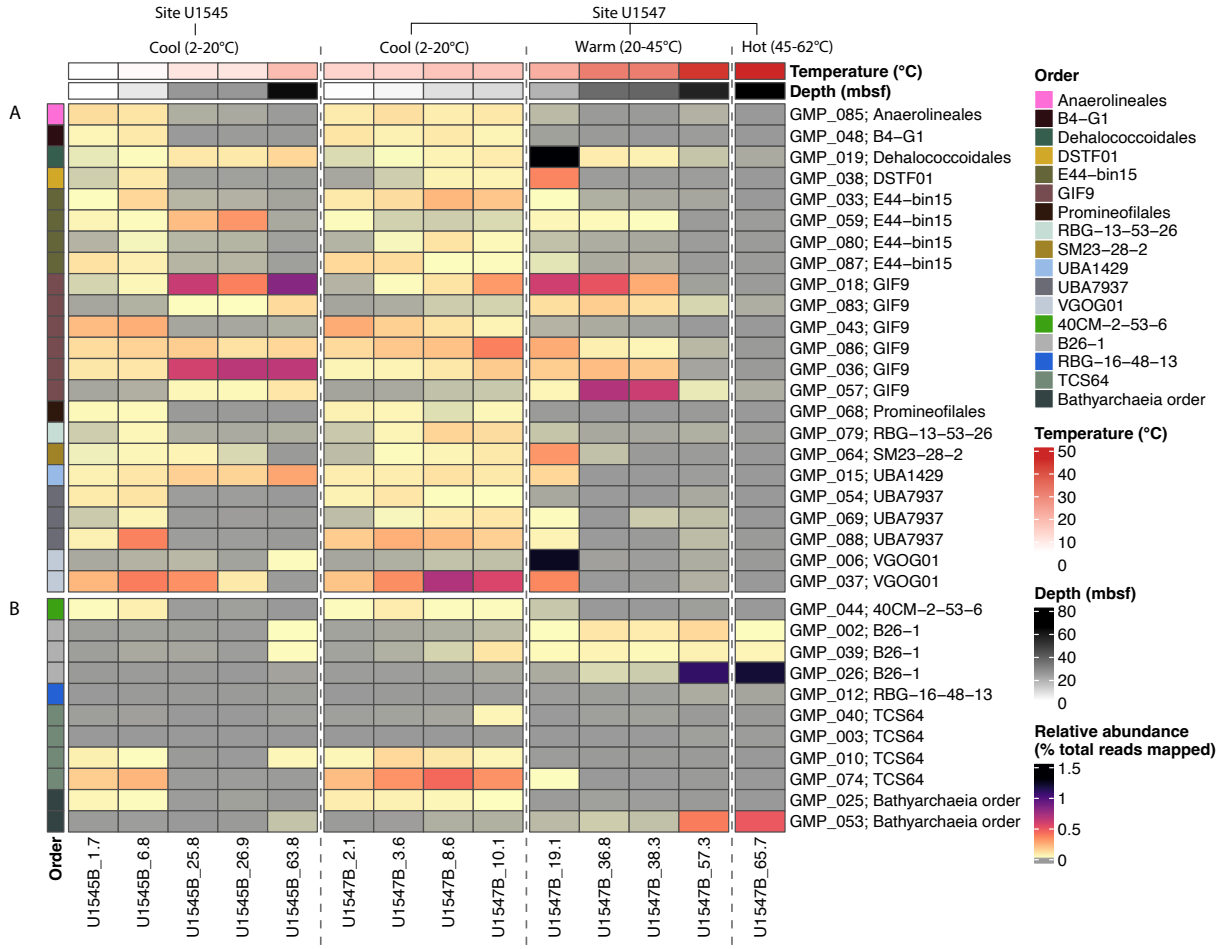
Supplementary Figure 4-3. Heatmap of read frequency for metabolic and cellular processes of Chloroflexi in Guaymas Basin metagenome samples. Metabolic and cellular processes were identified at the examined sites/depths using KofamScan. Processes involve pathways associated with energy-related metabolisms, genetic maintenance, survival strategies, and co-factor/vitamin biosynthesis. Sites and depths (mbsf) are given along the Y axis of the heatmap. Expression levels are normalized as log2 transformation of TPM+1 (a value of 1 was added to TPM values to avoid zeros). TPM: Transcripts Per Million.



Supplementary Figure 4-4. Heatmap of read frequency for metabolic and cellular processes of the subsurface microbial community in Guaymas Basin metagenome samples. Metabolic and cellular processes were identified at the examined sites/depths using KofamScan. Processes involve pathways associated with energy-related metabolisms, genetic maintenance, survival strategies, and co-factor/vitamin biosynthesis. Sites and depths (mbsf) are given along the Y axis of the heatmap. Expression levels are normalized as log2 transformation of TPM+1 (a value of 1 was added to TPM values to avoid zeros). TPM, Transcripts Per Million.



Supplementary Figure 4-5. Canonical Correlation Analysis (CCA) of subsurface MAGs and environmental parameters. The CCA plot depicts the correlation of Guaymas Basin MAG occurrence with in-situ environmental parameters. On the basis of Fisher's method for combining p-values, we show environmental variables with p-values < 0.05 resulting from a two-sided permutation test. Arrow direction indicates a positive or negative correlation of the environmental parameter with the ordination axes, for statistically significant ($p < 0.05$) environmental parameters (temperature, $p = 0.0001$; pH, $p = 0.0102$; salinity, $p = 0.0268$; hydrogen sulfide (H_2S), $p = 0.0053$; carbon monoxide (CO), $p = 0.004$; calcium carbonate ($CaCO_3$), $p = 0.0240$; total organic carbon (TOC), $p = 0.0084$; total nitrogen (TN), $p = 0.0103$). Arrow length reflects correlation strength between environmental parameter and MAG occurrence. Samples are color-coded by site, and their temperature regimes are indicated by shape (circles for 2-20°C; triangles for 20-45°C and squares for 45-62°C). Some samples are overlapping in their positions in the upper left corner of the plot.



Supplementary Figure 4-6. Relative abundance of Chloroflexota and Thermoproteota MAGs, identified by order, in metagenomic samples from drill sites U1545B and U1547B. The heatmap for Chloroflexota (**A**) and Thermoproteota (**B**) shows for each column the percentage of total pre-processed reads from a metagenomic sample that mapped to all Chloroflexota and Thermoproteota MAGs, respectively, in order of increasing sampling depth and temperature from left to right for samples from sites U1545B (left) and U1547B (right). MAG taxonomy is color coded to the left of the plot; taxonomic order names are given in the legend to the right of the heatmap. Temperatures and depths are color coded at the top of the plot.

Supplementary References

Altshuler, I., Raymond-Bouchard, I., Magnuson, E., Tremblay, J., Greer, C.W., Whyte, L.G. (2022). Unique high Arctic methane metabolizing community revealed through in situ 13CH4-DNA-SIP enrichment in concert with genome binning. *Scientific Reports*, *12*, 1160.

Anantharaman, K., Hausmann, B., Jungbluth, S.P., Kantor, R.S., Lavy, A., Warren, L. A., Rappé, M.S., Pester, M., Loy, A., Thomas, B.C. (2018). Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *The ISME Journal*, *12*, 1715–1728.

Baker, B.J., Lazar, C.S., Teske, A.P., Dick, G.J. (2015). Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome*, *3*, 1–12.

Bonato, P., Alves, L.R., Osaki, J.H., Rigo, L.U., Pedrosa, F.O., Souza, E.M., Zhang, N., Schumacher, J., Buck, M., Wassem, R. (2016). The NtrY–NtrX two-component system is involved in controlling nitrate assimilation in *Herbaspirillum seropedicae* strain SmR1. *The FEBS Journal*, *283*, 3919–3930.

Buessecker, S., Palmer, M., Lai, D., Dimapilis, J., Mayali, X., Mosier, D., Jiao, J.-Y., Colman, D.R., Keller, L. M., St. John, E. (2022). An essential role for tungsten in the ecology and evolution of a previously uncultivated lineage of anaerobic, thermophilic Archaea. *Nature Communications*, *13*, 3773.

Castelle, C.J., Hug, L.A., Wrighton, K.C., Thomas, B.C., Williams, K.H., Wu, D., Tringe, S.G., Singer, S.W., Eisen, J.A., Banfield, J.F. (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nature Communications*, *4*, 2120.

Chatterjee, M., Fan, Y., Cao, F., Jones, A. A., Pilloni, G., Zhang, X. (2021). Proteomic study of *Desulfovibrio ferrophilus* IS5 reveals overexpressed extracellular multi-heme cytochrome associated with severe microbiologically influenced corrosion. *Scientific Reports*, *11*, 1–11.

Chklovski, A., Parks, D.H., Woodcroft, B.J., Tyson, G.W. (2022). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *BioRxiv*, 2022–2027.

Colognori, D., Trinidad, M., Doudna, J.A. (2023). Precise transcript targeting by CRISPR-Csm complexes. *Nat Biotechnol*. 41, 1256-1264.

Cunin, R., et al.,"Biosynthesis and metabolism of arginine in bacteria." Microbiological reviews 50.3 (1986): 314-352.

DelVecchio, V.G., Kapatral, V., Elzer, P., Patra, G., Mujer, C. (2002). The genome of *Brucella melitensis*. *Veterinary Microbiology*, *90*, 587–592.

Deng, X., Okamoto, A. (2018). Electrode potential dependency of single-cell activity identifies the energetics of slow microbial electron uptake process. *Frontiers in Microbiology*, *9*, 2744.

Deng, X., Dohmae, N., Nealson, K.H., Hashimoto, K., Okamoto, A. (2018). Multi-heme cytochromes provide a pathway for survival in energy-limited environments. *Science Advances*, *4*, eaao5682.

Dodsworth, J. A., Gevorkian, J., Despujos, F., Cole, J.K., Murugapiran, S. K., Ming, H., Li, W.-J., Zhang, G., Dohnalkova, A., Hedlund, B.P. (2014). *Thermoflexus hugenholtzii* gen. nov., sp. nov., a thermophilic, microaerophilic, filamentous bacterium representing a novel class in the Chloroflexi, Thermoflexia classis nov., and description of

Thermoflexaceae fam. Nov. and Thermoflexales ord. nov. *International Journal of Systematic and Evolutionary Microbiology*, *64*, 2119–2127.
Dombrowski, N., Seitz, K.W., Teske, A.P., Baker, B.J. (2017). Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome*, *5*, 1–13.
Dombrowski, N., Teske, A.P., Baker, B.J. (2018). Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nature Communications*, *9*(1), 4999.
Dong, X., Greening, C., Rattray, J. E., Chakraborty, A., Chuvochina, M., Mayumi, D., Dolfing, J., Li, C., Brooks, J.M., Bernard, B.B. (2019). Metabolic potential of uncultured bacteria and archaea associated with petroleum seepage in deep-sea sediments. *Nature Communications*, *10*(1), 1816.
Edgcomb, V.P., Teske, A.P., Mara, P. (2022). Microbial hydrocarbon degradation in Guaymas Basin—exploring the roles and potential interactions of fungi and sulfate-reducing bacteria. *Frontiers in Microbiology* 13:831828, doi 10.3389/fmicb.2022.831828
Eisenhofer, R., Odriozola, I., Alberdi, A. Impact of microbial genome completeness on metagenomic functional inference. *ISME COMMUN.* 3: 12 (2023). https://doi.org/10.1038/s43705-023-00221-z
Evans, P.N., Boyd, J.A., Leu, A.O., Woodcroft, B.J., Parks, D.H., Hugenholtz, P., Tyson, G.W. (2019). An evolving view of methane metabolism in the Archaea. *Nature Reviews Microbiology*, *17*(4), 219–232.
Evans, P.N., Parks, D.H., Chadwick, G.L., Robbins, S.J., Orphan, V.J., Golding, S.D., Tyson, G.W. (2015). Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science*, *350*(6259), 434–438.
Farag, I.F., Biddle, J.F., Zhao, R., Martino, A.J., House, C.H., León-Zayas, R.I. (2020). Metabolic potentials of archaeal lineages resolved from metagenomes of deep Costa Rica sediments. *The ISME Journal*, *14*, 1345–1358.
Fincker, M., Huber, L.A., Orphan, V.J., Rappé, M.S., Teske, A., Spormann, A.M. (2020). Metabolic strategies of marine subseafloor Chloroflexi inferred from genome reconstructions. *Environmental Microbiology* 22, 3188-3203.
Flieder, M., Buongiorno, J., Herbold, C.W., Hausmann, B., Rattei, T., Lloyd, K.G., Loy, A., Wasmund, K. (2021). Novel taxa of Acidobacteriota implicated in seafloor sulfur cycling. *The ISME Journal*, *15*(11), 3159–3180.
Fullerton, H., Moyer, C.L. (2016). Comparative single-cell genomics of Chloroflexi from the Okinawa Trough deep-subsurface biosphere. *Applied and Environmental Microbiology*, *82*, 3000–3008.
Garber, A.I., Nealson, K.H., Okamoto, A., McAllister, S. M., Chan, C.S., Barco, R.A., Merino, N. (2020). FeGenie: a comprehensive tool for the identification of iron genes and iron gene neighborhoods in genome and metagenome assemblies. *Frontiers in Microbiology* 11:37.
Geller-McGrath, D., Konwar, K., Edgcomb, V.P., Pachiadaki, M., Roddy, J., Wheeler, T., McDermott, J.E. (2022). MetaPathPredict: A machine learning-based tool for predicting metabolic modules in incomplete bacterial genomes. *BioRxiv*, 2012–2022.
He, Y., Li, M., Perumal, V., Feng, X., Fang, J., Xie, J., Sievert, S.M., Wang, F. (2016). Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nature Microbiology*, *1*(6), 1–9.

