WIND CAVE: DIRECT ACCESS TO A DEEP SUBSURFACE AQUIFER REVEALS A DIVERSE MICROBIAL COMMUNITY AND UNUSUAL MANGANESE METABOLISM

A Dissertation

Presented to

The Graduate Faculty of The University of Akron

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

OLIVIA SUZANNE HERSHEY

December, 2021

WIND CAVE: DIRECT ACCESS TO A DEEP SUBSURFACE AQUIFER

REVEALS A DIVERSE MICROBIAL COMMUNITY

AND UNUSUAL MANGANESE METABOLISM

Olivia Suzanne Hershey

Dissertation

Approved:

____
Advisor
Dr. Hazel A. Barton

____
Committee Member
Dr. John M. Senko

____
Committee Member
Dr. R. Joel Duff

____
Committee Member
Dr. Michael C. Konopka

____
Committee Member
Dr. Zhong-Hui Duan

Accepted:

____
Program Director, Integrated Bioscience
Dr. Hazel A. Barton

____
Dean of the College
Dr. Mitchell S. McKinney

____
Interim Director of the Graduate School
Dr. Marnie M. Saunders

____
Date

ABSTRACT

Caves provide a unique environment for studying microbial ecology, providing a portal to the microbial communities of the terrestrial subsurface. Despite the geologic isolation and nutrient limitation of growth in the subsurface, caves contain remarkably diverse microbial communities, with unique adaptations that allow community subsistence and growth. At Wind Cave National Park, South Dakota, a series of lakes are formed at the intersection of Wind Cave and the regionally important Madison aquifer. These lakes (WCL), provide a rare natural window into the aquifer, and my research has demonstrated that they allow us to examine the microbial community of a karst aquifer without the sources of contamination often associated with surface drilling. Though the isolation (125 m below the surface) and long residence time (~25 years) of water *en route* to the lakes results in ultraoligotrophic conditions (0.29 mg L^{-1} TOC), the lakes support a stable and diverse community of microbes, albeit with cell numbers lower than almost any body of water on Earth (~2,300 cells mL^{-1}). This low biomass, combined with a reduced cell size as an adaptive strategy to survival in these nutrient limited conditions, made collecting sufficient cell mass for DNA based analyses problematic. I therefore optimized the standard techniques used to sample aquatic communities, using tangential flow filtration to filter more than 1,000 L of water from the Madison aquifer, through a 45 nm-pore size membrane allowing the capture of even the smallest cells within this microbial ecosystem. Metagenomic sequencing combined with comparative filtration revealed that WCL was enriched in ultrasmall cells, such as those found in the Patescibacteria and Nitrospirota. Evidence of integron-facilitated genetic plasticity suggests that metabolic flexibility is an important mechanism for adaptation and survival in WCL. Finally, our metagenomic and phylogenetic data suggest that manganese plays a central role in primary production and carbon turnover in WCL, and that primary production in the community is based on chemolithotrophic Mn(II) oxidation. Not only could this be the first Mn(II)-based microbial community described, but may provide a unique ecosystem in which to understand the important drivers of the Mn biogeochemical cycle within the subsurface.

ACKNOWLEDGEMENTS

First and foremost, I am extremely grateful to my advisor, Dr. Hazel Barton, for recognizing my strengths and potential when I could not see them myself, and for her patience while my self-confidence caught up with reality. She helped as much with my personal growth as she did with this research. I'd also like to extend a special thanks to Dr. John Senko for the valuable experience I've gained working with him over the past year. Thanks to Hazel and John for also being on my committee with Dr. Joel Duff, Dr. Michael Konopka, and Dr. Zhong-Hui Duan, all of whom provided insightful questions, comments, and advice regarding my research.

The list of people and organizations that made this research possible is extensive, and I can only name a few here. Many thanks to the staff at Wind Cave National Park that carried sampling equipment to and from the Wind Cave lakes, with a special thanks to Marc Ohms for his help organizing our sampling trips. Katey Bender collected many of my samples at the lakes, and I knew I could trust her to treat them as carefully as if it were her own research. Samples were sequenced with the help of Riley Drake, who was a joy to meet when she came to Wind Cave to help us collect samples and work on some of her own research. I

am also incredibly thankful for the funding I received from the National Park Service, the Cleveland Grotto Science Fund, the Choose Ohio First Bioinformatics Fellowship, and most significantly, the Cave Conservancy Foundation Academic Fellowship in Karst Studies.

Finally, the completion of this dissertation would not have been possible without the support of my friends, colleagues, and family. I owe my sanity to Kayla Calapa and her incredible ability to show me the objective side of any situation, and to Ceth Parker for showing me how to have fun and that there is always joy and wonder in science. To Charbel, Katey, David, Shagun, Reed, and all the other friends I cannot count, please know that every bit of advice, constructive criticism, and encouragement was deeply appreciated, and I could not have made it here without you. Last, but certainly not least—Nathan, my constant companion, I'd thank you with all my heart, but you've been holding it this whole time.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Prior to the advent of molecular biology, microbial ecology relied heavily on cultivation-dependent techniques to understand microbial activities in the environment. Early cave researchers used the same cultivation techniques as soil scientists and (somewhat unsurprisingly) found that caves were a weak reflection of the microbiology of surface soils (Hess, 1900; Scott, 1909; Høeg, 1946; Caumartin, 1963). The interpretation of microbial activity in caves was therefore limited, and it seemed to be of little to interest the scientific community, with less than 40 papers published prior to 1997 (Hershey & Barton, 2018). Yet, these papers defined our understanding of cave microbiology, suggesting that caves were essentially lifeless due to an absence of photosynthetic input, or simply home to transient microbial species introduced by the activity of animals or humans (Caumartin, 1963). When endemic cave microorganisms were putatively identified it was through unusual metabolisms that were (incorrectly) thought to distinguish them from soil species, such as iron-oxidation (Caumartin, 1963).

1.1 Early Molecular Phylogenetics

The primary limitation of cultivation-based approaches is that the vast majority (>99%) of environmental microorganisms cannot be cultured; as in other microbial environments, the ability to accurately describe microbial diversity within caves required cultivation-independent techniques (Amann *et al*., 1996). Some early non-cultivation approaches did support the idea that microbiology in caves was more complex than originally thought: Fliermans *et al*. (1977) used antibodies to identify non-culturable *Nitrobacter* in Mammoth Cave sediments; the microscopic techniques of Cunningham *et al*., (1995) demonstrated a rich structural diversity from samples deep within Lechuguilla Cave; and Gonzalez *et al.* (1999) demonstrated a rich diversity of actinobacteria in Spanish caves using fatty acid methyl ester (FAME) analyses (Fliermans & Schmidt, 1977; Cunningham *et al*., 1995; Gonzalez *et al*., 1999). Nonetheless, it wasn't until the use of molecular phylogenetics in the 1990s that the potential diversity of microorganisms in cave environments emerged (Hershey & Barton, 2018).

Molecular phylogenetics for bacteria and archaea relies on utilizing the highly conserved 16S small ribosomal subunit rRNA (16S rRNA) gene sequence as a genetic marker (Woese, 1987). The 16S rRNA sequence is present and similar in structure in both bacteria and archaea; however, differences in highly conserved regions of the sequence resulted in the two groups being distinguished as separate kingdoms (Woese & Fox, 1977; Woese, 1987). Lesser-conserved regions of the sequence can further distinguish different major groups from each other, and the combination of conserved, variable, and hypervariable regions across the whole 16S rRNA sequence are used to identify organisms at the species level (Soergel *et al*., 2012). Longer 16S rRNA sequences provide a more robust identification of microorganisms at the species level, but hypervariable regions enable the reliable placement of some microbes to its genus with a 400 bp sequence, or placement to class or order with as few as 150 bp (Ludwig *et al*., 1998; Wang *et al*., 2007; Jeraldo *et al*., 2011).

The use of the 16S rRNA gene was used to classify many cultured bacterial and archaeal isolates. In the 1980s, this gene became a revolutionary tool for identifying microorganisms in the environment when Pace *et al.* used it to distinguish previously uncultured species (Stahl *et al*., 1984; Pace *et al*., 1986). Using this method, 16S rRNA sequences were PCR amplified from an environmental DNA sample using primers complimentary to regions of the 16S rRNA sequence highly conserved among both bacteria and archaea, commonly called “universal” primers (Hugenholtz *et al*., 1998; Watanabe *et al*., 2001; Baker *et al*., 2003). The PCR products were cloned into *Escherichia coli* and sequenced via Sanger sequencing (Pace, 1997). Sequences could then be aligned and compared with 16S rRNA sequences from known, cultured microbes and classified accordingly (Pace *et al*., 1986). Pace and colleagues also carried out the first molecular analysis of a microbial cave community to examine the filamentous biofilms of a sulfidic stream within Sulfur-River Cave, Kentucky (Angert *et al*., 1998). This study revealed the surprising dominance of the *Epsiloproteobacteria*, which were previously seen only in deep, oceanic hydrothermal systems; it was also the first clue to the important influence that members of this phylum have within sulfidic cave environments (Campbell *et al*., 2006). Most importantly, the study also demonstrated that microbial cave communities could be remarkably distinct from their surface counterparts (Angert *et al*., 1998).

While cloning 16S rRNA gene sequences was an improvement on cultivation-based approaches to explore microbial diversity, and enabled the identification of many unculturable species, the number of clones generated that needed to be sequenced to fully characterize a microbial community limited the efficacy of this technique (Muyzer *et al*., 1993; Hughes *et al*., 2001). While PCR creates billions of amplicons, each step of cloning (ligation, transformation, screening) is inefficient: a single environmental sample may contain thousands of species, and the amplicons with higher abundance in the sample have a higher chance of being successfully cloned. Sanger sequencing is slow compared to modern sequencing methods, requiring forward and reverse priming, as well as internal priming, often in triplicate. To avoid wasteful sequencing of clones with identical DNA inserts, clones can be prescreened with restriction digest analyses, further increasing the time, labor, and reagents required to characterize the microbial community (Muyzer *et al*., 1993). Even with prescreening, rare species are often not identified until thousands of clones have been isolated and sequenced (Hughes *et al*., 2001).

The ability to analyze cave communities using these approaches has been further complicated by the low biomass of these environments (routinely <10^6 cells gram^{-1}), along with a complex geochemistry, both of which interfered with the ability to obtain sufficient DNA for analysis (Barton *et al*., 2006). Because of this, molecular phylogenetics in caves was limited to labs with both the molecular expertise and computing resources necessary to translate genetic difference into robust phylogenies, and by the end of the 1990s, only two labs were using such techniques in caves (Vlasceanu *et al*., 1997; Angert *et al*., 1998).

1.2 Next-Generation 16S rRNA Amplicon Sequencing

Among the many impacts of the Human Genome Project, the most powerful was the development of optically-based sequencing methods - collectively referred to as 'next-generation sequencing' (NGS) technologies (Ansorge, 2009; Lander, 2011). The dramatic increase in the number of bases that these technologies could sequence (>15 billion bases in as little as 4 hours) combined with their significant cost reductions, revolutionized the ability to sequence DNA (Snyder *et al*., 2009; Forde & O'Toole, 2013). Sogin *et al.* (2006) were the first to use NGS to identify environmental 16S rRNA; rather than restricting the identification of phylotypes within a community to a few hundred cloned 16S rRNA genes, NGS allowed Sogin and colleagues to sequence 120,000 PCR products directly (Sogin *et al*., 2006). The results were transformative and demonstrated that microbial ecosystems contained thousands of previously unidentified phylotypes (Sogin *et al*., 2006). Sogin *et al.* referred to this extensive collection of previously unidentified microorganisms as the 'rare biosphere' - organisms of sufficiently low number that they cannot be identified without deep-sequencing NGS approaches (Sogin *et al*., 2006).

The first study to apply NGS technology in cave environments were Ortiz *et al.* (2013) who used 454-pyrosequencing to examine ~400,000 PCR products from Kartchner Caverns, USA (Ortiz *et al*., 2013). Along with the 13 phyla already identified in caves by cloning approaches, Ortiz *et al.* demonstrated the presence of an additional 8 described and 12 candidate phyla, suggesting that caves also contained rare biosphere microorganisms (Ortiz *et al*., 2013). In a significant step forward, these researchers also used NGS to compare microbial communities in the cave with those in surface soils directly above. These data demonstrated that only 16% of the sequences were shared between the surface and the cave, confirming the uniqueness of microbial cave ecosystems (Ortiz *et al*., 2013). An analysis of several 16S rRNA datasets obtained from across multiple cave systems with broadly distributed geographical locations, including North America and Asia, confirm the robustness of the 13 dominant phyla already identified, along with another 14 phyla consistently represented in these populations (above a 0.1% threshold); these include the *Armatimonadetes* (OP10)*, Chlorobi, Cyanobacteria, Elusimicrobia, Spirochaetes* and the candidate phyla BRC1, GN04, NC10, OP3 (Ca*. Omnitrophica*), TM6 (Ca. *Dependentiae*), WS1 and WS3 (Ca.

Latescibacteria). Together these data support the existence of *rare biosphere* species within caves (Hershey & Barton, 2018).

While targeted PCR amplification using NGS makes it possible to rapidly screen the 16S rRNA sequences in the environment, it is also susceptible to significant technical issues, including primer and amplification biases that preferentially select for (or against) certain rRNA sequences for amplification (Chandler *et al*., 1997; Polz & Cavanaugh, 1998; DeSantis *et al*., 2007; Kembel *et al*., 2012). Notably, the standard V4 variable region primers used by the Earth Microbiome Project protocol (Caporaso *et al*., 2011) were modified to add degeneracy to the sequence to remove known biases against the *Thaumarchaeota* and SAR11 clades in 2016 (Apprill *et al*., 2015; Parada *et al*., 2016). While this sequence-based bias was resolved by primer modification, PCR bias from very high or very low GC content template DNA, copy number variation, or other PCR artifacts remain a limitation to amplification-based community analysis (Acinas *et al*., 2005; Krehenwinkel *et al*., 2017).

1.3 Metagenomics

Overcoming the limitations of PCR bias requires bypassing the PCR amplification step entirely and sequencing the sum of the genetic information in the environment (Miller *et al*., 2011). This process requires randomly fragmenting DNA into sizes appropriate for NGS sequencing (35 – 300 bp), either by mechanical means or using transposons (Adey *et al*., 2010). These fragments are then sequenced and the overlapping ends are computationally re-assembled back into a full-length DNA contigs, ranging in size from a few hundred to millions of bases - a technique referred to as shotgun sequencing due to the randomness of the initial DNA fragmentation (Sanger *et al*., 1977; Adey *et al*., 2010). Prior to the advent of NGS, shotgun methods were not possible using environmental DNA as the complexity of the samples reduced the likelihood of obtaining sufficient coverage for assembly (Venter *et al*., 2004). But the size of the data that could be obtained by NGS dramatically increased sequence coverage, making it possible to examine all of the genes in an environment rather than just one – a technique called metagenomics (Handelsman *et al*., 1998). Such metagenomic approaches allow the interactions that support microbial ecosystem dynamics to be identified through the functional gene composition of the community (Handelsman, 2004; Tyson *et al*., 2004; Venter *et al*., 2004).

Carrying out metagenomic approaches in oligotrophic caves has been problematic, primarily due to the significant amounts of DNA that were needed to create shotgun libraries in the past (from a minimum of a few hundred nanograms to multiple micrograms, depending on the method; Thomas *et al*., 2012). Library preparation with low amounts of DNA is prone to technical issues, including a greater potential for contamination, limited capture of DNA complexity, and a high percentage of read duplicates (Rinke *et al*., 2016). Despite these limitations, in 2014 Ortiz *et al.* were able to carry out metagenomic analyses of the microbial communities within Kartchner Caverns (Ortiz *et al*., 2014). Their data identified over 365,000 gene fragments from the microbial populations found on speleothems and walls within the cave and demonstrated the enrichment of genes involved carbohydrate metabolism and CO_2 fixation (Ortiz *et al*., 2014). The enrichment of these genes suggested that both heterotrophic and autotrophic metabolic activity were important in community growth and subsistence, along with potentially novel mechanisms of nutrient cycling, especially in regard to nitrogen (Ortiz *et al*., 2014).

Since Ortiz *et al.* (2014), significant advancements in library preparation techniques and technology have enabled shotgun sequencing of samples with sub-nanograms amounts of template DNA. For example, the protocol for the Nextera XT library preparation kit, originally intended for use with 1 ng of template DNA from small genomes, plasmids, or amplicons, can now be modified for use with as little as 0.1 pg template DNA from a microbial community. This technique was validated using a mock community of 54 bacterial and archaeal reference species, and then applied to DNA from samples of seawater ranging from 10μL – 1000 μL, with reproducible sequenced results (Rinke *et al*., 2016). The number of reads obtained from these low-input sequencing libraries is inadequate for all but the simplest microbial communities, especially if the goal of the sequencing is to assemble the short reads into longer contigs for identification of open reading frames, full gene sequences, and gene cassettes (Rinke *et al*., 2016). As community complexity increases, so does the amount of DNA required to achieve the high sequence coverage required for contig assembly from short reads; more DNA is also required to ensure sequencing of rare species in the community (Lynch & Neufeld, 2015; Rinke *et al*., 2016). Nonetheless, a low-input DNA library can still be informative through assignment of short sequences via taxonomic and functional signatures, enabling the characterization of low biomass environments that were previously only explored through 16S rRNA amplicon sequencing (Tringe *et al*., 2005).

The decreasing cost of sequencing, increased efficacy of library preparation, expansion of technology for computing in cloud and remote servers, and inexpensive data storage options, have made studying microbial ecology through metagenomics more accessible than ever (Teeling & Glockner, 2012; Chen *et al*., 2019). Pipelines and software used for bioinformatics analysis often differ between research groups based on preference and specific research questions, and open-sourced tools are constantly being created and updated. Despite differing software used, the basic elements of metagenomic analyses include, but are not limited to: assembly of sequencing reads, prediction of genes/proteins from these assemblies, and annotation of the predicted reads based on similarities to items in reference databases (Roumpeka *et al*., 2017). Researchers may choose to process data locally (or use supercomputer clusters for programs requiring significant memory), or use web application servers, or some combination of each (Wooley *et al*., 2010; Thomas *et al*., 2012; Nayfach & Pollard, 2016). Web-based servers such as MG-RAST and Integrated Microbial

Genomes and Microbiomes (IMG/M) also serve as repositories for metagenomic (and genomic) data, allowing researchers to easily access and compare microbial communities from different environments for comparative metagenomics (Wilke *et al*., 2016; Chen *et al*., 2019). While MG-RAST has a fairly shallow learning curve and allows users to quickly perform simple comparative analyses, IMG/M is much more flexible, allows for analyses at the level of individual genes, and is not limited to metagenomic datasets. IMG/M is also a product of the Department of Energy Joint Genome Institute (JGI), which prioritizes environmental sequencing projects related to understanding biogeochemistry, carbon cycling, climate change, and biofuels, among others (Chen *et al*., 2019); as such, IMG/M has a more extensive and diverse selection of environmental datasets than MG-RAST.

While metagenomic sequencing is more popular than ever, annotation of these datasets remains problematic. During annotation, short sequencing reads or predicted genes obtained through identification of open reading frames (ORFs) on assembled contigs are compared to reference databases or datasets. Reference databases are built from a collection of known data; that is, genomes obtained from cultivated microbes or empirically identified genes or proteins (Hiraoka *et al*., 2016; Nayfach & Pollard, 2016). The majority of genomes sequenced are from microbial isolates, and the majority of these isolates (80%) belong to one of three bacterial phyla-- *Proteobacteria, Firmicutes* and *Actinobacteria*; however, a phylogenetic analysis of 16S rRNA from a number of environmental metagenomic datasets suggests that more than 60 major lineages across the bacterial and archaeal domains (Hugenholtz & Kyrpides, 2009). More than half of these lineages have no cultured representatives and have often been referred to as *microbial dark matter*, or MDM (Rinke *et al.,* 2013). More than 35 of these phyla are part of a monophyletic group described as the *Candidate Phyla Radiation* (CPR), which generally have small genomes lacking many biosynthetic pathways (Brown *et al.,* 2015). While these organisms do not have any cultured representatives, some of their functions can often be inferred by identifying conserved motifs or using low thresholds matches to existing references (Baker *et al*., 2014). Nonetheless, sequences that demonstrate little to no similarity to reference sequences within the databases require more investigation to determine whether they are the result of sequencing error, or truly novel genes with unknown function.

The annotation of metagenomic datasets can be improved by adding new genomes to reference databases (Hiraoka *et al*., 2016; Nayfach & Pollard, 2016). Because the vast majority of microbes cannot be cultured, many new reference genomes are acquired through cultivation-independent techniques, primarily metagenomic binning and single cell genome sequencing, which are often used in a complementary approach (Rinke *et al.*, 2013). In metagenomic binning, assembled contigs are sorted into groups (or "bins") based on various genomic signatures, such as G+C content, tetranucleotide frequencies, codon usage, and read depth (Dick *et al*., 2009; Albertsen *et al*., 2013; Sedlar *et al*., 2017). With enough sequence coverage, these contigs can be further assembled into genomes, referred to as "metagenome-assembled genomes", or MAGs (Bowers *et al*., 2017). A number of candidate phyla genomes, some nearly complete, have been recovered using this metagenomic binning approach (Rinke *et al.*, 2013; Hedlund *et al*, 2014). Candidate phyla genome sequences have been obtained through single cell genome sequencing (Rinke *et al*., 2013; Hedlund *et al*., 2014); in this method, individual cells in a sample are sorted using fluorescence-activated cell sorting (FACS) or other microfluidics techniques into individual wells, lysed, and the genome is amplified using multiple displacement amplification (MDA) to obtain enough DNA for sequencing (Rodrigue *et al*., 2009). The resulting single amplified genome (SAG) is then sequenced, assembled, and analyzed (Rinke *et al.*, 2013; Hedlund *et al*., 2014, Marcy *et al.*, 2007). Assembling complete genomes from MAGs can be complicated by the presence of extended repeating regions, while SAGs are often fragmented and incomplete due to various technical challenges (uneven amplification or chimeric artifacts from MDA); however, when analyzed together, these two techniques can be used to resolve complete genomes (Hedlund *et al.*, 2014). Once validated, these genomes can be added to a reference dataset, and used to help analyze metagenomic datasets from other environments.

Observing a microbial community at the genome level through MAGs and SAGs can provide important context for the genes observed in a metagenomic dataset (Baker *et al*., 2014). For example, A MAG displaying a complete metabolic pathway might suggest that this uncultured organism plays a specific functional role in the microbial community, whereas an incomplete pathway that is functionally supported by geochemical analyses might indicate linkages between organisms in the community via metabolic handoffs (Baker *et al*., 2014; Anantharaman *et al*., 2016). The completeness of a genome can be estimated using single copy marker gene and can provide evidence that missing genes are not simply due to missing sequence data (Parks *et al*., 2015), but even in incomplete genomes, the presence of certain operons can provide functional insight into community interactions (Baker *et al*., 2014).

While cultivation-independent methods have exponentially expanded our knowledge of microbial diversity, they are not a replacement for cultivation-based techniques; interpretation of sequence annotations remains anchored by studies of gene function in cultivated microbes that are needed to form the foundation of reference sequence databases. Thus, as metagenomics reveals the presence of novel sequences with novel functions, connecting the genomic content to function requires demonstrating function in cultivated organisms. Nonetheless, facilitate the cultivation of novel organisms, information obtained from metagenomic data, MAGs, and SAGs can be used to improve cultivation techniques. In this way, cultivation-dependent and -independent techniques are cyclically linked, and improvements to both are required to fully profile a microbial community (Prakash *et al*., 2013).

One of the most under-explored environments for microbial interactions and processes is the oligotrophic subsurface. One portal into the subsurface are caves, which can extend over 2 km in depth allowing examination of the subsurface without the need for destructive drilling equipment or contamination (Lehman, 2007; Korbel *et al*., 2017). Nonetheless, most caves are epigenic (formed by meteoric water entering the subsurface) and are therefore hydrologically connected to surface processes (Palmer, 1990); as such, the microbiology of the cave is influenced by the route the water has taken. However, caves that formed through hypogenic conditions (ascending groundwater) are generally isolated from such meteoric water inputs and could provide an important conduit to study the microbiology of the deep subsurface including geochemistry and groundwater interactions (Palmer, 2011; Klimchouk *et al*., 2017; Kováč, 2018). The purpose of this work is to take advantage of the unique access to the Madison aquifer offered by Wind Cave. Where Wind Cave intersects the Madison aquifer, a series of lakes are formed that represent the potentiometric surface of the aquifer. The ability to travel through the cave to access this aquifer without disturbance provides a unique opportunity to investigate the microbial community within a deep and isolated karst aquifer.

CHAPTER II

A PRACTICAL GUIDE TO STUDYING THE MICROBIOLOGY OF KARST AQUIFERS

Published in *Karst Water Environment*:

Hershey, O.S., Kallmeyer, J., & Barton, H.A. (2019). A Practical Guide to Studying the Microbiology of Karst Aquifers. In: T. Younos, M. Schreiber, & K. Kosič Ficco. (Eds.). *The Handbook of Environmental Chemistry: Vol. 68, Karst Water Environment* (pp.191-207). Springer, Cham

2.1 Abstract

Examination of microbial communities within karst aquifers is an important aspect of determining the quality of the drinking water obtained from groundwater. While past work has been based on culture-based assays, a more complete view of the microbial community within karst aquifers can be achieved using molecular approaches based on DNA sequencing. Due to a reduced cell number when compared to surface environments, collecting sufficient microbial cells for analysis in karst aquifers can be problematic. In addition to issues of cell density, particulates due to the geologic location, technological limitations of equipment that

can be hand-carried and work for extended periods underground, and even the physical access to some of these subsurface sites, all contribute to making examination of the microbiology in karst aquifers a challenge. This chapter highlights some of the approaches we have used to successfully isolate microbial cells for DNA extraction from an aquifer accessed in a remote cave location. The methods we developed can aid other researchers to evaluate the microbiology of similar isolated karst aquifers.

2.2 Introduction

An important element of monitoring the quality of drinking water sourced from karst aquifers and other ground water resources is the examination of microbial communities within these environments. Although there is considerable debate how much microbes from subterranean environments contribute to the total number of microbes on Earth, karst aquifers might be an important contributor to total biomass of terrestrial subsurface microbial populations (Griebler & Lueders, 2009; Kallmeyer *et al*., 2012). The source of these subsurface microbes may be endemic or introduced from agriculture- or industry-associated activities. In pristine (or uncontaminated) aquifers, the microbial community present is likely a result of initial seeding of microorganisms from the surface environment (soils), and then adaptation of those microorganisms to the physical and chemical conditions found within the aquifer itself over potentially hundreds-of-thousands to millions of years

(Gray & Engel, 2013; Hug *et al*., 2015). In contaminated aquifers, the microbial community may be disturbed from its natural state by the addition of waste water from human and/or animal sources and activities, increasing the number of coliform bacteria or fecal indicator species, such as *Escherichia coli* (Cho & Kim, 2000; Pronk *et al*., 2005; Ohrel Jr & Register, 2006). Spills or wastewater containing chemical contaminants may also dramatically change the structure of the microbial community present (Dojka *et al*., 1998; Abed *et al*., 2002; North *et al*., 2004).

