NORTHWESTERN UNIVERSITY

Querying Microbial Community Structure and Function in the Shallow Subsurface:

Observations from Three North American Cave Systems

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Earth and Planetary Sciences

By

Matthew J. Selensky

EVANSTON, ILLINOIS

September 2023

Abstract

Querying Microbial Community Structure and Function in the Shallow Subsurface: Observations from Three North American Cave Systems

Caves are accessible windows into the shallow subsurface, serving as transitional ecosystems between the photosynthesis-dependent surface and the deeper subsurface. Without a source of solar energy to ultimately power carbon (C) fixation (photolithoautotrophy), cave ecosystems are either reliant on surface-derived nutrients, recovering reducing power from the microbial oxidation of redox-sensitive inorganic compounds for C fixation (chemolithoautotrophy), or some combination. Often regarded as detrital, surface-dependent ecosystems with few exceptions, caves are also beginning to be understood as hosts for highly dynamic communities of bacteria and archaea capable of a wide range of biogeochemically relevant metabolisms. As transitional ecosystems between two distinct endmembers, caves represent ideal natural laboratories to query the relationship between microbial community composition and biogeochemical functional potential.

Here, we employ diverse methodologies to investigate the microbial community structures and potential metabolisms of three distinct cave systems across different climatic regions of North America: 1) the semi-arid continental lava caves of Lava Beds National Monument (California, United States); 2) the submerged, humid tropical Sistema Sac Actun (Quintana Roo, Mexico); and 3) the enormous, humid subtropical Mammoth Cave (Kentucky, United States). Despite significant differences in host rock lithology and other abiotic factors, communities from each cave system are united by the presence of metabolically flexible microbiota.

Lipid-specific stable C isotope analysis suggests that Actinobacteriota in biofilms within lava caves at Lava Beds are abundant and actively fix C in the presence of surface-derived organic

C. In the anchialine Sac Actun cave system, a large-scale 16S rRNA gene tag survey demonstrates that metabolically flexible taxa such as *Comamonadaceae* dominate aquifer communities and co-occur with putatively hydrogen- and methane-oxidizing taxa. Throughout Mammoth Cave, several metagenome-assembled genomes classified as Actinobacteriota demonstrate the capacity for hydrogen, carbon monoxide, and/or thiosulfate oxidation to drive RuBisCO-based chemolithoautotrophy. We also provide novel network analysis software (BNGAL) to aid in the modeling and visualization of microbial niche space from complex taxonomic count data. This work identifies several microbial players involved in the mediation of key biogeochemical cycles in the studied cave systems. Surface-independent metabolisms such as chemolithoautotrophy may be a feature of, rather than an exception to, microbial communities in these shallow subsurface environments.

Acknowledgments

Nothing accomplished is ever done alone. This dissertation could not have been possible without the numerous family members, mentors, teachers, colleagues, peers, and friends around me who have encouraged my growth during my time at Northwestern.

To Prof. Magdalena Osburn, my advisor, it has been the thrill of a lifetime to become so familiar with ecosystems as stunning and hauntingly beautiful as caves. Thank you for giving me the opportunity to learn from you, helping me to grow into the scientist (and admittedly amateur caver!) I am today. I will especially treasure the memories of our long summer fieldwork days in California or Kentucky, escaping the heat by furiously surveying and sampling somewhere spectacular below the surface. I am very proud of what we have accomplished together.

I'd like to thank my other committee members for their always helpful advice and support. To Prof. Beddows, thank you for letting me tag along to do field work in the Riviera Maya, a beautiful part of the world I would not have been able to study without you. To Prof. Blair, thank you for always being willing to let me use your laboratory instruments and for being a constant source of good advice and ideas in our committee meetings.

To all members of the Osburn Lab, past and present – thank you for being the pillars of my time at Northwestern. To Dr. Caitlin Casar, thank you for being an unwavering source of advice, support, and friendship. I credit you for introducing me to programming as a scientist, which has profoundly influenced the direction of my research. To Dr. Jamie McFarlin, thank you for always being the one to stay late in the lab to show me how to use various instruments. You were an excellent teacher to a newcomer in OrgGeo like me. To Floyd Nichols, thank you for your reliability, friendship, and willingness to whiteboard anything out with me. To Dr. Nilou Sarvian, thank you for always being willing to talk about anything and make people laugh. To Dr. Andrew

Masterson, thank you for your patience and expert ability to simply explain complicated topics. I truly enjoyed learning isotope geochemistry in the lab from you. To Zoe Vincent, thank you for your expert field assistance and ability to capture any cave critter. To Abby Hsiao, thank you for your general brilliance and sharing your beautiful knitting skills. To Jackson Watkins, Meerah Shah, Bailey Nash, Aiden Burdick, Timothy Coston, Mia Thompson, Kyra Lin, Sohyun Lee, and Hannah Dion-Kirschner, I have sincerely appreciated the time I have spent working with and getting to know each of you. To Dr. Kaycee Morra, thank you for being a kind sharer of coffee. To Dr. Bradley Stevenson, thank you for your patience in showing me molecular biology in the lab and for being a friendly presence. To Dr. Lily Momper, thank you for your advice and fun attitude. To Drs. Fabrizio Sabba and Sarah Ben Maamar, thank you for your friendship and advice to seek opportunities in research computing. I also wish to acknowledge the many other exceptional undergraduates, graduate students, postdocs, and faculty in the Earth and Planetary Sciences Department for your brilliance and sense of community.

To the staff in the Earth and Planetary Sciences Department – especially Lisa Collins, Tia Ng, Negatwa Tewodros, Murielle Harris, and Grace Schellinger – thank you for your constant patience, positive attitudes, and advice on helping me navigate Northwestern over the years. I also wish to extend thanks to my colleagues at Research Computing Services, especially Janna Nugent, Dr. Scott Coughlin, and Dr. Alper Kinaci, who I have enjoyed working with and learning from for much of my graduate education. To my previous mentors at Montana State, Drs. Eric Boyd and Daniel Colman, thank you for giving me my first opportunity to study environmental microorganisms. Working in the Boyd Lab as an undergraduate student played a significant role in my pursuit of a PhD in this field.

I had the immense pleasure of working with dozens of fantastic cavers, scientists, and other explorers while in the field. Thank you to the amazing team I worked with in Mammoth Cave, including Jo, Bob, and Maggie Osburn, Thomas Brucker, Rickard Toomey, Elizabeth Winkler, Rick Olson, Aaron Addison, Sophia Roosth, Zoey Vincent, and Zack Neher. A special thanks goes to Bob Osburn for providing cave mapping and surveying expertise. I also wish to thank National Park Service staff and the Cave Research Foundation for facilitating our research in the park. I also wish to thank every member of the equally excellent team of scientists, cavers, and Park Service staff I worked with at Lava Beds National Monument, CA, especially Drs. Jennifer Blank, Diana Northup, and Brittany Kruger. Finally, I offer my sincere thanks to Patricia Beddows, Natalie Gibb, Alex Fraser, Nikolas Tkachenko, Vlada Dekina, Rory O'Keefe, Vincent Rouquette-Cathala, and Luis Leal for their cave diving and scientific expertise in the Yucatán. Chapter III of this dissertation would have been impossible without their efforts. I wish to additionally thank all Under the Jungle staff, Karyn DeFranco, Alex Farmer, and Kate Haile for their support and leadership in the field as well.

I certainly would not be here without my family, especially my parents, Rick and Amy Selensky. Thank you for teaching me to love and respect the Earth and all its lucky inhabitants. To my sister Christie, brother Jacob, and sister-in-law Stephanie, thank you for always believing in me and pushing me to be the best version of myself. Finally, I cannot thank my partner Zack Neher enough, who has been with me almost every step of the way on this journey, ferrying me to and from airports with duffle bags full of cave rocks and sampling equipment. I appreciate all of you.

This work was funded by the Cave Research Foundation (CRF), National Cave and Karst Research Foundation (NCKRI), and Biologic and Resource Analog Investigations in Low-Light Environments (BRAILLE).

Dedicated to Mom and Dad

They thought I was shooting for the Moon all those years ago,

but somehow, I landed in a cave.

Table of Contents

List of Figures

List of In-Text Tables

List of Online Supplemental Tables and Figures

Online Supplemental Table 2-1: Metadata for Lava Beds National Monument samples.

Online Supplemental Table 2-2: Lipid yields. Yields are reported across three extracted intact polar lipid fractions: glycolipids, phospholipids, and phosphatidylcholine lipids. All values are reported in μg/g TLE.

Online Supplemental Table 2-3: Binned lipid yields. Lipid yields are binned by sample type and lipid class. Column "summed_yield" is equal to the total lipid yield for each lipid class in a given sample type. Correspondingly, "binned_mean" is equal to the mean lipid yield, and "binned_sd" reports the standard deviation. All values are reported in μg/g TLE.

Online Supplemental Table 2-4: Compound-specific isotope values ($\delta^{13}C_{IPL}$). All samples were analyzed in duplicate. Mean values are reported in addition to the standard deviation.

Online Supplemental Table 2-5: Calculated Calvin cycle-based f_L values of lipids. A lipid with an f_L value equal to 1 is interpreted to source 100% of its carbon from *in situ* chemolithoautotrophy via the Calvin cycle, whereas f_L = 0 represents 100% of its carbon sourced from surface-dependent heterotrophy. Refer to the main text for general model assumptions.

Online Supplemental Table 2-6: Calculated reductive acetyl-CoA pathway f_L values of lipids. A lipid with an f_L value equal to 1 is interpreted to source 100% of its carbon from *in situ* chemolithoautotrophy via the reductive acetyl-CoA pathway, whereas f_L = 0 represents 100% of its carbon sourced from surface-dependent heterotrophy. Refer to the main text for general model assumptions.

Online Supplemental Table 3-1: Number of quality-controlled 16S rRNA gene-derived reads per sample.

Online Supplemental Table 3-2: ASV-level taxonomic abundance data rarefied to a sampling depth of 9,957.

Online Supplemental Table 3-3: Metadata for Yucatán aquifer samples.

Online Supplemental Table 3-4: Phylum-level abundance table.

Online Supplemental Table 3-5: Global network EBC prevalence table.

Online Supplemental Table 3-6: Global network EBC memberships for each ASV-level taxon.

Online Supplemental Table 3-7: List of taxa node edge between cluster (EBC) membership for each region. Saved as an Excel table with each tab corresponding to one region.

Online Supplemental Figure 3-1: Interactive global network, filled by edge betweenness cluster (EBC). Open file in a standard web browser to probe individual pairwise relationships. Refer to Figure 3-4 in the main text for a static version of this figure.

Online Supplemental Table 4-1: Concentrations of nitrate, nitrite, ammonia, and sulfate for samples from Mammoth Cave. Values for measured $\delta^{13}C_{org}$, $\delta^{15}N_{bulk}$, $\delta^{15}N_{NO3}$ and $\delta^{13}C_{DOC}$ compositions are also reported.

Online Supplemental Table 4-2: Metagenome-assembled genome (MAG) completeness and contamination as assessed by CheckM. Results are reported for dereplicated MAG set as produced via the drep software. MAG taxonomic classifications are reported from GTDBtk (v. 3).

Online Supplemental Table 4-3: Phylum-level classification key for the Silva (v. 138) and GTDBtk (v. 3) databases.

Online Supplemental Table 4-4: Silva (v. 138) classifications of representative metagenomic 16S rRNA genes obtained via RiboTagger.

Online Supplemental Table 4-5: Rarefied abundance table of 16S rRNA gene counts obtained via RiboTagger.

Online Supplemental Table 4-6: Functional gene annotations via METABOLIC for dereplicated MAGs.

Online Supplemental Table 4-7: Quantification estimates of metagenome-assembled genomes (MAGs) via metaWRAP's quant_bins module. This module calls the software Salmon to report MAG copies per million metagenomic reads.

Online Supplemental Table 5-1: Example ASV-level count table, rarefied to a sampling depth of 10,000.

Chapter I: *Introduction*

Bacteria and archaea are the microbial facilitators of global biogeochemistry. They inhabit every known niche on Earth, expressing diverse and often ancient metabolisms to cycle elements, such as carbon (C) and nitrogen (N), through various redox states to form major biogeochemical cycles that are critical for all life to exist (Falkowski et al., 2008; Strom, 2008). In one sense, microbial ecology is the science of probing the relationship between microbial community composition (structure) and its contributions to broader biogeochemistry (function) across the spectrum of habitable environments.

The querying of microbial community structure-function relationships and their distribution across space has been a fundamental undercurrent of thought since the inception of the field. Sergei Winogradsky was the first to observe the partitioning of different populations of microbes across geochemical gradients in sediment columns (Zavarzin, 2006). While introducing the technique of replicating specific environmental conditions to select for microbial communities expressing a desired function, Martinus Beijerinck stressed that so-called enrichment culturing yields, "… living organisms … appear[ing] under predetermined conditions, either because they alone can develop, or because they are more fit and win out over their competitors" (Schlegel & Jannasch, 1967; Van Niel, 1949). Inspired by Winogradsky and Beijerinck, Lourens Baas-Becking later declared his oft-quoted maxim, "Everything is everywhere, but, the environment selects" (Baas Becking, 1934), eloquently linking microbial community structure and function to spatial dispersal (Martiny et al., 2006). Constraining the identities and distributions of microbes involved in the cycling of elements, especially C and N, across different environments is the foundation for understanding the broader biosphere and its place within the Earth System.

On the modern surface Earth, RuBisCO-based photosynthesis is the dominant process that "fixes" C from an inorganic phase (CO_2 or HCO_3^-) to organic C for biomass building (photoautotrophy; Falkowski et al., 2008). This does not hold true for the subsurface where photosynthesis is inherently limited. Far from being sterile, however, underground habitats teem with diverse microbial communities living at a range of depths, including in deep mines, aquifers, and comparatively shallow caves (Barton & Northup, 2007; Brankovits et al., 2017; Casar et al., 2020; Lavoie et al., 2017; Purkamo et al., 2020; Stevens, 1997). Lacking local, Sun-driven primary productivity, microbial communities in the more oligotrophic subsurface are presented with two strategies to sustain energy and C demands: 1) heterotrophic consumption of allochthonous organic C derived from photosynthesis, or 2) the oxidation of inorganic compounds to power *in situ* C fixation through one of the seven known pathways (chemolithoautotrophy; Figueroa et al., 2018; Stevens, 1997). Microbial communities living in underground habitats that are relatively isolated from the surface, such as deep mines (Casar et al., 2020; Momper et al., 2017), aquifers (Hug et al., 2015), and the isolated, hypogenic Movile Cave (Sarbu et al., 1996) are known to rely on both processes. Considering its significant energetic demands, *in situ* autotrophy in the oligotrophic subsurface might be expected to occur only where organic C is limiting to heterotrophy.

Compared to other subsurface environments discussed above, caves are shallower and are much more intertwined with biogeochemical processes on the surface. Caves can receive significant amounts of photosynthetically sourced dissolved organic C (DOC), N, and other nutrients from incoming fluids (e.g., cave drips, underground rivers) or material brought in by cave macrofauna (Barton & Northup, 2007). Nevertheless, relative to many surface biomes, caves are still considered C- and energy-limiting, so alternative metabolic lifestyles supported by

chemolithoautotrophy may also be significant on an ecosystem scale (Barton & Northup, 2007; Ortiz et al., 2014; Tetu et al., 2013). In this way, caves are important transitional ecosystems between the surface and deeper, more isolated subsurface endmembers. As such, these environments facilitate the study of microbial community structure-function relationships pertaining to biogeochemically relevant metabolisms, such as C fixation and N cycling.

This work explores the structures and potential biogeochemical functions of microbial communities in three cave systems across North America, each representing distinct geological and climatic settings. A broad range of methodologies, from stable isotope analysis to metagenomics, are employed to query the key microbial facilitators of C and N cycling in these diverse cave habitats. From the terrestrial lava tubes of Lava Beds National Monument (California, United States), stable carbon isotope compositions from bacterial membrane lipids suggest the abundance of active chemolithoautotrophs, likely Actinobacteriota, in cave biofilms, despite surface inputs of organic C (**Chapter II**). In the anchialine karst cenotes and submerged conduits of the Sac Actun cave system (Quintana Roo, Mexico), taxonomic marker gene analysis suggests the ubiquitous presence of highly metabolically flexible taxa, such as putatively hydrogenotrophic *Comamonadaceae* spp., across the aquifer. This contrasts with more "specialist" taxa such as the sulfur oxidizing *Sulfurovum*, whose distribution is much more restricted (**Chapter III**). Meanwhile, in the terrestrial karst of Mammoth Cave (Kentucky, United States), functional gene annotations of draft genomes assembled from cave biofilms identify families of Actinobacteriota that are capable of oxidizing ambient concentrations of hydrogen and carbon monoxide gas to support RuBisCO-based chemolithoautotrophy (**Chapter IV**). Finally, to facilitate systems-level approaches that probe microbial structure-function relationships across environments or niche space, novel, open-source network analysis software is described (**Chapter V**).

Chapter II: *Stable Isotope Depletions in Lipid Biomarkers Suggest Subsurface Carbon Fixation in Lava Caves*

2.1 Abstract

Lava caves, formed through basaltic volcanism, are accessible conduits into the shallow subsurface and the microbial life residing there. While evidence for this life is widespread, the level of dependence of these microbial communities on surface inputs, especially that of organic carbon (OC) is a persistent knowledge gap, with relevance to both terrestrial biogeochemistry and the characterization of lava caves as Mars analog environments. Here, we explore carbon cycling processes within lava caves at Lava Beds National Monument, CA. We interrogate a range of cave features and surface soils, characterizing the isotopic composition ($\delta^{13}C$) of bulk organic and inorganic phases, followed by organic geochemical analysis of the distribution and $\delta^{13}C$ signatures of fatty acids derived from intact polar lipids (IPLs). From these data we estimate the carbon sources of different sample types, finding that surface soils and mineral-rich speleothems incorporate plant-derived biomass ($\delta^{13}C_{VPDB}$ ~ -30‰), whereas biofilms are dominated by strongly ^{13}C-depleted lipids (minimum $\delta^{13}C_{VPDB}$ -45.4‰) specific to bacteria, requiring a significant proportion of their biomass to derive from *in situ* fixation of inorganic carbon from previously respired OC. Based on the prevalence and abundance of these ^{13}C-depleted lipids, we conclude that biofilms here are fueled by *in situ* chemolithoautotrophy, despite relatively high concentrations of dissolved OC in collocated cave waters. This unexpected metabolic potential mirrors that found in other deep subsurface biospheres and has significant positive implications for the potential microbial habitability of the Martian subsurface.

2.2 Introduction

Subsurface microbial communities are found across a range of depths on Earth, spanning shallow environments such as caves and aquifers 1 to 100 m deep (Barton & Northup, 2007; Lavoie et al., 2017; Northup et al., 2011; Sarbu et al., 1996) to crustal environments at least 4.4 km deep (Casar et al., 2020; Kormas et al., 2003; Probst et al., 2018; Purkamo et al., 2020). Microbial communities in the subsurface are often sustained by a combination of surface-derived organic carbon (OC) and C fixed *in situ* via chemolithoautotrophy (hereafter lithoautotrophy; Anantharaman et al., 2016; Momper et al., 2017; Probst et al., 2018; Stevens, 1997), but the balance between these sources and how they change with depth remains unknown. The lithoautotrophic potential of deep microbial communities has been a major target of geobiological and astrobiological interest due to their inferred surface independence and implications to the habitational history of Earth (Lollar et al., 2006; Momper et al., 2017; Purkamo et al., 2020). For instance, the subsurface of Mars is potentially habitable for microorganisms, but to be viable this theoretical subterranean biome must persist using lithoautotrophy rather than photosynthesis, owing to the apparent absence of a photosynthetic surface biosphere. This potential may also not be limited to ultra-deep life, as evidence for diverse lithoautotrophic microbes has likewise been found in shallower subsurface biomes including caves, oceanic crust, and the deep critical zone, despite being more directly impacted by surficial processes and C fluxes (Barton & Northup, 2007; Engel et al., 2004; Lavoie et al., 2017; Momper et al., 2017; Orcutt et al., 2015; Purkamo et al., 2020).

Caves are portals through which we can explore the microbial ecology of the shallow subsurface. Cave habitats are generally oligotrophic, though they can receive enough dissolved OC (DOC) from the overlying surface and critical zone to sustain chemoheterotrophic populations

of bacteria, archaea, and fungi (Barton & Northup, 2007). Despite the clear connection between caves and the surface, microbiologists have noted the distinctiveness of cave microbiota for decades, which include chemolithotrophic and heterotrophic taxa (Barton & Northup, 2007; Fliermans & Schmidt, 1977; Lavoie et al., 2017; Northup et al., 2011). Lithotrophs are implicated in cycling crucial elements such as N and S in cave environments (Barton & Northup, 2007; Fliermans & Schmidt, 1977; Hathaway, Sinsabaugh, et al., 2014; Sarbu et al., 1996), but many such microbes have diverse metabolic potential and can couple catabolic energy generation to C fixation (lithoautotrophy), assimilate OC directly (lithoheterotrophy), or do both (mixotrophy). Due to the important role that lithotrophs play in cave nutrient cycling, it is imperative to understand their prevalence, activity, and importantly, C sources.

Evidence for lithotrophy in caves does exist, particularly among N- and S-cycling microbes, varying based on host rock lithology. For instance, a seminal study in Mammoth Cave, KY determined nitrifying *Nitrospira* sp. cell abundances were 100 times denser than those of overlying soils (Fliermans & Schmidt, 1977). Further, DNA sequencing data suggest that lithotrophs oxidize ammonia and nitrite within the karst caves below the Nullarbor Plain, Australia (Holmes et al., 2001; Tetu et al., 2013). Microbial sulfide oxidation, another important lithotrophic metabolism, is observed in caves around the world, including the acidic, sulfur-rich Frasassi Cave of Italy (Hamilton et al., 2015; Macalady et al., 2008) and Cueva de Villa Luz, Mexico (Hose et al., 2000). Although many of the lithotrophs studied in these caves exhibit genomic potential for C fixation, the demonstration of active, *in situ* lithoautotrophy in caves is rare. In a notable case, CH_4-, S-, and N-based lithoautotrophy is the foundation of a subterranean food web in the hypogenic Movile Cave, Romania (Sarbu et al., 1996).

Basaltic caves are especially relevant to planetary speleology due to their igneous origin and inferred presence elsewhere in the solar system (Greeley, 1971; Horz, 1985; Kaku et al., 2017; Keszthelyi et al., 2008; Léveillé & Datta, 2010). Lava caves form as the outer shell of erupting lava solidifies while the liquid interior continues to flow, creating hollow conduits. Microbial communities from these typically shallow caves are found in volcanic regions worldwide, such as the Cascades (Lavoie et al., 2017), the Canary Islands (Gonzalez-Pimentel et al., 2018), the Azores (de los Ríos et al., 2011), as well as Hawai'i and parts of the American Southwest (Northup et al., 2011). These volcanic caves are also thought to be present on Mars (Horz, 1985; Léveillé & Datta, 2010). Microbial biofilms are a prominent feature of many lava caves, with white, tan, and yellow hydrophobic coatings covering the walls in many sites (Northup et al., 2011). These communities harbor microbes distinct from overlying surface soils and include putative lithotrophic and heterotrophic taxa (Hathaway, Sinsabaugh, et al., 2014; Lavoie et al., 2017).

Shallow caves often receive relatively high fluxes of soil-derived OC (Saiz-Jimenez & Hermosin, 1999). If lava cave lithotrophs subsist on this flux of surface-derived OC, then the relevance of lava caves to astrobiology may be limited. There is evidence that lava cave biofilms contain surface-sourced OC. For instance, pyrolysis of microbial biofilms from lava caves in La Palma (Canary Islands) suggests that yellow biofilms acquire their distinctive hue from plant-derived lignin breakdown products (Gonzalez-Pimentel et al., 2018), signaling the incorporation of OC from the surface. However, other lines of evidence hint towards the presence of active lithoautotrophy in lava caves. For example, scanning electron microscopy (SEM) suggests an intimate association between brown cave coralloids and the iron-oxidizing lithoautotrophs *Gallionella* and *Leptothrix* in Azorean lava caves (de los Ríos et al., 2011). To our knowledge, only one lithoautotroph has been isolated from lava caves and described in the literature: an aerobic

Fe(II)-oxidizing member of the *Pseudomonas* (Popa et al., 2012). Direct evidence for *in situ*, quantitatively important lithoautotrophic activity in lava caves has yet to be demonstrated.

The caves of Lava Beds National Monument, CA (hereafter Lava Beds), are well-suited to studying the activity of lithoautotrophs in the shallow subsurface. Many terrestrial lava cave systems, such as those in Hawai'i or the Azores, are located beneath regions of high surficial primary productivity (de los Ríos et al., 2011; Northup et al., 2011). Conversely, Lava Beds is situated in a low-productivity, semi-arid surface environment, thereby more closely mimicking the desiccated conditions of the Martian surface. Caves at Lava Beds experience significant surficial DOC input relative to the deeper subsurface (e.g., Osburn et al., 2019). Nonetheless, the abundance of 16S rRNA genes from taxa capable of C fixation such as Actinobacteria and *Nitrospira* sp. in biofilms here (Lavoie et al., 2017) suggests the coexistence of lithoautotrophs and heterotrophs. Notably, members of the Actinobacteria have been demonstrated to compose at least 39% of the microbial communities in Lava Beds cave biofilms, at least twice as abundant as overlying surface soils (Lavoie et al., 2017). However, DNA amplicon sequencing alone cannot discriminate between lithoautotrophy and heterotrophy, as microbial phylogeny does not necessarily imply function and many taxa are metabolically flexible. While it is possible that such potential lithotrophs derive energy from the oxidation of host rock minerals and/or aqueous substances such as ammonium to power *in situ* C fixation, this remains to be demonstrated here, and requires a different methodological approach.

Stable isotopes can track biological C sources through the lens of biosynthetic fractionation (Hayes, 2001). In a mixed community such as a lava cave biofilm, bulk C isotopic measurements of organic matter integrate signals from all members proportional to their biomass abundance. The homogenization of this isotopic signal hinders the quantitative identification of specific metabolic

signatures. Compound-specific isotopic analysis (CSIA) interrogates individual organic compounds, yielding powerful information about particular organisms that produce diagnostic biomarkers (Close et al., 2014). Fatty acids (FAs) derived from intact polar lipids (IPLs) form the membranes of living bacterial and eukaryotic cells and are commonly used as chemotaxonomic tracers of C metabolism (Boschker et al., 1999; Budge et al., 2008; Schubotz et al., 2013; Schwab et al., 2017). The isotopic compositions of IPL-derived FAs ($\delta^{13}C_{IPL}$) can track C within a mixed community when compared to the isotopic ratios of possible C sources (e.g., Boschker et al. 1999). Furthermore, most IPLs degrade rapidly upon cell death (Logemann et al., 2011), suggesting that IPL quantification reflects the FA composition of extant biomass (Sturt et al., 2004). As such, $\delta^{13}C_{IPL}$ tracks the isotopic contributions from living biomass.

Here, we describe the C isotopic landscape of lava caves at Lava Beds by coupling organic geochemical analysis to bulk and compound-specific isotope geochemistry targeted at broad swath of biological, mineral, and sedimentary features. We first characterize the isotopic composition and abundances of organic and inorganic phases within these features. Next, we present IPL-derived FA distributions and their individual isotopic compositions. Finally, we estimate the relative contributions of C sourced from the surface vs. *in situ* C fixation with an isotope mass balance model. This work identifies abundant and active lithoautotrophy in the shallow subsurface, informing the use of the terrestrial subsurface as an extraterrestrial analog.

2.3 Materials and Methods

2.3.1. Field site and sampling approach

Samples (n = 91) were collected from nine caves and overlying soils from Lava Beds (Figure 2-1A). Lava Beds is located on the northern slope of the Medicine Lake volcanic shield in the semi-arid shrublands of northern California, USA (Waters et al., 1990). Since its genesis 500

kya, Medicine Lake volcano has erupted hundreds of times (Donnelly-Nolan et al., 2007), the flows of which are responsible for over 830 lava caves identified to date at Lava Beds. Most caves here were created by the Mammoth flow (Waters et al., 1990) which occurred 35 ± 5 kya, although some were formed during later flows throughout the Pleistocene and Holocene, with the latest eruption occurring only 950 years ago (Donnelly-Nolan et al., 2007). This work encompasses samples from caves B090, B140, L300, L480, L490, L780, L820, L853, and V460.

We systematically sampled a variety of putatively biogenic and abiogenic features from each cave (Figure 2-1B-E) to produce a broad overview of the lipid and isotopic landscapes of the area. Cave samples originated from a range of depths below the surface, horizontal distances from entrances, and relative humidity conditions. Surface soil samples were also taken above each cave. Solid samples were collected in pre-combusted glass jars using solvent-cleaned chisels and scoopulas. Cave drip water samples were collected where possible in pre-combusted glass vials and were frozen at -20°C within four hours.

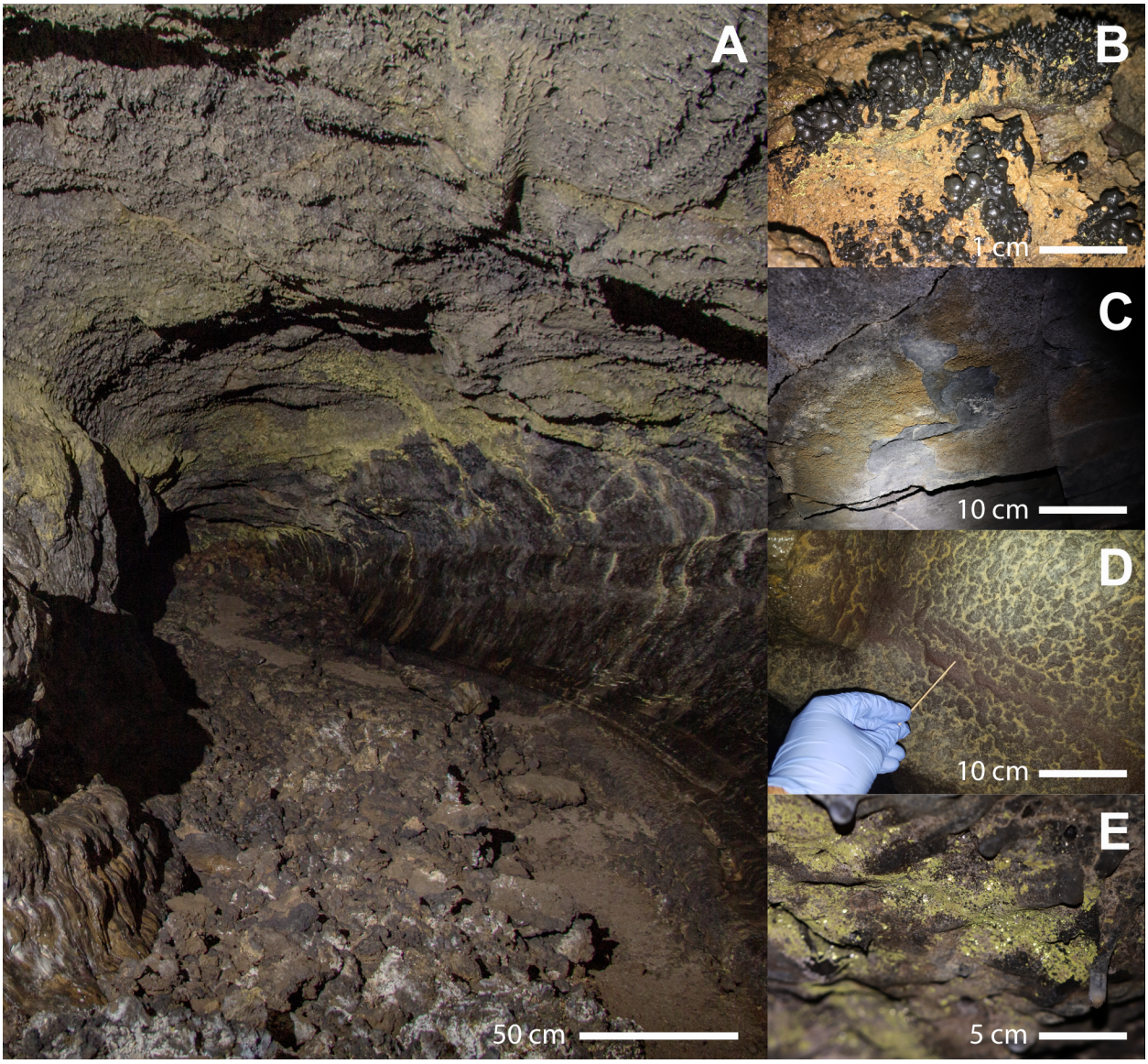


Figure 2-1: ***Lava caves and their features at Lava Beds National Monument, CA.*** A: A typical cave from Lava Beds. Microbial biofilms are visible as bright yellow and tan colors on the ceiling. B: A cluster of black cave polyps. C: Ceiling mineral crusts, an example of a mineral feature. D: Tan-colored ooze attached to a cave wall. E: A yellow biofilm covered in water droplets adhered to the surface of lava cave basalt. Approximate scale bars are provided in the bottom right corner of each photo.

2.3.2. Intact polar lipid (IPL) extraction

Samples for lipid analysis were kept frozen at -20°C until processing. Lipids were extracted using a modified Bligh Dyer method (Bligh & Dyer, 1959; Schubotz et al., 2013). Briefly, freeze-dried basalt samples were crushed with a solvent-washed mortar and pestle. Weighed aliquots of each sample (0.1-30g, depending on the sample type; Online Supplemental Table 2-1) were then sonicated in extraction solvents and centrifuged, retaining the supernatants. Extraction included 2X extractions with 4:5:10 10 mM phosphate buffer:dichloromethane (DCM):methanol (MeOH), 2X with 4:5:10 10 mM trichloroacetic acid (TCA) buffer:DCM:MeOH, and 1X with 1:2 DCM:MeOH to broadly target compounds with polar headgroups (Schubotz et al., 2013). Extracts were pooled then separated into two-phase solutions with additional water and DCM. The organic phase was collected into preweighed glass vials, washed with water, and dried down under a stream of dry $N_{2(g)}$, after which the total lipid extract (TLE) mass was measured.

2.3.3. IPL separation, derivatization, and characterization

The TLE was separated into fractions based on headgroup chemistries (Close et al., 2014). The TLE was applied to pre-combusted 60 Å silica gel columns, then eluted with the following solvent schedule: neutral hydrocarbons (3:1 hexanes:ethyl acetate); glycolipids (3:1 ethyl acetate:MeOH); phospholipids (MeOH); and residual polar lipids (4:1 MeOH:water). Lipids from the three IPL fractions (glyco-, phospho-, and residual polar lipids; GL, PL, and r-PL, respectively) were then derivatized to fatty acid methyl esters (FAMEs) using a base-catalyzed transesterification with anhydrous MeOH (Kopf et al., 2016). FAMEs were identified by GC-MS using a ThermoFisher Trace GC 1310 coupled to a flame ionization detector (FID) and an ISQ quadrupole mass spectrometer. A Zebron© ZB-5 capillary GC column (30 m, 0.25 mm ID, 25 μm film thickness) was used with He carrier gas flowing at 1.2 mL min^{-1}. The GC program initially

held at 100°C for 1 min, followed by a 14.5°C min^{-1} ramp to 180°C, a 2°C min^{-1} ramp to 240°C, and a 12°C min^{-1} to 320°C, with a final 4-minute hold. Compounds were quantified relative to an internal standard (palmitic acid *iso*-butyl ester) using FID peak areas.

To identify the position of the methyl moieties in branched saturated FAs after GC-FID-MS, we compared the abundances of fragment ions containing the tertiary carbon for each possible structure. Two methyl-branched FAs with 17 C atoms exhibited almost indistinguishable mass spectra although they had very different retention times (Supplemental Figure S2), with large *m/z* peaks at 241 ($[M-43]^+$) and 199 ($[M-85]^+$) consistent with *iso*- and 10-methyl branching, respectively. Comparison to an external 10-methyl $C_{17:0}$ FAME standard (Matreya Lipids and Biochemicals, State College, PA) confirmed the identity of the later-eluting FAME. We assign the other compound as 10,14-dimethyl pentadecanoic acid (10,14-DiMe $C_{17:0}$) on the basis of retention time and the increased abundance of an $[M-15]^+$ peak when compared to the standard.

2.3.4. Determination of double bond positions

The double bond positions of unsaturated FAMEs were determined by derivatization to dimethyl disulfide (DMDS) adducts. Unsaturated FAMEs were first separated from saturated compounds via silver thiolate chromatography (Aponte et al., 2012) from a subset of samples spanning all detected unsaturated FAMEs. The sample was transferred to silver thiolate columns in hexanes after column conditioning (three column volumes each of MeOH, DCM, and hexanes). Saturated FAMEs were collected with four column volumes of hexane eluent. Monoenes, dienes, and trienes were eluted in four column volumes of DCM then dried under a stream of dry $N_{2(g)}$. Unsaturated fractions were derivatized with DMDS to produce DMDS adducts (Shibamoto et al., 2016). Dried unsaturated fractions were resuspended in 0.1 mL of 1.3% iodine in DMDS solution and lightly shaken in a 30°C incubator for 24 hours. DMDS-treated samples were added then

eluted with 4 mL of 10% diethyl ether in hexanes on a $Na_2S_2O_3$(aq)-conditioned, pre-combusted Extrelut NT© column. Adducts were dried under $N_{2(g)}$ and then analyzed and identified using GC-MS.

2.3.5. IPL-specific $\delta^{13}C$

Compound-specific carbon isotope ratios of individual IPL-derived fatty acids ($\delta^{13}C_{IPL}$) were determined with a Thermo Trace 1310 GC coupled to a Thermo Delta V^+ isotope ratio MS via a combustion interface (GC-C-IRMS). The GC used a ZB-5MS capillary column (30 m, 0.1 mm ID, 1 μm film thickness), He carrier gas at 1.4 mL min^{-1}, and a temperature program identical to the one used with GC-FID-MS. Eluted FAMEs were combusted to $CO_{(g)}$ within a Pt, Ni, and 2Xcu wire bundle held at 940°C plumbed directly into the GC outlet. All samples were analyzed in duplicate, with a precision of approximately 0.2‰. Tank $\delta^{13}C$ values were converted to the Vienna Pee Dee belemnite (VPDB) scale through repeated analysis of the A6, A7, and B5 standards (A. Schimmelmann, Indiana University) of C_{16}-C_{30} *n*-alkanes, which were run every three sets of duplicates. We screened samples for low intensity peaks, only measuring the $^{13}C/^{12}C$ ratio for those with i_{45} greater than 100 mV. The range of sample peak intensities (0.1 V to 0.7 V) was always within that bracketed by the standards. We additionally screened for variability in individual compounds by removing peaks with a root mean square error (RMSE) greater than 0.3. The working dynamic range of peak intensities (0.1 V to 15 V) in the A7/A6/B5 standards (Appendix Figure 2-1) demonstrated that linearity produced <0.2‰ shifts ($\Delta\delta^{13}C$) over this range, and that it was appropriate for the reported $\delta^{13}C_{IPL}$ values.

2.3.6. Bulk carbon and nitrogen analyses

Total organic carbon (TOC) abundance as well as its stable carbon isotopic composition ($\delta^{13}C_{TOC}$) was determined using a Costech 4010 Elemental Analyzer coupled to a Thermo Delta

V^+ IRMS through a Conflo IV interface (EA-IRMS). Freeze-dried samples were powdered, acidified with 0.1 M HCl, rinsed in MilliQ water, then freeze dried again and weighed into tin capsules for analysis. The $\delta^{13}C$ values were ascertained by cross-checking with acetanilide and urea standards of known composition (-29.5‰ and -8.0‰ vs. VPDB, respectively for carbon Schimmelmann et al., 2009).

Dissolved inorganic carbon (DIC) concentrations and isotopic compositions ($\delta^{13}C_{DIC}$) of lava cave drip water were measured using a Thermo Gas Bench II coupled to a Thermo Delta V^+ IRMS. As described above, raw sample $CO_{(g)}$ mass peaks were compared to known $\delta^{13}C$ values of internal standards ($R^2 = 0.9997$): "NHC2" ($NaHCO_3$ Macron Chemicals), -2.70‰; "NUCLM1" (Carrara Lago Marble), +2.31‰; "NBS18" (IAEA calcite), -5.01‰. Dissolved organic carbon (DOC) concentrations and stable isotopic compositions ($\delta^{13}C_{DOC}$) of drip waters were measured at the UC Davis Stable Isotope Facility using an O.I. Analytical Model 1030 TOC Analyzer interfaced to a PDZ Europa 20-20 IRMS.

2.3.7. Statistical analyses and data visualization

All statistical analyses were performed using R version 3.6.1. Complete-linkage clustering was performed on a Manhattan distance matrix that was calculated from log_{10}-transformed relative lipid abundance data. Manhattan distance was calculated due to the large number of zero values in the lipid dataset. All dendrograms were visualized using base R and the packages ggtree (Yu et al., 2017) and ggplot2 (Wickham, 2016). Dimensionality reduction was performed via non-metric multidimensional scaling (NMDS) on a Bray-Curtis dissimilarity matrix that was calculated from relative lipid abundances using the package vegan (Oksanen et al., 2019). NMDS was chosen over other multivariate analyses such as principal component analysis (PCA), as NMDS handles sparse

data more effectively. All data and R scripts are available at https://zenodo.org/record/5016282 (DOI: 10.5281/zenodo.5016282).

2.4 Results

2.4.1. Field observations and sample descriptions

Lava cave samples (n = 91) were classified into four major types based on visual characteristics in the field: polyps/coralloids (n = 21), mineral features (n = 20), oozes (n = 8), and biofilms (n = 28; Figure 2-1B-E). Polyps (Figure 2-1B) are elongated accretionary mineral assemblages that are finely laminated in cross-section, resembling small stromatolitic structures observed in other caves (Daza Brunet & Bustillo Revuelta, 2014; Melim et al., 2001; Rossi et al., 2010). Polyps can be brown, white, or black in color, and are found on floors and walls throughout a wide range of depths and climatic conditions. Similar elongated mineral assemblages exhibiting a more dendritic morphology and lacking lamination are called coralloids but are binned with polyps here due to their similarities in form, environment, and resultant lipid compositions. Our "mineral features" category includes silica and calcite crusts, amorphous silica patinas, and other mineral structures (Figure 2-1C) which are often found in drier areas. Oozes are soft, white to ocherous in color, occasionally have a granular texture, and are found in notably damp areas (Figure 2-1D). Biofilms are thin (< 1 mm), biogenic hydrophobic veneers that coat basalt surfaces (Figure 2-1E) and are subclassified into tan (n = 7), yellow (n = 9), white (n = 9), and mixed (n = 3) types. When cave humidity is at 100%, these films are commonly found covered with beads of water. Subcategories of biofilms may be co-located but more typically cover distinct surfaces on the cave walls and ceilings. "Mixed biofilm" samples contain mixtures of biofilm colors. Soil samples were taken from above and within the caves. When compared to surface soils, cave soils are finer-grained and darker in color, often containing woody debris when located near a cave

entrance. Surface soils are lighter in color, sandier in texture, and are rich in pumice derived from the nearby Glass Mountain eruption which blanketed the area approximately 900 years ago (Donnelly-Nolan et al., 2007). Sparse vegetation near surface soil sampling sites included dryland shrubs, grasses, and conifers.

2.4.2. Distribution and abundance of IPL-derived fatty acids

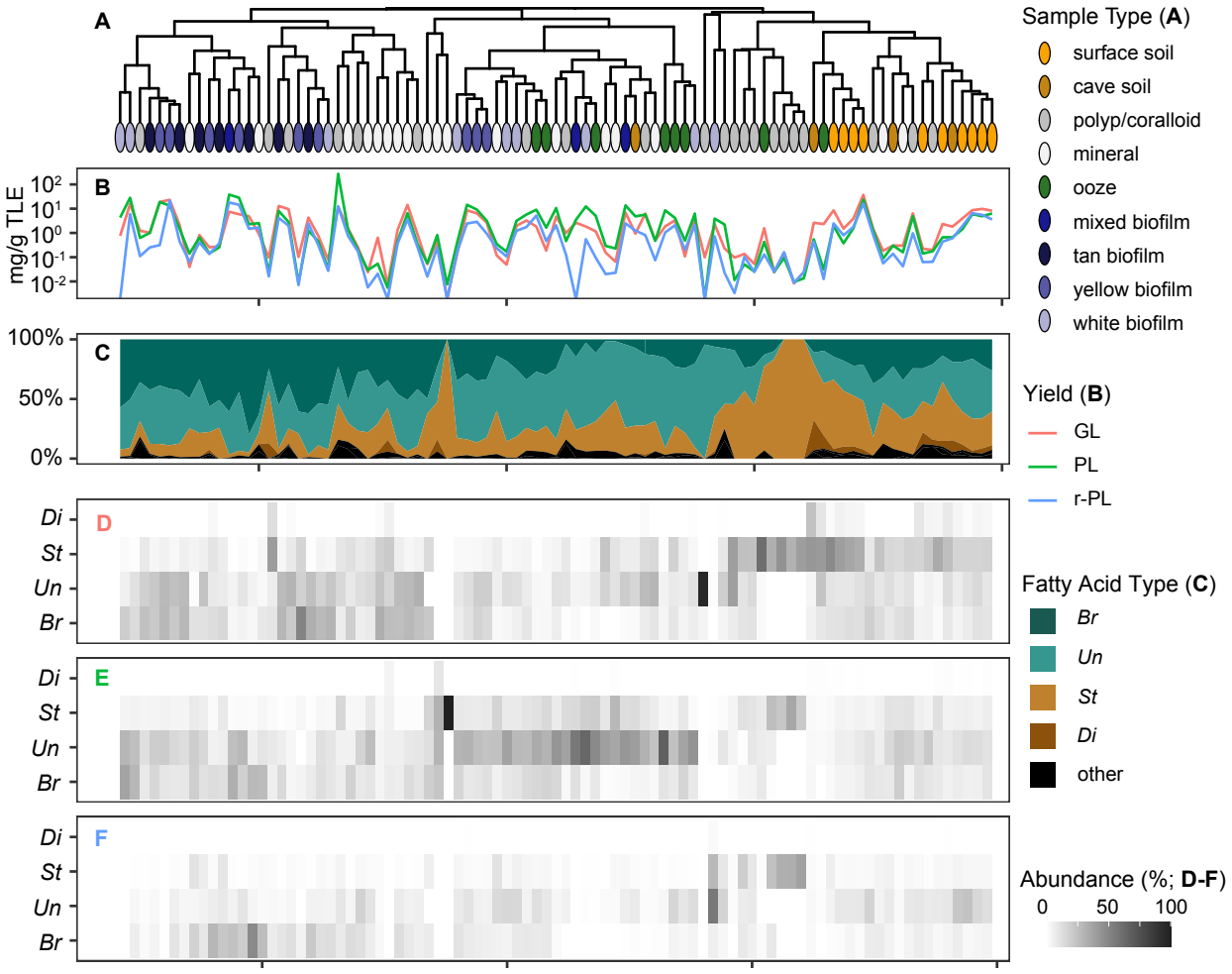


Figure 2-2: ***Lipid distributions across soils and lava cave features.*** Columns contain data from individual samples shown in panels. A: Complete-linkage hierarchical clustering analysis based on relative abundances of individual lipids. B: Total IPL-derived fatty acid yields for each sample (mg lipid/g total lipid extract, TLE) by IPL class (GL: glycolipids; PL: phospholipids; r-PL: residual polar lipids). C: Binned relative abundances of major fatty acid classes summed across IPL fractions (Br: branched saturated; Un: straight unsaturated; St: straight saturated; Di: diacid). D-F: Binned relative abundances of major fatty acid types from each IPL fraction (D: GL; E: PL; F: r-PL).

Of 106 unique FAs detected across three IPL fractions, we observe 36 methyl-branched FAs, 30 unsaturated FAs (26 monoenes, two dienes, and two polyenes), 21 straight chain FAs, and 19 diacid isomers (Online Supplemental Table 2-2). We also observe one $C_{17:0}$ alkenone, two

cyclic $C_{17:0}$ FAs, ricinoleic acid, one isoprenoid, six sterols, and 10 unknown peaks, for a total of 127 distinct compounds in this dataset (Online Supplemental Table 2-2). Methyl moieties are present in the *iso*, *anteiso*, C_2, and C_{10} positions in branched FAs. Monounsaturated FAs contain double bonds at the ω5, ω7, ω8, and ω9 positions. By examining the retention times of DMDS adducts, we determined that two monoenes ($C_{16:1\text{-}\omega7}$ and $C_{17:1\text{-}\omega8}$) occur in *cis* and *trans* conformations (Appendix Figure 2-2). Di- and polyene FAs were not abundant enough to determine double bond positions. Across the entire dataset, unsaturated FAs are the most abundant in terms of summed yield relative to the TLE (sum 415 mg/g), followed by methyl branched (334 mg/g), straight chain (275 mg/g), and diacids (8 mg/g). Both individual and binned FA yields vary considerably by sample type (Online Supplemental Tables 2-2 and 2-3).

To explore similarity between soils and cave features, we performed hierarchical clustering analysis using the relative abundance of individual IPL-derived FAs (Figure 2-2A). First-order branching reveals two clades, one containing most soil samples and cave polyps/coralloids, and another containing most cave biofilms, oozes, and minerals. In the soil-rich clade, second-order branching separates a white biofilm from soils and polyps/coralloids while third-order branching distinguishes soils from most polyps/coralloids. Some cave minerals, polyps/coralloids, and oozes appear throughout this clade (Figure 2-2A). Second-order branching in the biofilm-rich clade differentiates most cave biofilms and mineral features from oozes, polyps/coralloids, and a cave soil (Figure 2-2A). Despite these general patterns, it should be noted that cave minerals and polyps/coralloids are found in most major clusters. Similarly, oozes are also well-dispersed across the dendrogram (Figure 2-2A).

Total IPL yields are highly variable across the dataset, spanning 6 orders of magnitude ranging from 7×10^{-3} to 3×10^{3} mg/g TLE (Figure 2-2B). Outliers in each group cause this large

spread; for example, although most samples contain ~1-10 mg IPL/g TLE, one polyp contains greater than 10^3 mg IPL/g TLE (Figure 2-2B). When outliers are removed, mixed biofilms tend to have the highest average IPL yields (31 ± 29 mg/g), followed by yellow biofilms (22 ± 22), surface soils (17 ± 22), white biofilms (11 ± 15), tan biofilms (7 ± 8), oozes (6 ± 5), cave soils (4 ± 2), polyps/coralloids (3 ± 5), and mineral features (3 ± 6). Relative to other IPL fractions, glycolipid (GL) yield is highest in soils whereas nearly all cave features contain more phospholipids (PLs; Figure 2-2B). Residual polar lipid (r-PL) yields were consistently the lowest regardless of sample type (Figure 2-2B).

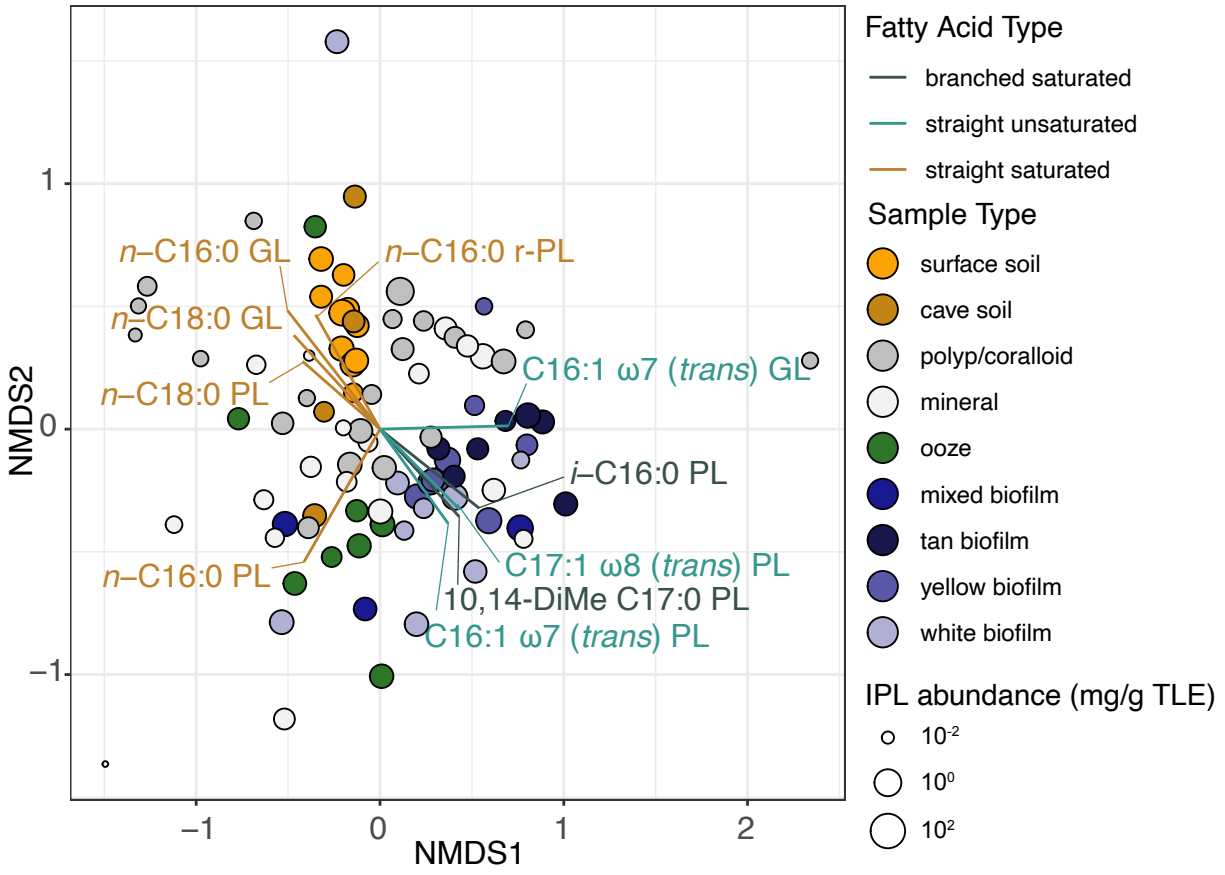


Figure 2-3: ***NMDS based on relative lipid abundance.*** Solution stress = 0.199. The top 10 IPL-derived FA loadings ($p < 0.01$) are plotted to visualize the main sources of variance in the dataset. Point sizes are proportional to the total lipid yield across IPL fractions (mg/g total lipid extract, TLE).