Hernández, V.M., Arteaga, A., Dunn, M.F. (2021). Diversity, properties and functions of bacterial arginases. *FEMS Microbiology Reviews*, *45*, fuab034.
Islam, Z.F., Cordero, P.R.F., Feng, J., Chen, Y.-J., Bay, S.K., Jirapanjawat, T., Gleadow, R.M., Carere, C.R., Stott, M.B., Chiri, E. (2019). Two Chloroflexi classes independently evolved the ability to persist on atmospheric hydrogen and carbon monoxide. *The ISME Journal*, *13*, 1801–1813.
Jiang, Y., Shi, M., Shi, L. (2019). Molecular underpinnings for microbial extracellular electron transfer during biogeochemical cycling of earth elements. *Science China Life Sciences*, *62*, 1275–1286.
Jeuken, L. J. C., Hards, K., Nakatani, Y. (2020). Extracellular electron transfer: respiratory or nutrient homeostasis? *Journal of Bacteriology*, *202*, e00029-20.
Kapili, B.J., Barnett, S.E., Buckley, D.H., Dekas, A.E. (2020). Evidence for phylogenetically and catabolically diverse active diazotrophs in deep-sea sediment. *ISME J* 14, 971–983.
King, G. M. (2006). Nitrate-dependent anaerobic carbon monoxide oxidation by aerobic CO-oxidizing bacteria. *FEMS Microbiology Ecology*, *56*, 1–7.
Kishida, K., Sohrin, Y., Okamura, K., Ishibashi, J. (2004). Tungsten enriched in submarine hydrothermal fluids. *Earth and Planetary Science Letters*, *222*, 819–827
Kletzin, A., Adams, M.W.W. (1996). Tungsten in biological systems. *FEMS Microbiology Reviews*, *18*, 5–63.
Knittel, K., Boetius, A. (2009). Anaerobic oxidation of methane: progress with an unknown process. *Annual Review of Microbiology*, *63*, 311–334.
Kochetkova, T.V, Rusanov, I.I., Pimenov, N.V, Kolganova, T.V, Lebedinsky, A.V, Bonch-Osmolovskaya, E.A., Sokolova, T.G. (2011). Anaerobic transformation of carbon monoxide by microbial communities of Kamchatka hot springs. *Extremophiles*, *15*, 319–325.
Kung, J. W., Löffler, C., Dörner, K., Heintz, D., Gallien, S., van Dorsselaer, A., Friedrich, T., Boll, M. (2009). Identification and characterization of the tungsten-containing class of benzoyl-coenzyme A reductases. *Proceedings of the National Academy of Sciences*, *106*(42), 17687–17692.
Lever, M.A., Alperin, M., Inagaki, F., Nakagawa, S., Steinsbu, B.O., Teske, A., and IODP Expedition 301 Scientists. (2006). Trends in basalt and sediment core contamination during IODP Expedition 301. *Geomicrobiology Journal* 23, 517-530.
Light, S.H., Su, L., Rivera-Lugo, R., Cornejo, J. A., Louie, A., Iavarone, A.T., Ajo-Franklin, C. M., Portnoy, D. A. (2018). A flavin-based extracellular electron transfer mechanism in diverse Gram-positive bacteria. *Nature*, *562*, 140–144.
Lin,Y.-S., Heuer, V., Goldhammer, T., Kellermann, M.Y., Zabel, M., Hinrichs, K.U. (2012a). Towards constraining H_2 concentration in subseafloor sediment: a proposal for combined analysis by two distinct approaches. *Geochim. Cosmochim. Acta* 77, 186–201.
Lin, X., Kennedy, D., Fredrickson, J., Bjornstad, B., Konopka, A. (2012b). Vertical stratification of subsurface microbial community composition across geological formations at the Hanford Site. *Environmental Microbiology*, *14*, 414–425.
Liu, R., Wei, X., Song, W., Wang, L., Cao, J., Wu, J., Thomas, T., Jin, T., Wang, Z., Wei, W. (2022a). Novel Chloroflexi genomes from the deepest ocean reveal metabolic strategies for the adaptation to deep-sea habitats. *Microbiome*, *10*(1), 1–17.

Liu, S., Yu, S., Lu, X., Yang, H., Li, Y., Xu, X., Lu, H., Fang, Y. (2022b). Microbial communities associated with thermogenic gas hydrate-bearing marine sediments in Qiongdongnan Basin, South China Sea. *Frontiers in Microbiology*, 13:1032851.

Magnabosco, C., Ryan, K., Lau, M., Kuloyo, O., Lollar, S., Kieft, T., van Heerden E., Onstott, T.C. (2016). A metagenomic window into carbon metabolism at 3 km depth in Precambrian continental crust. *ISME J* 10, 730–741.

McGonigle, J. M., Lang, S. Q., Brazelton, W. J. (2020). Genomic evidence for formate metabolism by Chloroflexi as the key to unlocking deep carbon in Lost City microbial ecosystems. *Applied and Environmental Microbiology*, *86*, e02583-19.

Momper, L., Jungbluth, S. P., Lee, M. D., Amend, J. P. (2017). Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community. *The ISME Journal*, *11*, 2319–2333.

Murphy, C. L., Biggerstaff, J., Eichhorn, A., Ewing, E., Shahan, R., Soriano, D., Stewart, S., VanMol, K., Walker, R., Walters, P. (2021). Genomic characterization of three novel Desulfobacterota classes expand the metabolic and phylogenetic diversity of the phylum. *Environmental Microbiology*, *23*, 4326–4343.

Palmer, M., Covington, J.K., Zhou, E.-M., Thomas, S.C., Habib, N., Seymour, C.O., Lai, D., Johnston, J., Hashimi, A., Jiao, J.-Y. (2023). Thermophilic Dehalococcoidia with unusual traits shed light on an unexpected past. *The ISME Journal* 17, 952-966.

Pawlowski, K., Klosse, U., De Bruijn, F.J. (1991). Characterization of a novel *Azorhizobium caulinodans* ORS571 two-component regulatory system, NtrY/NtrX, involved in nitrogen fixation and metabolism. *Molecular and General Genetics MGG*, *231*, 124–138.

Oelgeschläger, E., Rother, M. (2008). Carbon monoxide-dependent energy metabolism in anaerobic bacteria and archaea. *Archives of Microbiology*, *190*, 257–269.

Orsi, W.D., Edgcomb, V.P., Christman, G.D., Biddle, J. F. (2013). Gene expression in the deep biosphere. *Nature*, *499*, 205–208.

Qi, Y.-L., Evans, P. N., Li, Y.-X., Rao, Y.-Z., Qu, Y.-N., Tan, S., Jiao, J.-Y., Chen, Y.-T., Hedlund, B. P., Shu, W.-S. (2021). Comparative genomics reveals thermal adaptation and a high metabolic diversity in "Candidatus Bathyarchaeia. *Msystems*, *6*(4), e00252-21.

Rogers, Timothy J., et al.,"Chemolithoautotroph distributions across the subsurface of a convergent margin." The ISME Journal 17.1 (2023): 140-150.

Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cook, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* **12**, 87.

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

Sekowska, A., Dénervaud, V., Ashida, H., Michoud, K., Haas, D., Yokota, A., Danchin, A. (2004). Bacterial variations on the methionine salvage pathway. *BMC Microbiology*, *4*, 1–17.

Shi, L., Dong, H., Reguera, G., Beyenal, H., Lu, A., Liu, J., Yu, H.-Q., Fredrickson, J.K. (2016). Extracellular electron transfer mechanisms between microorganisms and minerals. *Nature Reviews Microbiology*, *14*, 651–662.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S.G., Banfield, J.F. (2018). Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology*, *3*, 836–843.

Sirajuddin, S., Rosenzweig, A.C. (2015). Enzymatic oxidation of methane. *Biochemistry*, *54*, 2283–2294.
Sokolova, T.G., Henstra, A.-M., Sipma, J., Parshina, S.N., Stams, A.J.M., Lebedinsky, A.V. (2009). Diversity and ecophysiological features of thermophilic carboxydotrophic anaerobes. *FEMS Microbiology Ecology*, *68*, 131–141.
Speirs, L.B.M., Rice, D.T.F., Petrovski, S., Seviour, R.J. (2019). The phylogeny, biodiversity, and ecology of the Chloroflexi in activated sludge. *Frontiers in Microbiology*, *10*, 2015.
Speth, D.R., Yu, F.B., Connon, S.A., Lim, S., Magyar, J.S., Peña-Salinas, M.E., Quake, S.R., Orphan, V.J. (2022). Microbial communities of Auka hydrothermal sediments shed light on vent biogeography and the evolutionary history of thermophily. *The ISME Journal*, *16*, 1750–1764.
Stolz, John F., and Partha Basu. "Evolution of nitrate reductase: molecular and structural variations on a common function." Chembiochem 3.2-3 (2002): 198-206.
Stout, S. A. (2016). Oil spill fingerprinting method for oily matrices used in the Deepwater Horizon NRDA. *Environ. Forensics* 17, 218–243.
Suominen, S., Dombrowski, N., Sinninghe Damsté, J.S., Villanueva, L. (2021). A diverse uncultivated microbial community is responsible for organic matter degradation in the Black Sea sulphidic zone. *Environmental Microbiology*, *23*, 2709–2728.
Therkildsen, Mette S., and Gary M. King. "Urea production and turnover following the addition of AMP, CMP, RNA and a protein mixture to a marine sediment." Aquatic Microbial Ecology 10.2 (1996): 173-179.
Timmers, P.H.A., Welte, C.U., Koehorst, J.J., Plugge, C.M., Jetten, M.S.M., & Stams, A.J.M. (2017). Reverse methanogenesis and respiration in methanotrophic archaea. *Archaea*, *2017*: 1654237
Von Damm, K.L., Edmond, J.M., Measures, C.I., Grant, B. (1985). Chemistry of submarine hydrothermal solutions at Guaymas Basin, Gulf of California. *Geochimica et Cosmochimica Acta*, *49*, 2221–2237
Vanwonterghem, I., Evans, P.N., Parks, D.H., Jensen, P.D., Woodcroft, B.J., Hugenholtz, P., Tyson, G.W. (2016). Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. *Nature Microbiology*, *1*, 1–9.
Venceslau, S.S., Stockdreher, Y., Dahl, C., Pereira, I.A.C. (2014). The "bacterial heterodisulfide" DsrC is a key protein in dissimilatory sulfur metabolism. *Biochimica et Biophysica Acta-Bioenergetics*, *1837*, 1148–1164.
Villanueva, L., von Meijenfeldt, F.A.B., Westbye, A.B., Yadav, S., Hopmans, E.C., Dutilh, B. E., Damsté, J.S.S. (2021). Bridging the membrane lipid divide: bacteria of the FCB group superphylum have the potential to synthesize archaeal ether lipids. *The ISME Journal*, *15*, 168–182.
Vuillemin, A., Kerrigan, Z., D'Hondt, S., Orsi, W.D. (2020). Exploring the abundance, metabolic potential and gene expression of subseafloor Chloroflexi in million-year-old oxic and anoxic abyssal clay. *FEMS Microbiology Ecology*, *96*, fiaa223.
Vorholt, J.A., Marx, C.J., Lidstrom, M.E., Thauer, R.K. (2000). Novel formaldehyde-activating enzyme in *Methylobacterium extorquens* AM1 required for growth on methanol. *Journal of Bacteriology*, *182*, 6645–6650.
Wegener, G., Laso-Pérez, R., Orphan, V. J., Boetius, A. (2022). Anaerobic degradation of alkanes by marine archaea. *Annual Review of Microbiology*, *76*, 553–577.

Wong, H.L., MacLeod, F.I., White, R.A., Visscher, P.T., Burns, B.P. (2020). Microbial dark matter filling the niche in hypersaline microbial mats. *Microbiome*, *8*, 1–14.
Xiong, L., Jian, H., Zhang, Y., Xiao, X. (2016). The two sets of DMSO respiratory systems of *Shewanella piezotolerans* WP3 are involved in deep sea environmental adaptation. *Frontiers in Microbiology*, *7*, 1418.
Zhang, W., Ding, W., Yang, B., Tian, R., Gu, S., Luo, H., Qian, P.-Y. (2016). Genomic and transcriptomic evidence for carbohydrate consumption among microorganisms in a cold seep brine pool. *Frontiers in Microbiology*, *7*, 1825.
Zhou, Z., St John, E., Anantharaman, K., Reysenbach, A.-L. (2022). Global patterns of diversity and metabolism of microbial communities in deep-sea hydrothermal vent deposits. *Microbiome*, *10*, 1–22.
Zumft, Walter G. "Cell biology and molecular basis of denitrification." Microbiology and molecular biology reviews 61.4 (1997): 533-616.

Chapter 5

MetaPathPredict-E: a machine-learning based tool for prediction of metabolic modules in incomplete eukaryotic genomes and transcriptomes

Statement of contribution

D.G.M. designed the research project, downloaded and extracted features from genomic and transcriptomic data, and trained machine learning models with input from Kishori M. Konwar, Harriet Alexander, and Virginia P. Edgcomb. D.G.M. developed the MetaPathPredict-E software. D.G.M. wrote the manuscript with feedback from Kishori M. Konwar, Harriet Alexander, and Virginia P. Edgcomb.

Abstract

Deciphering the metabolic potential within eukaryotic 'omics datasets recovered from environmental samples is challenging, and relative to prokaryotes, there are few high-quality representative genomes and transcriptomes of eukaryotic species in reference databases. With improvements in sequencing technology and assembly algorithms, genomes and transcriptomes of eukaryotes from environmental samples can now be reconstructed. However, this often yields partial eukaryotic genomes that give incomplete insight into their full metabolic potential. In computational pipelines designed to elucidate eukaryotic metabolism, the application of machine learning for the prediction of KEGG module presence or absence in incomplete genomes and transcriptomes is virtually unexplored. Here, we present MetaPathPredict-E, an extension of the MetaPathPredict software that utilizes machine learning models to predict the metabolic potential of incomplete eukaryotic genomes and transcriptomes. MetaPathPredict-E has a command line interface, can also be run as a Python module, and both formats can be utilized on a local operating system or on a computing cluster. In benchmarking of its classification models, MetaPathPredict-

E exhibited accurate predictions of KEGG module presence/absence within highly incomplete genomes and transcriptomes.

Introduction

Eukaryotic metabolism exhibits remarkable diversity across its kingdoms. This metabolic versatility is reflected by the adaptations and lifestyles of these organisms, with varying energy requirements and trophic modes (Weber et al., 2007; Harwood et al., 2009; Ginger et al., 2010; Edwards et al., 2023; Müller et al., 2012; Alexander et al., 2023). Protists are a diverse and eclectic group of single-celled eukaryotic organisms (Adl et al., 2019) encompassing member clades with a range of trophic modes (Lambert et al., 2022; Alexander et al., 2023). Many protists are obligate phototrophs or heterotrophs, though there are also mixotrophic protists that can switch between these two trophic modes through various mechanisms (Gast et al., 2006; Stoecker et al., 2009). Protists also exhibit both aerobic and anaerobic metabolisms (Muller et al., 1991; Ginger et al., 2010; Gawryluk et al., 2021; Fenchel and Finlay 1991; Fenchel 2011). The Archaeplastida, which contain within its member clades the Streptophyta and Chlorophyta, utilize the energy stored in glucose through the process of photosynthesis to fuel their cellular activities including growth, development, and reproduction (Leegood et al., 2006). Archaeplastida include taxa with remarkable metabolic flexibility, demonstrated by their ability to switch between photoautotrophy and mixotrophy, supplementing photosynthesis with the absorption of organic nutrients from the environment (Těšitel et al., 2018). Fungi play critical roles in nutrient cycling by breaking down complex organic matter into simpler molecules that can be utilized by other organisms (Harley, 1971; Finlay, 2004; Rashid et al., 2016). Under aerobic conditions, fungi utilize the ubiquitous pathway of oxidative phosphorylation to break down organic matter. However, in anaerobic environments, fungal species perform fermentative metabolisms and consort with chemoautotrophic prokaryotes (Kazda et al., 2014; Drake and Ivarsson, 2018).

The metabolic pathways of eukaryotic organisms encompassing processes involved in phototrophy, heterotrophy, and mixotrophy are encoded within their genomes. With the advent of high-throughput sequencing technologies and advances in computational biology, a suite of bioinformatics tools has emerged to unravel the complexities of eukaryotic metabolism and predict trophic modes. The process of identifying and functionally annotating genes within a genome is one of the first steps in predicting metabolism. Tools such as AUGUSTUS (Stanke et al., 2006),

the GeneMark software suite (Lukashin and Borodovsky, 1998), and the BRAKER3 pipeline (Gabriel et al., 2023) are designed primarily for predicting genes within eukaryotic genomes. Once genes are identified and proteins are predicted, their functional roles are often inferred with sequence alignment methods such as BLAST (Altschul et al., 1990) and DIAMOND (Buchfink et al., 2015), which compare predicted sequences to known protein databases. Annotation of protein families and domains using Hidden Markov Models (HMMs) is another common approach to gene annotation. Tools such as KofamScan (Aramaki et al., 2020), eggNOG Mapper (Huerta-Cepas et al., 2017), and InterProScan (Jones et al., 2014) utilize profile HMMs to annotate prokaryotic and eukaryotic proteins. The resulting gene annotations can then be mapped to metabolic pathways from databases including KEGG (Kanehisa, 2002) and MetaCyc (Karp et al., 2002). The presence of specific gene orthologs in a genome is thus thought to be indicative of an organism's metabolic potential.