The detection of pathogenic bacteria and viruses that can be harmful to human and animal health has been described in detail and can be successfully achieved with cultivation-dependent (requiring growth in nutrient agar) methods (Ashbolt *et al*., 2001; Ohrel Jr & Register, 2006). As traditional techniques in microbiology were historically developed to study human pathogens, many of the microorganisms that affect human health are identifiable using traditional laboratory techniques, enabling standardized methods for their use in water quality analysis (Abed *et al*., 2002); however, the assessment of karst aquifers for these human-associated pathogens does not provide a full understanding of the role of microorganisms within these environments. More than 99% of microbes cannot be cultivated using standard laboratory conditions (Amann *et al*., 1995), yet this uncultured majority carry out the majority of functions within the global biosphere (Gray & Head, 2001). Culture-based methods alone therefore fail to recognize the function of the majority of microorganisms that exist in the environment.

In aquifers that serve as an important source of drinking water, evaluating the natural microbial ecosystem enables understanding of microbial processes that help maintain water quality homeostasis (the maintenance of water chemistry through biotic processes; Iker *et al*., 2010). Indeed, in some cases, an evaluation of these processes is included as a component of drinking water quality standards. For example, the Swiss Water Protection Ordinance considers ecological goals to be an important part of groundwater protection, stating that *"The biotic community of underground waters shall: a) be close to nature and appropriate to the location; b) be specific to unpolluted or only slightly polluted waters*" ("Water Protection Ordinance," 1998; Goldscheider *et al*., 2006). These biocenoses account for both micro- and macro-organisms and their interactions, which have the potential to impact biogeochemical cycles maintaining water quality in karst aquifers (Iker *et al*., 2010). Despite this, the practical approaches needed to understand the microbial communities in karst aquifers can be technically challenging.

The low level of organic input into these environments results in a low biomass (low absolute cell numbers) within the environment. As most published water sampling methods for microbiology are typically used in surface environments with a high cell density, these approaches are unlikely to work in cave environments and karst aquifers. Evaluation of cell biomass is an important factor in determining which sampling method to use, and the sampling methods employed may be altered to obtain sufficient biomass for analysis. In extremely low biomass aquifer systems, commonly found in karst aquifers and exemplified by our study area in Wind Cave, Wind Cave National Park, USA (Figure 2.1), even the traditional DNA techniques used in environmental microbiology must also be altered from conventional methods. The main objective of this chapter is therefore to address the issues faced when sampling water from karst aquifers, particularly in clean, pristine groundwater sources and to provide potential solutions to researchers interested in this emerging field.

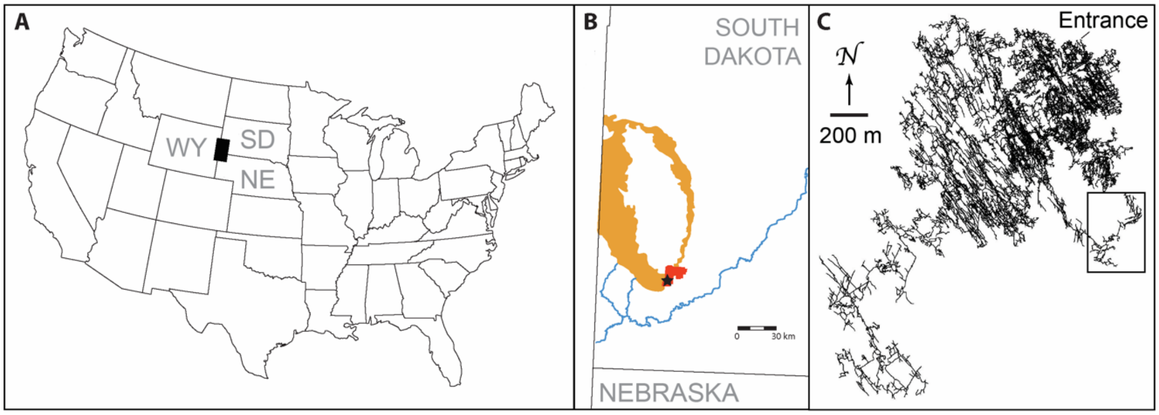

Figure 2.1. Location map of Wind Cave, South Dakota. A) Map of the United States, indicating the location of the Black Hills (solid black box and area shown in B). Black lines correspond to State boundaries, with the location of Wyoming, Nebraska and South Dakota indicated. B) The area where the Madison Limestone outcrops within the Black Hills is indicated (brown) which forms the majority of the aquifer recharge zone. The location of Wind Cave National Park (red), and the Wind Cave entrance(black star). The State boundaries are shown, along with the major rivers in the area. C) The survey line plot of the passages within Wind Cave. The area where the lakes are found is indicated by the black square. Arrow indicates true north (data compiled by, and with, permission of Wind Cave National Park).

2.3 Study Area

We have been studying the confined karst aquifer of the Madison Formation in South Dakota, where the aquifer is intersected by Wind Cave (Wind Cave National Park, South Dakota, USA; Figure 2.1). Due to the complex speleogenesis of Wind Cave, the deeper passages in the cave likely pre-date the current potentiometric surface of the Madison aquifer, although the speleogenesis of the whole cave is historically tied to this aquifer (Palmer & Palmer, 2000). Where the cave intersects the Madison aquifer, a series of lakes are created by the current potentiometric surface of the aquifer. The water found in the Wind Cave lakes consists primarily of local recharge water (96%) that has a residence time of ~25 years (Long & Valder, 2011); however, due to the complex structure of the aquifer, the lakes also contain water up to ~50,000 years old (Long & Valder, 2011). Accessing the lakes requires travelling for several hours from the cave entrance, through approximately 3 km of passages with an overall descent of over 200 m, with much of the cave requiring climbing and crawling, occasionally through passages less than 20 cm in width.

2.4 Issues with Standard Methods of Sampling Karst

In order to study the microbiology of karst aquifers, access is most often gained through wells and springs. Nonetheless, without careful sampling strategies, the microbial population examined through these routes of access can be influenced by the microbiology of the route itself (Lehman, 2007). In the case of wells, the casing can potentially provide sources of energy or electron acceptors for growth that would not normally be present (through concrete or iron) and can skew our understanding of microbial community energetics, while likewise in non-cased wells the geochemistry of the rock units or water leaking from shallower aquifers through which the well penetrates, can similarly affect community structure (Lehman, 2007). While springs may allow more direct access to karst aquifers, their exposure at the surface means that photosynthetic primary production can dramatically alter the native microbial population of the karst aquifer itself, including, for example, the presence of larger plant and animal species that in turn contain their own native microbial populations (Elshahed *et al*., 2003; Pinowska *et al*., 2007). Due to these limitations, studying karst aquifers through natural conduits (caves) can increase the accuracy of the microbiological interpretations. Unfortunately, the cave environment can have its own set of challenges, requiring specialized equipment and physical endurance: equipment must be durable enough to withstand bumping, scraping, squeezing, dropping, etc., as well as exposure to sediments and debris; it must be light enough to be carried long distances and small enough to traverse tight passages; finally, the time needed to transport samples back to the surface (and laboratory) must be considered.

2.5 Methods Developed for Sampling Karst Aquifers

Cell Enumeration

One of the most critical pieces of information needed in order to study the microbiology of any aqueous environment is to understand the size of the microbial population, which is done by direct counting the number of cells present. While this approach is laborious and requires specialized fluorescent microscopy, direct cell counting makes it possible to determine the volume of water that needs to be filtered for the DNA approaches used, which in turn, determines the sampling procedure.

In order to carry out cell counting, we collect three 10 mL samples on site using sterile syringes (Figure 2.2). These samples are then preserved on site in 2% paraformaldehyde by adding 1 mL of a 20% paraformaldehyde concentrate in phosphate buffered saline (pH 7.4). Such preservation is critical to ensure that cell numbers are not artificially lost through potential predation or inflated through growth. The samples we have been working with do not contain any grazing arthropod species; however, if such species are present in the collected samples, they can either be enumerated along with the microbial species (albeit at a much lower magnification) or an 8 µm pre-filter can be used to remove them for enumeration of non-eukaryotic species in the water column.

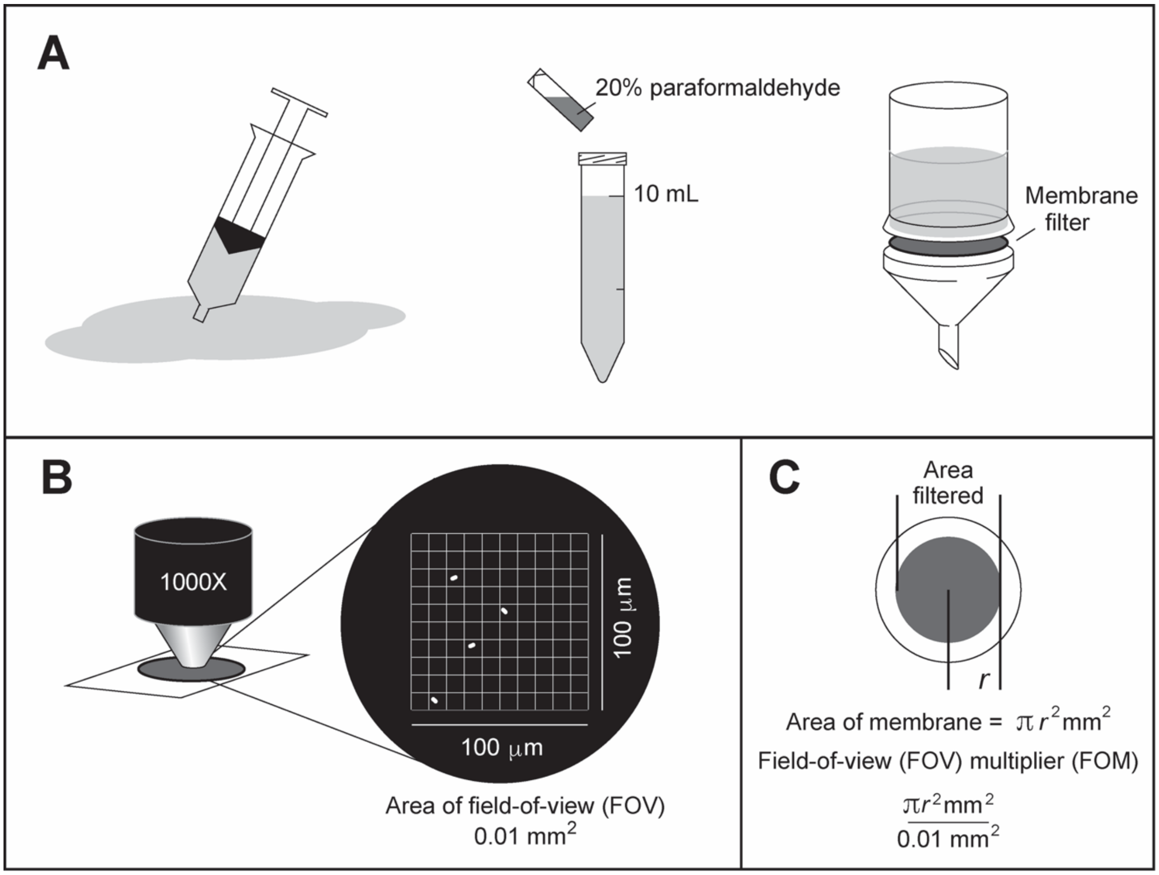
A
20% paraformaldehyde
10 mL
Membrane filter
B
1000X
100 μm
100 μm
Area of field-of-view (FOV)
0.01 mm^2
C
Area filtered
r
Area of membrane = πr^2mm^2
Field-of-view (FOV) multiplier (FOM)
$\frac{\pi r^2 \text{mm}^2}{0.01 \text{ mm}^2}$

Figure 2.2. Sampling and counting strategy for cell enumeration in karst aquifers. A) Three 10 mL samples are collected and preserved on site. The sample is stained with a DNA-binding fluorescent dye and filtered on to a 0.02 – 0.2 μm non-fluorescent filter. B) The sample is visualized on a fluorescence microscope (at 1000x total magnification) and the number of cells within a known area (field of view; FOV) are counted. C) The average number of cells per FOV (between 50 – 200 FOVs depending on the cell density) and area of the FOV are then calculated and multiplied by the total surface area of the membrane to obtain the total cell number (in 10 mL).

For cell enumeration, the entire sample of preserved water is stained with the SYBR Green I DNA-binding fluorescent dye (Sigma-Aldrich, St. Louis, MO) and filtered onto 0.02 – 0.10 μm pore membrane filters (such as Whatman Anodisc #9809-6002 or Whatman Cyclopore #7063-2502) depending on the target population (Figure 2.2); most microbial communities can be collected for counting with a 0.1 µm membrane, although a trained microscopist will be able to identify trapped bacteriophage on a 0.02 µm membrane. Total cell counts are then carried out using epifluorescence microscopy; between 50 and 200 fields of view at 1,000X total magnification are inspected to determine how many cells are observed on average in a field of view (FOV; Figure 2.2). The average number of cells is the determined by dividing the total cell number by the total FOV to determine the average number of cells in the (10 mL) sample, which is in turn used to calculate the total number of cells mL^{-1} of water (Figure 2.2).

Aggregation of cells to particulates or the formation of biofilms is potentially problematic in all water samples, especially from geologic environments, and may either reduce the number of visible cells (remove them from a FOV), or artificially increase the number of cells counted in any individual FOV, subsequently inflating the total count. In order to disrupt such aggregates and increase the accuracy of counting, the samples can be treated using the methods developed by Kallmeyer *et al*. (2008) without the need for NaCl that is normally used to adjust the salinity of samples from ocean environments. In past sample analysis, we have found that four different treatments have been effective in our samples, which can be modified based on the observed aggregation: (1) filtering the untreated sample straight onto the membrane; (2) sonicating the sample for 10 min (at 640 W) before filtering; (3) treating with a detergent solution (100 mM disodium EDTA dihydrate, 100 mM sodium pyrophosphate decahydrate, 1% vol/vol Tween 80) and methanol before vortexing for 30 minutes (to dissolve extracellular polymers) and filtering onto the membrane; or (4), a combination of methods 2 and 3. In the case of low biomass samples, blank counts using only reagents and membrane (no sample) should be included before and after each period of counting, with the average blank values subtracted from each cell count (Kallmeyer *et al*., 2008).

DNA Isolation for Identification

Modern environmental microbiology requires the use of molecular DNA techniques to identify microbial communities (Pace, 1997). There are two current approaches: the first relies on the 16S small subunit ribosomal RNA gene (16S rRNA), and the second requires sequencing the entirety of the DNA found in using a computer program (EMIRGE) to identify 16S rRNA sequences in this metadata (Miller *et al*., 2011; Mirete *et al*., 2016; Garrido-Cardenas & Manzano-Agugliaro, 2017). The first has the advantage of requiring only tiny amounts of DNA to generate community profiles through the polymerase chain reaction (PCR), although this approach can suffer from the bias toward certain amplified sequences by the primers used (Suzuki & Giovannoni, 1996; Pinto & Raskin, 2012). The latter does not suffer from such biases and can provide additional functional information, but requires significantly more DNA for the generation of the libraries necessary for sequencing (Head *et al*., 2014; Bowers *et al*., 2015).

Without sophisticated technology that can extract tiny levels of DNA with high efficiency from complex geochemical samples (such as the Boreal Genomics Aurora DNA extraction system, Boreal Genomics, Vancouver BC; So *et al*., 2010), disruption and kit-based DNA extraction protocols can be used. Such kit-based approaches are notoriously inefficient, with even the best protocols only able to isolate ~17% of total DNA from a sample (Claassen *et al*., 2013). Therefore, for most laboratories the ability to carry out PCR approaches to identify the microbial communities requires a minimum of 300 ng of purified DNA (conservatively this requires ~2 µg of total DNA within the bacteria collected for DNA extraction). Given that an average bacterium contains ~3 femtograms of DNA (assuming the 4 x 10^6 bases of the *Escherichia coli* genome as a standard and a molecular weight of 660 for each base pair), this means that a minimum of 6.6 x 10^8 bacterial cells must be isolated for analyses. Generally, this would not be difficult in most aquatic samples (Table 2.1), but isolated karst aquifers are anticipated to have microbial cell numbers equivalent to other oligotrophic freshwater sources, at ~2 x 10^6 cells mL^{-1} (compared to the open oceans of 1.5 x 10^7 cells mL^{-1}; Farnleitner *et al*., 2005; Newton & McMahon, 2011). In our analyses of the isolated Wind Cave Lakes, such high numbers are rarely seen. Thus, while samples from the open ocean may contain sufficient cell numbers for DNA extraction in as little as 7 mL of sample, the lakes in Wind Cave required ~300,000 mL of water to obtain the same cell density (Table 2.1).

Table 2.1. The average number of bacterial cells identified in different bodies of water and the comparable volume of water needed to isolate sufficient microbial cells for sufficient DNA extraction. Table A.1. is an extended version of this table

Aquatic Environment	Typical bacterial abundance (cells/mL)	Amount filtered for 2 μg DNA (L of water)	Reference
Open Ocean	1.0 x 10^4 – 1.0 x 10^7	0.007	Whitman *et al*. 1998
Eutrophic River Warnow	2.4 x 10^6	0.03	Freese *et al*. 2006
North Atlantic	8.2 x 10^5 – 2.4 x 10^6	0.3	Rowe *et al*. 2012
West Pacific	2.9 x 10^5 -1.2 x 10^6	0.5	Rowe *et al*. 2012
Crater Lake	2.0 x 10^5 – 1 x 10^6	0.7	Urbach *et al*. 2001
Sargasso Sea	4.6 - 8.8 x 10^5	0.8	Rowe *et al*. 2012
Swiss Cave Pool	5.2 x 10^5	1.3	Shabarova and Pernthaler, 2010
Limestone Karst Aquifer Spring	6.8 x 10^4	10	Farnleitner *et al*. 2005
Dolomite Karst Aquifer Spring	1.5 x 10^4	45	Farnleitner *et al*. 2005
Wind Cave Lakes	2.3 x 10^3	287	Hershey *et al*. unpublished
Lake Vostok	2.0 x 10^2 -1.0 x 10^3	660	Karl *et al*. 1999

2.6 Sample Collection

Samples Volumes Below 1 L

One of the most practical methods to collect DNA in high biomass waters (>10^6 cells mL^{-1}; Table 2.1) is using the 0.22 μm Sterivex filtration (Millipore #SVGV010RS) technique, which has the advantage of being inexpensive, is able to be used with most pumps (including syringe and hand-driven pumps), and has been used extensively in ocean sampling (Schauer *et al*., 2000; Francis *et al*., 2005). Nonetheless, water samples collected in karst environments, such as caves, often have significant clogging issues due to particulates, such as clay, silt, iron oxides and calcite rafts in the water. In order to overcome these limitations, we have used sterile cheesecloth (sterilized using an autoclave), which can be placed over the end of the water draw line. Such cheesecloth will not hinder the access of microbial cells, but is particularly effective at limiting the entry of particles that can quickly clog the Sterivex filter. To remove the bacterial cells for extraction, a syringe can be used to release the cells from the Sterivex filter using a reverse flow. This method is more appropriate for surface karst environments such as springs and wells, as the proximity to surface energy sources yields higher biomass, requiring less water to be sampled.

Sample Volumes Below 20 L

In environments that require larger volumes of water, we have used 0.2 μm aluminum oxide Anodisc filters (Whatman #6809-6022). These filters have an advantage over other membrane approaches in that the aluminum oxide portion of the filter can be directly crushed into DNA extraction buffers, with a chemistry that does not interfere with most DNA extraction protocols; however, the aluminum oxide composition of the filter means that the filters are extremely brittle and great care should be used, especially in mounting. The Anodisc filters can be set up for filtration using a Millipore stainless steel filter holder (Millipore #XX3002500; Figure 2.3) and can once again be used with most water pumps. An additional benefit of this approach is that the Millipore filter holder can also be used with traditional cellulose filters, allowing microbial cells to be filtered onto the membrane, which

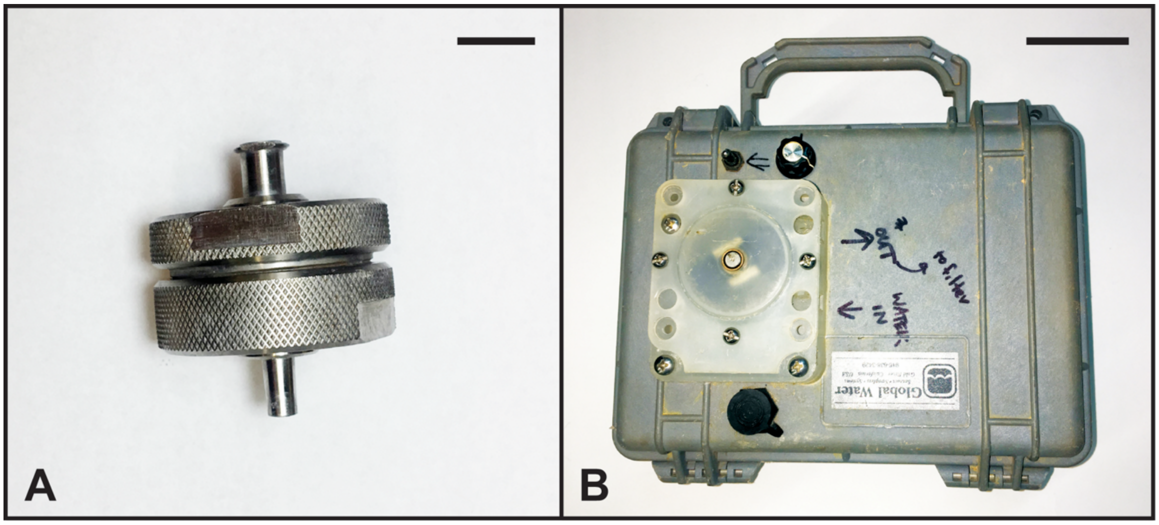

Figure 2.3. Examples of equipment used for sample collection and enumeration. A) A steel Millipore 25 mm filter holder (scale bar is 1 cm). The entire filter holder can be autoclaved and contains a Luer-locking mechanism for syringe or tubing attachments. The holder can accommodate a number of different filter membranes, which can be mounted so that water is either pushed or pulled through the device, allowing it to work with a number of different pumps. B) The Global Waters SP200 peristaltic pump, which is 23 cm by 19 cm and weighs 2 kg (scale is 5 cm). The peristaltic mechanism is exposed through the lid of its case, which is unscrewed for the peristaltic tubing to be attached. The pump is marked-up to ensure the correct attachments are made in the cave, while the power connection port is visible below the peristaltic mechanism.

can be transferred directly to selective growth media for cultivation. In this way, even in extremely low biomass samples with poor cultivation rates (<0.1%), larger concentrations of cells can be captured for growth and other functional studies. Nonetheless, due to connections using Luer locks with the filter holder, only narrow tubing can be used. Due to the increased velocity of water movement (due to the Bernoulli Principle), any clogging of the membrane by particulates can be particularly problematic, leading to increased pressure that can rupture membranes or cause tubing/connections to fail. Therefore, while the Anodisc method is very effective, it should not be used for water volumes >20 L or for samples with high particulates.

Sample Volumes Below 50 L

Samples in the range of 10^4-10^5 cells mL^{-1} require significantly larger volumes of water to be collected. This requires a high-capacity pump along with a filter with a large enough surface area to handle large volumes of water. Finding a pump capable of working with the weight and power constraints needed for transport into the cave that was capable of moving quite large volumes of water (in excess of 1 L min^{-1}) was a challenge. We found that the SP200 variable speed peristaltic pump (Global Water Instrumentation #CJ0000; Figure 2.3), which can pump up to 2 L min^{-1}, was the best commercially available product for our samples. Designed for environmental field work and built into a rugged Pelican case, in our experience the SP200 pump works well in even challenging caves (although the use of a desiccant to keep the electronics dry within the Pelican case during cave use is advisable). The pump was powered by a high capacity 12V rechargeable lithium-ion battery pack (BiXPower BP160), adapted for compatibility with the pump battery connector.

To filter volumes below 50 L we have used a Nalgene disposable 0.22 µm filter device, which has the advantage of being sterile (Gamma-irradiated) and can remain open in the water to avoid backpressure problems, even with high levels of particulates in the water column (Figure 2.4). Nonetheless, the use of a sterile gauze pre-filter to limit large particulates is useful (Figure 2.4). This filter can be suspended just below the surface in the water column and the water is pulled from the vacuum port. This approach has been very successful for filtering large volumes of water in caves; however, it is critical that the device is pulled from the water before the pump loses power, which would result in the loss of suction and allow cells trapped on the membrane surface to diffuse away due to Brownian motion. For DNA extraction, the membrane can be cut out of the filter unit and, due to its cellulose chemistry, can be extracted directly in a phenol/chloroform mix (the cellulose dissolves in phenol), leaving behind the DNA/cell mix for extraction.

Sample Volumes Above 50 L

In oligotrophic bodies of water with fewer than 10^4 cells mL^{-1}, potentially hundreds of liters of water may need to be collected for enough bacterial cell mass to allow DNA extraction. This can be achieved using a Nalgene filter for an extended period of time to allow enough water to pass through the filter (>24 h); however, a growing body of research has identified ultra-small bacteria in similarly oligotrophic groundwater environments (Miyoshi *et al*., 2005; Luef *et al*., 2015). Antarctic lake brine (Miteva & Brenchley, 2005), and marine environments (Ghai *et al*., 2013). These studies suggest that there is a significant bacterial population (>15%) able to pass through a 0.2 μm filter, which was previously considered the lower size limit for life (Maniloff, 1997). As small size yields a higher surface area to volume ratio in bacteria and increases the uptake of scarce nutrients, it is logical to assume that environments low in nutrients, such as pristine karst aquifers, there may be a higher relative contribution from ultra-small species to the microbial community.