To further explore the distributions of IPLs across cave morphologies and differentiate sample types on the basis of lipids, these data were compared via non-metric multidimensional scaling (NMDS; Figure 2-3). This analysis corroborates the patterns visible from hierarchical clustering, specifically that yellow and tan biofilms are distinct from soils and other cave morphologies (Figure 2-3). Plotted vector loadings reveal the contributions of the 10 most variant individual lipids ($p < 0.01$) in the dataset that are the main drivers of spatial distribution in the ordination. Here, the abundance of *trans*-unsaturated and branched saturated FAs tend to drive the separation of biofilms from other cave features and soils (Figure 2-3). With the exception of a GL-derived *trans*-$C_{16:1\text{-}\omega7}$ FA, these biofilm-associated FAs, specifically *i*-$C_{16:0}$, 10,14-DiMe $C_{17:0}$, *trans*-$C_{16:1\text{-}\omega7}$, and *trans*-$C_{17:1\text{-}\omega8}$, all derive from the PL fraction. *Trans*-$C_{16:1\text{-}\omega7}$ GL ordinates primarily with tan biofilms, while its PL counterpart ordinates with yellow and some white biofilms (Figure 2-3).

Soil samples cluster together and ordinate with GL- and r-PL-derived *n*-$C_{16:0}$ and GL- and PL-derived *n*-$C_{18:0}$ (Figure 2-3). Most oozes ordinate with the PL-derived *n*-$C_{16:0}$ FA. Unlike biofilms or soils, polyps and mineral features do not discernibly cluster in this ordination (Figure 2-3), consistent with hierarchical clustering (Figure 2-2A). These sample types do not cluster with others (e.g., biofilms) taken from the same cave (Appendix Figure 2-3). Variability within polyp/coralloid and mineral samples alone is also not explained by the cave from which each sample originated (Appendix Figure 2-4).

2.4.3. Bulk and IPL-specific Stable Carbon Isotopes

To constrain the reservoirs of C available to microbes at Lava Beds, we measured $\delta^{13}C$ values of bulk phases including DIC ($\delta^{13}C_{DIC}$) and DOC ($\delta^{13}C_{DOC}$) of cave waters, and total organic carbon of solid materials ($\delta^{13}C_{TOC}$). Mean $\delta^{13}C_{TOC}$ values are more negative in lava cave

biofilms relative to surface soils, although $\delta^{13}C_{TOC}$ values tend to be close to $\delta^{13}C_{DOC}$ compositions regardless of sample type (Figure 2-4). Mean $\delta^{13}C_{DIC}$ values of drip water was -6.1 ± 1.9‰ (n = 10; Figure 4), while the mean $\delta^{13}C_{DOC}$ value was -29.8 ± 2.0‰ (n = 10; Figure 4). Neither DOC nor DIC concentrations and isotope compositions correlated with location by cave or nearby sample types.

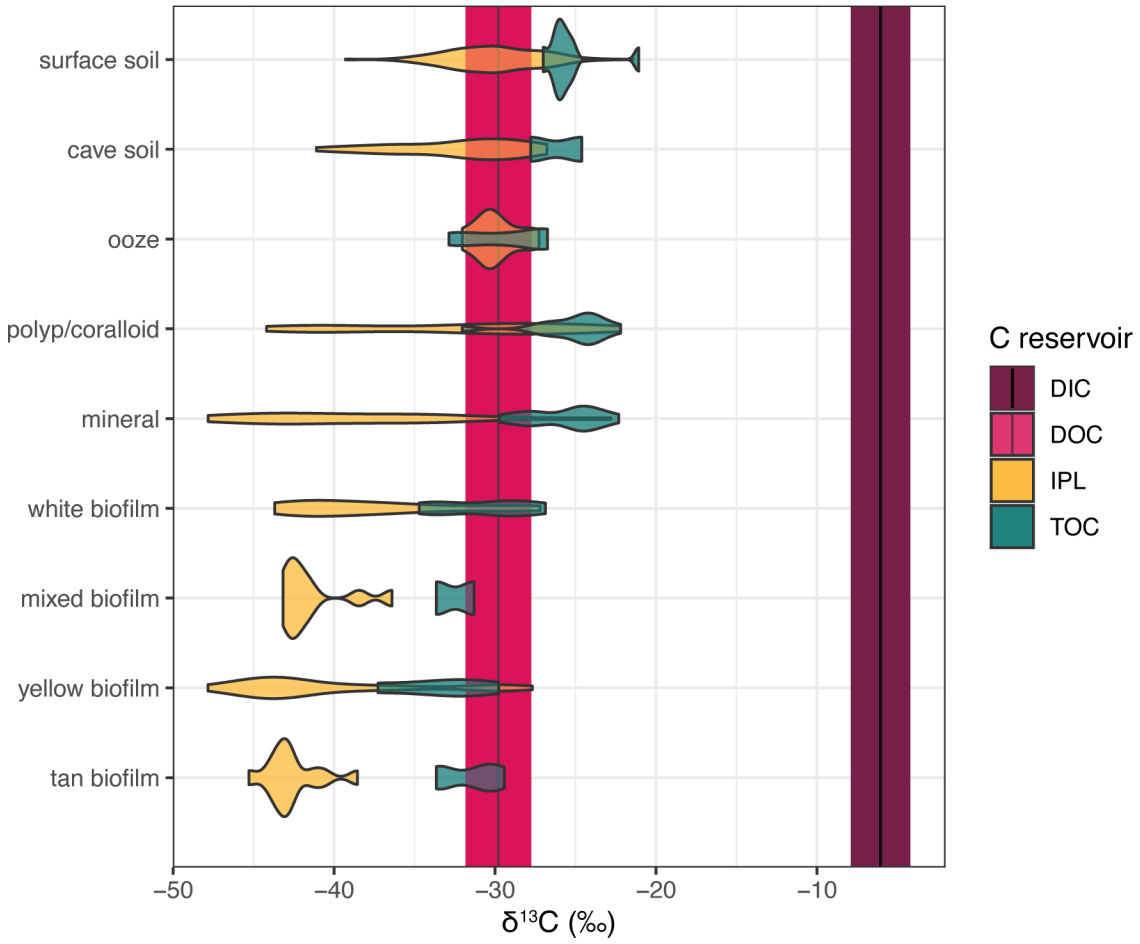


Figure 2-4: ***Stable carbon isotope compositions of C reservoirs at Lava Beds.*** Mean $\delta^{13}C_{DIC}$ and $\delta^{13}C_{DOC}$ values from cave drip waters are illustrated by vertical lines with respective standard deviations as filled rectangles. Individual $\delta^{13}C_{TOC}$ and $\delta^{13}C_{IPL}$ values for each sample type are represented by violin plots reporting their density distributions.

Compound-specific carbon isotope values of IPL-derived fatty acids ($\delta^{13}C_{IPL}$) were measured from samples with sufficiently high yield, in order to identify the C sources used by

living cave biomass. Most sample types contain lipids with mean $\delta^{13}C_{IPL}$ values that are more negative than $\delta^{13}C_{DOC}$ or corresponding mean $\delta^{13}C_{TOC}$ values, with the notable exception of oozes (Figure 2-4). However, we observe distinct distributions of $\delta^{13}C_{IPL}$ values by sample type, with strong ^{13}C depletions in biofilms compared to soils and oozes despite some overlap (Figure 2-4, Online Supplemental Table 2-4). In general, the spread of $\delta^{13}C_{IPL}$ values is lower in tan biofilms compared to soils and other cave features (Figure 2-4, Online Supplemental Table 2-4). By contrast, $\delta^{13}C_{IPL}$ distributions in yellow and white biofilms are bimodal (Figure 4). Yellow and white biofilms tend to have high abundances of *n*-$C_{16:0}$ and $C_{18:1\text{-}\omega7}$ FAs that are ^{13}C-enriched enriched relative to other biofilm FAs, whereas tan biofilms tend to lack these FAs (Figure 2-5).

Biofilms exhibit distinct and depleted $\delta^{13}C_{IPL}$ values compared to soils and most mineral features (Figures 2-4 and 2-5). For instance, mean $\delta^{13}C_{i\text{-}C16:0}$ from biofilms = -43.9 ± 1.3‰ compared to -31.0 ± 1.3‰ in soils. However, one lipid ($C_{18:1\text{-}\omega7}$) consistently exhibits more ^{13}C-enriched $\delta^{13}C_{IPL}$ values across cave morphologies (mean $\delta^{13}C_{18:1\omega7}$ = -30.9 ± 2.3‰ in biofilms and -29.0 ± 0.8‰ in surface soils). The long chain FAs *n*-$C_{20:0}$, *n*-$C_{22:0}$, and *n*-$C_{24:0}$ were only measured in the soil samples, where their yields were sufficient for analysis, and their mean $\delta^{13}C_{IPL}$ values were -31.5 ± 1.6‰, -30.6 ± 1.4‰, and -31.5 ± 1.3‰, respectively (Figure 2-5).

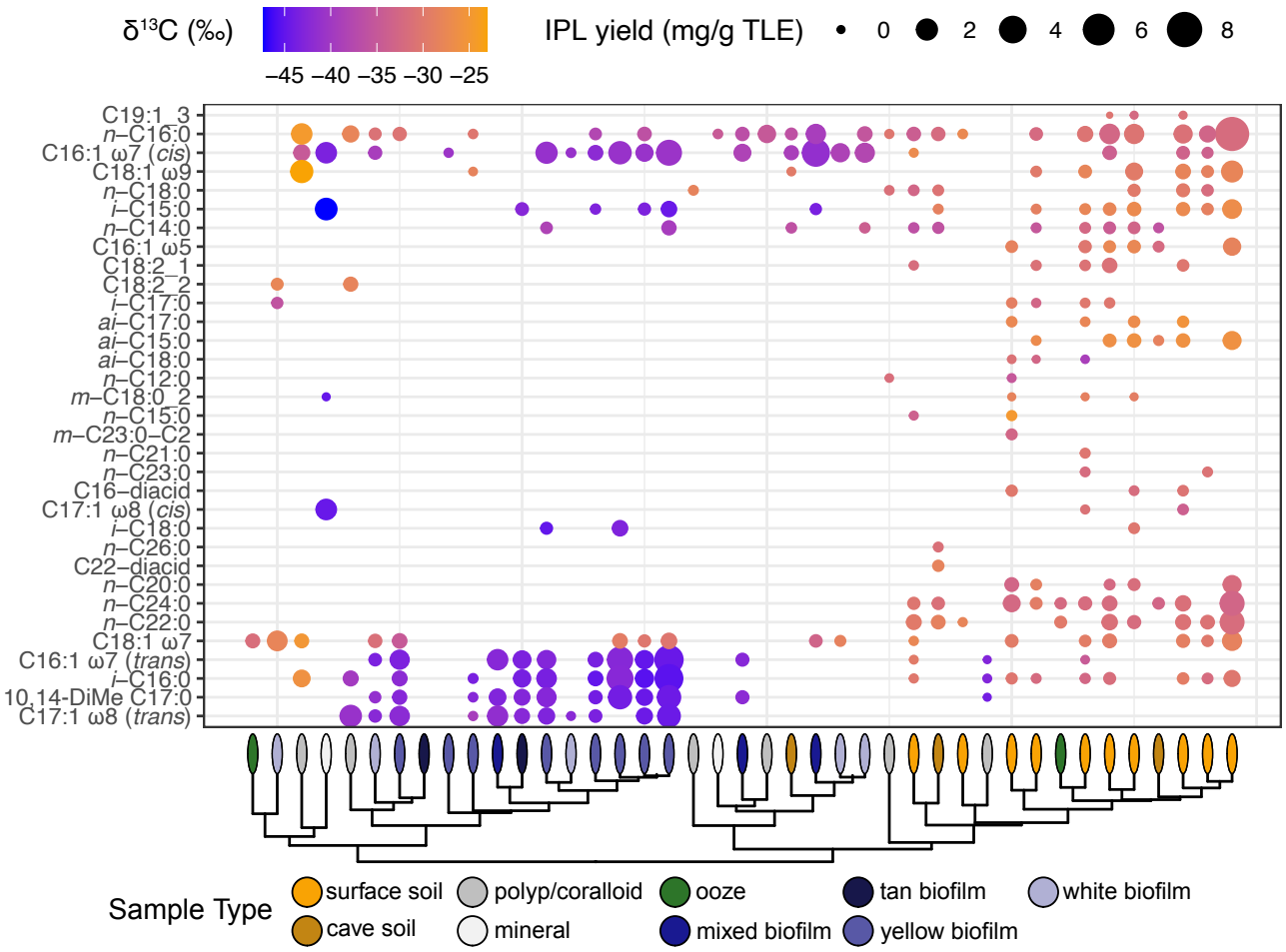


Figure 2-5: ***Cave biofilms are ^{13}C-depleted relative to soils and other cave features.*** The color of each point corresponds to the mean $\delta^{13}C_{IPL}$ of duplicate measurements of each glycolipid, while the size is proportional to the abundance of each glycolipid relative to the total lipid extract (TLE). Complete-linkage clustering was performed on a Manhattan distance matrix that was calculated from log-transformed relative lipid abundance data from the samples analyzed by GC-C-IRMS.

2.5 Discussion

2.5.1. A diversity of fatty acids

We report a diverse ensemble of IPL-derived FAs present within these samples (Figures 2-2, 2-3, and 2-5), particularly among branched and monounsaturated structures, totaling 106 individual fatty acid structures. Of the 36 branched saturated FAs, *iso*-, *anteiso*-, and midchain-methyl branched fatty acids are all abundant, especially in biofilms (Figures 2-2 and 2-3). Although *iso*- and *anteiso*-branched FAs are common bacterial lipids found across a range of environments (Kaneda, 1991), the abundance of midchain-branched FAs as well as the richness of branched FAs we observe are notable. Lipid profiles differ by sample type, which we will discuss here.

Biofilms: White and yellow biofilms are rich in lipids, with IPL yields (mg/g TLE) occasionally exceeding those of surface soils (Online Supplemental Table 2-2). Most cave samples were very tightly adhered to the host basalt (Figure 2-1B-E). As a result, significant portions of rock were included in some samples, thereby "diluting" absolute lipid yields. To overcome this, individual FA yields are reported relative to TLE abundance and so are not as affected by this dilution.

Cave biofilms are dominated by bacterial lipids, specifically branched saturated and unusual unsaturated FAs (Figures 2-2C-F and 2-3). Of the branched FAs, PL-derived *i*-$C_{16:0}$ and 10,14-DiMe $C_{17:0}$ ordinate closest to yellow and white biofilms, along with PL-derived *trans*-$C_{16:1-\omega7}$ and *trans*-$C_{17:1-\omega8}$ (Figure 2-3). Tan biofilms, however, are mainly associated with GL-derived *trans*-$C_{16:1-\omega7}$, suggesting differences in bacterial taxonomy. Both branched and *trans*-unsaturated FAs are bacterial biomarkers, the latter being especially prevalent in Gram-negative strains (Kaneda, 1991; Keweloh & Heipieper, 1996). The predominance of these FAs may also reflect

homeoviscous membrane adaptation to the relatively cold temperatures (4-8°C) of the cave walls (Siliakus et al., 2017), although temperatures are consistent between cave features.

One of the most abundant FAs from cave biofilms, 10,14-DiMe $C_{17:0}$ (Figure 2-5), is attributed to Actinobacteria, a diverse phylum of Gram-positive bacteria. To our knowledge, this double-branched FA has been thus far exclusively found in cultured representatives of the Actinobacterial genus *Pseudonocardia* (Reichert et al., 1998), although similar dimethyl acids are produced by other Actinobacteria such as *Gaiella* (Albuquerque et al., 2011) and *Crossiella* spp. (Labeda, 2001). Intriguingly, members of the family *Pseudonocardiacea* were previously demonstrated to be the dominant type of Actinobacteria in cave biofilms from Lava Beds (Lavoie et al., 2017), supporting an Actinbacterial origin of this lipid here. In these caves, 10,14-DiMe $C_{17:0}$ tends to co-occur with 10-Me $C_{17:0}$ (Online Supplemental Table 2-2), another biomarker for Actinobacteria (Willers et al., 2015). Though not diagnostic of this phylum, terminally-branched FAs such as *i*-$C_{16:0}$ are associated with Gram-positive strains in general (Willers et al., 2015) and are also abundant in biofilms, further supporting their production by Actinobacteria in these samples.

Soils: Lipid profiles of cave and surface soils are highly influenced by plants. These samples exhibit some of the highest yields and the highest proportions of saturated straight chain FAs and diacids (Figure 2-2). The FAs most associated with such soils are GL- and r-PL-derived *n*-$C_{16:0}$ and GL- and PL-derived *n*-$C_{18:0}$ (Figure 2-3). These common FAs are ubiquitous across bacteria and eukaryotes, limiting their chemotaxonomic interpretation. However, soils also contain the mid- and long chain FAs *n*-$C_{20:0}$, *n*-$C_{22:0}$, *n*-$C_{24:0}$, and *n*-$C_{26:0}$ (Supplemental Table S2), characteristic of epicuticular plant waxes (Harwood & Stumpf, 1971). The abundance of diacids

(Figure 2-2C), most commonly found in plant cutin and its breakdown products (Chefetz et al., 2002), further corroborate plant contributions.

Polyps, coralloids, and other features: Other sample types, such as oozes, polyps, coralloids, and mineral features, exhibit more variable IPL profiles (Figures 2-2 and 2-3). Oozes are associated with the PL-derived *n*-$C_{16:0}$ and $C_{18:1\text{-}\omega7}$ FAs (Figure 2-3), which are common, nonspecific bacterial biomarkers (Rontani et al., 2003; Willers et al., 2015). Polyps/coralloids and mineral features ordinate between biofilms and soils (Figure 2-3), appearing to lack distinguishable lipid profiles by group.

Intragroup variability of the polyps/coralloid and mineral features is not explained by the cave from which each sample originated (Appendix Figures 2-3 and 2-4). This is consistent with a previous 16S rRNA gene survey which concluded that a given cave did not significantly affect microbial community diversity or composition in biofilms (Lavoie et al., 2017). Our data suggest that the community composition of these other sample types is also highly controlled by site-specific variables, such as local geochemistry and mineralogy, that were outside the scope of this study. We note that polyps/coralloids and mineral features were found across a wide range of cave conditions (e.g., relative humidity, degree of airflow) that could contribute to differences in lipid compositions.

2.5.2. δ^{13}CIPL signatures differentiate lava cave biofilms from soils

The carbon isotopic composition of biomass records the C source(s) used by organisms modified by assimilatory biosynthetic fractionations. For heterotrophic assimilation of OC, this fractionation is minimal, enriching biomass relative to input carbon at maximum by ~1‰ (Hayes, 2001). By contrast, biological fixation of C can strongly discriminate against the incorporation of ^{13}C depending on the pathway. For instance, the Calvin Cycle used by plants produces biomass

that is 20-30‰ more depleted in ^{13}C than source CO_2. Conversely, the reductive acetyl-CoA (Wood-Ljungdahl) pathway, an alternative C metabolism used by acetogens and others, can produce extremely ^{13}C-depleted biomass, up to 52‰ lower than CO_2 (Hayes, 2001).

In sufficiently closed systems, OC respiration can add ^{13}C-depleted C to the local DIC reservoir. The subsequent fixation of C from this reservoir results in strongly depleted biomass that mimics the strong fractionations produced by the reductive acetyl-CoA pathway (Figure 2-6). To differentiate these potential mechanisms, we compare $\delta^{13}C_{IPL}$ to those of bulk C reservoirs, thereby linking C sources to individual communities based on the chemotaxonomy discussed above. For instance, if a given FA is produced by a heterotroph that assimilates DOC, that lipid will be isotopically similar to DOC, offset by ~3‰ on average due to lipid biosynthesis (Hayes, 2001). If the FA is instead produced by an autotroph, it will reflect the DIC pool modified by the fractionation arising from the C fixation pathway expressed. In other words, the lipids produced by autotrophs should be additively depleted in ^{13}C relative to both the DOC and DIC pools.

We expect plant- and soil-derived OC infiltrating from the surface to be the dominant source of DOC in cave drip waters (Saiz-Jimenez & Hermosin, 1999). Long chain FAs from plant waxes and diacids found in soils constrain this isotopic endmember to relatively ^{13}C-enriched values ($\delta^{13}C_{IPL}$ -28.3 ± 0.1 to -31.9 ± 0.3‰; Figure 2-5, Supplemental Online Supplemental Table 3-2), consistent with the expected range from C3 plant biomass (Hayes, 2001). The mean $\delta^{13}C_{DIC}$ value in drip fluids is -6.1 ± 1.9‰ (Figure 2-5), an expected value that reflects the abiotic dissolution of atmospheric $CO_{(g)}$ given a $CO_{(g)}$ $\delta^{13}C$ value of ~ -8‰ (Keeling et al., 2017). Drip fluids notably lack the ^{13}C-depleted DIC isotopic signatures that would imply OC remineralization

in overlying subsoils during transit to the cave environment. The mean value of $\delta^{13}C_{DOC}$ we measure is -29.8 ± 2.0‰, consistent with a plant-based origin.

Most cave features contain some IPL-derived FAs that reflect the isotopic composition of the DOC pool (Figure 4, 5). These lipids, notably the unsaturated FAs $C_{18:2}$ and $C_{18:1\omega7}$, are associated with fungi (Stahl & Klug, 1996) and heterotrophic bacteria (Rontani et al., 2003), respectively. The congruence between these $\delta^{13}C_{IPL}$ values and those of DOC suggest that heterotrophic microorganisms dependent on surface-derived OC produce these lipids.

By contrast, the significant ^{13}C depletions in the most abundant FAs from yellow, tan, and (to a lesser extent) white biofilms (Figures 2-4 and 2-5) cannot be explained by the incorporation of surface-derived OC. Instead, an additional fractionation step is required, most likely C fixation occurring within the caves. The bacterial FAs *i*-$C_{16:0}$, 10,14-DiMe $C_{17:0}$, *trans*-$C_{16:1\omega7}$, and *trans*-$C_{17:1\omega8}$ found primarily in tan and yellow biofilms, along with the more widespread *i*-$C_{18:0}$, exhibit the most negative $\delta^{13}C_{IPL}$ values in this dataset, ranging from -39.2 to -45.4‰ (Online Supplemental Table 2-4). This same trend is present, yet muted, in the $\delta^{13}C_{TOC}$ data, in which mean $\delta^{13}C_{TOC}$ values from biofilms are 6.6‰ more negative than soils (t-test, $p = 4.4\times10^{-8}$; Figure 5).

2.5.3. A model of carbon fixation within Lava Beds caves

Biofilm-derived lipids at Lava Beds are up to 39.3‰ ^{13}C-depleted relative to mean DIC (Figure 5), and we posit that this reflects a signal of *in situ* C fixation. If autotrophy were to occur in the overlying subsoils, the $\delta^{13}C_{DOC}$ values of drip water entering the caves should carry a depleted isotopic signature. If the subsoil were the site of most C fixation here, $\delta^{13}C_{DOC}$ would be expected to match the depleted $\delta^{13}C_{IPL}$ values from biofilms. The opposite is true here: $\delta^{13}C_{DOC}$ is ^{13}C-enriched relative to mean $\delta^{13}C_{IPL}$ in cave biofilms by an average of 10.2 ± 4.5‰ (Figure 2-4).

This requires *in situ* C fixation to occur within the biofilms themselves regardless of the autotrophic mechanism. Two possible mechanisms of generating these isotopic depletions include the reductive acetyl-CoA pathway and the Calvin cycle (Figure 2-6A-C).

The reductive acetyl-CoA pathway is most common in anaerobic acetogens, methanogens (Fuchs, 2011), and sulfate-reducing bacteria (Spormann & Thauer, 1988). Since methanogens are members of the Archaea, they do not produce membrane FAs in nearly the same quantities as Bacteria and are therefore not directly detectable with our methods. Nevertheless, it is possible they could be a source of ^{13}C-depleted biomass here by producing ^{13}C-depleted methane that is assimilated by bacterial methanotrophs or other OC assimilated by other bacteria whose FAs we measure. However, a previous 16S rRNA gene survey of Lava Beds cave biofilms reported low abundances (< 1%) of putative methane-cycling bacteria (Lavoie et al., 2017). Even if these few microbes were extremely active, they can only contribute ^{13}C-depleted isotopic signals proportional to their relative abundance. DOC produced by the reductive acetyl-CoA pathway would be expected to have a similar δ^{13}C value as those from IPLs (Figure 2-6B), which is not observed in biofilms (Figure 4). Although we acknowledge that the reductive acetyl-CoA pathway could still be an active autotrophic mechanism here, we focus our discussion below on the Calvin cycle due to our DOC data and previous taxonomic surveys of these caves (Lavoie et al., 2017).

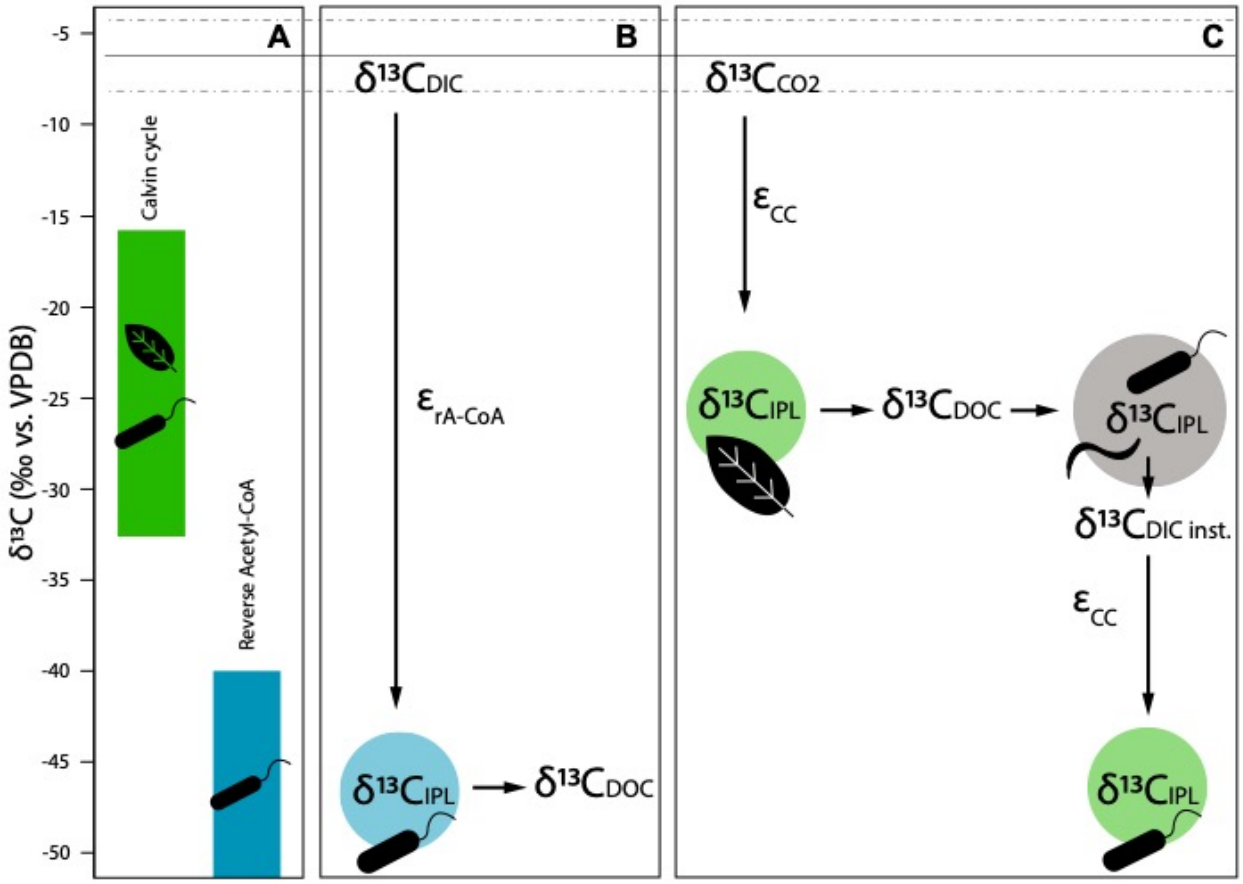


Figure 2-6: ***A conceptual model of carbon cycling at Lava Beds.*** Cartoons indicate which groups of organisms can express each metabolism (leaf = plants; bacillus = bacteria; worm = cave eukaryotes). 6A: Expected biomass $\delta^{13}C$ values from the fixation of inorganic C via two autotrophic metabolisms, the Calvin Cycle (CC) and the reductive Acetyl-CoA Pathway (rA-CoA). 6B: Expected biomass $\delta^{13}C_{IPL}$ and water drip $\delta^{13}C_{DOC}$ values if C is fixed via rA-CoA. 6C: Expected $\delta^{13}C$ values if C is fixed via "semi-closed" dynamics based on CC-dominated fixation. Estimated fractionations (ε) are from Hayes (2001). The mean $\delta^{13}C_{DIC}$ value measured from Lava Beds caves is represented with the horizontal grey bar, while the dotted lines represent one standard deviation.

Biofilms could generate the observed $\delta^{13}C$-depletions from remineralized DOC, invoking a symbiosis between heterotrophs and lithoautotrophs through an instantaneous DIC pool if the Calvin Cycle is the major autotrophic mechanism (Figure 2-6C). Carbon enters the caves in drip waters as either DOC or DIC in comparable abundance. Drip water DOC concentrations range

from 5.0 ± 0.1 to 23.3 ± 0.6 ppm and exhibit a mean $\delta^{13}C_{DOC}$ value of -29.8 ± 2.0‰ (Figure 2-4), whereas DIC concentrations range from 3.0 to 12.7 ppm with an observed mean $\delta^{13}C_{DIC}$ value of -6.0 ± 1.9‰. Heterotrophs assimilate and remineralize a portion of the inflowing DOC, excreting an isotopically similar instantaneous DIC pool (~ -30‰; Figure 2-6C). This DIC is then fixed by co-located lithoautotrophs, depleting lipids by another ~20‰ from the Calvin Cycle (Figure 2-6A,C). The abundance of ^{13}C-enriched FAs such as *n*-$C_{16:0}$ and $C_{18:1\omega7}$ co-occurring with depleted bacterial FAs (Figure 2-5) supports the co-occurrence of autotrophic and heterotrophic microbes (Figure 2-6C). Since IPL-derived FAs comprise ~10% of bulk TOC, this process also drives mean $\delta^{13}C_{TOC}$ towards more negative values in samples where such lipids are abundant (Figure 2-4). Lipid biosynthesis alone produces a 3‰ depletion in ^{13}C relative to TOC (Hayes, 2001), too small to explain the observed disparities. Some samples contain ^{13}C-enriched TOC relative to DOC (e.g., minerals and polyps/coralloids), which may reflect a highly divergent local DOC pool (which was unable to be sampled) or contributions from carbonate minerals embedded in the silica matrix which could not be removed prior to analysis.

Since most IPL-derived FAs are produced by multiple organisms, the $\delta^{13}C_{IPL}$ values we observe are likely the result of mixing distinct isotopic signatures from several unique populations exhibiting different C metabolisms. If we assume endmember values for lithoautotrophy and heterotrophy, a simple binary mixing isotope mass balance model (Equation 2-1) can be applied to estimate the relative fraction of lithoautotrophy necessary to produce individual IPL-derived lipids (Figure 2-7). In this model, *in situ* lithoautotrophy (*L*) and surface photosynthesis (*S*) are major C sources. Where δ_{IPL} is the measured $\delta^{13}C_{IPL}$ value, δ_L and δ_S are the endmember $\delta^{13}C$ values of biomass fixed via *L* and *S*, respectively, and f_L is the fraction of the observed δ_{IPL} signal attributable to *L*. For cave samples, δ_S is assumed to be equal to drip water $\delta^{13}C_{DOC}$ and δ_L is

equivalent to δ_S – 20‰ (fractionation of remineralized DOC by the Calvin Cycle, CC). For surface soils, δ_S is assumed to be equal to δ_{TOC}, the mean $\delta^{13}C_{TOC}$ value of surface soils. From this simple model, the degree of CC lithoautotrophic influence on $\delta^{13}C_{IPL}$ values (f_L) depends on C source isotopic compositions (Equation 2-1).

Equation 2-1

$$f_L = \frac{\delta_{IPL} - \delta_S}{\delta_L - \delta_S}$$

The application of this model to our observed $\delta^{13}C$ values demonstrates that f_L is significantly higher for bacterial IPL-derived FAs in cave biofilms compared to lipids from soils and other cave features (Figure 2-7). These FAs, specifically *i*-$C_{16:0}$, 10,14-DiMe-$C_{17:0}$, *trans*-$C_{16:1\text{-}\omega7}$, and *trans*-$C_{17:1\text{-}\omega8}$, require 67±6% (range 47-80%) of their C to be sourced from *in situ* CC-based lithoautotrophy to explain the observed $\delta^{13}C_{IPL}$ values (Figure 2-7; Online Supplemental Table 2-5). This contrasts with the plant wax-sourced *n*-$C_{20:0+}$ FAs and the more taxonomically ubiquitous *n*-$C_{18:0}$, which have significantly lower f_L values (Figure 2-7). These lipids require significantly less CC-sourced carbon, consistent with their photoautotrophic origin. If the reductive acetyl-CoA pathway is instead assumed to be the major autotrophic mechanism (where $\delta_L = \delta_S$ – 52‰, the largest fractionation reported for this pathway and therefore most conservative value for our model; Hayes 2001), these biofilm lipids require 26 ± 2% (range 18-30%) of C sourced from rA-CoA-based lithoautotrophy (Online Supplemental Table 2-6). This is still significantly more than *n*-$C_{20:0+}$ and *n*-$C_{18:0}$ (Online Supplemental Table 2-5).

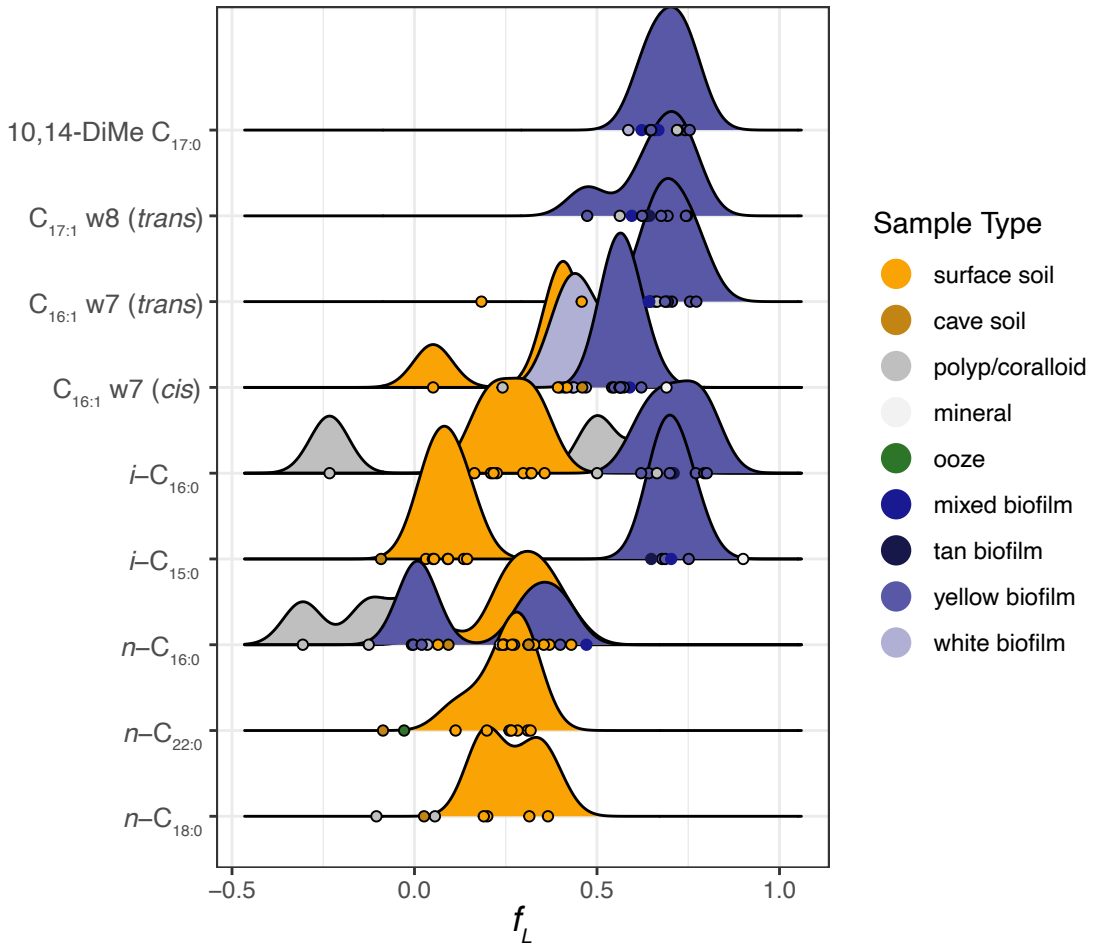


Figure 2-7: ***Fraction of Calvin Cycle-based lithoautotrophy required by isotopic mass balance.*** The term f_L is the fraction of the observed $\delta^{13}C_{IPL}$ signal attributable to Calvin cycle-based lithoautotrophy, calculated from Equation 2-1.

Some lipids are abundant across sample types but exhibit distinct $\delta^{13}C_{IPL}$ and resultant f_L values in specific types (Figures 2-5 and 2-7). For instance, *i*-$C_{15:0}$ is strongly ^{13}C-depleted in biofilms but is relatively ^{13}C-enriched in soils (Figure 2-5), producing mean f_L values of 0.62 ± 0.14 (n = 7) in biofilms and 0.11 ± 0.05 (n = 10) in soils (Figure 2-7). A similar variation in C source is also observed for *i*-$C_{16:0}$ and *cis*-$C_{16:1\text{-}\omega7}$ (Figure 2-7). As *i*-$C_{15:0}$ and *i*-$C_{16:0}$ are common

to Gram-positive bacteria (Kaneda, 1991; Willers et al., 2015), these isotopic patterns suggest different life strategies employed among this group of microorganisms in caves compared to those on the surface.

2.5.4 Potential source organisms for ^{13}C-depleted lipids at Lava Beds

While it is impossible to determine the precise bacterial sources of these lipids with our methods, we theorize likely source organisms based on literature on the microbial ecology of similar environments (Gonzalez-Pimentel et al., 2018; Lavoie et al., 2017; Northup et al., 2011). A 16S rRNA gene survey of Lava Beds biofilms identified microbes that are putatively capable of C fixation (Lavoie et al., 2017) including Actinobacteria (especially the order *Pseudonocardia*). Though many *Pseudonocardia* sp. are heterotrophs, there are lithoautotrophic strains capable of hydrogen (Grostern & Alvarez-Cohen, 2013) or ammonia oxidation (Z.-P. Liu et al., 2006). Other Actinobacteria, such as the uncultured T3 lineage, are capable of C fixation via nitrate-dependent iron oxidation (Kanaparthi et al., 2013). The predominant FAs produced by *Pseudonocardia* strains tend to be *i*-$C_{15:0}$ and *i*-$C_{16:0}$ (Y. Huang & Goodfellow, 2015), which are some of the most abundant and ^{13}C-depleted FAs we observe in biofilms (Figures 2-5 and 2-7).

The C isotopic compositions of the most diagnostic fatty acids particularly inform source organisms. For instance, 10,14-DiMe-$C_{17:0}$, a diagnostic biomarker for Actinobacteria (Willers et al., 2015), is one of the most prevalent membrane lipids in our biofilm samples (Figures 2-3 and 2-5) and is modeled to be the most influenced by lithoautotrophy on average (Figure 2-7). The sample types which contain abundant 10,14-DiMe-$C_{17:0}$ correspond to those with high relative abundance of Actinobacterial 16S rRNA gene sequences from Lavoie *et al.* (2017). Members of the Actinobacteria also dominate biofilms in other lava cave systems around the world (Gonzalez-Pimentel et al., 2018; Northup et al., 2011), with some being capable of fixing C via the Calvin

Cycle (Grostern & Alvarez-Cohen, 2013; Park et al., 2009). Based on our analysis, we propose that 10,14-DiMe-$C_{17:0}$ is a promising biomarker for Actinobacteria in lava cave environments.

Trans-unsaturated FAs are also abundant in Lava Beds biofilms (Figures 2-3 and 2-5), indicating the widespread presence of Gram-negative bacteria (Kaneda, 1991; Willers et al., 2015). Gammaproteobacteria, a diverse phylum of Gram-negative taxa, are also highly abundant according to DNA surveys from Lava Beds biofilms (Lavoie et al., 2017) and therefore could be a major source of these lipids. Members of the Gammaproteobacteria are capable of a variety of metabolic processes, including lithoautotrophy and N cycling. Regardless of the source organisms for the *trans*-unsaturated FAs, their depleted $\delta^{13}C_{IPL}$ values (Figure 2-4) suggest they fix C in these caves.

2.5.5. Potential sources of energy for lava cave lithoautotrophs

Based on lipid biomarker abundances and their isotopes, we conclude that lithoautotrophic bacteria are key members of cave biofilm communities at Lava Beds. This suggests that this terrestrial shallow subsurface environment supports lithoautotrophs despite significant DOC inputs. This begs the question of which catabolic energy sources fuel this C fixation. The basalt walls of lava caves themselves are replete with potential sources of energy from minerals, such as those bearing Fe(II), that lithoautotrophs could exploit. Additionally, drip fluids in Lava Beds caves carry soluble redox-sensitive species such as nitrate. Though total nitrogen levels are typically low in such cave water drips, nitrite and nitrate concentrations can reach 1 mg/L and 8 mg/L, respectively (Lavoie, 2017). Given these high concentrations, an abundance of taxa capable of N cycling (Lavoie, 2017), and the low $\delta^{13}C_{IPL}$ values we observe in biofilms from Lava Beds

relative to other C reservoirs (Figures 2-4 and 2-5), we posit that a source of lithoautotrophic energy is N-based.

2.5.6. Inferring the habitability of lava caves

We present organic geochemical evidence of an actively C-fixing biosphere occupying a basaltic system in the shallow subsurface. This work lends credence to the microbial habitability of similar environments elsewhere in the solar system. The surface regolith of Mars has been shown to carry nitrate salts (Shen et al., 2019; Stern et al., 2015), which, if transported to lava caves, could theoretically be an energy source for subsurface life. From a metabolic perspective, nitrate is both an attractive source of bioavailable N and an effective electron sink, particularly when coupled to reduced C or iron, thus providing a potential life strategy for putative lithoautotrophs living in Martian lava caves.

Our analysis indicates that lava caves host lithoautotrophs despite significant DOC fluxes from the surface. While this is seemingly paradoxical, we suggest that it may occur if lithoautotrophs are colonizing taxa in newly formed lava caves, prior to the establishment of surface vegetation. Once significant surface DOC input is established from developing surface biomes, thereby sustaining a larger population of heterotrophs, lithoautotrophs may maintain an ecological advantage by occupying an independent niche. Subsequent work focusing on community dynamics and colonization timescales of lava cave environments will be necessary to disentangle these interactions. Other work will be required to better understand and constrain the specific mechanisms and purveyors of C fixation in this system. For example, metagenomic- and/or transcriptomic-focused surveys could prove to be useful complementary studies by directly

tying taxonomy to the abundance and expression of C fixation-related genes by microbes in Lava Beds biofilms.

2.6 Conclusion

IPL-derived fatty acids from lava cave biofilms are structurally and isotopically divergent from those found in surface soils or speleothems. Key differentiating lipids include distinctive branched and *trans* unsaturated lipids, particularly those attributed to Actinobacteria (10,14-DiMe $C_{17:0}$), other Gram-positive bacteria (e.g., *i*-$C_{16:0}$), and Gram-negative bacteria (*trans*-$C_{16:1\text{-}\omega7}$, *trans*-$C_{17:1\text{-}\omega8}$). Large ^{13}C-isotope depletions observed in most biofilm lipids suggest widespread *in situ* chemolithoautotrophy, with lipid distributions implicating members of the Actinobacteria in C fixation. Our finding that lava caves host active C fixation has significant and positive implications for the search for life in extraterrestrial subsurface environments where surface biomes are likely absent.

2.7 Acknowledgments, Samples, and Data

This work was supported by Biologic and Resource Analog Investigations in Low-Light Environments (BRAILLE; NNH16ZDA001N), a project funded by the NASA Planetary Science and Technology Analog Research (PSTAR) program. Samples were collected under permit numbers LABE-2017-SCI-028, LABE-2018-SCI-0008, and LABE-2019-SCI-0012. We sincerely thank Randy Paynor, Katrina Smith, David Hayes, and Patricia Seiser for permitting and field support and all other park staff at Lava Beds National Monument for facilitating this study. MRO is a fellow in the CIFAR Earth 4D program. All data and R scripts written for data analysis and visualization are hosted at https://zenodo.org/record/5016282 (DOI: 10.5281/zenodo.5016282) and can be cited as:

"Matt Selensky. (2021, June 23). mselensky/Selensky2021_LavaBeds_Lithoautotrophy: Stable Carbon Isotope Depletions in Lipid Biomarkers Suggest Subsurface Carbon Fixation in Lava Caves (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.5016282".

The authors declare no conflict of interest pertaining to this work.

Chapter III: *Microbial Biogeography of the Eastern Yucatán Carbonate Aquifer*

3.1 Abstract

Constraining the spatial distribution of microorganisms and their ecological interactions is crucial for informing biogeochemistry. We explore horizontal and vertical patterns of microbial biogeography in the eastern Yucatán carbonate aquifer by examining the relative abundance of microbial taxa via 16S rRNA gene sequencing. As one of the largest anchialine groundwater systems on Earth, the density-stratified Yucatán aquifer consists of a freshwater lens overlying saline groundwater, with myriad sinkholes (cenotes) leading into a vast network of subsurface conduits. Several studies describe microbial communities within specific regions of the aquifer, yet fundamental questions remain regarding the ecology and distribution of biogeochemically relevant microbes. Our analysis demonstrates that the aquifer hosts a distinct microbiome from nearby seawater, with regionalism observed across both hydrological flow paths and vertical water column zones. We developed novel software (BNGAL) to construct taxonomic co-occurrence networks at different regional scales and categorize highly connected groups of taxa into potential niches. Our network analysis-based approach suggests that ubiquitous, metabolically flexible taxa such as the family *Comamonadaceae* act as ecological linchpins across several niches, often directly or indirectly co-occurring with taxa capable of anammox (e.g., *Gemmataceae*), methanotrophy (e.g., *Methyloparacoccus*), or organoheterotrophy. Further, communities from a deep, pit-like cenote open to the surface show the strongest niche partitioning between water column zones, differing from those encountered throughout the mostly dark and oligotrophic aquifer system. Our results suggest that the core microbiome could modulate different

biogeochemical regimes depending on location, acting as reservoirs of metabolic potential in disparate environments of this groundwater system.

3.2 Introduction

Lourens Baas-Becking famously stated, “Everything is everywhere, but, the environment selects,” (Baas Becking, 1934) to frame the concept of microbial biogeography, the study of the spatio-temporal distribution of microorganisms (Martiny et al., 2006). The advent of high-throughput DNA sequencing technologies has enabled microbial ecologists to directly survey microbial communities from a myriad of habitats to test this now-famous hypothesis, including but not limited to soils (Chu et al., 2020; S. Li et al., 2020), marine settings (Reintjes et al., 2019; Staley & Gosink, 1999; Villarino et al., 2022), and aquifer systems (Hug et al., 2015; Opalički Slabe et al., 2021). A major focus of the field of microbial ecology is the interplay between microbes and their immediate environments. Interface environments with steep geochemical gradients are particularly key areas in which these interactions can be observed and parsed.

Anchialine aquifers are those that contain subterranean connections to the ocean and are common features of coastal carbonate platforms. The anchialine carbonate aquifer permeating the Yucatán Peninsula is one of the largest in the world (Bauer-Gottwein et al., 2011) and is stratified with a meteoric freshwater lens floating on denser, marine-derived saline groundwater (Beddows et al., 2007). These distinct water masses display sub-stratifications in salinity (or conductivity), temperature, pH, redox potential, concentrations of most cations an anions, as well as dissolved oxygen, organic carbon, and inorganic carbon (Bauer-Gottwein et al., 2011; Beddows et al., 2007; Pohlman et al., 1997; Torres-Talamante et al., 2011). The depth of the halocline, the mixing zone between these water layers, increases with distance from the coast, increasing ~5-10 m below the

water table in Caribbean-adjacent sites to greater than 100 m in the middle of the peninsula (Bauer-Gottwein et al., 2011; Beddows, 2004a).

The Caribbean side of the Yucatán carbonate aquifer contains approximately 2,000 km of mapped submerged horizontal conduits (caves) that can connect to surficial processes through sinkholes (cenotes) formed primarily via cave roof collapse (Beddows, 2004b; QRSS, 2023). Most (~97%) of the groundwater volume in the Yucatán aquifer is held within the highly porous carbonate matrix, though the majority (>99%) of the flux occurs in the caves (Worthington et al., 2000), potentially allowing for horizontal dispersal of microbial communities across the aquifer. The extreme density stratification significantly limits vertical transfer of material, except in areas adjacent to rock debris (e.g., roof collapses), where turbulent mixing can transfer some saline groundwater into the base of the freshwater lens (Beddows, 2004a; Beddows et al., 2007; Stoessell, 1995). In addition to horizontal caves, vertically extensive submerged "pits" are occasionally found in the eastern Yucatán aquifer (Beddows et al., 2007; Kambesis & Coke, 2016), though they tend to be much more common in the northwestern regions of the peninsula around the Chicxulub impact crater (Bauer-Gottwein et al., 2011; Socki et al., 2002). These deep voids can be open or closed to the surface, but typically lack horizontal outflows at or below the halocline. This limits the lateral and horizontal transfer of material even further than is observed in horizontal conduits, isolating such pit environments from the rest of the aquifer (Beddows, 2004b; Smart et al., 2006).

These competing hydrological regimes position the eastern Yucatán carbonate aquifer as an ideal natural laboratory to explore biogeographic patterns of microbial community distribution in the shallow subsurface. However, due to the inherent difficulties of obtaining samples from submerged caves beyond the open water zone, previous surveys describing the biogeochemistry and microbial communities of the water column have tended to focus on those in the immediate

vicinity of the cenotes or deep pits (Brankovits et al., 2017; Escobar-Zepeda et al., 2021; L. Huang et al., 2021; Moore et al., 2020; Navarrete-Euan et al., 2021; Socki et al., 2002; Stoessell, 1992; Stoessell et al., 1993; Suárez-Moo et al., 2022; Torres-Talamante et al., 2011). Although such sites are more frequently sampled, the conduits represent the vast majority of the actively flowing aquifer habitat, connecting disparate regions of the groundwater system (Pohlman, 2011; Worthington et al., 2000).

Surface-derived organic matter entering open cenotes can accumulate in the halocline (Torres-Talamante et al., 2011) and can be detected in conduit sediments over 200 m downstream (van Hengstum et al., 2009). Its decomposition can create oxygen-poor conditions in the freshwater and halocline (Pohlman et al., 1997). By contrast, the less-studied conduits and other isolated portions of the aquifer are distinct, and can be comparably oligotrophic (Alcocer et al., 1998). The heterogeneity of the environments found within the Yucatán carbonate aquifer is reflected in the complex spatial distributions of the diverse microbiota it harbors (Escobar-Zepeda et al., 2021; L. Huang et al., 2021; Moore et al., 2020; Navarrete-Euan et al., 2021; Suárez-Moo et al., 2022). Small- and large-scale biogeographic patterns have been observed affected by factors such as water column zone, human influence (Moore et al., 2020; Navarrete-Euan et al., 2021), distance from the coast (Suárez-Moo et al., 2022), and proximity to intruding plant roots (Escobar-Zepeda et al., 2021). Not only do cenotes from the eastern and northwestern regions of the Yucatán Peninsula harbor distinct microbial communities, phylum-level differences were observed between freshwater, halocline, and saline groundwater, regardless of cenote (Moore et al., 2020). Intra-phylum microbial biogeography has also been observed throughout the Yucatán aquifer, with different groups of sulfur-cycling Campilobacterota predominating in different sites (L. Huang et al., 2021). Despite this foundational work, the distribution and abundance of these and other

biogeochemically relevant microbes throughout other environments within the eastern Yucatán carbonate aquifer remains an open question.

To understand the biogeochemical potential of the entire aquifer, we must first map the potentially uneven spatial distribution of taxa capable of diverse metabolic functions. Given the vast network of low-nutrient conduits that connect generally more eutrophic cenote entrances to the aquifer system (Pohlman, 2011), we hypothesize that a "core" microbiome exists throughout the aquifer. We posit that the local composition of such a core microbiome reflects site-specific environmental contexts, such as distance from the coast, fluid geochemistry, and position in the water column. We explore the biogeography of the Bacteria and Archaea that colonize the water column of the eastern Yucatán carbonate aquifer through analysis of the relative abundance of 16S rRNA taxonomic marker genes in duplicate from 66 sampling points representing 11 unique caves. To investigate the drivers of microbial distributions, we consider the environmental context of each community, including aquifer connectivity (pit vs. conduit), distance from the Caribbean coast, aqueous geochemistry, and position in the water column. We employ a correlational network analysis-based approach to explore the prevalence and abundance of key, biogeochemically relevant taxa throughout the chosen study sites. Our analysis emphasizes identifying members of a "core" microbiome common throughout these diverse aquifer habitats to inform future studies of the biogeochemical potential of the entire Yucatán carbonate aquifer and similar anchialine ecosystems.

3.3 Methods

Figure 3-1: ***Study sites.*** Bacterial and archaeal communities from 66 water samples spanning the freshwater, halocline, and saline groundwater layers in the aquifer were analyzed in duplicate and compared to Caribbean seawater. Communities were sampled from 11 aquifer and 3 surface seawater sites near Tulum, Quintana Roo, Mexico. Sites are colored by inferred hydrological 'region' according to previously mapped conduits (shaded lines within black inset; adapted from Kambesis & Coke, 2016). The meteoric freshwater typically flows towards the coast, though decoupled saline groundwater may alternatively flow coastward and inland based on sea level (Beddows et al., 2007). Refer to Table 3-1 in the main text for site descriptions and label IDs. Site Xel Ha (XeH) was sampled in two conduit branches, which we consider a single site due to their highly similar characteristics.