Genomic sequencing of eukaryotes has yielded a growing number of high-quality draft genomes, yet reconstructing eukaryotic genomes remains a significant challenge due to the inherent complexity of their genomic architecture (Tørresen et al., 2019). Since expansive noncoding and repetitive genomic regions pose challenges for sequence read assembly algorithms, transcriptomic sequencing presents a compelling alternative approach. This technique sequences only the RNA transcripts of an organism, bypassing these complex non-coding regions of the genome to greatly simplify the read assembly process. Software such as Trinity (Haas et al., 2013) is used to assemble transcripts from transcriptomic data, while TransDecoder (https://github.com/TransDecoder/) or GeneMarkS-T (Tang et al., 2015) can predict coding regions of transcripts that can then be functionally annotated. Transcriptomic data only provides sequencing information about actively transcribed genes, and does not reflect information about genes that were not expressed at the time that a sample was collected. The rapid degradation of RNA by cellular machinery presents an additional challenge that can bias transcriptomic sequencing results, confounding transcript quantification and other downstream analyses (Gallego Romero et al., 2014).

Machine learning models are also increasingly being deployed in bioinformatics tools to analyze eukaryotic datasets, such as to predict their trophic mode. By analyzing large genomic and transcriptomic datasets, learning algorithms can identify patterns and relationships that may not be readily apparent through manual methods. Machine learning architectures including random

forests have been used to successfully predict trophic modes within genomic and transcriptomic datasets with a high degree of accuracy (Lambert et al., 2022; Alexander et al., 2023). A major challenge, however, is that the coding regions of genomes and transcriptomes of eukaryotes reconstructed from environmental samples are often incompletely recovered. Machine learning has been shown to be useful in analyzing poorly characterized organisms and/or incompletely recovered proteomes (Lambert et al., 2022; Alexander et al., 2023).

Machine learning methods to predict metabolic pathway completeness in eukaryotes, however, are lacking. We aim to fill this gap with the development of a new tool, described here, that is designed to predict modules within eukaryotic proteomes. Building upon the open-source tool MetaPathPredict (Geller-McGrath et al., 2024), we introduce MetaPathPredict-E, a deep learning-powered extension for eukaryotic metabolic pathway prediction. MetaPathPredict-E addresses a gap in metabolic pathway reconstruction tools for eukaryotes that harnesses the utility and predictive power of genomic and transcriptomic data. It directly connects curated KEGG metabolic knowledge with machine learning, enabling the efficient reconstruction and prediction of KEGG modules within diverse but incomplete eukaryotic datasets including isolate genomes and transcriptomes, metagenome-assembled genomes (MAGs), and single amplified genomes (SAGs).

The models contained within MetaPathPredict-E were trained on gene annotation data from taxonomically diverse eukaryotic isolate genomes and transcriptomes obtained from the NCBI RefSeq (O'Leary et al., 2016), JGI GOLD (mirrored in NCBI GenBank; Clark et al., 2016), NCBI TSA (Wheeler et al., 2007), and MMETSP (Keeling et al., 2014) databases. The metabolic modules defined in the KEGG database serve as the tool's reference for both reconstructing and predicting the metabolic potential of input gene annotations. The KEGG database contains metabolic pathway information for all domains of life, including eukaryotes. KEGG modules are functional units within KEGG pathways that consist of ordered sequences of KEGG reactions. Examples of modules include individual carbon fixation and vitamin biosynthesis pathways, as well as modules encoding transporters. MetaPathPredict-E has a command line interface for execution on a local operating system or computing cluster and is available as part of the MetaPathPredict Python module on GitHub (https://github.com/d-mcgrath/MetaPathPredict-E).

A schematic of the MetaPathPredict-E pipeline is shown in Figure 1. The tool accepts gene annotations from one or more genomes or transcriptomes with associated KEGG ortholog (KO)

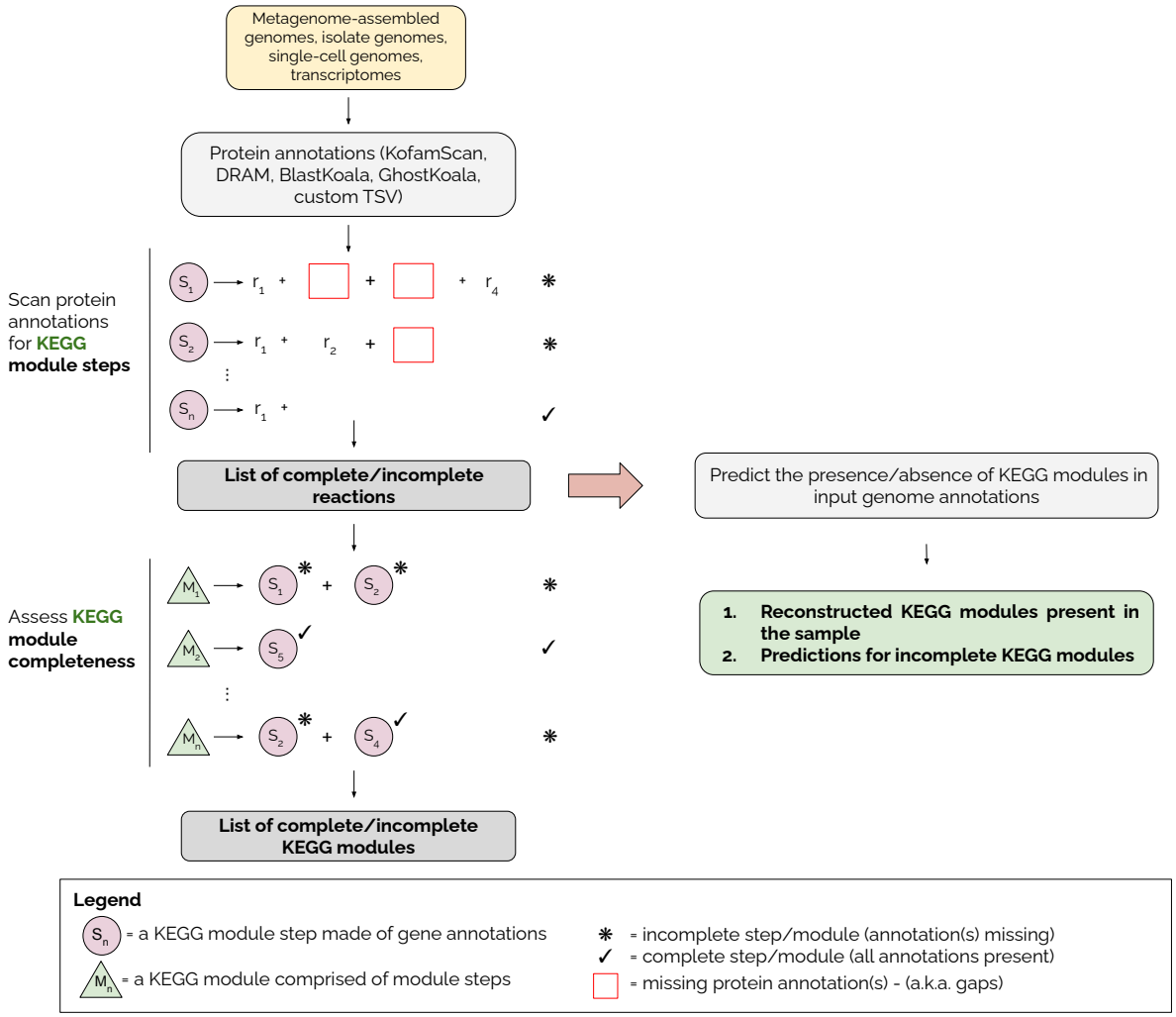


Figure 5-1. Overview of the MetaPathPredict-E pipeline. Input gene annotations are first read into MetaPathPredict-E, then the data are scanned for any complete KEGG modules and are formatted into a feature matrix. Predictions are then made for all incomplete modules (or modules specified by the user). A summary and detailed reconstruction and prediction output, along with gapfilling options are returned as the final output.

gene identifiers. Due to the potential incompleteness of input datasets, MetaPathPredict-E uses a dual-pronged approach. First, it maps input annotations to KEGG modules, and reconstructs both complete and fragmented modules within the input data. Subsequently, it predicts the presence or absence of all incomplete modules. Input files to MetaPathPredict-E can be from the output of tools including KofamScan (Aramaki et al., 2020), DRAM (Shaffer et al., 2020), blastKOALA (Kanehisa et al., 2016), ghostKOALA (Kanehisa et al., 2016), or a custom tab-separated values (TSV) file of KO gene identifier annotations. The tool contains individual multi-label classification models trained specifically for metabolic predictions of Fungi, Streptophyta, Chlorophyta, Excavata, Stramenopiles, Alveolates, Rhizaria, and Metazoa (as taxonomically defined by Adl. et

al., 2012). MetaPathPredict-E also contains a general model that was trained to predict KEGG modules that were prevalent in the genome and transcriptome data across all the aforementioned eukaryotic supergroups, plus clades that lacked sufficient representation in the downloaded datasets used for model creation. The distribution of module classes for each model is shown in Supplementary Figure 1. Each of its deep learning models contained distinct network architectures determined individually through Bayesian optimization. MetaPathPredict-E made predictions with a macro F1 score (a summary metric for multi-label predictive performance) of at least 0.81 on all held-out test datasets of proteomes containing at least 30% of their original gene content. The models were trained to classify the presence or absence of KEGG modules that were present in at least 25% of the proteomes within each of the training datasets. False negative predictions were uncommon during model evaluation, while false positives increased when predictions were made on highly incomplete (<30% completeness) gene annotation information. We believe that MetaPathPredict-E will help facilitate studies of eukaryotic genomes and transcriptomes from environmental samples and will help to decipher their metabolic potential.

Materials and Methods

Filtering genome database metadata, downloading genomes and transcriptomes, and annotating genes

A total of 554 genomes (and their predicted gene sequences, where available) were downloaded from the NCBI RefSeq database using the ncbi-genome-download tool with the flags "--formats all", "--section refseq". The "advanced search" option was used to extract the information for all fungal genomes present in the JGI GOLD database, including mirror links to download the genomes from the NCBI GenBank database. The database metadata file for release v.9 was downloaded (https://gold.jgi.doe.gov/downloads) and filtered to keep the information for eukaryotic genomes that were extracted through searching the database. A total of 1,838 eukaryotic genomes from JGI GOLD with mirror links to NCBI were downloaded using the ncbi-genome-download command line tool (Blin 2023). The NCBI RefSeq and JGI GOLD genomes (extracted from NCBI GenBank) were cross-referenced, and in instances of duplicate representatives the RefSeq version of the genome was retained. A total of 2,473 genomes were further downloaded from NCBI using the NCBI Datasets CLI (Sayers et al., 2021) with the "datasets download genome

taxon" command including the "--dehydrated" and "--include genome,protein,seq-report" flags followed by the "datasets rehydrate" command. A set of 1,562 assembled transcriptomes from the NCBI TSA database and protein-coding sequences of 638 improved MMETSP (Johnson et al., 2019) transcriptome assemblies and their predicted gene content (https://zenodo.org/records/3247846) were manually downloaded. Protein-coding sequences were predicted for the NCBI TSA transcriptomic assemblies using the "TransDecoder.LongOrfs" and "TransDecoder.Predict" scripts from TransDecoder (https://github.com/TransDecoder/). Only select genomes downloaded from NCBI had available gene predictions, and therefore BRAKER3 version 3.0.7 with default parameters (Gabriel et al., 2023) was utilized to predict genes (without RNA-Seq/protein evidence) for all the genomes extracted from NCBI. For genomes with both pre-existing gene predictions and BRAKER3 gene predictions, the BUSCO scores were compared and the set of protein sequences with the highest BUSCO score was retained to avoid incorporating duplicate proteomes into training datasets. The completeness of all genomes and transcriptomes was assessed using BUSCO v5.4.7 (Simão et al., 2015). All proteomes with ≥ 80% BUSCO completeness (measured as the total percentage of single copy, duplicate, and fragmented BUSCO genes present) were retained for downstream model training.

The KofamScan command line tool (Aramaki et al., 2020) was utilized to assign functional annotations to all downloaded predicted protein coding regions of genomes and transcriptomes. Leveraging the KEGG database's KOfam HMM collection (Kanehisa et al., 2002; ftp://ftp.genome.jp/pub/db/kofam/), KofamScan assigned KO gene identifiers to genome and transcriptome predicted protein sequences. Only KofamScan-derived annotations surpassing the HMM's adaptive scoring threshold defined by KofamScan were retained for downstream model training. These HMM scoring thresholds were more robust than the use of fixed e-value cutoffs and minimized the inflation of our training and assessment datasets with potentially inaccurate annotations.

Formatting gene annotation data, fitting KEGG module classification models

One general model was trained on the entire dataset of available eukaryotic genomes and transcriptomes that had at least 80% BUSCO completeness. For this model, the full dataset (n = 5,184) was split such that 75% of genomes and transcriptomes were used for training with the remaining 25% reserved as a test dataset. The training dataset was further split into 80%

training/20% validation sets. Each sample in the three datasets contained a vector of KO gene identifier (protein family) presence/absence encoded by ones and zeroes, respectively. The labels for the training data were the presence or absence of complete KEGG modules. Additional training, validation and test datasets were curated for major groups of eukaryotes for which sufficient high-quality proteomes were available, including Metazoa, Fungi, Chlorophyta, Streptophyta, Stramenopiles, Alveolata, Rhizaria, and Excavata.

Training and test datasets contained both complete and incomplete gene annotations of eukaryotic proteomes from a diverse array of phyla. The number of genomes and transcriptomes in each training dataset, as well as the number of features and labels are presented in Supplementary Table 1. Taxonomic groups for which very little data (< 100 genomes and/or transcriptomes) were available included the Alveolata and Rhizaria. The taxonomic distribution of all genome and transcriptome data used for model training is shown in Supplementary Figure 2. The incomplete annotations were constructed from proteome gene annotation samples with a BUSCO completeness of at least 80% that were randomly down-sampled to retain 5-95% of the original gene content, while the KEGG modules labels associated with the samples remained unchanged. Each proteome was downsampled 10 times at each sampling rate with a distinct random number generator seed and all complete and down-sampled versions were retained for training and testing. Model training datasets varied in sample size, depending on taxon and availability of genomic/transcriptomic data (see Supplementary Table 1).

The gene copy number data of the downloaded proteomes was formatted into a matrix containing protein family presence/absence (1 or 0, respectively) in columns and proteome samples in rows. The multiple labels in each dataset were the presence/absence (1 or 0) of KEGG modules in the samples as determined using the KEGG modules downloaded from the KEGG database and the Anvi'o Python module (Eren et al., 2015). The "unroll_module_definition" function from the Anvi'o module was utilized with downloaded KEGG module data to create a list of all possible KEGG Ortholog combinations to complete each module. For the module to be categorized as present, at least one possible (non-redundant) combination of every step of the module had to be present in a proteome, otherwise it was designated as absent. The models were constructed to predict KEGG modules for which at least 25% of all proteomes in the training data contained the full module. The models were trained using the gene annotation data of the genomes and transcriptomes consolidated from the NCBI RefSeq, JGI GOLD, NCBI TSA, and MMETSP

databases. The constructed models classify the multi-label presence or absence of complete KEGG modules based on the gene annotations of a genome or transcriptome.

A deep learning classification approach was chosen to model the relationship between protein family presence/absence and the presence/absence of metabolic modules. MetaPathPredict-E is built using the Keras deep learning library (Chollet et al., 2015). L2-regularization was utilized to adjust hidden unit weights during training, with a learning rate of 0.001. Features used in the training datasets were protein family presence/absence in the form of KO gene identifiers. The deep learning architecture of each of MetaPathPredict-E's multi-label models was determined using a Bayesian optimization hyperparameter tuning method from the Keras Tuner library (O'Malley et al., 2019). The number of hidden layers and associated hidden units varied by model but were fully connected in each case. The output layer of the models contained a set of nodes to predict a multi-label output of KEGG module presence/absence.