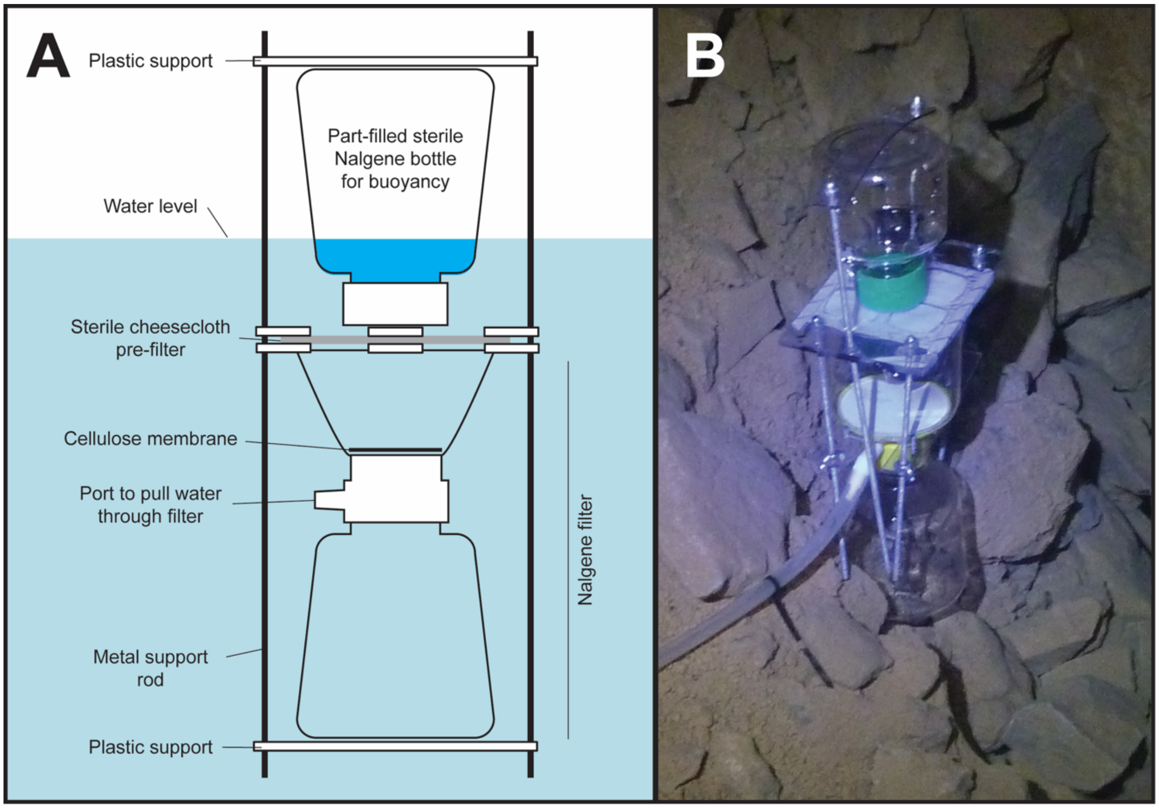
A
Plastic support
Part-filled sterile Nalgene bottle for buoyancy
Water level
Sterile cheesecloth pre-filter
Cellulose membrane
Port to pull water through filter
Nalgene filter
Metal support rod
Plastic support
B

Figure 2.4. The Nalgene filter set-up we have used in caves, which has the advantage of being inexpensive, commercially available, sterile (so they can be placed in the water column), and not susceptible to back-pressure problems. A) Schematic of our set-up, showing the vertical support rods and the plastic supports to hold the device together. The partially-filled Nalgene bottle allows the position of the device in the water column to be determined (and can be used to collect a water sample for chemical analysis), while the cheesecloth pre-filter traps clays and other large particulates. All of these materials are autoclaved and carried into the cave sterile for assembly. B) The sampling device being used to sample in a karst aquifer accessed via a cave conduit (Photo courtesy of M. Carnol).

To collect these smaller cells, it is possible to use a Nalgene filter membrane with a smaller pore size (such as 0.1μm); however, the relatively small surface area of the membrane combined with the reduced pore size generates significant resistance to water flow, which drastically increases the time needed to sample. When a Nalgene filter was reduced to 0.1 µm pore size, even using the maximum speed of the SP200 pump, the flow rate through the membrane was reduced to 0.1 L min^{-1}. Assuming a constant rate of flow, it would therefore take approximately 50 hours to sample 300 L of water. The increased sampling times may be prohibitive for a variety of reasons, especially in a remote cave location that make it impractical to return to the surface multiple times: in the absence of a source of electricity, the pump must be operated by multiple batteries for extended runs; the batteries must be lightweight and small enough for transport through the cave, limiting the Amp hours and duration of available power; batteries may not perform optimally in the cooler conditions and high humidity of the cave; and over extended duration collection trips, researchers need to bring sufficient clothing (for warmth) and food. When sampling exceeds 24 hours, researchers may need to camp in the cave, requiring camping supplies in the addition to research equipment, necessitating large, cumbersome, and heavy packs needed to carry this additional equipment. Therefore, all efforts should be taken to reduce the likelihood of long sampling events.

To resolve these many issues, a different filtration method is required. All the described methods so far use "dead-end" filtration of water, with flow perpendicular to the membrane. Due to this method of flow, caking of materials on the membrane will also reduce flow and eventually cause clogging (Li & Li, 2015). Alternatively, tangential flow filtration (TFF) allows water to flow tangentially across the membrane, allowing water and particles smaller than the membrane pores to permeate and exit the filter, while retaining larger particles (Petrusevski *et al*., 1995). This process also allows pores to be very small (on the nanometer scale) without reducing flow rate and limiting the potential for caking and clogging.

A compact and durable TFF device that we have successfully used is the Large Volume Concentration column (LVC kit; Innovaprep, Drexel, MO; Figure 2.5). This filter utilizes tangential flow across 2.5 m^2 of a 45 nm pore polysulfone hollow fiber membrane. Despite the exceedingly small pore size, the high surface area of the membrane allows filtration through the device to continue at 1 L min^{-1}. This is a major improvement from the previous sampling strategies, with the LVC kit filtering 300L in 5 hours and can be run longer for larger samples. While tangential flow demonstrates a significant advantage over other membrane technologies, the columns themselves are expensive and samples must be released from the filtration column with a provided buffer. In our assays, removing all of the cells and particulates trapped by the membrane has taken over 400 mL of a 0.075% Tween 20/PBS elution fluid; the total volume of the collected sample must then be centrifuged to concentrate the cells for further processing. Despite this extra step, an additional benefit to this collection method is that the 45 nm pore size also enables a large portion of the bacteriophage population (size range 25 - 200 nm) to be collected.

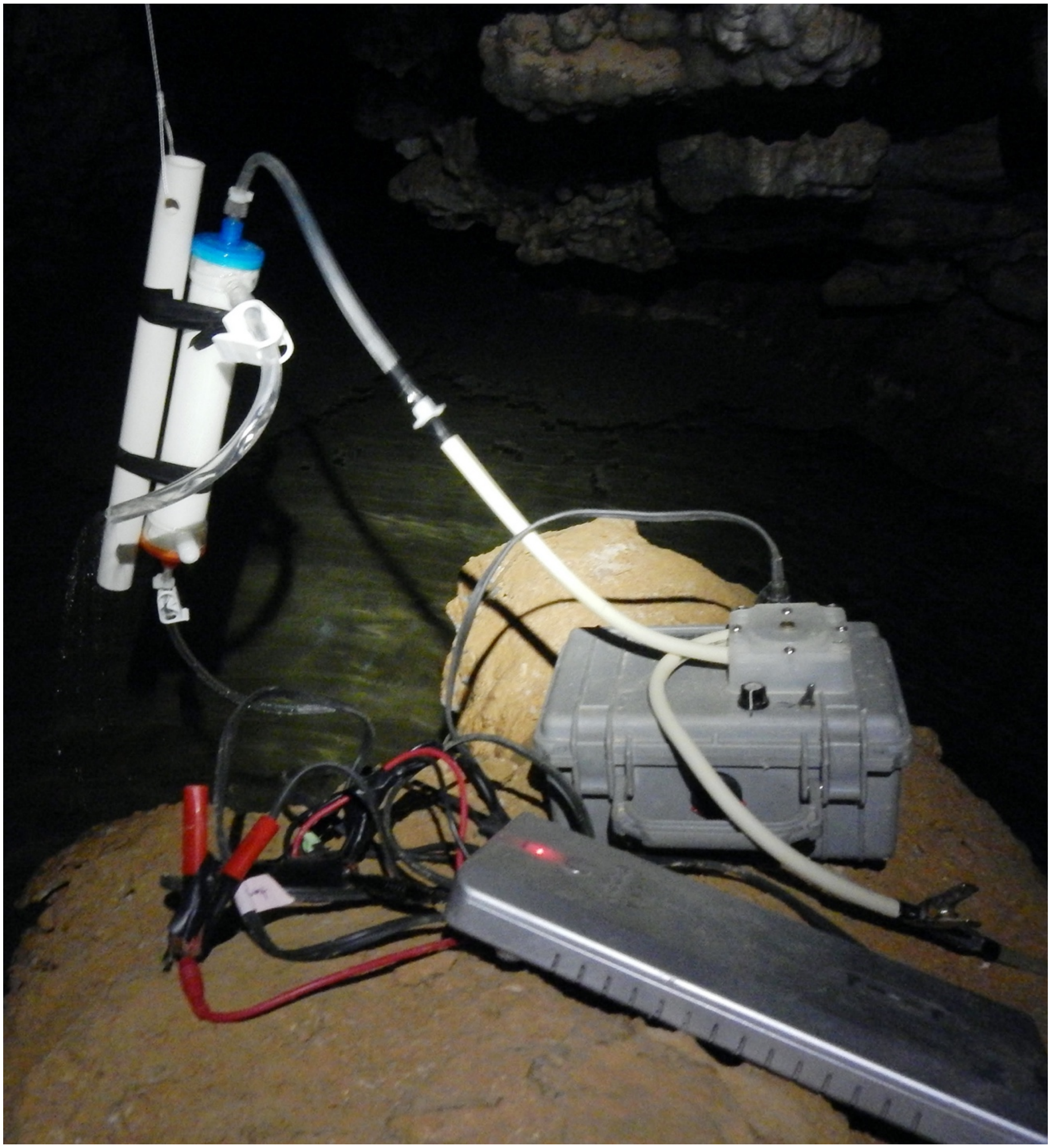

Figure 2.5. The large volume concentration (LVC) column sampling the lakes in Wind Cave, along with the Global Waters SP200 pump and a high capacity 12V/19V rechargeable lithium-ion battery pack (BiXPower BP160). The column does not need to be immersed within the water; however, a mechanism needs to be set up that allows the column to remain vertical during sampling. In the cave we use a piece of rigid plastic tubing to support the column.

2.7 Limitations of Microbial Sampling in Karst

The use of the described techniques allows for the isolation of sufficient DNA for the identification of microorganisms in karst aquifers, either through 16S rRNA sequence analysis or through metagenomics approaches; however, recent work has suggested that even these approaches have their limitations. Our preliminary data suggest that microbial communities from karst aquifers may contain significant populations of the recently described '*microbial dark matter*' (Marcy *et al*., 2007; Rinke *et al*., 2013; Head *et al*., 2014). This 'dark matter' includes microbial species (particularly within the bacteria) that cannot be identified using current databases. Many of these may include representative operational taxonomic units (OTUs) of the recently identified Candidate Phyla Radiation (CPR; Brown *et al*., 2015). The CPR represents a number of phyla in the tree of life that have no known cultured representatives and have only been observed through DNA collected from environmental surveys. Microbes with these unusual sequences do not match to reference databases and may be wrongfully discarded from environmental microbial surveys as sequencing artifacts, when they actually represent novel divergent microbial lineages (Hug *et al*., 2015).

A survey of a groundwater aquifer in Rifle, Colorado estimated that the CPR may represent more than 15% of the species present and have been under-sampled in most culture-independent surveys due to their divergent 16S rRNA sequences and collection methodologies (Brown *et al*., 2015). The Rifle survey also predicts that CPR representatives are likely to be ultra-small based on the small size of their putative genomes, providing a potential explanation for the lack of available information on these microbes. This phenomenon can be observed in the different contributions of these unclassified OTUs in samples collected using different pore-sized membranes (0.2 μm and 0.1 μm) from the Wind Cave lakes (Figure 2.6). The difference in the collected communities appears to be based primarily on the size of the filter, and the majority of this difference consists of OTUs that remain unassigned to previously described taxa. This suggests that a significant portion of the community is indeed excluded when relying on a 0.2 μm filter for microbial collection. This excluded portion of the community, the presumptive "microbial dark matter", despite being composed primarily of ultra-small cells, must not be discounted when determining the overall metabolic processes that are ongoing in karst aquifers. These contributions to the community metabolism may include steps or enzymes in various nutrient cycles, production of secondary metabolites that may be used by other members of the community, or it may be possible that small cell size is indicative of a reduced genome size, which would indicate metabolic dependence by distinct phyla on the rest of the microbial community, rather than a contribution to ecosystem processes (Anantharaman *et al*., 2016).

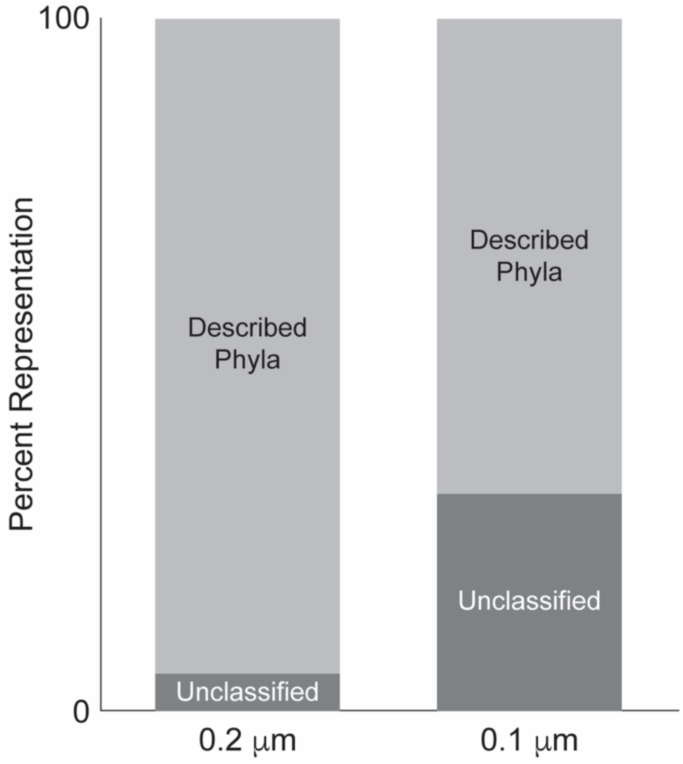

Figure 2.6. The relative contribution of microorganisms in the Wind Cave Lakes that cannot be identified using traditional taxonomic databases based on the pore size of the filters used to collect the sample. The described phyla are the percentage of Illumina DNA sequences that can be placed in known phyla using the most recent SILVA reference database, while the unclassified sequences are those that cannot be assigned to any known phyla.

2.8 Conclusions

The information obtained from understanding the contributions of the known bacterial and archaeal communities, the potential contributions by currently unclassified community, combined with metagenomic data, can reveal the presence of biogeochemical pathways driven by microbial activities in groundwater, and generalizations can be made about what process maintain the quality of drinking water (Iker *et al*., 2010). Likewise, by understanding the metabolic interactions between microorganisms in these environments can reveal important information about the ability of a microbial community to function in its environment, or their contribution to cave forming processes (Tringe *et al*., 2005; Engel & Randall, 2011; Gray & Engel, 2013). By improving the collection methods used to obtain microbial communities from karst aquifer systems, we can better understand the processes that affect these systems, the implications for water quality in deviations from pristine aquifer conditions, and the importance of the microbial community to groundwater health and quality (Ward *et al*., 2005). The aim of this chapter was to share the methods we have developed (over many years) with other karst researchers, with an aim of promoting microbiology in what remains a relatively under-explored microbial environment. Given the preliminary work of ourselves and other investigators, it is likely that this will soon be an exciting and emergent field (Hug *et al*., 2015; Luef *et al*., 2015).

CHAPTER III

HIGH MICROBIAL DIVERSITY DESPITE EXTREMELY LOW BIOMASS IN A DEEP KARST AQUIFER

Published in *Frontiers in Microbiology*:

Hershey, O. S., Kallmeyer, J., Wallace, A., Barton, M. D., & Barton, H. A. (2018). High Microbial Diversity Despite Extremely Low Biomass in a Deep Karst Aquifer, *Frontiers in Microbiology*, *9*, 2823.

3.1 Abstract

Despite the importance of karst aquifers as a source of drinking water, little is known about the role of microorganisms in maintaining the quality of this water. One of the limitations in exploring the microbiology of these environments is access, which is usually limited to wells and surface springs. In this study, we compared the microbiology of the Madison karst aquifer sampled via the potentiometric lakes of Wind Cave with surface sampling wells and a spring. Our data indicated that only the Streeter Well (STR), which is drilled into the same hydrogeologic domain as the Wind Cave Lakes (WCL), allowed access to water with the same low biomass (1.56 - 9.25 x 10^3 cells mL^{-1}). Filtration of ~300 L of water from both of these sites through a 0.2 µm filter allowed the collection of sufficient cells for DNA extraction, PCR amplification of 16S rRNA gene sequences, and identification through pyrosequencing. The results indicated that bacteria (with limited archaea and no detectable eukaryotic organisms) dominated both water samples; however, there were significant taxonomic differences in the bacterial populations of the samples. The STR sample was dominated by a single phylotype within the *Gammaproteobacteria* (Order *Acidithiobacillales*)*,* which dramatically reduced the overall diversity and species richness of the population. In WCL, despite less organic carbon, the bacterial population was significantly more diverse, including significant contributions from the *Gammaproteobacteria, Firmicutes, Chloroflexi, Actinobacteria, Plantomycetes, Fusobacter* and *Omnitrophica* phyla. Comparisons with similar oligotrophic environments suggest that karst aquifers have a greater species richness than comparable surface environs. These data also demonstrate that Wind Cave provides a unique opportunity to sample a deep, subterranean aquifer directly, and that the microbiology of such aquifers may be more complex than previously anticipated.

3.2 Introduction

Karst is a term used to denote landscapes formed by water within soluble rock, often limestone ($CaCO_3$). Aquifers that flow through karst landscapes often do so through solutionally enlarged conduits (caves), which provide multiple, direct connections between meteoric water and groundwater systems (White, 1988; Ford & Williams, 2013). They also provide an important source of drinking water, with 25% of the world's fresh water flowing through karst aquifers (Ford & Williams, 2013). Recent work has demonstrated that subsurface aquifers play an important, but overlooked, role in the global hydrologic cycle, contributing as much as four times the freshwater discharge into oceans as rivers and streams (Taniguchi *et al*., 2002; Kwon *et al*., 2014). Areas with significant karst landscapes, such as the Mediterranean Sea, can receive up to 75% of their freshwater input from karst springs rather than surface run-off (Garcia-Solsona *et al*., 2010). Aquifers also contribute to terrestrial nutrient input into the oceans (Garcia-Solsona *et al*., 2010; Weinstein *et al*., 2011; Kwon *et al*., 2014; McCormack *et al*., 2014). The importance of karst aquifers in geochemical cycles and as human water sources has led to an increase in research aimed at investigating indigenous microbial species in such groundwater (Farnleitner *et al*., 2005; Goldscheider *et al*., 2006; Griebler & Lueders, 2009; Pronk *et al*., 2009; Wilhartitz *et al*., 2009; Iker *et al*., 2010; Gray & Engel, 2013); however, the microbial ecology of these environments remains poorly understood, as the vast majority of karst groundwater research is concerned with point source contamination rather than native microbial community structure (Dojka *et al*., 1998; Boyer & Pasquarell, 1999; Cho & Kim, 2000; Ashbolt *et al*., 2001; North *et al*., 2004; John & Rose, 2005; Sinreich *et al*., 2014).

Wind Cave is found within the Mississippian age limestone of the Madison Formation along the eastern flank of the Black Hills South Dakota, USA (Figure

3.1). Originally designated a US National Park in 1903, Wind Cave is one of the longest (at 218 km) and oldest caves in the world, with initial cave forming processes (speleogenesis) occurring during the mid-Carboniferous period [~350 million years ago (Mya; Palmer and Palmer, 2000)]. The most recent speleogenetic processes began during the Eocene Epoch (~ 50 Mya), when the Black Hills tilted toward the southeast, causing groundwater flow to accelerate passage enlargement and begin a close association between the aquifer and the cave (Bakalowicz *et al*., 1987). At a depth of -122 m, the cave intercepts the Madison aquifer, providing the only direct physical access (other than through drilled wells or discharge springs) to an enormous aquifer that underlies four US states and two Canadian provinces (Long *et al*., 2012).

At the point where Wind Cave intersects the aquifer a series of lakes are created (Figures 3.1C-D). The lack of an obvious discharge route out of the lakes and their relationship to the local potentiometric surface suggests that the surface of the lakes represents the local surface of the Madison aquifer (Back, 2011; Long *et al*., 2012). Measurements of stable isotopes in calcite precipitates near the lakes site suggest that they have existed in this region of Wind Cave for ~1.14 (±0.13) Myr, where they have remained isolated from diurnal or seasonal variation and under permanent aphotic conditions (Ford *et al*., 1993). Their geologic isolation also means that they remain separated from the surficial hydrologic cycle, with recharge water taking an estimated 25 years to reach the lakes (Back, 2011; Long & Valder, 2011). The lakes themselves sit in a region of the Madison aquifer

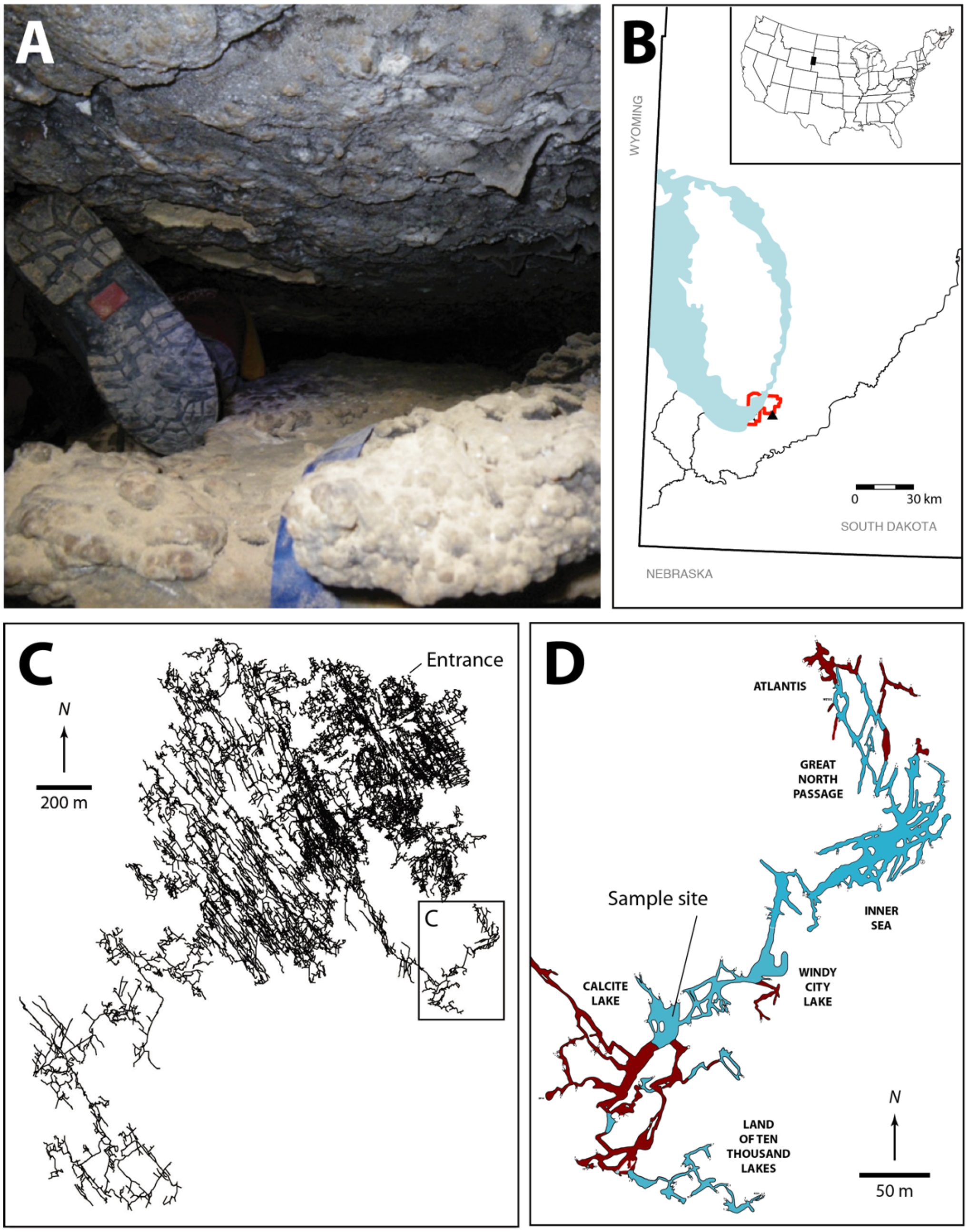

Figure 3.1. Images illustrating the route to the Wind Cave lakes. A) Image illustrating some of the passages that must be traversed *en route* to the Wind Cave Lakes (it should be noted that this is not the smallest passage that researchers must navigate with equipment). B) Location map of South Dakota, the Black Hills and Wind Cave. The exposed Madison limestone, where some of the Madison aquifer water recharge occurs is indicated in blue, the location of Wind Cave National Park (red), Wind Cave (black star), and Streeter well (black triangle).

C) The survey line plot of the passages within Wind Cave, demonstrating the location of the lakes in relation to the natural entrance to the cave. D) Location of the sample site within the lakes area. The location of the lakes is indicated in blue, along with dry cave passages (brown). The named areas of the cave are indicated. All arrows indicate true north. Cave data compiled by, and with permission of, Wind Cave National Park; adapted from Hershey and Barton, 2018.

containing groundwater flow paths in a complex aquifer pattern, with 39% recharge from the Precambrian rocks of the Black Hills, while 33% and 25% comes from ancient recharge basins (>10,000 years) flanking the eastern and western slopes of the hills, respectively (Figure 3.1B; Long *et al*., 2012). This hydrology may explain the relative stability of the lake water chemistry; sampling over the past 40 years has demonstrated little variation in pH, electrical conductivity, temperature, nutrients (N and P), and dissolved O_2 (Griebler & Lueders, 2009; Back, 2011). While presenting a variety of technical challenges for sample collection due to the significant distances from the surface, technical climbs and constricted passageways (Figure 3.1A), and depth underground, these lakes provide a rare and valuable window for directly accessing this region of the Madison aquifer.

Given the unique opportunity that Wind Cave provides to directly access an important aquifer, we examined the microbial diversity of the Wind Cave lakes and compared it to microbial communities sampled from the aquifer by surrounding wells and springs. Our results suggest that the Wind Cave lakes contain a unique ultra-oligotrophic, deep subterranean lake ecosystem dominated by bacteria, with cell numbers below those previously observed in similar freshwater environments, which cannot be directly sampled via regional wells and discharge springs.

3.3 Results

There are a number of surface springs and wells near Wind Cave that provide access to water from the Madison aquifer (Long & Valder, 2011), of which two wells (PW2 and STR) and one spring (BCS), are found in in the same hydrogeologic domain as Wind Cave (as evidenced by the similarity in stable isotope values and chemistry to lakes found in the cave; Back, 2011). In order to determine how much water would need to be filtered to collect sufficient cells for DNA extraction, we carried out cell counts (Table 3.1). The lakes at WCL contained an average of 2.93 x 10^3 cells mL^{-1}, while STR and PW2 were slightly higher, at 6.30 x 10^3 and 9.30 x 10^4 cells mL^{-1}, respectively (Table 3.1). These cell numbers are much lower than anticipated, as previous observations of other karst springs have cell counts in the range of 10^5 - 10^6 cells mL^{-1} (Farnleitner *et al*., 2005; Newton *et al*., 2011; Smith *et al*., 2012), lower than other measured bodies of water (Table A.1). Indeed, these numbers are more comparable to much deeper (~1,500 m) fracture fluids (McMahon & Parnell, 2014). The higher cell numbers seen in BCS (8.84 x 10^4 cells mL^{-1}; Table 3.1) may be due to its use as a water source for a large herd of cattle that periodically wander into the spring. Culture-based analyses on selective media indicated that this microbial population included members of the *Enterobacteriaceae* (data not shown), making it impossible to distinguish native from contaminant microorganisms, and no additional sampling was carried out at BCS.