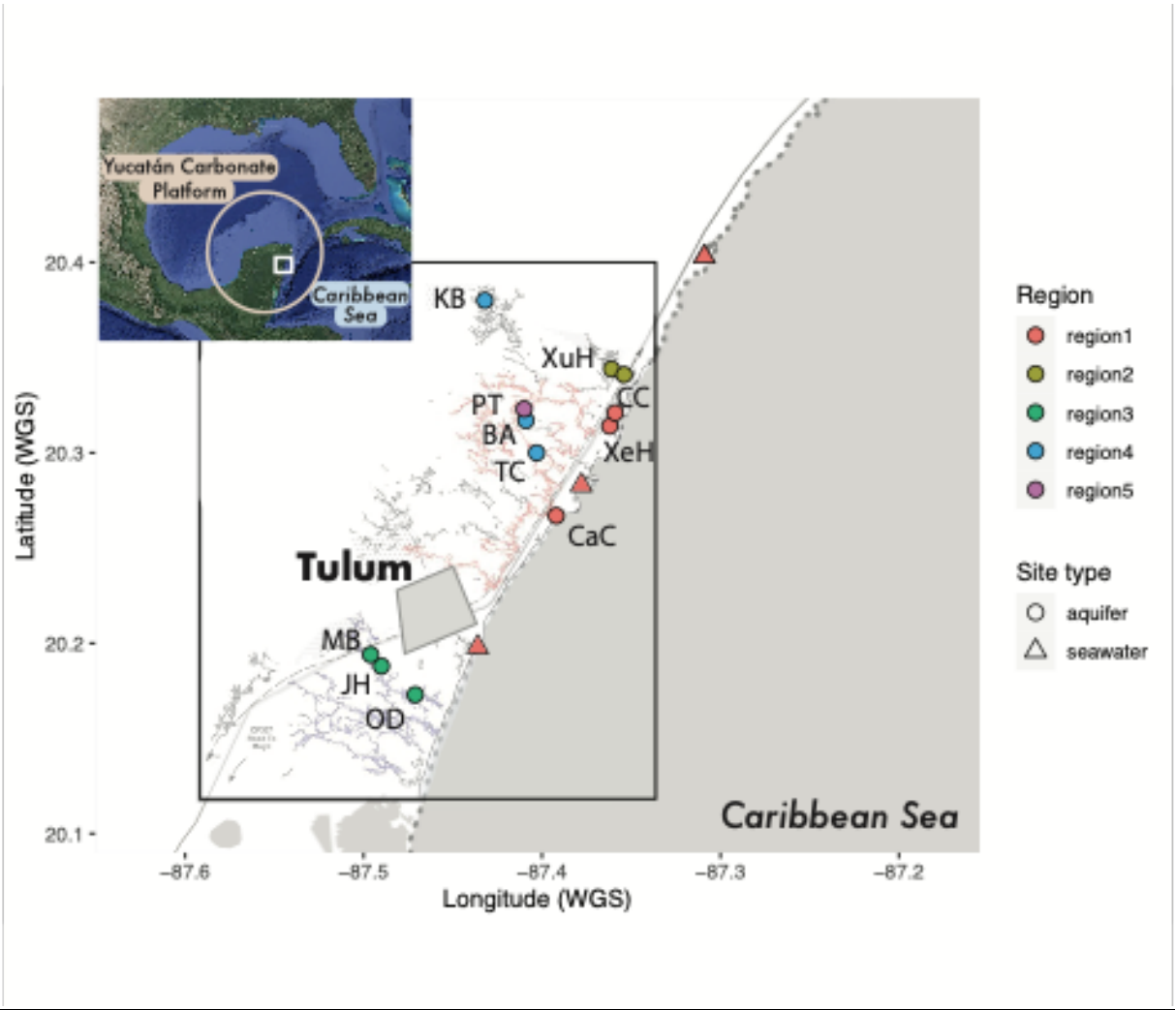


Table 3-1: Characteristics of sites sampled in the Eastern Yucatán carbonate aquifer.

Region	Label	Site Name	Sample Site Type	Entrance cenote distance inland (km)	Max. sampling depth (m)	Visible sulfide layer in halocline	Sequenced samples	Unique locations
1	SW	Seawater	Surface seawater	-	0.5	N	6	3
	CaC	Casa Cenote	Narrow, open coastal inlet	0.2	7.5	N	7	4
	XeH	Xel Ha	Wide, open coastal inlet	1	4.4	N	7	4
2	CC	Chan Chemuyil	Closed conduit	0.8	15.2	N	6	4
	XuH	Xunaan Ha	Closed conduit	1.5	14.4	N	7	5
3	JH	Jailhouse	Closed conduit	4.7	18.9	N	14	7
	MB	Maya Blue	Closed conduit	5.5	18.5	N	4	3
	OD	Odyssey	Closed conduit	2.2	15.5	N	8	6
4	BA	Blue Abyss	Closed pit	4.7	60	Y	10	8
	KB	K'oox Baal	Closed conduit	9.9	18.6	N	7	4
	TC	Tikim Chi	Closed conduit	3.1	19.2	N	10	6
5	PT	The Pit	Open pit	5.7	60	Y	20	12

3.3.1. Field sites and sampling

In August 2019, a team of experienced cave divers led by P. Beddows obtained water samples (n = 66) from anchialine aquifer sites (n = 11) near the Caribbean coast in Quintana Roo, Mexico (Figure 3-1; Table 3-1). Endmembers and control samples include 3 surface seawater samples from ~10 m offshore and waist depth, a field control, and a purified drinking water sample. Samples were collected in ethanol-rinsed 1L glass autoclavable bottles. Divers first partially filled the sampling bottles with cenote water to reduce buoyancy, then once at a given sampling depth, they opened, inverted, and completely voided the bottle with compressed air using a nozzle whip. Divers then swam upstream, inverted and filled the bottle with undisturbed upstream groundwater to minimize contamination. Sampling depth was recorded from wrist-mounted dive computers. The field control was collected during the dive to site JH (Figure 3-1; Table 3-1) by descending with an ethanol-rinsed sampling bottle filled with purified drinking water, which remained unopened until filtering at the field laboratory. To ultimately identify potentially cross-contaminating DNA during transit in the field, a corresponding sample of purified drinking water that did not leave the laboratory was also filtered for comparison (discussed further in section

3.3.5). When feasible, the water column was characterized by multiparameter probes to measure depth, water temperature, pH, and conductivity. Divers descended at a rate of 2 cm/s to allow for thermal equilibration of the probe. Samples were placed in a cooler on ice until transport to the field laboratory for filtering the same day.

3.3.2. Water filtering

Prior to filtering and further analysis, the conductivity of an aliquot of unfiltered water was compared to field-based conductivity measurements to assess whether collection bottles were properly voided of surface water. Subsequently, 1L of each checked water sample was then filtered using a 0.2 μm Sterivex filter. Tubing was thoroughly pre-flushed with unfiltered sample to prevent cross-contamination. Filtered water for aqueous geochemical analysis (see Section 3.3.3) was stored without headspace in the dark at 4˚C until analysis. Filters were temporarily stored at -20˚C in the field and at -80˚C at Northwestern University prior to DNA extraction.

3.3.3. Geochemical analysis

Total alkalinity was determined by Gran titration with standardized 0.02 N HCl on the 0.2 μm filtered samples within 1-2 days of collection. Daily titrations of a ~3000 μeq/L Na_2CO_3 standard ensured measurement uncertainty remained below 5%. The pH probes were calibrated daily using pH 7, 4, and 10 solutions. Concentrations of chloride, sulfur, and sulfate were measured on the ion chromatograph at QBIC on a Thermo Scientific Dionex ICS-5000+ and using a Dionex IonPac AS22 column. The analysis was run using an eluent of 4.5 mM sodium carbonate and 1.4 mM sodium bicarbonate and a Dionex AERS 500 Carbonate 4 mm Electrolytically Regenerated Suppressor.

3.3.4. DNA extraction

DNA was extracted from filters using a previously described protocol (Colman et al., 2016). Thawed filters were first split as duplicates for downstream analyses. Split filters were then shredded into fine pieces using sterile, DNA-free scissors. To lyse cells, shredded filters were vortexed at high speed for 10 minutes in tubes containing sterile alumina beads and 2 mg mL^{-1} lysozyme buffer solution (25 mM Tris HCl, pH 8.0 and 2.5 mM EDTA, pH 8.0 in DNA-free deionized water). To remove protein contaminates, the lysed cells were incubated at 55°C with 20 mg mL^{-1} proteinase k in TESC buffer for 25 minutes. Samples were then incubated on ice for 15 minutes and were pelleted by centrifugation at 14,000 RCF for 10 minutes. A phenol:chloroform:isoamyl alcohol solution (25:24:1) was then thoroughly mixed with the supernatant. After centrifugation at 14,000 RCF for 10 minutes, the aqueous supernatant was then transferred to a sterile low-bind microcentrifuge tube. The supernatant was treated with ice-cold isopropyl alcohol (100%) and incubated at room temperature for 5 minutes to precipitate DNA, which was subsequently pelleted by centrifugation. The supernatant was then discarded, and the pellet was washed twice with ice-cold molecular-grade ethanol (70%) in between rounds of centrifugation. The DNA pellet was allowed to completely dry until it was resuspended in DNA-free deionized water and stored at -80°C until sequencing.

3.3.5. DNA sequencing and quality control

Frozen DNA extracts were sent to the Environmental Sample Preparation and Sequencing Facility at Argonne National Laboratory (Lamont, IL). “Universal” bacterial and archaeal primers (515F/806R) were used to amplify the hypervariable V4 region of the 16S rRNA gene (Caporaso et al., 2012). Libraries were sequenced via Illumina MiSeq. The resultant 6,874,763 paired-end reads were imported into the QIIME2 (version 2020.6) software environment for processing

(Bolyen et al., 2019). After demultiplexing using the *demux* command, a total of 6,260,931 reads passed the denoising and chimera check steps and were assigned into amplicon sequence variants (ASVs) with the DADA2 algorithm (Callahan et al., 2016). Taxonomies were assigned to representative sequences using a pretrained Silva (v.138) classifier in QIIME2 (Bokulich et al., 2018; Quast et al., 2012; Robeson et al., 2020). In addition to singleton taxa, contaminating ASVs were removed from the subsequent ASV-level taxonomic abundance table (Lee et al., 2015). Briefly, the abundance of each ASV was normalized by the number of environmental samples. ASVs found in the negative controls from each DNA extraction batch were then identified in their corresponding samples. Such an ASV was considered a contaminant if its normalized abundance in each batch of samples was less than one order of magnitude than the normalized abundance in its corresponding negative control (Lee et al., 2015). A total of 54 contaminating ASVs, mainly Gammaproteobacteria, were identified and removed from the dataset with this method. Following the removal of contaminants, we observed a total of 4,183 unique ASVs in our dataset. The number of quality-controlled reads per sample ranged from 6,139 to 148,590 (Online Supplemental Table 3-1). Based on the inflection point of an alpha rarefaction plot, we rarefied the ASV table to a sampling depth of 9,957 for downstream statistical analyses, with a total of 106 samples included in the dataset.

3.3.6. Data analysis

Beta diversity was estimated at the ASV level by calculating Bray-Curtis dissimilarity on a matrix representing the relative abundance of ASV-level taxa (Online Supplemental Table 3-2) and visualized in R by ordination via non-metric multidimensional scaling (NMDS) using the vegan package (Oksanen et al., 2019) and custom scripts. We estimated alpha diversity for each sample via the Shannon index (Shannon, 1948) using a custom R script.

We developed "Biological Network Analysis and Learning" ("BNGAL"; https://github.com/mselensky/bngal), a custom R package and associated command-line tool to pre-process data and construct pairwise associative networks of ASVs with relevant metadata parameters. Networks are composed of "nodes" connected to each other by "edges", with graph theory examining the underlying substructures that may provide useful biological and/or ecological insight (Steele et al., 2011). There are many methods that can be used to construct networks from biological data (Matchado et al., 2021), each with their own strengths and limitations. We constructed a network model of that computes a correlation coefficient between each possible pair of ASVs, with the identity of a node corresponding to a single ASV (or environmental variable) and edges corresponding to the strength of its Spearman correlation coefficient with another node. Further description of the construction of the network model is provided below.

In the BNGAL pipeline, rarefied ASV counts of non-singleton taxa are first rescaled with environmental values chosen based on data density (water depth, distance from the coast, conductivity, and concentrations of sulfate, chloride, total sulfur, and alkalinity). To avoid the inclusion of spurious correlations in the network, we impose an "observational threshold" (n) such that $n \geq 5$ for every pairwise relationship. A Spearman correlation matrix is then calculated, from which all possible network nodes and edges are identified. Edges are filtered to only include pairwise relationships with absolute Spearman correlation coefficients greater than 0.6 and p-values less than 0.05. BNGAL employs the 'igraph' software package (Csardi & Nepusz, 2006) to build the network and classifies ASVs into subnetworks based on edge betweenness clustering. "Edge betweenness" is the number of shortest possible paths that pass through a given edge (Girvan & Newman, 2002). "Edge betweenness clusters" (EBCs) are computed by iteratively

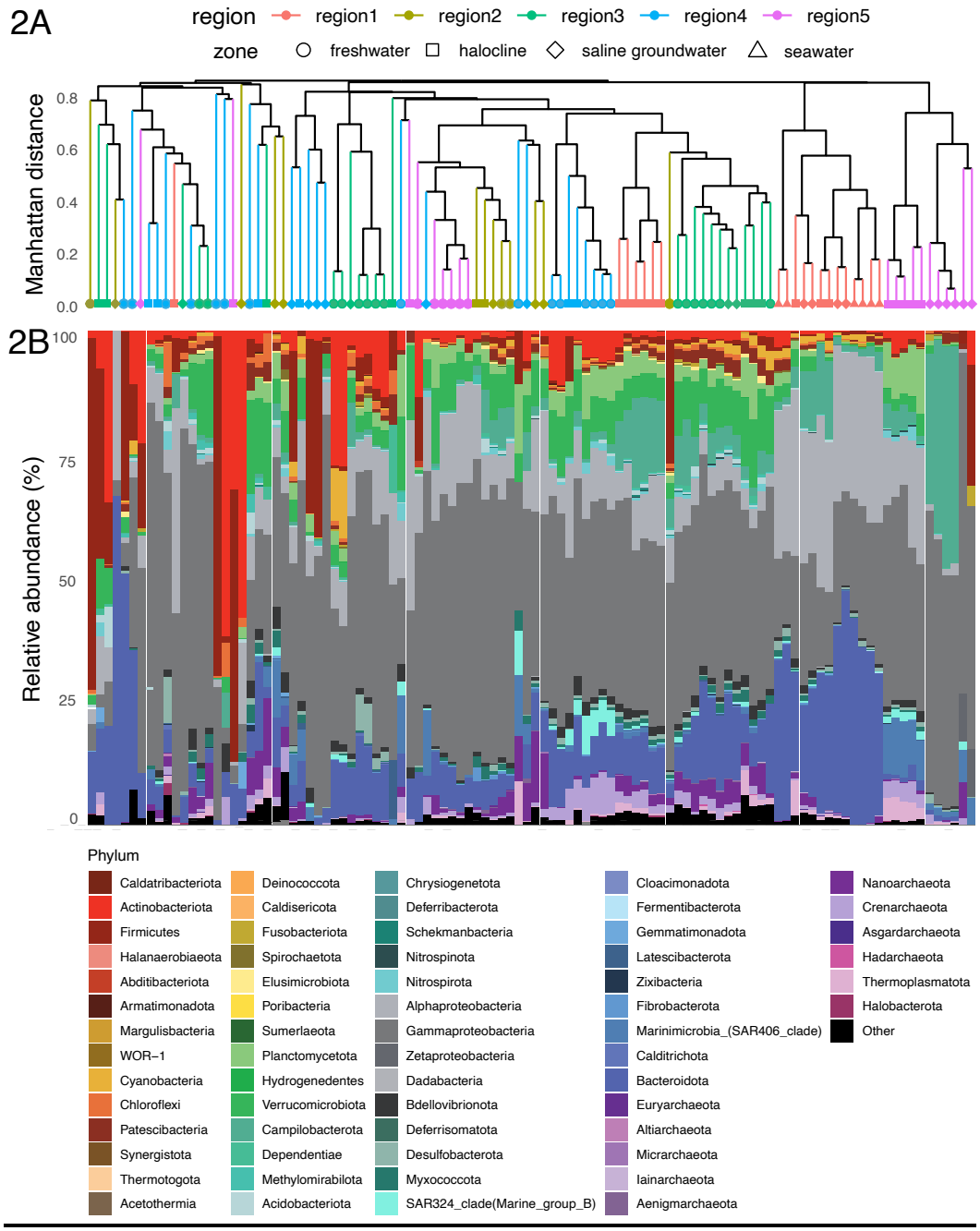
removing edges with the highest such betweenness values until the network is divided into several densely intra-connected subnetwork clusters, revealing the internal structure of the entire network (Girvan & Newman, 2002). Networks are statically visualized with ‘igraph’, while interactive plots are generated with the ‘igraph’-based javascript library ‘visNetwork’.

3.3.7. Data availability statement

Demultiplexed sequence data for all samples are freely available from the Sequence Read Archive (SRA) at the National Center for Biotechnology Information (NCBI). Data are referenced under BioSample accession numbers SAMN33725555-SAMN33725702 (BioProject PRJNA943420).

3.4 Results and Discussion

Figure 3-2: ***Microbial community composition of the eastern Yucatán carbonate aquifer.*** 2A: "Complete" hierarchical cluster from a Manhattan distance matrix generated from an ASV-level taxonomic relative abundance table. Dendrogram ends are colored by inferred hydrological region (Figure 1) and are shaped by water column zone. 2B: Bacterial and archaeal community compositions, filled by phylum.



We survey the microbial communities present in 11 sites in the eastern Yucatán carbonate aquifer by 16S rRNA gene sequencing and compared them to three nearby surface seawater sites (Figure 3-1, Table 3-1). We grouped caves based on known and inferred flow paths, into "regions" along the Caribbean coastline (Table 3-1). Due to the anchialine nature of the aquifer, cave samples are further classified using the measured conductivity into their water column zone of freshwater, halocline, or saline groundwater (Online Supplemental Table 3-3). We define the halocline as portions of the water column where conductivity increases 40-50 mS/cm over a depth of 10m or less, thereby separating groundwater with near-normal salinity from the overlying meteoric freshwater lens. Especially at sites within 1 km from the coast, the "freshwater" can be brackish (5-8 mS/cm; Appendix Figure 3-1). We classify Xel Ha and Casa Cenote as "halocline" communities due to their overall brackish characteristics. From the rarefied ASV table, we identify 4,183 unique ASVs representing 82 total phyla, 202 classes, 538 orders, 917 families, and 1,953 genera. Of these, 145 ASVs were identified as Archaea from 12 unique phyla (Online Supplemental Table 3-2).

3.4.1. Microbial community compositions throughout the eastern Yucatán carbonate aquifer

Across the dataset, Proteobacteria (especially the classes Alpha- and Gammaproteobacteria) is the dominant phylum, representing 49.5% of sequences (Online Supplemental Table 3-4). Other abundant phyla include Bacteroidota (11.4%), Verrucomicrobiota (5.6%), Firmicutes (5.4%), Actinobacteriota (4.7%), Campilobacterota (4.7%), Planctomycetota (2.9%), and Nanoarchaeota (2.1%). All other individual phyla represent less than 2% of the total observed reads each (Online Supplemental Table 3-4). These numbers vary considerably by sample type and location. Verrucomicrobiota compose a notably higher proportion of the aquifer

microbiome regardless of region or water column zone (Online Supplemental Table 3-4). By contrast, coastal seawater and heavily marine-influenced aquifer communities ("Region 1") are enriched in sequences of Alphaproteobacteria and Bacteroidota relative to other sites (Online Supplemental Table 3-4). We also observe variable distributions of taxa within aquifer sites. Notably, halocline and saline groundwater communities from The Pit ("Region 5") harbor significantly more of the SAR406 clade (Marinimicrobia) compared to the rest of the dataset (Online Supplemental Table 3-4).

To further explore how the microbial taxonomic landscape differs across sampling sites, we grouped communities at the ASV level via hierarchical clustering and then examined how those from different geographic regions and water column zones cluster based on $\log_{10}$-transformed relative compositions (Figure 3-2). First-order branching divides microbial communities into two clusters, one whose shallower branches are highly dissimilar and a much larger one that exhibits clear regional patterns (Figure 3-2). Most notably, marine-influenced communities from sites Casa Cenote and Xel Ha ("Region 1"; Table 3-1) cluster closely with surface seawater and are distinct from saline groundwater (Figure 3-2). Communities from site Xel Ha, a shallow and very wide open anchialine inlet on the Caribbean coast (Figure 3-1), clusters much more closely with surface seawater communities than others in the aquifer (Figure 3-2). Alongside surface seawater, communities from site Xel Ha are marked by particularly high abundances of ASVs mapped to the phyla Bacteroidota and Alphaproteobacteria compared to the rest of the aquifer (Figure 3-2; marine mean 2.96×10^{-3} vs. aquifer mean of 8.54×10^{-3}; $p = 1.12 \times 10^{-7}$).

Intriguingly, none of the communities from Casa Cenote cluster with other Region 1 communities despite their near-marine conductivity values (Appendix Figure 3-1, Online Supplemental Table 3-3). Instead, they group primarily with halocline and freshwater communities

from non-marine regions (Figure 3-2). This pattern aligns with fundamental differences in the hydrology and geometry of the marine-influenced sites Casa Cenote and Xel Ha, which both circulate large volumes (~10^7m/year) of seawater (Beddows, 2004a; Beddows et al., 2007). However, because site Casa Cenote is partially roofed and comparatively narrower, the tidal wedge of marine water is only ~200m from the coast there, compared to 1-1.5km distance in the wide, open site of Xel Ha (Beddows, 2004a; Beddows et al., 2007). Nevertheless, the relatively high compositions of the SAR_324 clade (Marine Group B) and Crenarchaeota, phyla commonly encountered in marine ecosystems (Karner et al., 2001; T. D. Wright et al., 1997), distinguish Casa Cenote communities from others in the aquifer (Figure 3-2), likely reflecting the influence of some intruding seawater.

By contrast, freshwater communities tend to harbor relatively high proportions of Verrucomicrobiota, Nanoarchaeota, Planctomycetota, and Gammaproteobacteria (Figure 3-2; Online Supplemental Table 3-4). Freshwater communities in particular tend to be dominated by a single unclassified Gammaproteobacteria mapped to the family *Comamonadaceae*, which reaches 61.0% of the community at its most abundant (Online Supplemental Table 3-2).

3.4.2. Diversity

We further investigated the alpha and beta diversities of communities in the eastern Yucatán carbonate aquifer and compared them to nearby seawater (Figure 3-3). Non-metric multidimensional scaling (NMDS) analysis of a presence-absence matrix at the ASV level (adapted from Online Supplemental Table 3-2) demonstrates that groundwater communities with the highest Shannon diversity in our dataset cluster away from seawater and marine-influenced communities (Figure 3-3A). This suggests that these highly diverse groundwater microbial communities, primarily from Regions 2, 3, and 4, are distinct from those found in seawater or the

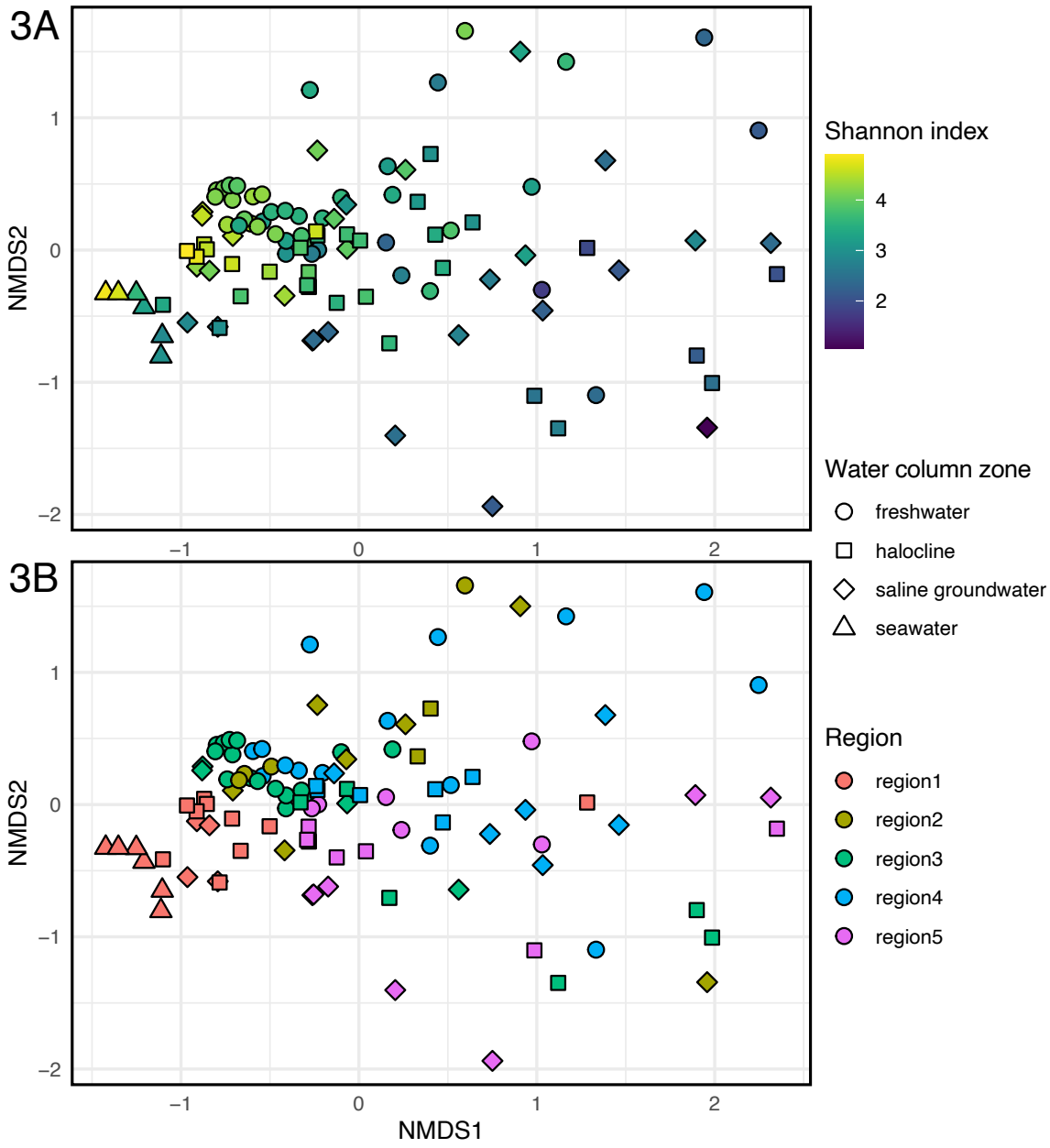
seawater-adjacent sites in Region 1, agreeing with relative abundance-based hierarchical clustering results (Figure 3-2). Furthermore, despite some variability, halocline and saline groundwater microbial communities from Region 5, solely comprised of site The Pit, cluster away from other regions in NMDS space (Figure 3-3B), suggesting that this site hosts a distinct microbiome.

Regions 2, 3, and 4 encompass sites in closed conduits sampled away from cenote openings. Since caves from these three regions are geographically well-distributed on the Caribbean coast (Figure 3-1), we interpret the groupings observed in hierarchical clustering and NMDS results (Figures 3-2 and 3-3) as evidence of the presence of a diverse core microbiome in the eastern Yucatán carbonate aquifer. By contrast, Region 5 is represented solely by The Pit, which is a 60m-deep pit cenote open to the surface that contains a visible layer of hydrogen sulfide at the halocline (Table 3-1; Appendix Figure 3-1). Though The Pit receives shallow flow of freshwater from caves from Region 4, the distribution of microbial taxa in the halocline and saline groundwater here is distinct (Figure 3-2B), with a marked increase in the abundance of sulfur-cycling microbes such as the SUP05 cluster of the Gammaproteobacteria (Glaubitz et al., 2013) as well as the families *Sulfurovaceae* and *Sulfurospirillaceae* of the Campilobacterota (formerly classified under the Epsilonproteobacterota; Waite et al., 2017). As such, we distinguish The Pit from adjacent caves in the flow path of Region 4. Notably, communities from another deep sulfidic pit in Region 4, Blue Abyss (Figure 3-1; Table 3-1), do not share these characteristics with The Pit. The most obvious environmental difference between The Pit and Blue Abyss is that the latter is closed to the surface, preventing surficial organic matter and sunlight from directly reaching the water column.

Microbial community composition in the eastern Yucatán carbonate aquifer is influenced by both water column zone and inferred hydrological region (Figures 3-2 and 3-3). Although

aquifer communities tend to cluster by region, water column zone also appears to drive some minor clustering within a given region (Figure 3-2). Further, a number of key taxa are shared among groundwater sites regardless of these variables. To identify the taxa that comprise this putative "core microbiome" throughout the aquifer, we created a co-occurrence network model and examined the abundances of major co-occurring groups throughout the aquifer at different regional scales.

Figure 3-3: ***Patterns of diversity and regionalism displayed by microbial communities in the Eastern Yucatán carbonate aquifer.*** Non-metric multidimensional scaling (NMDS) analysis on a binary (presence/absence) matrix of bacterial and archaeal taxa binned at the ASV level (Online Supplemental Table 3-2; solution stress = 0.189). 3A: Communities with higher Shannon index values, indicating higher alpha diversity, tend to be more akin to each other than those with lower values. 3B: Freshwater aquifer communities tend to cluster regardless of inferred hydrological region (Figure 3-1) while those from the halocline and saline groundwater exhibit more variability. Communities from Regions 1 and 5 appear to be the most distinct overall.



3.4.3. A global co-occurrence network model of microbial niche space in the eastern Yucatán carbonate aquifer

We employ network analysis to model complex interactions within microbial communities, which has previously revealed ecologically relevant associations between groups of taxa (Dohlman & Shen, 2019; Faust et al., 2012; Jiang et al., 2019; Layeghifard et al., 2017; Steele et al., 2011). Our network theory-based approach applied to the entire dataset demonstrates the presence of several interconnected subnetworks of co-occurring taxa (Figure 4A). The prevalence, abundance, and diversity of these subnetworks, termed "EBCs" after the algorithm used to identify them, can strongly vary by both geographic region and water column zone (Figure 3-4B). Under our network model parameters, we observed 1,002 nodes (997 ASVs and 5 environmental variables) and 16,046 edges, representing 24% of unique ASVs (Online Supplemental Table 3-2). We discuss the distribution and potential ecological significance of network node taxa from key EBCs in the following subsections.

3.4.3.1 Ubiquitous, universally abundant subnetworks in the eastern Yucatán carbonate aquifer

We interpret EBCs with high prevalence as those that harbor members of the "core microbiome" in the eastern Yucatán carbonate aquifer. Present in 98% of samples, EBC3 is the subnetwork with the highest observed prevalence (Online Supplemental Table 3-5). The mean relative abundance of ASVs mapped to EBC3 is 23.0% across all groundwater samples and 0.86% in surface seawater samples (Figure 4B; $p < 2.2\text{x}10^{-16}$ t-test; Online Supplemental Table 3-5). As such, we interpret EBC3 to represent a component of the core microbiome in the aquifer. EBC3 is comprised of 40 ASVs, primarily represented by uncultured members of the Nanoarchaeota (o. Woesearchaeales), Verrucomicrobiota (g. *Candidatus* Omnitrophicus), Campilobacterota (g.

Sulfurimonas), and novel bacterial and archaeal ASVs only classifiable to the domain level (Online Supplemental Table 3-6). Two ASV-level groupings of the phylum Methylomirabilota are also members of EBC3 (Online Supplemental Table 3-6), though they are noted to only be present in low abundance (< ~1%) in the freshwater or upper halocline layers of most aquifer communities.

Uncultivated taxa compose some of the most abundant and ubiquitous taxa in the eastern Yucatán carbonate aquifer. For instance, an unclassified *Comamonadaceae* bin from EBC3 is both the most prevalent (found in 91% of all samples) and most abundant bin in the Yucatán carbonate aquifer by median relative abundance (7.7%; Online Supplemental Table 3-2). We examine median abundance when considering ubiquitous taxa to avoid the undue influence of high- or low-abundance outliers. The *Comamonadaceae* are a highly metabolically flexible family of Gammaproteobacteria commonly found in aquatic environments, with cultured representatives capable of aerobic heterotrophy, iron reduction, hydrogen oxidation, and denitrification (Willems, 2014). Despite its abundance, this *Comamonadaceae* bin only correlates with an uncultured marine taxon mapped to the NS9 marine group of Bacteroidota (Figure 3-4A).

EBC2 is similarly prevalent as EBC3, with its taxa being present in 87% of communities, though it is noticeably less abundant in communities from Regions 1 and 5 (marine-influenced and The Pit, respectively; Figure 3-4). EBC2 is a diverse subnetwork, harboring 49 ASVs from 13 distinct phyla (Online Supplemental Table 3-6) and tends to be more abundant in freshwater compared to other water column zones (Figure 3-4B). A node classified to the hgcI clade of the Actinobacteriota family *Sporichthyaceae* exhibits the highest degree in EBC2, co-occurring with 32 taxa from several subclusters (Figure 3-4A). Another taxon mapped to the same clade (hgcI_clade;uncultured_bacterium) co-occurs with 26 other nodes, although the only external EBC connection it makes with another subcluster is with unclassified *Methyloparacoccus* (Figure 3-

4A). The high number of both internal and external EBC connections exhibited by this clade of Actinobacteriota suggests that it is associated with multiple niches throughout freshwater portions of the aquifer (Girvan & Newman, 2002). Two Planctomycetota ASVs, classified as uncultured *Gemmataceae* and *Phycisphaeraceae* (CL500-3), are the most prevalent and abundant ASVs from EBC2.

Taxa from the subnetwork EBC2 often co-occur with those from EBC15, a diverse and highly interconnected group of 118 taxa of mostly low abundance (< 1%), present in 64% of communities (Online Supplemental Table 3-6). Within EBC15, unclassified *Methyloparacoccus* (f. *Methylococcaceae*) is the taxonomic bin with highest prevalence (n = 24) with a median relative abundance of 0.075%, directly co-occurring with several taxa from EBC2, including the two nodes representing members of the hgcI clade of *Sporichthyaceae* (Figure 3-4A). EBC15 is the most abundant in the freshwater of Odyssey cave (Figure 3-4B) and contains 118 taxa, represented by taxa capable of cycling C1 compounds such as *Methylococcaceae*, *Methylomonadaceae*, *Methylophilaceae*, and hydrogen oxidizers such as other *Hydrogenophilaceae* strains. Communities with an abundance of EBC2 taxa tend to also host others that comprise EBC4, which is present in 84% of samples (Online Supplemental Table 3-5). The most prevalent taxonomic bin from EBC4, unclassified *Rhodobacteraceae* (Online Supplemental Table 3-6), only co-occurs with unclassified *Sphingomonadaceae* in the same subnetwork ($\rho = 0.68$; Figure 3-4A). The only other ASV in EBC4, an ASV identified as *Candidatus planktoluna* (f. *Microbacteriaceae*), connects EBC4 to EBC2 through *Nitrosarchaeum*, *Sediminibacterium*, and unclassified *Gemmataceae* ($\rho =$ 0.65, 0.62, and 0.67, respectively).

Although the five taxa comprising EBC7 tend to be encountered in low (≤ 1%) abundance, this sparsely interconnected subcluster is present in 73% of communities. Such taxa include an

unclassified NS9 marine group taxon, *Candidatus Peribacteria*, and Campilobacterota (including two unclassified *Sulfurimonas* nodes and an uncultured *Arcobacteraceae*; Online Supplemental Table 3-2). Of these taxa, the unclassified NS9 marine group and uncultured *Arcobacteraceae* are the most prevalent and abundant (Online Supplemental Table 3-6). This subcluster is particularly abundant in Casa Cenote and Xel Ha, marine-influenced sites within Region 1 (Figure 3-4B).

3.4.3.2 Regionally abundant subnetworks elucidate microbial biogeography

Some groups of "core microbiome" taxa exhibit high prevalence, being present in >50% of samples, but are only abundant in specific communities (Figure 3-4B), suggesting that they proliferate only under appropriate environmental conditions. For example, taxa from EBC21 and EBC34 are present in 70% and 63% of communities respectively (Online Supplemental Table 3-6) but are by far the most abundant in halocline and saline groundwater communities from the deep and open site The Pit ("Region 5"). EBC21 contains 49 co-occurring taxa from 14 phyla, such as four bins from the SAR406 clade, Planctomycetota (especially OM190 and *Phycisphaeraceae*), and Gammaproteobacteria such as an uncultured bacterium bin of the SUP05 cluster (f. *Thioglobaceae*). By contrast, EBC34 is less diverse, comprising 9 nodes, of which five are Campilobacterota, represented by the families *Arcobacteraceae* (unclassified), *Sulfurimonadaceae* (g. *Thiovulum*), *Sulfurospirillaceae* (g. *Sulfurospirillum*), and *Sulfurovaceae* (g. *Sulfurovum*). Three other nodes in this subcluster are classified as Gammaproteobacteria, which represent uncultured *Ectothiorhodospiraceae*, the WHC3-3 group of *Nitrinocolaceae*, and *Thioglobaceae* (an unclassified SUP05 cluster bin; Online Supplemental Table 3-2). Notably, EBC21 acts to "bridge" the subclusters EBC34 and EBC11, which are most abundant in the saline groundwater of Region 5 (The Pit) and Region 1 (specifically Xel Ha and seawater), respectively (Figure 3-4). With 16 significant co-occurrences, the unclassified *Sulfurovum* bin is a particularly

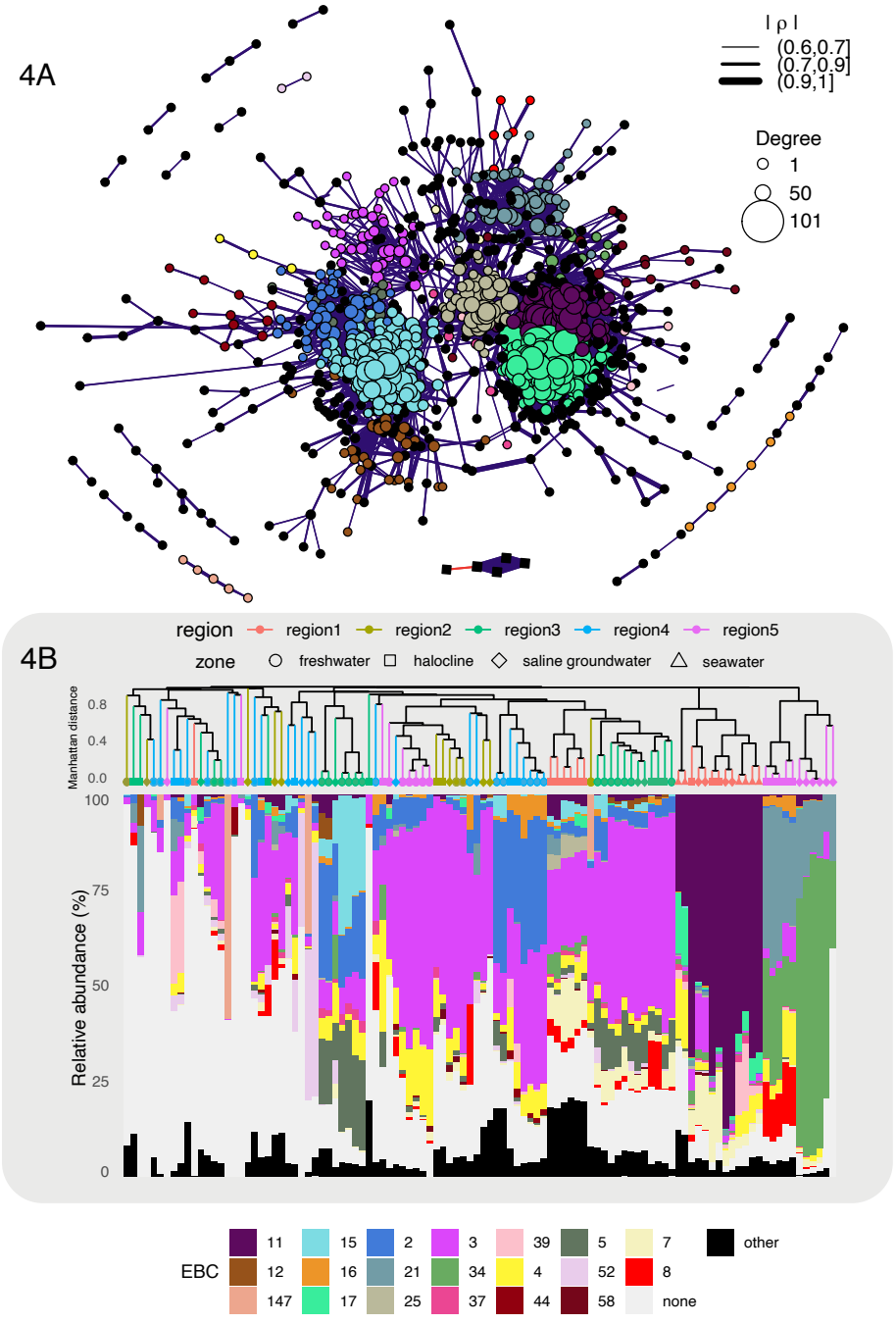
central node in EBC21, exhibiting especially strong correlations with two members of the SAR406_clade from EBC34 ($\rho = 0.66$ and 0.63), an unclassified *Spongiibacter* ($\rho = 0.69$), and a *Marinobacterium* metagenome from EBC11 ($\rho = 0.73$; Figure 3-4, Online Supplemental Table 3-2). Notably, two ASVs from the SUP05 cluster, a genus of sulfur-oxidizing Gammaproteobacteria (Glaubitz et al., 2013), are highly abundant in halocline and saline groundwater communities from The Pit. These ASVs are prevalent throughout the aquifer and seawater samples, but only reach above 5% of the community in The Pit, slightly below the halocline at the zone of sulfide accumulation (Appendix Figure 3-1). One node classified to the SUP05 cluster is a part of EBC21, while another node from the same group is a member of EBC34 (Online Supplemental Table 3-6).

Taxa from EBC5 are similarly prevalent (in 64% of communities) but are instead most abundant in the freshwater of Region 3 (Figure 3-4B). This subcluster contains 11 microbial nodes, with the most prevalent and abundant taxonomic bin representing unclassified *Hydrogenophilaceae*, a family mainly consisting of chemolithoautotrophs that are capable of various forms of sulfur oxidation as well as hydrogen oxidation (Orlygsson & Kristjansson, 2014). This node connects EBC5 to the much more ubiquitous subcluster EBC2 via the unclassified hgcI clade node of *Sporichthyaceae* ($\rho = 0.63$, Figure 3-4A). EBC5 taxa also tend to positively correlate with members of EBC15, which is represented by taxa capable of cycling C1 compounds such as *Methylococcaceae*, *Methylomonadaceae*, *Methylophilaceae*, and hydrogen oxidizers including other *Hydrogenophilaceae* strains (Figure 3-4A; Online Supplemental Table 3-2).

Most co-occurring ASVs from seawater communities or marine-influenced coastal aquifer sites (Region 1) are neither abundant nor present in other parts of the aquifer (Figure 3-4B, Online Supplemental Table 3-2). This trend is illustrated by the distribution of taxa from EBC11, a large

subnetwork composing 81 ASVs that are mainly found in marine environments (Online Supplemental Table 3-6). Marine Alphaproteobacteria such as the HIMB11 group of *Rhodobacteraceae* (Pujalte et al., 2014) or the SAR116 clade (Mullins et al., 1995), Bacteroidota such as *Cryomorphaceae* (Bowman, 2014) or the photoheterotrophic NS5 marine group of *Flavobacteraceae* (Priest et al., 2022), and Gammaproteobacteria such as *Litoricolaceae* (Webb et al., 2014) only exceed 3% relative abundance in marine samples and are either absent or < 0.5% in more inland aquifer communities, including the saline groundwater (Online Supplemental Table 3-2). This implies that saline groundwater communities harbor a microbiome that is distinct from those in nearby coastal marine settings.

Figure 3-4: ***Global co-occurrence network of bacterial and archaeal ASVs from the eastern Yucatán carbonate aquifer.*** 4A: Network colored by EBCs. Nodes are sized by degree (total number of co-occurrences) while the width of each edge corresponds to the strength of the Spearman correlation coefficient (ρ). Refer to Online Supplemental Figure 3-1 for an interactive version of this figure to probe individual relationships. 4B: Relative abundance of EBCs. Samples are ordered via ASV-level hierarchical clustering as in Figure 3-2A. (EBC=edge betweenness cluster)



3.4.3.3 Regional networks reveal site-specific co-occurrence patterns

While regional biogeography is present among co-occurring groups of microbes in the global network (Figure 3-4B), environmental variables (depth, conductivity, and concentrations of sulfate, total sulfur, chloride, and alkalinity; Figure 3-4A, squares) do not strongly correlate with any microbial nodes (Figure 3-4A, circles). This is likely because significant variation in these environmental values between regions obscures potential pairwise trends (Appendix Figure 3-1). Distance from the coast did not correlate with any other nodes and is thus excluded from the global network (Figure 3-4A). To probe more granular individual microbe-microbe and microbe-environmental variable co-occurrence relationships, we computed separate networks for each hydrological region (Appendix Figure 3-2). Descriptions of each regional network and subsequent subcluster abundance are presented below.

The marine-influenced communities in Region 1 strongly cluster by site, with surface seawater and Xel Ha communities grouping together, apart from those within Casa Cenote (Appendix Figure 3-3A). Here, microbial community composition (Appendix Figure 3-3A) and individual co-occurrence patterns (Appendix Figure 3-2A) are broadly consistent with variation in hydrological regimes. The wide, open Xel Ha channel allows for recirculating seawater to reach 1 – 1.5 km inland, while the bedrock topography in Casa Cenote prevents the tidal seawater wedge from penetrating more than 200 m inland, resulting in the aqueous physicochemical characteristics of Casa Cenote to resemble the aquifer more closely (Beddows, 2004a). Unlike other sites, mangroves surround Casa Cenote, which could also affect community composition in that site. Notably, the subcluster EBC1_R1 is abundant in the coastal aquifer sites (especially Casa Cenote) but is not in surface seawater (Appendix Figure 3-3A). This subcluster contains many taxa prevalent throughout other areas of the aquifer, such as the uncultured *Comamonadaceae*,

Campilobacterota such as *Arcobacteraceae* and *Sulfurimonas*, Omnitrophales, and others (Online Supplemental Table 3-7). Alkalinity and water column depth have strong positive correlations with many members of EBC1_R1, consistent with the subcluster's highest abundance in Casa Cenote and Xel Ha, which were sampled somewhat more deeply than the surface seawater samples (Table 3-1). As terminal outflows of the freshwater layer of the aquifer into the Caribbean, Casa Cenote and Xel Ha exhibit higher alkalinity values than surface seawater, presumably due to the dissolution of carbonates from the rock matrix of the Yucatán platform (Stoessell et al., 1989).

Most aquifer communities exhibit more variable clustering patterns. Communities in Region 2, represented by the near-coast caves Chan Chemuyil and Xunaan Ha (Figure 3-1), show some weak depth-dependent clustering, though they also tend to group by cave (Appendix Figure 3-3B). The uncultured *Comamonadaceae* ASV only co-occurs with a highly central *Candidatus omnitrophicus* node in this region (Appendix Figure 3-2B). By contrast, Region 3 comprises the sites Jailhouse, Maya Blue, and Odyssey (Figure 3-1) and displays clustering primarily by site, though water column zone also drives some clustering (Appendix Figure 3-3C). Communities from Odyssey are well-dispersed throughout the hierarchical cluster, though they tend to be most similar to the surface-most freshwater of Jailhouse. Most Jailhouse communities contain high abundances (>39%) of EBC5_R3, which is dominated by the unclassified *Comamonadaceae* ASV (Online Supplemental Table 3-7), a highly central node in this network (Appendix Figure 3-2C). Here, this ASV shows a very strong positive correlation ($\rho > 0.8$) with six other ASVs, namely a *Hydrogenophaga* sp. (also within the *Comamonadaceae*), a *Sulfurimonas* sp., three mapped to the NS9 marine group, HdN1 of the *Halomonadaceae*, an uncultured *Cryomorphaceae*, and an *Arcobacteraceae* (Campilobacterota). Communities from the sites Blue Abyss, K'oox Baal, and Tikim Chi (Region 4; Figure 3-1), cluster by cave and water column zone (Appendix Figure 3-

3D). The unclassified *Comamonadaceae* bin is part of EBC2_R4, which is most abundant in freshwater communities from Region 4 (Appendix Figure 3-3D) and contains the unclassified *Comamonadaceae*, a taxon mapped to the OM190 group of the Planctomycetota, and two unclassified Woesearchaeles nodes (Online Supplemental Table 3-7). Notably, the unclassified *Comamonadaceae* co-occurs with an unclassified *Sediminibacterium* bin (f. *Chitinophagaceae*; ρ = 0.65), which is a highly central node in the EBC1_R4 subnetwork.

Compared to other aquifer communities, the Region 5 co-occurrence network structure and EBC abundances demonstrate evidence for strong niche partitioning by water column zone that is clearly distinct from other regions (Appendix Figures 3-1 and 3-2E). Region 5 is solely comprised of The Pit (Figure 3-1) and microbial communities here are clearly distinguished by water column zone, with those in the freshwater lens clustering separately from halocline and saline groundwater communities (Appendix Figure 3-3E). Here, the unclassified *Comamonadaceae* bin is a member of EBC13_R5 and is most abundant in the freshwater lens, connecting 25 other nodes from multiple subclusters. It most strongly positively correlates with unclassified *Sediminibacterium* (ρ = 0.87) and the SH3-11 group of Verrucomicrobiota (ρ = 0.87). Total alkalinity positively correlates with the unclassified *Comamonadaceae* bin (ρ = 0.78; Appendix Figure 3-2E). All other environmental variables included in the network (total sulfur, sulfate, chloride, conductivity, and water column depth) negatively correlate with the unclassified *Comamonadaceae* node, reflecting its freshwater prevalence. Below the freshwater, an unclassified bin related to the SUP05 cluster of S-cycling Gammaproteobacteria (Glaubitz et al., 2013) is the sole member of EBC20_R5 and negatively correlates with the unclassified *Comamonadaceae* bin (ρ = -0.62), consistent with the highest abundance of this taxon in the halocline and saline groundwater of The Pit (Figure 3-5). This node exhibits relatively high degree (27) and positively correlates with other high-degree,

single-node subclusters such as EBC26_R5 (another SUP05 cluster bin, ρ = 0.84), EBC38_R5 (uncultured *Sulfurospirillum*, ρ = 0.70), and EBC7_R5 (unclassified *Sulfurovum*, ρ = 0.76). The abundance of EBC26_R5 is roughly even in the halocline and saline groundwater, while EBC38_R5 and EBC7_R5 are most abundant in the saline groundwater. Consistent with these spatial distributions, EBC20_R5 negatively correlates with alkalinity (ρ = -0.62) and positively correlates with conductivity, chloride, and total sulfur (ρ = 0.66, 0.64, 0.63, respectively).

3.4.4. Microbiota in the Eastern Yucatán Carbonate Aquifer inhabit distinct regional niches

Our analysis of ASV-level co-occurrence patterns across this dataset provides clear evidence for the uneven distribution of microbial communities (Figure 3-2) and co-occurrences (Figure 3-4B) throughout the aquifer. Co-occurrence patterns of key ASVs differ significantly between the global and regional networks (Figure 3-4B, Appendix Figure 3-2), suggesting that members of the "core microbiome" are inherently flexible in their co-occurrence partners. This could be either due to the same taxon being involved in different biogeochemical processes across the aquifer, or different partners filling the same niche space in different environments.

For instance, members of a metabolically flexible family of Gammaproteobacteria, *Comamonadaceae*, are some of the most ubiquitous and abundant taxa in the Yucatán carbonate aquifer (Figure 3-2B), though which taxa they co-occur with varies by hydrological region. A single unclassified member of this family dominates freshwater communities and is a member of EBC3 in the global network (Figure 3-4). Here, this ASV only co-occurs with an unclassified member of the NS9 marine group of Bacteroidota. This lack of network edge connections implies that this node does not represent a "keystone species" (Girvan & Newman, 2002) in the aquifer despite its abundance and ubiquity. However, our regional-scale network analysis demonstrates

instead that this node co-occurs with several different groups of microbes and exhibits distinct network node properties depending on the hydrological region (Appendix Figure 3-2; Appendix Figure 3-4).

Members of the *Comamonadaceae* have been previously recognized as abundant "keystone species" in the surface-most portions of a schist aquifer (Ben Maamar et al., 2015). In that setting, *Comamonadaceae* ASVs positively correlated with nitrate concentrations, though their role in heterotrophic denitrification was determined to be unlikely due to limited organic carbon in the groundwater system (Ben Maamar et al., 2015). Genomic evidence suggests that many members of the *Comamonadaceae* are also capable of thiosulfate oxidation (Deja-Sikora et al., 2019). In the Yucatán carbonate aquifer, ASVs from this family tend to be most abundant in freshwater and positively correlate with taxa thought to be capable of anammox, hydrogen oxidation, and methylotrophy, though the specific taxa involved varies by region (Appendix Figure 3-2; Appendix Figure 3-4).