Stratified sampling is a sampling method that ensures that all groups within the training and test data are represented in the same proportion as they are in the population as a whole. A multi-label stratified sampling method approach was used to generate 75% train/25% test dataset splits that each contained data observations with preserved proportions of positive ('KEGG module present') and negative ('KEGG module absent') classes for each of the training datasets. The training datasets were further separated using the same method into 80% train/20% validation dataset splits to fit deep learning models. Sample weights were also applied to each of the training datasets, to penalize misclassification more harshly during training of samples containing modules that were less prevalent in the training data.

The binary cross entropy loss function was used in tandem with the Adaptive Moment Estimation (Adam) optimizer. The input and hidden layers utilized the rectified linear unit (ReLu) activation function; the output layer contained a sigmoid activation function. Dropout (Srivastava et al., 2014) was applied to all hidden layers (with dropout probabilities determined through Bayesian optimization) to avoid overfitting the training data. Batch normalization (Ioffe et al., 2015) was also applied at all layers except the final layer to prevent overfitting the training data and to speed up model convergence.

Evaluating MetaPathPredict-E on test genomes and transcriptomes randomly down-sampled to simulate varying degrees of proteome incompleteness

MetaPathPredict-E's models were validated on held-out test sets consisting of a combination of near-complete to complete and simulated incomplete genomes and transcriptomes, and the performance metrics were extracted using the Scikit-learn (Pedregosa et al., 2011) Python module. The annotations in each test set were created by randomly downsampling near-complete/complete proteomes to simulate recovered gene annotations from incomplete genomes. Five percent to 95% of genes from each annotation set were randomly retained (in increments of 5%) and used as input for MetaPathPredict-E predictions of KEGG module presence/absence. The performance metrics used in evaluating the model were macro precision, macro recall, and macro F1 score (Table 1).

Metric	Definition
Precision	true positive/(true positive + false positive)
Recall	true positive/(true positive + false negative)
F1 score	2 × ((precision × recall)/(precision + recall))

Table 5-1. Definitions of machine learning model performance metrics used to assess the MetaPathPredict-E model. The macro F1 score, precision, and recall are the average F1 score, precision, and recall across all labels, with each label weighted uniformly.

Testing MetaPathPredict-E with a set of high-completeness MAGs from built environment metagenomes

MetaPathPredict-E was also validated on a test set of gene annotations extracted from MAGs recovered from hospital room and infant gut metagenomes (Olm et al., 2019). MAG protein sequences were downloaded (https://github.com/MrOlm/InfantEukaryotes/tree/master) and re-annotated using KofamScan. The completeness of the MAG protein datasets was assessed using BUSCO. The set of protein annotations were filtered to retain 8 MAGs with a BUSCO completeness of at least 80%. The method for the assessment of MetaPathPredict-E's models was the same as was described above on held-out test datasets during model training.

Testing MetaPathPredict-E with a set of low completeness genomes from the NCBI, JGI GOLD, and MMETSP databases

Further validation of MetaPathPredict-E utilized gene annotations from the low completeness (<80% BUSCO completeness) proteomes acquired from the NCBI, JGI, and MMETSP databases. MetaPathPredict-E's performance was evaluated on these MAGs using the detection rate method, which assessed the percentage of fully present KEGG modules in the data that MetaPathPredict-E was able to correctly classify. The low completeness proteomes were separated by eukaryotic supergroup, with each model making predictions for its associated group. The general model was used to make predictions on the comprehensive set of low completeness proteomes.

Gapfilling for incomplete modules predicted as present

MetaPathPredict-E provides enzyme gapfilling options for KEGG modules predicted as present by suggesting putative KO gene annotations missing from an input proteome's gene annotations that could fill in missing gaps in predicted modules.

Results and Discussion

MetaPathPredict-E is a machine learning-based tool designed to predict the presence of metabolic modules in eukaryotic genomes and transcriptomes that are missing gene annotation information due to incomplete sequencing or annotation of a proteome. Its models were trained on both complete and down-sampled eukaryotic proteomes, where protein annotations were randomly removed in increasing increments. High-completeness proteomes (BUSCO completeness score of at least 80%, calculated as the proportion of single copy, duplicated, and fragmented BUSCO genes present; see Figure 2) were labelled as containing KEGG modules if all the genes necessary to non-redundantly complete the reaction steps of a KEGG module were present, otherwise the module was labeled as absent. All labels were preserved in down-sampled versions of the high-completeness proteomes used during model training. To assess MetaPathPredict-E's performance, we implemented a variety of benchmarking experiments in which we tested the tool on simulated and unmodified proteome datasets.

MetaPathPredict-E's performance metrics on held-out test data from Chlorophyta, Streptophyta, Alveolata, Rhizaria, Stramenopiles, Excavata, Fungi, and Metazoa proteomes suggest its models predict with high fidelity (macro F1 score ≥ 0.81) when at least 30% of gene

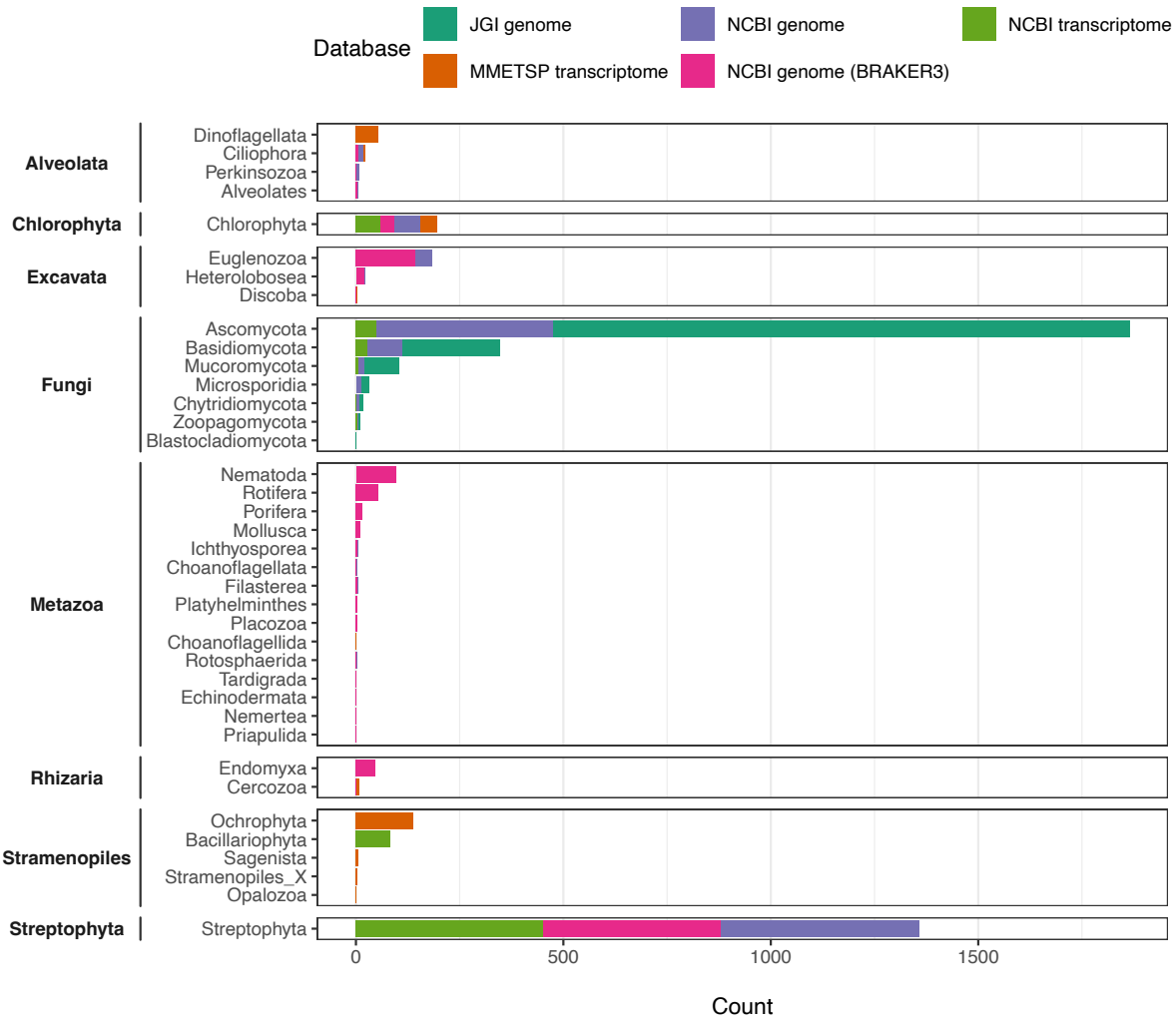


Figure 5-2. Bar chart showing the distribution of phyla of all genomes and transcriptomes utilized for model training. Eukaryotic groups are displayed in bold text to the far left. One model was trained for each cluster of phyla indicated in the y-axis (one model per group). Bars are colored to show the distribution of databases the training genomes and transcriptomes were downloaded from. All genomes and transcriptomes had a BUSCO completeness score ≥ 80%. An addition general model was trained to classify KEGG modules prevalent across all eukaryotic data that was at least 80% complete. See Supplementary Figure 1 for the full distribution of proteomes used to train the general model; see Supplementary Table 1 for more training data information.

annotations are present (Figure 3). The efficacy of MetaPathPredict-E's models was assessed using artificially incomplete gene annotation data, unmodified low-BUSCO completeness proteomes, and MAGs recovered from built environment metagenomes.

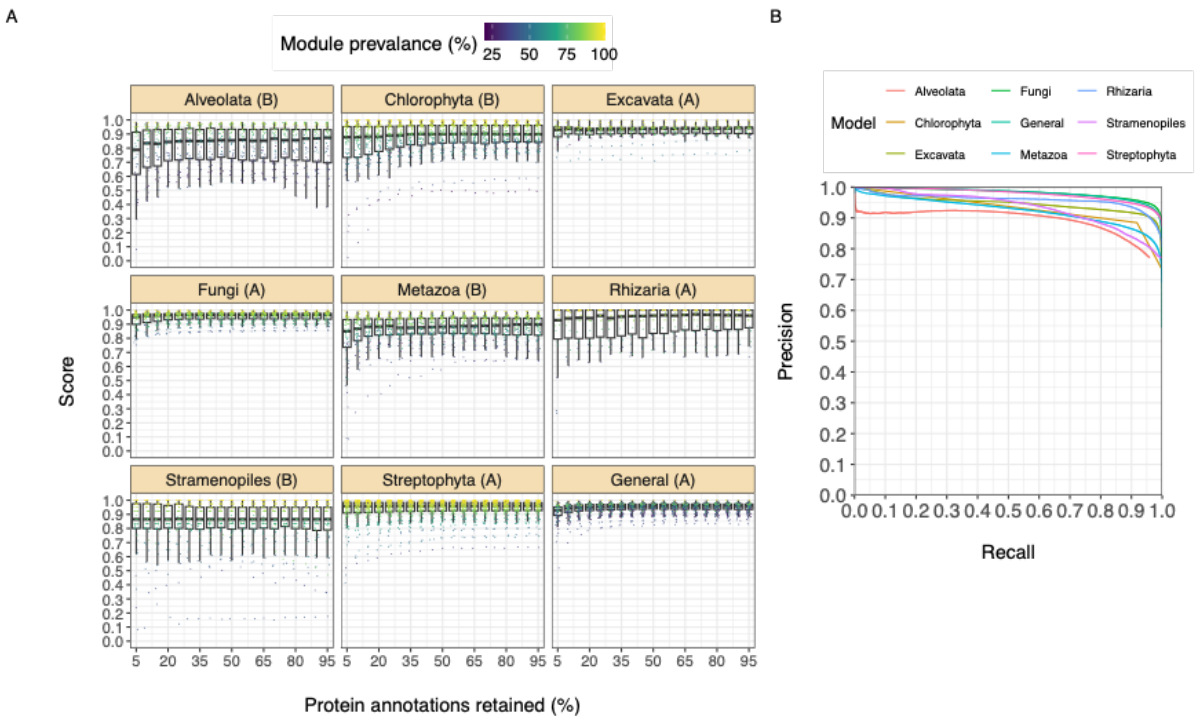


Figure 5-3. Panel A: Precision-recall (PR) curves for MetaPathPredict-E's models (n = 9). Each PR curve was calculated for predictions made on all test data (including all down-sampled test set genomes and transcriptomes) for each model. Panel B: Boxplots of the distribution of the macro F1 score for all labels in each of the 19 down-sampled test datasets, faceted by model (facet labels correspond to specific models). The test datasets each contain different versions of gene annotations. In the 95% dataset, 5% of protein families present in the samples were randomly removed; this process was repeated in 5% increments, down to 5% on the far-left side of the plots. Each point drawn over the boxplots is the F1 score for a particular label from the multi-label prediction outputs. The module prevalence coloring of the labels corresponds to the percentage of test dataset samples containing each label predicted by the models. "Group A" models (described in the text) correspond to the Excavata, Fungi, Rhizaria, Streptophyta, and General models. The "Group B" models are the Alveolata, Chlorophyta, Metazoa, and Stramenopiles models.

Benchmarking MetaPathPredict-E on held-out, down-sampled NCBI, JGI GOLD, MMETSP, and NCBI TSA data

The performance of MetaPathPredict-E's deep learning models were first evaluated on held-out test datasets comprised of genomes from NCBI RefSeq and JGI GOLD, as well as transcriptomes from the NCBI TSA and MMETSP databases. All proteomes had a BUSCO completeness score of at least 80%. A set of 19 test datasets was created by randomly downsampling the data to retain 5% to 95% of gene annotations (in 5% increments). In all 19 datasets, each proteome was randomly downsampled 10 times using a distinct random number generation seed for each iteration.

MetaPathPredict-E's deep learning strategy exhibited a strong performance on this held-out test data. For the Streptophyta, Fungi, Excavata, Rhizaria and general model (group A models), the macro F1 score ranged from 0.90-0.95 (average 0.94; Figure 3A) when predicting on test datasets

in which 30% or more of the original gene annotations had been retained (Figure 3B). For the Alveolata, Chlorophyta, Metazoa, and Stramenopiles models (group B models) the macro F1 score ranged from 0.81-0.89 (average 0.85; Figure 3B). The average of all the macro recall scores for the 19 test datasets was 0.93 and the lowest individual macro recall score was 0.82 (Metazoa model, 5% of original gene annotations retained; Figure 3A). For the group A models, MetaPathPredict-E rarely made false negative predictions based on data from highly incomplete gene annotation sets. MetaPathPredict-E also did not misclassify most positive class (module "present") proteome labels; the average macro precision of MetaPathPredict-E's models was 0.87 for all test datasets. Macro precision was lowest for the group B models, ranging from 0.73-0.84 when 5% of the original gene annotations were retained. MetaPathPredict-E's capacity to attain high macro recall score on datasets with substantially reduced gene annotations indicates that it offsets the limited data by more frequently labeling a module as 'present,' resulting in decreased precision. The reduced performance of the group B models relative to the group A models is most likely attributed to the relative inadequacy of available training datasets in covering the taxonomic diversity and metabolic variation for these eukaryotic groups (Supplementary Figure 2; Supplementary Table 1). A low number of training samples can lead to a reduced capability for a model to learn meaningful patterns from the training data for that group. The Excavata and Rhizaria models also outperformed the group B models despite also having small training datasets. This could be due to the fact that the training and test datasets covered only a small fraction of the taxonomic diversity of these two large groups (Supplementary Figure 2) as well as a low number of training data features and labels (Supplementary Table 1), and a relatively lower amount of variance in module presence/absence which yielded simpler, less comprehensive training datasets. Taken together, the Alveolata, Stramenopiles, Metazoa, and Chlorophyta models (as well as the Rhizaria and Excavata models) would benefit from larger training datasets as more genome and transcriptome data for these groups becomes available.