Table 3.1. Comparative direct cell counts for the sites included in this study.

Sample	Treatment	Fields of View	Volume filtered (mL)	Cells/mL	Std Dev (%)
All sample sites					
CTL	None	200	5	5.2 x 10^1	28.8
PW2	None	100	10	9.3 x 10^4	4.2
BCS	None	100	10	8.8 x 10^4	52.5
STR	None	100	10	6.3 x 10^3	12.1
WCL	None	100	10	2.9 x 10^3	8.2
Calcite Lake					
WCL	None	100	10	3.48 x 10^3	10.8
WCL	None	100	10	1.47 x 10^3	16.7
WCL	US	50	10	6.84 x 10^3	24.3
WCL	US	100	10	1.88 x 10^3	14.7
WCL	US	100	10	1.56 x 10^3	16.2
WCL	Det. Mix	50	10	9.25 x 10^3	20.9
WCL	Det. Mix	50	10	8.85 x 10^3	21.3
WCL	Det. Mix + US	50	10	6.84 x 10^3	24.3
WCL	Det. Mix + US	50	10	8.44 x 10^3	21.8
FISH					
WCL	EUB338I/II/III	100	10	2.43 x 10^3	18.3
WCL	CREN499/ARC915	100	10	3.48 x 10^1	46.1
STR	CREN499/ARC915	100	10	*bdl*[a]	46.1
Nucleated cells					
WCL	None	100	10,000	0	0

The comparatively high arsenic concentration in PW2 (Table 3.2), suggested that the water in this well was contaminated by a significant intrusion of water from the shallower Minnelusa aquifer (Figure 3.2; Back, 2011), complicating the interpretation of the microbiology (PW2 was subsequently sealed and made inaccessible by WCNP due to high arsenic levels). We therefore focused our work on the samples obtained via the cave lake (WCL) and the accessible surface well (STR). To determine the contribution of archaea to these populations, we combined FISH with our cell counting methods. The results (Table 3.1) indicate

Table 3.2: Water chemistry of the sample sites used within this study.

Site	**pH**	**Temp**	**SO_4 (mg/L)**	**As (∝g/L)**	**NO_2+NO_3 (mg/L)**	**PO_4^- (mg/L)**	**TOC (mg/L)**	**Glucan (pg/mL)**
BCS	7.3	18.0	1,290	1.7	0.39	-	38.80	-
PW2	7.8	14.9	38.4	26.7	0.21	3.37	-	-
STR	7.6	12.8	11.4	8.5	0.51	2.44	34.58[a]	-
WCL	7.75	13.8	9.57	12.8	2.38	0.41	0.29[a]	*bdl*[b]

[a] Average values of three separate samples

[b] bdl – below detection limit

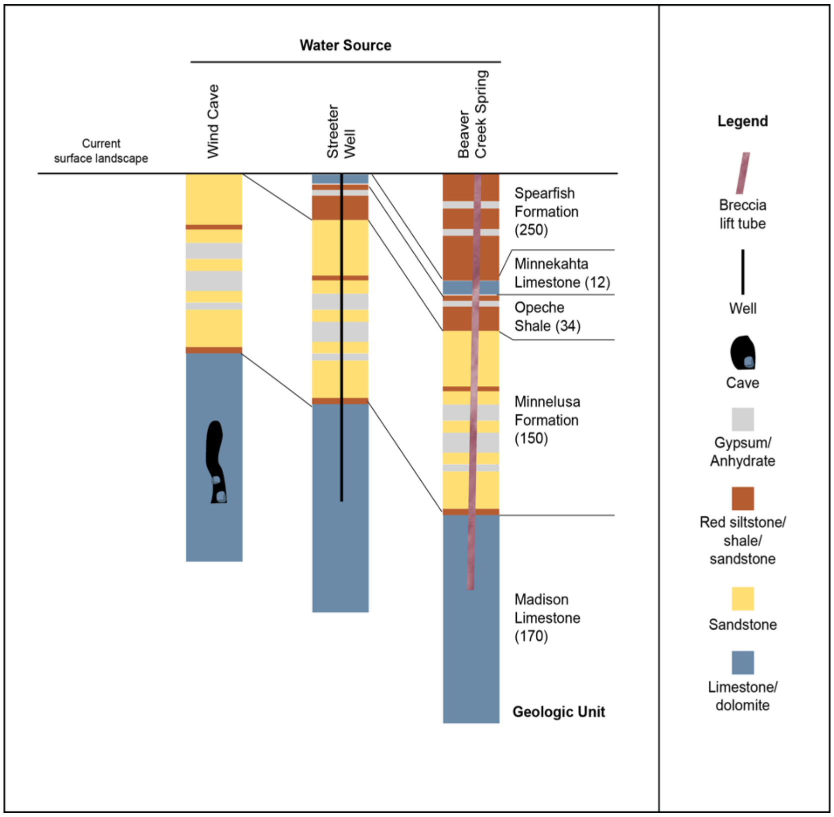

Figure 3.2. Geologic profile of sample sites. Stratigraphic profile demonstrating the overlying rocks that confine the Madison aquifer. WICA is located within the Madison Limestone Formation. STR is completed to the Madison aquifer and passes through the Minnelusa Formation (sandstone and shale), as well as the Opeche Shale. STR is a partially cased well. Beaver Creek Spring (west of WICA and STR) is located in the Spearfish Formation, with water originating from the Madison aquifer. Adapted from Epstein 2001, and Long and Valder, 2011.

that the Archaea were present in the WCL samples, albeit at a low rate (~2%). We were unable to count archaea in the STR sample above background control values (Table 3.1). During our bacterial cell counts, no eukaryotes were visible in the WCL water samples. To confirm this, we counted cells in 10 L samples (Table 3.1). Although scant fungal spores (~8 – 12 μm in diameter) were occasionally observed outside of the counting frame (data not shown), there were no identifiable eukaryotic/protozoan cells in WCL (Table 3.1). This lack of a eukaryotic population was supported by our inability to amplify any 18S rRNA gene sequences via PCR using the EK-1F/EK-1520 primer set (López-García *et al*., 2001; data not shown).

To determine whether the very low planktonic cell numbers in WCL represented a statistical outlier, samples were collected for counting from WCL over the course of six years and analyzed in two independent laboratories, using a variety of separation techniques to obtain more robust counting data. Our data (Table 3.1) indicated that cell numbers within WCL are relatively constant, with an average of 2.48 x 10^3 cells mL^{-1}, although slightly higher numbers could be observed following efforts to disrupt biofilms (7.64 x 10^3 cells mL^{-1}; Table 3.1). The results from all methods across three separate sampling periods ranged from 1.56 - 9.25 x 10^3 cells ml^{-1}, with untreated and sonicated samples being on the low end of the spectrum and detergent mix and detergent mix/ultrasonic treated samples being slightly higher (Table 3.1). As a secondary analysis to confirm these low cell numbers, we tested the water for total (1,3)-β-D-glucan, a common polysaccharide of both with Gram negative and positive bacteria. Our data demonstrates that the concentration of this polymer was below the limit of detection (<31 pg mL^{-1}). To convert this value to cell number, we compared it to (1,3)-β-D-glucan production in the model species, *Pediococcus sp.,* which generates an average level of 1.2 pg cell^{-1} (Dueñas *et al*., 2003), indicating a detection limit of 31 ng equates to 2,500 cells mL^{-1}. Thus, while quantitative (1,3)-β-D-glucan production by environmental species is unknown, our inability to detect this polymer supports the cell count data of low biomass (Table 3.1).

Cell counts from WCL and STR were used to determine the volume of water necessary to collect sufficient DNA for examining diversity in the samples: assuming approximately 3 fg DNA per cell and 2.93 x 10^{3} cells mL^{-1} in WCL, and a DNA extraction efficiency of 17% (Claassen *et al*., 2013), we determined that DNA analysis would require filtering a minimum of 200 L of water. We were able to filter ~300 L of water from STR and WCL through a 0.2 µm filter and obtained ~5 ng of DNA from each site after extraction. In order to examine the microbial community profile at each site, we used 454-pyrosequencing of 16S rRNA PCR amplicons. The number of OTUs generated for analysis were distributed as follows: CL1: 5,566; CL2: 5,817; CL3: 4,340; SW1: 5,266; SW2: 5,706; SW3: 8,945; and PCTL: 3,983.

A rarefied analysis of the sequenced products (Figure 3.3; Chao, 1984) suggested that the WCL community demonstrates a significantly higher species richness and diversity than that found at STR, despite the 100-fold reduction in the amount of available organic carbon (0.29 mg L^{-1} versus 34.58 mg L^{-1}). Indeed, the rarefaction curve for WCL suggests that the lakes contain hundreds of unique species, while STR phylotypes are more broadly represented within the analyzed data (as expected, there is limited diversity in the preparation control; Figure 3.3).

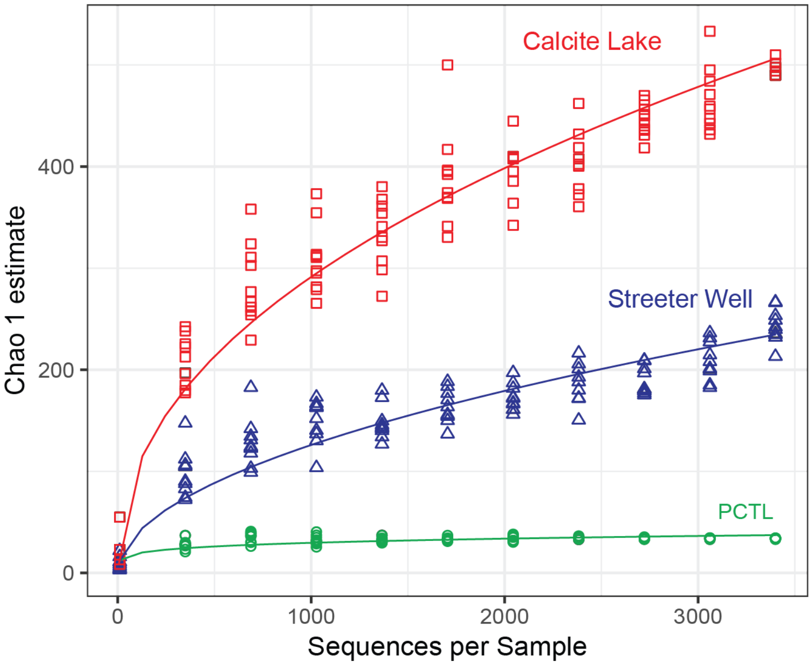

Figure 3.3: Chao 1 estimate of community species richness of WCL samples (red), STR samples (blue), and PCTL (green). Chao 1 points were generated using median abundance values of sample replicates, randomly subsampled (with replacement) at even sampling depths over 10 iterations. A best fit curve was generated using a robust linear model.

Analysis of the identified taxa revealed a broad phylum-level diversity within both the WCL and STR samples, with 14 and 11 represented phyla, respectively (Figure 3.4). The dominant phyla in WCL are similar to that observed in other cave biomes (Figure 3.4), with dominance by the *Proteobacteria* (averaged across three samples; 31%), *Firmicutes* (14%), *Chloroflexi* (12%) and *Actinobacteria* (9%), along with significant contributions by members of the *Plantomycetes* (3.5%; Hershey & Barton, 2018). The WCL samples also contained a greater contribution by members of the *Fusobacter* (6%), *Omnitrophica* (7%), *Nitrospirae* (2%) and unclassified bacterial sequences (5%) compared to other cave samples (Figure 3.4). Sequencing demonstrated an average of 4% archaea within the WCL population, primarily consisting of the *Thaumarchaeota,* in support of the microscopic data (Table 3.1 and Figure 3.4).

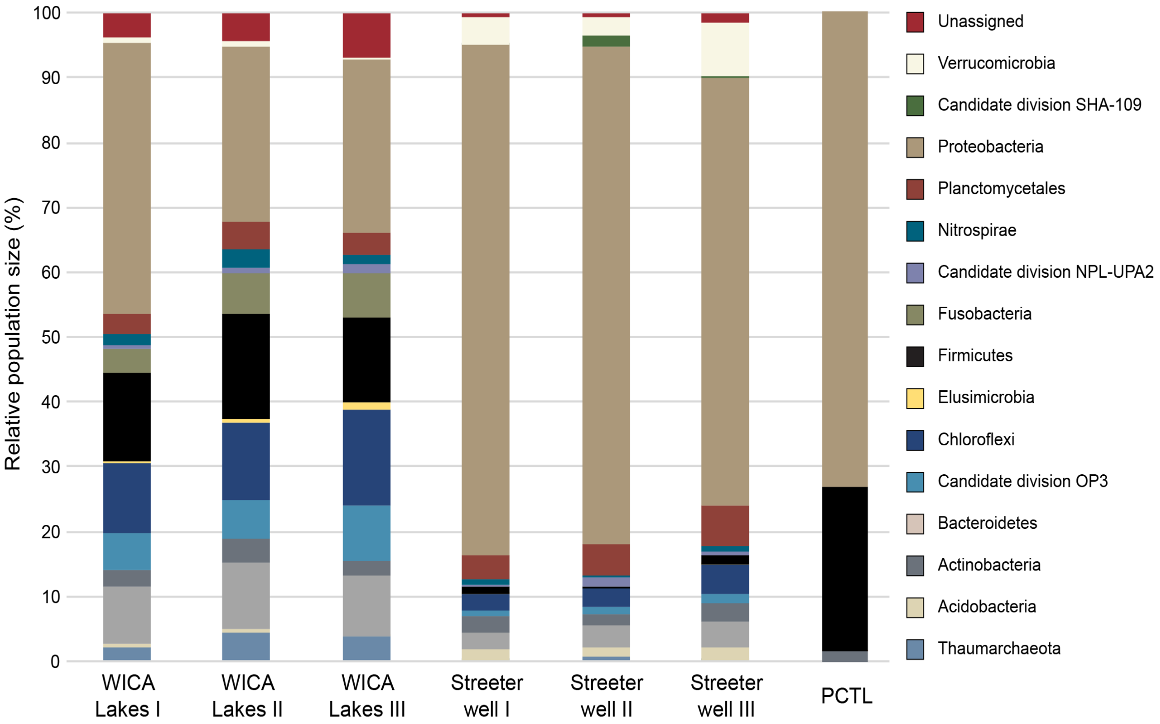

Figure 3.4: Phylum level 454-pyrosequencing data from the Wind Cave (WICA) Lakes, Streeter Well, and the processing control (PCTL). Three PCR replicates were performed for each sample, generating 5,566 (WICA lakes I), 5,817 (WICA Lakes II), 4,340 (WICA Lakes III), 5,266 (Streeter Well I), 5,706 (Streeter Well II), 8,945 (Streeter Well III), and 3,983 (PCTL) sequences after processing. Relative contribution of each phyla is shown; phyla composing less than 1% of the microbial community were not included for clarity (a more detailed breakdown of all represented clades at the Order level is shown in Figure 3.5).

Despite the much lower available organic carbon in the WCL sample, the community appeared evenly distributed across several phyla, while STR is dominated by one phylotype (Figures 3.4 and 3.5). In order to determine whether species richness was statistically different between samples, we used the reciprocal Simpson index to quantify the average proportional abundance of each taxon in the sample. The results (Figure 3.6A) suggest that there were more taxa represented in the WCL sample, with a higher proportional representation (richness) in the population (Interlandi & Kilham, 2001). This analysis confirms the observation that STR is dominated by one species. Indeed, even the minimal DNA amplified from the PCTL control appears more diverse than the community in STR (Figure 3.6A), suggesting that access to the organic carbon in STR may be selecting for this phylotype within the *Acidithiobacillales* (Figures 3.5 and 3.6).

Population differences between the sample sites were further visualized using a weighted Unifrac principle coordinate analysis (PCoA; Figure 3.6B). The PCoA plot demonstrates that the WCL community is indeed significantly different from the STR (and PCTL) populations, with very little in-group difference between the WCL samples, compared with those for STR (Figure 3.6B). Thus, while WCL and STR had comparable, low cell numbers, the available organic carbon and/or geochemistry appears to have selected for quite distinct community compositions (Figures 3.4 and 3.6).

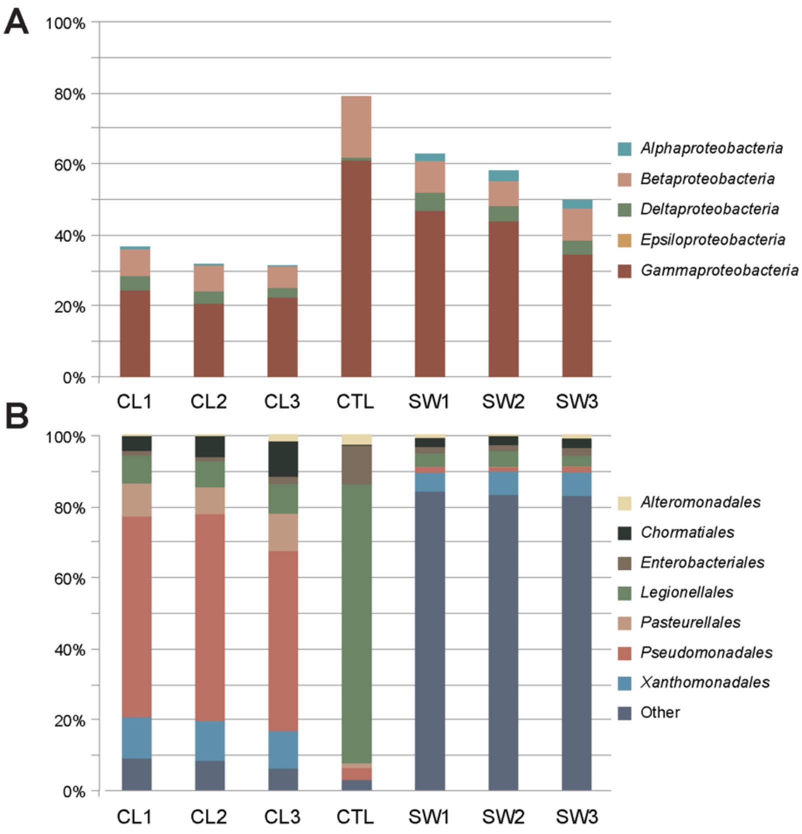

Figure 3.5. The results of the 454-pyrosequencing analysis of the *Proteobacterial* populations within Wind Cave Lakes (WCL), Streeter Well (STR), and processing control (CTL) samples. A) The total population(%) of the *Proteobacteria* within the samples, along with the sub-division level distribution within the *Alpha-, Beta-, Delta-, Epsilo-* and *Gammaproteobacteria.* B) The relative distribution of orders within the *Gammaproteobacterial* populations demonstrated in A.

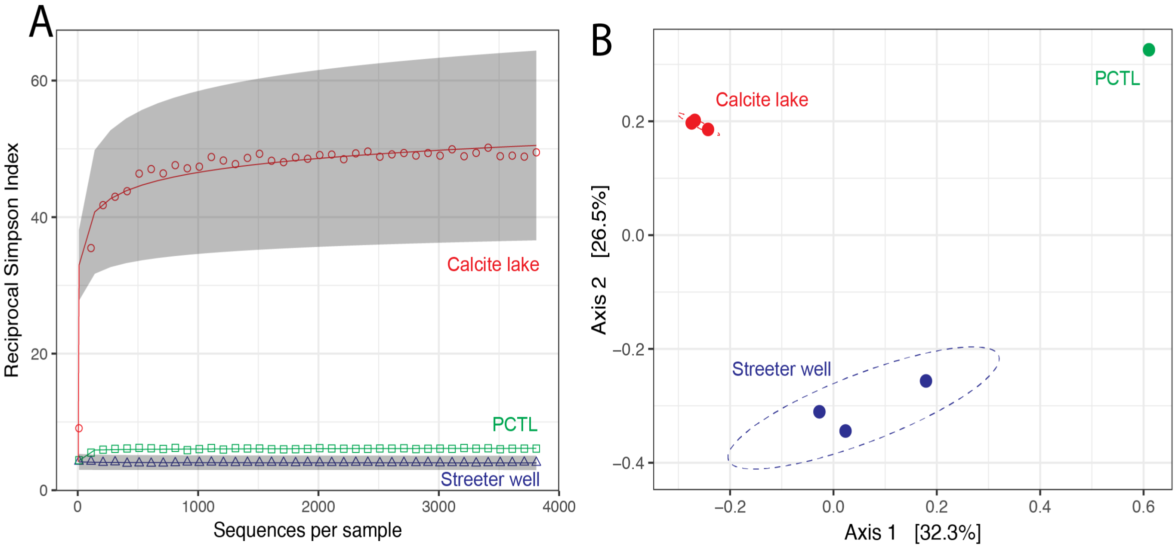

Figure 3.6. Alpha and beta diversity metrics of WCL, STR and PCTL. A) Reciprocal Simpson rarefaction plot of WCL, STR, PCTL, demonstrating community evenness. Data were randomly subsampled (with replacement) at even intervals to generate reciprocal Simpson curve. The 95% intervals are indicated by gray shading. B) Principle Coordinate Analysis (PCoA) of unweighted Unifrac distances between WCL (red), STR (blue), and PCTL (green). Ellipses depict the 95% confidence interval of the sample cluster, with centroids determined by Source mean values.

To determine whether the WCL and STR populations shared similarity with other karst communities (despite being distinct from each other), we expanded our PCoA analysis to include comparable environments (Figure 3.7A). These sites were chosen based on oligotrophic conditions, sampling of pelagic microbial communities, and 454-pyrosequencing. These comparative samples included: the Edwards karst aquifer of Texas (Edwards aquifer; SRP010407; Engel & Randall, 2011); a shallow karst aquifer in Germany (Limestone aquifer; ERP020663; Herrmann *et al*., 2017); stream water in a Kentucky epigenic cave (Cascade Cave system; SRP058014; Brannen-Donnelly & Engel, 2015); and a surface lake open to photosynthetic input (the oligotrophic Lake Brienz; SRP021556; Köllner *et al*., 2013; Figure 3.7A). Analyses are therefore based on sequencing method (454-pyrosequencing), PCR primers and average number of amplicons.

Three different distance methods (normalized-weighted Unifrac, unweighted Unifrac, and the non-phylogenetic Bray-Curtis metric) were used to generate distance matrices for principle coordinate analyses to visualize differences between the samples in this study and the data from the comparable environments. All metrics used generated similar grouping patterns in a principle coordinate analysis; due to the varying samples sizes used among studies we used the unweighted Unifrac metric (which is qualitative, rather than quantitative; Figure 3.7A). This PCoA demonstrates that the karst aquifer/spring samples exhibit similarity in the taxa identified, grouping together and remaining distinct from both the epigenic cave stream and surface lake samples (Figure 3.7A); however, even with this grouping the WCL and STR populations remain distinct from other samples, although differences in the primer sets used (V1/V3 versus V3/V4) may influence the overall result. In order to determine whether there were similarities in species richness between the karst aquifers, we again we used a comparable Reciprocal Simpson index, with a best fit projection for samples with fewer available sequences (Figure 3.7B). The results demonstrated that the site that shares the closest hydrogeological similarity to WCL, the German limestone aquifer, also demonstrated high species richness. In all cases, STR had the lowest overall diversity, despite being structurally similar to WCL in terms of geologic isolation and aphotic conditions (Figure 3.7B).

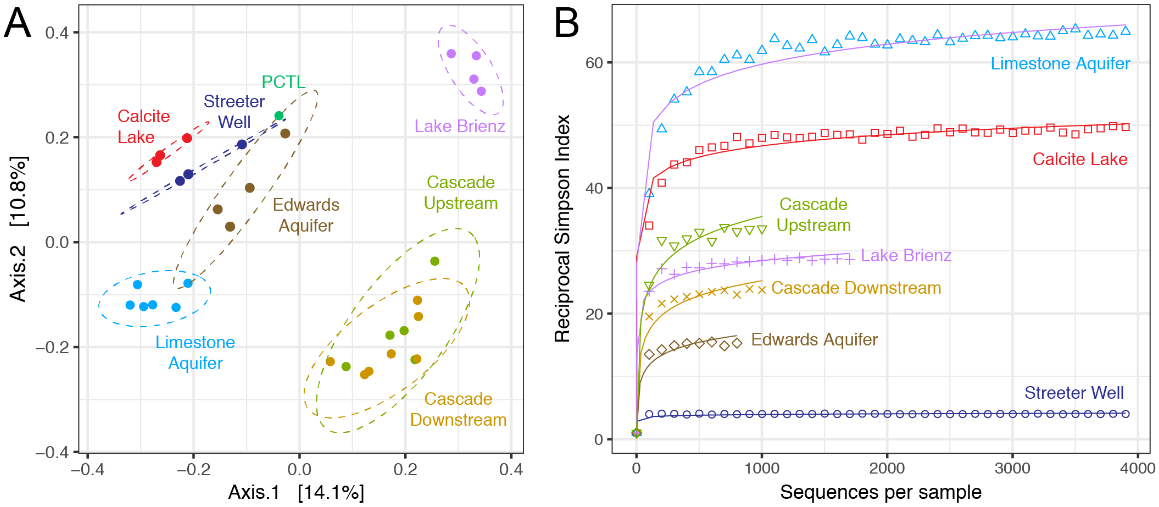

Figure 3.7. Alpha and beta diversity metrics for WCL and comparative environments. A) Principle Coordinate Analysis plot of unweighted Unifrac distances between WCL, STR, PCTL, and similarly oligotrophic environments. Comparative data include the Edwards aquifer (SRP010407), pelagic microbial populations from oligotrophic Lake Brienz, Switzerland (SRP021556), a freshwater spring (SRP144147), a limestone aquifer (ERP020663) and the oxygen minimum zone off the Gulf of California (SRP058343). Ellipses depict the 95% confidence interval. B) Comparative reciprocal Simpson rarefaction plot of WCL and STR, with the additional identified freshwater environments (as described in A), demonstrating community diversity.

3.4 Discussion

Wind Cave is one of the most geologically complex caves in the world, the structure of which provides the rare opportunity to travel into the subsurface and directly sample the microbiology of the Madison aquifer (Figure 3.1; Palmer & Palmer, 2000). Given the physical and technical challenges of examining the aquifer via the cave (Figure 3.1A), we wanted to examine how the microbiology sampled through the cave compared to the more easily accessible surface wells (PW2 and STR) and spring (BCS). Both PW2 and BCS contained a higher cell number than either WCL or STR (Table 3.1); however, high levels of As and SO_4^{2-} in PW2 and BCS indicated that these sites were heavily influenced by the shallower Minnelusa aquifer (Table 3.2 and Figure 3.5).

Due to the chemistry at PW2 and BCS, STR has traditionally been used by hydrologists to sample the Madison aquifer (Back, 2011). Despite purging more than three volumes (2,000 liters) of water from STR (as recommended in past protocols; Hose & Lategan, 2012; Korbel *et al*., 2017), our data demonstrated differences between the STR and WCL samples (Figures 3.4 and 3.6). These included a higher level of TOC (a variable not examined in previous hydrologic studies) and SO_4^{2-} at STR compared to WCL. We believe that these higher values may be due to minimal casing of the STR well (Ohms, M., personal communication, 2018), which would allow water from the shallower Minnelusa Formation to be drawn by the well (Back, 2011; Long & Valder, 2011). Thus, while the main volume of water from STR may be from the same hydrogeologic domain as Wind Cave, it is mixing with the higher organic content, SO_4^{2-}, and microbiology of the Minnelusa (Figure 3.2; Atlas *et al*., 1991; Naus *et al*., 2001).