To illustrate, the unclassified *Comamonadaceae* bin positively correlates with taxa capable of anammox (namely Planctomycetota such as *Phycisphaeraceae, Gemmataceae*, and the OM190 clade) in every regional network other than Region 2 (Appendix Figure 3-4). A taxon highly related to a metagenome representing the NS9 marine group of Flavobacteriales also co-occurs with unclassified *Comamonadaceae* in Regions 1, 3, and 5, which is also the only significant relationship displayed by the unclassified *Comamonadaceae* bin in the global network (Figure 3-4B, Appendix Figure 3-4). In the same regional networks, the unclassified *Comamonadaceae* bin also correlates with *Hydrogenophaga*, a genus of hydrogen oxidizers in the same family. Interestingly, the *Comamonadaceae* bin only correlates with methylotrophic bacteria (*Methylophilaceae*) in Regions 1 and 5 (Appendix Figure 3-4). In Region 2, the unclassified

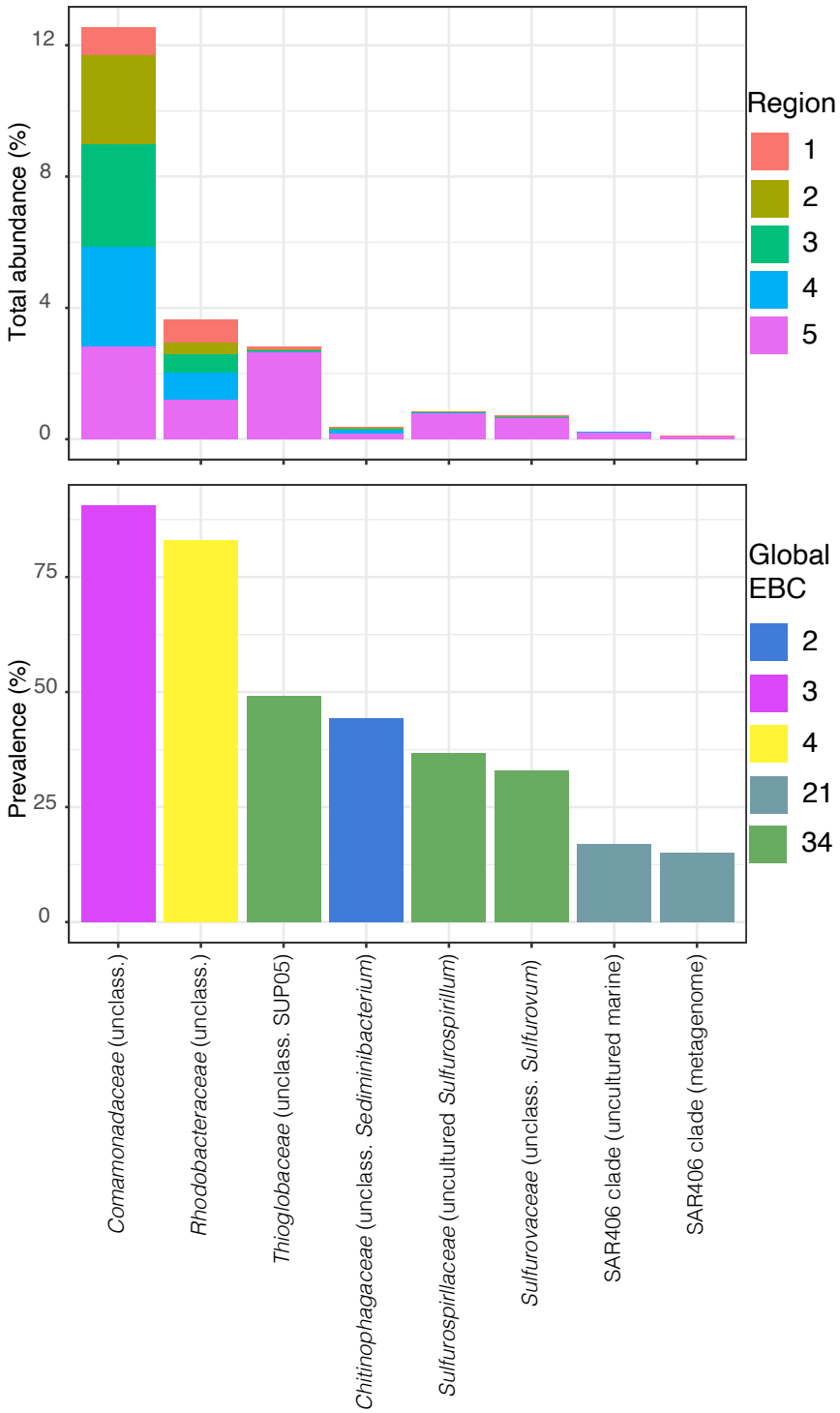
Comamonadaceae bin only correlates with a taxon classified to *Candidatus omnitrophus* (Appendix Figure 3-4), an uncultured lineage that is common in other subsurface environments whose members are capable of various sulfur and nitrogen redox metabolisms (Momper et al., 2017; Perez-Molphe-Montoya et al., 2022).

Based on its varied set of co-occurrences with taxa capable of several different functions (Appendix Figure 3-4), we speculate that this unclassified *Comamonadaceae* bin acts as a reservoir of metabolic potential in the Yucatan carbonate aquifer. Nitrate reduction to nitrite is a commonly reported metabolism of cultured representatives of the *Comamonadaceae* (Willems, 2014). We therefore speculate that members of this abundant family could supply nitrite for anammox metabolisms fueled by ammonia produced via organic matter remineralization in the water column, especially in sub-oxic regions or microenvironments. Intriguingly, heterotrophic members of the order Flavobacterales have been previously implicated in the creation of sub-oxic microenvironments from organic matter degradation in an exposed ice sheet containing sulfurous deposits (K. E. Wright et al., 2013). Considering its co-occurrence patterns with the unclassified *Comamonadaceae* bin (Appendix Figure 3-4), members of the NS9 marine group could fill a similar ecological role in areas of the Yucatán aquifer containing sufficient organic matter.

The unclassified *Comamonadaceae* bin also co-occurs with an unclassified *Sediminibacterium* bin in networks from Regions 1, 4, and 5 (Appendix Figure 3-2; Appendix Figure 3-4). Although methane metabolism has not been observed in *Sediminibacterium* spp. to date, members of this organoheterotrophic genus of Bacteroidota have been previously observed to associate with methane-oxidizing communities in engineered biofilms and are suggested to utilize organic substrates produced from methane oxidation (Limbri et al., 2014; van der Ha et al., 2013). In our study sites, the unclassified *Sediminibacterium* bin co-occurs with the

methylotrophic genus *Methylotenera* in Regions 2 and 4 (Appendix Figure 3-2). Unlike the much more ubiquitous *Comamonadaceae* bin, the unclassified *Sediminibacterium* is absent in seawater and all saline groundwater communities except for one in Blue Abyss (Online Supplemental Table 3-2), suggesting a distinctly terrestrial origin.

Figure 3-5: ***Prevalence and abundance of key global network taxa.*** The prevalence of each taxon is filled by EBC membership defined in Figure 4. The total relative abundance of each node is binned by the regions defined in Figure 1. EBC = edge betweenness cluster.



Where surface-derived organic carbon accumulates at the halocline, oxygen is rapidly consumed, and methane is produced by microbial methanogenesis (Brankovits et al., 2017, 2018). This methane represents a major source of both energy and carbon in such areas. Isotopic mass balance from a previous survey estimates that 21% of the average cave shrimp diet in nearby Cenote Bang (1.8 km upstream of Maya Blue) is ultimately sourced from methanotrophic bacteria, even for those sampled in the caves downstream from the open Bang sinkhole (Brankovits et al., 2017). In sub-oxic microenvironments in the aquifer, methylotrophs and methanotrophs could produce hydrogen (Jo et al., 2020; Nandi & Sengupta, 1998) available for oxidation by *Comamonadaceae* spp. or others.

In the global network presented in this study, two nodes classified to the hgcI clade of Actinobacteriota co-occur with methylotrophic and methanotrophic bacteria in EBC2 (Figure 3-4A), which likely represents a distinct niche from EBC3, the subcluster containing the unclassified *Comamonadaceae* (Online Supplemental Table 3-6). Notably, both Actinobacteria link their subcluster, which contains two prevalent bins mapped to an uncultured *Gemmataceae* and *Phycisphaeraceae* (CL500-3) ASVs, to EBC15 through an unclassified bin mapped to the genus *Methyloparacoccus* (Figure 4A). *Methyloparacoccus* is a group of methanotrophic Gammaproteobacteria (Hoefman et al., 2014) while *Phycisphaeraceae* is associated with anammox in other settings (Rios-Del Toro et al., 2018). As such, the hgcI clade of Actinobacteriota may serve to bridge methane- and ammonia-cycling communities in the Yucatán carbonate aquifer.

Our results suggest that open pit cenotes with direct surface organic matter and sunlit haloclines host distinct communities with fundamentally different co-occurrence patterns than closed conduits and closed pits (e.g., Blue Abyss) in the Yucatán carbonate aquifer (Figures 3-4

and 3-5; Appendix Figure 3-2). We note that members of the SAR406 clade (Marinimicrobia), a group putatively capable of sulfur and nitrite oxidation (Thrash et al., 2017; J. J. Wright et al., 2014), are most abundant in the halocline of The Pit and are either absent or in low abundance elsewhere (Figure 3-5). In the global network (Figure 3-4A), several nodes belonging to the SAR406 clade co-occur with other putative sulfur oxidizers such as an unclassified bin of the SUP05 cluster, *Sulfurovum* (Giovannelli et al., 2016; Glaubitz et al., 2013; Mori et al., 2018) as well as an uncultured *Sulfurospirillum*, a genus whose cultured members exhibit widespread capacity for elemental sulfur reduction and hydrogen oxidation (Kruse et al., 2018; Schumacher et al., 1992; Waite et al., 2017). One cultured isolate of the SUP05 cluster reduces nitrate to nitrite coupled to thiosulfate oxidation (Shah et al., 2017), while *Sulfurovum* likely reduces nitrate coupled to sulfide oxidation in another setting in the Yucatán carbonate aquifer (L. Huang et al., 2021). These nodes are members of EBC21 and EBC34, which are most abundant in portions of The Pit at or below the halocline (Figure 3-4A) where sulfur begins to accumulate (Appendix Figure 3-1).

The unclassified SUP05 cluster bin is found in 49% of all communities (Figure 5; Table S7), including nearly all halocline and saline groundwater samples from every system except Maya Blue cave. Intriguingly, this pattern is not as clear for its Campilobacterota counterparts, as the *Sulfurovum* and *Sulfurospirillum* bins are less prevalent (33 and 37%, respectively; Figure 3-5; Online Supplemental Table 3-6). *Sulfurovum* is commonly encountered in highly sulfidic cave habitats (Macalady et al., 2008; Porter & Engel, 2008). Campilobacterota (specifically *Sulfurimonas* and *Sulfurovum*) have been previously observed to dominate Cenote Siete Bocas, a cenote ~15 km from the Caribbean coast that contains abundant surface organic matter, with sunlight penetrating the water column through several small (< 5 m) openings (L. Huang et al.,

2021). In that site, geochemical and genetic evidence suggests *Sulfurovum* likely oxidizes sulfide via nitrate reduction (Giovannelli et al., 2016; Han & Perner, 2015; L. Huang et al., 2021; Mori et al., 2018). In our global network, the unclassified *Sulfurovum* from EBC34 positively correlates with an ASV closely related to an uncultured *Marinobacterium* metagenome ($\rho = 0.73$) and an unclassified *Spongiibacter* bin ($\rho = 0.70$) from EBC11, thereby connecting the subclusters EBC34 and EBC11, which are most abundant in the saline groundwater of The Pit and marine-influenced sites, respectively (Figure 3-4A). Notably, some cultured *Marinobacterium* spp. are capable of nitrate reduction as well (Huo et al., 2009), suggesting that communities in The Pit may follow a similar ecological paradigm.

Intriguingly, communities from Blue Abyss (Region 4) and The Pit (Region 5) are distinct (Figure 3-2) even though they are both vertically extensive pit cenotes that contain accumulated sulfide and surface-derived organic matter, can exhibit oxygen-deficient haloclines and some saline groundwater, and lie along the same inferred flow path (Figure 3-1; Table 3-1). The major differences between these sites are the lack of both sunlight and the direct input of surficial organic matter entering Blue Abyss, which is completely closed to the surface (Table 3-1). Despite the lack of a direct connection to the surface, some cave divers report the presence of bats in an air bell above the water of Blue Abyss. Bat guano could provide a distinct source of organic matter that could affect microbial community composition in Blue Abyss, though the quantity of guano entering the aquifer in this site is unconstrained. Nevertheless, the surface-most freshwater community in Blue Abyss (6.2 m depth) is unusually dominated by Actinobacteriota, especially the genus *Kocuria* (Online Supplemental Table 3-2), where the relative abundance of this phylum reaches 62% (Online Supplemental Table 3-4). Though Actinobacteriota are commonly found in soils, they are also major components of the bat guano microbiome (De Leon et al., 2018; Selvin

et al., 2019). Bat guano input could be an unaccounted lever on the microbial community composition of Blue Abyss that requires further investigation.

Most putative sulfur-cycling microbes in The Pit, namely the SUP05 cluster, *Sulfurovum*, and *Sulfurospirillum*, are either absent or are in low abundance in Blue Abyss (Online Supplemental Table 3-2). Instead, sub-halocline communities in Blue Abyss tend to be dominated by various members of the phylum Firmicutes, as well as *Acinetobacter venetianus* and *Pseudomonas pachastrellae* (Online Supplemental Table 3-2). The latter two members of the Gammaproteobacteria order Pseudomonadales are both commonly found in marine settings and are capable of aerobic degradation of "recalcitrant" organic matter such as *n*-alkanes or other lipids (Di Cello et al., 1997; Fondi et al., 2016; Romanenko et al., 2005). In the global network, these are the sole members of EBC52, which does not correlate with any nodes in the larger network (Figure 3-4A). As such, we speculate that *A. venetianus* and *P. pachastrellae* live within a restricted niche (Girvan & Newman, 2002) reliant on the consumption of recalcitrant organic matter. The hydrology of the Yucatán aquifer is marked by active saline groundwater circulation, with the shallowest saline water shuttling back-and-forth. Although inflows from warm Caribbean surface water thermally equilibrate approximately 10 km inland (Beddows et al., 2007), deeper saline groundwater flows are significant and likely continuous inland (Beddows, 2004a). We posit that the residence time of the saline water is theoretically long enough that the degradation of any organic matter trapped at the density interface could lead to progressive anoxia with distance inland, especially in inland sites with less overall horizontal saline flow such as Blue Abyss (Beddows et al., 2007; Kambesis & Coke, 2016). Additionally, since >99% of groundwater flow in the Yucatán aquifer occurs in the conduits (Worthington et al., 2000), we speculate that these hydrological factors contribute to a lack of widely available oxidants such as nitrate in Blue Abyss

to fuel the oxidation of sulfur compounds by populations of *Sulfurovum* and/or the SUP05 cluster (L. Huang et al., 2021; Shah et al., 2017). This would correspond to the abundance of mostly anaerobic, fermentative Firmicutes in the saline groundwater of this site (Figure 3-2; Online Supplemental Table 3-2), none of which show significant co-occurrences patterns with other taxa and are thus excluded from the global and Region 4 networks (Figure 3-4A; Appendix Figure 3-2D).

Despite the presence of sunlight, sulfide, and organic matter, The Pit lacks large, well-described anoxygenic photosynthesizing taxa in its halocline and saline groundwater. Nonetheless, an unclassified bin of *Rhodobacteraceae*, a diverse family of Alphaproteobacteria often implicated in the sulfur cycle (Pujalte et al., 2014), is present in 83% of samples (Online Supplemental Table 3-6) and is most abundant (~19%) in the freshwater and halocline of The Pit, where sunlight is present (Appendix Figure 3-5). Although we cannot assess its physiology with our methods, this family of Alphaproteobacteria is noted to host several genera of photosynthetic purple non-sulfur bacteria (Pujalte et al., 2014). This near-ubiquitous taxon also reaches almost 10% relative abundance in some communities near the sunless halocline of Blue Abyss (Appendix Figure 3-5). As such, this bin could represent a more flexible group of sulfur cyclers present across many environmental conditions in the aquifer in a similar manner as the unclassified *Comamonadaceae* bin (Figure 3-5). *Sulfurovum*-dominated communities have been previously observed in surface glacial waters receiving sulfide from a cold seep despite the thermodynamic favorability of other redox metabolisms such as anoxygenic photosynthesis (K. E. Wright et al., 2013). There, intense aerobic heterotrophy is thought to establish sub-oxic microenvironments to allow *Sulfurovum* to oxidize sulfide with nitrate. The creation of such sub-oxic microenvironments could be enhanced in open sites lacking horizontal outflows such as The Pit. In cenotes that are more directly

connected to outflowing conduits (such as Jailhouse cenote in Region 3), suspended and dissolved organic matter can better disperse as it transits downstream, theoretically preventing the establishment of such conditions.

Our data suggest that differing hydrological regimes significantly affect the composition and interactions present in the sulfur-cycling communities in the Yucatán carbonate aquifer. Interestingly, the *Sulfurovum*, *Sulfurospirillum*, and SUP05 cluster nodes within EBC21 and EBC34 in the global network are only found in abundances higher than 1% in the halocline and saline groundwater of Odyssey cave (Online Supplemental Table 3-2; Appendix Figure 3-5). This conduit is directly downstream of Jailhouse cenote; it is possible that smaller, mobile sub-oxic microenvironments created from the degradation of surface detritus that move downstream can support much lower abundances of the same sulfur cyclers as those found in The Pit, eventually dissipating. However, Odyssey cave contains much higher abundances of EBC15 in the global network (Figure 3-4), which contains many methanogenic taxa such as *Methyloparacoccus* (C. Li et al., 2021), suggesting that even non-pit cenotes could host fundamentally different biogeochemistries and community structures than their deeper counterparts.

3.5 Conclusions

The Yucatán carbonate aquifer is one of the largest anchialine ecosystems on the planet, harboring a diverse microbiome that colonizes disparate groundwater habitats. Our analysis of the distribution of 16S rRNA genes in this well-connected aquifer demonstrate that regionalism exists across geographic distance, water column depth, and cenote type, owning to the heterogeneous nature of the anchialine system. Significant compositional differences are noted between microbial communities in the saline groundwater and open seawater, despite the active circulation of Caribbean surface water ~10 km inland. Network analysis suggests that ubiquitous and

metabolically flexible taxa such as *Comamonadaceae* act as "keystone species," serving as reservoirs of metabolic potential throughout several potential niches. Further, we observe biogeographic patterns in the distribution and co-occurrence patterns of key ASVs, in which communities within a given hydrological region tend to be more similar to each other than those from other regions. We also find that communities living in The Pit, a deep cenote that is open to the surface, are not representative of those encountered throughout the rest of the aquifer, whose habitat space is mostly constituted by dark, oligotrophic conduits. The Pit also demonstrates the strongest vertical stratification as well as niche partitioning between the freshwater, halocline, and saline groundwater layers among the chosen study sites. We conclude the Yucatán carbonate aquifer hosts a diverse and flexible core microbiome whose members proliferate under specific environmental or hydrological conditions. We speculate that this leads to the expression of distinct biogeochemical paradigms in different areas of the aquifer.

3.6 Acknowledgements

We extend our sincere thanks to the Under The Jungle Dive Shop Family (Quintana Roo, Mexico), especially Natalie Gibb, whose efforts in the field made this research possible. We also thank the team of experienced technical cave divers who supported the July-August 2019 sampling expedition, including Alex Fraser, Nikolas Tkachenko, Vlada Dekina, Rory O'Keefe, and Vincent Rouquette-Cathala. Deep diving was undertaken by Luis Leal with safety support by Natalie Gibb. Hearty sustenance buoyed the field team from Jen, Bart, and the team at Turtle Bay Café & Bakery, as well as Imelda and family from her Ecococina. We additionally thank personnel at the Environmental Sample Preparation and Sequencing Facility at Argonne National Laboratory (Lamont, IL) for amplifying and sequencing DNA, as well as Rebecca Sponenburg and team for elemental analysis performed at the Quantitative Bio-element Imaging Center (Northwestern

University). This project has been supported with grants to MRO from the David and Lucille Packard Foundation and the Canadian Institute for the Advancement of Research – Earth 4D Program.

Chapter IV: *A Metagenomic View of Microbial Functional Potential throughout Mammoth Cave, KY*

4.1 Introduction

Terrestrial caves host diverse microbial communities and are windows into the biogeochemistry and microbial ecology of the shallow subsurface (Barton & Northup, 2007; Hathaway, Garcia, et al., 2014; Lavoie, 2017; Sarbu et al., 1996; Selensky et al., 2021). Shielded from photosynthetic primary production beyond the twilight zone, caves are accessible conduits into the transitionary zone between surface biomes and the deep subsurface biosphere. Microbial communities in caves subsist on a combination of surface-derived organic matter and *in situ* chemolithoautotrophy to meet carbon and energy demands (Barton & Northup, 2007; Sarbu et al., 1996; Selensky et al., 2021; Tomczyk-Żak & Zielenkiewicz, 2016). Interface environments such as these empower microbial ecologists to probe questions regarding the relative influence of surface or *in situ* nutrients on extant microbiota and their contributions to biogeochemistry.

Constraining the functional potential of cave microbiota is crucial for informing shallow subsurface biogeochemistry. Cave microbes have been traditionally thought to depend primarily on surficial processes and inputs, colonizing these dark habitats mainly as organoheterotrophs adapted to oligotrophic conditions (Barton et al., 2004; Laiz et al., 1999; Simon et al., 2003). However, our collective understanding of these enigmatic communities is shifting. For instance, Movile Cave (Romania) is a model shallow subsurface ecosystem that is almost entirely isolated from surface inputs; isotopic modeling suggests that sulfur-based chemolithoautotrophy there partially sustains populations of macroscopic fauna such as cave shrimp (Sarbu et al., 1996). Even in shallower cave systems that are in closer contact with the surface, such as Lava Beds National Monument (United States), common microbiota in soils such as Actinobacteria are implicated in

chemolithoautotrophy within biofilms, despite an abundance of surface-derived organic carbon entering the caves (Selensky et al., 2021). Methane-based carbon metabolisms are thought to partially sustain macroscopic communities in submerged caves in the Yucatán carbonate aquifer (Mexico; Brankovits et al., 2017). A consortium of carbon-fixing, ammonia-oxidizing, and nitrite-oxidizing bacteria and archaea in carbonate sediments within Pindal Cave (Spain) are thought to facilitate the net consumption of methane and production of nitrite in such sediments, forming amorphous carbonate "moonmilk" deposits as indirect byproducts (Martin-Pozas et al., 2022). Further, populations of nitrite-oxidizing bacteria have been noted to be significantly higher in carbonate sediments within Mammoth Cave (United States) compared to overlying surface soils (Fliermans & Schmidt, 1977). Considering these and other examples, *in situ* primary productivity in caves may be relatively common, though it remains underexplored. Mammoth Cave in particular presents as an ideal natural laboratory to probe microbial functional capacities in a shallow subsurface environment spanning multiple habitats.

As the longest mapped cave system on the planet, Mammoth Cave, KY (United States) hosts a diverse range of subterranean microhabitats due to its complex geological setting. The Green River sets the base level of the Mammoth Cave system, where new passages continue to form, while more upland areas contain ancient caves. (Palmer, 2017). Beneath areas where the Green River or its tributaries have not eroded away the impermeable Fraileys Shale Member associated with the Big Clifty Sandstone cap, conduits formed within the underlying limestone strata tend to be very dry, even so far as to accumulate highly soluble sulfate minerals such as mirabilite and delicate gypsum flowers that would otherwise dissolve (J. W. Hess & White, 1989). By contrast, caves under areas without the cap rock are much wetter, marked by active carbonate speleothem formation. Oxic surface waters entering these regions of the cave system through drips,

waterfalls, or underground rivers carry nutrients that can fuel microbial growth. Further, in areas with open surface entrances and appealing conditions, bat populations can be established and bring in significant amounts of nitrogenous organic matter in the form of guano that can be harnessed by microbial communities (W. Hess, 1900). The distinct physicochemical landscape found within this complex cave system has strong potential to affect microbial community composition and functional potential. However, this environment has not been widely queried to date with modern sequencing approaches.

The nitrogen cycle has immense ecological, cultural, and economic importance in Mammoth Cave, as it has served as a major historical source of raw saltpeter, $Ca(NO_3)_2$ (Fliermans & Schmidt, 1977; W. Hess, 1900; Hill, 1981). Copious amounts of nitrate (up to several weight percent by volume) have been observed to accumulate in sediments from Mammoth Cave and other carbonate caves despite the lack of a clear consensus on its origins (Fliermans & Schmidt, 1977; W. Hess, 1900; Hill, 1981; Pace, 1971). A seminal microbiological survey in Mammoth Cave observed that cell abundances of nitrite-oxidizing *Nitrobacter* spp. were significantly higher in cave sediments than those in overlying soils (Fliermans & Schmidt, 1977). Further, after thoroughly leaching *in situ* sediments of nitrate, *Nitrobacter* cell abundances did not change while nitrate concentrations returned to natural levels within a few years of extraction, requiring the presence of active oxidative nitrogen cycling processes (Fliermans & Schmidt, 1977). Nevertheless, it is still unknown whether such processes are primarily abiotic or biotic in nature. Further, assuming microbes mediate the oxidation of nitrite, the sources of more reduced nitrogenous compounds, such as ammonia, are ultimately still unknown in Mammoth Cave. Bat guano could be a readily available source of ammonia, but nitrate also actively accumulates in areas of the system beyond the reach of modern or historical bat populations (Fliermans &

Schmidt, 1977). Additionally, nitrate and more reduced nitrogen compounds could deposit directly into cave sediments from surface-derived fluids (Hill, 1981; Pace, 1971), though it remains unclear whether this could produce the concentrations observed *in situ*. Consortia of carbon-fixing, ammonia-oxidizing, and nitrite-oxidizing bacteria and archaea have been found in carbonate cave sediments elsewhere (Martin-Pozas et al., 2022) and could also contribute to nitrate accumulation from reduced nitrogen inputs in this system. Identifying microbial players capable of mediating these carbon and nitrogen cycling processes is one of the first steps in constraining these possibilities and the overall nitrogen balance of Mammoth Cave.

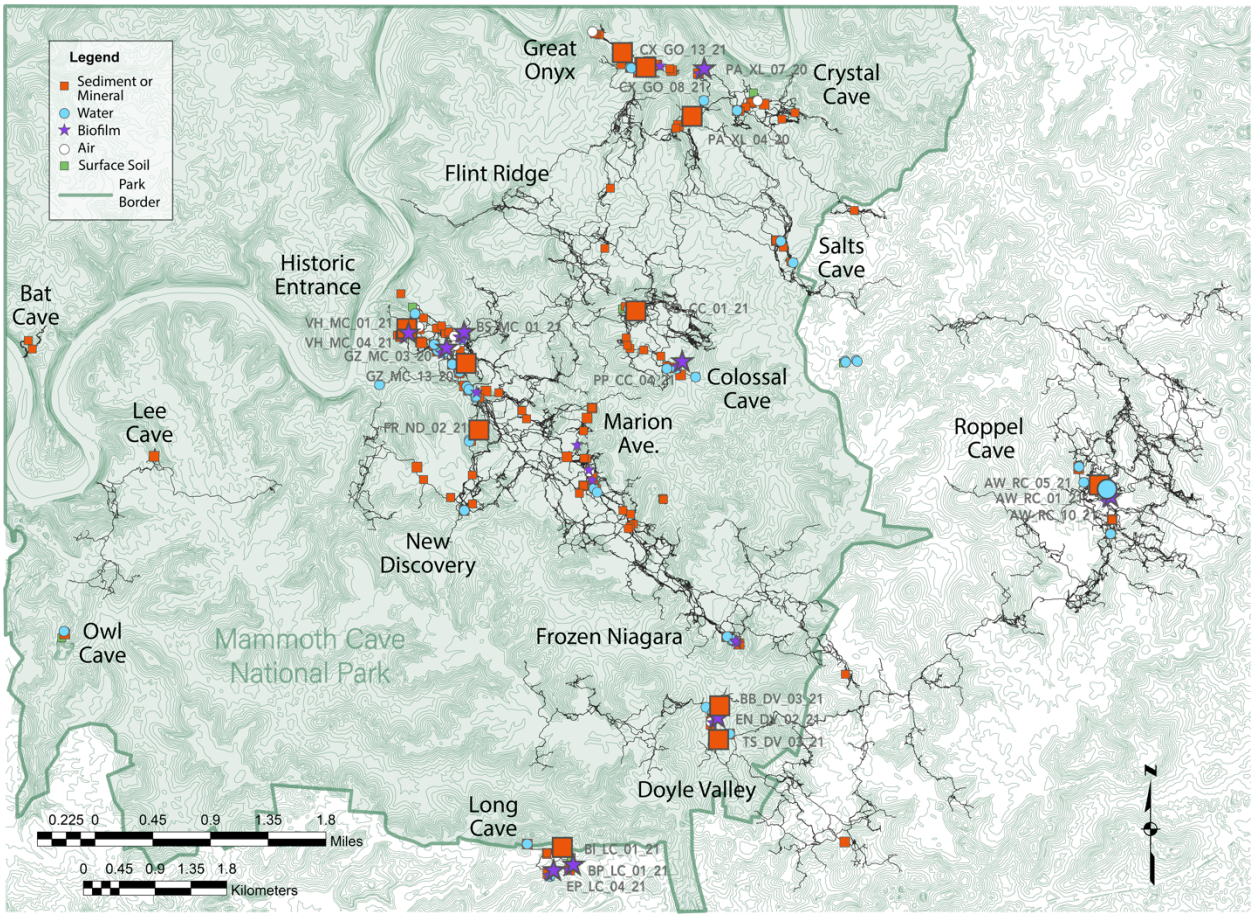
Here, we identify potential microbial players in the nitrogen and carbon cycles across a range of microhabitats in Mammoth Cave, KY. We focus our analysis on identifying genes involved in cycling carbon and nitrogen associated with metagenome-assembled genomes (MAGs) from 21 sites across the cave systems representing various cave habitat endmembers. We pair these metagenomes to stable carbon and nitrogen isotope data from 186 sites representing different source reservoirs as well as concentrations of nitrate, nitrite, ammonia, and sulfate to constrain potential biogeochemical processes in the cave system. Our analysis emphasizes the twin roles that carbon and nitrogen cycling both play in the microbial ecology of Mammoth Cave. To our knowledge, we also provide the first draft genomes classified to the *Egibacteraceae* and *Pseudonocardiaceae* from a cave environment. These two families of Actinobacteriota are highly abundant, yet poorly characterized, community members of biofilms in both limestone and volcanic caves worldwide (Cuezva et al., 2012; Farda et al., 2022; Gonzalez-Pimentel et al., 2018; Hathaway, Garcia, et al., 2014; Lavoie et al., 2017; Riquelme et al., 2015). These MAGs are shown to support autotrophic, hydrogen-driven lifestyles supplemented by carbon monoxide oxidation

and potential syntrophy with other phyla, providing an unexpected perspective on the nature of microbial functional potential in cave environments.

4.2 Methods

4.2.1 Field site description and sampling approach

Figure 4-1. ***Sample sites throughout the wider Mammoth Cave region.*** A total of 186 sites representing cave sediments, surface soils, biofilms, and mineral features were taken from various parts of the Mammoth Cave system. A total of 40 water samples were also collected for geochemical, isotope, and DNA analysis. Enlarged points represents sites from which the metagenomes presented here originate. Map was created by M. and B. Osburn.



Mammoth Cave is situated within the drainage basin of the Green River in central Kentucky, where its 686 km of mapped passages cut through three major limestone strata deposited 340-330 million years ago (Palmer, 1981, 2017; Pohl, 1970). The deepest portions of the cave

system are within the dolomitic St. Louis Limestone, which often contains nodules of chert or shale and is the current hydrologic base level and horizon of new cave formation. Above the St. Louis Limestone lies the thinly bedded Ste. Genevieve Limestone, which also commonly includes dolomite and shale and includes the majority of Mammoth Cave passages (Palmer, 2017). The uppermost and youngest stratum of Mammoth Cave is the Girkin Formation, which contains several layers of less consolidated limestone and is capped by the impermeable Big Clifty Sandstone (Palmer, 1981, 2017). Though the limestones were deposited during the Mississippian Period, Mammoth Cave itself begun forming approximately 30 million years ago after the Green River eroded through the Big Clifty Sandstone to preferentially flow through pre-existing cracks and bedding planes, forming conduits via dissolution (Fliermans & Schmidt, 1977; Palmer, 2017). A large-scale monocline structure exists beneath much of the karst in south-central Kentucky, where euxinic fluids may rise to the surface from deeper petroleum and sulfide bearing units, such as Phantom Waterfall in adjacent Parker Cave (Angert et al., 1998; Rick Olson, 2013).

A total of 186 samples were collected across the cave system from a range of environments for geochemical and stable C and N isotopic analysis (Figure 4-1; Online Supplemental Table 4-1). Of these, 21 also underwent metagenomic sequencing (Figure 4-1). Four metagenome samples were collected during a December 2020 field season, while the remaining 17 were obtained in July and August 2021. Sediments, mineral features, and biofilms (solid samples) were aseptically sampled with ethanol-cleaned spatulas and chisels (when necessary). All samples were taken in accordance with the limitations of our research permit and in a careful manner consistent with the preservation of the National Park. We collected minimal material from distributed areas hidden from obvious view to minimize environmental disturbance. Aqueous samples were collected by first thoroughly rinsing a filter apparatus with *in situ* water to prevent cross-contamination. After

rinsing, 1.3 L of water was passed through a sterile 0.2 μm filter. Filtered water was stored at 4°C until geochemical analysis in the laboratory. Samples destined for DNA sequencing were stored in liquid nitrogen in the field until receipt at the Osburn Geobiology Lab at Northwestern University, at which point they were stored at -80°C until DNA extraction. Corresponding solid samples were also collected for geochemical and stable isotope analysis and were stored at -20°C until laboratory analysis.

4.2.2 Geochemical analyses

Concentrations of water soluble nitrate, nitrite, ammonia, and sulfate were measured in triplicate from cave sediments and biofilms with previously published spectrophotometric methods that we adapted for miniature volumes in a 96-well plate (Kolmert et al., 2000; Krom, 1980; Miranda et al., 2001). We first successively leached samples with a 2M KCl solution to extinction (Kaneko et al., 2010) and pooled individual leachates from each sample. For each individual reaction, we added 100 μL of pooled leachate (or 0.2 μm-filtered cave water, depending on the sample type) to a 96-well plate containing a range of standards of the given analyte. Absorptions of wavelengths 543, 660, and 420 nm were measured on a Gen5^{TM} 96-well plate Spectrophotometer (BioTek) and compared to standards for nitrite, ammonia, and sulfate, respectively, to determine leachate concentrations of each analyte. Nitrate concentrations were determined by subtracting the difference between fluids reduced with vanadium chloride (nitrate and nitrite) and those without such reduction (nitrite only; Miranda et al., 2001). Analyte concentrations from sediments and biofilms are reported in parts per million (Supplementary Table 1).

4.2.3. Bulk $\delta^{13}C_{org}$ and $\delta^{15}N$ analysis

Total organic carbon (TOC) abundance and its ^{13}C composition ($\delta^{13}C_{org}$), as well as total nitrogen (TN) abundance and its ^{15}N composition ($\delta^{15}N$), were measured for each solid sample on a Costech 4010 Elemental Analyzer coupled to a Thermo Delta V+ IRMS through a Conflo IV interface (EA-IRMS). Freeze-dried samples were homogenized, acidified with 0.1 M HCl, rinsed in MilliQ water to remove residual acid, and weighed into tin capsules for EA-IRMS. Previous tests determined water rinsing does not significantly affect C or N isotope compositions (data not shown). Acetanilide and urea standards of known isotopic composition (Schimmelmann et al., 2009) were used to report sample isotope compositions relative to VPDB ($\delta^{13}C_{org}$) and atmospheric air ($\delta^{15}N$). A total of 11 fluids and 14 sediment leachates were sent to the Stable Isotope Facility at the University of California-Davis for measurement of nitrate-specific $\delta^{15}N$ values ($\delta^{15}N_{NO3}$) using the denitrifier method (Casciotti et al., 2002). The concentration and stable isotope composition of DOC ($\delta^{13}C_{DOC}$) from 40 cave waters representing pools, drips, and underground rivers were analyzed at the UC Davis Stable Isotope Facility using an O.I. Analytical Model 1030 TOC Analyzer connected to a PDZ Europa 20-20 IRMS.

4.2.4. DNA sequencing and metagenomic analysis

We extracted DNA for shotgun metagenomics from 21 samples and a negative control with a commercial kit (ZymoBIOMICS™ DNA MiniPrep) according to the manufacturer's instructions. We note that the water filters sample was first shredded into small strips using scissors thoroughly cleaned with decontaminant solution (Thermo Scientific™ DNA Away™) before extraction. If a given extract contained less than 1 micrograms of DNA as determined by a PicoGreen® double-stranded DNA assay (Thermo Scientific™), the sample was re-extracted and pooled together until that amount could be recovered. A total of 21 extracts were sent to the NUSeq

Core Facility (Northwestern University) for shotgun metagenomic sequencing on a NovaSeq 6000 SP flow cell (Illumina). Illumina adapters were trimmed from the resultant paired-end reads with version 0.39 of the *trimmomatic* software (Bolger et al., 2014). Metagenome contigs and scaffolds were assembled from the trimmed reads via *metaSPAdes* (v. 3.14.1; Nurk et al., 2017) from a consensus of *k*-mers of lengths 15, 21, 33, 55, and 77.

Metagenome-assembled genomes (MAGs) were generated by first separately binning each assembly with three different programs (*maxbin2*, *CONCOCT*, and *metabat2*) and combining them with *metaWRAP* (Uritskiy et al., 2018), with a final manual curation step in Anvio' (v.7, Eren et al., 2020). Briefly, we used the default parameters for the *maxbin2* and *CONCOCT* modules within *metaWRAP*, while MAGs were generated separately in *metabat2* across a range of minimum contig lengths (1.5, 2, 2.5, and 3kbp). MAG completeness and contamination were assessed with *CheckM* (v. 1.0.7; Parks et al., 2015) for each minimum contig length iteration for the *metabat2* MAGs; bins assembled from contigs at least 2 kbp were determined to be the least contaminated and most complete overall. As such, we combined and refined bins generated from *maxbin2*, *CONCOCT*, and *metabat2* (>2kbp) with the bin_refinement module in *metaWRAP* to generate consensus MAGs that were at least 50% complete and 50% contaminated for downstream curation. These prefiltered consensus MAGs were then imported into new Anvio' (Eren et al., 2020) profiles in which built-in hidden Markov models (HMMs) detected the presence of single-copy genes to estimate completeness and contamination as an complementary quality check step to *CheckM*. In addition to these metrics, GC content was considered when manually refining low-quality MAGs (i.e., contamination >10%) to reduce contamination and increase completeness when possible.

The resulting curated set of consensus MAGs was then exported from Anvio' and assessed for quality with *CheckM*. The abundance of each curated consensus MAG (bin copies per million

metagenomic reads) was estimated using the quant_bins module in *metaWRAP*. To further improve quality, MAGs were then reassembled against their respective trimmed metagenomic reads via the reassemble_bins module in *metaWRAP*, which leverages *CheckM* to only include reassembled MAGs in the final set if bin quality increases. To quantify the abundance of a particular MAG more accurately in reference to the broader community, we stress that the quant_bins module was applied to the nonreassembled representative for each bin (Uritskiy et al., 2018), while subsequent taxonomically-resolved analysis was considered on the dereplicated representative.

Genes from the reassembled MAG set was annotated with *METABOLIC* (Zhou et al., 2022) using gene calls generated from a built-in version of *prodigal* (Hyatt et al., 2010). The reassembled MAGs were subsequently dereplicated with *dRep* (Olm et al., 2017) to enable MAG-specific abundance comparisons across metagenomic libraries. Taxonomy was assigned to each dereplicated MAG using *GTDB-Tk* (v. 2.2.0; Chaumeil et al., 2019), which leverages *pplacer* to place MAGs on domain-specific reference trees (Matsen et al., 2010). Wider taxonomy in the metagenomes was assessed by extracting and counting copies of the V4 region of the 16S rRNA gene using RiboTagger (Xie et al., 2016). The extracted sequences (n = 415) were then manually classified with Silva version 138 (Quast et al., 2012). The 16S rRNA gene copies from each community were then binned by taxonomic class, after which the table was rarefied to a depth of 49 to facilitate cross-community comparisons. All code written to process metagenomic reads into assemblies, the final set of dereplicated MAGs, and representative 16S rRNA gene sequences is available at https://github.com/mselensky/mammoth-metagenomes.

4.3 Results

4.3.1. Sample descriptions, geochemistry, and C/N isotopic compositions

A total of 186 solid samples representing the Mammoth Cave system (Figure 4-1) were analyzed for geochemistry (specifically nitrate, nitrite, ammonia, and sulfate concentrations) as well as bulk $\delta^{13}C_{org}$ and $\delta^{15}N_{bulk}$ compositions (Figure 4-2; Online Supplemental Table 4-1). Samples are characterized into the following categories: surface soil, cave sediment, mineral features, bat guano, or cave biofilms. Surface soils represent surface endmembers and tend to be characterized by the presence of plant detritus and humus, and therefor relatively high organic matter content. By contrast, the characteristics of cave sediment are highly site-specific, often reflecting local geological conditions. Sediments can be dry or wet, and rich in quartz/carbonate sand, carbonate mud, gypsum, or some combination of the three. Mineral features encompass a variety of small speleothem features encountered in Mammoth Cave and other similar systems, such as polyps, crusts, or coatings, and can be black, white, or brown in color. Cave biofilms often carry a characteristic "petrichor" odor and tend to be white or tan in color, though some yellow biofilms are also observed in the cave system. Biofilms can be thin, continuous films on cave surfaces, or can comprise a collection of separate circular colonies some millimeters in diameter.

Surface soils throughout the region exhibit a relatively narrow spread of $\delta^{13}C_{org}$ values with a more variable distribution in $\delta^{15}N_{bulk}$ values (Figure 2), with mean values of -28.0 $\pm$ 1.6‰ and +1.7 $\pm$ 2.1‰, respectively. By contrast, despite some overlap, cave biofilms tend to be the most depleted in both $\delta^{13}C_{org}$ and $\delta^{15}N_{bulk}$ values on average (-31.7 $\pm$ 4.9‰ and 0.0 $\pm$ 3.4‰), respectively). Conversely, cave sediments and mineral features are consistently enriched in both isotopes relative to biofilms ($\delta^{13}C_{org}$ = -25.1 $\pm$ 2.7‰, -25.7 $\pm$ 1.9‰, and $\delta^{15}N_{bulk}$ = 2.7 $\pm$ 3.8‰, 3.1 $\pm$ 2.5‰, respectively, Figure 4-2). Although only one direct bat guano sample is included in the

dataset, it exhibits relatively enriched $\delta^{13}C_{org}$ and $\delta^{15}N_{bulk}$ values of -24.1 and +7.5‰, respectively. The subset of 21 samples analyzed for metagenomes are highlighted in the figure and their mean $\delta^{13}C_{org}$ and $\delta^{15}N_{bulk}$ values grouped by sample category are similar to those from the larger dataset (Table 4-1).

Figure 4-2: ***Comparison of δ¹³C and δ¹⁵N values by sample type.*** The shaded boxes represent the spread of isotope values for each element (mean plus one standard deviation in either direction). The spread of $\delta^{15}N_{NO3}$ and $\delta^{13}C_{DOC}$ values are reported for n = 14 water samples (shaded in blue; points not shown), while the remaining points reflect $\delta^{15}N_{bulk}$ and $\delta^{13}C_{org}$ values. Samples with metagenome representatives are marked as large diamonds (note that both $\delta^{13}C$ and $\delta^{15}N$ values could not be measured in every metagenome).

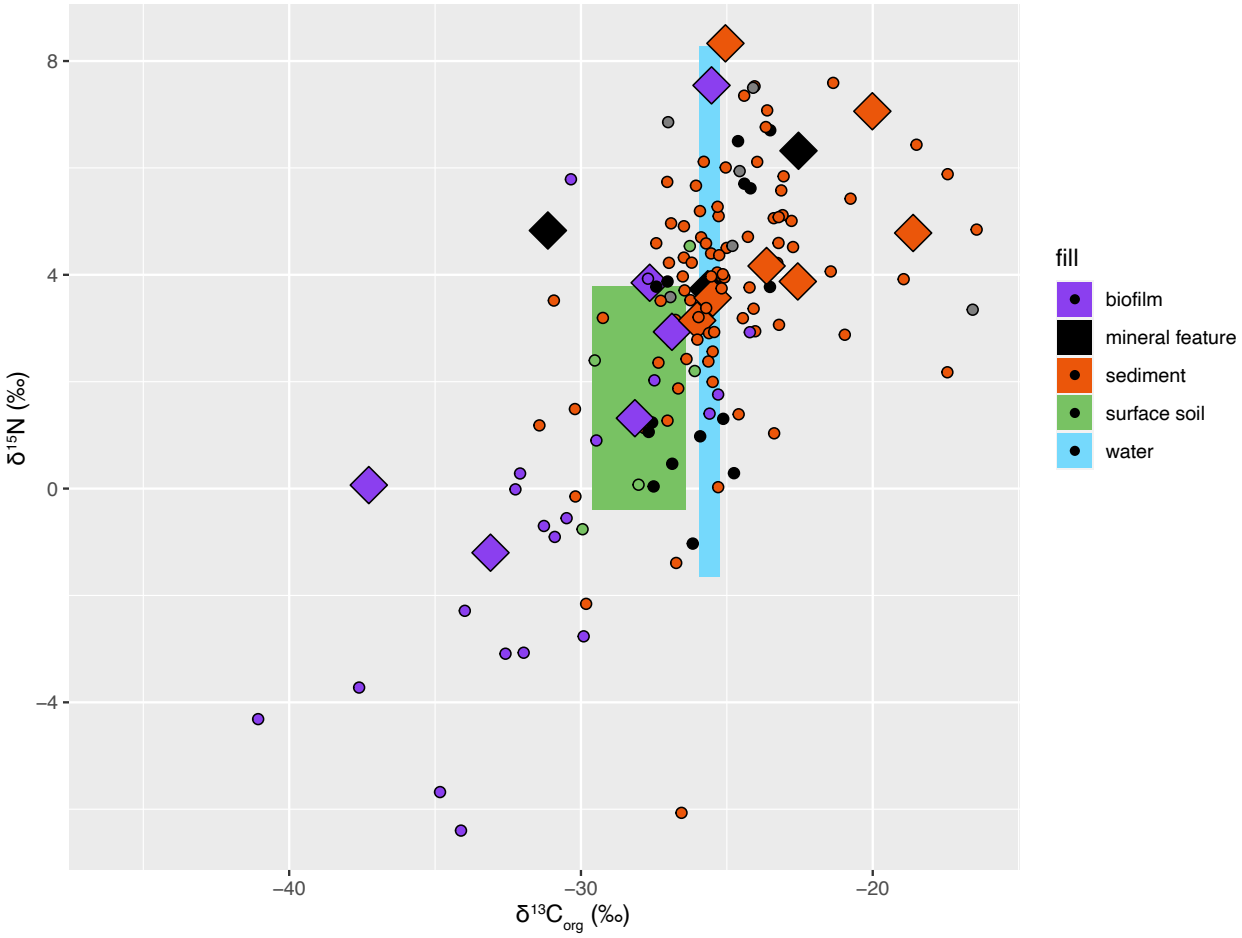
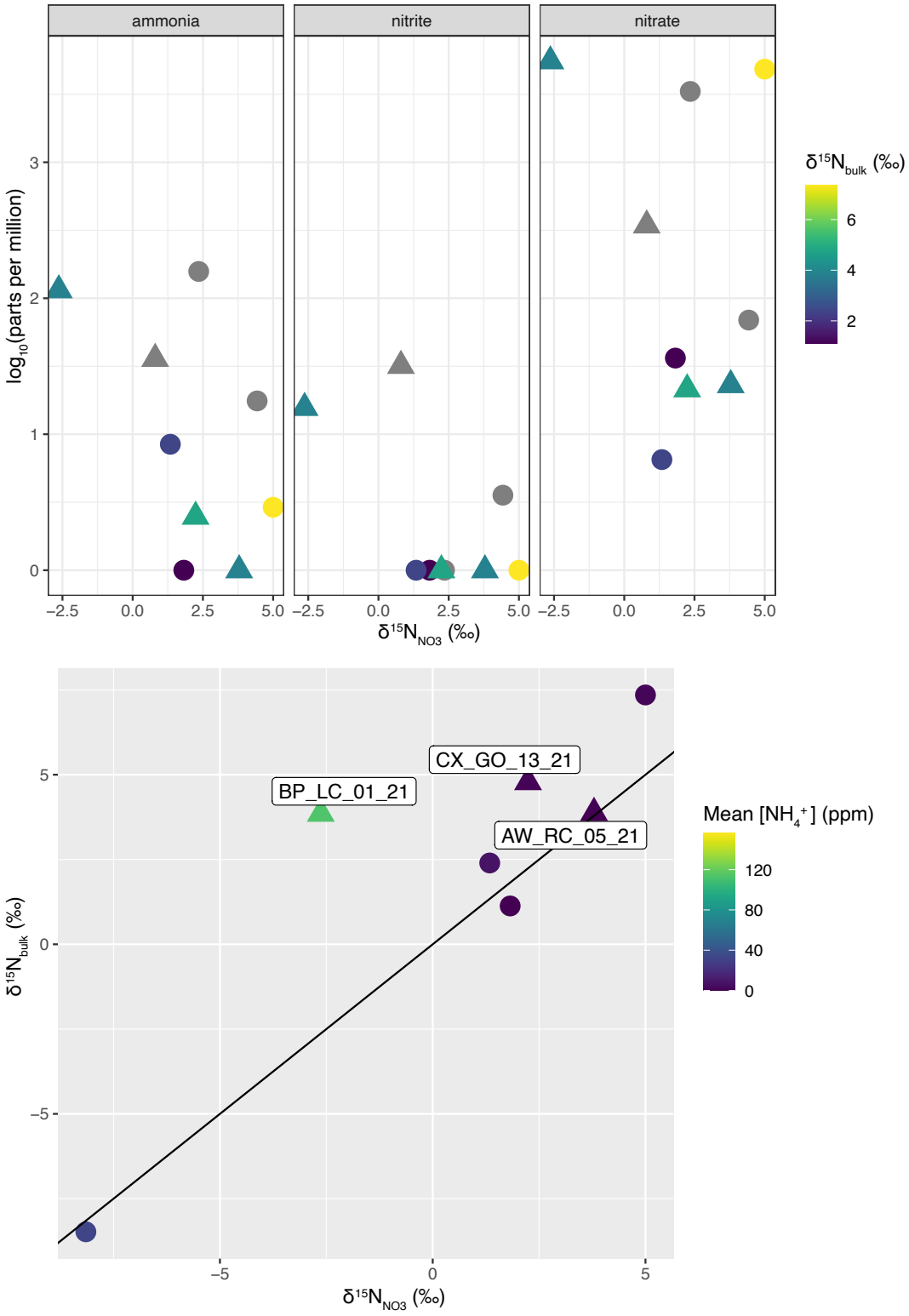


Table 4-1: Characteristics of 21 metagenomes collected throughout Mammoth Cave, KY.

Metagenome	Date	Sample Type	Number of Scaffolds	Number of Reads	Number of Dereplicated MAGs	$\delta^{13}C_{org}$	$\delta^{15}N_{bulk}$	$\delta^{15}N_{NO3}$	$[NH_4^+]$	$[NO_3^-]$	$[SO_4^{2-}]$	$[NO_2^-]$
						(‰)			*parts per million*			
AW_RC_01_21	27-Jul-21	water	1929074	12102512	11	-24.9	-	-	-	3±0.1	486.9±9.6	0.1±0.1
AW_RC_05_21	27-Jul-21	sediment	1407951	11499158	15	-22.6	3.9	3.8	-	22±0.5	3030.4±0	0±0
AW_RC_10_21	27-Jul-21	biofilm	969497	14225800	21	-33.4	-	0.8	34.9±23.1	342.6±10	1014.8±241.3	30.9±20.1
BB_DV_03_21	28-Jul-21	mineral feature	2153928	11583204	5	-22.5	6.3	-	11.2±0	58.2±2.6	348.7±43.2	-
BL_LC_01_21	1-Aug-21	sediment	511072	12370885	5	-20.0	7.1	-	66±5.3	2359.2±351.5	133.9±11.2	-
BP_LC_01_21	1-Aug-21	biofilm	991583	10185561	6	-27.6	3.9	-2.6	113.6±9.3	5528.7±33.8	1269.8±76.2	14.6±5.7
BS_MC_01_21	31-Jul-21	mineral feature	2045461	12976858	8	-25.7	3.7	-	21.1±5.6	16.1±0.7	138.6±37.6	-
CX_GO_08_21	2-Aug-21	sediment	1754354	15479230	12	-25.9	-	-	-	5.6±0.7	1015±0	-
CX_GO_13_21	2-Aug-21	sediment	1070495	8250747	8	-18.6	4.8	2.2	1.5±0.2	20.4±0.2	174.7±6.5	-
EN_DV_02_20	9-Dec-20	biofilm	2935603	15123155	3	-26.9	2.9	-	0.9±0	6.2±0.7	-	-
EP_LC_04_21	1-Aug-21	biofilm	997872	12059505	18	-31.1	4.8	-	6.9±0	9492.2±293	5386.1±94.5	6.4±0.8
FR_ND_02_21	26-Jul-21	sediment	1563474	11904240	12	-23.6	4.2	-	3±0	99.4±0.6	92.7±9.7	-
GZ_MC_03_20	6-Dec-20	biofilm	1595797	12312917	4	-37.3	0.1	-	12±0.5	289.3±2.5	515.5±29.7	-
GZ_MC_13_20	6-Dec-20	sediment	485336	5867610	8	-26.0	3.1	-	0.6±0	8298.2±317.9	10326.6±302.5	-
PA_XL_04_20	8-Dec-20	sediment	2293923	11874353	3	-25.5	3.6	-	-	38.5±4.8	-	-
PA_XL_07_20	8-Dec-20	biofilm	2353733	12055325	1	-28.2	1.3	-	1±0	5±0.6	101.5±12.9	-
PP_CC_04_21	30-Jul-21	biofilm	522373	12784851	10	-33.1	-1.2	-	22.2±5.8	239.2±38.9	1599.6±55.9	20.2±5.5
TS_DV_03_21	28-Jul-21	mineral feature	1959600	10920166	0	-	6.0	-	2.7±0.2	2.7±0	77.7±0	-
VD_CC_01_21	30-Jul-21	mineral feature	2628743	14354898	3	-	0.3	-	-	-	88.6±17.1	-
VH_MC_02_21	31-Jul-21	mineral feature	526967	12638691	16	-25.0	8.3	-	6.1±0.3	7.3±0.5	18160.2±1140.4	-
VH_MC_04_21	31-Jul-21	biofilm	185318	10165086	12	-25.5	7.5	-	9.6±2.1	8660.6±772	48700.9±950.3	4.9±0.4

To better constrain the isotopic composition of potential N pools available to cave communities, we measured $\delta^{15}N_{NO3}$ from a total of 14 cave waters (Figure 4-1) and 11 leachates from solid samples (not displayed; Online Supplemental Table 4-1), including four with representative metagenomes (triangles; Figure 4-3). Though the relationship is weak, as ammonia concentrations decrease, $\delta^{15}N_{NO3}$ values generally becomes more enriched (Figure 3A). There are no clear trends between the concentrations of nitrite or nitrate and $\delta^{15}N_{NO3}$ values in our dataset (Figure 3). Further, sample BP_LC_01_21 is notably more ^{15}N-enriched (+3.85‰) in bulk values compared to $\delta^{15}N_{NO3}$ values (-2.63‰), in contrast to most other samples whose bulk nitrogen and nitrate isotope values are more similar (Figure 4-3B).