Benchmarking MetaPathPredict-E against built environment MAGs

MetaPathPredict-E was further tested on gene annotations from a set of 8 eukaryotic MAGs recovered from built environment metagenomes (Olm et al., 2019; Figure 4). This dataset contained 8 high-completeness (≥ 80% BUSCO completeness) fungal MAGs affiliated with the Ascomycota and Basidiomycota (Figure 4B) that were recovered from infant gut microbiomes and

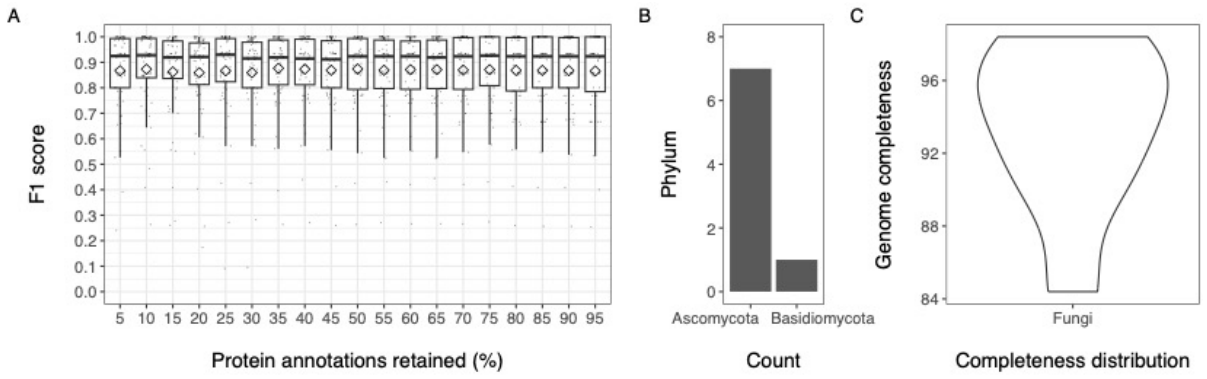


Figure 5-4. Panel A: Boxplots displaying the macro F1 score distribution of MetaPathPredict-E's label predictions for fungal MAGs. Down-sampled gene annotations of high completeness MAGs (≥ 80% BUSCO completeness) used in this held-out test set are from built environment metagenomes. Each boxplot displays predictions of randomly down-sampled versions of the gene annotation test set in increments of 5% (95% down to 5%; from right to left). Panel B: Distribution of phyla of the MAGs utilized in this test set. Panel C: Violin plot of the BUSCO completeness distribution of the 8 fungal MAGs.

neonatal intensive care hospital rooms. This analysis facilitated the benchmarking of MetaPathPredict-E on MAGs. A set of 19 test datasets was created by randomly downsampling the gene annotations of these MAGs to retain 5% to 95% of annotations (in 5% increments) as was done for the genomic test set. MetaPathPredict-E classified the presence/absence of KEGG modules in each MAG. Overall, results were comparable to the performance on the NCBI and JGI GOLD genomes. This could be partially due to the near-complete nature of this set of MAGs. The models accurately predicted the presence or absence of KEGG modules in all 19 test datasets regardless of the percentage of gene annotations randomly retained. The average macro F1 score across all the datasets for KEGG module predictions was 0.87 (Figure 4A). Predictions were still reliable when 20% or less gene annotation data was retained, with a mean macro F1 score of 0.86. The robust performance of MetaPathPredict-E on built environment fungal MAGs even at low gene annotation retention levels was at the cost of decreased precision, however. The average macro precision was 0.83, while the mean macro recall was 0.95. This indicates MetaPathPredict-E could consistently detect positive class labels (KEGG module presence) at the cost of a slightly increased rate of misclassifying some negative labels as positive.

Benchmarking MetaPathPredict-E against incomplete genomes and transcriptomes

In addition to model assessments made through down-sampling protein annotations of high-completeness proteomes, we evaluated the set of all genomes and transcriptomes downloaded from NCBI, JGI GOLD, MMETSP, and NCBI TSA that had a BUSCO completeness score lower than

A

Model	Detection rate	Number of genomes
General	94.6	1,933
Fungi	91.9	145
Metazoa	87.4	87
Streptophyta	94.4	972
Stramenopiles	91.5	178
Chlorophyta	94.1	78
Rhizaria	97.1	22
Excavata	86.1	22

B

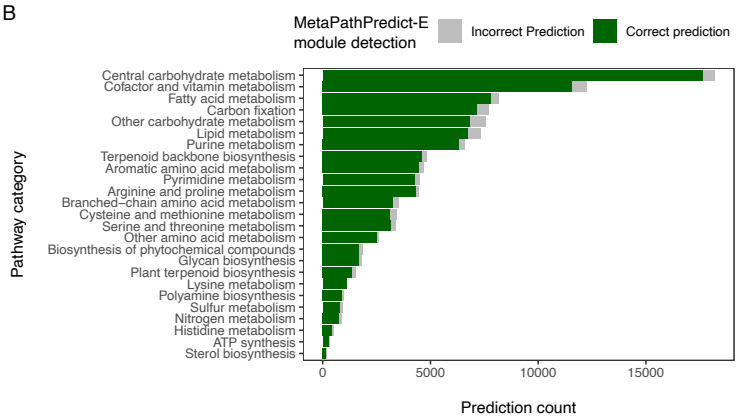


Figure 5-5. Panel A: Detection rates of complete KEGG modules in test set genomes and transcriptomes that had less than 80% BUSCO completeness for 8 of MetaPathPredict-E's models (Alveolata not shown; no low completeness proteomes). Panel B: Stacked bar charts showing MetaPathPredict-E's detection rates of KEGG module categories contained within the test set genomes. Dark green coloring indicates MetaPathPredict-E detected the module; the grey coloring is for complete modules that were not detected.

80%. The majority of low completeness proteomes were affiliated with the Streptophyta, Fungi and Stramenopiles (see "Number of genomes" in Figure 5A and Supplementary Figure 3). MetaPathPredict-E's performance on this test set resembled the previous results on simulated and unmodified proteome annotation data (Figure 5A,B). The mean detection rate of positive class labels was 92.1% across all test dataset proteomes. However, the median BUSCO completeness score for genomes and transcriptomes in this test dataset was 61%, and they were 52% complete on average. Despite the samples in the test skewing towards a moderate level of completeness, the high detection rate of modules suggested a robust performance similar to the previous benchmarking tests. All of the models predicted the presence of central carbon metabolisms, purine and fatty acid metabolisms, as well as cofactor and vitamin metabolisms (Figure 5B) in instances where KEGG modules were missing or incomplete in the data. Predictions for carbon fixation modules were prevalent in Streptophyta, Rhizaria, Stramenopiles, and Alveolata proteomes in which these modules were only partially recovered or missing. Overall, MetaPathPredict-E achieved a high detection rate of module labels in all test datasets on proteomes that ranged from a low to moderately high degree of completeness and made sensible predictions for modules that were incompletely recovered or missing in the proteome data.

The Fungi, Streptophyta, and general model training datasets contained significantly larger numbers of taxonomically diverse genome and transcriptome representatives compared to the other modelled groups. The volume of training examples for these group models likely contributed to their more robust macro precision and recall on test datasets. MetaPathPredict-E's Excavata and

Rhizaria models also exhibited robust performance metrics, likely due to the relatively low taxonomic and metabolic diversity present in these training datasets.

Conclusion

Building upon the open-source tool MetaPathPredict, MetaPathPredict-E expands the scope of that software to predict the functional potential of eukaryotic genomes and transcriptomes. MetaPathPredict-E both identifies complete KEGG modules and predicts the presence or absence of fragmented or missing ones. Leveraging annotations of individual protein sets in the format of KO gene identifiers, MetaPathPredict-E integrates with existing KEGG-based annotation tools like KofamScan (Aramaki et al., 2020), DRAM (Shaffer et al., 2020), blastKOALA (Kanehisa et al., 2016), and ghostKOALA (Kanehisa et al., 2016). It is possible to also utilize a custom set of KO gene annotations in one or more TSV files as input. MetaPathPredict-E also suggests putative KO gene annotations that could potentially fill in the missing steps of KEGG modules that the tool predicts are present.

MetaPathPredict-E further validates the utilization of protein family presence or absence as a feature for the prediction of metabolic potential in eukaryotes. Performance metrics assessed on genome and transcriptome datasets from the NCBI RefSeq, JGI GOLD, MMETSP, and NCBI TSA databases as well as MAGs from the built environment demonstrate the effectiveness of deep learning for the accurate prediction of KEGG module presence/absence across a spectrum of eukaryotic gene annotation completeness, ranging from sparse to nearly-complete proteomes. Our performance tests on MetaPathPredict-E revealed an interesting dynamic between data completeness and prediction accuracy. While macro recall remained remarkably robust even when downsampling proteins to retain just 5-10% of annotations, a decrease in macro precision was observed. This trend was characterized by an increase in false positive predictions (i.e., modules incorrectly classified as present) in all our tests and suggests that the model compensates for limited sequence data by adopting a more aggressive "presence" calling strategy. This overconfident positive class prediction issue emerged only when 25% or less gene annotation data was retained from the original gene content, indicating a potential area for future model improvement. While severely incomplete inputs (< 30% completeness) may not be optimal for use with MetaPathPredict-E, this finding highlights an avenue for enhancing the model's macro precision at lower data thresholds. Based on these observations, we recommend using

MetaPathPredict-E on datasets containing at least 30% of recovered gene annotation information for optimal performance.

MetaPathPredict-E facilitates more comprehensive reconstruction of the metabolic potential encoded within fragmented eukaryotic genomes and transcriptomes from environmental samples. The models for group B (Alveolata, Chlorophyta, Metazoa, and Stramenopiles) had a macro F1 score of at least 0.81 and less than 0.90 on test datasets containing 30% or more gene annotations. This performance was less robust compared to the group A models (general model, Fungi, Streptophyta, Rhizaria, Excavata), for which the macro F1 score was at least 0.90 for datasets retaining 30% or more of their original gene annotations. This was likely caused by small training datasets available for those groups. The tool's machine learning model will enhance the ability to infer what metabolic processes eukaryotes are capable of, even when analyzing proteomes that are highly incomplete. MetaPathPredict-E integrates the field of machine learning with the expanding use of 'omics sequencing techniques and will help facilitate the metabolic analysis of environmental eukaryotic proteome datasets. Its precision and accuracy will continue to improve as genomes and transcriptomes become available for eukaryotic groups for which little data exists at present.

Data Availability

Genomic and transcriptomic data used for creation of MetaPathPredict-E's models is available from the NCBI RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/), the MMETSP improved assemblies database (https://zenodo.org/records/3247846), the NCBI TSA database (https://www.ncbi.nlm.nih.gov/genbank/tsa/), and the JGI GOLD database (https://gold.jgi.doe.gov/). Fungal MAG data used for benching marking is available on GitHub (https://github.com/MrOlm/InfantEukaryotes).

Code Availability

The MetaPathPredict-E Python module is available from the following GitHub repository: https://github.com/d-mcgrath/MetaPathPredict-E.

References

Adl, Sina M., et al., "The revised classification of eukaryotes." Journal of eukaryotic microbiology 59.5 (2012): 429-514.

Adl, Sina M., et al.,"Revisions to the classification, nomenclature, and diversity of eukaryotes." Journal of Eukaryotic Microbiology 66.1 (2019): 4-119.

Alexander, Harriet, et al., "Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton." mBio (2023): e01676-23.

Altschul, Stephen F., et al., "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.

Aramaki, Takuya, et al., "KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold." *Bioinformatics* 36.7 (2020): 2251-2252.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. "Fast and sensitive protein alignment using DIAMOND." *Nature methods* 12.1 (2015): 59-60.

Clark, Karen, et al., "GenBank." Nucleic acids research 44.D1 (2016): D67-D72.

Delmont, Tom O., et al., "Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean." Cell Genomics 2.5 (2022).

Drake, Henrik, and Magnus Ivarsson. "The role of anaerobic fungi in fundamental biogeochemical cycles in the deep biosphere." Fungal Biology Reviews 32.1 (2018): 20-25.

Finlay, Roger D. "Mycorrhizal fungi and their multifunctional roles." Mycologist 18.2 (2004): 91-96.

Gabriel, Lars, et al., "BRAKER3: Fully automated genome annotation using RNA-Seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA." Biorxiv (2023).

Gallego Romero, Irene, et al.,"RNA-seq: impact of RNA degradation on transcript quantification." BMC biology 12 (2014): 1-13.

Gast, Rebecca J., et al., "Kleptoplasty in an Antarctic dinoflagellate: caught in evolutionary transition?." Environmental Microbiology 9.1 (2007): 39-45.

Gawryluk, Ryan MR, and Courtney W. Stairs. "Diversity of electron transport chains in anaerobic protists." Biochimica et Biophysica Acta (BBA)-Bioenergetics 1862.1 (2021): 148334.

Geller-McGrath, D., Konwar, K. M., Edgcomb, V. P., Pachiadaki, M., Roddy, J. W., Wheeler, T. J., & McDermott, J. E. (2024). Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict. Elife, 13, e85749.

Ginger, Michael L., et al., "Intermediary metabolism in protists: a sequence-based view of facultative anaerobic metabolism in evolutionarily diverse eukaryotes." Protist 161.5 (2010): 642-671.

Haas, Brian J., et al., "De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis." *Nature protocols* 8.8 (2013): 1494-1512.

Haas, Brian J. https://github.com/TransDecoder/TransDecoder.

Huerta-Cepas, Jaime, et al., "Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper." *Molecular biology and evolution* 34.8 (2017): 2115-2122.

Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. pmlr, 2015.