There appears to be microbiological evidence for this water mixing within the STR well. While total cell numbers were comparable between the two sites (6.3 x 10^3 cells mL^{-1} at STR versus 1.56 - 9.25 X 10^3 cells mL^{-1} at WCL; Table 3.1), taxonomic comparisons demonstrated that the microbial populations were significantly different (Figure 3.4). While both the WCL and STR samples were both dominated by members of the *Gammaproteobacteria,* at the order level there were important differences (Figure 3.5). Within WCL, the *Gammaproteobacteria* (18% of identified OTUs) were dominated by members of the *Pseudomonadales, Xanathomonadales,* and *Chromatiales.* In STR, the much higher *Gammaproteobacterial* population (~73%), comprised a single OTU within the *Acidithiobacillales* (Figures 3.4 and 3.5). This OTU had the greatest sequence identity to an uncultured *Acidithiobacillales* clone KCM-B-112, which has been found in environments with high hydrocarbon content (Garrity *et al*., 2005; Marti *et al*., 2017). While this suggests that hydrocarbon breakdown may drive community energetics within STR, the closest cultured species, *Thioalkalivibrio sulfidiphilus,* is a chemolithotrophic, sulfur-oxidizer, suggesting that S-cycling may also be important (Muyzer *et al*., 2011). Nonetheless, *T. sulfidiphilus* only has 89% 16S sequence identity to KCM-B-112, and the low SO_4^{2-}-levels in STR make such metabolic inference difficult without additional biogeochemical support. The presence of similar members of the *Acidithiobacillales* at low abundance (<0.2%) in the WCL samples, suggests that STR may be seeded by microorganisms from the aquifer, but are selected for by the geochemistry of the well environment or grow within the Minnelusa Formation (Figure 3.4 and 3.5). In contrast, WCL is completely contained within the Madison limestone formation and not exposed to the organic content of the Minnelusa Formation (Figure 3.2). Indeed, the TOC in WCL is much lower (0.29 mg L^{-1}) compared to STR at 34.59 mg L^{-1}, although the SO_4^{2-} is comparable (Table 3.2). Our work supports the findings of Korbel *et al*., (2017), who determined that despite purging of stagnant water from wells and boreholes, microbial community analyses can still be influenced by variances in well casings, geologic strata through which wells pass, and even the purging strategies used (Hose & Lategan, 2012; Korbel *et al*., 2017). Thus, while STR may be a more easily accessible sample site for access to the water chemistry of the Madison aquifer, accurate microbiological investigations of this important water source require sampling via the cave. Consequently, the samples collected through the cave provide the most accurate window into the microbiology of the Madison aquifer.

At the phylum level, in addition to *Gammaproteobacteria,* the WCL community contained members of the *Actinobacteria* and *Firmicutes,* and significant contributions from the *Planctomycetales, Nitrospirae, Fusobacteria, Chloroflexi* and *Omnitrophica* (Figure 3.4). The *Actinobacteria* and *Firmicutes* are dominated by members of the *Actinomycetales, Bacillaceae, Pseudomonadales,* associated with carbon turnover in soils (Barton, 2015). The high levels of NO_2^- $+NO_3^-$ present in the lakes may support the growth of autotrophic nitrite oxidizing *Nitrospirae* (Daims *et al*., 1999; Lucker *et al*., 2010). It is unclear as to a potential source of this NO_3^- in WCL; however, the nearby South Dakota Badlands are comprised of caliche, which are nitrate-rich paleosols deposited when conditions were more arid in this semi-arid region. Given the close association of the Badlands with a local recharge zone or the potential for deep burial, it may be that the waters pass through such deposits *en route* to WCL. Members of the *Nitrospirae* are primarily associated with nitrite oxidation, with some members capable of complete nitrification of ammonia to nitrate (Daims *et al*., 2015). The *Nitrospirae* have also been shown to be slow growing and metabolically flexible, with the ability to grow mixotrophically under microaerophilic or anaerobic conditions (Fujitani *et al*., 2014; Koch *et al*., 2015). Together with the presence of ~2% ammonia-oxidizing *Thaumarchaeota,* this may indicate an active nitrogen cycle driving autotrophic growth within the lakes (Figure 3.4).

With limited or no cultured representatives, the metabolic activity of the *Planctomycetales, Chloroflexi* and *Omnitrophica* make their activity in the environment difficult to predict. Nonetheless, these phyla are often highly represented in oligotrophic environments, particularly caves, where they appear to grow using oxidized-carbon compounds under extreme nutrient-limited conditions (Fuerst & Sagulenko, 2011; Barton, 2015). Members of the *Omnitrophica* are widely distributed in subsurface environments (Rinke *et al.,* 2013), but their potential metabolic role is poorly understood (Glöckner *et al*., 2010; Shabarova & Pernthaler, 2010; Momper *et al*., 2017). Recent work suggests that members of the *Omnitrophica, Nitrospirae* and *Planctomycetales,* are involved in iron-oxidation, and may play important roles in the geochemical cycling of iron and sulfur in the environment (Lefèvre, 2016; Lin *et al*., 2017; Momper *et al*., 2017). These data, along with the observation of other iron-cycling genera, such as *Geothrix,* suggest that iron (which is abundant within Wind Cave) may be an important driver of autotrophic growth (Kolinko *et al*., 2012; Lefèvre *et al*., 2013). These data suggest that autotrophic growth within WCL may be driven by iron, nitrogen and carbon cycling, similar to that observed in similar German karst aquifer and the nearby deep Sandford Underground Research Facility (Herrmann *et al*., 2017; Momper *et al*., 2017).

At the phylum level, the taxa observed in WCL appeared to be similar to that seen in other karst environments (aquifers and springs) and remain distinct from communities influenced by surface nutrient sources, such as Lake Brienz (Figure 3.7A). The WCL samples cluster together with the other karst aquifers but remain distinct from the epigenic cave stream (created where surface waters flow into the subsurface; Figure 3.7A). The STR sample intersects within the 95% confidence interval of the Edwards aquifer data, which itself intersect with the German limestone aquifer. This is likely due to similar sampling strategies, with the German limestone aquifer and Edwards aquifer using well access and similar primer sets; however, humanly accessible caves do not intersect these aquifers at depth, preventing in-cave sampling (Engel & Randall, 2011; Herrmann *et al*., 2017), while the limited access to WCL has prevented re-sampling with the same primer sets to confirm this grouping. In the PCoA plot, the PCTL sample was close to the confidence intervals of two of the sampled sites (Figure 3.7A). Although this clustering may be due to resolution (when compared to WCL and STR alone, the PCTL sample was quite distinct), it suggests that control sample data from low biomass samples remain an important dataset to demonstrate the absence of sample/sequencing contamination (Barton *et al*., 2006). The PCTL itself was dominated by members of the *Legionella,* which generally live inside amoebae in the natural environment (Harf & Monteil, 1988). We were unable to detect eukaryotic microorganisms within the lake via PCR amplification or direct microscopic observation, and the dominance of *Legionella* in the preparation control may reflect contamination of purchased buffers (DNase- and RNase-free laboratory water controls did not contain such contamination). Such contamination has been shown to be present in commercial laboratory reagents, further emphasizing the importance of preparation controls when working with low biomass samples (Shen *et al*., 2006).

One of the most notable observations from this study was that the WCL samples, under the lowest nutrient conditions, had amongst the highest observed diversity (Figure 3.7B). This has been observed before: described by Hutchinson as the "*paradox of the plankton*" (Hutchinson, 1961). Microbial ecologists have attempted to describe the differences that account for this phenomenon, arguing that environmental heterogeneity, mixing and even predation prevent the establishment of true competitive exclusion, allowing for high species richness even in the most starved environments (Czárán *et al*., 2002; Kerr *et al*., 2002; Torsvik *et al*., 2002; Cadier *et al*., 2017). Such obvious mechanisms of heterogeneity are absent at the lakes within Wind Cave: the cave has stable environmental conditions, with no diurnal, seasonal or annual variation; the cave entrance is miles from the sample site, and there is no airflow sufficient to facilitate lake water mixing; and no potential eukaryotic predators/grazers were detected (Table 3.1). An alternate explanation is that of Interlandi and Kilham (2001), who examined planktonic diatom diversity in oligotrophic lake systems. Their work demonstrated a correlation between the number of limited resources, biomass and diversity (measured using the Simpson's index; Interlandi & Kilham, 2001). These investigators argued that resource competition itself was driving high diversity: despite the bulk water chemistry being homogenous, at the scale of individual microbial cells, there is sufficient resource variation to promote competition between species (even in low biomass environments; Interlandi & Kilham, 2001). The absence of light, low biomass, ultra-oligotrophic conditions, and limiting P may therefore contribute to the high diversity observed within both WCL and the German limestone aquifer (Figure 3.7B; Herrmann *et al.,* 2017).

This work represents the first study of the microbiology of a deep, geologically isolated aquifer sampled directly via access through a cave. Our data suggest that Wind Cave provides a unique access point to study the microbial community of an aquifer without the confounding variables introduced via well or spring sampling (Yanagawa *et al*., 2013). The data suggest very different community energetics than would have been suggested through sampling a well and indicate that the microbial communities have a much higher potential to influence the quality of drinking water than would have been determined from well-based analyses (for example, denitrification; Herrmann *et al.,* 2017). Nonetheless, the growing population and development of the Black Hills region means that demand on the Madison aquifer has increased. Recent permits have been approved to increase the amount of water drawn from the Streeter well to

~91,000,000 L year^{-1}, with additional wells in the Black Hills and Fall Rivers water districts predicted to draw an additional ~1,420,000,000 L year^{-1}. If these wells draw water faster than the local recharge rates, this could result in a dramatic drop in the potentiometric surface of the Madison aquifer (Greene, 1993). If the aquifer were to drop below the current level of the cave, access via Wind Cave (and the unique ability to sample the microbiology independent of well access) will be gone. It is therefore critical to carry out as many microbiological analyses of this important aquifer before the site is lost.

3.5 Acknowledgements

We would like to thank the staff at Wind Cave National Park for in cave assistance and carrying equipment to the sample sites, in particular Marc Ohms, for his never-ending patience with the repeated sampling expeditions over the years, and Jeff Hughes for his faith in the interesting microbiology of the Wind Cave lakes. We would also like to thank Juan Giarrizzo for the initial development of the sampling equipment that allowed us to filter 300 L of water under the weight and size constraints of the cave, DNA extraction, PCR analysis and sequencing. We would also like to thank the contributions of the anonymous reviewers, whose comments greatly increased the quality of this paper. This research was funded in part by the US National Park Service (to HAB), the Cleveland Grotto Science and Choose Ohio First Bioinformatics Fellowship to (OSH).

CHAPTER IV

MANGANESE AUTOTROPHY AND HIGHLY MOBILE GENETIC INFORMATION CHARACTERIZE THE MICROBIOLOGY OF A DEEP SUBSURFACE AQUIFER

4.1 Abstract

Subsurface aquifers are an important source of freshwater; however, our understanding of the biogeochemical processes that support microbial ecosystem dynamics remains limited due to the difficulty of accessing these environments. The lakes in Wind Cave (WCL) provide access for directly sampling the Madison aquifer, a regionally significant aquifer that contains one of the lowest cell numbers of any water on Earth. Despite this low biomass, the microbial community found there is remarkably stable and diverse. The low number of microbial cells in this environment has made sampling challenging; however, using tangential flow filtration, we were able to collect enough cells for metagenomics to study the community interactions and processes that supported life under the ultraoligotrophic conditions (0.29 mg L^{-1} TOC). Using comparative filtration, our data demonstrate that the WCL community is enriched in ultrasmall cells, particularly within the Patescibacteria and Nitrospirota. Metagenomic analyses suggest an increase in genotypes associated with an oligotrophic lifestyle, including carbohydrate metabolism and amino acid scavenging. A reduction in evident prophage, along with a significant increase in integron density, suggests that integrons provide the genetic plasticity and metabolic flexibility necessary for adaptative evolution and survival. Finally, primary productivity in the ecosystem appears to be driven by chemolithoautotrophic manganese- and nitrite-oxidation, while evidence of redox active Mn-oxides may in-turn aid in the oxidation of recalcitrant organic carbon to release low molecular weight organics for growth. These suggest that due to the unique geochemical conditions of karst aquifers, the Mn geochemical cycle may play a principle role in subsurface primary productivity, while subsurface Mn-reduction may play an important, but as yet underappreciated role in the carbon cycle.

4.2 Introduction

Subsurface aquifers play an important, but often overlooked, role in the global hydrologic cycle, contributing as much as four times the freshwater discharge of rivers and streams (Taniguchi *et al*., 2002; Kwon *et al*., 2014). Areas such as the Mediterranean Sea can receive up to 75% of their freshwater input from karst groundwater, while many US states are highly dependent upon such groundwater for irrigation and drinking water supplies (Maclay, 1995; Hudak, 2000; Garcia-Solsona *et al*., 2010). There is significant literature examining the role of indigenous microorganisms in geochemical cycling to maintain water quality in karst aquifers (Farnleitner *et al*., 2005; Goldscheider *et al*., 2006; Griebler & Lueders, 2009; Pronk *et al*., 2009; Wilhartitz *et al*., 2009; Iker *et al*., 2010; Gray & Engel, 2013); however, most research is concerned with point source contamination (Dojka *et al*., 1998; Boyer & Pasquarell, 1999; Cho & Kim, 2000; Ashbolt *et al*., 2001; North *et al*., 2004; John & Rose, 2005; Sinreich *et al*., 2014). There remains very little data on the role of microorganisms in maintaining water quality in deep karst aquifers, such as the regionally significant Madison aquifer, which is the main water supply for five US states and three Canadian provinces.

Subsurface aquifers are typically accessed for sampling via springs or wells/boreholes drilled from the surface. Nonetheless, these methods are prone to contamination from well-casing bacteria or communities residing in the rock units through which the well is drilled (Basso *et al*., 2005; Korbel *et al*., 2017; Hershey *et al*., 2018; Hershey *et al*., 2019). To avoid such confounding contamination, we have been studying the microbial diversity of the Madison aquifer by travelling through Wind Cave (Wind Cave National Park, South Dakota, USA), where at -120 m a series of lakes are found where the cave intersects the potentiometric surface of the aquifer (Hershey *et al*., 2018). Our data demonstrated that the microbial community profile of the aquifer sampled through the cave is significantly different from a nearby well, which had been traditionally used to sample water chemistry; the well samples were dominated by microbial species that utilize hydrocarbons from the overlying Minnelusa formation, even after extensive purging of the well before sampling (Hershey *et al*., 2018). These data, combined with additional hydrogeological and chemical evidence, demonstrated that Wind Cave provides the only direct means of assuredly sampling the microbiology of the Madison aquifer (Back, 2011; Long & Valder, 2011; Long *et al*., 2012; Hershey *et al*., 2018). While the Wind Cave lakes (WCL) provide a valuable sampling location, accessing the site involves crawling, climbing and squeezing (sometimes through gaps ~20 cm wide) for over 3 km through the cave, while hand-carrying all necessary equipment and power, making research efforts logistically challenging (Hershey *et al*., 2018; Hershey *et al*., 2019).

Our past results indicated that isolated in the deep subsurface from photosynthetic input, the Madison aquifer is an ultra-oligotrophic environment, containing less than 0.3 mg L^{-1} available organic C (Hershey *et al*., 2018). The water also contains one of the lowest cell numbers of any body of water tested, measured at 2.3 x 10^3 cells mL^{-1}; the only places where cell numbers have been measured below this value are inside deep (>3,600 m) South African gold mines and in accretion ice from Lake Vostok, Antarctica (Karl *et al*., 1999; Borgonie *et al*., 2011; Hershey *et al*., 2018). This low biomass makes sampling additionally challenging, and while our initial study filtered 400 L of water, we were only able to collect sufficient material for comparative amplicon-based 16S rRNA sequencing. These samples revealed that despite the low biomass, the microbial community of these lakes demonstrated a high species richness (Hershey *et al*., 2018), and direct cell counting and fluorescent in situ hybridization (FISH) demonstrated that the microbial community is dominated by Bacteria (~90%), with the remaining population represented by members of the Archaea (Hershey *et al*., 2018). Despite multiple attempts, we have not identified any protists, with the only detected Eukarya found as scant fungal spores (Hershey *et al*., 2018). We believe that this lack of Eukarya is unique and due to the extremely low biomass of this system.

While 16S rRNA data can be used to infer microbial ecosystem dynamics in well-studied environments, such as soils and the human microbiome, the WICA lakes contain a significant proportion of OTUs from phyla with few or entirely uncultured representatives, making such metabolic inference challenging (Langille *et al*., 2013; Rinke *et al*., 2013; Douglas *et al*., 2020). Metagenomic sequencing can overcome this limitation by allowing community metabolic interactions to be estimated through functional gene content; however, obtaining sufficient DNA for the volume of sequencing necessary requires substantially more material than we were able to obtain using our traditional filtration techniques (Hershey *et al*., 2018; Hershey *et al*., 2019). In this study we used tangential flow filtration to filter >1,000 L of the Madison water for DNA extraction and comparative metagenomic analyses. These analyses suggest that the structure of the unusual microbial community of the Madison aquifer is influenced by Mn redox chemistry driven by Mn(II) chemolithoautotrophy, while the presence of integrons suggests a high level of horizontal gene transfer between species. We believe these adaptations are critical to ecosystem energetics and organic carbon turnover, supporting microbial community growth and subsistence isolated from surface nutrient input.

4.3 Results

Our previous work collected samples from the Wind Cave Lakes (WCL) using a 0.22 µm filter; however, single-cell whole genome sequencing (SCWGS) suggested that the WCL also contained a significant population of ultra-small cells (Hershey *et al*., 2018; Beam *et al*., 2020). To test the relative abundance and population structure, we filtered ~1,400 L of lake water using a large volume concentration (LVC) tangential flow filter with a 45 nm pore polysulfone membrane during two sampling campaigns (one in 2017 and one in 2018). The cells were collected from the filter on site using an elution buffer, yielding ~400 mL of concentrated cells. To examine the different population sizes, the collected cell sample was split into two 200 mL volumes on site; one volume was immediately preserved with ethanol for transport, with the remaining cells filtered through a 0.22 µm pore disposable Nalgene filter. To collect the >0.22 µm cell fraction, the Nalgene membrane was cut from the filter using a sterile scalpel and stored in 70% ethanol, while the filtrate was preserved with ethanol. This provided three separate sample types at each sampling event: total cell population (Full), cells large enough to be trapped on a 0.22 µm filter (>0.2 µm), and the cell population able to pass through 0.22 µm (<0.2 µm). DNA was extracted from all three sample types;

however, the <0.2 μm fraction was challenging to work with as the cell pellet was very small and began diffusing immediately following centrifugation. We were therefore able to extract sufficient DNA for 16S rRNA Illumina iTag sequencing from all samples, but only enough DNA for metagenomic analyses from the Full and >0.2 μm cell fractions.

Community Structure

Illumina iTag sequencing demonstrated that the WCL (Full) community was structurally different than that previous collected using a 0.22 μm filter, including an increase in the phyla *Verrucomicrobiota, Nitrospirota, Elusimicrobiota* and *Methlomirabilota* (Figure 4.1). To determine whether these increases were related to population capture, we compared the <0.2 μm and >0.2 μm cell size fractions to the Full population (Figure 4.1). Unsurprisingly, there appeared to be a much larger contribution of phyla known to produce large cell sizes, such as the Proteobacteria in the >0.22 μm sample, with a simultaneous reduction in this group within the <0.22 μm fraction (Figure 4.1). The relative population changes between the two fractions (Figure 4.2) demonstrated that there was an increase in phyla known to include small cell populations in the <0.2 μm fraction, including the Patescibacteria (*Parcubacteria*/OD1, *Microgenomates*/OP11), Verrucomicrobia and *Nanoarchaeota*; however, there was also an enrichment in the population of *Nitrospirota*, along with the *Firmicutes* and remaining unassigned bacteria. The increase in *Firmicutes* may be due to spores, but the dramatic increase in the

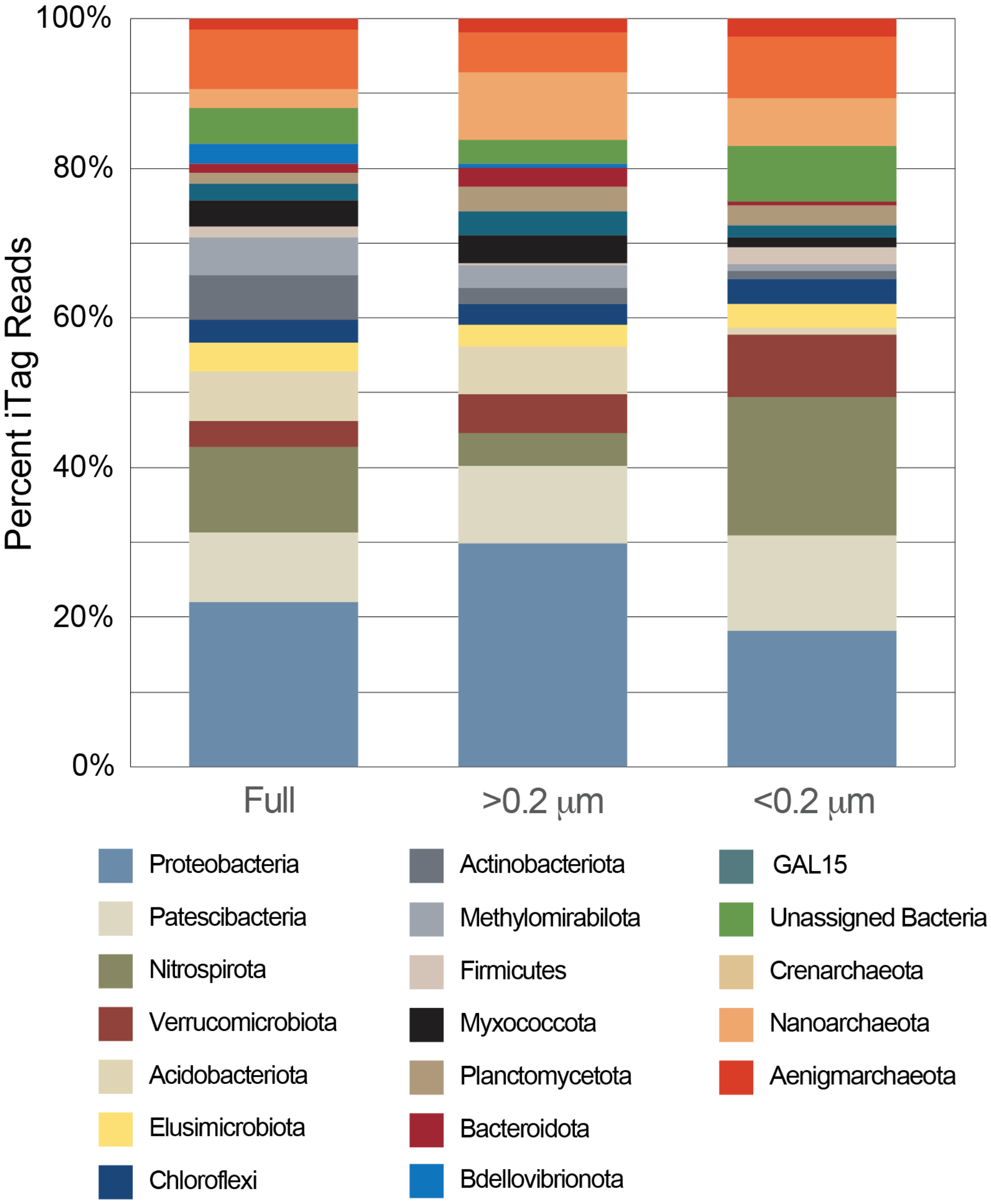

Figure 4.1. 16S rRNA Illumina iTag data from the Wind Cave Lakes (WCL), demonstrating the relative contribution of each phylum in the total cell population (Full), the cell fraction captured on a 0.22 µm membrane filter (>0.2 µm), and the cell fraction able to pass through a 0.22 µm filter (<0.2 µm). Phyla composing less than 0.1% were not included for clarity.

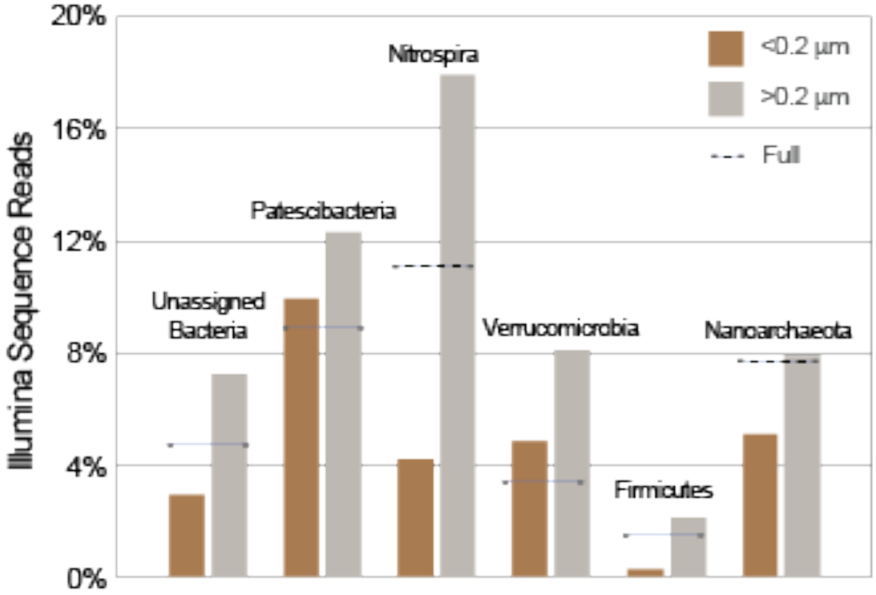

Figure 4.2. Select phyla, including those known to contain small cell populations, and their relative abundances in the large (>0.22 µm) and small (<0.22 µm) fractions of the WICA microbial community. The dashed line represents the relative abundance of each phylum in the total population sample.

Nitrospirota suggests that this population may make a significant contribution to community structure that was previously unobserved due to a very small adaptive cell size (Miyoshi *et al*., 2005; Kuhn *et al*., 2014; Luef *et al*., 2015).

Analysis of 16S rRNA sequences extracted from metagenomic contigs also showed a high proportion of unclassified bacteria and archaea (~25%) which provides support that the unassigned OTUs in the 16S rRNA iTag data were not artifacts. Up to 25% of the metagenomic 16S sequences belong to *Parcubacteria* and *Microgenomates*, both of which are included in the *Patescibacteria* superphyla of the Candidate Phyla Radiation (CPR), further supporting that the WICA lakes are richer in MDM than previously thought. Conversely, the proportion of *Nitrospirota* (~4%) and *Proteobacteria* (~7%) were much lower than what was observed in the 16S iTag sequencing, likely due to amplification bias (Suzuki & Giovannoni, 1996; Pinto & Raskin, 2012; Kennedy *et al*., 2014).