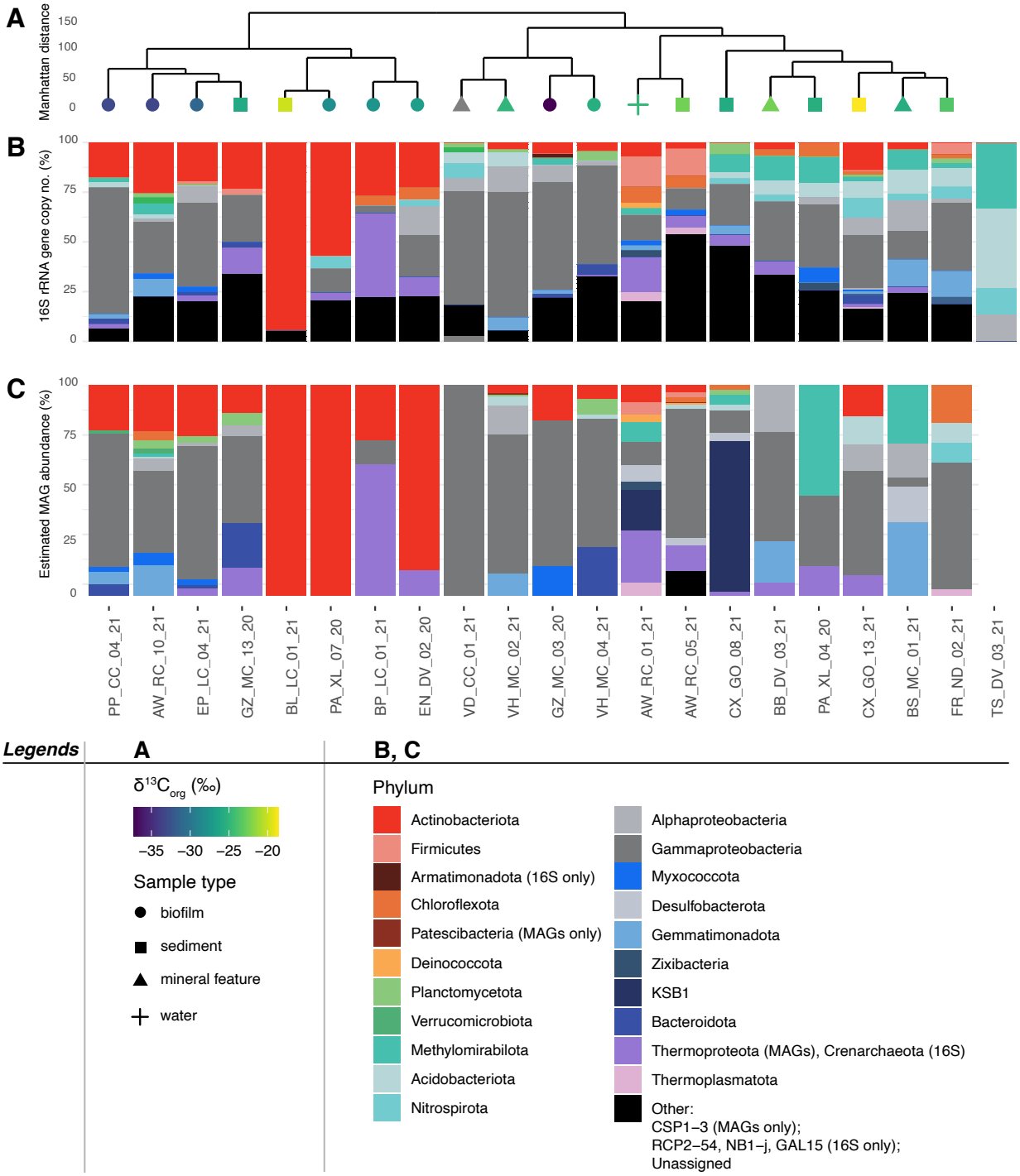
Figure 4-3: ***Nitrate isotope composition against nitrogen reservoir sizes.*** Metagenome representatives are marked by triangles, while data from other samples (Online Supplemental Table 4-1) are marked as circles.



4.3.2. Assembly, recovery, and annotation of metagenome-assembled genomes (MAGs)

From 21 metagenomes, we obtained a total of 250,734,752 paired-end reads separately assembled into 30,882,154 contigs (Table 4-1). After consensus binning, manual curation, and reassembly (see Methods), we assembled 361 MAGs with genome completeness ≥ 30% as estimated by CheckM (Parks et al., 2015). Of these, 181 MAGs were deemed "high-quality" (completeness ≥ 50%, contamination ≤ 10%) and were subsequently annotated for the presence of key functional genes using the METABOLIC software (Zhou et al., 2022). This "redundant" set of high-quality MAGs was dereplicated (see Methods) into 166 unique, high-quality MAGs (Table 4-1) representing 20 bacterial and 2 archaeal phyla (Figure 4-4; Online Supplemental Table 4-2). Metagenome TS_DV_01_21 did not yield any MAGs after quality control and is thus excluded from our MAG-centric analyses (Figures 4-5).

Figure 4-4. ***Annotated MAGs broadly represent the most abundant phyla in Mammoth Cave metagenomes.*** MAG abundance is visualized at the phylum level and reported in relative abundance (%). 4A. Complete hierarchical clustering was performed on a Manhattan distance matrix calculated from a class-level relative abundance table of rarefied 16S rRNA gene counts pulled from whole metagenome shotgun data (see Methods). 4B. Relative abundance of taxa visualized at the phylum level from Silva (v. 138) classified 16S rRNA gene (V4) sequences. The total number of 16S reads in metagenome TS_DV_02_21 was less than the rarefied sampling depth and is thus excluded from the dendrogram. 4C. The relative abundance of MAGs were quantified via Salmon and classified with GTDB-tk (v. 2.2.0). Note the variable number of MAGs obtained per metagenome in Table 4-1. Phylum-level classifications are linked between the Silva and GTDB databases according to Online Supplemental Table 4-3.


A
Manhattan distance
150
100
50
0
B
16S rRNA gene copy no. (%)
100
75
50
25
0
C
Estimated MAG abundance (%)
100
75
50
25
0
PP_CC_04_21
AW_RC_10_21
EP_LC_04_21
GZ_MC_13_20
BL_LC_01_21
PA_XL_07_20
BP_LC_01_21
EN_DV_02_20
VD_CC_01_21
VH_MC_02_21
GZ_MC_03_20
VH_MC_04_21
AW_RC_01_21
AW_RC_05_21
CX_GO_08_21
BB_DV_03_21
PA_XL_04_20
CX_GO_13_21
BS_MC_01_21
FR_ND_02_21
TS_DV_03_21
Legends
A
δ13Corg (‰)
−35 −30 −25 −20
Sample type
biofilm
sediment
mineral feature
water
B, C
Phylum
Actinobacteriota
Firmicutes
Armatimonadota (16S only)
Chloroflexota
Patescibacteria (MAGs only)
Deinococcota
Planctomycetota
Verrucomicrobiota
Methylomirabilota
Acidobacteriota
Nitrospirota
Alphaproteobacteria
Gammaproteobacteria
Myxococcota
Desulfobacterota
Gemmatimonadota
Zixibacteria
KSB1
Bacteroidota
Thermoproteota (MAGs), Crenarchaeota (16S)
Thermoplasmatota
Other:
CSP1–3 (MAGs only);
RCP2–54, NB1–j, GAL15 (16S only);
Unassigned

Microbial community composition as estimated by rarefied 16S rRNA gene counts within metagenomes (Xie et al., 2016) demonstrates class-level similarities among metagenomes from similar sample types (Figure 4-4A,B), despite some variability. Further, independent abundance estimates from our high quality MAG set (Figure 4-4C; Patro et al., 2017; Uritskiy et al., 2018) corroborate overall compositions inferred from 16S rRNA gene abundance (Figure 4-4B). We focus the reporting of our results on the distributions of key MAGs (Figure 4-4C) within major class-level community clusters (Figure 4-4A).

Cluster 1 represents the biofilms PP_CC_04_21, AW_RC_10_21, and EP_LC_04_21, which exhibit some of the most depleted $\delta^{13}C_{org}$ values in our dataset and form a distinct cluster with GZ_MC_13_20 (Figure 4-4A). The 16S rRNA gene-based community profiles of the biofilms show a high proportion of hits mapped to Gammaproteobacteria (Figure 4-4B; Online Supplemental Tables 4-4, 4-5). Correspondingly, the most abundant individual MAGs from PP_CC_04_21 and AW_RC_10_21 are classified to an uncultured clade of Gammaproteobacteria, JACCXJ01 (Figure 4-4C; Online Supplemental Table 4-2; Ortiz et al., 2021). A notable exception is the purple biofilm EP_LC_04_21, which is dominated by a MAG classified to the Gammaproteobacterial order Xanthomonadales (Online Supplemental Table 4-2). Additionally, the most abundant MAG in GZ_MC_13_20 is instead a novel member of the recently described family of Bacteroidota, *Salinibacteraceae* (Online Supplemental Table 4-2; Viver et al., 2018). The 16S rRNA gene- and MAG-informed views of composition consistently rank Actinobacteriota as comprising roughly one fifth of these communities (Figure 4-4B,C), with the class Actinomycetia being notably identified in both sets. Other, lower abundance phyla included in both sets include Gemmatimomonadota, Methylomirabilota, and Myxococcota. Metagenomes PP_CC_04_21 and AW_RC_10_21 notably both harbor MAGs classified to the Actinobacteriota

families *Egibacteraceae* and *Pseudonocardiaceae* (c. Actinomycetia; Online Supplemental Table 4-2). The Archaeal families *Nitrosomonadaceae* and *Nitrosopumilaceae*, within the phyla Thermoproteota (GTDBtk) or Crenarchaeota (Silva), are notably absent or low abundance in both views of community composition in the biofilms PP_CC_04_21 and AW_RC_10_21, unlike the other two in Cluster 1 (Figure 4-4B,C).

The metagenomes BL_LC_01_21, PA_XL_07_20, BP_LC_01_21, and EN_DV_02_20 form Cluster 2, which is mostly comprised of biofilms (Figure 4-4A). The proportion of Actinobacteriota here is markedly higher compared to other communities (Figure 4-4B,C). In both views of community composition, the sediment BL_LC_01_21 contains the highest proportion of Actinobacteriota (Figure 4-4B,C). The five MAGs from BL_LC_01_21 represent the families *Nocardioidaceae*, *Streptosporangiaceae*, *Streptomycetaceae*, *Micromonosporaceae*, and *Rubrobacteraceae* (Online Supplemental Table 4-2). Only one quality MAG (f. *Pseudonocardiaceae*) was recovered from PA_XL_07_20 (Online Supplemental Table 4-2). In addition to an Actinobacteriota MAG (f. *Miltoncostaeaceae*), BP_LC_01_21, a bat-impacted white biofilm (Table 4-1), is particularly dominated by the families of ammonia-oxidizing Archaea (AOA) *Nitrososphaeraceae* (Tourna et al., 2011). Another family of AOA, *Nitrosopumilaceae*, is present in this biofilm and EN_DV_02_20 (Online Supplemental Table 4-2).

Deeply branching from Clusters 1 and 2, the metagenomes VD_CC_01_21, VH_MC_02_21, GZ_MC_03_20, and VH_MC_04_21 comprise Cluster 3. These communities harbor abundant MAGs mapped to the Gammaprotebacterial orders Burkholderiales, Steroidobacterales, or Xanthomonadales (Figure 4-4, Online Supplemental Table 4-2). The Actinobacteriota MAGs that are present in these metagenomes are classified to orders Acidimicrobiales, Nitriliruptorales, or class Acidimicrobiia (Online Supplemental Table 4-2).

GZ_MC_03_20, a white biofilm interspersed with a black coating, is the most ^{13}C-depleted sample ($\delta^{13}C_{org}$ = -37.27‰) and notably clusters away from the other ^{13}C-depleted biofilms (Figure 4-4A). The historically bat-impacted metagenomes VH_MC_02_21 and VH_MC_04_21 lack AOA-related 16S sequences and MAGs, instead harboring various abundant Xanthomonadales MAGs and (Figure 4-4B,C). The latter metagenome is also noted to harbor the same *Salinibacteraceae* MAG described from GZ_MC_13_20 (Online Supplemental Table 4-2).

The metagenomes AW_RC_01_21 (a standing pool of water adjacent to gypsum deposits) and AW_RC_05_21 (gypsum-rich sediment) are distinguished by the presence of Firmicutes, Chloroflexota, and the Archaeal order Nitrososphaerales (Figure 4-4). The uncultured phylum KSB1 (Q. Li et al., 2022) is an abundant MAG within AW_RC_01_21, while the MAG community in AW_RC_05_21 is dominated by the JACCXJ01 group of Gammaproteobacteria and contains MAGs classified to the uncharacterized phylum CSP1-3 (Figure 4-4C; Hug et al., 2015). The remaining communities (CX_GO_08_21, BB_DV_03_21, PA_XL_04_20, CX_GO_13_21, GZ_MC_13_20, BS_MC_01_21, and FR_ND_02_21) are all sediments or mineral features (Table 4-1), exhibit relatively enriched $\delta^{13}C_{org}$ values (Figure 4-4A), and are marked by the co-occurrence of the phyla Methylomirabilota and Acidobacteriota (Figure 4-4B). The silty sediment CX_GO_08_21 stands apart, being dominated by two MAGs classified to the KSB1 phylum (Figure 4-4C). The phyla CSP1 -3 and KSB1 are both be absent from the Silva database (Figure 4-4C). Additionally, BS_MC_01_21, a black biovermiculation, contains two MAGs classified to the phylum Desulfobacterota (Online Supplemental Table 4-2). The metagenome TS_DV_02_21 contained significantly fewer 16S reads than the rarefied sampling depth and is excluded from the dendrogram. No quality MAGs could be assembled from this metagenome either, but we retain

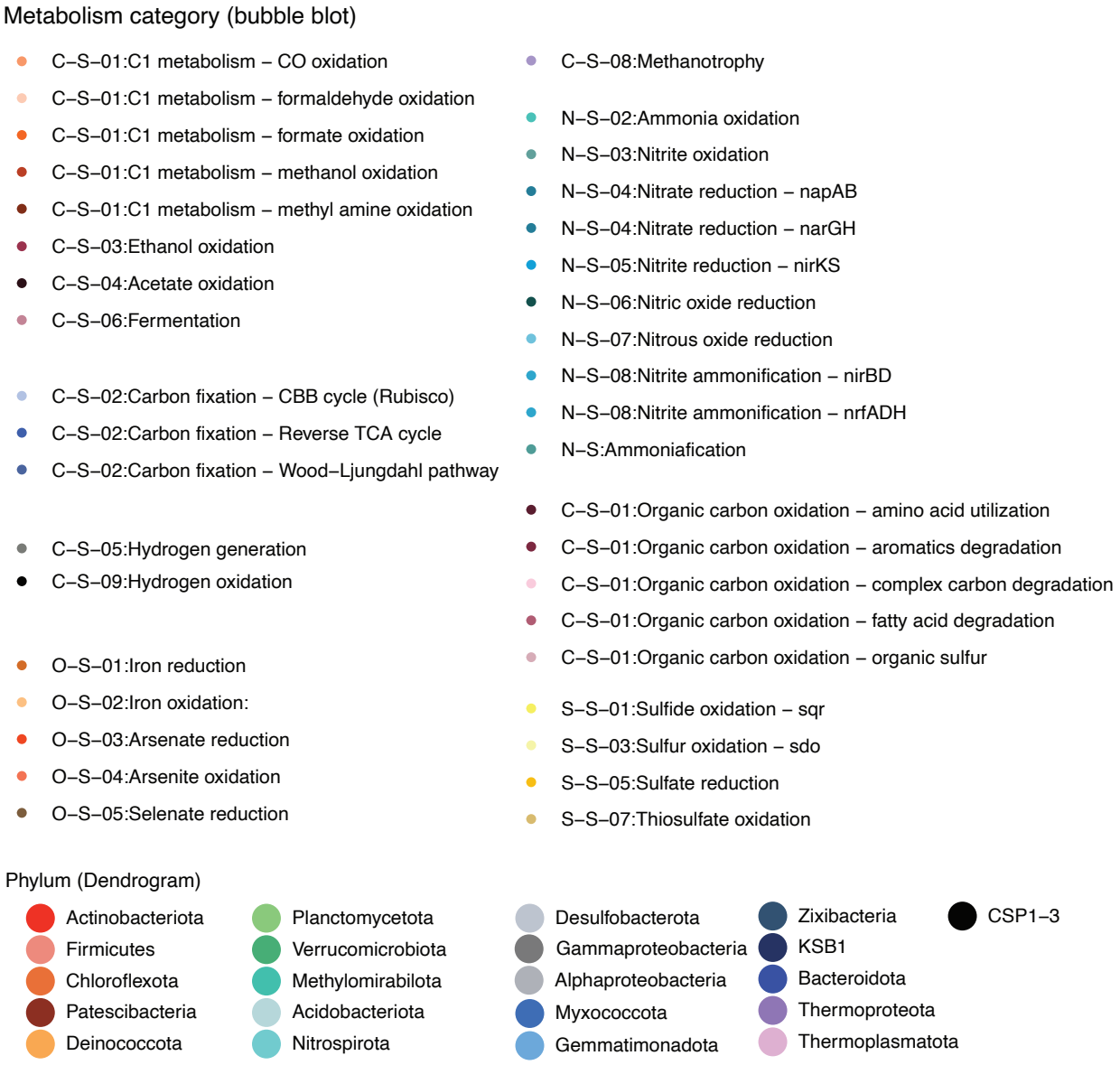
the unrarefied 16S community composition as a qualitative comparison to similar sediments (Figure 4-4B).

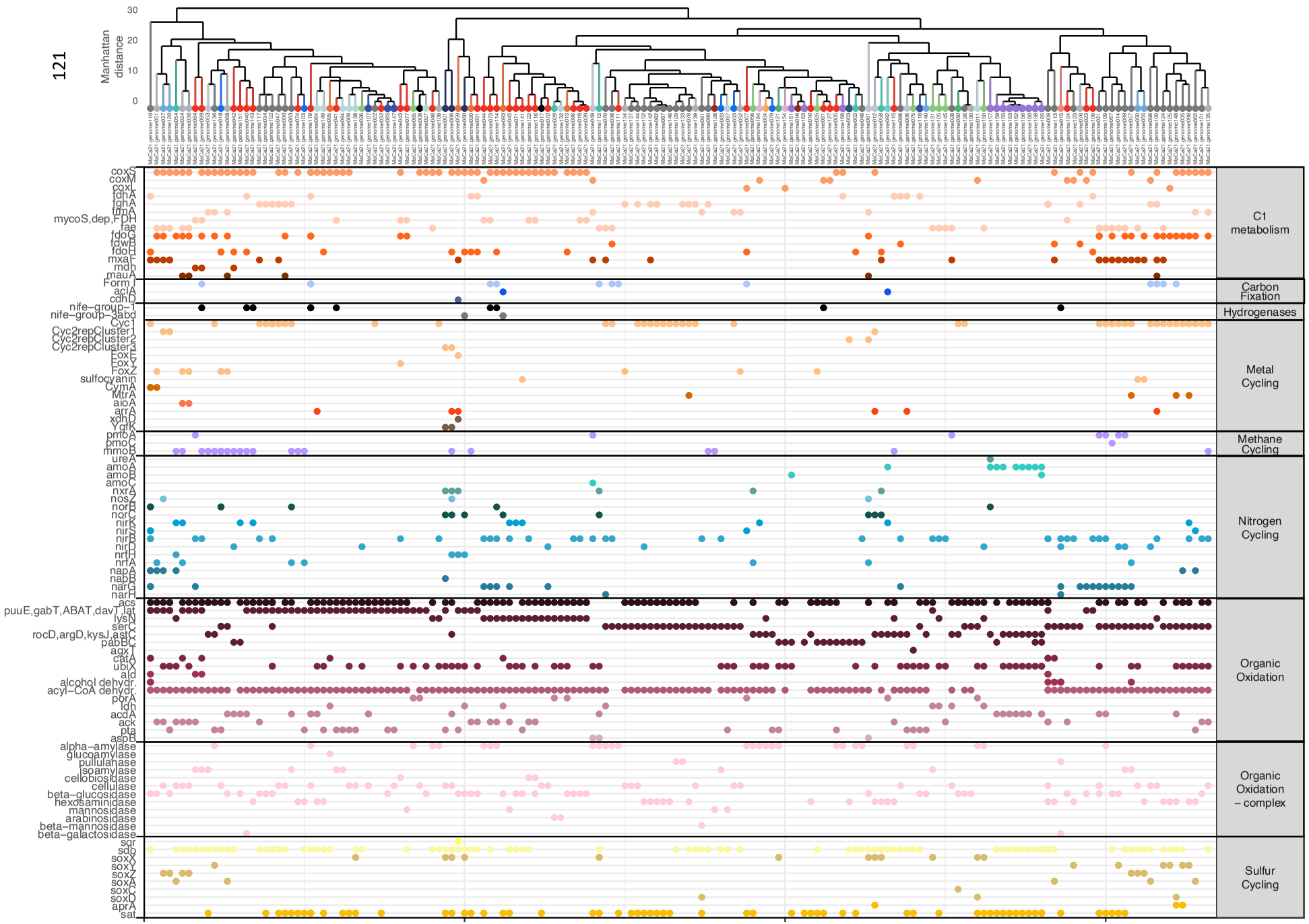
The biofilm communities EP_LC_04_21 and VH_MC_04_21 both originate from actively bat-impacted cave environments (Table 4-1) and form another subcluster, harboring relatively high proportions of reads mapped to Bacteroidota and the Gammaproteobacterial order Xanthomonadales (Figure 4-4B), in general agreement with MAG-based compositions (Figure 4-4C; Online Supplemental Table 4-2). Both samples have similarly high concentrations of ammonia, nitrate, sulfate, and nitrite (Online Supplemental Table 4-1).

The only metagenome obtained from an aqueous environment, AW_RC_01_21 is one of the most diverse communities in the dataset and contains Chloroflexota, Myxococcota, Methylomirabilota, Crenarchaeota (order *Nitrosphaerales*), in general agreement with MAG-based compositions, though we lack a Myxococcota MAG from here (Online Supplemental Table 4-2). With the exception of CX_GO_13_21, the remaining four metagenomes that form a cluster with AW_RC_01_21 are marked by a lack of Actinobacteriota (Figure 4-4B,C), while all four contain relatively high proportions of Acidobacteriota and Nitrospirota sequences in the 16S rRNA view of composition (Figure 4-4B). These metagenomes, namely PA_XL_04_20, VD_CC_01_21, CX_GO_13_21, and FR_ND_02_21, are all sediment or mineral features that contain relatively low (< 175 ppm) concentrations of sulfate and nitrate (Online Supplemental Table 4-1).

4.3.3. Functional genes associated with MAGs

Figure 4-5: ***Functional gene profiles of metagenome-assembled genomes (MAGs) from Mammoth Cave.*** Dereplicated MAGs are organized by complete hierarchical clustering performed on a Manhattan distance matrix calculated from a presence-absence matrix of functional gene profiles determined via METABOLIC (Online Supplemental Table 4-6). Gene annotations were performed on the redundant MAG set to enable direct read mapping. As such, a gene is determined to be present in a given dereplicated MAG if it is in at least one of its identified cohort.




121
Manhattan distance
30
20
10
0
coxS
coxM
coxL
fdhA
fghA
fmdA
mycoS,dep,FDH
fae
fdoG
fdwB
fdoH
mxaF
mdh
mauA
Form I
aclA
cdhD
nife-group-1
nife-group-3abd
Cyc1
Cyc2repCluster1
Cyc2repCluster2
Cyc2repCluster3
FoxE
FoxY
FoxZ
sulfocyanin
CymA
MtrA
aioA
arrA
xdhD
YgfK
pmoA
pmoC
mmoB
ureA
amoA
amoB
amoC
nxrA
nosZ
norB
norC
nirK
nirS
nirB
nirD
nrfH
nrfA
napA
napB
narG
narH
acs
puuE,gabT,ABAT,davT,lat
lysN
serC
rocD,argD,kysJ,astC
pabBC
agxT
catA
ubiX
ald
alcohol dehydr.
acyl-CoA dehydr.
porA
ldh
acdA
ack
pta
aspB
alpha-amylase
glucoamylase
pullulanase
isoamylase
cellobiosidase
cellulase
beta-glucosidase
hexosaminidase
mannosidase
arabinosidase
beta-mannosidase
beta-galactosidase
sqr
sdo
soxX
soxY
soxZ
soxA
soxC
soxD
aprA
sat
C1 metabolism
Carbon Fixation
Hydrogenases
Metal Cycling
Methane Cycling
Nitrogen Cycling
Organic Oxidation
Organic Oxidation – complex
Sulfur Cycling

To explore taxonomic patterns of functional gene distributions, we narrow our analysis to the presence or absence of genes assembled into MAGs (Figure 4-5). Such a MAG-centric view may limit our ability to detect genes from low-abundance or rare community members (Albright & Louca, 2023), but we posit that our MAG set is representative of the most abundant, and arguably most functionally significant, community members of these otherwise understudied microenvironments. We observe variable distributions in genetic functional potential based on both MAG phylogeny and the environment from which it was recovered (Figure 4-5; Online Supplemental Table 4-1). We focus on reporting the distribution of key taxonomically resolved genes involved in the carbon, nitrogen, and sulfur cycles.

4.3.3.1. Carbon fixation, hydrogen, and C_1 metabolism genes.

Genes involved in the Calvin Cycle, reverse TCA cycle, and Wood Ljungdahl pathway (Form I RuBisCO, *aclA*, and *cdhD*, respectively) are detected in 16 MAGs, with Form I RuBisCO being the most prevalent (13/16), followed by *aclA* (2/16), and *cdhD* (1/16). Form I RuBisCO is notably detected in six Actinobacteria MAGs (Figure 4-5) representing three orders (Acidimicrobiales, Euzebyales, and Mycobacteriales). Other MAGs containing Form I RuBisCO include the CPS1-5 clade of Methylomirabilota, as well as the Gammaproteobacterial orders Arenicellales, Burkholderiales, and Steroidobacterales (Figure 4-5). Five of the six Actinobacteria MAGs containing gene profiles for Form I RuBisCO also harbor genes for hydrogen oxidation, specifically *nife-group-1* (Figure 4-5; Berney et al., 2014; Greening et al., 2016). Alternative carbon fixation pathways are rarer; *aclA* (rTCA cycle) is only present in two MAGs from the Chloroflexota (cluster VGOW01) and Nitrospirota (f. *Nitrospiraceae*; Figure 4-5). The *cdhD* gene (Wood Ljungdahl pathway; (Maupin-Furlow & Ferry, 1996)) is only detected in a single Chloroflexota MAG (cluster A4b; Figure 4-5). In contrast to the relatively restricted distribution

of autotrophic genes, those involved in C_1 metabolism are much more widespread (Figure 4-5). The *coxS* gene, involved in carbon monoxide oxidation (Cunliffe, 2011; King, 2003; Moran et al., 2004), is the most prevalent in the MAG set (77/166 dereplicated MAGs), being most common in Actinobacteriota (n = 30) and Proteobacteria (n = 21). Genes related to methanotrophy include those encoding soluble and particulate methane monooxygenase (*mmoB* and *pmoA, pmoC*, respectively; Murrell et al., 2000) are detected in 28 MAGs, from Proteobacteria (n = 12), Actinobacteriota (n = 8), Desulfobacterota (n = 2), Myxococcota (n = 2), and single representatives of Chloroflexota, KSB1, Methylomirabilota, and Patescibacteria. We note that *pmoA* or *pmoC* are almost exclusively found in MAGs classified to the uncharacterized order of Gammaproteobacteria JACCXJ01 (Ortiz et al., 2021).

4.3.3.*2. Nitrogen cycling genes.*

Genes involved in the synthesis of ammonia monooxygenase (*amoA*, *amoB*, or *amoC*), genetic markers for ammonia oxidizing bacteria (AOB) and archaea (AOA; Rotthauwe et al., 1997), are only found in 11 MAGs (Figure 4-5), nine of which are classified as either *Nitrosopumilaceae* or *Nitrosophaeraceae* (Figure 4-5). Other than these families of AOA, one Desulfobacterota and one Nitrospirota MAG also contain copies of one of these gene sets. The nitrite oxidoreductase gene (*nxr*), a marker for nitrite oxidation, is only detected in six MAGs, from the phyla Chloroflexota (n = 1), KSB1 (n = 2), Methylomirabilota (n = 2), and Planctomycetota (n = 1).

By contrast, genes involved in reductive nitrogen metabolisms are more common across our MAG set (Figure 4-5). Various types of the nitrite reductase-encoding *nir* gene (*nirB*, *nirD, nirK,* and *nirS*; Philippot, 2002) are especially prevalent, at least one of these genes is present in 53/166 dereplicated MAGs representing nine phyla (Figure 4-5; Online Supplemental Table 4-6).

The *nirK* and *nirS* genes encode nitric oxide-producing nitrite reductase as part of canonical denitrification, while *nirB* and *nirD* are associated with nitrite reduction to ammonia as part of DNRA (Canfield et al., 2010). Genes encoding nitrate reductase (*narG*, *narH*; Canfield et al., 2010) are present in 25 MAGs, primarily Gammaproteobacteria (n = 11), Actinobacteriota (n = 8), and Alphaproteobacteria (n = 2), though these genes also appear in single MAG representatives of Firmicutes, Gemmatimonadota, KSB1, and Methylomirabilota. A total of 13 dereplicated MAGs contain *norB*, which encodes nitric oxide reductase (Canfield et al., 2010), including the phyla Thermoproteota (n=1), Actinobacteriota (n=1), Chloroflexota (n=1), Desulfobacterota (n=2), KSB1 (n=3), Methylomirabilota (n=2), Myxococcota (n=1), and Gammaproteobacteria (n=2). Nitrogenase (encoded by *nif*) is notably absent in all MAGs (Figure 4-5).

4.3.3.*3. Sulfur cycling genes.*

The sulfur dioxygenase gene (*sdo*), associated with sulfide oxidation (H. Liu et al., 2014), is one of the most ubiquitous in our MAG-resolved dataset, being present in 80/166 dereplicated MAGs (Figure 5). Genes involved in thiosulfate oxidation, namely those within the *sox* cluster (Meyer et al., 2007), were identified within 35 dereplicated MAGs, representing Proteobacteria (n = 17), Gemmatimoadota (n = 4), KSB1 (n = 3), Methylomirabilota (n = 3), Acidobacteriota (n = 2), Desulfobacterota (group B; n =2), Planctomycetota (n = 2), and Verrucomicrobiota (n = 2). A marker for dissimilatory sulfate reduction, the sulfate adenyltransferase gene (*sat*; Pereira et al., 2011) is also widespread in our MAG set (64/166), being especially prevalent in Proteobacteria (n = 19), Thermoproteota (n = 12), Actinobacteriota (n = 9), Planctomycetota (n = 5), and Acidobacteriota (n = 5).

4.4 Discussion

4.4.1 Cave biofilms depend on distinct sources of carbon and nitrogen.

Bulk organic carbon and stable nitrogen isotope data suggest that cave biofilms depend on a distinct source (or sources) of these elements compared to surface soils, sediments, and mineral features (Figure 4-2). Cave biofilms are consistently the most ^{13}C- and ^{15}N-depleted samples in both the full dataset (-31.7 $\pm$ 4.9‰ and 0.0 $\pm$ 3.4‰, respectively) and in those with metagenome representatives (mean $\delta^{13}C_{org}$ = -30.4 $\pm$ 4.0‰ and 2.8 $\pm$ 3.0‰, respectively). By contrast, cave sediments and mineral features exhibit mean $\delta^{13}C_{org}$ values of -25.1 $\pm$ 2.7 and -25.7 $\pm$ 1.9‰, and $\delta^{15}N_{bulk}$ values of +3.8 $\pm$ 2.13‰ and +3.1 $\pm$ 2.5‰, respectively (Figure 4-2). Given a mean $\delta^{13}C_{DOC}$ value of -25.6 $\pm$ 0.4‰ in cave pools, underground rivers, and percolating waters (Online Supplemental Table 4-1), these patterns are consistent with the bulk heterotrophic consumption of surface-derived biomass in such sediments (Hayes, 2001; Selensky et al., 2021). Biofilms are instead depleted in both $\delta^{13}C_{org}$ and $\delta^{15}N_{bulk}$ compared to mean surface soil and water DOC (Figure 4-2). Although we lack compound-specific isotope data to confirm the incorporation of ^{13}C-depleted carbon into living biomass, *in situ* carbon fixation via the Calvin Benson Bassham (CBB) cycle or Wood Ljungdahl (WL) pathway in biofilms could explain the depletions in biofilm $\delta^{13}C_{org}$ values compared to incoming cave water $\delta^{13}C_{DOC}$ and bulk surface soil $\delta^{13}C_{org}$ values (Hayes, 2001; Selensky et al., 2021). Based on the stable isotope data discussed above, we hypothesize that some bacteria and/or archaea within cave biofilms are either actively fixing inorganic carbon or are incorporating an alternative pool of ^{13}C-depleted organic carbon, such as methane (Barker & Fritz, 1981).

In the absence of sunlight, the oxidation of inorganic, redox-sensitive compounds could instead be an important source of energy to drive the endergonic process of carbon fixation forward

(Stevens, 1997). A comparison of $\delta^{15}N_{bulk}$ values, $\delta^{15}N_{NO3}$ values, and nitrogen reservoir concentration data (Figure 4-3) suggests the active cycling of nitrogen in many communities. Microbially mediated nitrogen cycling processes, such as ammonia oxidation, could be a major source of energy for cave chemolithoautotrophs in certain communities. In our dataset, there is a negative relationship between ammonia concentration, $[NH_4^+]$ (ppm), and the isotopic composition of nitrate (Figure 4-3). Microbial ammonia oxidation to nitrite strongly discriminates against $^{15}NH_4^+$ ($\varepsilon \sim$ -35‰; Mariotti et al., 1981) while nitrite oxidation to nitrate imparts a positive fractionation when nitrite accumulates (ε up to +15‰; Ryabenko, 2013).

As such, when the ammonia pool is net consumed by oxidative processes, the product nitrate pool is expected to become progressively depleted in ^{15}N compared to bulk nitrogen (Mariotti et al., 1981; Mooshammer et al., 2020; Sigman & Casciotti, 2001). We observe the largest such offset between $\delta^{15}N_{bulk}$ and $\delta^{15}N_{NO3}$ values in BP_LC_01_21, a biofilm located within an active bat hibernarium containing abundant guano deposits (Figure 4-3, Table 4-1). Two abundant MAGs classified to the ammonia-oxidizing family of Archaea, *Nitrososphaeraceae* (Tourna et al., 2011), were recovered from this biofilm (Figure 4C) and contain the ammonia monooxygenase gene, *amoA* (Rotthauwe et al., 1997). According to this paradigm, ammonia oxidation is also expected to be a dominant nitrogen cycle process in the sediment CX_GO_13_21 (Figure 3), which also hosts an abundant MAG with *amoA* classified to the family *Nitrosopumilaceae* (Figures 4-4, 4-5; Online Supplemental Tables 4-2, 4-6). We focus the remaining discussion on exploring relationships between bulk environmental measurements and community composition and functional potential inferred via metagenomics (Figures 4-4, 4-5).

4.4.2. MAGs associated with ammonia and nitrite oxidation are uncommon.

Our set of 166 dereplicated MAGs represent several microbial lineages (Figure 4-4) that are capable of a diverse array of biogeochemically relevant metabolisms (Figure 4-5). When tied to bulk geochemical and isotopic data from the cave environment (Figure 4-3), the estimated abundance and functional potentials of these MAGs reveal key insights about the potential carbon and energy sources exploited by the foundational members of various microbial niches within Mammoth Cave.

Genes involved in oxidative nitrogen cycle processes (*amo*, *nxr*) are relatively uncommon in our MAG set (Figure 4-5). Metagenome-derived 16S rRNA gene sequences suggest the ubiquitous presence, albeit low-abundance, of ammonia-oxidizing Archaea (o. Nitrososphaerales) and nitrite-oxidizing Bacteria (o. Nitrospirales) in most communities (Figure 4-4B). The *amoA* gene is present in 10 of the 12 MAGs classified to the AOA families *Nitrosomonadaceae* and *Nitrosopumilaceae* within this order (Figure 4-5). Although we observe 16S rRNA sequences and/or MAGs classified to these families in many metagenomes (Figure 4-4B,C), MAG quantification (Patro et al., 2017; Uritskiy et al., 2018) and the abundance of 16S rRNA gene tags mapped to these families (Quast et al., 2012; Xie et al., 2016) corroborate that these families are the most abundant in metagenome BP_LC_01_21 (Figure 4-4B,C), which was taken from a white biofilm on a cave wall hanging over guano-laden sediment within an active bat hibernarium (Table 4-1). The high abundance of these known ammonia oxidizing archaea (Qin et al., 2017; Tourna et al., 2011) in this metagenome also corresponds with the highest concentrations of ammonia (114 $\pm$ 9 ppm) in our dataset, alongside relatively high concentrations of nitrate, nitrite, sulfate (Online Supplemental Table 4-1), and qualitative field observations of a distinct ammonia-like odor. Intriguingly, these two Archaeal families appear to be ubiquitous, low-abundance members of

most communities based on 16S rRNA data (Figure 4-4B, Online Supplemental Table 4-4), save for the biofilms dominated by Actinobacteriota, Myxococcota, and Gemmatimamonadota (three left-most metagenomes; Figure 4-4B,C).

The near-ubiquity (albeit low abundance) of Nitrospirales in our 16S rRNA dataset (Figure 4-4B) recalls an early microbiological survey noting higher cell counts of Nitrospirota, a phylum harboring many known nitrite-oxidizing bacteria, in Mammoth Cave sediments compared to surface soils (Fliermans & Schmidt, 1977). Conflicting with the previously assumed ubiquity of this phylum, we only assembled one Nitrospirota MAG, from metagenome FR_ND_01_21 (Online Supplemental Table 4-2). Despite its singleton status, this MAG encodes a highly flexible metabolism, containing genes associated with the reverse TCA cycle (*aclA*), sulfur oxidation (*sdo*), nitrite reduction (*nirK*), and ammonia oxidation (*amoA*; Online Supplemental Table 4-6). Intriguingly, we did not observe the nitrite oxidoreductase gene (*nxr*) for nitrite oxidation in this MAG, though taxonomically resolved *nxr* was detected in six MAGs from various phyla in the metagenomes AW_RC_01_21, CX_GO_08_21, and PA_XL_04_20 (Online Supplemental Table 4-6).

Several *amoA*-containing MAGs mapped to the order Nitrososphaerales are also present in the same metagenomes (Figure 4-4C, 4-5). This suggests that some members of these communities can obtain energy from more reduced pools of nitrogen, thereby mediating nitrate accumulation. However, we lack $\delta^{15}N_{NO3}$ measurements to assess the relative balance of ammonia oxidation in these specific metagenomes. Further, whether oxidative nitrogen cycle processes in Mammoth Cave are mostly mediated by AOA such as Nitrososphaerales or bacteria such as Nitrospirota in communities lacking fresh guano inputs remains an open question. AOA have been found to be several hundred times more abundant than ammonia-oxidizing bacteria (AOB) in similar carbonate

cave sediments elsewhere, presumably because AOA can sustain growth with lower concentrations of ammonia (Zhao et al., 2017). Additionally, sediment metagenomes from another carbonate cave system associate *nxr* genes with Nitrospirota (Ortiz et al., 2014). The facts that the only recovered Nitrospirota MAG (Online Supplemental Table 4-2) lacks *nxr* (Online Supplemental Table 4-6), and the abundance of 16S rRNA sequences mapped to this phylum are generally low (Figure 4-4B), suggest that Nitrospirota spp. may be rarer cave community members in Mammoth Cave overall and may not be involved in nitrite oxidation as was previously assumed (Fliermans & Schmidt, 1977).

MAGs with genes related to reductive nitrogen cycle processes, namely *nir* (encoding nitrite reductase), *nar*, and/or *nap* (encoding nitrate reductase; Canfield et al., 2010), are much more abundant in our MAG set (Figure 4-5), especially in metagenomes without active guano input. We hypothesize that this reflects the major pools of N used by microbial communities in cave sediments overall. Although *amoA*, *amoB*, and/or *amoC* are present in MAGs from various sediment and mineral feature metagenomes (Online Supplemental Table 4-2), such MAGs are lower in abundance compared to MAGs containing *nir, nar*, and/or *nap*. This distribution of functional potential is consistent with the widespread input of oxidized nitrogen species, primarily nitrate, entering Mammoth Cave through waters percolating from the surface (W. Hess, 1900; Hill, 1981). Our geochemical measurements of cave waters (n = 40) demonstrate the near absence of ammonia in cave fluids (Online Supplemental Table 4-1). If bat guano were instead to be the major source of nitrogen throughout the cave system, we posit that MAGs with oxidative nitrogen metabolisms such *Thermoproteota* spp. should be more abundant. Instead, we speculate that because of the high concentrations of nitrate relative to ammonia in incoming cave waters, the relative abundance of key N-cycling taxa (Figure 4-4), and the observed patterns of nitrogen-based

functional potential (Figure 4-5), any oxidative nitrogen metabolisms are ultimately supported by the reducing power generated from a microbially mediated reductive nitrogen cycle outside areas with active guano deposition.

4.4.3. Actinobacteriota in cave biofilms exhibit highly flexible carbon metabolisms.

In contrast to the Thermoproteota and Nitrospirota, Actinobacteriota are both highly prevalent and abundant across different types of samples (Figure 4-4), consistent with previous surveys of this and similar cave systems (Barton & Northup, 2007; Lavoie, 2017). Actinobacteriota have been noted to be important community members in caves around the world, though much remains unknown regarding their metabolism (Barton & Northup, 2007; Cuezva et al., 2012; Farda et al., 2022; Gonzalez-Pimentel et al., 2018; Hathaway, Garcia, et al., 2014; Lavoie et al., 2017; Ortiz et al., 2013; Porca et al., 2012; Riquelme et al., 2015). Our MAG set thus provides a novel opportunity to probe the metabolic potential of this ecologically important yet enigmatic phylum of bacteria. Of the 36 dereplicated MAGs mapped to Actinobacteriota, 30 contain hits for the *coxS*, *coxM*, or *coxL* cluster of genes encoding for aerobic carbon monoxide dehydrogenase (Figure 4-5; Cunliffe, 2011; King, 2003; Moran et al., 2004). Further, 21 Actinobacteriota MAGs also contain at least one hit to other genes involved in the oxidation of other single-carbon compounds (*fdoH*, *fdoG*, *fdhA*, *mycoS/dep/FDH*, *mdh*, and/or *fae*; Figure 4-5), suggesting that C_1 metabolism could be an especially important carbon and energy source for this phylum in Mammoth Cave.

Six MAGs representing the Actinobacteriota are capable of carbon fixation via the CBB cycle based on the presence of Form I RuBisCO (Figure 4-5). These MAGs are classified to the families *Egibacteraceae* (n = 1; o. Euzebyales) and *Pseudonocardiaceae* (n = 3; o. Mycobacteriales) and are especially abundant in metagenomes PP_CC_04_21 and AW_RC_10_21 (Online Supplemental Table 4-2), which are both ^{13}C-depleted white biofilms

associated with active speleothem formation (Table 4-1). These dereplicated *Egibacteraceae* and *Pseudonocardiaceae* MAGs all contain the NiFe Group 1 hydrogenase gene (*nife-group-1*) required for hydrogen oxidation (Greening et al., 2016). This gene has been shown to support hydrogenotrophic, CBB-based autotrophy in cultured isolates from the family *Pseudonocardiaceae* (Grostern & Alvarez-Cohen, 2013) and is present in other groups of Actinobacteriota that instead fix carbon via the Wood-Ljungdahl pathway (Jiao et al., 2021). Like most of the Actinobacteriota, the *Pseudonocardiaceae* and *Egibacteraceae* MAGs are also capable of carbon monoxide oxidation indicated by the presence of the *coxS* gene (Figure 4-5). When considering potential inorganic electron acceptors for energy metabolism, the three *Pseudonocardiaceae* MAGs containing Form I RuBisCO also exhibit widespread potential for reductive nitrogen cycle processes as evidenced by the presence of the *nirB*, *narG*, and/or *norB* genes (Figure 4-5). By contrast, the *Egibacteraceae* MAG is instead capable of sulfate reduction via the *sat* gene (Figure 4-5).

4.4.4. Metagenomics provide novel insights into the microbial ecology of cave biofilms and sediments.

Of the putative autotrophs in our MAG set, only the Actinobacterial families *Egibacteraceae* and *Pseudonocardiaceae* encode *nife-group-1* (Figure 4-5). Each of the six dereplicated Proteobacteria MAGs with RuBisCO instead contain the *cyc1* gene associated with Fe(II) oxidation (Online Supplemental Table 4-6) and are found in mineral features or reddish cave sediment that are relatively ^{13}C-enriched (Table 4-1; Online Supplemental Table 4-2). We hypothesize that this reflects the major inorganic energy sources that are available to these microbial communities.

To our knowledge, we report the first draft genomes representing the *Egibacteraceae* and *Pseudonocardiaceae* from a subterranean environment (Online Supplemental Table 4-2). Previous taxonomic marker gene surveys demonstrate these families are highly abundant in biofilms found within basaltic and limestone caves around the world, though their genomic potential has remained unassessed (Cuezva et al., 2012; Farda et al., 2022; Gonzalez-Pimentel et al., 2018; Hathaway, Garcia, et al., 2014; Lavoie et al., 2017; Ortiz et al., 2013; Porca et al., 2012; Riquelme et al., 2015). We posit that the *Egibacteraceae* and *Pseudonocardiaceae* are cornerstone members of Mammoth Cave and other cave ecosystems worldwide due to their capacity for obtaining carbon and energy from diverse, yet scarce, sources (Figure 4-5; Online Supplemental Table 4-6).

The closest relative of the *Egibacteraceae* MAG as assessed by GTDB-tk (Online Supplemental Table 4-2) originates from an Antarctic desert soil metagenome in which photosynthesis is unlikely to occur (Ji et al., 2017). There, the oxidation of atmospheric concentrations of CO and H_2 gas were shown to sustain autotrophy in several clades, including relatives of the *Egibacteraceae* and *Pseudonocardiaceae* MAGs in our dataset (Ji et al., 2017; Ortiz et al., 2021). From those soils, the same sets of genes encoding enzymes for CO oxidation, H_2 oxidation, and CBB-based autotrophy observed in *Egibacteraceae* and *Pseudonocardiaceae* from Mammoth Cave (*coxS*, *nife-group-1*, *Form I*, respectively; Figure 4-5) were detected in the same families (Ji et al., 2017). The genes *coxS* and *nife-group-1* encode enzymes with high substrate affinity, facilitating biomass growth even at ambient atmospheric concentrations (Berney et al., 2014; Cordero et al., 2019; Greening et al., 2016).

We speculate that *Egibacteraceae* and *Pseudonocardiaceae* live as mixotrophs in cave biofilms, capable of exploiting hydrogen and carbon monoxide oxidation to fix carbon under limiting conditions (Figure 4-5). Bulk isotope data suggests the incorporation of ^{13}C-depleted

carbon into biomass from the biofilms AW_RC_10_21 and PP_CC_04_21 (Table 4-1; Online Supplemental Table 4-1). Bulk organic matter from these biofilms carry $\delta^{13}C_{org}$ values of -33.35 and -33.10‰, respectively (Table 4-1). Meanwhile, the mean $\delta^{13}C_{DOC}$ value of waters taken from drips, pools, and streams across Mammoth Cave is -25.60 $\pm$ 0.35‰ (n = 40; Supplemental Table 4-5), suggesting the input of significant amounts of ^{13}C-depleted carbon into these biofilms (Hayes, 2001). One plausible source is methane produced via methanogenesis, which strongly discriminates against ^{13}C (Barker & Fritz, 1981). Methanotrophs assimilating this carbon would then be expected to incorporate this depleted signal into their own biomass, driving $\delta^{13}C_{org}$ towards more negative values (Hayes, 2001). Methanotrophic MAGs are abundant in both communities; *pmoA*-containing MAGs classified to the uncharacterized clade of Gammaproteobacteria JACCXJ01 comprise approximately 53 and 22% of the MAG-based communities in PP_CC_04_21 and AW_RC_10_21 (Figure 4-5; Supplemental Table 8). Intriguingly, JACCXJ01 was first detected in the same Antarctic soils discussed previously, where it was determined to be an uncharacterized group of methanotrophic Chromatiales (Ji et al., 2017), another order commonly encountered in caves (Porca et al., 2012). *mmoB*-containing MAGs mapped to Myxococcota (f. *Haliangiaceae*) and Actinobacteriota (o. Solirubrobacterales) are also present in these communities, albeit at lower abundances (Figure 4-5; Online Supplemental Table 4-7).

In a sub-oxic microenvironment such as an established biofilm, the oxidation of C_1 compounds, including methane, can produce localized hydrogen gas (Jo et al., 2020; Nandi & Sengupta, 1998), thereby supplementing energy demands for putative I-fixing Actinobacteriota such as *Pseudonocardiaceae* and *Egibacteraceae*. In an oligotrophic cave environment, such a symbiosis could lessen overall competition for organic carbon substrates. There is precedent for these specific families fixing carbon in cave biofilms while other, co-localized microbial

populations assimilate carbon via heterotrophy (Lavoie et al., 2017; Selensky et al., 2021). Lipid-specific $\delta^{13}C_{org}$ values from Actinobacteriota-dominated biofilms in lava tubes demonstrate that 10,14-dimethyl pentadecanoic and 10-methyl hexadecanoic acids (10,14-DiMe $C_{17:0}$ and 10-Me $C_{17:0}$, respectively), diagnostic biomarkers for this phylum, are highly abundant and significantly ^{13}C-depleted relative to bioavailable pools of carbon and fatty acids attributed to other phyla (Selensky et al., 2021). Remarkably, 10,14-DiMe $C_{17:0}$ is only known to be produced by cultured members of the *Pseudonocardiaceae* (Reichert et al., 1998) and 10-Me $C_{17:0}$ is the major membrane fatty acid of the type *Egibacteraceae* strain (Zhang et al., 2016). Unraveling the wider distribution of and potential interactions between putative methanotrophs, hydrogen oxidizing bacteria, and autotrophs in these and other biofilms and sediments will further elucidate the mechanisms that allow the *Pseudonocardiaceae* and *Egibacteraceae* to be successful in caves worldwide. Ongoing analyses aim to determine whether these fatty acids are also present in Mammoth Cave, are ^{13}C-depleted, and co-occur with these families of Actinobacteriota to further constrain their unique metabolic lifestyles.

The biofilm GZ_MC_03_20 exhibits the most negative $\delta^{13}C_{org}$ value in our dataset (-37.27‰; Table 1) but lacks MAG-resolved genes associated with carbon fixation (Figure 4-5; Online Supplemental Table 4-2). These discrepancies should be considered in the context of the environment from which the metagenomes were sampled (Table 4-1). GZ_MC_03_20, a biofilm overlying a black mineral crust (Table 4-1), lacks populations of *Egibacteraceae* and *Pseudonocardiaceae*, instead harboring high numbers of putatively heterotrophic *Steroidobacteraceae*, Xanthomonadales, and *mmoB*-containing Myxococcota (Online Supplemental Table 4-2, 4-7). Although highly ^{13}C-depleted organic carbon could enter this community through *mmoB*-mediated methanotrophy (Hayes, 2001), we also note the distinctive abundance of taxa

such as Xanthomonadales and Burkholderiales in metagenomes associated with black coatings on cave surfaces. Given the community-level similarities between GZ_MC_03_20, the black coating VD_CC_01_21, and the bat-impacted metagenomes from Vespertilio Hall, we interpret this biofilm to represent a distinct niche from AW_RC_01_21 and PP_CC_04_21 (Table 4-1). Xanthomonadales in particular is associated with the degradation of complex organic matter (Lueders et al., 2006); indeed, six of the 12 MAGs associated with this order contain gene hits for hexosamidase, a chitin-degrading enzyme (Figure 4-5, Online Supplemental Table 4-2). Future work should aim to constrain the sources and sinks of carbon to microbial communities in these putatively manganese-rich black coatings (White et al., 2009) throughout the cave system.

Our analysis impresses the importance of metabolic flexibility and potential syntrophy in shaping cave microbiomes. These two traits are generally not associated with oligotrophic or "extreme" ecosystems, in which genomic streamlining/specification and inter-species competition are instead often discussed as shaping genome and microbial community structure, respectively (Giovannoni et al., 2005; Logue et al., 2012; Mende et al., 2017). The distributions of metabolic potentials in the diverse MAGs recovered from Mammoth Cave (Figures 4-4, 4-5) challenge this paradigm. Instead, we propose that when organic carbon is inherently limited due to a lack of local photosynthesis, taxa reliant on point-sources of reducing power, such as bat guano, are restricted in their spatial distribution. The Nitrosophaerales order of AOA fall under this framework in our analysis. By contrast, phyla such as Actinobacteriota are highly successful in caves because they gain or maintain diverse genomic potential for obtaining energy from non-point sources, such as ambient H_2 or CO gas. Other genes involved in substrate scavenging from this phylum have been previously demonstrated to undergo positive selection in oligotrophic conditions, supporting this notion (Props et al., 2019). Additionally, microbial syntrophic interactions in low-nutrient

environments can push otherwise endergonic metabolisms to become energetically favorable, including in H_2-driven subsurface environments (Lau et al., 2016; Morris et al., 2013). However, further phylogenomic and experimental evidence of the taxa described here is required. By considering the utilization of both point and non-point sources of electron donors, subsequent study of cave metagenomes can inform our understanding of ecosystem development as a whole.

4.5. Conclusions

Here, we explore the functional potential and community structure of microbial communities throughout various habitats within Mammoth Cave, KY, through the lens of metagenome-assembled genomes (MAGs). Bulk stable isotope analysis reveals that certain biofilms in the cave system rely on a distinct source of carbon and/or nitrogen compared to sediments and mineral features. Notably, the ammonia-oxidizing archaeal order Nitrososphaerales is especially abundant in a biofilm within an active bat hibernarium compared to other cave sediments. Our results lead us to speculate that the oxidation of nitrogenous compounds, such as ammonia, is volumetrically important only in cave areas with active bat populations. Further, to our knowledge, we assembled the first draft genomes representing the families *Egibacteraceae* and *Pseudonocardiaceae* from a subsurface setting. Based on previous 16S rRNA gene surveys, these Actinobacteriota are ubiquitous in biofilms found within both karst and volcanic caves around the world, and our analysis sheds the first light on their physiological and ecological potential. MAGs from these families are exclusively present in ^{13}C-depleted biofilms and exhibit remarkable metabolic flexibility, including hydrogen oxidation, carbon monoxide oxidation, and carbon fixation. These MAGs are also found in communities containing potential methanotrophs, namely an uncharacterized order of Chromatiales-like Gammaproteobacteria, JACCXJ01, and the

MAGs from the phylum Myxococcota. We conclude that taxa capable of obtaining energy from non-point sources are inherently successful in low-nutrient cave environments.