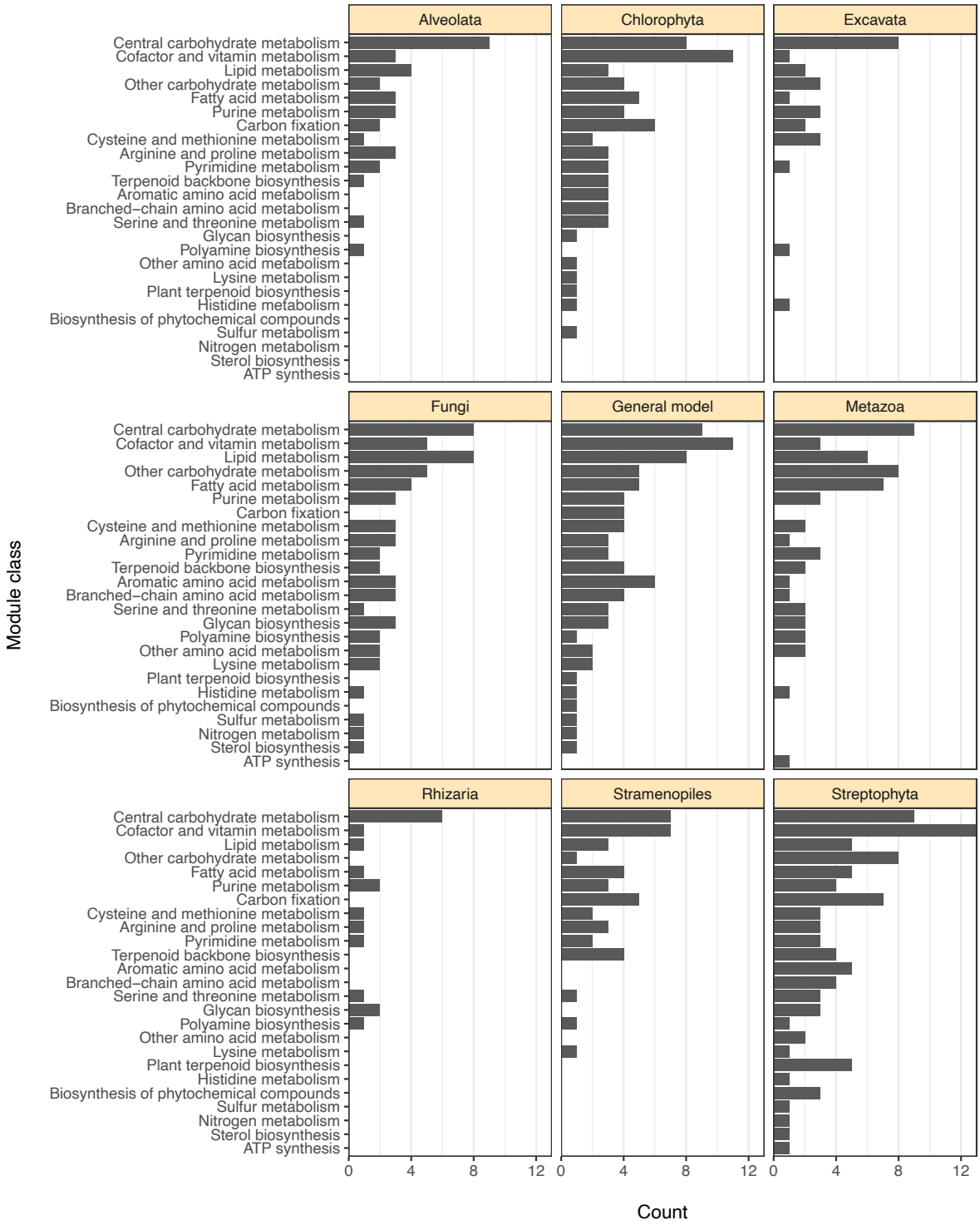
Johnson, Lisa K., Harriet Alexander, and C. Titus Brown. "Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes." Gigascience 8.4 (2019): giy158.
Jones, Philip, et al., "InterProScan 5: genome-scale protein function classification." *Bioinformatics* 30.9 (2014): 1236-1240.
Kanehisa, Minoru. "The KEGG database." *'In silico' simulation of biological processes: Novartis Foundation Symposium 247*. Vol. 247. Chichester, UK: John Wiley & Sons, Ltd, 2002.
Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. "BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences." Journal of molecular biology 428.4 (2016): 726-731.
Karp, Peter D., et al., "The metacyc database." *Nucleic acids research* 30.1 (2002): 59-61.
Kazda, Marian, Susanne Langer, and Frank R. Bengelsdorf. "Fungi open new possibilities for anaerobic fermentation of organic residues." Energy, Sustainability and Society 4 (2014): 1-9.
Keeling, Patrick J., et al., "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing." PLoS biology 12.6 (2014): e1001889.
Lambert, Bennett S., et al.,"The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics." Proceedings of the National Academy of Sciences 119.7 (2022): e2100916119.
Leegood, Richard C., Thomas D. Sharkey, and Susanne Von Caemmerer, eds. Photosynthesis: physiology and metabolism. Vol. 9. Springer Science & Business Media, 2006.
Lukashin, Alexander V., and Mark Borodovsky. "GeneMark. hmm: new solutions for gene finding." *Nucleic acids research* 26.4 (1998): 1107-1115.
Muller, M., G. H. Coombs, and M. J. North. "Energy metabolism of anaerobic parasitic protists." biochemical Protozoology (1991): 80-91.
O'Leary, Nuala A., et al., "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation." Nucleic acids research 44.D1 (2016): D733-D745.
Olm, Matthew R., et al., "Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms." Microbiome 7 (2019): 1-16.
O'Malley, Tom, et al., KerasTuner. 2019, https://github.com/keras-team/keras-tuner.
Rashid, Muhammad Imtiaz, et al., "Bacteria and fungi can contribute to nutrients bioavailability and aggregate formation in degraded soils." Microbiological research 183 (2016): 26-41.
Sayers, Eric W., et al., "Database resources of the national center for biotechnology information." Nucleic acids research 49.D1 (2021): D10.
Shaffer, Michael, et al., "DRAM for distilling microbial metabolism to automate the curation of microbiome function." Nucleic acids research 48.16 (2020): 8883-8900.
Simão, Felipe A., et al., "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." Bioinformatics 31.19 (2015): 3210-3212.
Srivastava, Nitish, et al., "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.
Stanke, Mario, et al., "AUGUSTUS: ab initio prediction of alternative transcripts." *Nucleic acids research* 34.suppl_2 (2006): W435-W439.

Stoecker, Diane K., et al., "Acquired phototrophy in aquatic protists." Aquatic Microbial Ecology 57.3 (2009): 279-310.
Weber, Andreas PM, et al., "Metabolism and metabolomics of eukaryotes living under extreme conditions." International review of cytology 256 (2007): 1-34.
Wheeler, David L., et al., "Database resources of the national center for biotechnology information." Nucleic acids research 35.suppl_1 (2007): D5-D12.
Harwood, John L., and Irina A. Guschina. "The versatility of algae and their lipid metabolism." Biochimie 91.6 (2009): 679-684.
Edwards, Kyle F., et al., "Trophic strategies explain the ocean niches of small eukaryotic phytoplankton." Proceedings of the Royal Society B 290.1991 (2023): 20222021.
Müller, Miklós, et al., "Biochemistry and evolution of anaerobic energy metabolism in eukaryotes." Microbiology and Molecular Biology Reviews 76.2 (2012): 444-495.
Těšitel, Jakub, et al., "Mixotrophy in land plants: why to stay green?." Trends in Plant Science 23.8 (2018): 656-659.
Tørresen, Ole K., et al.,"Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases." Nucleic acids research 47.21 (2019): 10994-11006.
Fenchel, Tom, and Bland J. Finlay. "The biology of free-living anaerobic ciliates." European journal of protistology 26.3-4 (1991): 201-215.
Fenchel, Tom. "Anaerobic eukaryotes." Anoxia: Evidence for eukaryote survival and paleontological strategies (2011): 3-16.Tang, Shiyuyun, Alexandre Lomsadze, and Mark Borodovsky. "Identification of protein coding regions in RNA transcripts." Nucleic acids research 43.12 (2015): e78-e78.
Geller-McGrath, David, et al., "Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict." Elife 13 (2024): e85749.
Pedregosa, Fabian, et al., "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.
Chollet, Francois. Deep learning with Python. Simon and Schuster,
Eren, A. Murat, et al., "Anvi'o: an advanced analysis and visualization platform for 'omics data." PeerJ 3 (2015): e1319.
Harley, J. L. "Fungi in ecosystems." Journal of Ecology 59.3 (1971): 653-668.
Blin, K. (2023). ncbi-genome-download (0.3.3). Zenodo. https://doi.org/10.5281/zenodo.8192486
Alexander, Harriet et al., "Eukaryotic TOPAZ MAGs." OSF, 21 Sept. 2023. Web.

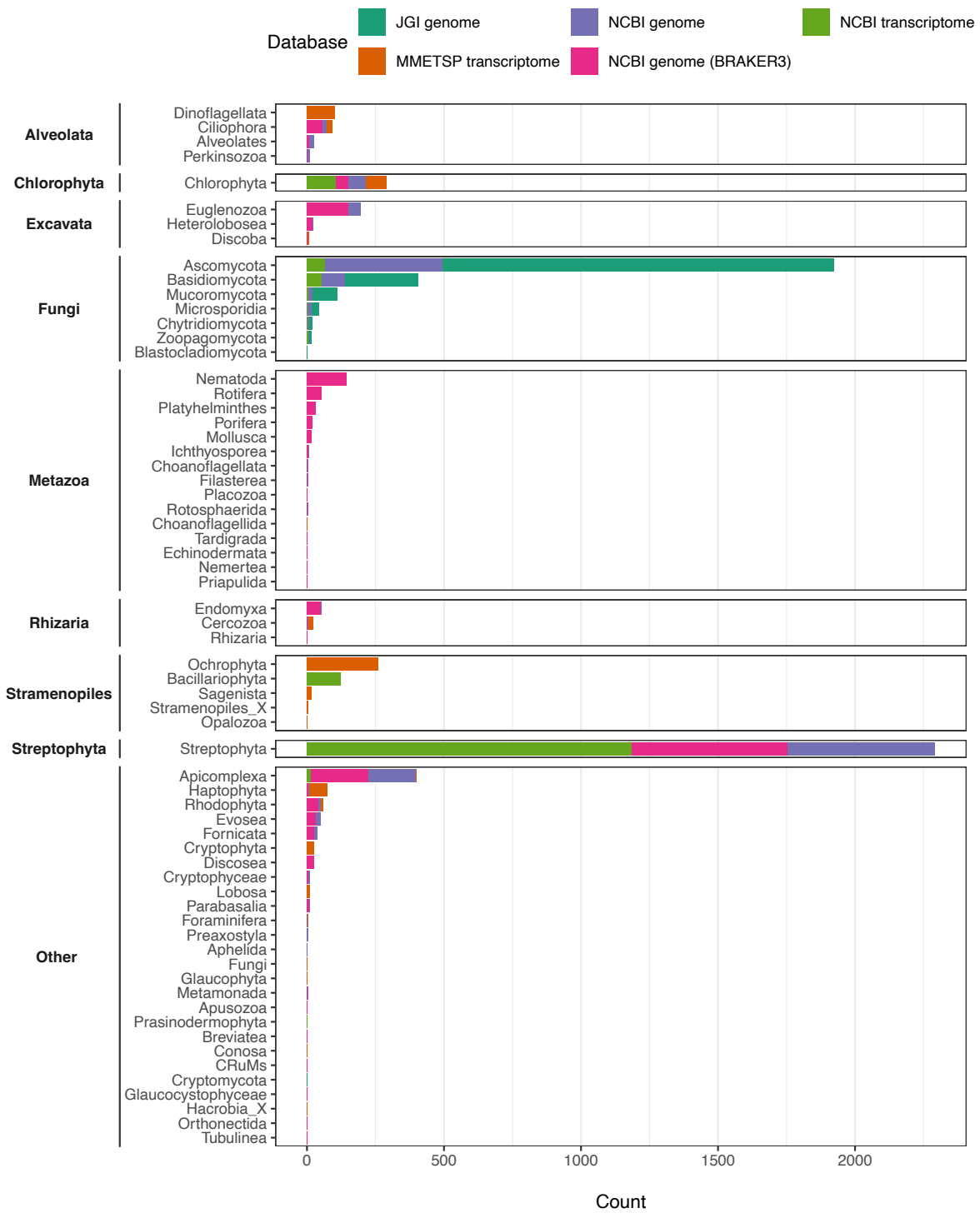
Acknowledgements

D. Geller-McGrath acknowledges funding from the Academic Programs Office at the Woods Hole Oceanographic Institution (WHOI). We would like to thank Andy Solow (WHOI) for helpful discussions about statistical methods during the development of this project.

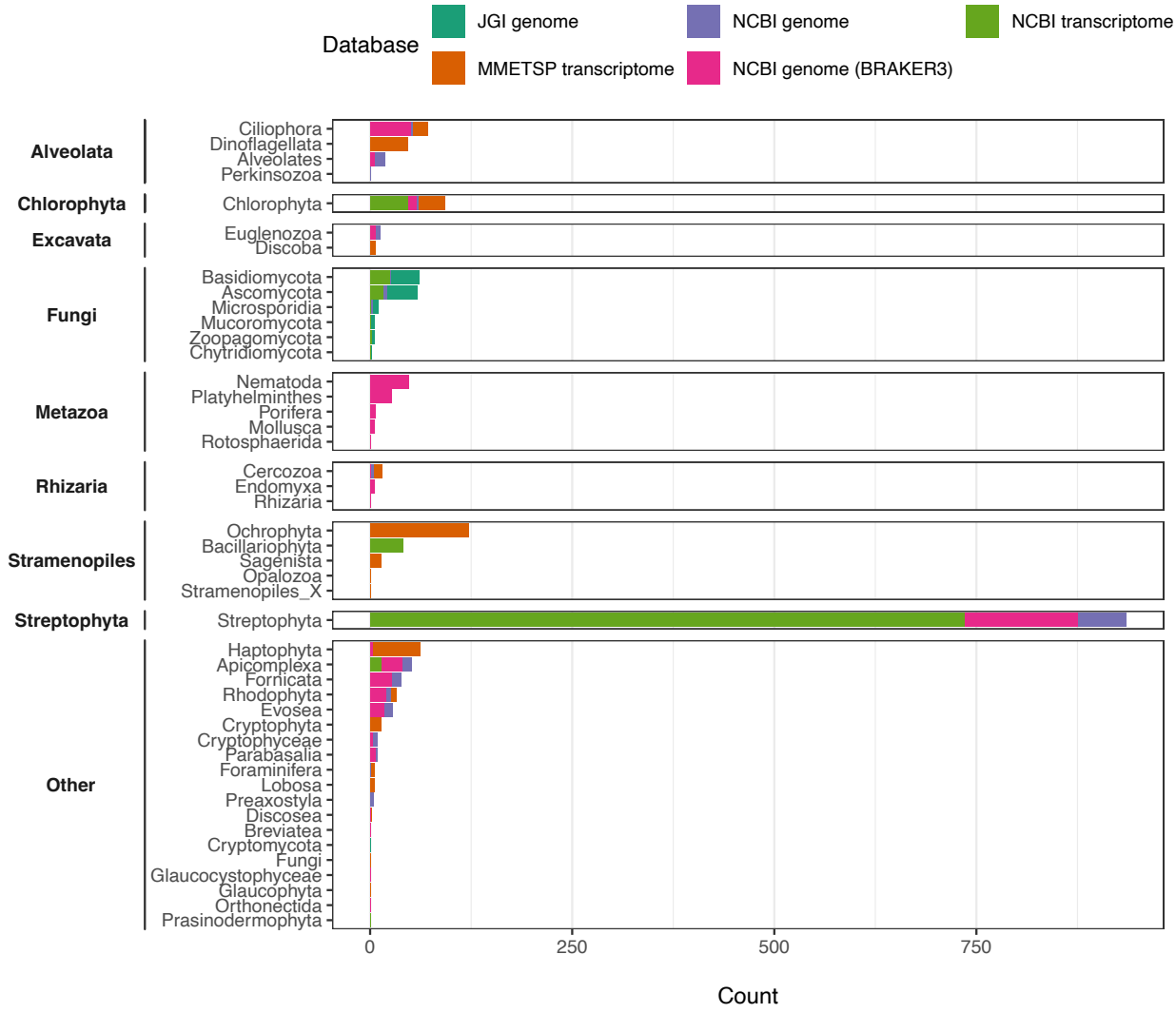
Supplementary Figures



Supplementary Figure 5-1. Bar charts of the distribution of KEGG module classes for all nine of MetaPathPredict-E's models. Module classes are labelled along the *y*-axis, and the count for each module class is labelled along the *x*-axis. Each bar chart corresponds to the specific model that is labelled above it. Model bar charts are displayed in the following order from top to bottom, left to right: Alveolata, Chlorophyta, Excavata, Fungi, General model, Metazoa, Rhizaria, Stramenopiles, and Streptophyta. Modules had to be present in 25% of the training samples for a module to be included in a model's set of labels (KEGG module presence/absence).



Supplementary Figure 5-2. Bar chart showing the distribution of phyla of all genomes and transcriptomes with at least 80% BUSCO completeness downloaded from NCBI, JGI, and MMETSP databases. Eukaryotic groups are displayed in bold text to the far left. Bars are colored to show the distribution of databases the proteomes were downloaded from. One model was trained for each group, and an additional general model was trained to classify KEGG modules prevalent across all groups including the "Other" category.