Comparative metagenomics

To identify the key microbial community activities within WCL, we carried out comparative metagenomics using publicly available metagenomes from similar environments where the planktonic microbial community had been sampled; this included groundwater, the deep oceanic subsurface, and freshwater lakes (Table A.2 – A.7). We also included other sampled cave environments to rule out the general influence of cave physiochemical conditions. Using Prodigal for gene recognition resulted in 497,033 predicted genes within the WCL metagenome, which were annotated with the SEED database using SUPER-FOCUS to identify a metabolic role. Of the genes identified, we observed an enrichment in protein (11.6%), carbohydrate (11.4%), and amino acid metabolism (11.1%; Figure 4.3). Unsurprisingly, given the geologic isolation and absence of light in the WCL, photosynthesis (<0.05%) was the least abundant functional category (Figure 4.3). However, only 13% of the predicted genes were matched to a SEED function, so we additionally annotated these genes using the Cluster of Orthologous Groups (COG) database, which assigned ~55% genes to a COG entry. A significant portion of the COG assignments (10.4%) remained poorly characterized (Class R general function and prediction, and Class S unknown); however, the most common functional COG classes were translation, ribosomal structure, and biogenesis (9.5%), cell wall/membrane/envelope biogenesis (8.1%), and amino acid transport and metabolism (7.8%).

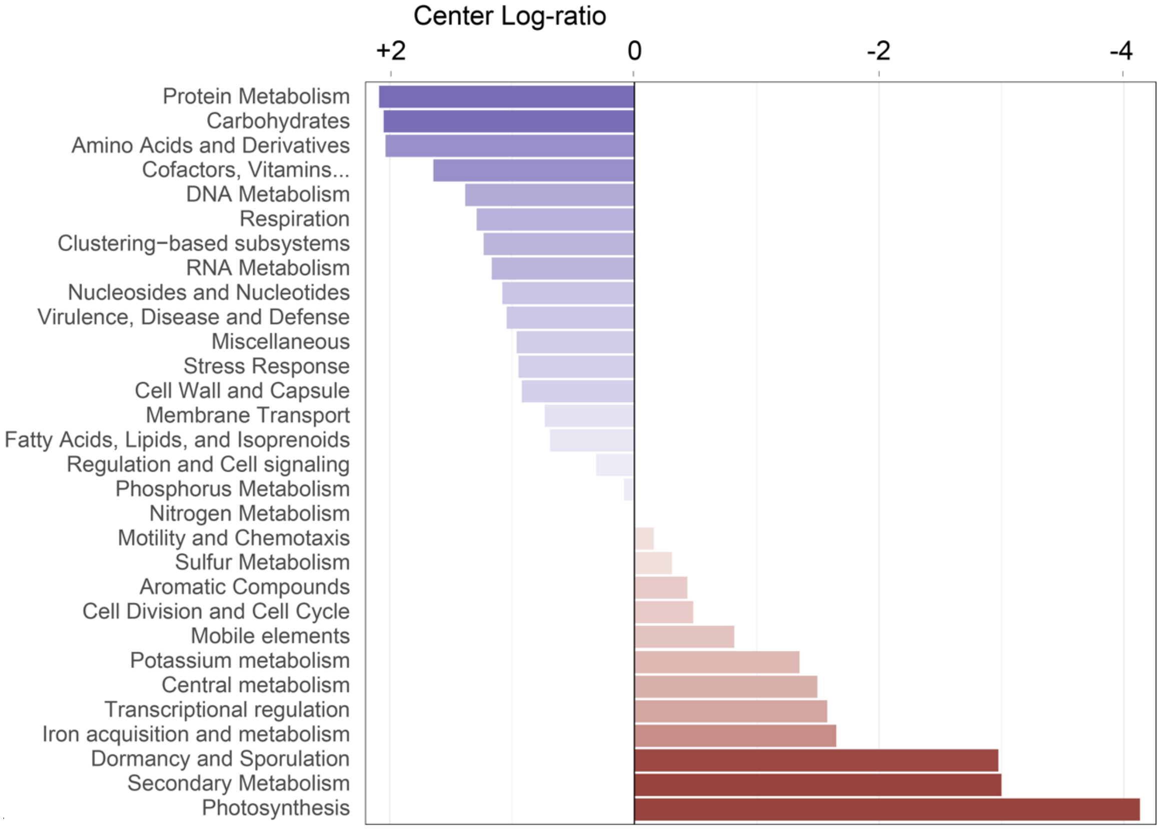

Figure 4.3. Centered log-ratio (clr) of SEED Level 1 functional gene categories in the WCL metagenome. Values represent the ration between the counts for each feature and the geometric mean count for all features in the sample.

In order to determine the variance in functional gene content between the comparative environments examined, a scaled PCA plot (Figure 4.4) generated from the compositional metagenome dataset. This PCA plot demonstrated that the WCL samples were functionally distinct from metagenomes collected from groundwater, freshwater, and deep ocean subsurface samples, although it did group with other caves, possibly through shared geochemical conditions (pH, Ca^{2+}, low TOC). Nonetheless, the WCL metagenome was in a distinct functional group that was closer in identified gene function to the Mariana and Kermadec deep ocean trenches (Figure 4.4). Interestingly, both groundwater samples obtained from underground research facilities (Sanford and Honorobe) did not group closely with our groundwater aquifer samples.

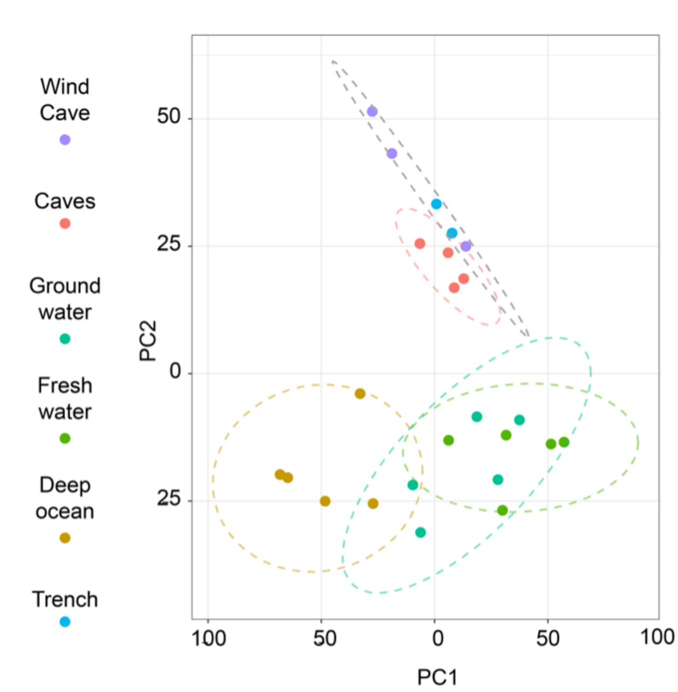

Figure 4.4. A scaled principle component analysis (PCA) plot of the variance of functional gene composition between each biome type examined.

SEED annotations were examined using differential abundance analysis to identify any metabolic pathways or processes that may distinguish WCL from other biome types. This analysis showed that sulfur metabolism, nitrogen fixation, and several carbohydrate utilization pathways (including monosaccharide and sugar alcohol catabolism, and fermentation) were underrepresented in WCL compared to the other biomes examined. Genes involved in nitrate and nitrite ammonification were present at levels similar to those found in other cave, freshwater, and groundwater biomes, but were considerably higher than the deep ocean subsurface and oceanic trench biomes examined.

Few major catabolic or anabolic pathways were more abundant in WCL than in the other biome types. Instead, cellular functions, such as protein folding and chaperone/usher pathways, DNA recombination and repair, and osmoregulation, were prevalent in WCL. Peptide methionine sulfoxide reductase, an enzyme that repairs oxidative damage to methionine peptides in proteins, was significantly increased in WCL over the other biomes, while other genes for coping with oxidative stress were not significantly elevated (Singh *et al*., 2018). Overall, ATP synthases in WCL were increased, with WCL having more FoF1-type ATP synthases than freshwater, groundwater, and oceanic trenches, and more V-type ATP synthases than oceanic trenches and other cave biomes.

Integrons

One of the notable features during our metagenomic comparisons was the abundance of integron integrase (*intI*) genes. *IntI* is the catalytic component of the site-specific integron gene cassette assembly platform (Mazel, 2006; Cambray *et al*., 2010), which was a significantly overrepresented SEED function in WICA compared to all other biomes examined (Figure 4.5). Because integrons often contain numerous repetitive sequences, they are difficult to assemble from metagenomic sequence data; however, by identifying the *attC* insertion sites for integrase and identifying the upstream ORFs, the number and function of integron-associated genes can be estimated (Buongermino Pereira *et al*., 2020). The average number of *attC* sites identified in WCL metagenome varied between 0.84 and 1.44 per Mb of assembled sequence, which was significantly higher than all the other metagenomes examined (0.04 to 1.25 copies per million bases; Figure 4.6A). The predicted number of integron-associated ORFs was also higher in WCL (1.06 to 1.95 integron-associated ORFs per million bases) when compared to the other samples (0.06 to 1.72 ORFs per million bases).

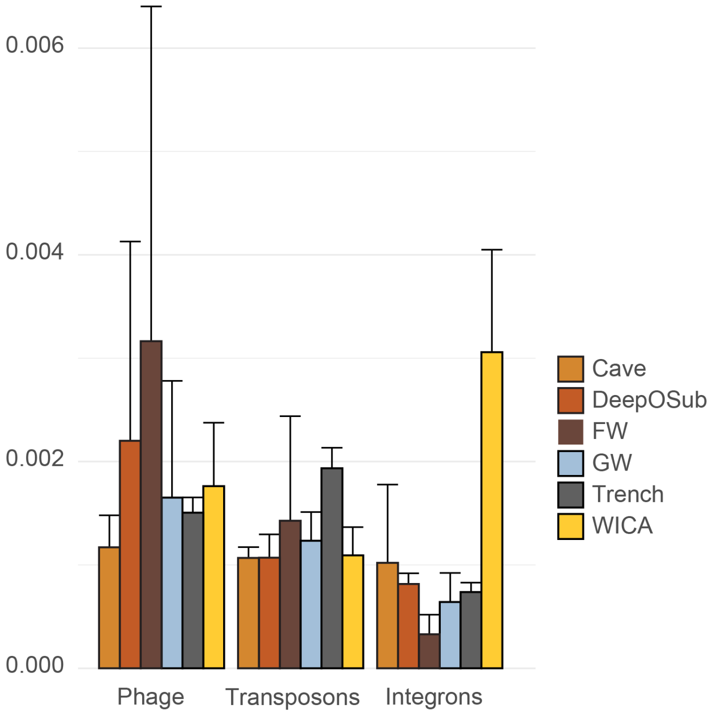

Figure 4.5. The relative abundance of SEED functional genes in the phages, prophages, transposable elements, and plasmids SEED subsystem.

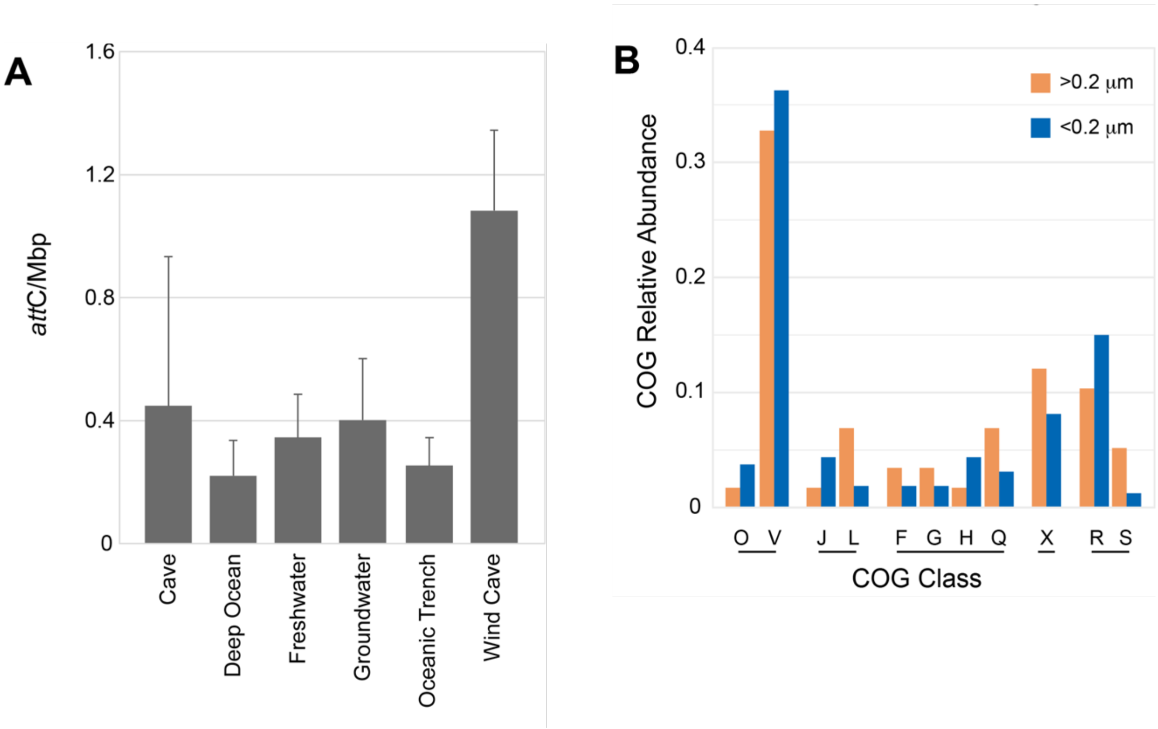

Figure 4.6. The relative abundance of *attC* sequences and integron-associated genes assigned to COG Classes. A) Averaged relative abundance of *attC* insertion sites (per Mbp) in each biome type examined, identified using Migfinder. B) Relative abundance of COG assignments for integron-associated ORFs identified in the large cell fraction (>0.2 μm) and Full microbial community. For clarity, only some COG classes are shown.

ORF assignment to SEED functions was limited due to the shorter length of the extracted integron-associated genes, with only 49/1373 (3.2%) genes assigned. Assignment to COGs was more successful, with 446/1373 (32.5%) of genes assigned, of which 373 had a known function (Figure 4.6B). The genes for defense mechanisms (COG Class V, 33.9%) were the most abundant of the integron-associated genes (Figure 4.6B), although nearly 75% of these were components of toxin-antitoxin systems, which have been shown to play a role in stabilizing large gene integron cassette arrays (Mazel, 2006; Cambray *et al.*, 2010). Genes for these toxin-antitoxin systems were also significantly more prominent in WCL SEED data, with several separate toxin-antitoxin pairs (*Phd-Doc, YdcE-YdcD, MazEF*, and others) represented. Genes involved in transcription (Class K, 6.1%) and mobilome-related genes (Class X, 9.4%) were the next most abundant COG Classes identified in the integron-associated genes (Figure 3B). When compared to genes identified in the whole WCL metagenome, Class V (defense) and Class X (mobilome), as well as the poorly characterized Class R and S genes were overrepresented in integron-associated genes. Comparatively, nearly all of the COG Classes related to metabolism and information processing associated with integrons were either underrepresented or not significantly different from the total metagenome. The only metabolic exception was Class Q genes, involved in secondary metabolite production (Figure 4.6B).

Integrons are often classified as mobile or chromosomal based on their association with mobile genetic elements (MGEs), such as transposons and recombinases (Mazel, 2006; Boucher *et al*., 2007; Gillings, 2014). To estimate the proportion of mobile integrons in WCL, we classified integrons found within 2000 bases of an MGE as putative mobile integrons, while the remainder were considered to be chromosomal. Of the 623 integron gene arrays identified, only 65 (10.4%) were putatively mobile, carrying 234 ORFs. Of these ORFs, 102 (43.5%) were assigned to a COG, of which 31% belonged to Class V, 21.9% belong to Class X, and 11% were poorly characterized. In contrast, the putative chromosomal integrons contained 1139 ORFs. Only 30.2% of these ORFs were assigned to a COG, and 20% were poorly characterized. Class V was the most abundant class in these integrons as well (33.4%). In general, chromosomal integrons contained more COGs in the metabolic and cellular processing and signalling classes.

Ecosystem energetics

The WCL are deep underground with ultraoligotrophic conditions (0.29 mg L^{-1} TOC; Hershey *et al*., 2018), making it unclear as to the drivers of ecosystem energetics. Nonetheless, there are a number of potential energy sources that could support autotrophic growth, including NO_2^-/NO_3^- at 2.4 mg L^{-1} and extensive Fe deposits, although dissolved Fe(II) is only 0.1 mg L^{-1} (Hershey *et al*., 2018). To identify any drivers of community energetics we examined the metagenome for the relative abundance of genes involved in chemolithoautotrophic metabolism (Anantharaman *et al*., 2016; Garber, 2020). Rather than nitrate- or iron-associated autotrophic pathways, the dominant genes were associated with Mn-oxidation, including two Mn oxidizing multicopper oxidases (*McoA* and *MnxG*), and a Mn oxidizing peroxidase (*MopA*; Figure 4.7; Tebo *et al*., 2005; Sujith & Bharathi, 2011). This was unexpected, although there are substantive Mn-oxide deposits throughout the cave. Another potential source for chemolithotrophic growth was sulfide, although sulfur deposits are extremely rare in Wind Cave. Hydrogenases were the most abundant oxidase observed; however, with such an extensive role in cellular functioning, we do not consider H_2 a likely source for ecosystem energetics (data not shown). While not significantly different from the other biomes, the TCA cycle was dominant among the central carbohydrate metabolic pathways, followed by acetogenesis from pyruvate and PEP-pyruvate anaplerotic reactions. The presence of 2-oxaloglutarate reductase (EC 1.2.7.3), fumarate reductase (EC 1.3.5.4), and ATP-citrate synthase (EC 2.3.3.8) indicate the presence of a functional reductive TCA (rTCA) cycle, which has been proposed as a mechanism for chemolithoautotrophic growth in the *Nitrospira* (Hugler & Sievert, 2011; Yu & Leadbetter, 2020).

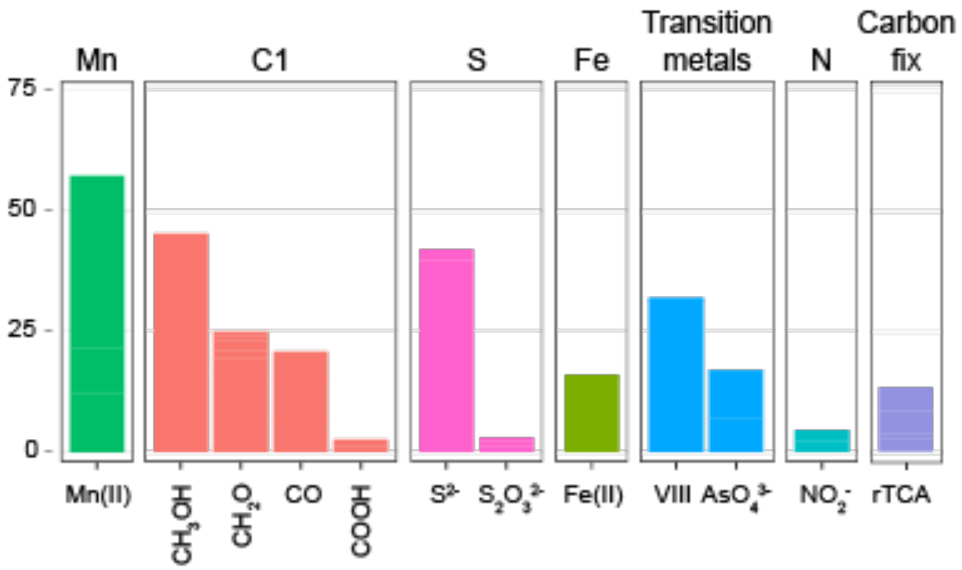

Figure 4.7. Average abundance of select genes indicating chemolithoautotrophic processes identified from assembled WICA contigs using Lithogenie. Chemolithoautotrophic processes are grouped by element and include (from left to right): manganese (Mn(II) oxidation), C1 compounds (methanol, formaldehyde, CO, and formate oxidation), sulfur (sulfur oxidation, thiosulfate oxidation), iron (Fe(II) oxidation), Transition metals (breakdown of halogenated compounds, arsenate reduction), nitrogen (nitrite oxidation), and carbon fixation (reductive TCA cycle).

The enrichment of genes in the metagenome involved in Mn chemolithotrophic growth was unexpected. Nonetheless, there are a number of features within the cave that are rich in Mn-oxides, including the presence of extensive Mn-oxide floor deposits and wall coatings (LaRock & Cunningham, 1995). While the analysis of these Mn-oxides to determine a biogenic origin is challenging, the cave also contains one of the world's largest collections of subaqueous helictites; the origin of these helictites is unclear, although they are only found in ancient lake basins and appear similar in structure to subaqueous biofilms being agitated in a water current (Figure 4.8A). The helictites are covered in a thick calcite coating, which would have been precipitated in the ancient lake basins, preserving the internal structure. Sectioning the helictites demonstrated a consistent brown, inner core comprised of Mn-oxides, as determined both using EDAX and mineral structure (Figure 4.8B-F). Notably the Mn-oxides were associated with numerous micro-fossils (Figure 4.8C).

Recent work by Yu and Leadbetter (2020) demonstrated that members of the Nitrospirota, which represent as high as 18% of the total community structure as determined by iTag sequencing, have the capacity to carry out autotrophic Mn(II) oxidation (manganese oxidizing bacteria; MOB; Parro & Moreno-Paz, 2003; Mi *et al*., 2011; Yu & Leadbetter, 2020). In order to identify whether such species were present in WCL, we used metaxa2 to extract 16S rRNA sequences from the metagenome for the dominant *Nitrospirota* found within the lakes. The results (Figure 4.1C) demonstrated that a large proportion of the unclassified species in

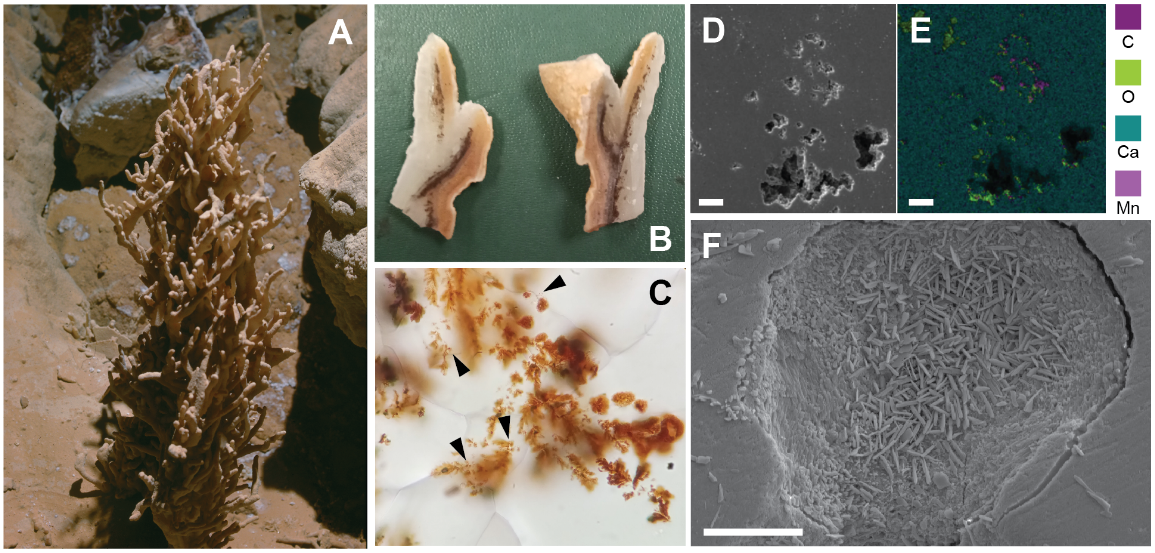

Figure 4.8. The chemical structure of helictite bushes within Wind Cave. A) A typical helictite structure, showing the vent-like structure. Approximate scale of image is 60 cm. B) The helictite bush had been heavily calcified in a lake basin; however, cutting the helictite in half demonstrated a brown-residue core. C) A thin-section demonstrating brown material embedded within the calcite matrix and associated microfossils (indicated with arrows; approximate width of image is 100 μm). D) SEM image of thin-section surface shown in C, with small pits where the row material fell out of the thin-section, creating small pits. E) EDAX mapping of these pits revealed an enrichment of Mn (scale bar is 50 μm). F) A close-up of a Mn-rich pit in the thin-section demonstrates minerals with a similar morphology to Mn-oxides (scale bar is 20 μm).

WCL were found within the *Candidatus* Class Trogloglia, which includes the Mn-oxidizing chemolithoautotroph *Candidatus* Manganitrophus noduliformans, and *Nitrospira marina*, a nitrite-oxidizing chemolithoautotroph (Yu & Leadbetter, 2020; Bayer *et al*., 2021). Notably, the closest phylotypes to our assembled 16S rRNA sequence within the Trogloglia were found in karst systems, which have similar chemistries and environmental conditions to WCL. While these data do not conclusively indicated Mn(II) and nitrite autotrophy as a source of community energetics it does suggest that Mn has been available in these lakes since at least the deposition of these helictites (>1 million years based on isotopic dating of associated calcite deposits; Bakalowicz *et al*., 1987; LaRock & Cunningham, 1995).

4.4 Discussion

Manganese is an abundant transition metal, and can be found in the environment as dissolved Mn(II) ions and microcrystalline Mn(III)/Mn(IV) oxides (MnO), the latter of which are produced through biotic and abiotic processes. Due to the slow rate of Mn(II) autooxidation, most biogenic MnO deposits (such as those seen at hydrothermal vents, metalliferous sediments, ferromanganese deposits and desert varnish; Tebo *et al*., 2005) are thought to be biogenically created by the reactive oxygen species ($O_2^{\cdot -}$ and H_2O_2) generated during microbial metabolism reactions(Tebo *et al*., 2005; Learman *et al*., 2011; Geszvain *et al*., 2012; Johnson *et al*., 2016). Nonetheless, Mn deposits are geochemically complex, and production likely involves the intertwined redox chemistry of Fe(II)/Fe(III) and Mn(II)/Mn(III)/Mn(IV).

Mn(II) can serve as a source of electrons for oxidation, but historically Mn(II)-oxidation was thought promote growth through the highly reactive Mn(III) oxides produced. These redox-active Mn(III) oxides can reduce recalcitrant organic polymers, such as humic acids and humus, releasing low molecular weight compounds that in turn support microbial growth (Nealson *et al*., 1988; Francis *et* *al*., 2001). In the WCL metagenome, this may be supported by the observed enrichment in genes that utilize C1-C2 carbon sources (Figures 4.3 and 4.7). The water entering WCL is old, with some coming from the ancient Archaean basalts at the center of the Black Hills, at >50,000 years, which would be highly oxidized organic carbon. Alternatively, the shales associated with the surrounding rock units may similarly provide a source of recalcitrant carbon. If Mn(III) provides a means of breaking down recalcitrant organic carbon, this may stimulate microbial growth without the need for Mn(II) chemolithoautotrophy.