Chapter V: *Biologic Network Graph Analysis and Learning (BNGAL): A Novel Tool to Model Microbial Niche Space from Taxonomic Count Data*

5.1 Introduction

The advent of high-throughput taxonomic marker gene sequencing directly from the environment pioneered by Carl Woese and Norman Pace (Barns et al., 1996; Pace et al., 2012) has empowered microbial ecologists to query entire communities of bacteria and archaea, the majority of which cannot be cultured using standard laboratory methods (Vartoukian et al., 2010). The "gold standard" taxonomic marker gene for bacteria and archaea encodes ribosomal RNA (rRNA) from the small subunit (16S) of ribosomes, which foment polypeptide synthesis and are inherent to all biology on Earth. The widespread sequencing of so-called 16S rRNA genes from the environment has revolutionized microbiology, revealing a wealth of previously unknown microbial diversity with relevance to the fields of medicine, environmental health, biogeochemistry, and others (Cho, 2021; Wecke & Mascher, 2011).

A major challenge facing microbial ecologists today is to characterize communities of bacteria and archaea comprehensively and accurately. Environmental DNA sequencing has demonstrated that microbial communities across a range of habitats are highly complex, often encompassing thousands of unique individual species-level designations. Further, microbial "species" (defined by DNA sequence similarity cutoffs) interact with each other and their environment to establish ecological niches. Additionally, a single sequenced sample may harbor several separate microenvironments which can each support distinct niches. Considering all of this, holistic, systems-level analytical tools are critical to grasp the true complexity and biogeochemical potential of environmental microbial communities.

To fill this gap, network analysis has become more common in the field of microbial ecology to examine the potential relationships between microbial taxa and their environments (Dohlman & Shen, 2019; Faust et al., 2012; Layeghifard et al., 2017; Proulx et al., 2005; Steele et al., 2011). Despite the power of these approaches, there remains a lack of accessible network analysis software tools available for microbial ecologists to query both taxonomic interactions and environmental niche space from DNA sequencing data. Additionally, network analysis is prone to error and noise if the input data are not properly filtered or otherwise pre-processed (Faust et al., 2012; Goberna & Verdú, 2022). One approach is to filter out individual taxa below a certain abundance threshold, though this risks the removal of "rare" community members who may be ecologically relevant despite their low abundance. Another approach is to remove individual taxa that are observed a flat number of times in the dataset. However, without subsequent filtering, the latter risks the inclusion of potentially spurious relationships in the network due to the inherent sparsity of most taxonomic datasets (Deng et al., 2012; Faust, 2021; Matchado et al., 2021). More sophisticated models exist to overcome these limitations while reducing false positives (Faust et al., 2012), though their implementation may be inaccessible to non-experts.

Here, I introduce Biological Network Graph Analysis and Learning (BNGAL), novel open-source software written in the R programming language that constructs high-quality networks from microbial taxonomic abundance data to explore relationships between microbial taxa and their environment. BNGAL filters input data based on a threshold representing the number of times a given *pair* of taxa are observed, not the prevalence of a single taxon. This approach inherently excludes possibly spurious relationships arising from data sparsity. To illustrate, given a simple taxonomic presence-absence table as an example (Table 5-1), if the so-called "observational threshold" is set to 3, then correlations will only be inferred from relationships with greater than

three observations. Under this guideline, the Taxon1-Taxon2 and Taxon1-Taxon3 relationships are conserved, but Taxon2-Taxon3 are not, as only one relationship is observed. Without additional processing, the taxonomic prevalence and/or abundance-based filtering methods described above would include this suspect relationship, likely lowering the quality of the output network. In this way, BNGAL attempts to maximize the number of high-quality relationships while minimizing potential statistical noise included in the final network.

Table 5-1: Example presence-absence matrix of *N* unique taxa across five 16*S* rRNA gene libraries.

TagLibrary	Taxon1	Taxon2	Taxon3	…	Taxon*N*
Library1	1	1	1	-	1
Library2	1	1	0	-	1
Library3	1	1	0	-	1
Library4	1	0	1	-	1
Library5	1	0	1	-	1

Based on co-occurrence patterns of the filtered relationships, BNGAL categorizes groups of taxa (and environmental variables, if provided by the user) into subnetworks that represent individual niches. The abundance of taxa from each subnetwork is also mapped back to each sample to explore potential biogeographic patterns of niche distribution across habitats within a given dataset. The mission of BNGAL is to make network analysis and visualization more accessible to microbiome researchers analyzing taxonomic marker gene (e.g., 16S rRNA) count data. In the following sections, I describe the backend functions within BNGAL and apply the pipeline to a novel 16S rRNA gene dataset of Hawaiian lava tube communities to demonstrate the utility of modeling and visualizing taxonomic abundance data in this manner.

5.2 Software design and description

BNGAL is written in the R programming language and is available as both a standalone R package ("bngal") and a command line utility ("bngal-cli"). The R package comprises several functions to filter, reformat, summarize, and visualize networks constructed from taxonomic count data at various levels of classification. For ease of use, bngal-cli contains two executable wrapper scripts that organize the appropriate bngal functions to into pipelines to construct, visualize, and summarize relevant data from co-occurrence networks. In this way, the primary purpose of the bngal R package is to essentially serve as the backend to the more accessible bngal-cli interface. Descriptions of the bngal functions included in each bngal-cli executable pipeline (bngal-build-nets and bngal-summarize-nets) are discussed in the following subsections.

5.2.1. Input data filtering and network construction

The first bngal-cli wrapper (bngal-build-nets) filters and formats input data for network construction, constructs co-occurrence networks at a specified level of classification, and writes processed network data as well as static and interactive network visualizations (PDF and HTML formats, respectively) to a specified output directory. The following are input arguments for bngal-build-nets (required arguments are marked with an asterisk): asv_table* (taxonomic count table, can be collapsed at any level of classification), metadata* (sample metadata), taxonomic_level (defaults to "ASV"-level), output (output directory), correlation (type of correlation to calculate; "spearman" or "pearson" accepted), corr_columns (metadata column(s) to include in pairwise comparisons), corr_cutoff (absolute correlation coefficient cutoff for pairwise comparisons; default = 0.6), p_value (*p*-value cutoff; default = 0.05), abun_cutoff (relative abundance cutoff by which to exclude taxa from network; default = 0), cores (number of CPUs), subnetworks (metadata column by which to split input data to create separate networks), transformation (numeric

transformation to apply to input data; default = NULL), direction (direction for abun_cutoff; can be 'greaterThan' or 'lessThan'; default = 'greaterThan'), sign (type of pairwise relationship; can be one of 'positive', 'negative', or 'all'; default = 'all'), obs_threshold (minimum number of unique observations required for a given pairwise relationship; default = 5), graph_layout (type of igraph layout for network visualizations; default = 'layout_nicely'). Given this long list of options, I stress that BNGAL essentially runs in two modes that determined by the –subnetworks flag. If –subnetworks is absent, a "global" network of the entire input matrix will be constructed. If instead –subnetworks is set to the name of a column of categorical data in the provided metadata, then separate networks will be constructed based on those categories. This affords BNGAL a high level of flexibility in terms of input data and research questions that can be asked from its use.

The first bngal function for bngal-build-nets is bin_taxonomy, which takes a taxonomic count table and summarizes abundances at a user-specified level of classification for network construction. Taxa above or below an optional abundance cutoff will be removed at this step if the –abun_cutoff and –direction arguments are provided. Next, prepare_network_data reformats the binned taxonomic abundance data for downstream functions. If the –subnetwork argument is provided, data are split into each unique category so separate networks are built downstream. Its output then feeds prepare_corr_data, which additionally filters pairwise taxonomic relationships below the observational threshold, the number of times a pairwise relationship must be observed in the prepared dataset to be included (default = 5). The prepared taxonomic data, in addition to any user-provided numeric metadata passed through the –corr_columns argument, are normalized on a scale of 0-1 before correlation matrix construction. The function corr_matrix is a wrapper around rcorr from the Hmisc R package (https://hbiostat.org/R/Hmisc), which computes Spearman or Pearson correlation coefficients from the output of prepare_corr_data.

Afterwards, all possible nodes and edges from the prepared data are identified with the functions get_node_ids and generate_edges, respectively. The last filtering function in the bngal-build-nets pipeline, prepare_net_features, returns a finalized list of nodes and edges for each network filtered by p-value and correlation coefficient strength cutoffs, as well as the direction of each relationship (positive, negative, or all). The pw_summary function then writes pairwise summary statistics for the specified taxonomic level to an output directory. This function primarily summarizes the number of network nodes and edges that are present in each sample after each filtering step. The bngal functions get_igraph, get_edge_betweenness, and get_ebc_member_ids wrap external functions from the igraph software package to construct the initial network from the quality-controlled data, calculate edge betweenness centrality for each network node, and return a dataframe of nodes mapped to their edge betweenness clusters (EBCs; Girvan & Newman, 2002), respectively. The color_nodes function then creates a node color scheme for visualization by 1) node phylum, 2) node EBC membership, and 3) taxonomic functional grouping from a curated, non-comprehensive list of bacterial and archaeal families. Finally, the plot_networks function produces and exports static (PDF) and interactive (HTML) scientific publication-quality network visualizations in PDF and HTML formats. The final step of bngal-build-nets is export_network_data, which exports node and edge data for a defined level of taxonomic classification to the given output directory as RDS files.

5.2.2. Network statistics summarization and biogeographic visualization

The second and final bngal-cli wrapper, bngal-summarize-nets, summarizes and produces more useful visualizations from the bngal-build-nets output. The following are input arguments for bngal-summarize-nets (required options are marked with an asterisk*): asv_table*; metadata*; network_dir* (input network data; equivalent to bngal-build-nets output folder); taxonomic_level;

subnetworks; fill_ebc_by (metadata column by which to fill EBC composition plots); interactive (whether or not output plots are interactive HTMLs [TRUE] or static PDFs [FALSE]; default = FALSE); cores (number of CPUs); output (optional separate output directory).

The first bngal function in the bngal-summarize-nets pipeline is load_network_data, which imports the network data outputs written from bngal-build-nets. After reformatting taxonomic abundance data at a given classification level with bin_taxonomy, the function extract_node_data returns all nodes from the imported data for downstream manipulation. Next, get_alpha.div() calls the diversity() function from the vegan R package (Oksanen et al., 2019) to return the Shannon index value for each sample at each level of classification to estimate alpha diversity. The ebc_compositions() function then calculates the EBC composition of each sample for each level of classification and joins alpha diversity results for downstream functions. Biogeographic patterns of taxonomic abundance within a given EBC are visualized and explored by the plot_core_comp function, which exports a plot summarizing the relative contribution of each categorical variable from a user-provided metadata column (through the –fill_ebc_by argument) to each EBC alongside the coreness of each node. Next, export_ebc_taxa_summary() exports three CSV files that summarize 1) the abundance of each EBC per sample; 2) the spread of the relative abundance of each taxon, their coreness value, and EBC membership; and 3) the prevalence of each EBC across samples for a given network. To prepare for the summary visualizations, build_dendrograms() constructs hierarchical clusters (i.e., dendrograms) for samples included in a given network at a specified level of taxonomic classification. Finally, build_taxa.barplot() takes the outputs from ebc_compositions() and build_dendrograms() and creates clustered taxonomic barplots at a given level of classification. Dendrogram ends can be colored by a user-defined metadata variable for qualitative exploration of metadata contribution to community variability

through the –fill_ebc_by argument. By default, bngal-summarize-nets creates separate clustered barplots filled by phylum and EBC membership and exports publication-quality PDF figures to the defined output directory.

The last outputs of bngal-summarize-nets originate from the summarize_cooccurrence() backend function, which calculates a number of statistics for each network node that is exported to the file "connectivity_plots/[x]_connections_data.csv", where [x] is the provided taxonomic level. With each network node as a row, this file contains the following columns of data for the user (Table 5-2).

With the optional –query argument, users may also provide a space-delimited list of individual taxonomic names environmental variables of interest. This allows for the exploration of more granular co-occurrence patterns across subnetworks or within the global network, depending on whether the if –subnetworks option is provided.

Table 5-2: Column descriptions of the "connections_data.csv" output file for a given taxonomic level of classification.

Column name	Description
`tot_xions`	The total number of direct connections (primary edges) associated with the given node.
`pos_xions`	The number of positive primary edges.
`inter_ebc_xions`	The number of primary edges connecting the node to a different EBC subcluster.
`inter_phy_xions`	The number of primary edges connecting the node to a different phylum.
`unique_inter_phy_xions`	The number of unique phyla connected by primary edges.
`unique_inter_ebc_xions`	The number of unique EBCs connected by primary edges.
`n_taxa_per_phy`	The number of taxa associated with the given node's phylum in the network.
`total_xions_k2`	The summed number of primary edges from all nodes directly connected to the given node (secondary edges)
`inter_ebc_k2`	The number of secondary edges that connect to a different EBC subcluster.
`inter_phy_k2`	The number of secondary edges that connect to a different phylum.

5.2.3. Distribution and installation

The backend BNGAL R package (bngal) and its associated command-line utility (bngal-cli) are both publicly available as GitHub repositories (https://github.com/mselensky/bngal and https://github.com/mselensky/bngal-cli, respectively). The Anaconda package manager is required to be installed separately to manage the several R package and system-level dependencies called by bngal. The automatic bash setup script will install bngal from the cloned repository as summarized below and has been tested on the Linux (Red Hat) and MacOS (Darwin) operating systems.

```
$ git clone https://github.com/mselensky/bngal-cli
$ cd bngal-cli
$ bash bngal-setup.sh
```

5.2.4. Description of example 16S rRNA gene dataset

BNGAL is applied on the output taxonomically resolved (collapsed) count tables from standard 16S rRNA gene analysis pipelines such as Qiime2 (Bolyen et al., 2019). As such, only a brief and incomplete description how the taxonomic counts are generated here is provided to focus on the demonstration of general use cases with BNGAL. A total of 72 16S rRNA gene libraries representing 24 samples (three replicates per site) from three undisclosed lava caves in Hawaii were provided by collaborators (Diana Northup and Jennifer Hathaway, University of New Mexico), who also sent corresponding sample splits for stable isotope and lipid analysis (Table 5-3). Since the libraries were amplified with the 515F/806R primers (Caporaso et al., 2012), the same Qiime2- and Silva (v. 138)-based processing pipeline (Bolyen et al., 2019; Quast et al., 2012) as described in Chapter III was performed on this dataset. The final collapsed taxonomic count table was rarefied to a depth of 10,000 using the vegan package in R (Online Supplemental Table 5-1; Oksanen et al., 2019). The corresponding sample splits have also been analyzed for the stable isotope compositions of organic carbon ($\delta^{13}C_{org}$) and bulk nitrogen ($\delta^{15}N_{bulk}$) following the methods described in Chapters II and IV. These values, alongside percent C and N, represent the numeric environmental variables (Table 5-3) in the networks discussed below. Categorical variables include cave (“KK”, “MA”, “AK”) and sample type (“biofilm”, “ooze”, “mineral crust”, “coralloids”, “moonmilk”, and “spheroids”; Table 5-3).

Table 5-3: Sample metadata from Hawaiian lava tubes provided by collaborators.

Sample name	Cave	Sample type	Total N (%)	Organic C (%)	$\delta^{15}N_{bulk}$ (‰)	$\delta^{13}C_{org}$ (‰)
HLT-4	MA	ooze	0.4	2.1	-0.1	-36.2
HLT-24	KK	biofilm	0.0	0.1	-4.8	-35.7
HLT-20	KK	ooze	4.9	27.7	1.7	-35.5
HLT-23	MA	biofilm	0.2	1.0	-1.8	-34.2
HLT-17	KK	biofilm	0.0	0.2	-6.0	-33.9
HLT-1	MA	mineral crust	0.0	0.1	-3.0	-33.6
HLT-18	MA	biofilm	0.1	0.6	-3.7	-33.5
HLT-2	AK	biofilm	0.0	0.1	-3.5	-32.7
HLT-3	MA	moonmilk	0.4	2.6	-11.8	-32.4
HLT-8	KK	moonmilk	0.0	0.2	-8.9	-31.8
HLT-15	AK	moonmilk	0.4	2.7	-6.3	-31.6
HLT-9	KK	mineral crust	0.0	0.1	-2.2	-29.9
HLT-22	KK	spheroids	0.1	0.8	-2.3	-29.8
HLT-10	AK	mineral crust	0.0	0.0	-1.2	-29.6
HLT-25	AK	biofilm	0.0	0.1	-1.0	-29.5
HLT-13	AK	spheroids	0.2	0.9	-1.2	-28.9
HLT-11	MA	coralloids	0.3	2.0	-4.3	-28.2
HLT-16	KK	mineral crust	0.0	0.1	-1.5	-28
HLT-5	AK	coralloids	0.0	0.2	2.0	-26.5
HLT-6	MA	mineral crust	0.1	1.2	1.8	-25.5
HLT-12	AK	coralloids	0.1	3.4	1.0	2.2
HLT-21	KK	coralloids	0.1	2.5	-3.0	-
HLT-19	AK	mineral crust	0.0	0.3	-1.9	-
HLT-7	KK	spheroids	0.1	2.4	-2.8	-
HLT-14	MA	spheroids	0.2	3.9	-1.9	-

5.3 Example applications of BNGAL

To demonstrate the flexibility and utility of BNGAL, I apply its two main modes (determined by the --subnetworks flag) to the example Hawaii 16S rRNA dataset (Online Supplemental Table 5-1) and associated numeric metadata (Table 5-3). I first perform a holistic co-occurrence analysis across the entire dataset to showcase how visualizing the abundance of taxa within a given EBC can reveal biogeographic patterns. Guided by results from the "global" network, I then perform a more granular co-occurrence analysis by examining separate networks by sample type (Table 5-2).

Figure 5-1: ***BNGAL output visualizing global community composition clustered at the ASV level, summarized by EBC membership.***

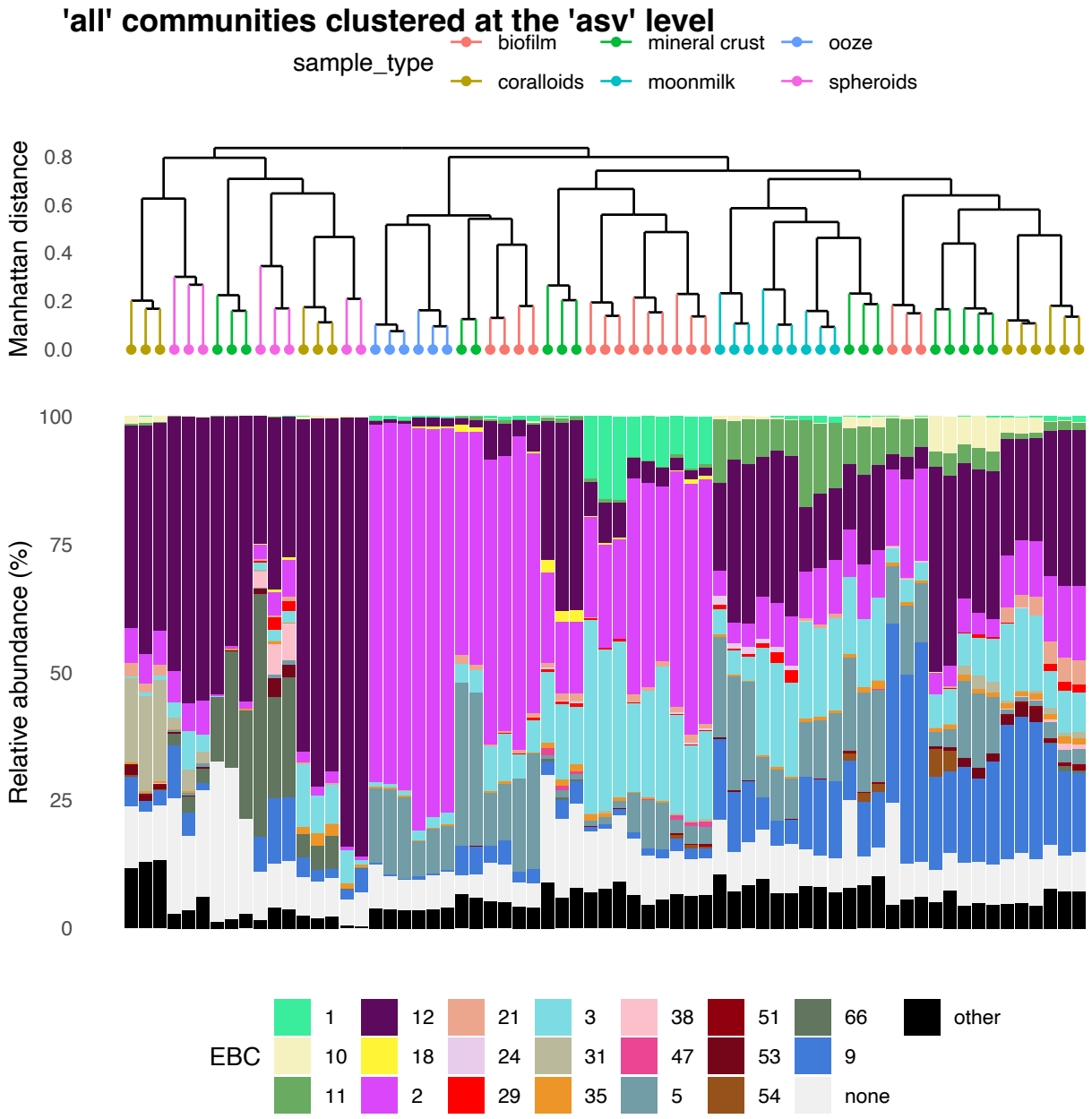
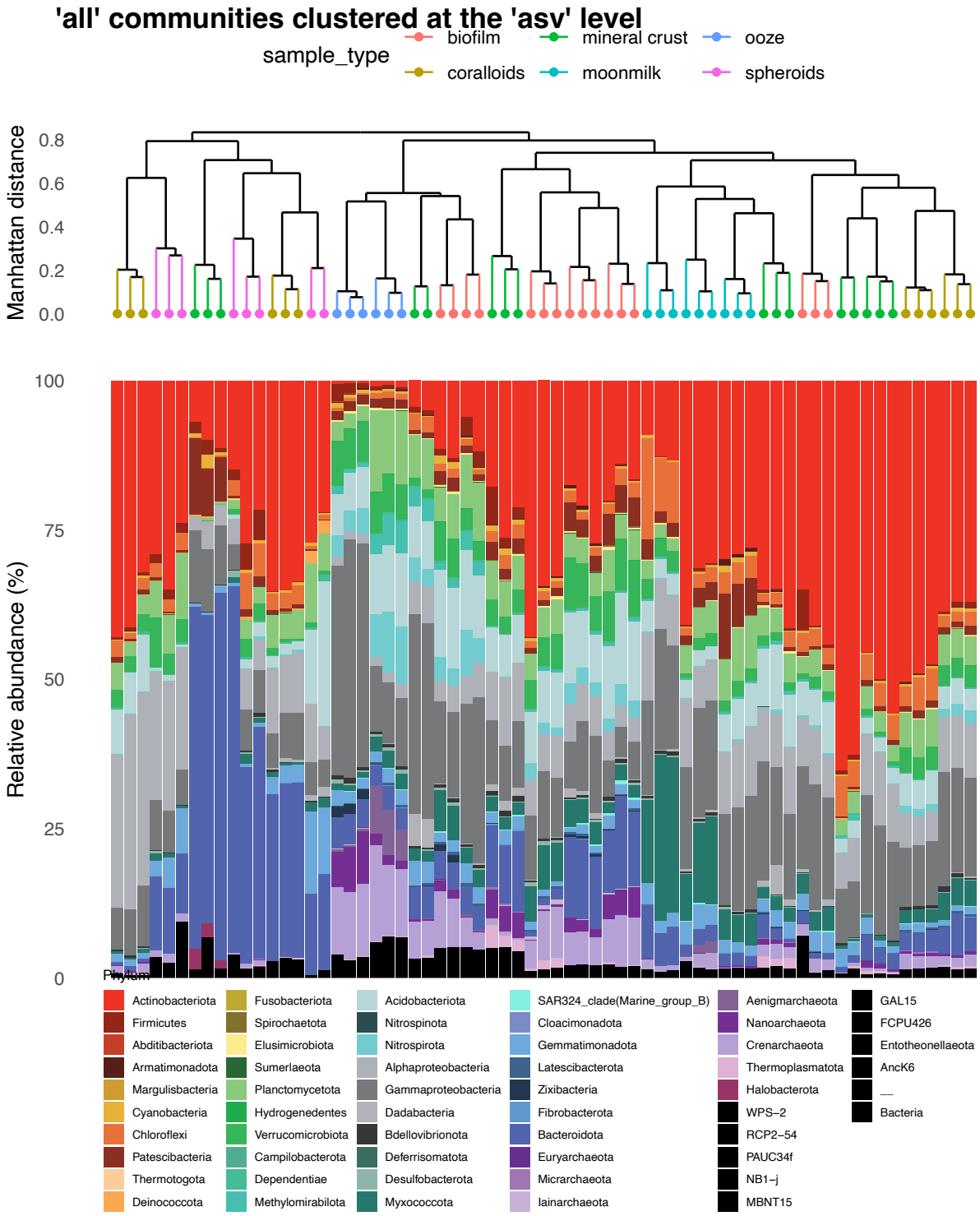


Figure 5-2: ***BNGAL output visualizing global community composition clustered at the ASV level, summarized by phylum.***



The global ASV-level hierarchical cluster output from BNGAL demonstrates some community similarity among similar sample types (dendrograms; Figures 5-1 and 5-2). Clear distributional patterns of co-occurring taxa can be observed when summarizing community structure by EBC membership (Figure 5-1). Such trends are much less evident when considering abundance by taxonomic phylum (Figure 5-2), highlighting the utility of BNGAL's implementation of edge betweenness clustering in exploring community structure (Csardi & Nepusz, 2006; Girvan & Newman, 2002). EBCs are interpreted as potential niches when applied to taxonomic count data (Girvan & Newman, 2002). It may be therefore reasonable to assume that the taxa comprising EBCs 1 and 12 are adapted to different environments, with the former being most abundant in biofilms while the latter is most abundant elsewhere (Figure 5-1). A visualization of the resultant network corroborates that taxa from these two EBCs do not co-occur, positively or negatively (Figure 5-3).

Given the evidence for niche partition among sample type (Figures 5-1 and 5-3), separate networks by sample type were then constructed to explore whether co-occurrence patterns within a sample type change by cave (Table 5-3; networks not shown). From this output, the host cave does not appear to drive the co-occurrence patterns observed in biofilms, but it may play a larger role for communities living in mineral crusts (Figure 5-4).

Figure 5-3: ***BNGAL output visualizing a global co-occurrence network.*** Nodes are colored by EBC membership, while edges are colored by the direction of the co-occurrence relationship (blue = positive, red = negative).

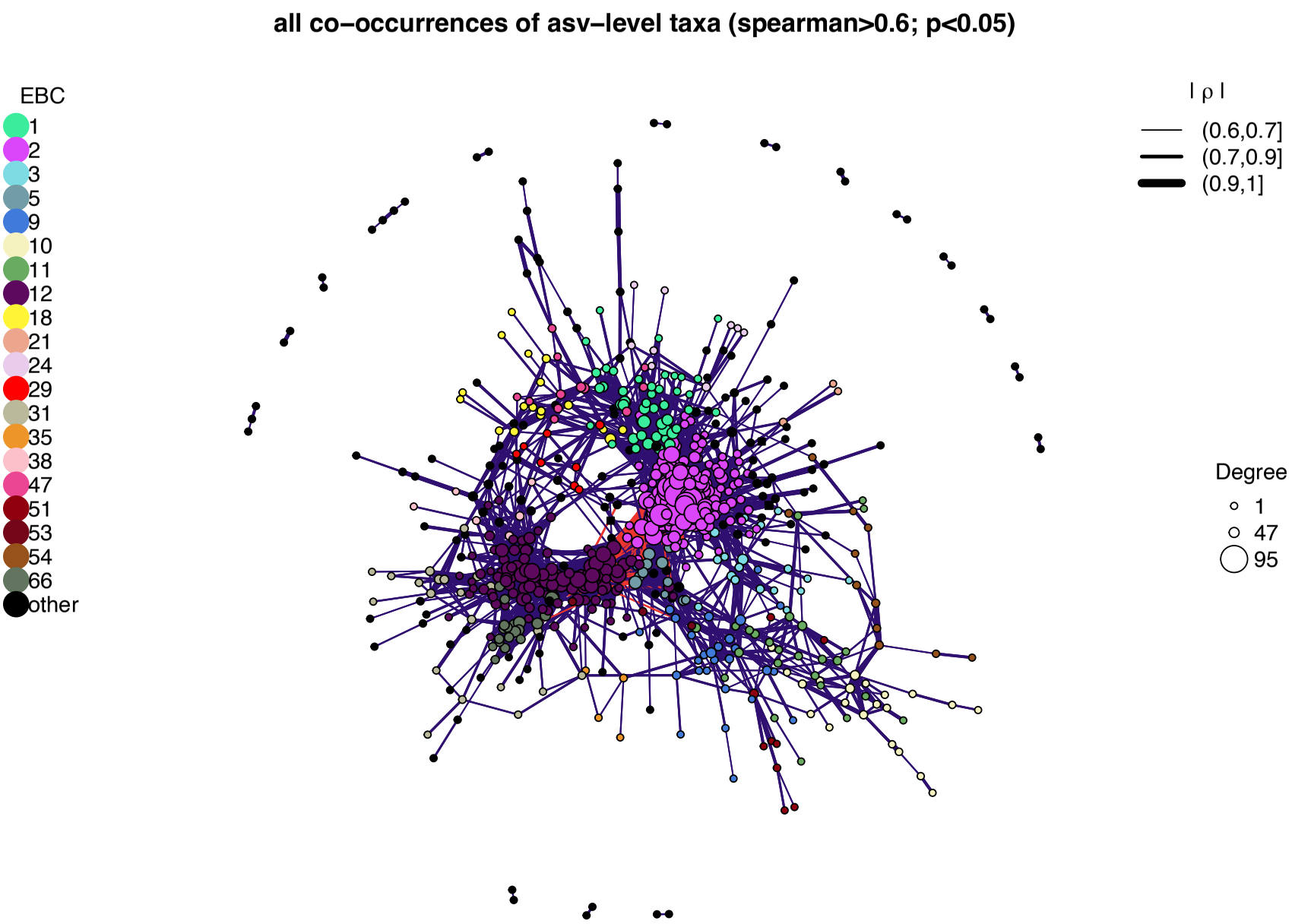
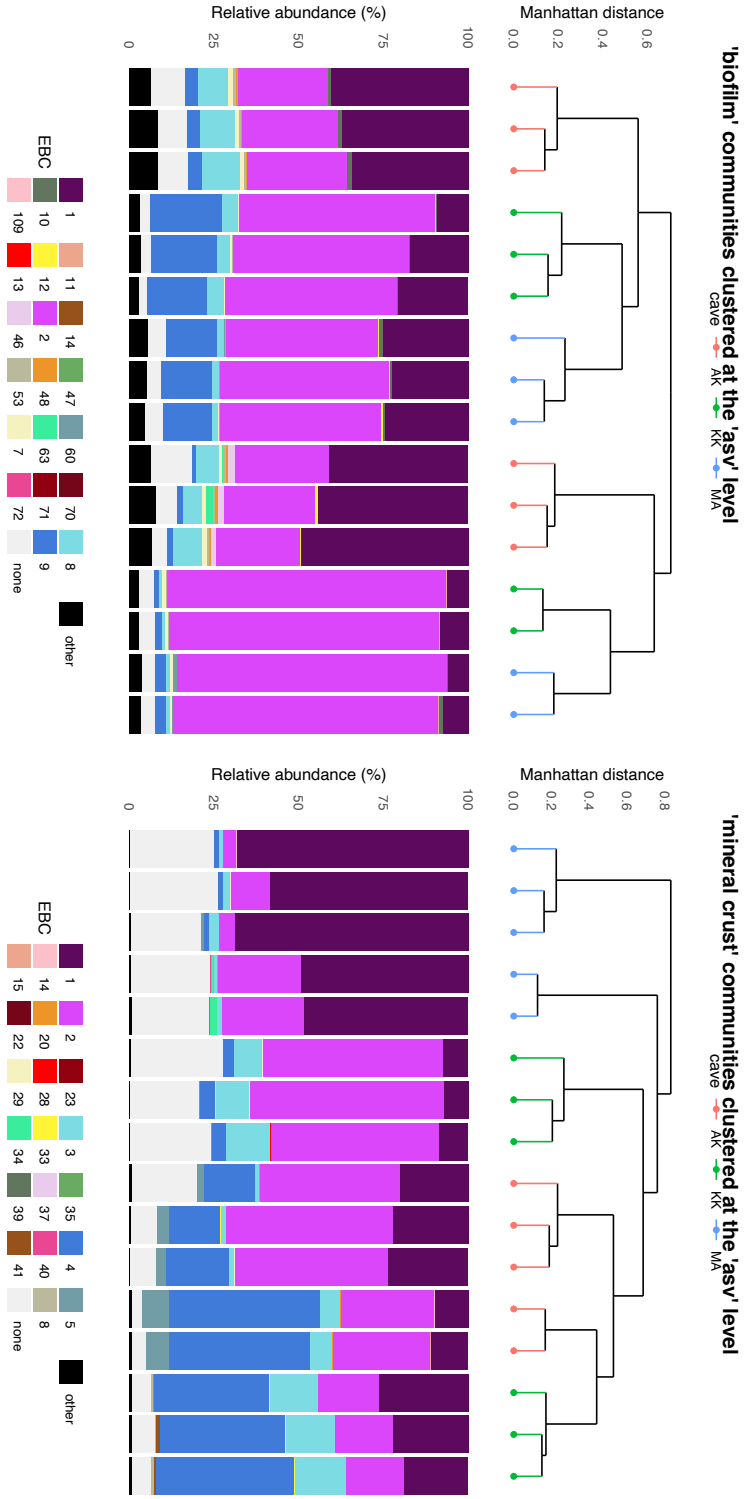

all co−occurrences of asv−level taxa (spearman>0.6; p<0.05)
EBC
1
2
3
5
9
10
11
12
18
21
24
29
31
35
38
47
51
53
54
66
other
| ρ |
(0.6,0.7]
(0.7,0.9]
(0.9,1]
Degree
1
47
95

Figure 5-4: ***BNGAL output visualizing community composition by the "biofilm" and "mineral crust" sample types clustered at the ASV level, summarized by EBC membership.*** Note that EBC membership is specific to each network and should not be cross compared.



Upon examination of the taxa_spread.csv file in the network_summaries output folder, the three most abundant taxa in the biofilm EBC 1 are an uncultured clade of Crossiella (f. *Pseudonocardiaceae*), an unclassified *Euzebyaceae*, and an uncharacterized Gammaproteobacteria, PLTA13. Given the ubiquitous presence of the Actinobacteriota families in these and other biofilms (Lavoie et al., 2017; Selensky et al., 2021), co-occurrence analysis was performed on the unclassified *Euzebyaceae* to explore its relationships with other taxa and environmental variables across other sample types (Appendix Figure 5-1), showing a high amount of variability by sample type. Further investigation into the correlational and/or distributional patterns of this and other taxa with BNGAL will elucidate the broader ecology of this sample set.

To demonstrate these two typical BNGAL use cases, the bash script to produce the global ASV-level network with bngal-cli, as well as those by sample type, is provided below:

```
# activate bngal conda environment
conda activate bngal
# build global network
asv_table=rarefied-asv-table.csv
meta_data=sample-metadata.csv
out_dr=`pwd`/global-network
mkdir -p $out_dr

bngal-build-nets \
  --asv_table=$asv_table \
  --metadata=$meta_data \
  --output=${out_dr} \
  --taxonomic_level="asv" \
  --obs_threshold=5 \
  --corr_columns="d15N_AIR,d13C_VPDB,pct_C,pct_N" \
  --graph_layout="layout_nicely"
bngal-summarize-nets  \
  --asv_table=$asv_table \
  --metadata=$meta_data \
  --network_dir=$out_dr \
  --output=${out_dr} \
  --taxonomic_level="asv" \
  --cores=1 \
  --fill_ebc_by="sample_type" \
  --interactive=FALSE
```

```
# build separate networks by sample type in another output directory
# bolded text indicates metadata column by which data are separated
# into separate networks
out_dr=`pwd`/sampleType-network
mkdir -p $out_dr

bngal-build-nets \
  --asv_table=$asv_table \
  --metadata=$meta_data \
  --output=$out_dr \
  --taxonomic_level="asv" \
  --obs_threshold=5 \
  --subnetworks="sample_type" \
  --corr_columns="d15N_AIR,d13C_VPDB,pct_C,pct_N" \
  --graph_layout=$GRAPH_LO
bngal-summarize-nets  \
  --asv_table=$asv_table \
  --metadata=$meta_data \
  --network_dir=$out_dr \
  --output=${out_dr} \
  --taxonomic_level="asv" \
  --cores=1 \
  --subnetworks="sample_type" \
  --fill_ebc_by="cave" \
  --interactive=FALSE

# re-run summarize-nets to produce node-specific co-occurrence plot of
queried nodes across networks
bngal-summarize-nets  \
  --asv_table=$asv_table \
  --metadata=$meta_data \
  --network_dir=$out_dr \
  --output=${out_dr} \
  --taxonomic_level="genus" \
  --cores=1 \
  --subnetworks="sample_type" \
  --fill_ebc_by="cave" \
  --interactive=FALSE \
  --query "Pseudonocardiaceae;Crossiella Euzebyaceae;uncultured
Gaiellaceae;Gaiella d13C_VPDB d15N_AIR pct_C pct_N" \
  --skip_plotting=TRUE
```

Chapter VI: *Conclusion*

The wide range of methodologies applied to the three distinct cave habitats investigated here provides several insights into the potential structure-function relationships among microbial communities in the shallow subsurface. Metabolic flexibility appears to be a hallmark of the taxa queried in these dark, oligotrophic environments. The phylum Actinobacteriota deserves particular attention as a consistently encountered, successful colonizer of these and other terrestrial cave ecosystems worldwide (Cuezva et al., 2012; Gonzalez-Pimentel et al., 2018; Lavoie et al., 2017; Porca et al., 2012; Tomczyk-Żak & Zielenkiewicz, 2016). The abundance of highly ^{13}C-depleted 10,14-DiMe $C_{17:0}$ and 10-Me $C_{17:0}$ lipid biomarkers relative to other C reservoirs implicates this phylum in C fixation in terrestrial lava caves (**Chapter II**). In a completely distinct terrestrial carbonate cave, novel draft genomes of the Actinobacteriota families *Egibacteraceae* and *Pseudonocardiaceae* exhibit potential for carbon monoxide, hydrogen, and occasionally thiosulfate oxidation, in addition to RuBisCO-based C fixation (**Chapter IV**). The dominant cell membrane fatty acids of these families are 10,14-DiMe $C_{17:0}$ and 10-Me $C_{17:0}$, respectively (Reichert et al., 1998; Zhang et al., 2016). It is notable that these methodologically distinct observations, several years and thousands of kilometers apart, converge on the identification of the *Egibacteraceae* and *Pseudonocardiaceae* as some of the key facilitators of cycling C in terrestrial caves.

This work also informs the biogeographic patterns in the community structure of a submerged cave ecosystem by examining the distribution of microbiota across a vertically stratified aquifer (**Chapter III**). Contrasting with its terrestrial counterparts, the anchialine caves accessing the Yucatán carbonate aquifer generally do not harbor abundant populations of Actinobacteriota (**Chapter III**). Instead, the 16S rRNA gene tag survey identifies an

uncharacterized member of the *Comamonadaceae* as being highly prevalent and abundant. The application of custom-written network analysis software (**Chapter V**) to this dataset demonstrates that this family exhibits distinct co-occurrence patterns with diverse taxa dependent on hydrological region. Despite compositional differences between the communities inhabiting terrestrial caves and those in the Yucatán carbonate aquifer, taxa putatively involved in the oxidation of hydrogen, methane, and/or carbon monoxide are abundant across the three studied cave systems (**Chapters II-IV**).

These and other "non-point" energy sources (**Chapter IV**) have been shown to support microbiota in non-photosynthetic ecosystems elsewhere (Ortiz et al., 2014, 2021; Stevens, 1997). By contrast, in the caves studied here, populations of sulfur-oxidizing bacteria such as *Sulfurovum* only widely proliferate in the Yucatán carbonate aquifer where decomposing organic matter accumulates, such as deep pit cenotes, which are limited in their geographic distribution (**Chapter II**). Likewise, MAGs representing the ammonia-oxidizing archaea Nitrosophaerales are by far the most abundant in areas of Mammoth Cave actively receiving nitrogen-rich bat guano (**Chapter IV**). In both cave systems, such "specialist" taxa reliant on "point" energy sources are less successful in terms of overall abundance and prevalence.

The fact that Actinobacteriota produce hyphae has been attributed to their success in caves, as they can serve as nucleation points for mineral or biofilm formation (Barton & Northup, 2007; Martin-Pozas et al., 2022). Our results suggest that Actinobacteriota in terrestrial cave biofilms can also obtain carbon and energy from diverse sources (**Chapters II, IV**). Their affinity for forming biofilms and potential for hydrogen-driven chemolithoautotrophy (**Chapter IV**) position them as globally distributed keystone taxa in caves. Syntrophic interactions between Actinobacteriota and other microbiota should be examined to further elucidate the nature of

microbially driven biogeochemistry in caves. Given the wealth of genome-resolved functional gene annotations now available (**Chapter IV**) and other ongoing ‘omics analyses performed by the Osburn Lab in Mammoth Cave, these and other fundamental questions pertaining to the structure-function relationships of these resilient ecosystems can continue to be examined and tested.

As always, more work is to be done.

References

Albright, S., & Louca, S. (2023). Trait biases in microbial reference genomes. *Scientific Data*, *10*(1), 84. https://doi.org/10.1038/s41597-023-01994-7

Albuquerque, L., França, L., Rainey, F. A., Schumann, P., Nobre, M. F., & da Costa, M. S. (2011). Gaiella occulta gen. Nov., sp. Nov., a novel representative of a deep branching phylogenetic lineage within the class Actinobacteria and proposal of Gaiellaceae fam. Nov. And Gaiellales ord. Nov. *Systematic and Applied Microbiology*, *34*(8), 595–599. https://doi.org/10.1016/j.syapm.2011.07.001

Alcocer, J., Lugo, A., Marín, L. E., & Escobar, E. (1998). Hydrochemistry of waters from five cenotes and evaluation of their suitability for drinking-water supplies, northeastern Yucatan, Mexico. *Hydrogeology Journal*, *6*(2), 293–301. https://doi.org/10.1007/s100400050152

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., Brodie, E. L., Williams, K. H., Hubbard, S. S., & Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*(1), 13219. https://doi.org/10.1038/ncomms13219

Angert, E. R., Northup, D. E., Reysenbach, A.-L., Peek, A. S., Goebel, B. M., & Pace, N. R. (1998). Molecular phylogentic analysis of a bacterial community in Sulphur River, Parker Cave, Kentucky. *American Mineralogist*, *83*(11-12 Part 2), 1583–1592. https://doi.org/10.2138/am-1998-11-1246

Aponte, J. C., Dillon, J. T., Tarozo, R., & Huang, Y. (2012). Separation of unsaturated organic compounds using silver–thiolate chromatographic material. *Journal of Chromatography A*, *1240*, 83–89. https://doi.org/10.1016/j.chroma.2012.03.082

Baas Becking, L. G. M. (1934). *Geobiologie of inleiding tot de milieukunde*. W.P. Van Stockum & Zoon.

Barker, J. F., & Fritz, P. (1981). Carbon isotope fractionation during microbial methane oxidation. *Nature*, *293*(5830), 289–291. https://doi.org/10.1038/293289a0

Barns, S. M., Delwiche, C. F., Palmer, J. D., & Pace, N. R. (1996). Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proceedings of the National Academy of Sciences*, *93*(17), 9188–9193. https://doi.org/10.1073/pnas.93.17.9188

Barton, H. A., & Northup, D. (2007). Geomicrobiology in cave environments: Past, current and future perspectives. *Journal of Cave and Karst Studies*, *69*, 163–178.

Barton, H. A., Taylor, M. R., & Pace, N. R. (2004). Molecular Phylogenetic Analysis of a Bacterial Community in an Oligotrophic Cave Environment. *Geomicrobiology Journal*, *21*(1), 11–20. https://doi.org/10.1080/01490450490253428

Bauer-Gottwein, P., Gondwe, B. R. N., Charvet, G., Marín, L. E., Rebolledo-Vieyra, M., & Merediz-Alonso, G. (2011). Review: The Yucatán Peninsula karst aquifer, Mexico. *Hydrogeology Journal*, *19*(3), 507–524. https://doi.org/10.1007/s10040-010-0699-5

Beddows, P. A. (2004a). *Groundwater Hydrology of a Coastal Conduit Carbonate Aquifer: Caribbean Coast of the Yucatán Peninsula, México*. University of Bristol.

Beddows, P. A. (2004b). Yucatán Phreas, Mexico. In *Encyclopedia of Caves and Karst Science*. New York Pages.

Beddows, P. A., Smart, P. L., Whitaker, F. F., & Smith, S. L. (2007). Decoupled fresh–saline groundwater circulation of a coastal carbonate aquifer: Spatial patterns of temperature and specific electrical conductivity. *Journal of Hydrology*, *346*(1–2), 18–32. https://doi.org/10.1016/j.jhydrol.2007.08.013

Ben Maamar, S., Aquilina, L., Quaiser, A., Pauwels, H., Michon-Coudouel, S., Vergnaud-Ayraud, V., Labasque, T., Roques, C., Abbott, B. W., & Dufresne, A. (2015). Groundwater Isolation Governs Chemistry and Microbial Community Structure along Hydrologic Flowpaths. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.01457

Berney, M., Greening, C., Hards, K., Collins, D., & Cook, G. M. (2014). Three different [NiFe] hydrogenases confer metabolic flexibility in the obligate aerobe *M ycobacterium smegmatis*. *Environmental Microbiology*, *16*(1), 318–330. https://doi.org/10.1111/1462-2920.12320

Bligh, E. G., & Dyer, W. J. (1959). A rapid method of total lipid extraction and purification. *Canadian Journal of Biochemistry and Physiology*, *37*, 911–917.

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Boschker, H. T. S., de Brouwer, J. F. C., & Cappenberg, T. E. (1999). The contribution of macrophyte-derived organic matter to microbial biomass in salt-marsh sediments: Stable carbon isotope analysis of microbial biomarkers. *Limnology and Oceanography*, *44*(2), 309–319. https://doi.org/10.4319/lo.1999.44.2.0309

Bowman, J. P. (2014). The Family Cryomorphaceae. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes* (pp. 539–550). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38954-2_135

Brankovits, D., Pohlman, J. W., Ganju, N. K., Iliffe, T. M., Lowell, N., Roth, E., Sylva, S. P., Emmert, J. A., & Lapham, L. L. (2018). Hydrologic Controls of Methane Dynamics in Karst Subterranean Estuaries. *Global Biogeochemical Cycles*, *32*(12), 1759–1775. https://doi.org/10.1029/2018GB006026

Brankovits, D., Pohlman, J. W., Niemann, H., Leigh, M. B., Leewis, M. C., Becker, K. W., Iliffe, T. M., Alvarez, F., Lehmann, M. F., & Phillips, B. (2017). Methane- and dissolved organic carbon-fueled microbial loop supports a tropical subterranean estuary ecosystem. *Nature Communications*, *8*(1), 1835. https://doi.org/10.1038/s41467-017-01776-x

Budge, S., Wooller, M., Springer, A., Iverson, S. J., McRoy, C., & Divoky, G. (2008). Tracing carbon flow in an arctic marine food web using fatty acid-stable isotope analysis. *Oecologia*, *157*(1), 117–129.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Canfield, D. E., Glazer, A. N., & Falkowski, P. G. (2010). The Evolution and Future of Earth's Nitrogen Cycle. *Science*, *330*(6001), 192–196. https://doi.org/10.1126/science.1186120

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J. A., Smith, G., & Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, *6*(8), 1621–1624. https://doi.org/10.1038/ismej.2012.8

Casar, C. P., Kruger, B. R., Flynn, T. M., Masterson, A. L., Momper, L. M., & Osburn, M. R. (2020). Mineral-hosted biofilm communities in the continental deep subsurface, Deep Mine Microbial Observatory, SD, USA. *Geobiology*, *18*(4), 508–522. https://doi.org/10.1111/gbi.12391

Casciotti, K. L., Sigman, D. M., Hastings, M. G., Böhlke, J. K., & Hilkert, A. (2002). Measurement of the Oxygen Isotopic Composition of Nitrate in Seawater and Freshwater Using the Denitrifier Method. *Analytical Chemistry*, *74*(19), 4905–4912. https://doi.org/10.1021/ac020113w

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2019). GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, btz848. https://doi.org/10.1093/bioinformatics/btz848

Chefetz, B., Tarchitzky, J., Deshmukh, A. P., Hatcher, P. G., & Chen, Y. (2002). Structural characterization of soil organic matter and humic acids in particle-size fractions of an agricultural soil. *Soil Science Society of America Journal*, *66*, 129–141. https://doi.org/10.2136/sssaj2002.1290

Cho, J.-C. (2021). Omics-based microbiome analysis in microbial ecology: From sequences to information. *Journal of Microbiology*, *59*(3), 229–232. https://doi.org/10.1007/s12275-021-0698-3

Chu, H., Gao, G.-F., Ma, Y., Fan, K., & Delgado-Baquerizo, M. (2020). Soil Microbial Biogeography in a Changing World: Recent Advances and Future Perspectives. *MSystems*, *5*(2), e00803-19. https://doi.org/10.1128/mSystems.00803-19

Close, H. G., Wakeham, S. G., & Pearson, A. (2014). Lipid and 13C signatures of submicron and suspended particulate organic matter in the Eastern Tropical North Pacific: Implications for the contribution of Bacteria. *Deep Sea Research Part I: Oceanographic Research Papers*, *85*, 15–34. https://doi.org/10.1016/j.dsr.2013.11.005

Colman, D. R., Feyhl-Buska, J., Robinson, K. J., Fecteau, K. M., Xu, H., Shock, E. L., & Boyd, E. S. (2016). Ecological differentiation in planktonic and sediment-associated chemotrophic microbial populations in Yellowstone hot springs. *FEMS Microbiology Ecology*, *92*(9), fiw137. https://doi.org/10.1093/femsec/fiw137

Cordero, P. R. F., Bayly, K., Man Leung, P., Huang, C., Islam, Z. F., Schittenhelm, R. B., King, G. M., & Greening, C. (2019). Atmospheric carbon monoxide oxidation is a widespread mechanism supporting microbial survival. *The ISME Journal*, *13*(11), 2868–2881. https://doi.org/10.1038/s41396-019-0479-8

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, *1695*(5), 1–9.

Cuezva, S., Fernandez-Cortes, A., Porca, E., Pašić, L., Jurado, V., Hernandez-Marine, M., Serrano-Ortiz, P., Hermosin, B., Cañaveras, J. C., Sanchez-Moral, S., & Saiz-Jimenez, C. (2012). The biogeochemical role of Actinobacteria in Altamira Cave, Spain. *FEMS Microbiology Ecology*, *81*(1), 281–290. https://doi.org/10.1111/j.1574-6941.2012.01391.x

Cunliffe, M. (2011). Correlating carbon monoxide oxidation with cox genes in the abundant Marine Roseobacter Clade. *The ISME Journal*, *5*(4), 685–691. https://doi.org/10.1038/ismej.2010.170

Daza Brunet, R., & Bustillo Revuelta, M. Á. (2014). Exceptional silica speleothems in a volcanic cave: A unique example of silicification and sub-aquatic opaline stromatolite formation (Terceira, Azores). *Sedimentology*, *61*(7), 2113–2135. https://doi.org/10.1111/sed.12130

De Leon, M., Montecillo, A. D., Pinili, D. S., Siringan, M. A. T., & Park, D.-S. (2018). Bacterial diversity of bat guano from Cabalyorisa Cave, Mabini, Pangasinan, Philippines: A first report on the metagenome of Philippine bat guano. *PLOS ONE*, *13*(7), e0200095. https://doi.org/10.1371/journal.pone.0200095

de los Ríos, A., Bustillo, M. A., Ascaso, C., & Carvalho, M. R. (2011). Bioconstructions in ochreous speleothems from lava tubes on Terceira Island (Azores). *Sedimentary Geology*, *236*(1), 117–128. https://doi.org/10.1016/j.sedgeo.2010.12.012

Deja-Sikora, E., Gołębiewski, M., Kalwasińska, A., Krawiec, A., Kosobucki, P., & Walczak, M. (2019). Comamonadaceae OTU as a Remnant of an Ancient Microbial Community in Sulfidic Waters. *Microbial Ecology*, *78*(1), 85–101. https://doi.org/10.1007/s00248-018-1270-5

Deng, Y., Jiang, Y.-H., Yang, Y., He, Z., Luo, F., & Zhou, J. (2012). Molecular ecological network analyses. *BMC Bioinformatics*, *13*(1), 113. https://doi.org/10.1186/1471-2105-13-113

Di Cello, F., Pepi, M., Baldi, F., & Fani, R. (1997). Molecular characterization of an n-alkane-degrading bacterial community and identification of a new species, Acinetobacter venetianus. *Research in Microbiology*, *148*(3), 237–249. https://doi.org/10.1016/S0923-2508(97)85244-8

Dohlman, A. B., & Shen, X. (2019). Mapping the microbial interactome: Statistical and experimental approaches for microbiome network inference. *Experimental Biology and Medicine*, *244*(6), 445–458. https://doi.org/10.1177/1535370219836771

Donnelly-Nolan, J. M., Nathenson, M., Champion, D. E., Ramsey, D. W., Lowenstern, J. B., & Ewert, J. W. (2007). Volcano hazards assessment for Medicine Lake Volcano, northern California. *U.S. Geological Survey Scientific Investigations Report*, *2007-5174-A*(1–26).