Supplementary Figure 5-3. Bar chart showing the distribution of phyla of all genomes and transcriptomes with less than 80% BUSCO completeness downloaded from NCBI, JGI, and MMETSP databases. Eukaryotic groups are displayed in bold text to the far left. Bars are colored to show the distribution of databases the proteomes were downloaded from. KEGG module predictions were for these proteomes by all of MetaPathPredict-E's models.

Supplementary Tables

Eukaryotic model	Number of training samples	Number of features	Number of labels
General (all training data)	5,184	10,670	87
Fungi	2,374	8,499	63
Streptophyta	1,357	9,331	96
Stramenopiles	231	5,208	44
Excavata	208	3,895	26
Metazoa	206	8,307	56
Chlorophyta	196	5,700	67
Alveolata	90	5,158	35
Rhizaria	54	3,844	18

Supplementary Table 5-1. A table containing the number of training genomes/transcriptomes (column 2) for each model (column 1), in addition to the number of features used for training (column 3) and the number of labels each model was trained to predict (column 4). The general model contains 468 additional genomes and/or transcriptomes in its training dataset that do not fall into one of the groups listed here. See Supplementary Figure 2 for the full taxonomic distribution of all the data used for training. Features correspond to the presence or absence of protein families (KEGG orthologs), one per feature; the labels are the presence or absence of KEGG modules.

Chapter 6

Conclusion

In this thesis, I aimed to pair two studies that focus on aspects of microbial metabolism in marine environments to the development of two bioinformatic tools to facilitate this type of analysis. Chapter 2 presented a genome mining study of biosynthetic gene clusters (BGCs) encoding secondary metabolites and an evaluation of their expression profiles along an oxycline water column. Chapter 4 investigated the primary and secondary metabolism of prokaryotic microbial communities inhabiting hydrothermally-influenced deep biosphere sediments. The computational analyses in these chapters involved the use of existing bioinformatics tools to examine metagenomic and metatranscriptomic datasets. For both of Chapters 2 and 4, I recovered metagenome-assembled genomes (MAGs) from metagenomic co-assemblies and analyzed their gene content. I quantified genome abundance and gene expression profiles in samples from distinct habitats by mapping reads from metagenomic and metatranscriptomic datasets to the MAGs. Chapter 3 introduced a novel method for predicting primary metabolism in bacteria when in the case of metagenome analysis of environmental samples, there is incomplete genomic data for many/most taxa in the sample. Chapter 5 built on Chapter 3 by developing an approach to predict primary metabolism for eukaryotes. For theses tool developments, I trained deep learning models to predict the presence or absence of metabolic modules (as defined in the KEGG database) in bacterial genomes (Chapter 3), and in eukaryotic genomes and transcriptomes (Chapter 5). The breadth of these chapters is entirely intentional; as "metabolism" encompasses such a broad network of biochemical reactions and pathways, necessitating a multifaceted approach for prokaryotic and eukaryotic datasets.

There are two major observations that emerged from this thesis. These are highly relevant to my personal research interests and hold importance for the fields of marine biology, bioinformatics, and computational biology. The first is the need for integrated data analyses to develop a more comprehensive understanding of metabolic processes. Much of the utility of the work presented in this thesis is derived from the integration of metagenomic and

metatranscriptomic data types as well as associated geochemical data, which together provided greater insight and confirmation of the findings than they would have separately. Chapters 2 and 4 integrated metagenomic and metatranscriptomic datasets to facilitate the analysis of genes involved in primary and secondary metabolisms as well as patterns in their expression. The examination of these two data types together facilitated more informative analyses of how microorganisms interact with their environments. Chapter 5 utilized genomic and transcriptomic datasets from eukaryotic organisms that together provided a larger breadth of taxonomic coverage in the training data than either data type would have by itself. The second major observation from this thesis is the value of machine learning methods for enhancing the analysis of microbial communities. The creation of new metabolic prediction tools in Chapters 3 and 5 highlighted two of the many avenues for the use of machine learning algorithms that can be trained on the plethora of biological data that has become available since the advent of next-generation sequencing. As research projects worldwide continue to generate enormous streams of data, it will be imperative to take advantage of machine learning algorithms to help inform future studies.

Several research questions emerged as I explored the various aspects of my thesis. In my second chapter, I aimed to assess the potential for organisms whose MAGs were recovered from the Cariaco Basin oxycline water column to produce secondary metabolites. One of my main questions was whether there were differences in biosynthetic gene cluster expression between particle-associated and free-living microorganisms inhabiting this environment. This investigation inspired several additional scientific questions that propelled the rest of the research in this thesis:

1. How do environmental parameters influence the microbial diversity and the metabolism of microorganisms inhabiting subsurface sedimentary habitats, specifically, the hydrothermal deep biosphere of Guaymas Basin? This deep biosphere environment consists of largely anoxic sediments and samples were obtained from sites with distinct environmental features. The sediments of Guaymas Basin also contain abundant hydrocarbons in addition to vast reservoirs of organic and inorganic compounds. Investigation of deep drill core samples contributes to the understanding of how microorganisms survive in in this deep biosphere.

2. What are the current limitations of bioinformatics tools for assessing the metabolic potential of environmental bacterial genomes, and how are they limited by the sequencing coverage of a

sample? If machine learning methods could help elucidate the metabolic potential encoded in bacterial genomes for incomplete genomic data, this could facilitate a more informed and robust understanding of microbial metabolism and functionality in future environmental studies.

3. Could the same or a similar machine learning approach be applied to eukaryotic genomes and transcriptomes? Exploring this question could open new avenues for the analysis of eukaryotic genomes, which are more complex and often even less complete when recovered from environmental samples than those of prokaryotes. Additionally, most available tools have been optimized for prokaryotic data. Machine learning models could also facilitate the analysis of incomplete eukaryotic data, helping to make inferences about their metabolic capabilities.

Chapter 2 leveraged metagenomes and metatranscriptomes recovered from Cariaco Basin over the course of two seasons. Cariaco Basin is a permanently stratified water column situated off the coast of Venezuela that receives a restricted flow of oxygenated water from the Caribbean Sea due to the presence of a sill that is only ~150 meters below the sea surface. To date, microbiological studies of the Cariaco Basin redoxcline have included the characterization of bacterial and archaeal communities based on marker genes, metagenomes, metatranscriptomes, qPCR, incubation studies, and microscopy, as well as studies of viral elements and protists (Mara et al., 2020; Edgcomb et al., 2011; Taylor et al., 2001; Madrid et al., 2001; Taylor et al., 2018; Suter et al., 2021). As done in this chapter of the thesis, a previous study compared data from particle-associated (PA) and free-living (FL) microbial communities inhabiting the redoxcline (Suter et al., 2017). In this chapter, I examined biosynthetic gene clusters (BGCs) encoded in prokaryotic MAGs recovered from FL and PA metagenomes from the Cariaco Basin redoxcline. Evidence for the expression of biosynthetic transcripts was also investigated by mapping metatranscriptome reads to biosynthetic genes. Samples were collected along the Cariaco Basin's redoxcline that included oxic, suboxic, dysoxic, anoxic, and euxinic depths. The samples were filtered using filters with two different pore sizes: a larger filter (>2.7 μm) to capture the PA microbial communities, and a smaller filter (0.2-2.7 μm) for the FL fraction.

Mining of biosynthetic gene clusters (BGCs) from environmental genomes has become a popular bioinformatics approach for analyzing the potential interactions that can occur within microbial communities. A recent study analyzed ~40,000 BGCs from over 1,000 marine

metagenomes collected globally from the world's oceans (Paoli et al., 2022). However, this massive genome mining effort did not cover sulfidic end-members, leaving the potential for microorganisms to produce secondary metabolites in these habitats virtually unexplored. I mined BGCs from MAGs recovered from a co-assembly of PA and FL metagenomes sampled from the Cariaco Basin redoxcline. I then examined the potential for microbial communities in each fraction to synthesize secondary metabolites. The abundance patterns of BGC-containing MAGs and BGC expression patterns were compared between water column fractions. There was enhanced expression of BGCs by Myxococcota, Desulfobacterota, Omnitrophota, Planctomycetota, and Gammaproteobacteria in the PA fraction, while in the FL fraction BGC expression was predominantly from Alphaproteobacteria, Omnitrophota, SAR324, and Desulfobacterota. The higher expression of BGCs in PA samples suggests a more common release of secondary metabolites within particles than in FL microbial populations, though the exact benefits to FL microbes remain unclear and may involve intracellular roles.

This study could have been enhanced by also including a comparison of biosynthetic potential of microorganisms occupying other contrasting habitats to facilitate direct comparisons of BGC content and expression profiles found in Cariaco Basin to those reported globally in marine (and other) environments. Additional recovered MAGs could have been analyzed for BGCs they encoded and expressed; the MAGs analyzed in this study had an estimated CheckM completeness of at least 75%, with an estimated contamination of 5% or less. Lower completeness MAGs with 5% contamination of less could have been included, with the caveat that inclusion of less complete MAGs confounds interpretation due to missing information. I also could have annotated BGCs from the entire metagenomic co-assembly using tools such as BiosyntheticSPAdes (Meleshko et al., 2019) which annotates BGCs from assembly graphs. This could have helped recover a significantly larger proportion of biosynthetic potential from the Cariaco redoxcline metagenomes. The metagenomic and metatranscriptomic reads could have been mapped to the BGCs detected in the whole co-assembly for a more comprehensive analysis of BGC abundance and expression profiles. Additional taxonomic and functional information would be available within contigs annotated to BGCs from the co-assembly that were not contained within the MAGs. Contigs not included in MAGs can be taxonomically classified using tools such as CAT (Bastiaan von Meijenfeldt et al., 2019) to infer which lineages contained them. In addition, BiosyntheticSPAdes can identify BGCs spanning multiple contigs. One of the limitations of using version 6.0 of

antiSMASH is that detection of BGCs was limited to those encoded within a single contig of a genomic assembly. This is a particular limitation for MAGs containing very short contigs. In addition, the number and type of BGCs that antiSMASH can detect increases with each new version of the tool. AntiSMASH version 7.0 was published in July of 2023 and supports the prediction of 81 cluster types, up from 71 supported by version 6.0 (Blin et al., 2023). The analyses performed in this chapter can therefore be re-run as newer and more robust BGC detection methods are made available.

The conclusions reached in Chapter 2 helped guide the research question addressed in Chapter 3. In my second chapter, I discarded all genomes from downstream analysis that had an estimated completeness less than 75%. This "quality control" measure removed a significant number of moderately to highly incomplete MAGs from this study. It is common practice in metagenomics studies to discard medium- to low-quality MAGs (with 50-75% completeness or less), due to difficulties in discerning the metabolic potential of these genomes when much of the gene content is missing. In addition, bioinformatics methods for discerning metabolism have suboptimal performance on highly incomplete genomes. I addressed these deficiencies in Chapter 3 with the development of a new tool, MetaPathPredict, that makes accurate predictions for highly incomplete bacterial genomes. I utilized the massive amount of sequencing data available in public databases to train deep learning models that learned to predict the metabolic capacity present within incomplete bacterial genomes that contained as little as 30% of their total gene content. Importantly, MetaPathPredict consistently made predictions with a high degree of macro precision and recall, indicating its ability to accurately predict an array of various metabolic modules. This method is targeted toward environmental metagenomic or single-cell datasets and can be run on highly incomplete genomes that would often be discarded prior to further analyses. MetaPathPredict could continue to expand its prediction capacity to more modules which will improve its predictions as more sequencing data is made available for use as training data.

MetaPathPredict could be expanded to make predictions for archaea in addition to bacteria. The data used to train and validate its models could also be improved upon. Currently, the training data has been annotated with KofamScan (Aramaki et al., 2020). To make initial improvements to the protein annotations used for training and validation, more than one annotation tool could be used to annotate the training data. This could expand the number of genes in training genomes that get functionally characterized. After improvements to the training data through the incorporation

of multiple annotation methods, the training genomes could be further assessed for the presence of complete KEGG modules using gapfilling methods. All genomes used for training and testing had a CheckM completeness score of 100, and a contamination estimation of 0. These metrics are estimates at best, and there is a possibility that some coding potential is missing from their gene annotations. Tools such as KEMET (Palù et al., 2022) could be used on the training genomes to further characterize unannotated coding regions or to predict the presence of genes that are missing from the DNA sequences of the training genomes. Pathway prediction approaches such as Gapseq (Zimmerman et al., 2021) could be applied to predict KEGG pathways (sets of interconnected KEGG modules), which may help improve training labels (KEGG module presence/absence) that are present in the complete training genomes, even if one or more enzymes are missing or unannotated after the aforementioned gapfilling strategies have been implemented. This layered approach would facilitate the creation of a more robust set of training data by reducing the occurrence of false negative labels. In addition, MetaPathPredict could be expanded to include the predictive capacity of metabolic processes documented in other metabolic pathway databases that are not contained within the KEGG database. Metabolic pathways from databases including SEED (Overbeek et al., 2005) and MetaCyc (Karp et al., 2002) not found in KEGG could be included in future releases of MetaPathPredict.

In the fourth chapter, I aimed to investigate the environmental factors driving microbial community structure in the hydrothermally-influenced deep subsurface sediments of Guaymas Basin. Previous studies of Guaymas Basin sediments prior to IODP Expedition 385 were largely limited to surficial sediments. These studies have included the analyses of 16S rDNA amplicon and metagenomic samples for archaeal and bacterial microbial lineages, as well as *Beggiatoa* and other sulfur-oxidizing bacterial mats on the sediment surface (Ramírez et al., 2020; Teske et al., 2002; Dhillon et al., 2003; Dombrowski et al., 2017; Teske et al., 2019; Vigneron et al., 2014). Additionally, the 18S rDNA sequences of eukaryotic taxa have been characterized from the surficial sediments (Edgcomb et al., 2002). The deepest samples collected from the subsurface of Guaymas Basin prior to IODP 385 (down to 25 mbsf) were used to measure microbial methanogenesis activity as a part of the Deep Sea Drilling Program Leg 64 (Oremland et al., 1982). The work in this chapter was part of a larger set of aims of the International Ocean Discovery Program Expedition 385, which for the first time collected deep drill core samples over a range of sites and depths spanning up to hundreds of meters below the seafloor. Metagenomes and PCR-

based amplicons of the *mcrA* methanogenesis marker gene from IODP 385 samples, analyzed in separate studies, have provided insights into the sparse populations of methanogens inhabiting the deep subsurface of the Guaymas Basin (Bojanova et al., 2023; Hinkle et al., 2023). The IODP 385 metatranscriptomes were also analyzed individually in a separate study, which found that a significant proportion of transcripts were annotated to cellular maintenance and repair functions, indicating that subsurface cells invest a substantial amount of energy into these processes (Mara et al., 2023), as has been hypothesized previously (Lever et al., 2015; Inagaki et al., 2015). This chapter aimed to broadly characterize genomes of the microbial communities inhabiting the deep subsurface of Guaymas Basin, and to determine the environmental factors that played a role in shaping the taxonomic diversity of communities at different depths and sites.