There are three known Mn cytochrome oxidases, all of which have been identified within the WCL metagenome. *MnxG* and *MofA* are multicopper oxidases, with *MofA* demonstrating extracellular activity regulated by Fe(II); *MoxA* is a Ca^{2+}-binding heme peroxidase with activity that is upregulated by Ca^{2+} ions, which are abundant in the WCL water. The growth of the Mn(II) oxidizing *Erythrobacter* sp. str. SD-21, which expresses *moxA* along with the CO_2 fixation RubisCO pathway, is not affected by light; however, the Mn(II)-oxidation rate was dramatically reduced by light (Francis *et al*., 2001). This correlates with the presence of Mn-oxides in the environment, which are enriched in aphotic sediments (Nealson *et al*., 1988; Tebo *et al*., 2005). It has been thought that Mn(III)/Mn(IV) particulates are photoreduced by light in the open ocean, maintaining higher Mn(II) levels in the near-surface, although the role of light inhibition of MOB has not been investigated (Sunda & Huntsman, 1988). If this is indeed the case, cave environments have the potential to have enhanced Mn(II) oxidation due to their aphotic conditions.

Despite the established paradigm that environmental Mn(II) could not directly serve as an electron donor, Mn(II) chemolithoautotrophic growth was recently demonstrated in members of the Nitrospirota (Yu & Leadbetter, 2020). The autotrophic growth of the Nitrospirota used $MnCO_3$ as the Mn(II) source and O_2 as the terminal electron acceptor, and although growth rates were not affected by the presence of light, Mn(II) oxidation rates were not measured in its presence/absence (J. Leadbetter, personal communication, 2021). The Madison aquifer is housed within the Madison limestone, which is saturated with respect to CO_3^{2-}, with high levels of dissolved Mn(II); at ~1 μM, these levels are far above the amount in the open ocean, (~5 nM), where Mn(II)-oxidation reactions are an important component of ecosystem energetics (Sunda & Huntsman, 1988). Interestingly, the water within WCL contains 2 mg L^{-1} NO_3^{2-}. While the geochemistry of deep oceanic sediments suggests that NO_3^- serves as an important electron acceptor in the environment, no isolates capable of coupling Mn(II) oxidation/NO_3- reduction have been described, although the WCL ecosystem could be examined for the presence of such species (Tebo *et al.*, 2005).

Our metagenomic analyses found that reversable reactions within the rTCA cycle are the most abundant of the central carbohydrate metabolism pathways, along with ATP-dependent citrate lyase, an indicator gene for the presence of a functional rTCA cycle. Yu and Leadbetter (2020) similarly demonstrated that the rTCA was used to assimilate CO_2 for Mn(II)-oxidative autotrophic growth. Given this similarity, we assembled full-length 16S rRNA from the dominant species within the WCL metagenome (Figure 4.9). These data demonstrated that the dominant species were similarly within the Nitrospirota, including one phylotype (OSH-1082) that was taxonomically closely related to the chemolithotrophic species identified, *Candidatus* Manganitrophus noduliformas (Yu & Leadbetter, 2020). Together with the dramatic enrichment of Nitrospirota within the WCL lake samples, we believe this is indicative of Mn(II) oxidation driving primary productivity within the WCL ecosystem.

In addition to extant Mn(II)-oxidation, we believe that this microbial activity has been ongoing within the Wind Cave system for over 1 million years. Throughout Wind Cave there are also large formations with an obvious structural morphology called helictite bushes, which can approach 2 m in diameter (Palmer, 1981; LaRock & Cunningham, 1995; White *et al*., 2009; Palmer, 2017). There are hundreds of these bushes throughout the cave, but they are all found along a single fault that intersects the lakes area and WCL (LaRock & Cunningham, 1995). The mechanisms of formation for these helictites remain a mystery; however, they are only found below the remnant edges of ancient lake basins and are often 'sprinkled' with calcite rafts that form at the water/air interface of lakes (Palmer, 1981; Davis, 1991; LaRock & Cunningham, 1995). Thin-sections of these bushes demonstrates a MnO core associated with significant microfossils (LaRock & Cunningham, 1995). The Madison limestone overlays the Deadwood Formation, which contains significant phyllosilicates that release Mn(II) upon dissolution (DeWitt *et al*., 1986). It is our hypothesis that the fault running through Wind Cave provides a preferential flow path for water in the Madison aquifer, which picks up Mn(II) from the Deadwood Formation. As Madison water intersects into Wind Cave, it provides a source of Mn for Mn(II) oxidation and growth, generating hydrothermal-like biofilms at the infeeding source of the Mn. Over time these MnO biofilms are fossilized by $CaCO_3$, preserving these structures as helictite bushes (LaRock & Cunningham, 1995). The calcite rafts that have caught within these helictite bushes when the lake basins dried, and dated at >1 million years in age.

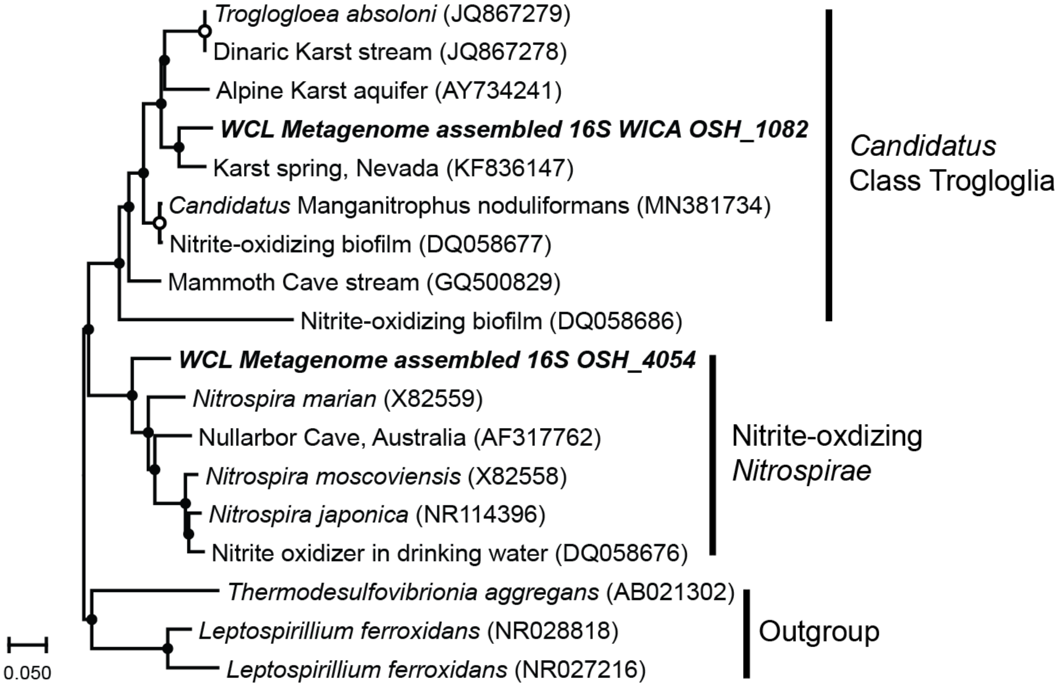
Troglogloea absoloni (JQ867279)
Dinaric Karst stream (JQ867278)
Alpine Karst aquifer (AY734241)
WCL Metagenome assembled 16S WICA OSH_1082
Karst spring, Nevada (KF836147)
Candidatus Manganitrophus noduliformans (MN381734)
Nitrite-oxidizing biofilm (DQ058677)
Mammoth Cave stream (GQ500829)
Nitrite-oxidizing biofilm (DQ058686)
WCL Metagenome assembled 16S OSH_4054
Nitrospira marian (X82559)
Nullarbor Cave, Australia (AF317762)
Nitrospira moscoviensis (X82558)
Nitrospira japonica (NR114396)
Nitrite oxidizer in drinking water (DQ058676)
Thermodesulfovibrionia aggregans (AB021302)
Leptospirillium ferroxidans (NR028818)
Leptospirillium ferroxidans (NR027216)
Candidatus Class Trogloglia
Nitrite-oxdizing Nitrospirae
Outgroup
0.050

Figure 4.9. Phylogenetic analysis of *Nitrospirae* 16S rRNA genes extracted from assembled metagenomic contigs. Phylogenetic trees were generated using 1,236 positions with the NJ and ML tree-building algorithms. The trees generated similar topologies and the optimal NJ tree computed using the Tamura 3-parameter evolutionary model with a gamma distribution parameter of 0.26 (closed circles indicate >70% bootstrap support and open circles indicating >50% support from 1,000 generated trees).

Not all the cells within the WCL are autotrophic species, and there are other adaptations to nutrient limitation-- we also saw an enrichment in amino acid and co-factor transport and a reduction in cell size. Amino acid transport and co-factor assimilation are commonly associated with ultraoligotrophy, and increased scavenging mechanisms for amino acids can reduce the number of inorganic elements that play a critical role in enzymatic activity, such as amino acid synthesis – it therefore becomes energetically more advantageous to import amino acids rather than create the enzymes using unavailable co-factors. Similarly, the increase in co-factor transport is aimed to relieve this nutrient limitation. The other genes seen in higher abundance, such as DNA and RNA metabolism may reflect carbon turnover of microbial polymers from other microbial species within the environment. A significant population of WCL appears to comprise ultra-small cells (<0.22 μm), which we examined using differential-filtration (Beam *et al*., 2020). The size-differentiated metagenomes demonstrated a significant enrichment in members of the *Nitrospirota* (and in particular the Mn(II) oxidizing *Leptospirillia*), and ultra-small *Patescibacteria* and *Nanoarchaeota*, phyla that are known to contain ultrasmall cells. SCWGS confirmed these identifications, while also revealing that the *Patescibacteria* and *Nanoarchaeota* populations lack terminal electron transport chains (Brown *et al*., 2015; Castelle *et al*., 2017; Vigneron *et al*., 2020) suggesting both structural and energetic traits that may provide an adaptive advantage for these species in the lakes (Beam *et al*., 2020).

Our metagenomic analyses also revealed other potentially important adaptations, including a remarkable potential for horizontal gene transfer (Mazel, 2006). In addition to regular mobile elements (such as transposons, phage), the WCL metagenome contained ~3-fold more integrons-related sequences than comparative ecosystems (Mazel, 2006; Cambray *et al*., 2010; Gillings, 2014; Buongermino Pereira *et al*., 2020). Integrons are normally associated with the proliferation of antibiotic resistance cassettes in clinical settings, but are now recognized as a mechanism that can capture and shuffle gene cassettes with many potential functions within and between microbial chromosomes (Boucher *et al*., 2007; Gillings, 2014; Abella *et al*., 2015; Vit *et al*., 2021). The integron mechanism is such that newly inserted genes are placed adjacent to a promoter and immediately expressed, while older genes are pushed further from the promoter (Mazel, 2006; Boucher *et al*., 2007; Cambray *et al*., 2010; Gillings, 2014; Buongermino Pereira *et al*., 2020). Because gene expression is negatively correlated with distance from the promoter, new gene arrangements are quickly subjected to selective forces (Mazel, 2006; Boucher *et al*., 2007; Gillings, 2014; Engelstadter *et al*., 2016). A number of gene cassettes encoding toxin-antitoxin (TA) systems and restriction modification systems (RMS) were also present in the integron arrays in WCL. These addiction modules typically contain their own promoters and can prevent the loss of older gene cassettes in the integron array that were advantageous under previous conditions (Mazel, 2006; Cambray *et al*., 2010). This mechanism may provide a reservoir of genes that can be reshuffled, adding genetic plasticity to the microbial community that is essential for quickly responding and adapting to environmental stressors (Starikova *et al*., 2012; Engelstadter *et al*., 2016).

In WCL, the integrons were associated with some metabolic genes, including those used in carbon turnover (lipid and carbohydrate metabolism, protein turnover and amino acid utilization), as well as a substantial number of gene cassettes containing genes with no known homolog and potentially novel functions (Stokes & Hall, 1989). We believe that these data hint at an evolutionary adaptation to survival under extreme nutrient-limitation and provides a rapid mechanism for the dissemination of new adaptive traits under nutrient stress (Stokes & Hall, 1989; Labbate *et al*., 2009; Raz & Tannenbaum, 2010; Ghaly *et al*., 2020).

Carbon and energy limitations are a significant and distinct challenge in the WICA lakes when compared to other aphotic environments that are part of more open systems; for example, the deep ocean typically receives some amount of nutrient input from sinking particulate organic carbon, and groundwater can have an influx of nutrients from shallow, permeable regions and at recharge zones (Schütz *et al*., 2010; Glud *et al*., 2013). The isolation of the WICA lakes from surface processes and the age of the incoming Madison groundwater makes influxes of carbon unlikely. Instead, our data suggests that C in the system may be dominated by autochthonous carbon fixed by autotrophic growth. Taken together, primary production within the WCL could be driven by Mn(II) oxidation, making it the first such community described .

CHAPTER V

CONCLUSIONS

Karst aquifers provide up to 25% of the world's freshwater. Understanding the role microbial communities play in biogeochemical cycles in karst is particularly important in understanding the factors that maintain clean drinking water. Advancements in sequencing technology and computational analysis have greatly improved our ability to examine microbial interactions and processes in such environments, without the need for cultivation; however, accessing these environments can be problematic, and obtaining adequate samples without the introduction of contamination (chemical or microbial) from the surface is even more challenging. Wind Cave is one of the longest and oldest caves in the world and at a depth of 125 m intersects the Madison aquifer, creating a series of lakes that provide a unique portal to study the microbiology of this important karst aquifer.

My initial work in the Wind Cave lakes (WCL) compared the microbial communities from the Madison aquifer accessed through the cave with a nearby surface well, thought to supply water from the Madison aquifer. The microbial communities sampled between the two sites were distinct and suggested that chemistry from the overlying Minnelusa formation changed the diversity that could be sampled via the well (Chapter 3). We also found that the community in WCL, despite having one of the lowest measured cell numbers for water, was surprisingly diverse. The number of microbes that remained unassigned to known phyla in our analysis made understanding the geochemical interactions that supported microbial subsistence challenging end required a deeper metagenomic analysis.

The low cell counts in WCL made sampling difficult, requiring >1,000 L of water to be filtered to collect enough microbial biomass to extract adequate DNA for metagenomic sequencing; a problem that was further compounded by the fact that the WCL microbial community appeared to contain numerous cells capable of passing through a standard 0.2 μm filter. Simply switching to a 0.1 μm filter was not a solution to this problem—clogging and flow restriction increased the filtration time considerably, making sample collection unfeasible with the limited battery power we were able to carry into the cave. Ultimately, this led us to optimize filtration using a tangential flow filter with a 0.45 nm pore size (Chapter 2).

Using tangential flow filtration, I discovered that the Wind Cave lakes contained a microbial community that was more complex than our initial study suggested, with a significant population (~18%) of ultrasmall cells (<0.2 μm diameter; Chapter 4). These included representatives from poorly understood phyla, such as those found in the Candidate Phyla Radiation and DPANN groups. This small cell fraction was also enriched in members of the Nitrospirota, which appear to play a central role in primary production in WCL via chemolithotrophic

Mn(II)-oxidation. Such findings suggest that the WCL could be an important site to study the Mn biogeochemical cycle in more detail, such as the role of Mn(III) in the breakdown of recalcitrant organic compounds. My data also suggested that an unusual level of integrons within the metagenome provides a means for the microbial community to quickly adapt to environmental changes under such extreme conditions. These mechanisms of genetic plasticity may work in conjunction with other adaptations typically observed in oligotrophic environments (scavenging mechanisms, reduced cell size) to support the broad diversity that is observed in the WCL ecosystem.

While further investigation using cultivated representatives and metatranscriptomics may be necessary to determine active processes in the Madison aquifer, these data suggest that the WCL may represent the first described microbial community supported primarily by a manganese biogeochemical cycle. The literature on the potential for Mn oxides to serve as a biomarker is limited, even though the geologic record on Earth contains 2.2 Gya sedimentary Mn oxide deposits, such as the Kalahari manganese field in South Africa. A variety of Mn oxide deposits have been found throughout Wind Cave, with extensive deposits also in the nearby Jewel Cave, which may be directly related to past microbial Mn(II)-oxidation. Mn-rich rocks were also recently discovered in the Gale Crater on Mars, and this Mn appears to have accumulated in groundwater before deposition, reaching up to 35 wt % in the exposed rock. As such, my work should have broad applications beyond understanding the microbiology of Wind Cave.

Finally, the impetus for this work is rooted in conservation. Due to the growing population and development of the Black Hills region, water demand on the Madison aquifer has increased. New wells for the Southern Black Hills and Fall Rivers water districts have been recently permitted, with a combined requested draw of 1.42 billion L year^{-1}. Together these wells would draw water at rates that would greatly exceed the local recharge rates of the Madison aquifer and would result in a drop in the potentiometric surface at Wind Cave in excess of 10 meters, exceeding the current depth of the WCL lakes (~7 m). The resulting loss of the lakes would not only have significant negative effects on the WCL ecosystem but would also result in the loss of direct access to this important sample site. This research provides strong evidence that the conservation of this karst aquifer system is vital for assessing the true extent of microbial diversity and understanding the processes that enable life to exist under some of the most extreme nutritional limitations.

CHAPTER Vi

EXPERIMENTAL METHODS

3.1 Sample sites and sampling

Given their depth (-122 m) and distance (>3 km) from the entrance, the Wind Cave lakes (WCL) are subject to constant temperature (13.7°C air temperature; 13.8°C water temperature; Back, 2011), with no variation in relative humidity (99.9% measured using a RH300 Digital Psychrometer, Extech Instruments, Waltham, MA). Park Well #2 (PW2), located within Wind Cave National Park (WCNP), was drilled as a source of drinking water, reaching a depth of -208 m and drawing water from the Madison aquifer. Streeter Well (STR) is a private well that is similarly drilled down into the Madison Formation, to a depth of -283 m (Long & Valder, 2011; Figure 3.2). Before sampling, resident water was removed from the wells by flushing three well volumes (>2,000 L) of water using hydrostatic pressure (Hose & Lategan, 2012; Smith *et al*., 2012). Beaver Creek Spring (BCS), is on private land to the south of WCNP was also sampled as part of this study (Long *et al*., 2012).

Given the remoteness of the cave sample site and the narrow size of cave passages traversed, all equipment had to be battery powered with its largest dimension no greater than the narrowest passage height (20 cm; Figure 31A). We therefore collected cells via filtration through a Nalgene disposable 0.2 μm filter unit using a SP200 variable speed peristaltic pump (Global Water Instrumentation, Gold River, CA) with a pump rate of 1 L min^{-1} (Hershey *et al*., 2018). To filter at the wells, samples were collected using the same Nalgene filter set-up, with well water flowing into a 18 L, sterile, acid-washed bucket in which the filter unit was floated. To collect cells, the filter membranes were cut out of the filtration unit at the sample site using a sterile scalpel and stored in 70% alcohol for transport. Samples were stored at -80°C in the lab until processing.

3.2 Chemistry

Inorganic water chemistry from each of the sample sites has been compiled from raw data provided by the National Park Service (USA) and Back (2011; Table 3.2). Water pH was measured in the field using an Accumet AP61 portable pH meter (Fisher Scientific, Pittsburg, PA, USA). Total organic carbon (TOC) analysis was carried out by WATERS laboratories (Western Kentucky University, Kentucky, USA) using high temperature combustion method (SM5310B) in a Shimadzu total organic carbon analyzer TOC-V series. Due to the very low biomass observed in the lakes we decided to also measure (1-3)-β-5-Glucan, a common polysaccharide of both Gram negative and positive bacteria, to assess total microbial biomass (Duenas *et al*., 2003). Quantification was carried out via a chromogenic method using the Glucatell reagent in a BioTek ELx808 microplate reader, with measurements averaged from samples collected over a two-year period.

3.3 Cell Counting

During four separate sampling campaigns (2009-2015), 10 mL samples of water were collected using aseptic technique and immediately preserved by the addition of 1 mL 20% paraformaldehyde, to a final concentration of 2%. Samples were kept on ice for transportation and were stored at 4°C until processing (generally <1 week). For cell enumeration, samples were filtered onto a 25 mm diameter, 0.2 μm membrane filter (Anodisc or Cyclopore; Whatman, Piscataway, NJ) and stained with SYBR Green I (unless stated otherwise, all chemicals were obtained from Sigma-Aldrich, St. Louis, MO). Total cell counts were carried out using epifluorescence microscopy at 1000X, using either a Leica DMZ2500 (Leica, Wetzlar, Germany), or Olympus BX53 fluorescent microscope (Olympus America Inc., Center Valley, PA). Blank counts were determined at the beginning and the end of each working day, and the average blank value (52 cells mL^{-1}) was subtracted from each cell count. Between 50 and 200 fields of view were inspected in order to achieve a total count of at least 20 separate samples. To reduce cell aggregation and more accurately count cell numbers in the samples we used a number of techniques, including: (a) filtering the untreated sample straight onto the membrane, (b) immersing the sample vial into a sonication bath (Bandelin Sonorex, 10 min at 640 W) prior to filtration, (c) adding a detergent solution (100 mM disodium EDTA dihydrate, 100 mM sodium pyrophosphate decahydrate, 1 % vol/vol Tween 80) and methanol to the sample and vortexing it for 30 minutes to dissolve extracellular polymers; and (d) a combination of methods (b) and (c) (Kallmeyer *et al*., 2008). For detection of eukaryotic species, 10 L volumes of lake water were collected in triplicate and filtered onto a 25 mm diameter 0.2 µm or 8.0 µm filters, before staining with calcofluor white M2R and DAPI fluorescent stains (Sigma-Aldrich, St. Louis, MO) following the manufacturers recommended protocol. Nucleated cells were counted for 50 fields of view on an epifluorescence microscope at 400X, as described. The absence of observable eukaryotic organisms was confirmed by scanning the membrane at 100X magnification.

For domain-specific cell counts, fluorescence in situ hybridization (FISH) was performed using the bacterial-specific primers EUB338, EUB338II and EUB338III labeled with Cy3, and the archaeal-specific primers CREN499 and ARC915 labeled with Cy5. Details on the nucleotide sequence, specificities and formamide concentrations for the primers were used as described in probeBase (Loy *et al*., 2007). Briefly, 10 mL of paraformaldehyde-fixed water was filtered onto a 25 mm 0.2 µm Anodisc filter (GE Healthcare Bio-Sciences, Pittsburgh, PA) and stained with 50 ng mL^{-1} of each of the fluorescent probes as previously described (Daims *et al*., 1999). The filters were then enumerated on the Olympus BX53 fluorescent microscope. Samples were examined at a 1000X magnification with final counts being estimated from the average of 100 fields-of-view.

3.4 Molecular Techniques

All DNA protocols were carried out in a laminar-flow hood using aerosol resistant pipette tips to reduce the likelihood of contamination, along with preparation controls (PCTL). These preparation controls were created by subjecting all the sampling equipment to experimental processing in the absence of sample, including assembly and processing within the lab and at the field site. The preparation controls were extracted in parallel with the samples throughout the DNA extraction, PCR amplified and pooled prior to pyrosequencing. Cells were dislodged from filters via vortexing in 10 mL of Buffer A [200 mM tris(hydroxymethyl)aminomethane pH 8.0, 50 mM ethylenediaminetetraacetic acid (EDTA), and 200 mM NaCl] with 0.3% wt/vol sodium dodecyl sulfate (SDS) as a surfactant, followed by centrifugation at 13,000 x *g* (Barton *et al*., 2006). Genomic DNA was extracted from the pelleted cells using a low biomass bead beating protocol (Barton *et al*., 2006).

For pyrosequencing the DNA was PCR amplified using the universal primers 515F and 806R (Walters *et al*., 2011) in three separate reactions. Each sample was labeled with a unique barcode (Kozich *et al*., 2013) and the amplified DNA was sequenced using a Roche/454 pyrosequencer at University of Kentucky

Advanced Genetic Technologies Center (UK-AGTC; http://www.uky.edu/Centers/AGTC). The resulting read data were analyzed using the QIIME software (Caporaso *et al*., 2010b) with nucleotide sequences from each site separated using the PCR-encoded barcode. Poor quality reads were filtered out using the QIIME 454 denoising pipeline (Reeder & Knight, 2010), and chimeric sequences were identified and removed using UCHIME (Edgar *et al*., 2011) *de novo* and reference-based detection. OTUs were identified using UCLUST (Edgar, 2010) and classified using the SILVA QIIME 16S database (SILVA 119; Quast *et al*., 2013). Sequences were aligned to the SILVA 119 core alignment using the QIIME PyNAST alignment function and used to generate an approximate maximum likelihood phylogenetic tree using the default FastTree parameters (Caporaso *et al*., 2010a; Price *et al*., 2010). Raw sequences for this study were uploaded to the NCBI sequence read archive (SRA) with the acquisition number SRP147561. Comparative 454-sequencing data sets from similar oligotrophic aquatic environments were obtained from the NCBI SRA, including accession numbers SRP010407, SRP058014, SRP021556, and ERP020663.

For alpha-diversity metrics, rarefied OTU tables were generated from sequence data in QIIME using a step size of 100 sequences and an even sampling depth of 3,900 sequences. Species richness for each sample was calculated using the Chao 1 non-parametric estimator over 10 randomized iterations of sampling, with the median values used for a best fit line generated in R (Chao, 1984; Wickham, 2007, 2011; Venables & Ripley, 2013). The species richness of each sample was calculated using average tables (including standard error) for Simpson's Reciprocal Index generated in QIIME (Simpson, 1949; Caporaso *et al*., 2010b). Normalized OTU tables were generated for beta-diversity analyses utilizing weighted- and unweighted-Unifrac metrics for community comparisons and Principle Coordinate Analyses (PCoA) in R (Lozupone & Knight, 2005; McMurdie & Holmes, 2013).

4.1 Sample sites and sampling

At the Wind Cave Lakes (WCL) site, the air temperature is a constant 13.7°C with a water temperature of 13.8°C (Back, 2011). There is no variation in relative humidity at 99.9% measured using a RH300 Digital Psychrometer (Extech Instruments, Waltham, MA). Samples were collected in September 2017 and October 2018 using a SP200 variable speed peristaltic pump (Global Water Instrumentation, Gold River, CA), modified for reduced weight, and the Large Volume Concentration (LVC) kit (Innovaprep, Drexel, MO). Lake water was filtered at a rate of approximately 1 L min^{-1} for approximately 24 hours. Cells were collected from the LVC filter at the sampling location using the provided elution buffer canister (0.075% Tween 20/PBS), into a sterile 500 mL polycarbonate bottle. To ensure that all cells were released from the filter, multiple elution canisters were used (until the collected fluid came out clear), yielding ~400 mL of concentrated cells that were preserved with ethanol to a final concentration of 70% prior to transport out of the cave and stored at 4°C for transport. Preparation controls (PCTL) were also created for all samples by handling all collection filters and equipment at the field site and throughout laboratory processing in the absence of sample.