Engel, A. S., Stern, L. A., & Bennett, P. C. (2004). Microbial contributions to cave formation: New insights into sulfuric acid speleogenesis. *Geology*, *32*(5), 369–372. https://doi.org/10.1130/G20288.1

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., Trigodet, F., Watson, A. R., Esen, Ö. C., Moore, R. M., Clayssen, Q., Lee, M. D., Kivenson, V., Graham, E. D., Merrill, B. D., … Willis, A. D. (2020). Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology*, *6*(1), 3–6. https://doi.org/10.1038/s41564-020-00834-3

Escobar-Zepeda, A., Rosas-Escobar, P., Marquez Valdelamar, L., de la Torre, P., Partida-Martinez, L. P., Remegaldo, R., Sanchez-Flores, A., & Vergara, F. (2021). Distinctive prokaryotic microbiomes in sympatric plant roots from a Yucatan cenote. *BMC Research Notes*, *14*(1), 333. https://doi.org/10.1186/s13104-021-05746-x

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, *320*(5879), 1034–1039. https://doi.org/10.1126/science.1153213

Farda, B., Djebaili, R., Vaccarelli, I., Del Gallo, M., & Pellegrini, M. (2022). Actinomycetes from Caves: An Overview of Their Diversity, Biotechnological Properties, and Insights for Their Use in Soil Environments. *Microorganisms*, *10*(2), 453. https://doi.org/10.3390/microorganisms10020453

Faust, K. (2021). Open challenges for microbial network construction and analysis. *The ISME Journal*, *15*(11), 3111–3118. https://doi.org/10.1038/s41396-021-01027-4

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Computational Biology*, *8*(7), e1002606. https://doi.org/10.1371/journal.pcbi.1002606

Figueroa, I. A., Barnum, T. P., Somasekhar, P. Y., Carlström, C. I., Engelbrektson, A. L., & Coates, J. D. (2018). Metagenomics-guided analysis of microbial chemolithoautotrophic phosphite oxidation yields evidence of a seventh natural CO_2 fixation pathway.

Proceedings of the National Academy of Sciences, *115*(1). https://doi.org/10.1073/pnas.1715549114

Fliermans, C. B., & Schmidt, E. L. (1977). *Nitrobacter* in Mammoth Cave. *International Journal of Speleology*, *9*(1), 1–19. https://doi.org/10.5038/1827-806X.9.1.1

Fondi, M., Maida, I., Perrin, E., Orlandini, V., La Torre, L., Bosi, E., Negroni, A., Zanaroli, G., Fava, F., Decorosi, F., Giovannetti, L., Viti, C., Vaneechoutte, M., Dijkshoorn, L., & Fani, R. (2016). Genomic and phenotypic characterization of the species Acinetobacter venetianus. *Scientific Reports*, *6*(1), 21985. https://doi.org/10.1038/srep21985

Fuchs, G. (2011). Alternative pathways of carbon dioxide fixation: Insights into the early evolution of life? *Annual Review of Microbiology*, *65*(1), 631–658. https://doi.org/10.1146/annurev-micro-090110-102801

Giovannelli, D., Chung, M., Staley, J., Starovoytov, V., Le Bris, N., & Vetriani, C. (2016). Sulfurovum riftiae sp. Nov., a mesophilic, thiosulfate-oxidizing, nitrate-reducing chemolithoautotrophic epsilonproteobacterium isolated from the tube of the deep-sea hydrothermal vent polychaete Riftia pachyptila. *International Journal of Systematic and Evolutionary Microbiology*, *66*(7), 2697–2701. https://doi.org/10.1099/ijsem.0.001106

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappé, M. S., Short, J. M., Carrington, J. C., & Mathur, E. J. (2005). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science*, *309*(5738), 1242–1245. https://doi.org/10.1126/science.1114057

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Glaubitz, S., Kießlich, K., Meeske, C., Labrenz, M., & Jürgens, K. (2013). SUP05 Dominates the Gammaproteobacterial Sulfur Oxidizer Assemblages in Pelagic Redoxclines of the Central Baltic and Black Seas. *Applied and Environmental Microbiology*, *79*(8), 2767–2776. https://doi.org/10.1128/AEM.03777-12

Goberna, M., & Verdú, M. (2022). Cautionary notes on the use of co-occurrence networks in soil ecology. *Soil Biology and Biochemistry*, *166*, 108534. https://doi.org/10.1016/j.soilbio.2021.108534

Gonzalez-Pimentel, J. L., Miller, A. Z., Jurado, V., Leonila, L., Pereira, M. F. C., & Saiz-Jimenez, C. (2018). Yellow coloured mats from lava tubes of La Palma (Canary Islands, Spain) are

dominated by metabolically active Actinobacteria. *Scientific Reports*, *8*(1944). https://doi.org/10.1038/s41598-018-20393-2

Greeley, R. (1971). Lava tubes and channels in the lunar Marius Hills. *The Moon*, *3*(3), 289–314. https://doi.org/10.1007/BF00561842

Greening, C., Biswas, A., Carere, C. R., Jackson, C. J., Taylor, M. C., Stott, M. B., Cook, G. M., & Morales, S. E. (2016). Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. *The ISME Journal*, *10*(3), 761–777. https://doi.org/10.1038/ismej.2015.153

Grostern, A., & Alvarez-Cohen, L. (2013). RubisCO-based CO2 fixation and C1 metabolism in the actinobacterium Pseudonocardia dioxanivorans CB1190. *Environmental Microbiology*, *15*(11), 3040–3053. https://doi.org/10.1111/1462-2920.12144

Hamilton, T. L., Jones, D. S., Schaperdoth, I., & Macalady, J. L. (2015). Metagenomic insights into S(0) precipitation in a terrestrial subsurface lithoautotrophic ecosystem. *Frontiers in Microbiology*, *5*(756). https://doi.org/10.3389/fmicb.2014.00756

Han, Y., & Perner, M. (2015). The globally widespread genus Sulfurimonas: Versatile energy metabolisms and adaptations to redox clines. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00989

Harwood, J. L., & Stumpf, P. K. (1971). Fat metabolism in higher plants: XLIII. Control of fatty acid synthesis in germinating seeds. *Archives of Biochemistry and Biophysics*, *142*(1), 281–291. https://doi.org/10.1016/0003-9861(71)90285-2

Hathaway, J. J. M., Garcia, M. G., Balasch, M. M., Spilde, M. N., Stone, F. D., Dapkevicius, M. D. E. L. N. E., Amorim, I. R., Gabriel, R., Borges, P. A. V., & Northup, D. E. (2014). Comparison of bacterial diversity in Azorean and Hawai'ian lava cave microbial mats. *Geomicrobiology Journal*, *31*(3), 205–220. PubMed. https://doi.org/10.1080/01490451.2013.777491

Hathaway, J. J. M., Sinsabaugh, R. L., Dapkevicius, M. D. L. N. E., & Northup, D. E. (2014). Diversity of Ammonia Oxidation (amoA) and Nitrogen Fixation (nifH) Genes in Lava Caves of Terceira, Azores, Portugal. *Geomicrobiology Journal*, *31*(3), 221–235. PubMed. https://doi.org/10.1080/01490451.2012.752424

Hayes, J. (2001). Fractionation of carbon and hydrogen isotopes in biosynthetic processes. *Reviews in Mineralogy & Geochemistry*, *43*, 225–277. https://doi.org/10.2138/gsrmg.43.1.225

Hess, J. W., & White, W. B. (1989). Chemical Hydrology. In W. B. White & E. L. White (Eds.), *Karst Hydrology* (pp. 145–174). Springer US. https://doi.org/10.1007/978-1-4615-7317-3_6

Hess, W. (1900). The origin of nitrates in cavern earths. *Journal of Geology*, *8*(2), 129–134.

Hill, C. A. (1981). Origin of Cave Saltpeter. *The Journal of Geology*, *89*(2), 252–259. https://doi.org/10.1086/628584

Hoefman, S., van der Ha, D., Iguchi, H., Yurimoto, H., Sakai, Y., Boon, N., Vandamme, P., Heylen, K., & De Vos, P. (2014). Methyloparacoccus murrellii gen. Nov., sp. Nov., a methanotroph isolated from pond water. *International Journal of Systematic and Evolutionary Microbiology*, *64*(Pt_6), 2100–2107. https://doi.org/10.1099/ijs.0.057760-0

Holmes, A. J., Tujula, N. A., Holley, M., Contos, A., James, J. M., Rogers, P., & Gillings, M. R. (2001). Phylogenetic structure of unusual aquatic microbial formations in Nullarbor caves, Australia. *Environmental Microbiology*, *3*(4), 256–264. https://doi.org/10.1046/j.1462-2920.2001.00187.x

Horz, F. (1985). *Lava tubes-potential shelters for habitats*. 405–412.

Hose, L. D., Palmer, A. N., Palmer, M. V., Northup, D. E., Boston, P. J., & DuChene, H. R. (2000). Microbiology and geochemistry in a hydrogen-sulphide-rich karst environment. *Chemical Geology*, *169*(3), 399–423. https://doi.org/10.1016/S0009-2541(00)00217-5

Huang, L., Bae, H., Young, C., Pain, A. J., Martin, J. B., & Ogram, A. (2021). *Campylobacterota* dominate the microbial communities in a tropical karst subterranean estuary, with implications for cycling and export of nitrogen to coastal waters. *Environmental Microbiology*, *23*(11), 6749–6763. https://doi.org/10.1111/1462-2920.15746

Huang, Y., & Goodfellow, M. (2015). Pseudonocardia. In M. E. Trujillo, S. Dedysh, P. DeVos, B. Hedlund, P. Kämpfer, F. A. Rainey, & W. B. Whitman (Eds.), *Bergey's Manual of Systematics of Archaea and Bacteria* (pp. 1–32). https://doi.org/10.1002/9781118960608.gbm00184

Hug, L. A., Thomas, B. C., Brown, C. T., Frischkorn, K. R., Williams, K. H., Tringe, S. G., & Banfield, J. F. (2015). Aquifer environment selects for microbial species cohorts in sediment and groundwater. *The ISME Journal*, *9*(8), 1846–1856. https://doi.org/10.1038/ismej.2015.2

Huo, Y.-Y., Xu, X.-W., Cao, Y., Wang, C.-S., Zhu, X.-F., Oren, A., & Wu, M. (2009). Marinobacterium nitratireducens sp. Nov. And Marinobacterium sediminicola sp. Nov.,

isolated from marine sediment. *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*, *59*(5), 1173–1178. https://doi.org/10.1099/ijs.0.005751-0

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), 119. https://doi.org/10.1186/1471-2105-11-119

Ji, M., Greening, C., Vanwonterghem, I., Carere, C. R., Bay, S. K., Steen, J. A., Montgomery, K., Lines, T., Beardall, J., Van Dorst, J., Snape, I., Stott, M. B., Hugenholtz, P., & Ferrari, B. C. (2017). Atmospheric trace gases support primary production in Antarctic desert surface soil. *Nature*, *552*(7685), 400–403. https://doi.org/10.1038/nature25014

Jiang, D., Armour, C. R., Hu, C., Mei, M., Tian, C., Sharpton, T. J., & Jiang, Y. (2019). Microbiome Multi-Omics Network Analysis: Statistical Considerations, Limitations, and Opportunities. *Frontiers in Genetics*, *10*, 995. https://doi.org/10.3389/fgene.2019.00995

Jiao, J.-Y., Fu, L., Hua, Z.-S., Liu, L., Salam, N., Liu, P.-F., Lv, A.-P., Wu, G., Xian, W.-D., Zhu, Q., Zhou, E.-M., Fang, B.-Z., Oren, A., Hedlund, B. P., Jiang, H.-C., Knight, R., Cheng, L., & Li, W.-J. (2021). Insight into the function and evolution of the Wood–Ljungdahl pathway in Actinobacteria. *The ISME Journal*, *15*(10), 3005–3018. https://doi.org/10.1038/s41396-021-00935-9

Jo, S. Y., Rhie, M. N., Jung, S. M., Sohn, Y. J., Yeon, Y. J., Kim, M.-S., Park, C., Lee, J., Park, S. J., & Na, J.-G. (2020). Hydrogen Production from Methane by Methylomonas sp. DH-1 under Micro-aerobic Conditions. *Biotechnology and Bioprocess Engineering*, *25*(1), 71–77. https://doi.org/10.1007/s12257-019-0256-6

Kaku, T., Haruyama, J., Miyake, W., Kumamoto, A., Ishiyama, K., Nishibori, T., Yamamoto, K., Crites, S. T., Michikami, T., Yokota, Y., Sood, R., Melosh, H. J., Chappaz, L., & Howell, K. C. (2017). Detection of Intact Lava Tubes at Marius Hills on the Moon by SELENE (Kaguya) Lunar Radar Sounder. *Geophysical Research Letters*, *44*(20), 10,155-10,161. https://doi.org/10.1002/2017GL074998

Kambesis, P., & Coke, J. G. (2016). The Sac Actun system, Quintana Roo, Mexico. *Boletin Geologico y Minero*, *127*, 177–192.

Kanaparthi, D., Pommerenke, B., Casper, P., & Dumont, M. G. (2013). Chemolithotrophic nitrate-dependent Fe(II)-oxidizing nature of Actinobacterial subdivision lineage TM3. *The ISME Journal*, *7*(8), 1582–1594. https://doi.org/10.1038/ismej.2013.38

Kaneda, T. (1991). Iso- and anteiso-fatty acids in Bacteria: Biosynthesis, function, and taxonomic significance. *Microbiological Reviews*, *55*(2), 288–302. PubMed.

Kaneko, S., Inagaki, M., & Morishita, T. (2010). *A simple method for the determination of nitrate in potassium chloride extracts from forest soils*. 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia.

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

Keeling, R. F., Graven, H. D., Welp, L. R., Resplandy, L., Bi, J., Piper, S. C., Sun, Y., Bollenbacher, A., & Meijer, H. A. J. (2017). Atmospheric evidence for a global secular increase in carbon isotopic discrimination of land photosynthesis. *Proceedings of the National Academy of Sciences*, *114*(39), 10361. https://doi.org/10.1073/pnas.1619240114

Keszthelyi, L., Jaeger, W., McEwen, A., Tornabene, L., Beyer, R. A., Dundas, C., & Milazzo, M. (2008). High Resolution Imaging Science Experiment (HiRISE) images of volcanic terrains from the first 6 months of the Mars Reconnaissance Orbiter Primary Science Phase. *Journal of Geophysical Research: Planets*, *113*(E4). https://doi.org/10.1029/2007JE002968

Keweloh, H., & Heipieper, H. J. (1996). Trans unsaturated fatty acids in Bacteria. *Lipids*, *31*(2), 129–137. https://doi.org/10.1007/BF02522611

King, G. M. (2003). Molecular andCulture-Based Analyses of Aerobic Carbon Monoxide OxidizerDiversity†. *Applied and Environmental Microbiology*, *69*(12), 7257–7265. https://doi.org/10.1128/AEM.69.12.7257-7265.2003

Kolmert, Å., Wikström, P., & Hallberg, K. B. (2000). A fast and simple turbidimetric method for the determination of sulfate in sulfate-reducing bacterial cultures. *Journal of Microbiological Methods*, *41*(3), 179–184. https://doi.org/10.1016/S0167-7012(00)00154-8

Kopf, S. H., Sessions, A. L., Cowley, E. S., Reyes, C., Van Sambeek, L., Hu, Y., Orphan, V. J., Kato, R., & Newman, D. K. (2016). Trace incorporation of heavy water reveals slow and heterogeneous pathogen growth rates in cystic fibrosis sputum. *Proceedings of the National Academy of Sciences*, *113*(2), E110. https://doi.org/10.1073/pnas.1512057112

Kormas, K. A., Smith, D. C., Edgcomb, V., & Teske, A. (2003). Molecular analysis of deep subsurface microbial communities in Nankai Trough sediments (ODP Leg 190, Site 1176). *FEMS Microbiology Ecology*, *45*(2), 115–125. https://doi.org/10.1016/S0168-6496(03)00128-4

Krom, M. D. (1980). Spectrophotometric determination of ammonia: A study of a modified Berthelot reaction using salicylate and dichloroisocyanurate. *Analyst*, *105*(1249), 305–316. https://doi.org/10.1039/AN9800500305

Kruse, S., Goris, T., Westermann, M., Adrian, L., & Diekert, G. (2018). Hydrogen production by Sulfurospirillum species enables syntrophic interactions of Epsilonproteobacteria. *Nature Communications*, *9*(1), 4872. https://doi.org/10.1038/s41467-018-07342-3

Labeda, D. P. (2001). Crossiella gen. Nov., a new genus related to Streptoalloteichus. *International Journal of Systematic and Evolutionary Microbiology*, *51*(4), 1575–1579. https://doi.org/10.1099/00207713-51-4-1575

Laiz, L., Groth, I., Gonzalez, I., & Saiz-Jimenez, C. (1999). Microbiological study of the dripping waters in Altamira cave (Santillana del Mar, Spain). *Journal of Microbiological Methods*, *36*(1–2), 129–138. https://doi.org/10.1016/S0167-7012(99)00018-4

Lau, M. C. Y., Kieft, T. L., Kuloyo, O., Linage-Alvarez, B., Van Heerden, E., Lindsay, M. R., Magnabosco, C., Wang, W., Wiggins, J. B., Guo, L., Perlman, D. H., Kyin, S., Shwe, H. H., Harris, R. L., Oh, Y., Yi, M. J., Purtschert, R., Slater, G. F., Ono, S., … Onstott, T. C. (2016). An oligotrophic deep-subsurface community dependent on syntrophy is dominated by sulfur-driven autotrophic denitrifiers. *Proceedings of the National Academy of Sciences*, *113*(49). https://doi.org/10.1073/pnas.1612244113

Lavoie, K. H. (2017). Mammoth Cave Microbiology. In H. H. Hobbs III, R. A. Olson, E. G. Winkler, & D. C. Culver (Eds.), *Mammoth Cave* (pp. 235–250). Springer International Publishing. https://doi.org/10.1007/978-3-319-53718-4_16

Lavoie, K. H., Winter, A. S., Read, K. J. H., Hughes, E. M., Spilde, M. N., & Northup, D. E. (2017). Comparison of bacterial communities from lava cave microbial mats to overlying surface soils from Lava Beds National Monument, USA. *PloS One*, *12*(2), e0169339–e0169339. PubMed. https://doi.org/10.1371/journal.pone.0169339

Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, *25*(3), 217–228. https://doi.org/10.1016/j.tim.2016.11.008

Lee, M. D., Walworth, N. G., Sylvan, J. B., Edwards, K. J., & Orcutt, B. N. (2015). Microbial Communities on Seafloor Basalts at Dorado Outcrop Reflect Level of Alteration and Highlight Global Lithic Clades. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.01470

Léveillé, R. J., & Datta, S. (2010). Lava tubes and basaltic caves as astrobiological targets on Earth and Mars: A review. *Planetary and Space Science*, *58*(4), 592–598. https://doi.org/10.1016/j.pss.2009.06.004

Li, C., Hambright, K. D., Bowen, H. G., Trammell, M. A., Grossart, H., Burford, M. A., Hamilton, D. P., Jiang, H., Latour, D., Meyer, E. I., Padisák, J., Zamor, R. M., & Krumholz, L. R. (2021). Global co-occurrence of methanogenic archaea and methanotrophic bacteria in *Microcystis* aggregates. *Environmental Microbiology*, *23*(11), 6503–6519. https://doi.org/10.1111/1462-2920.15691

Li, Q., Zhou, Y., Lu, R., Zheng, P., & Wang, Y. (2022). Phylogeny, distribution and potential metabolism of candidate bacterial phylum KSB1. *PeerJ*, *10*, e13241. https://doi.org/10.7717/peerj.13241

Li, S., Wang, P., Chen, Y., Wilson, M. C., Yang, X., Ma, C., Lu, J., Chen, X., Wu, J., Shu, W., & Jiang, L. (2020). Island biogeography of soil bacteria and fungi: Similar patterns, but different mechanisms. *The ISME Journal*, *14*(7), 1886–1896. https://doi.org/10.1038/s41396-020-0657-8

Limbri, H., Gunawan, C., Thomas, T., Smith, A., Scott, J., & Rosche, B. (2014). Coal-Packed Methane Biofilter for Mitigation of Green House Gas Emissions from Coal Mine Ventilation Air. *PLoS ONE*, *9*(4), e94641. https://doi.org/10.1371/journal.pone.0094641

Liu, H., Xin, Y., & Xun, L. (2014). Distribution, Diversity, and Activities of Sulfur Dioxygenases in Heterotrophic Bacteria. *Applied and Environmental Microbiology*, *80*(5), 1799–1806. https://doi.org/10.1128/AEM.03281-13

Liu, Z.-P., Wu, J.-F., Liu, Z.-H., & Liu, S.-J. (2006). Pseudonocardia ammonioxydans sp. Nov., isolated from coastal sediment. *International Journal of Systematic and Evolutionary Microbiology*, *56*(3), 555–558. https://doi.org/10.1099/ijs.0.63878-0

Logemann, J., Graue, J., Köster, J., Engelen, B., Rullkötter, J., & Cypionka, H. (2011). A laboratory experiment of intact polar lipid degradation in sandy sediments. *Biogeosciences*, *8*(9), 2547–2560. https://doi.org/10.5194/bg-8-2547-2011

Logue, J. B., Langenheder, S., Andersson, A. F., Bertilsson, S., Drakare, S., Lanzén, A., & Lindström, E. S. (2012). Freshwater bacterioplankton richness in oligotrophic lakes depends on nutrient availability rather than on species–area relationships. *The ISME Journal*, *6*(6), 1127–1136. https://doi.org/10.1038/ismej.2011.184

Lollar, B. S., Lacrampe-Couloume, G., Slater, G. F., Ward, J., Moser, D. P., Gihring, T. M., Lin, L. H., & Onstott, T. C. (2006). Unravelling abiogenic and biogenic sources of methane in

the Earth's deep subsurface. *Chemical Geology*, *226*(3), 328–339. https://doi.org/10.1016/j.chemgeo.2005.09.027

Lueders, T., Kindler, R., Miltner, A., Friedrich, M. W., & Kaestner, M. (2006). Identification of Bacterial Micropredators Distinctively Active in a Soil Microbial Food Web. *Applied and Environmental Microbiology*, *72*(8), 5342–5348. https://doi.org/10.1128/AEM.00400-06

Macalady, J. L., Dattagupta, S., Schaperdoth, I., Jones, D. S., Druschel, G. K., & Eastman, D. (2008). Niche differentiation among sulfur-oxidizing bacterial populations in cave waters. *The ISME Journal*, *2*(6), 590–601. https://doi.org/10.1038/ismej.2008.25

Mariotti, A., Germon, J. C., Hubert, P., Kaiser, P., Letolle, R., Tardieux, A., & Tardieux, P. (1981). Experimental determination of nitrogen kinetic isotope fractionation: Some principles; illustration for the denitrification and nitrification processes. *Plant and Soil*, *62*(3), 413–430. https://doi.org/10.1007/BF02374138

Martin-Pozas, T., Cuezva, S., Fernandez-Cortes, A., Cañaveras, J. C., Benavente, D., Jurado, V., Saiz-Jimenez, C., Janssens, I., Seijas, N., & Sanchez-Moral, S. (2022). Role of subterranean microbiota in the carbon cycle and greenhouse gas dynamics. *Science of The Total Environment*, *831*, 154921. https://doi.org/10.1016/j.scitotenv.2022.154921

Martiny, J. B. H., Bohannan, B. J. M., Brown, J. H., Colwell, R. K., Fuhrman, J. A., Green, J. L., Horner-Devine, M. C., Kane, M., Krumins, J. A., Kuske, C. R., Morin, P. J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V. H., & Staley, J. T. (2006). Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology*, *4*(2), 102–112. https://doi.org/10.1038/nrmicro1341

Matchado, M. S., Lauber, M., Reitmeier, S., Kacprowski, T., Baumbach, J., Haller, D., & List, M. (2021). Network analysis methods for studying microbial communities: A mini review. *Computational and Structural Biotechnology Journal*, *19*, 2687–2698. https://doi.org/10.1016/j.csbj.2021.05.001

Matsen, F. A., Kodner, R. B., & Armbrust, E. V. (2010). pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, *11*(1), 538. https://doi.org/10.1186/1471-2105-11-538

Maupin-Furlow, J., & Ferry, J. G. (1996). Characterization of the cdhD and cdhE genes encoding subunits of the corrinoid/iron-sulfur enzyme of the CO dehydrogenase complex from Methanosarcina thermophila. *Journal of Bacteriology*, *178*(2), 340–346. https://doi.org/10.1128/jb.178.2.340-346.1996

Melim, L. A., Shinglman, K. M., Boston, P. J., Northup, D. E., Spilde, M. N., & Queen, J. M. (2001). Evidence for Microbial Involvement in Pool Finger Precipitation, Hidden Cave, New Mexico. *Geomicrobiology Journal*, *18*(3), 311–329. https://doi.org/10.1080/01490450152467813

Mende, D. R., Bryant, J. A., Aylward, F. O., Eppley, J. M., Nielsen, T., Karl, D. M., & DeLong, E. F. (2017). Environmental drivers of a microbial genomic transition zone in the ocean's interior. *Nature Microbiology*, *2*(10), 1367–1373. https://doi.org/10.1038/s41564-017-0008-3

Meyer, B., Imhoff, J. F., & Kuever, J. (2007). Molecular analysis of the distribution and phylogeny of the soxB gene among sulfur-oxidizing bacteria – evolution of the Sox sulfur oxidation enzyme system. *Environmental Microbiology*, *9*(12), 2957–2977. https://doi.org/10.1111/j.1462-2920.2007.01407.x

Miranda, K. M., Espey, M. G., & Wink, D. A. (2001). A rapid, simple spectrophotometric method for simultaneous detection of nitrate and nitrite. *Nitric Oxide*, *5*(1), 62–71. https://doi.org/10.1006/niox.2000.0319

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*(10), 2319–2333. PubMed. https://doi.org/10.1038/ismej.2017.94

Moore, A., Lenczewski, M., Leal-Bautista, R. M., & Duvall, M. (2020). Groundwater microbial diversity and antibiotic resistance linked to human population density in Yucatan Peninsula, Mexico. *Canadian Journal of Microbiology*, *66*(1), 46–58. https://doi.org/10.1139/cjm-2019-0173

Mooshammer, M., Alves, R. J. E., Bayer, B., Melcher, M., Stieglmeier, M., Jochum, L., Rittmann, S. K.-M. R., Watzka, M., Schleper, C., Herndl, G. J., & Wanek, W. (2020). Nitrogen Isotope Fractionation During Archaeal Ammonia Oxidation: Coupled Estimates From Measurements of Residual Ammonium and Accumulated Nitrite. *Frontiers in Microbiology*, *11*, 1710. https://doi.org/10.3389/fmicb.2020.01710

Moran, M. A., Buchan, A., González, J. M., Heidelberg, J. F., Whitman, W. B., Kiene, R. P., Henriksen, J. R., King, G. M., Belas, R., Fuqua, C., Brinkac, L., Lewis, M., Johri, S., Weaver, B., Pai, G., Eisen, J. A., Rahe, E., Sheldon, W. M., Ye, W., … Ward, N. (2004). Genome sequence of Silicibacter pomeroyi reveals adaptations to the marine environment. *Nature*, *432*(7019), 910–913. https://doi.org/10.1038/nature03170

Mori, K., Yamaguchi, K., & Hanada, S. (2018). Sulfurovum denitrificans sp. Nov., an obligately chemolithoautotrophic sulfur-oxidizing epsilonproteobacterium isolated from a

hydrothermal field. *International Journal of Systematic and Evolutionary Microbiology*, *68*(7), 2183–2187. https://doi.org/10.1099/ijsem.0.002803

Morris, B. E. L., Henneberger, R., Huber, H., & Moissl-Eichinger, C. (2013). Microbial syntrophy: Interaction for the common good. *FEMS Microbiology Reviews*, *37*(3), 384–406. https://doi.org/10.1111/1574-6976.12019

Mullins, T. D., Britschgi, T. B., Krest, R. L., & Giovannoni, S. J. (1995). Genetic comparisons reveal the same unknown bacterial lineages in Atlantic and Pacific bacterioplankton communities. *Limnology and Oceanography*, *40*(1), 148–158. https://doi.org/10.4319/lo.1995.40.1.0148

Murrell, J. C., McDonald, I. R., & Gilbert, B. (2000). Regulation of expression of methane monooxygenases by copper ions. *Trends in Microbiology*, *8*(5), 221–225. https://doi.org/10.1016/S0966-842X(00)01739-X

Nandi, R., & Sengupta, S. (1998). Microbial Production of Hydrogen: An Overview. *Critical Reviews in Microbiology*, *24*(1), 61–84. https://doi.org/10.1080/10408419891294181

Navarrete-Euan, H., Rodríguez-Escamilla, Z., Pérez-Rueda, E., Escalante-Herrera, K., & Martínez-Núñez, M. A. (2021). Comparing Sediment Microbiomes in Contaminated and Pristine Wetlands along the Coast of Yucatan. *Microorganisms*, *9*(4), 877. https://doi.org/10.3390/microorganisms9040877

Northup, D. E., Melim, L. A., Spilde, M. N., Hathaway, J. J. M., Garcia, M. G., Moya, M., Stone, F. D., Boston, P. J., Dapkevicius, M. L. N. E., & Riquelme, C. (2011). Lava cave microbial communities within mats and secondary mineral deposits: Implications for life detection on other planets. *Astrobiology*, *11*(7), 601–618. https://doi.org/10.1089/ast.2010.0562

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

Oksanen, J. F., Blanchet, G., Friendly, M., Kindt, R., Legendre, D. M., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2019). *vegan: Community Ecology Package* (2.5-6). https://CRAN.R-project.org/package=vegan

Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, *11*(12), 2864–2868. https://doi.org/10.1038/ismej.2017.126

Opalički Slabe, M., Danevčič, T., Hug, K., Fillinger, L., Mandić-Mulec, I., Griebler, C., & Brancelj, A. (2021). Key drivers of microbial abundance, activity, and diversity in karst spring waters across an altitudinal gradient in Slovenia. *Aquatic Microbial Ecology*, *86*, 99–114. https://doi.org/10.3354/ame01956

Orcutt, B., Sylvan, J., Rogers, D., Delaney, J., Lee, R., & Girguis, P. (2015). Carbon fixation by basalt-hosted microbial communities. *Frontiers in Microbiology*, *6*(904). https://doi.org/10.3389/fmicb.2015.00904

Orlygsson, J., & Kristjansson, J. K. (2014). The Family Hydrogenophilaceae. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes* (pp. 859–868). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30197-1_244

Ortiz, M., Legatzki, A., Neilson, J. W., Fryslie, B., Nelson, W. M., Wing, R. A., Soderlund, C. A., Pryor, B. M., & Maier, R. M. (2014). Making a living while starving in the dark: Metagenomic insights into the energy dynamics of a carbonate cave. *The ISME Journal*, *8*(2), 478–491. https://doi.org/10.1038/ismej.2013.159

Ortiz, M., Leung, P. M., Shelley, G., Jirapanjawat, T., Nauer, P. A., Van Goethem, M. W., Bay, S. K., Islam, Z. F., Jordaan, K., Vikram, S., Chown, S. L., Hogg, I. D., Makhalanyane, T. P., Grinter, R., Cowan, D. A., & Greening, C. (2021). Multiple energy sources and metabolic strategies sustain microbial diversity in Antarctic desert soils. *Proceedings of the National Academy of Sciences*, *118*(45), e2025322118. https://doi.org/10.1073/pnas.2025322118

Ortiz, M., Neilson, J. W., Nelson, W. M., Legatzki, A., Byrne, A., Yu, Y., Wing, R. A., Soderlund, C. A., Pryor, B. M., Pierson, L. S., & Maier, R. M. (2013). Profiling Bacterial Diversity and Taxonomic Composition on Speleothem Surfaces in Kartchner Caverns, AZ. *Microbial Ecology*, *65*(2), 371–383. https://doi.org/10.1007/s00248-012-0143-6

Osburn, M. R., Kruger, B., Masterson, A. L., Casar, C. P., & Amend, J. P. (2019). Establishment of the Deep Mine Microbial Observatory (DeMMO), South Dakota, USA, a Geochemically Stable Portal Into the Deep Subsurface. *Frontiers in Earth Science*, *7*, 196. https://doi.org/10.3389/feart.2019.00196

Pace, N. R. (1971). Caves and saltpeter: A novel hypothesis for saltpeter formation. *Caving in the Rockies*, 7–9.

Pace, N. R., Sapp, J., & Goldenfeld, N. (2012). Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life. *Proceedings of the National Academy of Sciences*, *109*(4), 1011–1018. https://doi.org/10.1073/pnas.1109716109

Palmer, A. N. (1981). *A Geological Guide to Mammoth Cave National Park*. Zephyrus Press.

Palmer, A. N. (2017). Geology of Mammoth Cave. In H. H. Hobbs III, R. A. Olson, E. G. Winkler, & D. C. Culver (Eds.), *Mammoth Cave* (pp. 97–109). Springer International Publishing. https://doi.org/10.1007/978-3-319-53718-4_6

Park, S. W., Hwang, E. H., Jang, H. S., Lee, J. H., Kang, B. S., Oh, J. I., & Kim, Y. M. (2009). Presence of duplicate genes encoding a phylogenetically new subgroup of form I ribulose 1,5-bisphosphate carboxylase/oxygenase in Mycobacterium sp. Strain JC1 DSM 3803. *Research in Microbiology*, *160*(2), 159–165. https://doi.org/10.1016/j.resmic.2008.12.002

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. https://doi.org/10.1101/gr.186072.114

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pereira, I. A. C., Ramos, A. R., Grein, F., Marques, M. C., Da Silva, S. M., & Venceslau, S. S. (2011). A Comparative Genomic Analysis of Energy Metabolism in Sulfate Reducing Bacteria and Archaea. *Frontiers in Microbiology*, *2*. https://doi.org/10.3389/fmicb.2011.00069

Perez-Molphe-Montoya, E., Küsel, K., & Overholt, W. A. (2022). Redefining the phylogenetic and metabolic diversity of phylum Omnitrophota. *Environmental Microbiology*, *24*(11), 5437–5449. https://doi.org/10.1111/1462-2920.16170

Philippot, L. (2002). Denitrifying genes in bacterial and Archaeal genomes. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, *1577*(3), 355–376. https://doi.org/10.1016/S0167-4781(02)00420-7

Pohl, E. R. (1970). Upper Mississippian deposits of south-central Kentucky. *Transactions of the Kentucky Academy of Science*, *31*, 1–15.

Pohlman, J. W. (2011). The biogeochemistry of anchialine caves: Progress and possibilities. *Hydrobiologia*, *677*(1), 33–51. https://doi.org/10.1007/s10750-011-0624-5

Pohlman, J. W., Iliffe, T. M., & Cifuentes, L. A. (1997). A stable isotope study of organic cycling and the ecology of an anchialine cave ecosystem. *Oceanographic Literature Review*, *45*(2).

Popa, R., Smith, A. R., Popa, R., Boone, J., & Fisk, M. (2012). Olivine-respiring bacteria isolated from the rock-ice interface in a lava-tube cave, a Mars analog environment. *Astrobiology*, *12*(1), 9–18. https://doi.org/10.1089/ast.2011.0639

Porca, E., Jurado, V., Žgur-Bertok, D., Saiz-Jimenez, C., & Pašić, L. (2012). Comparative analysis of yellow microbial communities growing on the walls of geographically distinct caves indicates a common core of microorganisms involved in their formation. *FEMS Microbiology Ecology*, *81*(1), 255–266. https://doi.org/10.1111/j.1574-6941.2012.01383.x

Porter, M. L., & Engel, A. S. (2008). Diversity of Uncultured *Epsilonproteobacteria* from Terrestrial Sulfidic Caves and Springs. *Applied and Environmental Microbiology*, *74*(15), 4973–4977. https://doi.org/10.1128/AEM.02915-07

Priest, T., Heins, A., Harder, J., Amann, R., & Fuchs, B. M. (2022). Niche partitioning of the ubiquitous and ecologically relevant NS5 marine group. *The ISME Journal*, *16*(6), 1570–1582. https://doi.org/10.1038/s41396-022-01209-8

Probst, A. J., Ladd, B., Jarett, J. K., Geller-McGrath, D. E., Sieber, C. M. K., Emerson, J. B., Anantharaman, K., Thomas, B. C., Malmstrom, R. R., Stieglmeier, M., Klingl, A., Woyke, T., Ryan, M. C., & Banfield, J. F. (2018). Differential depth distribution of microbial function and putative symbionts through sediment-hosted aquifers in the deep terrestrial subsurface. *Nature Microbiology*, *3*(3), 328–336. https://doi.org/10.1038/s41564-017-0098-y

Props, R., Monsieurs, P., Vandamme, P., Leys, N., Denef, V. J., & Boon, N. (2019). Gene Expansion and Positive Selection as Bacterial Adaptations to Oligotrophic Conditions. *MSphere*, *4*(1), e00011-19. https://doi.org/10.1128/mSphereDirect.00011-19

Proulx, S., Promislow, D., & Phillips, P. (2005). Network thinking in ecology and evolution. *Trends in Ecology & Evolution*, *20*(6), 345–353. https://doi.org/10.1016/j.tree.2005.04.004

Pujalte, M. J., Lucena, T., Ruvira, M. A., Arahal, D. R., & Macián, M. C. (2014). The Family Rhodobacteraceae. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes* (pp. 439–512). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30197-1_377

Purkamo, L., Kietäväinen, R., Nuppunen-Puputti, M., Bomberg, M., & Cousins, C. (2020). Ultradeep microbial communities at 4.4 km within crystalline bedrock: Implications for habitability in a planetary context. *Life*, *10*(2). https://doi.org/10.3390/life10010002

Qin, W., Heal, K. R., Ramdasi, R., Kobelt, J. N., Martens-Habbena, W., Bertagnolli, A. D., Amin, S. A., Walker, C. B., Urakawa, H., Könneke, M., Devol, A. H., Moffett, J. W., Armbrust,

E. V., Jensen, G. J., Ingalls, A. E., & Stahl, D. A. (2017). Nitrosopumilus maritimus gen. Nov., sp. Nov., Nitrosopumilus cobalaminigenes sp. Nov., Nitrosopumilus oxyclinae sp. Nov., and Nitrosopumilus ureiphilus sp. Nov., four marine ammonia-oxidizing archaea of the phylum Thaumarchaeota. *International Journal of Systematic and Evolutionary Microbiology*, *67*(12), 5067–5079. https://doi.org/10.1099/ijsem.0.002416

QRSS. (2023). *QRSS - Quintana Roo Speleological Survey* [Map].

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Reichert, K., Lipski, A., Pradella, S., Stackebrandt, E., & Altendorf, K. (1998). Pseudonocardia asaccharolytica sp. Nov. And Pseudonocardia sulfidoxydans sp. Nov., two new dimethyl disulfide-degrading actinomycetes and emended description of the genus Pseudonocardia. *International Journal of Systematic and Evolutionary Microbiology*, *48*(2), 441–449. https://doi.org/10.1099/00207713-48-2-441

Reintjes, G., Arnosti, C., Fuchs, B., & Amann, R. (2019). Selfish, sharing and scavenging bacteria in the Atlantic Ocean: A biogeographical study of bacterial substrate utilisation. *The ISME Journal*, *13*(5), 1119–1132. https://doi.org/10.1038/s41396-018-0326-3

Rick Olson. (2013). *Potential Effects of Hydrogen Sulfide and Hydrocarbon Seeps on Mammoth Cave Ecosystems*. *28*.

Rios-Del Toro, E. E., Valenzuela, E. I., Ramírez, J. E., López-Lozano, N. E., & Cervantes, F. J. (2018). Anaerobic Ammonium Oxidation Linked to Microbial Reduction of Natural Organic Matter in Marine Sediments. *Environmental Science & Technology Letters*, *5*(9), 571–577. https://doi.org/10.1021/acs.estlett.8b00330

Riquelme, C., Marshall Hathaway, J. J., Enes Dapkevicius, M. D. L. N., Miller, A. Z., Kooser, A., Northup, D. E., Jurado, V., Fernandez, O., Saiz-Jimenez, C., & Cheeptham, N. (2015). Actinobacterial Diversity in Volcanic Caves and Associated Geomicrobiological Interactions. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.01342

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. (2020). *RESCRIPt: Reproducible sequence taxonomy reference database management for the masses* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.10.05.326504

Romanenko, L. A., Uchino, M., Falsen, E., Frolova, G. M., Zhukova, N. V., & Mikhailov, V. V. (2005). Pseudomonas pachastrellae sp. Nov., isolated from a marine sponge. *International Journal of Systematic and Evolutionary Microbiology*, *55*(2), 919–924. https://doi.org/10.1099/ijs.0.63176-0

Rontani, J.-F., Koblížek, M., Beker, B., Bonin, P., & Kolber, Z. S. (2003). On the origin of cis-vaccenic acid photodegradation products in the marine environment. *Lipids*, *38*(10), 1085–1092. https://doi.org/10.1007/s11745-006-1164-z

Rossi, C., Lozano, R. P., Isanta, N., & Hellstrom, J. (2010). Manganese stromatolites in caves: El Soplao (Cantabria, Spain). *Geology*, *38*(12), 1119–1122. https://doi.org/10.1130/G31283.1

Rotthauwe, J. H., Witzel, K. P., & Liesack, W. (1997). The ammonia monooxygenase structural gene amoA as a functional marker: Molecular fine-scale analysis of natural ammonia-oxidizing populations. *Applied and Environmental Microbiology*, *63*(12), 4704–4712. https://doi.org/10.1128/aem.63.12.4704-4712.1997

Ryabenko, E. (2013). Stable isotope methods for the study of the nitrogen cycle. *Topics in Oceanography*, 1–40.

Saiz-Jimenez, C., & Hermosin, B. (1999). Thermally assisted hydrolysis and methylation of dissolved organic matter in dripping waters from the Altamira Cave. *Journal of Analytical and Applied Pyrolysis*, *49*(1), 337–347. https://doi.org/10.1016/S0165-2370(98)00112-0

Sarbu, S. M., Kane, T. C., & Kinkle, B. K. (1996). A chemoautotrophically based cave ecosystem. *Science*, *272*(5270), 1953. https://doi.org/10.1126/science.272.5270.1953

Schimmelmann, A., Albertino, A., Sauer, P. E., Qi, H., Molinie, R., & Mesnard, F. (2009). Nicotine, acetanilide and urea multi-level ^{2}H-, ^{13}C- and ^{15}N-abundance reference materials for continuous-flow isotope ratio mass spectrometry. *Rapid Communications in Mass Spectrometry*, *23*(22), 3513–3521. https://doi.org/10.1002/rcm.4277

Schlegel, H. G., & Jannasch, H. W. (1967). Enrichment Cultures. *Annual Review of Microbiology*, *21*(1), 49–70. https://doi.org/10.1146/annurev.mi.21.100167.000405

Schubotz, F., Meyer-Dombard, D. R., Bradley, A. S., Fredricks, H. F., Hinrichs, K. U., Shock, E. L., & Summons, R. E. (2013). Spatial and temporal variability of biomarkers and microbial diversity reveal metabolic and community flexibility in streamer biofilm communities in the Lower Geyser Basin, Yellowstone National Park. *Geobiology*, *11*(6), 549–569. https://doi.org/10.1111/gbi.12051

Schumacher, W., Kroneck, P. M. H., & Pfennig, N. (1992). Comparative systematic study on ?Spirillum? 5175, Campylobacter and Wolinella species: Description of ?Spirillum? 5175 as Sulfurospirillum deleyianum gen. Nov., spec. Nov. *Archives of Microbiology*, *158*(4), 287–293. https://doi.org/10.1007/BF00245247

Schwab, V. F., Herrmann, M., Roth, V. N., Gleixner, G., Lehmann, R., Pohnert, G., Trumbore, S., Küsel, K., & Totsche, K. U. (2017). Functional diversity of microbial communities in pristine aquifers inferred by PLFA- and sequencing-based approaches. *Biogeosciences*, *14*(10), 2697–2714. https://doi.org/10.5194/bg-14-2697-2017

Selensky, M. J., Masterson, A. L., Blank, J. G., Lee, S. C., & Osburn, M. R. (2021). Stable Carbon Isotope Depletions in Lipid Biomarkers Suggest Subsurface Carbon Fixation in Lava Caves. *Journal of Geophysical Research: Biogeosciences*, *126*(7). https://doi.org/10.1029/2021JG006430

Selvin, J., Lanong, S., Syiem, D., De Mandal, S., Kayang, H., Kumar, N. S., & Kiran, G. S. (2019). Culture-dependent and metagenomic analysis of lesser horseshoe bats' gut microbiome revealing unique bacterial diversity and signatures of potential human pathogens. *Microbial Pathogenesis*, *137*, 103675. https://doi.org/10.1016/j.micpath.2019.103675

Shah, V., Chang, B. X., & Morris, R. M. (2017). Cultivation of a chemoautotroph from the SUP05 clade of marine bacteria that produces nitrite and consumes ammonium. *The ISME Journal*, *11*(1), 263–271. https://doi.org/10.1038/ismej.2016.87

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shen, J., Zerkle, A. L., Stueeken, E., & Claire, M. W. (2019). Nitrates as a Potential N Supply for Microbial Ecosystems in a Hyperarid Mars Analog System. *Life (Basel, Switzerland)*, *9*(4), 79. PubMed. https://doi.org/10.3390/life9040079

Shibamoto, S., Murata, T., & Yamamoto, K. (2016). Determination of double bond positions and geometry of methyl linoleate isomers with dimethyl disulfide adducts by GC/MS. *Lipids*, *51*(9), 1077–1081. https://doi.org/10.1007/s11745-016-4180-7

Sigman, D. M., & Casciotti, K. L. (2001). Nitrogen Isotopes in the Ocean. In *Encyclopedia of Ocean Sciences* (pp. 1884–1894). Elsevier. https://doi.org/10.1006/rwos.2001.0172

Siliakus, M. F., van der Oost, J., & Kengen, S. W. M. (2017). Adaptations of archaeal and bacterial membranes to variations in temperature, pH and pressure. *Extremophiles : Life under Extreme Conditions*, *21*(4), 651–670. PubMed. https://doi.org/10.1007/s00792-017-0939-x

Simon, K. S., Benfield, E. F., & Macko, S. A. (2003). FOOD WEB STRUCTURE AND THE ROLE OF EPILITHIC BIOFILMS IN CAVE STREAMS. *Ecology*, *84*(9), 2395–2406. https://doi.org/10.1890/02-334

Smart, P. L., Beddows, P. A., Coke, J., Doerr, S., Smith, S., & Whitaker, F. F. (2006). Cave development on the Caribbean coast of the Yucatan Peninsula, Quintana Roo, Mexico. In R. S. Harmon & C. M. Wicks, *Perspectives on Karst Geomorphology, Hydrology, and Geochemistry—A Tribute Volume to Derek C. Ford and William B. White*. Geological Society of America. https://doi.org/10.1130/2006.2404(10)

Socki, R. A., Perry, E. Jr. C., & Romanek, C. S. (2002). Stable isotope systematics of two cenotes from the northern Yucatan Peninsula, Mexico. *Limnology and Oceanography*, *47*(6), 1808–1818. https://doi.org/10.4319/lo.2002.47.6.1808

Spormann, A. M., & Thauer, R. K. (1988). Anaerobic acetate oxidation to CO2 by Desulfotomaculum acetoxidans. *Archives of Microbiology*, *150*(4), 374–380. https://doi.org/10.1007/BF00408310

Stahl, P. D., & Klug, M. J. (1996). Characterization and differentiation of filamentous fungi based on fatty acid composition. *Applied and Environmental Microbiology*, *62*(11), 4136–4146. PubMed.

Staley, J. T., & Gosink, J. J. (1999). Poles Apart: Biodiversity and Biogeography of Sea Ice Bacteria. *Annual Review of Microbiology*, *53*(1), 189–215. https://doi.org/10.1146/annurev.micro.53.1.189

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., Chow, C.-E. T., Sachdeva, R., Jones, A. C., Schwalbach, M. S., Rose, J. M., Hewson, I., Patel, A., Sun, F., Caron, D. A., & Fuhrman, J. A. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal*, *5*(9), 1414–1425. https://doi.org/10.1038/ismej.2011.24

Stern, J. C., Sutter, B., Freissinet, C., Navarro-González, R., McKay, C. P., Archer, P. D., Jr., Buch, A., Brunner, A. E., Coll, P., Eigenbrode, J. L., Fairen, A. G., Franz, H. B., Glavin, D. P., Kashyap, S., McAdam, A. C., Ming, D. W., Steele, A., Szopa, C., Wray, J. J., … Team, M. S. L. S. (2015). Evidence for indigenous nitrogen in sedimentary and aeolian deposits from the Curiosity rover investigations at Gale crater, Mars. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(14), 4245–4250. PubMed. https://doi.org/10.1073/pnas.1420932112

Stevens, T. (1997). Lithoautotrophy in the subsurface. *FEMS Microbiology Reviews*, *20*(3–4), 327–337. https://doi.org/10.1111/j.1574-6976.1997.tb00318.x

Stoessell, R. K. (1992). Effects of Sulfate Reduction on CaCO3 Dissolution and Precipitation in Mixing-Zone Fluids. *SEPM Journal of Sedimentary Research*, *Vol. 62*. https://doi.org/10.1306/D4267A00-2B26-11D7-8648000102C1865D

Stoessell, R. K. (1995). Dampening of Transverse Dispersion in the Halocline in Karst Limestone in the Northeastern Yucatan Peninsula. *Ground Water*, *33*(3), 366–371. https://doi.org/10.1111/j.1745-6584.1995.tb00291.x

Stoessell, R. K., Moore, Y. H., & Coke, J. G. (1993). The Occurrence and Effect of Sulfate Reduction and Sulfide Oxidation on Coastal Limestone Dissolution in Yucatan Cenotes. *Ground Water*, *31*(4), 566–575. https://doi.org/10.1111/j.1745-6584.1993.tb00589.x

Stoessell, R. K., Ward, W. C., Ford, B. H., & Schuffert, J. D. (1989). Water chemistry and CaCO3 dissolution in the saline part of an open-flow mixing zone, coastal Yucatan Peninsula, Mexico. *GSA Bulletin*, *101*(2), 159–169. https://doi.org/10.1130/0016-7606(1989)101<0159:WCACDI>2.3.CO;2

Strom, S. L. (2008). Microbial Ecology of Ocean Biogeochemistry: A Community Perspective. *Science*, *320*(5879), 1043–1045. https://doi.org/10.1126/science.1153527

Sturt, H. F., Summons, R. E., Smith, K., Elvert, M., & Hinrichs, K.-U. (2004). Intact polar membrane lipids in prokaryotes and sediments deciphered by high-performance liquid chromatography/electrospray ionization multistage mass spectrometry—New biomarkers for biogeochemistry and microbial ecology. *Rapid Communications in Mass Spectrometry*, *18*(6), 617–628. https://doi.org/10.1002/rcm.1378

Suárez-Moo, P., Remes-Rodríguez, C. A., Márquez-Velázquez, N. A., Falcón, L. I., García-Maldonado, J. Q., & Prieto-Davó, A. (2022). Changes in the sediment microbial community structure of coastal and inland sinkholes of a karst ecosystem from the Yucatan peninsula. *Scientific Reports*, *12*(1), 1110. https://doi.org/10.1038/s41598-022-05135-9

Tetu, S. G., Breakwell, K., Elbourne, L. D. H., Holmes, A. J., Gillings, M. R., & Paulsen, I. T. (2013). Life in the dark: Metagenomic evidence that a microbial slime community is driven by inorganic nitrogen metabolism. *The ISME Journal*, *7*(6), 1227–1236. https://doi.org/10.1038/ismej.2013.14

Thrash, J. C., Seitz, K. W., Baker, B. J., Temperton, B., Gillies, L. E., Rabalais, N. N., Henrissat, B., & Mason, O. U. (2017). Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico "Dead Zone." *MBio*, *8*(5), e01017-17. https://doi.org/10.1128/mBio.01017-17

Tomczyk-Żak, K., & Zielenkiewicz, U. (2016). Microbial Diversity in Caves. *Geomicrobiology Journal*, *33*(1), 20–38. https://doi.org/10.1080/01490451.2014.1003341

Torres-Talamante, O., Alcocer, J., Beddows, P. A., Escobar-Briones, E. G., & Lugo, A. (2011). The key role of the chemolimnion in meromictic cenotes of the Yucatan Peninsula, Mexico. *Hydrobiologia*, *677*(1), 107–127. https://doi.org/10.1007/s10750-011-0746-9

Tourna, M., Stieglmeier, M., Spang, A., Könneke, M., Schintlmeister, A., Urich, T., Engel, M., Schloter, M., Wagner, M., Richter, A., & Schleper, C. (2011). *Nitrososphaera viennensis* , an ammonia oxidizing archaeon from soil. *Proceedings of the National Academy of Sciences*, *108*(20), 8420–8425. https://doi.org/10.1073/pnas.1013488108

Uritskiy, G. V., DiRuggiero, J., & Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*, *6*(1), 158. https://doi.org/10.1186/s40168-018-0541-1

van der Ha, D., Vanwonterghem, I., Hoefman, S., De Vos, P., & Boon, N. (2013). Selection of associated heterotrophs by methane-oxidizing bacteria at different copper concentrations. *Antonie van Leeuwenhoek*, *103*(3), 527–537. https://doi.org/10.1007/s10482-012-9835-7

van Hengstum, P. J., Reinhardt, E. G., Beddows, P. A., Schwarcz, H. P., & Gabriel, J. J. (2009). Foraminifera and testate amoebae (thecamoebians) in an anchialine cave: Surface distributions from Aktun Ha (Carwash) cave system, Mexico. *Limnology and Oceanography*, *54*(1), 391–396. https://doi.org/10.4319/lo.2009.54.1.0391

Van Niel, C. B. (1949). THE "DELFT SCHOOL" AND THE RISE OF GENERAL MICROBIOLOGY. *Bacteriological Reviews*, *13*(3), 161–174. https://doi.org/10.1128/br.13.3.161-174.1949

Vartoukian, S. R., Palmer, R. M., & Wade, W. G. (2010). Strategies for culture of 'unculturable' bacteria: Culturing the unculturable. *FEMS Microbiology Letters*, no-no. https://doi.org/10.1111/j.1574-6968.2010.02000.x

Villarino, E., Watson, J. R., Chust, G., Woodill, A. J., Klempay, B., Jonsson, B., Gasol, J. M., Logares, R., Massana, R., Giner, C. R., Salazar, G., Alvarez-Salgado, X. A., Catala, T. S., Duarte, C. M., Agusti, S., Mauro, F., Irigoien, X., & Barton, A. D. (2022). Global beta diversity patterns of microbial communities in the surface and deep ocean. *Global Ecology and Biogeography*, *31*(11), 2323–2336. https://doi.org/10.1111/geb.13572

Viver, T., Orellana, L., González-Torres, P., Díaz, S., Urdiain, M., Farías, M. E., Benes, V., Kaempfer, P., Shahinpei, A., Ali Amoozegar, M., Amann, R., Antón, J., Konstantinidis, K. T., & Rosselló-Móra, R. (2018). Genomic comparison between members of the

Salinibacteraceae family, and description of a new species of Salinibacter (Salinibacter altiplanensis sp. Nov.) isolated from high altitude hypersaline environments of the Argentinian Altiplano. *Systematic and Applied Microbiology*, *41*(3), 198–212. https://doi.org/10.1016/j.syapm.2017.12.004

Waite, D. W., Vanwonterghem, I., Rinke, C., Parks, D. H., Zhang, Y., Takai, K., Sievert, S. M., Simon, J., Campbell, B. J., Hanson, T. E., Woyke, T., Klotz, M. G., & Hugenholtz, P. (2017). Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed Reclassification to Epsilonbacteraeota (phyl. Nov.). *Frontiers in Microbiology*, *8*, 682. https://doi.org/10.3389/fmicb.2017.00682

Waters, A. C., Donnelly-Nolan, J. M., & Rogers, B. W. (1990). *Selected caves and lava tube systems in and near Lava Beds National Monument, California* (Report No. 1673; Bulletin, p. 116). USGS Publications Warehouse. https://doi.org/10.3133/b1673

Webb, H. K., Nguyen, S. H., & Ivanova, E. P. (2014). The Families Hahellaceae and Litoricolaceae. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes* (pp. 319–323). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38922-1_287

Wecke, T., & Mascher, T. (2011). Antibiotic research in the age of omics: From expression profiles to interspecies communication. *Journal of Antimicrobial Chemotherapy*, *66*(12), 2689–2704. https://doi.org/10.1093/jac/dkr373

White, W. B., Vito, C., & Scheetz, B. (2009). The mineralogy and trace element chemistry of black manganese oxide deposits from caves. *Journal of Cave and Karst Studies*, *71*(2), 136–143.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org

Willems, A. (2014). The Family Comamonadaceae. In E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, & F. Thompson (Eds.), *The Prokaryotes* (pp. 777–851). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-30197-1_238

Willers, C., Jansen van Rensburg, P. J., & Claassens, S. (2015). Phospholipid fatty acid profiling of microbial communities–a review of interpretations and recent applications. *Journal of Applied Microbiology*, *119*(5), 1207–1218. https://doi.org/10.1111/jam.12902

Worthington, S., Ford, D., & Beddows, P. A. (2000). In *Porosity and permeability enhancement in unconfined carbonate aquifers as a result of dissolution* (pp. 220–223). National Speleological Society of America.