Analysis of metagenomes and associated environmental parameters revealed that in situ temperature was the main factor influencing microbial diversity. The proximity of an active magmatic sill at one of the sites (Ringvent) resulted in a higher rate of temperature increase with depth compared to cold seep and off-axis sites. Ringvent exhibited a more rapid depth-wise decrease in abundance of microbial lineages in the transition from warm (20-45 °C) to hot (45-62 °C) sediments, and some thermotolerant lineages persisted in the hot sediments at all sampled sites. Key marker genes detected in the metagenomic datasets were indicative of methane, short-chain alkane, hydrogen, sulfur, and nitrogen metabolisms in addition to carbon fixation, fermentation, and hydrocarbon degradation. Despite the variation in microbial community profiles, there was a high degree of similarity in expressed metabolic genes within samples from cool (2-20 °C) and warm sediments including an abundance of expressed genes involved in organoheterotrophic processes. In the hot sediments, there was significantly reduced expression of many metabolic pathways as temperature increased, and observed shifts in certain metabolisms. Trends in diverse deep biosphere communities show that microbial abundance and diversity decrease downcore (Kallmeyer et al., 2012; Starnawski et al., 2017; Kirkpatrick et al., 2019). The MAGs recovered in this analysis were dominated by the Chloroflexota and Thermoproteota phyla, which have been detected in 16S rDNA amplicon studies of the surficial sediments of Guaymas Basin (Vigneron et al., 2014; Teske et al., 2019). The Hadarchaeota, which were most abundant in the hottest samples, have also been detected in surficial sediments from studies of 16S rDNA amplicons (Teske et al., 2019). However, the results of this chapter suggest that only certain lineages of these phyla persist as they are subjected to long-term burial in the deep, hot biosphere.

In addition to the examination of recovered MAGs, this study could have provided additional information by characterizing the gene content of the metagenomes as a whole by examining the metagenomic abundance patterns of all the genes from the metagenomic co-assembly in more detail. Additionally, the gene expression patterns of different protein coding regions of the MAGs (or the entire metagenomic co-assembly) could have been analyzed by mapping metatranscriptomic reads to the genes in the MAGs/metagenomic co-assembly. This would have provided a more refined view of the transcribed genes in the deep subsurface than the assessment done in this chapter, where only the percentages of metatranscriptome reads that mapped to each MAG was calculated.

The fifth chapter involved the development of MetaPathPredict-E, an extension of MetaPathPredict that increased its functionality to include metabolic module predictions for eukaryotes. Pipelines for the recovery and analysis of eukaryotic genomes from environmental samples are now established and actively supported, and there is a steady development of new bioinformatics tools geared towards eukaryotes that continue to improve on and expand upon existing methods (Alexander et al., 2023; Krinos et al., 2021; Neely et al., 2021; Gabriel et al., 2024). As bioinformatic methods and next-generation sequencing technologies continue to improve, they will facilitate the capture of more eukaryotic diversity from environmental samples.

MetaPathPredict-E is a set of deep learning models that have been trained to make predictions for groups of eukaryotes including the Fungi, Streptophyta, Chlorophyta, Excavata, Stramenopiles, Alveolata (not including Apicomplexa) and Rhizaria. MetaPathPredict-E also contains a model trained on all 9 combined training datasets that can make predictions for modules prevalent across all eukaryotic groups in the combined training dataset. MetaPathPredict-E made robust KEGG module predictions for eukaryotic genomes and transcriptomes that were at least 30% complete and is designed to handle incomplete annotation information. As metagenomic and single cell sequencing studies continue to recover additional eukaryotic organisms, MetaPathPredict-E will enable the analysis of genomes and transcriptomes from a broader taxonomic selection.

MetaPathPredict-E could be expanded and improved by incorporating many of the same approaches as described above for MetaPathPredict. The functional predictions could be expanded to include those from additional metabolic pathway databases in addition to KEGG. The annotations of the genomes and transcriptomes used for model training and validation could be

annotated with additional tools and gapfilled to build a more robust set of features (gene annotations) and labels (KEGG module presence/absence) for model creation. The models within MetaPathPredict-E could benefit especially from more training data, as additional eukaryotic genome and transcriptome sequences are made publicly available. Models covering more groups of eukaryotes can trained in the future as datasets for currently sparsely represented groups are made available.

The chapters of this thesis address multiple knowledge gaps in the understanding of marine habitats. The thesis findings have implications for the discovery of novel natural products and showcase evidence of survival and communication strategies for free-living and particle-occupying microbes under varying oxygen concentrations. In addition, this work contributes to the understanding of environmental factors shaping microbial communities in the hydrothermal deep biosphere. Drill core samples provided evidence of the genomic adaptations of microbes inhabiting the hydrothermal deep biosphere and their ability to remain transcriptionally active at depth. The bioinformatics tools developed as part of this work are designed to accurately predict the metabolic potential of environmental genomic and transcriptomic datasets. These tools are adept at handling sparse datasets, making them particularly useful for analyzing data from challenging environments where sampling may be limited. Taken together, this thesis enhances the understanding of microbial ecology in diverse marine settings and provides robust methodologies to aid future studies, thereby contributing to the broader fields of biological oceanography, microbial ecology, and bioinformatics.

Methods in computational biology and bioinformatics continue to expand the ability to recover genomes from environmental samples. The EukHeist pipeline (Alexander et al., 2023) is an automated workflow that recovers both prokaryotic and eukaryotic MAGs from large metagenomic datasets. The EukMetaSanity (Neely et al., 2021) and EUKelele (Krinos et al., 2021) pipelines are scalable workflows to predict eukaryotic gene content and to taxonomically classify eukaryotic genomes or transcriptomes, respectively. These tools can be run in tandem with prokaryotic gene prediction tools such as Prodigal (Hyatt et al., 2010) and annotation tools and pipelines such as Prokka (Seemann 2014), RAST (Aziz et al., 2008), and the NCBI PGAP pipeline (Tatusova et al., 2016) to recover and annotate more genomes encompassing taxa from all domains of life. These pipelines that facilitate the extraction of genomes from environmental datasets continue to expand the contents of genomic databases. This wealth of data enhances the genomic

coverage of diverse, environmentally relevant prokaryotes and eukaryotes and enables more detailed studies of their ecological roles, evolutionary relationships, and potential biotechnological and industrial applications. MetaPathPredict and MetaPathPredict-E complement these methods by providing metabolism predictions for moderately to highly incomplete genomes and transcriptomes extracted from environmental samples with current bioinformatics tools and pipelines.

Large sequence databases can also be harnessed to facilitate methods to infer protein function and structure, which can help inform metabolic pathway prediction models. A deep learning model, LookingGlass, was trained on millions of DNA sequencing reads to accurately classify oxidoreductase reads, predict enzyme optimal temperature, and recognize open reading frames (Hoarfrost et al., 2022). Another deep learning model, ProtENN (Bileschi et al., 2022), was trained on Pfam (Bateman et al., 2004) seed sequences and used in tandem with Pfam HMMs to assign protein family annotations to millions of unclassified proteins in the Pfam database. Protein structure prediction is another area of research that has made substantial progress in the last decade. The third version of the AlphaFold (Jumper et al., 2021) protein structure prediction tool, AlphaFold3 (Abramson et al., 2024), can predict the structure and interactions of DNA, RNA, ligands, and ions in addition to proteins. AlphaFold3 uses a deep learning architecture similar to that of the transformer (a.k.a. generative AI). The transformer is a deep learning architecture based on the multi-head attention mechanism (Vaswani et al., 2017) that is rapidly gaining a technological foothold in computational biology. Generative AI has already been trained for a variety of uses such as to generate functional protein sequences (ProteinGAN; Repecka et al., 2021), antibiotics (Swanson et al., 2024), and antibodies (Hie et al., 2024). Taken together, these emerging methods in the field of machine learning hold promising potential to help inform future analyses of marine habitats.

The integration of diverse data types and the development of novel predictive methodologies are fundamental to advancing our understanding of microbial metabolism in marine habitats. By combining metagenomic, metatranscriptomic, and environmental parameter measurements, researchers can achieve a more comprehensive view of how microbial communities function and interact within their environments. Future research stands to benefit greatly from expanding the scope of comparative studies across different marine ecosystems, ranging from coastal waters to deep-sea hydrothermal vents. Employing advanced computational

tools from the fields of machine learning and systems biology will be essential to unravel the complexities of metabolic processes and microbial interactions. The need for innovative data integration and analytical methods will remain critical as researchers strive to understand the vast network of metabolic processes and microbial interactions present in marine ecosystems. This understanding is not only crucial for marine science but also has practical implications for achieving a more complete understanding of biogeochemical cycling, impacts of climate change mitigation strategies, and the discovery of novel biotechnologically relevant compounds. As marine environments continue to change due to anthropogenic impacts, our ability to accurately predict and manage these ecosystems will depend heavily on the advancements made in data integration and predictive methodologies. I look forward to further investigating the intricate roles of microbial communities in diverse marine ecosystems, their potential for biotechnological applications, and their impact on global biogeochemical cycles.

References

Abramson, Josh, et al., "Accurate structure prediction of biomolecular interactions with AlphaFold 3." *Nature* (2024): 1-3.

Alexander, Harriet, et al., "Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton." *mBio* 14.6 (2023): e01676-23.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., & Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, *36*(7), 2251-2252.

Aziz, Ramy K., et al., "The RAST Server: rapid annotations using subsystems technology." *BMC genomics* 9 (2008): 1-15.

Bateman, Alex, et al., "The Pfam protein families database." *Nucleic acids research* 32.suppl_1 (2004): D138-D141.

Bileschi, Maxwell L., et al., "Using deep learning to annotate the protein universe." *Nature Biotechnology* 40.6 (2022): 932-937.

Blin, Kai, et al.,"antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation." Nucleic acids research 51.W1 (2023): W46-W50.

Bojanova, Diana P., et al.,"Well-hidden methanogenesis in deep, organic-rich sediments of Guaymas Basin." The ISME Journal 17.11 (2023): 1828-1838.

Dhillon, Ashita, et al.,"Molecular characterization of sulfate-reducing bacteria in the Guaymas Basin." Applied and environmental microbiology 69.5 (2003): 2765-2772.

Dombrowski, Nina, Andreas P. Teske, and Brett J. Baker. "Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments." Nature communications 9.1 (2018): 4999.

Edgcomb, Virginia P., et al.,"Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment." Proceedings of the National Academy of Sciences 99.11 (2002): 7658-7662.

Edgcomb, Virginia, et al.,"Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness." The ISME journal 5.8 (2011): 1344-1356.

Gabriel, Lars, et al.,"BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA." Genome Research (2024).

Hinkle, John E., et al.,"A PCR-Based Survey of Methane-Cycling Archaea in Methane-Soaked Subsurface Sediments of Guaymas Basin, Gulf of California." Microorganisms 11.12 (2023): 2956.

Hoarfrost, A., et al., "Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter." *Nature communications* 13.1 (2022): 2606.

Hyatt, Doug, et al., "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC bioinformatics* 11 (2010): 1-11.

Inagaki, Fumio, et al.,"Exploring deep microbial life in coal-bearing sediment down to~ 2.5 km below the ocean floor." Science 349.6246 (2015): 420-424.

Jumper, John, et al., "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.

Kallmeyer, Jens, et al.,"Global distribution of microbial abundance and biomass in subseafloor sediment." Proceedings of the National Academy of Sciences 109.40 (2012): 16213-16216.
Karp, Peter D., et al.,"The metacyc database." Nucleic acids research 30.1 (2002): 59-61.
Kirkpatrick, John B., Emily A. Walsh, and Steven D'Hondt. "Microbial selection and survival in subseafloor sediment." Frontiers in Microbiology 10 (2019): 956.
Krinos et al., (2021). EUKulele: Taxonomic annotation of the unsung eukaryotic microbes. Journal of Open Source Software, 6(57), 2817, https://doi.org/10.21105/joss.02817
Lever, Mark A., et al.,"Life under extreme energy limitation: a synthesis of laboratory-and field-based investigations." FEMS microbiology reviews 39.5 (2015): 688-728.
Madrid, Vanessa M., et al.,"Phylogenetic diversity of bacterial and archaeal communities in the anoxic zone of the Cariaco Basin." Applied and environmental microbiology 67.4 (2001): 1663-1674.
Mara, Paraskevi, et al.,"Microbial gene expression in Guaymas Basin subsurface sediments responds to hydrothermal stress and energy limitation." The ISME journal 17.11 (2023): 1907-1919.
Meleshko, Dmitry, et al.,"BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs." Genome research 29.8 (2019): 1352-1362.
Mara, Paraskevi, et al.,"Viral elements and their potential influence on microbial processes along the permanently stratified Cariaco Basin redoxcline." The ISME Journal 14.12 (2020): 3079-3092.
Neely, Christopher J., et al., "The high-throughput gene prediction of more than 1,700 eukaryote genomes using the software package EukMetaSanity." *bioRxiv* (2021): 2021-07.
Oremland, R. S. et al.,Initial Reports of the Deep-Sea Drilling Project, 64: Washington, DC (U.S. Government Printing Office), 759–762.
Overbeek, Ross, et al.,"The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes." Nucleic acids research 33.17 (2005): 5691-5702.
Palù, M., Basile, A., Zampieri, G., Treu, L., Rossi, A., Morlino, M. S., & Campanaro, S. (2022). KEMET–A python tool for KEGG Module evaluation and microbial genome annotation expansion. *Computational and Structural Biotechnology Journal*, *20*, 1481-1486.
Paoli, Lucas, et al., "Biosynthetic potential of the global ocean microbiome." *Nature* 607.7917 (2022): 111-118.
Ramírez, Gustavo A., et al.,"The Guaymas Basin subseafloor sedimentary archaeome reflects complex environmental histories." IScience 23.9 (2020).
Repecka, Donatas, et al., "Expanding functional protein sequence spaces using generative adversarial networks." *Nature Machine Intelligence* 3.4 (2021): 324-333.
Seemann, Torsten. "Prokka: rapid prokaryotic genome annotation." *Bioinformatics* 30.14 (2014): 2068-2069.
Starnawski, Piotr, et al.,"Microbial community assembly and evolution in subseafloor sediment." Proceedings of the National Academy of Sciences 114.11 (2017): 2940-2945.
Suter, Elizabeth A., et al.,"Free-living chemoautotrophic and particle-attached heterotrophic prokaryotes dominate microbial assemblages along a pelagic redox gradient." Environmental microbiology 20.2 (2018): 693-712.
Suter, Elizabeth A., et al.,"Diverse nitrogen cycling pathways across a marine oxygen gradient indicate nitrogen loss coupled to chemoautotrophic activity." Environmental microbiology 23.6 (2021): 2747-2764.

Swanson, Kyle, et al., "Generative AI for designing and validating easily synthesizable and structurally novel antibiotics." *Nature Machine Intelligence* 6.3 (2024): 338-353.
Tatusova, Tatiana, et al., "NCBI prokaryotic genome annotation pipeline." *Nucleic acids research* 44.14 (2016): 6614-6624.
Taylor, Gordon T., et al.,"Chemoautotrophy in the redox transition zone of the Cariaco Basin: a significant midwater source of organic carbon production." Limnology and Oceanography 46.1 (2001): 148-163.
Taylor, Gordon T., et al.,"Temporal shifts in dominant sulfur-oxidizing chemoautotrophic populations across the Cariaco Basin's redoxcline." Deep Sea Research Part II: Topical Studies in Oceanography 156 (2018): 80-96.
Teske, Andreas, et al.,"Microbial diversity of hydrothermal sediments in the Guaymas Basin: evidence for anaerobic methanotrophic communities." Applied and Environmental Microbiology 68.4 (2002): 1994-2007.
Teske, Andreas, et al.,"Characteristics and evolution of sill-driven off-axis hydrothermalism in Guaymas Basin–the Ringvent site." Scientific Reports 9.1 (2019): 13847.
Vaswani, Ashish, et al., "Attention is all you need." *Advances in neural information processing systems* 30 (2017).
Vigneron, Adrien, et al.,"Phylogenetic and functional diversity of microbial communities associated with subsurface sediments of the Sonora Margin, Guaymas Basin." PloS one 9.8 (2014): e104427.
Von Meijenfeldt, FA Bastiaan, et al.,"Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT." Genome biology 20 (2019): 1-14.
Zimmermann, J., Kaleta, C., & Waschina, S. (2021). Gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. Genome biology, 22, 1-35.