Wright, J. J., Mewis, K., Hanson, N. W., Konwar, K. M., Maas, K. R., & Hallam, S. J. (2014). Genomic properties of Marine Group A bacteria indicate a role in the marine sulfur cycle. *The ISME Journal*, *8*(2), 455–468. https://doi.org/10.1038/ismej.2013.152

Wright, K. E., Williamson, C., Grasby, S. E., Spear, J. R., & Templeton, A. S. (2013). Metagenomic evidence for sulfur lithotrophy by Epsilonproteobacteria as the major energy source for primary productivity in a sub-aerial arctic glacial deposit, Borup Fiord Pass. *Frontiers in Microbiology*, *4*. https://doi.org/10.3389/fmicb.2013.00063

Wright, T. D., Vergin, K. L., Boyd, P. W., & Giovannoni, S. J. (1997). A novel delta-subdivision proteobacterial lineage from the lower ocean surface layer. *Applied and Environmental Microbiology*, *63*(4), 1441–1448. https://doi.org/10.1128/aem.63.4.1441-1448.1997

Xie, C., Goi, C. L. W., Huson, D. H., Little, P. F. R., & Williams, R. B. H. (2016). RiboTagger: Fast and unbiased 16S/18S profiling using whole community shotgun metagenomic or metatranscriptome surveys. *BMC Bioinformatics*, *17*(S19), 508. https://doi.org/10.1186/s12859-016-1378-x

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zavarzin, G. A. (2006). Winogradsky and modern microbiology. *Microbiology*, *75*(5), 501–511. https://doi.org/10.1134/S0026261706050018

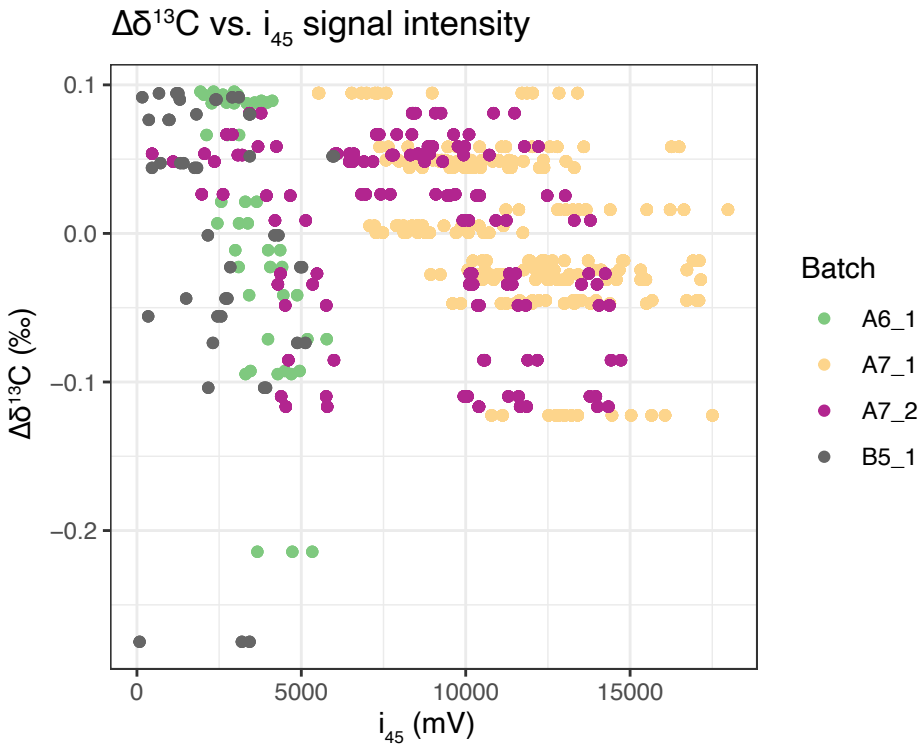
Zhang, Y.-G., Wang, H.-F., Yang, L.-L., Zhou, X.-K., Zhi, X.-Y., Duan, Y.-Q., Xiao, M., Zhang, Y.-M., & Li, W.-J. (2016). Egibacter rhizosphaerae gen. Nov., sp. Nov., an obligately halophilic, facultatively alkaliphilic actinobacterium and proposal of Egibaceraceae fam. Nov. And Egibacterales ord. Nov. *International Journal of Systematic and Evolutionary Microbiology*, *66*(1), 283–289. https://doi.org/10.1099/ijsem.0.000713

Zhao, R., Wang, H., Yang, H., Yun, Y., & Barton, H. A. (2017). Ammonia-Oxidizing Archaea Dominate Ammonia-Oxidizing Communities within Alkaline Cave Sediments. *Geomicrobiology Journal*, *34*(6), 511–523. https://doi.org/10.1080/01490451.2016.1225861

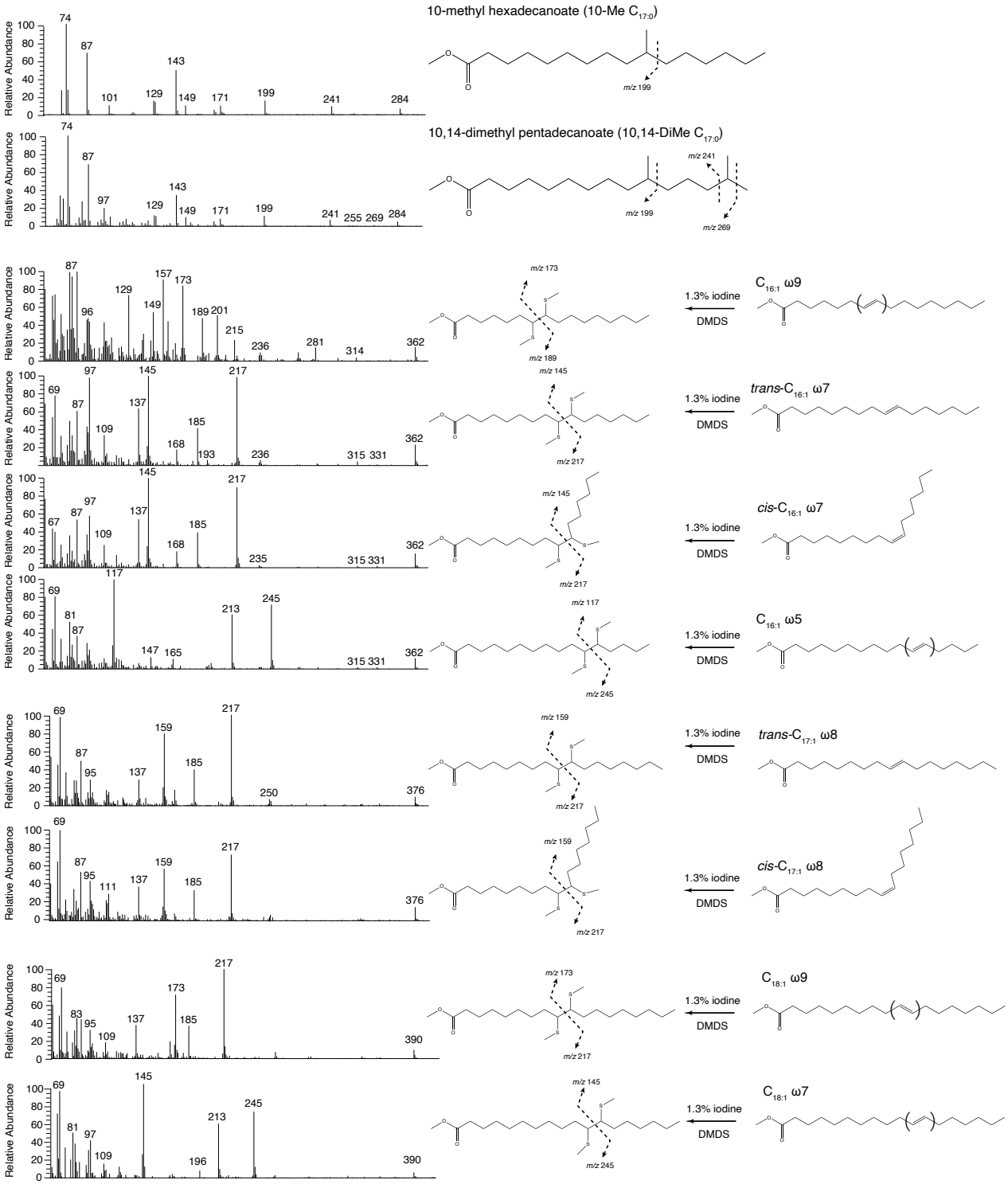
Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U., & Anantharaman, K. (2022). METABOLIC: High-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome*, *10*(1), 33. https://doi.org/10.1186/s40168-021-01213-8

Appendix

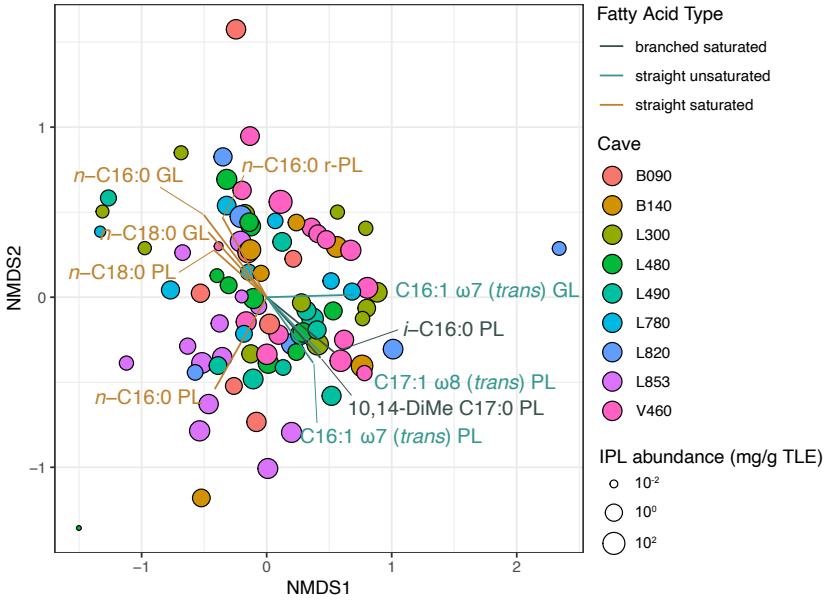
Appendix Figure 2-1: Dynamic range of GC-C-IRMS i_{45} signal intensity values. No clear relation was observed between signal peak intensity (i_{45}), measured in mV, and the residual values from the linear isotope scale conversion model ($\Delta\delta^{13}C$). Points are colored by analytical batch. A6, A7, and B5 external standard mixtures of C_{16}-C_{30} *n*-alkanes were obtained from A. Schimmelmann at Indiana University.



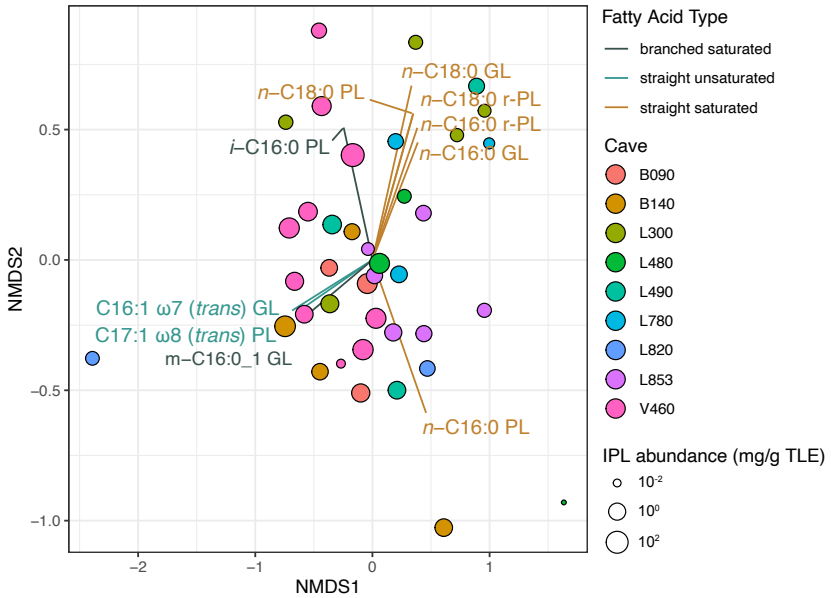
Appendix Figure 2-2: Mass spectra of selected fatty acid methyl esters and dimethyl disulfide (DMDS) derivatives. Unknown double bond conformations are represented with parentheses in the structure of the parent compound.



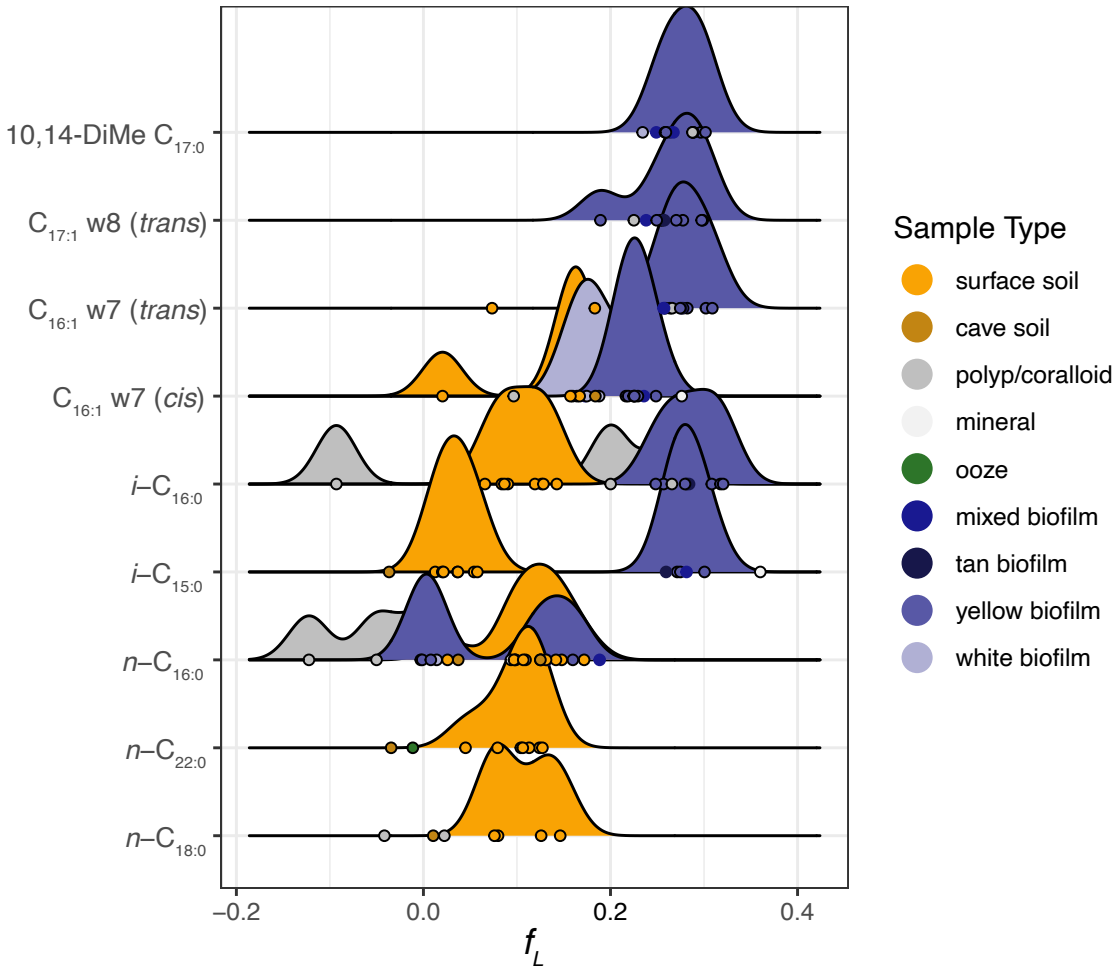
Appendix Figure 2-3: NMDS based on relative abundance, colored by cave. Solution stress = 0.199. The top 10 IPL-derived FA loadings ($p < 0.01$) are plotted to visualize the main sources of variance in the dataset. Point sizes are proportional to the total lipid yield across IPL fractions (mg/g total lipid extract, TLE).



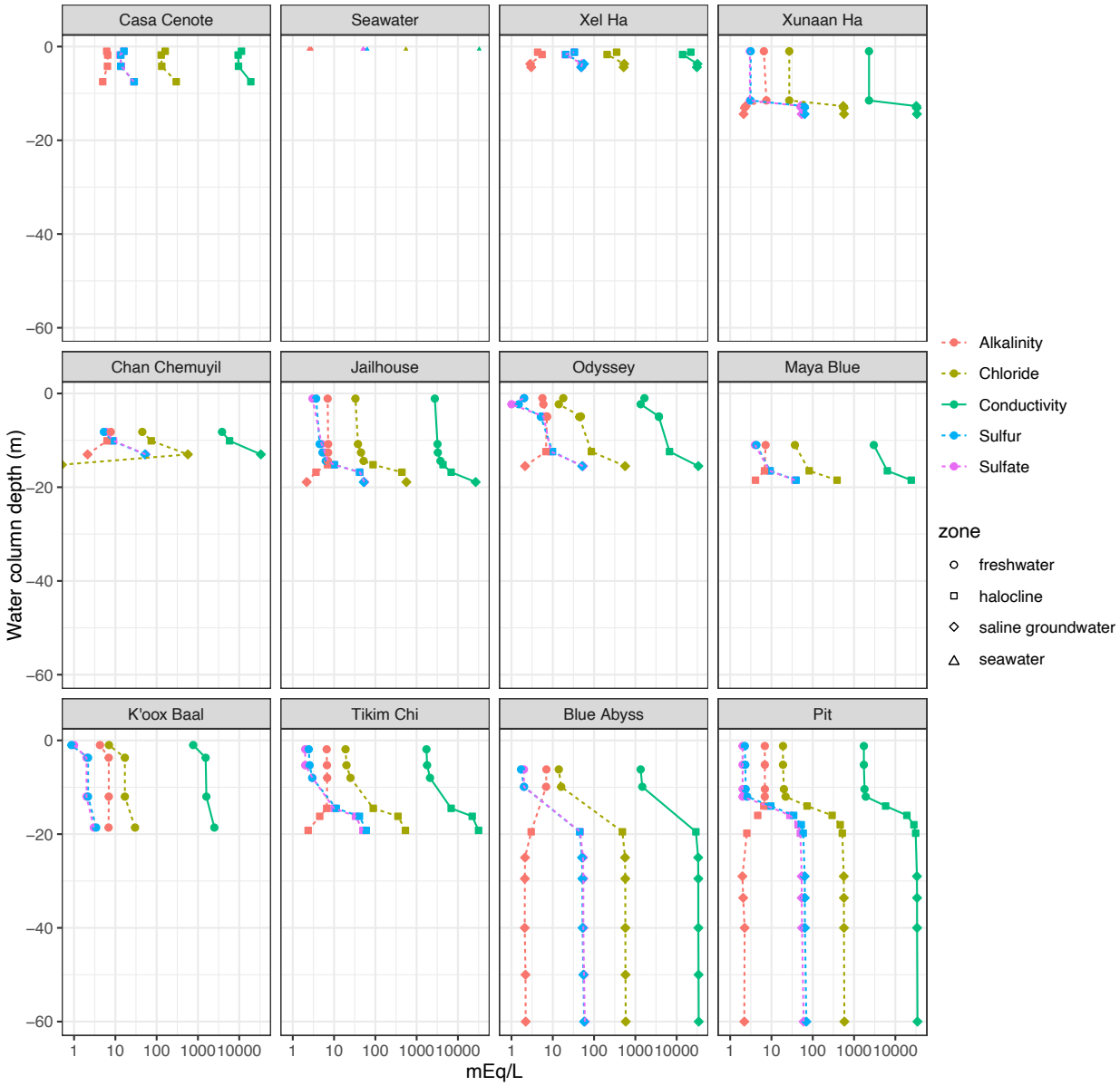
Appendix Figure 2-4: Relative abundance NMDS of polyp/coralloid and mineral samples only, colored by cave. Solution stress = 0.154. The top 10 IPL-derived FA loadings ($p < 0.01$) are plotted to visualize the main sources of variance in the dataset. Point sizes are proportional to the total lipid yield across IPL fractions (mg/g total lipid extract, TLE).



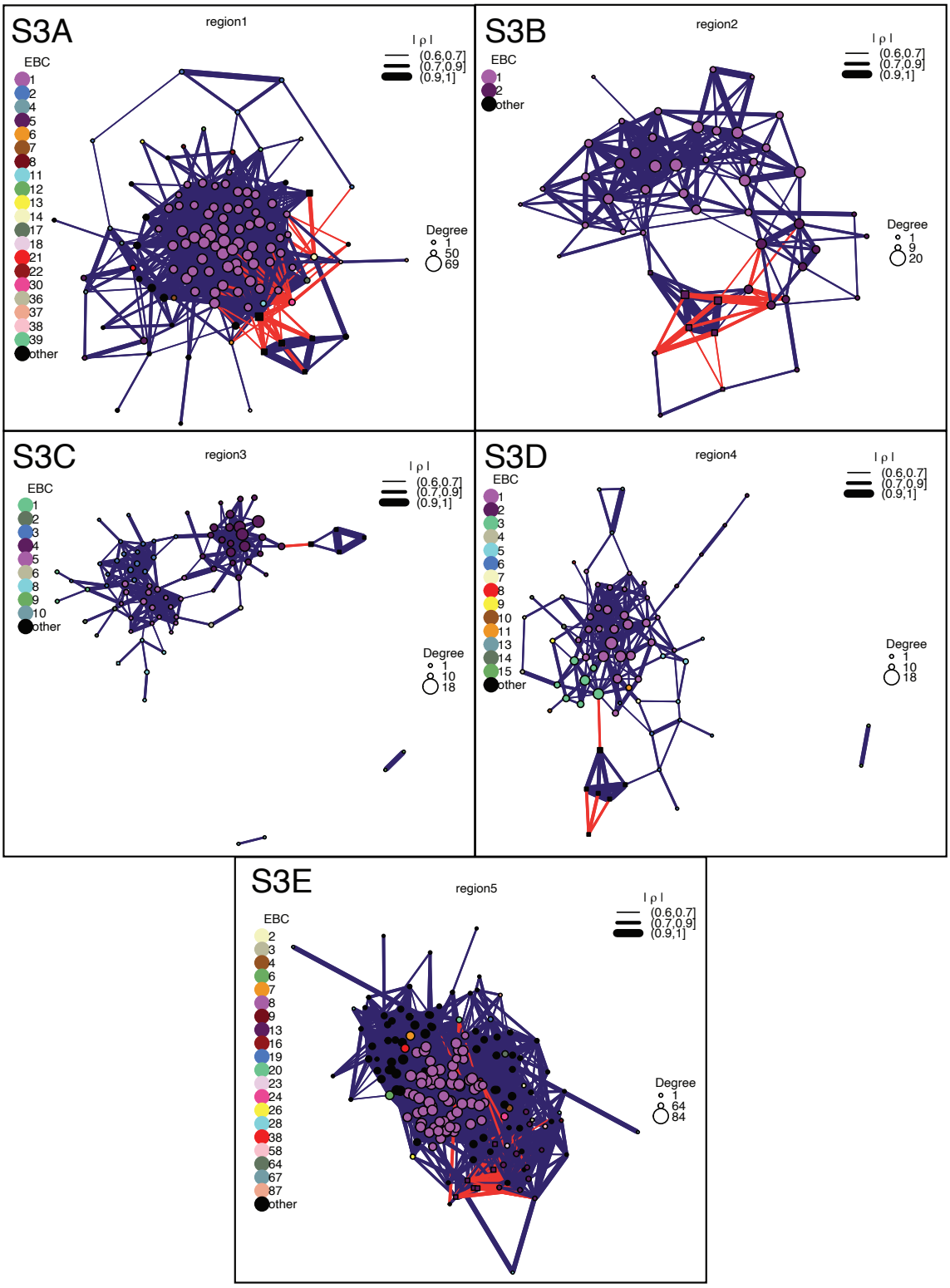
Appendix Figure 2-5: Fraction of reductive acetyl-CoA pathway-based lithoautotrophy required by isotopic mass balance. The term f_L is the fraction of the observed $\delta^{13}C_{IPL}$ signal attributable to reductive acetyl-CoA pathway-based lithoautotrophy, calculated from Equation 1 assuming a fractionation of 52‰.



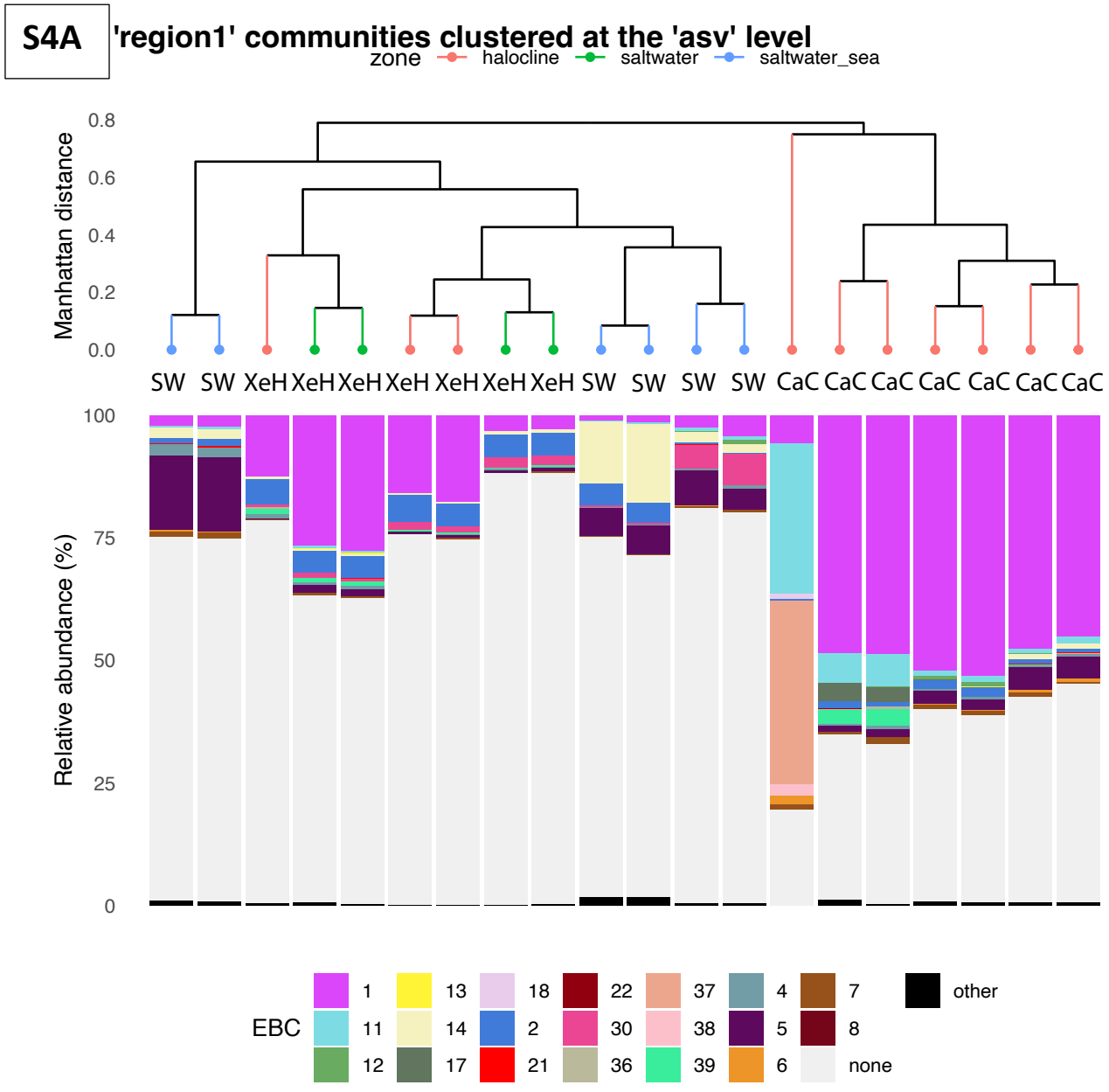
Appendix Figure 3-1: Geochemical variables across study sites. All values reported in milliequivalents per liter (mEq/L). Electrical conductivity is represented by a solid green line. Units of conductivity (originally in mS/cm) are converted to mEq/L by multiplying by 640 to scale with other measured variables in this plot.



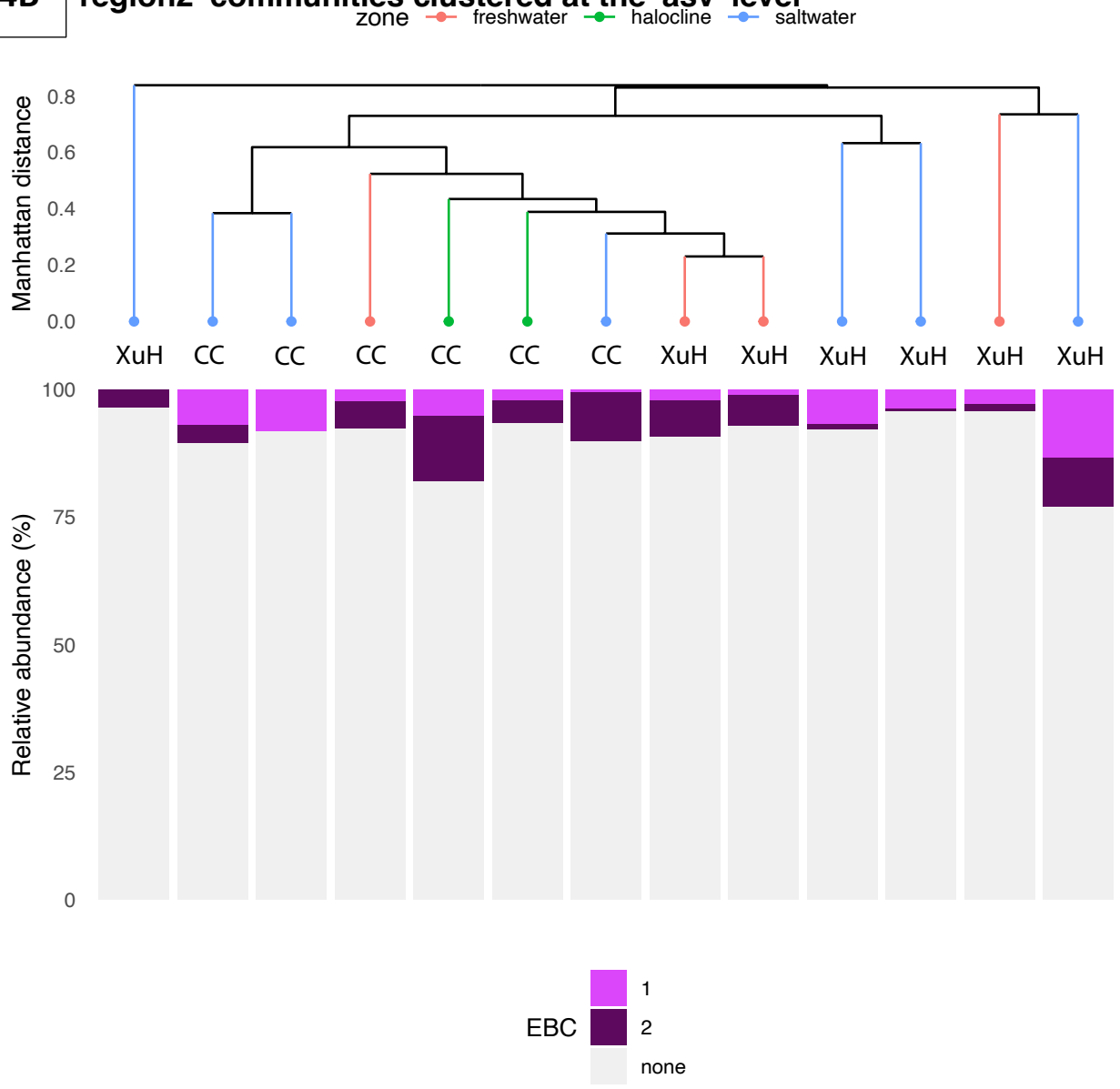
Appendix Figure 3-2: Regional co-occurrence networks. S3A: Region1. S3B: Region2. S3C: Region3. S3D: Region4. S3E: Region5. Refer to the main text for discussion.



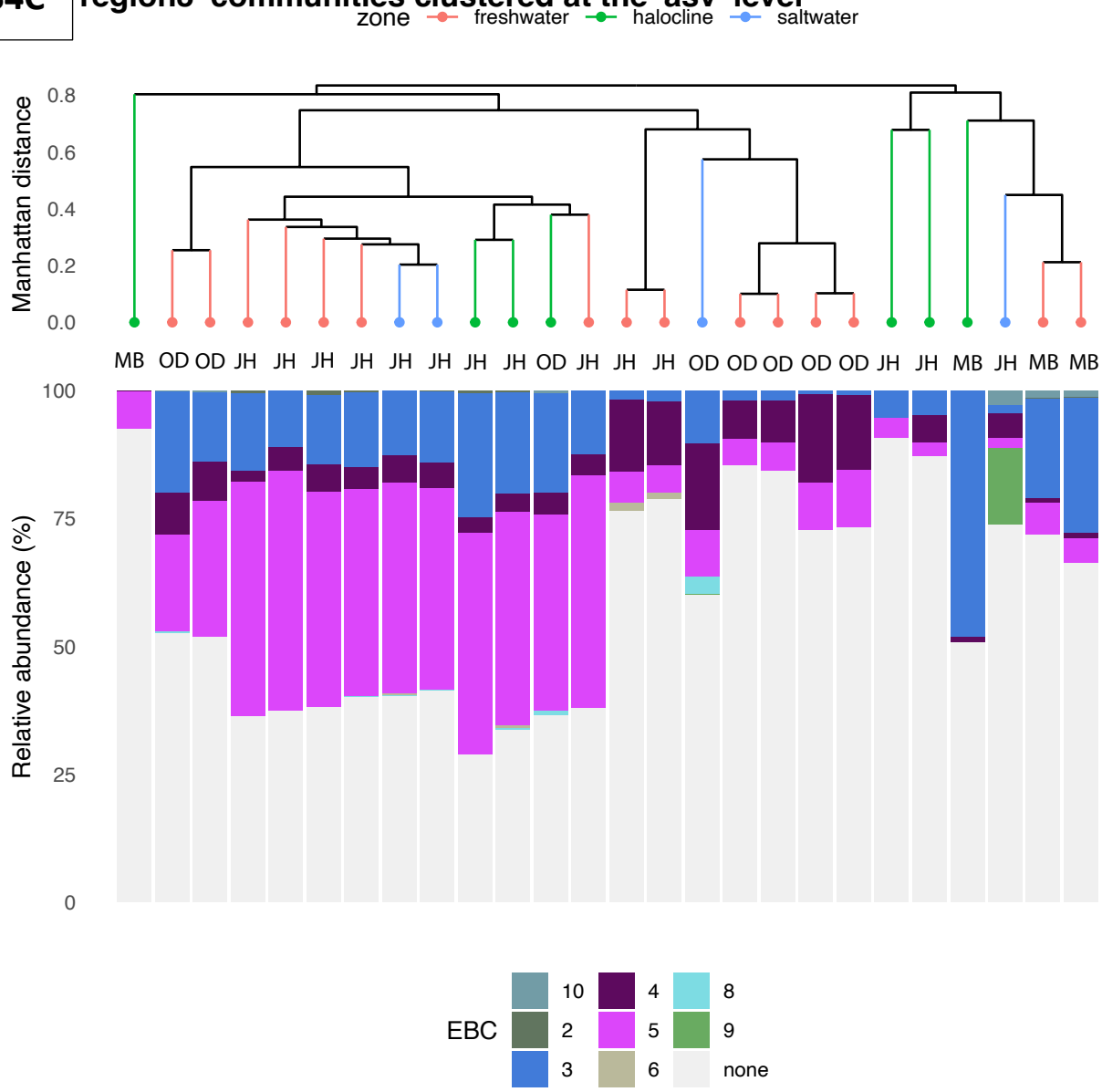
Appendix Figure 3-3: Relative abundance of regional network node EBCs. EBCs = edge betwenness clusters. Cave-specific codes are as follows (each plot is set on a separate page): S4A - Region 1: CaC = Casa Cenote, XeH = Xel Ha, SW = Seawater; S4B - Region 2: CC = Chan Chemuyil, XuH = Xunaan Ha; S4C - Region 3: MB = Maya Blue, OD = Odyssey, JH = Jailhouse; S4D - Region 4: KB = K'oox Baal, BA = Blue Abyss, TC = Tikim Chi; S4E - Region 5: PT = The Pit



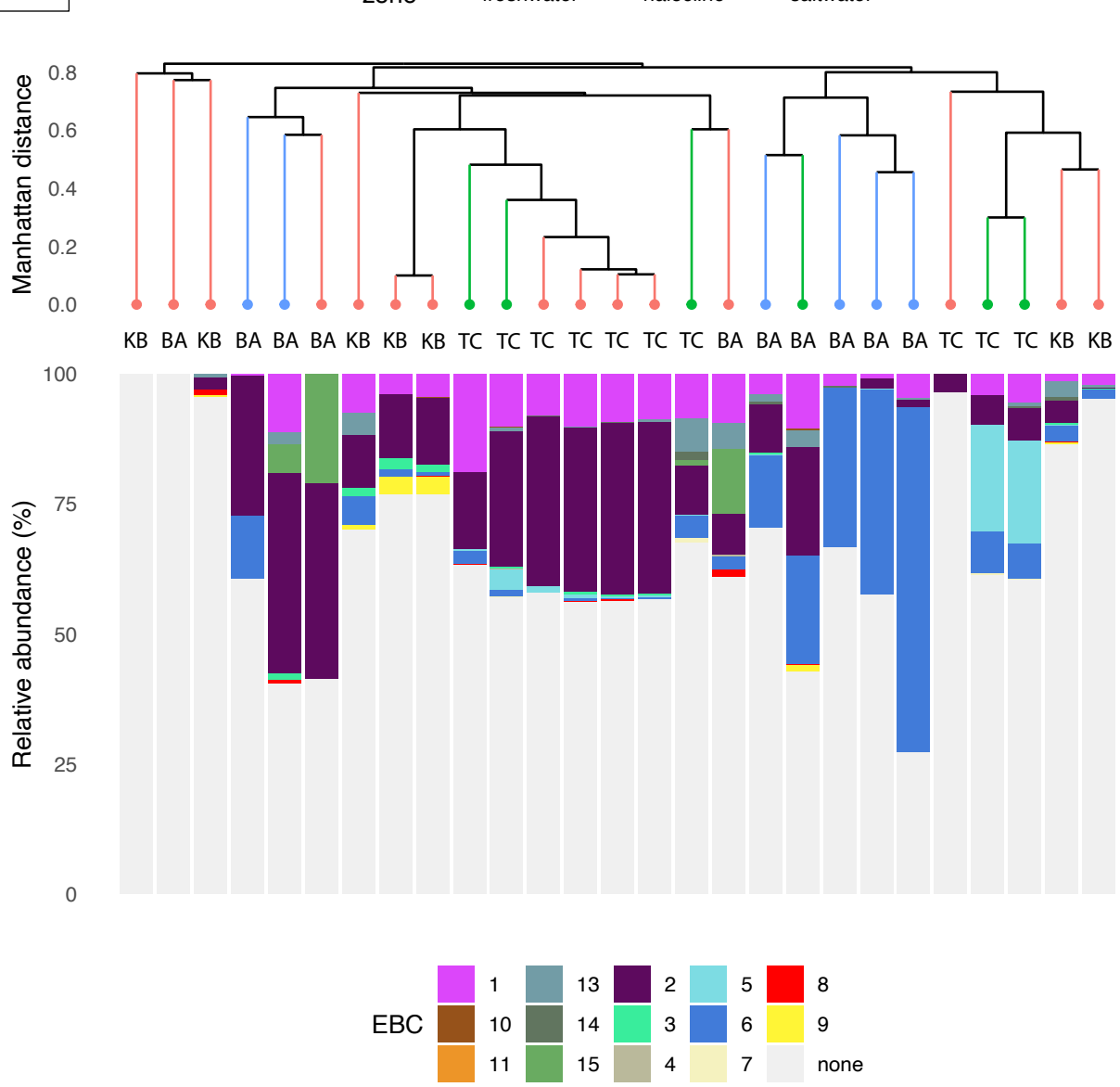
S4B 'region2' communities clustered at the 'asv' level



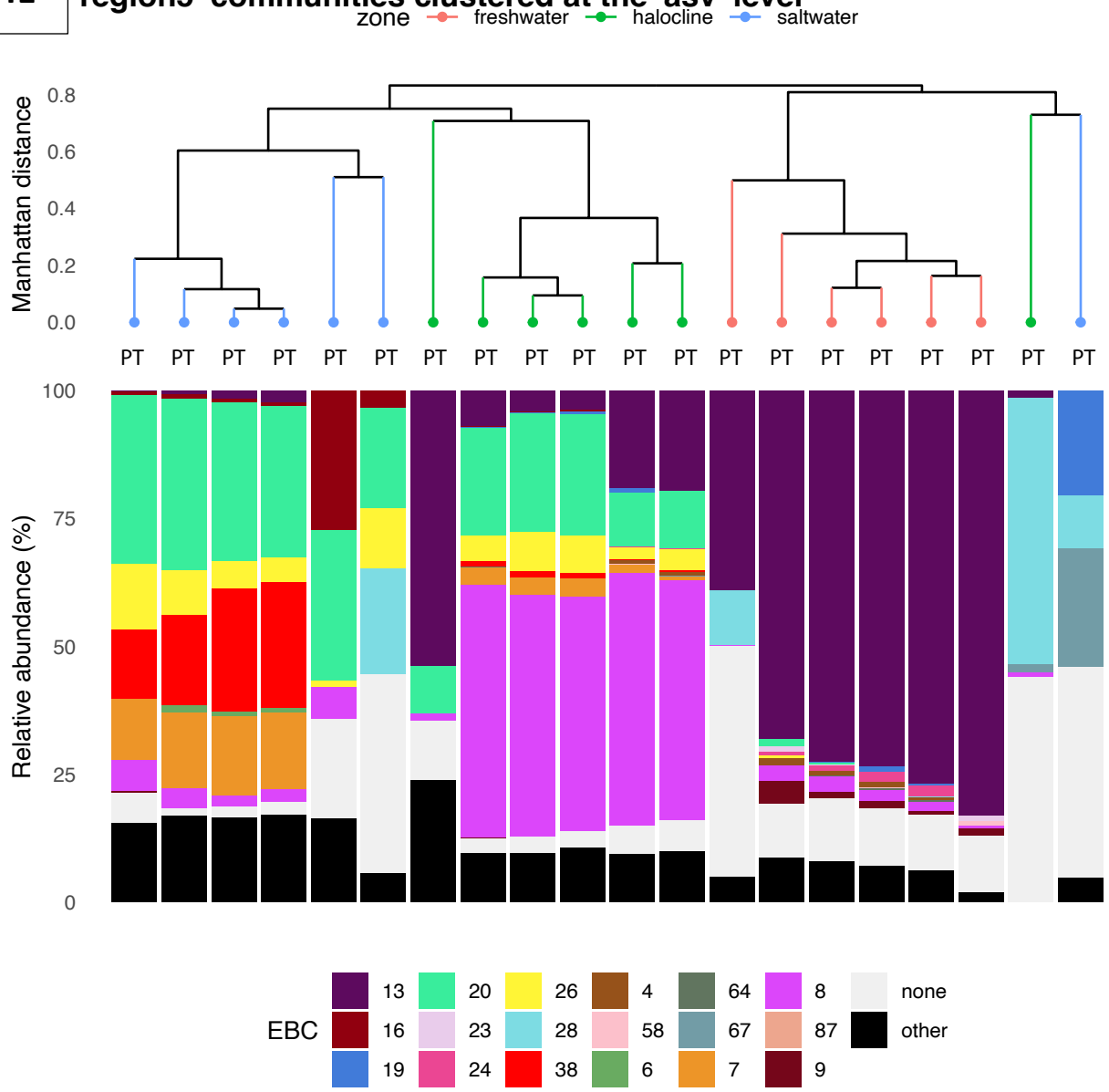
S4C 'region3' communities clustered at the 'asv' level



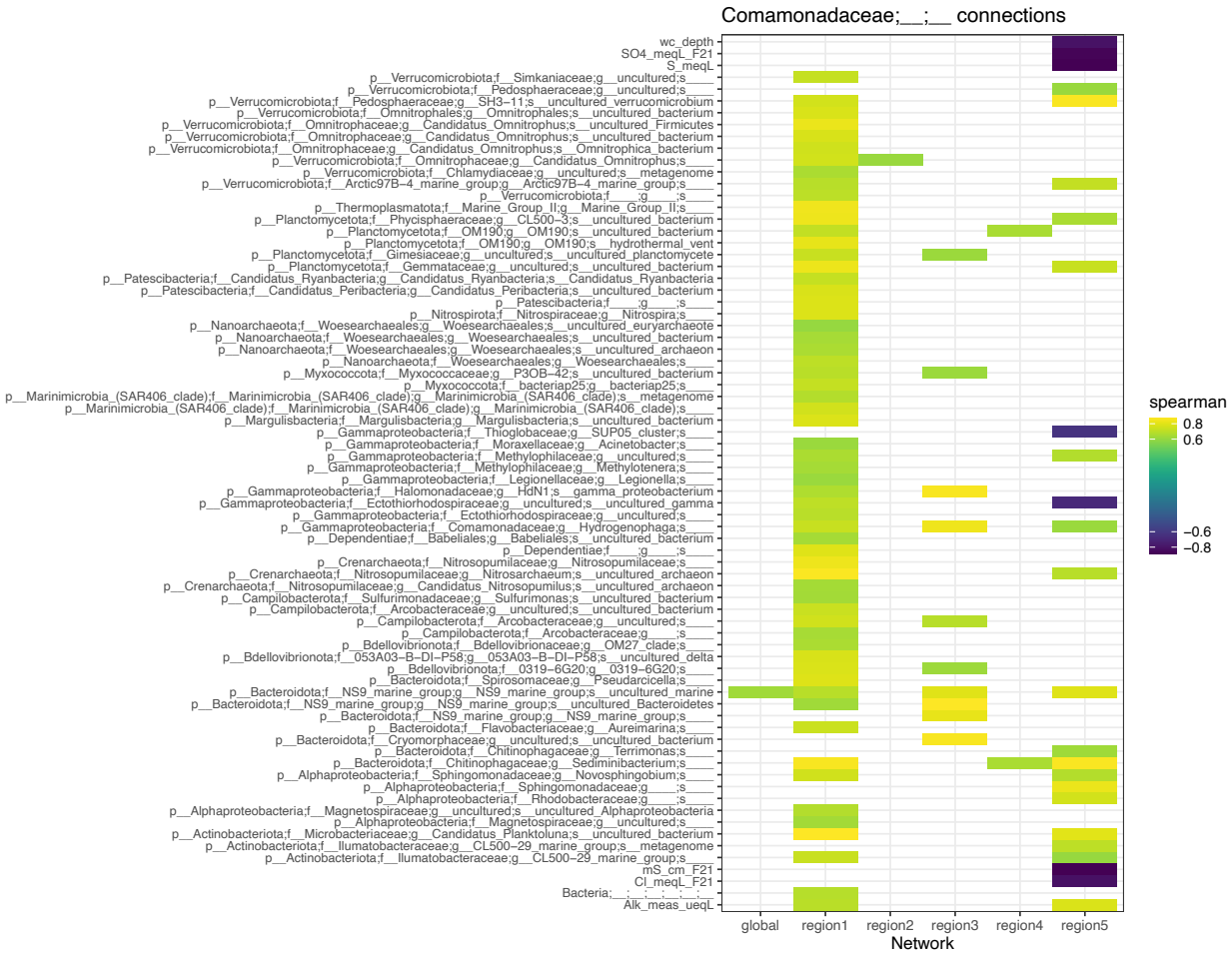
S4D 'region4' communities clustered at the 'asv' level



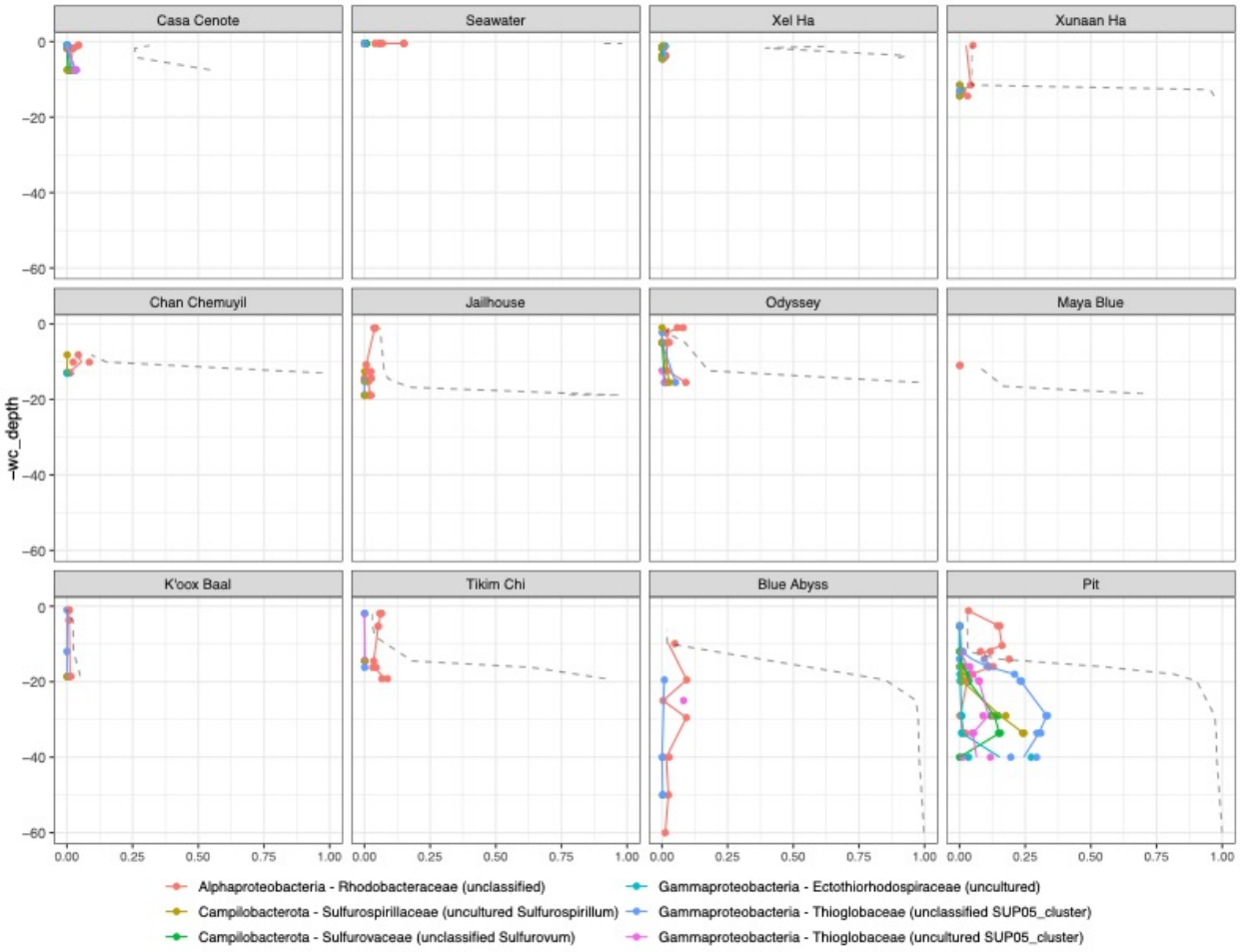
S4E 'region5' communities clustered at the 'asv' level



Appendix Figure 3-4: Co-occurrences of the unclassified* Comamonadaceae *bin with other network nodes across the global and regional networks.



Appendix Figure 3-5: Relative abundance of selected taxa putatively capable of sulfur cycling in the eastern Yucatán carbonate aquifer. Conductivity is represented as a dimensionless dashed line to visualize density stratification.



Appendix Figure 5-1: Spearman relationships of an uncultured* Euzebyaceae *ASV across separate networks with various taxa and environmental variables.