4.2 DNA extraction

All DNA protocols were carried out in a laminar-flow hood using aerosol resistant pipette tips to reduce the likelihood of contamination, along with preparation controls (PCTL). The unfractionated concentrated cell samples and the ultrasmall filtrate samples were further concentrated via centrifugation at 25,000 xg, after which the alcohol was removed, and the cell pellets were washed with PBS. Cells collected on membrane filters were dislodged via vortexing and then centrifuged to form a pellet (Barton *et al*., 2006). The pellets were resuspended in 500 μl 2x buffer A [200 mM NaCl, 200 mM Tris-HCl (pH 8.3), 20 mM EDTA], treated with 3 mg mL^{-1} lysozyme, incubated at 37°C for 30 min, after which 1.2 mg mL^{-1} Proteinase K and 10 μl 20% sodium dodecyl sulfate were added (Barton *et al*., 2006). The solution was incubated at 50°C for 30 min, vortexing occasionally. Genomic DNA was extracted using phenol-chloroform-isoamyl alcohol as previously described (Barton *et al*., 2006) and quantified using the Qubit 2.0 Fluorometer and dsDNA High Sensitivity Assay Kit (Life Technologies, Waltham, MA); The average yield of total genomic DNA was ~275 ng from each sample collection trip.

4.3 16S rRNA Illumina Sequencing and read processing

For each sample and PCTL, the bacterial and archaeal 16S rRNA gene V3/V4 variable region was amplified using the universal primers 515F (GTGCCAGCMGCCGC GGTAA) and 806R (GGACTACHVGGGTWTCTAAT) containing Illumina iTag sequences for multiplexed sequencing runs (Caporaso *et al*., 2011). Each PCR reaction contained 25 µL Taq Master mix (New England Biolabs Inc, Ipswich, MA), 100 µM of the indexed primers, 4 ng of template DNA, and were brought to a final volume of 50 µL using molecular grade water. The PCR was carried out as described in (Caporaso *et al*., 2011). Amplified products were visualized using gel electrophoresis, excised with a sterile scalpel, and purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research, Irvine, CA). The concentration of the PCR product was determined using the Qubit dsDNA High Sensitivity Assay Kit. Amplicons were sequenced using the Illumina Miseq (Illumina, Inc., San Diego, CA) with 2x250 read length at the University of Kentucky Genomics Core Laboratory (UK Healthcare, https://ukhealthcare.uky.edu/genomics-core-laboratory). Read denoising, quality filtering, and chimera checking using DADA2 and statistical analyses were carried out in QIIME2 (McKinney, 2010; Pedregosa *et al*., 2011; Callahan *et al*., 2016;

Bolyen *et al*., 2019). The remaining reads were then classified using a Naïve Bayes classifier pre-trained on the 515F/806R region from 99% clustered OTUS in the SILVA 138 reference database (Pruesse *et al*., 2007; Quast *et al*., 2012; Yilmaz *et al*., 2014; Bokulich *et al*., 2018; Robeson *et al*., 2020).

4.4 Metagenomic Sequencing, Read Processing, and Assembly

Potential upstream contaminants were removed from the extracted genomic DNA using 2.2 volumes of AMPure XP beads (Beckman Coulter A63881) per the manufacturer recommendations, followed by 3 washes with 80% ethanol on a tube magnet (Eppendorf 12321D). After drying, the genetic material was eluted off the AMPure beads using ultrapure water (Gibco 15230001). Each sample was then dual-indexed with unique 8bp indices using the Nextera XT DNA Library Preparation Kit (Illumina FC-131-1096). Following quality control analysis by the Qubit 4.0 fluorometer (Q33238) and Agilent 4200 Tape Station System, the amplified DNA library was sequenced (Paired End Sequencing, 2x 308 Cycles) on an Illumina MiSeq using a MiSeq 600 Cycle Reagent Kit v3 (Illumina MS-102-3003) at Koch Institute BioMicro Center.

Quality trimming and filtering were done using several publicly available software packages. First, reads were trimmed for quality using Trimmomatic version 0.36 with the following parameters: LEADING:3 TRAILING:3 HEADCROP:5 SLIDINGWINDOW:4:15 MINLEN:36 (Bolger *et al*., 2014). Reads were then processed using a combination of bbtools version 38.0 scripts similar to the rqcfilter used by the Joint Genome Institute (Bushnell, 2018). Briefly, reads were further trimmed for quality and length, followed by the removal of reads containing adapter sequences, phiX sequences, and short sequences known to be Illumina sequencing artifacts. Then, reads that mapped to masked contaminant sequences (human, dog, cat, and mouse) were removed using bbmap (Bushnell *et al*., 2019).

Sequences present in the PCTL controls were removed from the WICA samples sequences to account for contamination that may have been introduced during the sampling process. To do this, regions of low entropy and repeating sequences were masked, and the sample reads were mapped to the masked control references. Reads that mapped to the controls were removed, and the remaining unmapped reads were corrected using a bbtools bloom filter error correction tool (Heo *et al*., 2014). Read pairs were then merged to reduce the computational resources needed for assembly. Reads sets were assembled *de novo* with metaSPAdes version 3.13.1 with kmer sizes 33,44,77,99,127, with no additional error correction during assembly (Bankevich *et al*., 2012). The quality of the resulting assemblies was evaluated with QUAST version 5.0.22 (Gurevich *et al*., 2013).

4.5 Metagenomic Annotation

Metaxa2 version 2.2.3 was used to identify and extract partial 16S rRNA sequences from the assembled contigs (Bengtsson-Palme *et al*., 2015). For comparison to the 16S rRNA amplicon results above, classification of the extracted reads was carried out in QIIME2 using a full-length SILVA 138 16S rRNA classifier (Pruesse *et al*., 2007; Quast *et al*., 2012; Yilmaz *et al*., 2014; Bokulich *et al*., 2018; Robeson *et al*., 2020). Open reading frames (ORFs) were predicted from the assembled contigs using Prodigal version 2.6.3 with -p meta and defaults for all other parameters (Hyatt *et al*., 2010). The resulting ORF nucleotide sequences were searched against the SUPER-FOCUS version 0.34 (Silva *et al*., 2016) DB_100 reference dataset, which contains curated protein sequences from the SEED database (Overbeek *et al*., 2005) clustered at 100% sequence identify for highest functional resolution. The DIAMOND aligner option was used with the SUPER-FOCUS default parameters (minimum sequence identity 60%, minimum alignment length 15 amino acids, e-value 0.00001), resulting in a metagenomic profile organized into the SEED subsystem hierarchy (where Subsystem Level 1 is the most general classification and Function Level is the most specific classification). Annotation using the Clusters of Orthologous groups were performed using WebMGA (http://weizhong-lab.ucsd.edu/webMGA/server/cog/) with -e-value 0.001 parameter (Wu *et al*., 2011). Specific genes involved in the chemolithoautotrophic microbial metabolism (including C, H, O, N, S, Mn and others) were identified performing an hmmsearch against a set of curated HMM profiles with LithoGenie using contig_source meta and defaults for all other parameters (Anantharaman *et al*., 2016; Garber, 2020).

Integron-associated genes were identified by locating and validating *attC* sites in the assembled contigs, then searching for ORFs located less than 500 bases from the *attC* site using the default parameters in MIG-finder (Buongermino Pereira *et al*., 2020). Antibiotic resistance genes associated with integrons were identified using the Resistance Gene Identifier (RGI) software using the Comprehensive Antibiotic Resistance Database (CARD) version 3.1.2, with parameters -include_loose and -low_quality enabled (Alcock *et al*., 2020). The integron-associated genes were also processed using SUPER-FOCUS for SEED functional annotation and WebMGA for COG annotation as previously described.

4.6 Comparative Statistical analysis

The WICA metagenome was compared to 22 publicly available metagenomes obtained from the Integrated Microbial Genomes and Microbiomes (IMG/M) database, as summarized in Tables A.2 - A.7 (Chen *et al*., 2019). This dataset represents 5 different biomes: caves, groundwater (GW), freshwater lakes (FW), the deep ocean subsurface (DeepOSub), and deep ocean trenches (Trench). The FASTA files containing the nucleotide sequences of predicted ORFs were processed with SUPER-FOCUS and WebMGA using the same parameters as WICA.

All statistical analyses were performed in RStudio version 1.4.1106 (R Core Team, 2020). A metabolic profile for each comparative sample was made using SEED Function level assignments, with features present at an abundance less than 0.001% in all samples removed (Gloor, 2016). The datasets were converted from feature counts to ratios using a center log-ratio (clr) transformation using the CoDaSeq version 0.99.6 package (Gloor *et al*., 2017). Distance was calculated from the Euclidean distance of the clr transformation. The resulting Aitchison distance matrix was used to generate a principal component analysis (PCA) plot to visualize variance between sample groups and generate a dendrogram of individual samples (Gloor *et al*., 2017). The ALDEx2 package version 1.23.2 (Gloor, 2015) was used for differential abundance analysis to identify any genes for metabolic function that were over- or under-represented (|effect size| > 1.5) in the WICA metagenome compared to the other biomes, including other groundwater sites, the deep oceanic subsurface, and freshwater lakes examined. We also included other sampled cave environments to rule out the general influence of cave physiochemical conditions. A generalized linear model was used to compare all biome groups with aldex.glm and aldex.glm.effect; each biome was compared to WICA separately with the Welch's t, Wilcoxon rank test and effect-size estimations performed with the aldex wrapper function. Plots were generated in ggplot2 (Wickham, 2009) in addition to CoDaSeq and ALDEx2 plotting functions.

4.7 Thin Section preparation

For microbiological observations, double polished thin-sections were prepared from lengths along the longitudinal axis of speleothem sample EMP1 to a thickness approaching 30 μm, or until crystal grain loss began to occur from the action of grinding. Specimens were trimmed to size with a Buehler Isomet low speed saw (Lake Bluff, Illinois, United States) to fit the dimensions of petrography glass slides and to preserve excess material for SEM analysis. Trimmed and polished specimens were dried on a hotplate for 24 hours to minimize air pocket formation, before being mounted on Hillquist 26 mm x 46 mm thin section glass slides using Hillquist Thin Section Epoxy A-B resin and cured at 79°C for 30 minutes. A Hillquist Thin Section Machine (Hudson, New Hampshire, United States) was used to grind sections to within a thickness of 200 μm. Silicon carbide powder (600 and 1000 grit) and aluminum oxide powder (1200 grit) were then used sequentially to further grind the samples by hand to the desired final thickness between 30-50 μm. The surface was brought to a polish with Buehler Micropolish 0.3 μm Alpha Alumina powder on a Buehler Low Speed polishing wheel. Between each grit size, specimens were sonicated to remove excess grit powder. Bacterial cells and manganese-associated microstructures within the thin sections were observed using an Olympus BX53 light microscope (Shinjuku City, Tokyo, Japan) with QImaging Exi Aqua camera. Images were captured using Q Capture Pro 7 software.

4.8 SEM/EDX Analysis

Thin-sections for SEM/EDAX analysis did not undergo extensive grinding. Instead, samples were coated with a 3 nm layer of platinum (which can be penetrated via EDAX) using a Leica EM ACE600 high vacuum sputter coater (Wetzlar, Germany). SEM imaging and EDAX mapping was carried out using a Tescan Lyra 3 XMU scanning electron microscope (Brno, Czech Republic) equipped with EDAX module. Elemental mapping was performed using EDAX TEAM EDS software (Mahwah, New Jersey, United States). Initial X-ray counts per second were low during shorter duration scans and detection of elements besides C, O and Ca was minimal. Scanning was subsequently carried out for 1 hour at 15KV under high vacuum with a working distance of 10 mm to obtain the EDX data presented for manganese detection.

REFERENCES

Abed, R. M., Safi, N. M., Koster, J., de Beer, D., El-Nahhal, Y., Rullkotter, J., & Garcia-Pichel, F. (2002). Microbial diversity of a heavily polluted microbial mat and its community changes following degradation of petroleum compounds. *Appl Environ Microbiol*, *68*(4), 1674-1683. https://doi.org/10.1128/AEM.68.4.1674-1683.2002

Abella, J., Bielen, A., Huang, L., Delmont, T. O., Vujaklija, D., Duran, R., & Cagnon, C. (2015). Integron diversity in marine environments. *Environ Sci Pollut Res Int*, *22*(20), 15360-15369. https://doi.org/10.1007/s11356-015-5085-3

Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., & Polz, M. F. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol*, *71*(12), 8966-8969. https://doi.org/10.1128/AEM.71.12.8966-8969.2005

Adey, A., Morrison, H. G., Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., & Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, *11*(12), 1-17. https://doi.org/10.1186/gb-2010-11-12-r119

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, *31*(6), 533-538.

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., Huynh, W., Nguyen, A. V., Cheng, A. A., Liu, S., Min, S. Y., Miroshnichenko, A., Tran, H. K., Werfalli, R. E., Nasir, J. A., Oloni, M., Speicher, D. J., Florescu, A., Singh, B., Faltyn, M., Hernandez-Koutoucheva, A., Sharma, A. N., Bordeleau, E., Pawlowski, A. C., Zubyk,

H. L., Dooley, D., Griffiths, E., Maguire, F., Winsor, G. L., Beiko, R. G., Brinkman, F. S. L., Hsiao, W. W. L., Domselaar, G. V., & McArthur, A. G. (2020). CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res*, *48*(D1), D517-D525. https://doi.org/10.1093/nar/gkz935

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic Identification and in-Situ Detection of Individual Microbial-Cells without Cultivation. *Microbiological reviews*, *59*(1), 143-169. https://doi.org/Doi 10.1128/Mmbr.59.1.143-169.1995

Amann, R. I., Snaidr, J., Wagner, M., Ludwig, W., & Schleifer, K. H. (1996). *In situ* visualization of high genetic diversity in a natural community. *Journal Bacteriology*, *178*, 3496-3500.

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. *Nat Commun*, *7*, 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 phylogenetic analysis of a bacterial community in Sulphur River, Parker Cave, Kentucky. *Am. Mineralogist.*, *83*, 1583-1592.

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *N Biotechnol*, *25*(4), 195-203. https://doi.org/10.1016/j.nbt.2008.12.009

Apprill, A., McNally, S., Parsons, R., & Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology*, *75*(2), 129-137. https://doi.org/10.3354/ame01753

Ashbolt, N. J., Grabow, W. O., & Snozzi, M. (2001). Indicators of microbial water quality. *Water quality: Guidelines, standards and health*, 289-316.

Atlas, R. M., Horowitz, A., Krichevsky, M., & Bej, A. K. (1991). Response of microbial populations to environmental disturbance. *Microbial Ecology*, *22*(1), 249-256.

Back, J. (2011). Geochemical investigation of the Madison Aquifer, Wind Cave National Park, South Dakota. National Park Service. *Natural Resources Program Center, Fort Collins, Co*.

Bakalowicz, M. J., Ford, D., Miller, T., Palmer, A., & Palmer, M. (1987). Thermal genesis of dissolution caves in the Black Hills, South Dakota. *Geological Society of America Bulletin*, *99*(6), 729-738.

Baker, G. C., Smith, J. J., & Cowan, D. A. (2003). Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods*, *55*(3), 541-555. https://doi.org/10.1016/j.mimet.2003.08.009

Baker, P. A., Fritz, S. C., Dick, C. W., Eckert, A. J., Horton, B. K., Manzoni, S., Ribas, C. C., Garzione, C. N., & Battisti, D. S. (2014). The emerging field of geogenomics: constraining geological problems with genetic data. *Earth-Science Reviews*, *135*, 38-47.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, *19*(5), 455-477. https://doi.org/10.1089/cmb.2012.0021

Barton, H. A. (2015). Starving Artists: Bacterial Oligotrophic Heterotrophy in Caves. In *Microbial life of cave systems* (pp. 79-104). De Gruyter.

Barton, H. A., Taylor, N. M., Lubbers, B. R., & Pemberton, A. C. (2006). DNA extraction from low-biomass carbonate rock: an improved method with reduced contamination and the low-biomass contaminant database. *J Microbiol Methods*, *66*(1), 21-31. https://doi.org/10.1016/j.mimet.2005.10.005

Basso, O., Lascourreges, J. F., Jarry, M., & Magot, M. (2005). The effect of cleaning and disinfecting the sampling well on the microbial communities of deep subsurface water samples. *Environ Microbiol*, *7*(1), 13-21. https://doi.org/10.1111/j.1462-2920.2004.00660.x

Bayer, B., Saito, M. A., McIlvin, M. R., Lucker, S., Moran, D. M., Lankiewicz, T. S., Dupont, C. L., & Santoro, A. E. (2021). Metabolic versatility of the nitrite-oxidizing bacterium Nitrospira marina and its proteomic response to oxygen-limited conditions. *ISME J*, *15*(4), 1025-1039. https://doi.org/10.1038/s41396-020-00828-3

Beam, J. P., Becraft, E. D., Brown, J. M., Schulz, F., Jarett, J. K., Bezuidt, O., Poulton, N. J., Clark, K., Dunfield, P. F., Ravin, N. V., Spear, J. R., Hedlund, B. P., Kormas, K. A., Sievert, S. M., Elshahed, M. S., Barton, H. A., Stott, M. B., Eisen, J. A., Moser, D. P., Onstott, T. C., Woyke, T., & Stepanauskas, R. (2020). Ancestral Absence of Electron Transport Chains in Patescibacteria and DPANN. *Front Microbiol*, *11*, 1848. https://doi.org/10.3389/fmicb.2020.01848

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular ecology resources*, *15*(6), 1403-1414.

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). 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., Cope, E. K., Da Silva, R., Diener, C., Dorrestein, P. C., Douglas, G. M., Durall, D. M., Duvallet, C., Edwardson, C. F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J. M., Gibbons, S. M., Gibson, D. L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G. A., Janssen, S., Jarmusch, A. K., Jiang, L., Kaehler, B. D., Kang, K. B., Keefe, C. R., Keim, P., Kelley, S. T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M. G. I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B. D., McDonald, D., McIver, L. J., Melnik, A. V., Metcalf, J. L., Morgan, S. C., Morton, J. T., Naimey, A. T., Navas-Molina, J. A., Nothias, L. F., Orchanian, S. B., Pearson, T., Peoples, S. L., Petras, D., Preuss, M. L., Pruesse, E., Rasmussen, L. B., Rivers, A., Robeson, M. S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S. J., Spear, J. R., Swafford, A. D., Thompson, L. R., Torres, P. J., Trinh, P., Tripathi, A., Turnbaugh, P. J., Ul-Hasan, S., Van Der Hooft, J. J. J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., Von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K. C., Williamson, C. H. D., Willis, A. D., Xu, Z. Z., Zaneveld, J. R., Zhang, Y., Zhu, Q., Knight, R., & 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

Borgonie, G., García-Moyano, A., Litthauer, D., Bert, W., Bester, A., van Heerden, E., Möller, C., Erasmus, M., & Onstott, T. C. (2011). Nematoda from the terrestrial deep subsurface of South Africa. *Nature*, *474*(7349), 79-82.

Boucher, Y., Labbate, M., Koenig, J. E., & Stokes, H. W. (2007). Integrons: mobilizable platforms that promote genetic diversity in bacteria. *Trends Microbiol*, *15*(7), 301-309. https://doi.org/10.1016/j.tim.2007.05.004

Bowers, R. M., Clum, A., Tice, H., Lim, J., Singh, K., Ciobanu, D., Ngan, C. Y., Cheng, J. F., Tringe, S. G., & Woyke, T. (2015). Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics*, *16*, 856. https://doi.org/10.1186/s12864-015-2063-6

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., Weinstock, G. M., Garrity, G. M., Dodsworth, J. A., Yooseph, S., Sutton, G., Glockner, F. O., Gilbert, J. A., Nelson, W. C., Hallam, S. J., Jungbluth, S. P., Ettema, T. J. G., Tighe, S., Konstantinidis, K. T., Liu, W. T., Baker, B. J., Rattei, T., Eisen, J. A., Hedlund, B., McMahon, K. D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G. W., Rinke, C., Genome Standards, C., Lapidus, A., Meyer, F., Yilmaz, P., Parks, D. H., Eren, A. M., Schriml, L., Banfield, J. F., Hugenholtz, P., & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*, *35*(8), 725-731. https://doi.org/10.1038/nbt.3893

Boyer, D. G., & Pasquarell, G. C. (1999). Agricultural Land Use Impacts on Bacterial Water Quality in a Karst Groundwater Aquifer. *JAWRA Journal of the American Water Resources Association*, *35*(2), 291-300.

Brannen-Donnelly, K., & Engel, A. S. (2015). Bacterial diversity differences along an epigenic cave stream reveal evidence of community dynamics, succession, and stability. *Front Microbiol*, *6*, 729. https://doi.org/10.3389/fmicb.2015.00729

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208-211. https://doi.org/10.1038/nature14486

Buongermino Pereira, M., Osterlund, T., Eriksson, K. M., Backhaus, T., Axelson-Fisk, M., & Kristiansson, E. (2020). A comprehensive survey of integron-associated genes present in metagenomes. *BMC Genomics*, *21*(1), 495. https://doi.org/10.1186/s12864-020-06830-5

Bushnell, B. (2018). BBTools: a suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. *Joint Genome Institute*.

Bushnell, B., Egan, R., Copeland, A., Foster, B., Clum, A., & Sun, H. (2019). BBMap: a fast, accurate, splice-aware aligner. 2014. *Available: sourceforge. net/projects/bbmap*.

Cadier, M., Gorgues, T., LHelguen, S., Sourisseau, M., & Memery, L. (2017). Tidal cycle control of biogeochemical and ecological properties of a macrotidal ecosystem. *Geophysical Research Letters*, *44*(16), 8453-8462.

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

Cambray, G., Guerout, A. M., & Mazel, D. (2010). Integrons. *Annu Rev Genet*, *44*, 141-166. https://doi.org/10.1146/annurev-genet-102209-163504

Campbell, B. J., Engel, A. S., Porter, M. L., & Takai, K. (2006). The versatile epsilon-proteobacteria: key players in sulphidic habitats. *Nat Rev Microbiol*, *4*(6), 458-468. https://doi.org/10.1038/nrmicro1414

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., & Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, *26*(2), 266-267. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2804299/pdf/btp636.pdf

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Pena, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., & Knight, R. (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*, *7*(5), 335-336. https://doi.org/10.1038/nmeth.f.303

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A*, *108 Suppl 1*, 4516-4522. https://doi.org/10.1073/pnas.1000080107

Castelle, C. J., Brown, C. T., Thomas, B. C., Williams, K. H., & Banfield, J. F. (2017). Unusual respiratory capacity and nitrogen metabolism in a Parcubacterium (OD1) of the Candidate Phyla Radiation. *Scientific reports*, *7*(1), 1-12.

Caumartin, V. (1963). Review of the microbiology of underground environments. *Natl. Speleol. Soc. Bull.*, *25*, 1-14.

Chandler, D. P., Fredrickson, J. K., & Brockman, F. J. (1997). Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. *Mol. Ecol.*, *6*, 475-482.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of statistics*, 265-270.

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., Huntemann, M., Varghese, N., White, J. R., & Seshadri, R. (2019). IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic acids research*, *47*(D1), D666-D677.

Cho, J.-C., & Kim, S.-J. (2000). Increase in bacterial community diversity in subsurface aquifers receiving livestock wastewater input. *Applied and Environmental Microbiology*, *66*(3), 956-965.

Claassen, S., du Toit, E., Kaba, M., Moodley, C., Zar, H. J., & Nicol, M. P. (2013). A comparison of the efficiency of five different commercial DNA extraction kits for extraction of DNA from faecal samples. *J Microbiol Methods*, *94*(2), 103-110. https://doi.org/10.1016/j.mimet.2013.05.008

Cunningham, K., Northup, D., Pollastro, R., Wright, W., & LaRock, E. (1995). Bacteria, fungi and biokarst in Lechuguilla cave, Carlsbad Caverns National Park, New Mexico. *Environmental Geology*, *25*(1), 2-8.

Czárán, T. L., Hoekstra, R. F., & Pagie, L. (2002). Chemical warfare between microbes promotes biodiversity. *Proceedings of the National Academy of Sciences*, *99*(2), 786-790. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC117383/pdf/pq0202000786.pdf

Daims, H., Brühl, A., Amann, R., Schleifer, K.-H., & Wagner, M. (1999). The domain-specific probe EUB338 is insufficient for the detection of all Bacteria: development and evaluation of a more comprehensive probe set. *Systematic and applied microbiology*, *22*(3), 434-444.

Davis, D. G. (1991). Wind Cave helictite bushes as a subaqueous speleothem - further observations. *Geo2*, *19*, 13-15.

DeSantis, T. Z., Brodie, E. L., Moberg, J. P., Zubieta, I. X., Piceno, Y. M., & Andersen, G. L. (2007). High-Density Universal 16S rRNA Microarray Analysis Reveals Broader Diversity than Typical Clone Library When Sampling the Environment. *Microbial ecology*, *53*(3), 371-383. https://doi.org/10.1007/s00248-006-9134-9

DeWitt, E., Redden, J. A., Burack Wilson, A., & Buscher, D. (1986). *Mineral resource potential and Geology of the Black Hills National Forest, South Dakota and Wyoming.* (US Geological Survey Bulletin, Issue.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol*, *10*(8), R85. https://doi.org/10.1186/gb-2009-10-8-r85

Dojka, M. A., Hugenholtz, P., Haack, S. K., & Pace, N. R. (1998). Microbial diversity in a hydrocarbon-and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Applied and Environmental Microbiology*, *64*(10), 3869-3877.

Douglas, G. M., Maffei, V. J., Zaneveld, J., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., & Langille, M. G. (2020). PICRUSt2: An improved and customizable approach for metagenome inference. *BioRxiv*, 672295.

Dueñas, M., Munduate, A., Perea, A., & Irastorza, A. (2003). Exopolysaccharide production by Pediococcus damnosus 2.6 in a semidefined medium under different growth conditions. *International journal of food microbiology*, *87*(1-2), 113-120.

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

R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194-2200. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3150044/pdf/btr381.pdf

Elshahed, M. S., Senko, J. M., Najar, F. Z., Kenton, S. M., Roe, B. A., Dewers, T. A., Spear, J. R., & Krumholz, L. R. (2003). Bacterial diversity and sulfur cycling in a mesophilic sulfide-rich spring. *Appl Environ Microbiol*, *69*(9), 5609-5621. https://doi.org/10.1128/AEM.69.9.5609-5621.2003

Engel, A. S., & Randall, K. W. (2011). Experimental evidence for microbially mediated carbonate dissolution from the saline water zone of the Edwards Aquifer, Central Texas. *Geomicrobiology Journal*, *28*(4), 313-327.

Engelstadter, J., Harms, K., & Johnsen, P. J. (2016). The evolutionary dynamics of integrons in changing environments. *ISME J*, *10*(6), 1296-1307. https://doi.org/10.1038/ismej.2015.222

Farnleitner, A. H., Wilhartitz, I., Ryzinska, G., Kirschner, A. K., Stadler, H., Burtscher, M. M., Hornek, R., Szewzyk, U., Herndl, G., & Mach, R. L. (2005). Bacterial dynamics in spring water of alpine karst aquifers indicates the presence of stable autochthonous microbial endokarst communities. *Environmental Microbiology*, *7*(8), 1248-1259.

Fliermans, C., & Schmidt, E. (1977). Nitrobacter in Mammoth Cave. . *International Journal of Speleology*, *9*, 1-19.

Ford, D., & Williams, P. D. (2013). *Karst hydrogeology and geomorphology*. John Wiley & Sons.

Ford, D. C., Lundberg, J., Palmer, A., Palmer, M., Dreybrodt, W., & Schwarcz, H. (1993). Uranium-series dating of the draining of an aquifer: The example of Wind Cave, Black Hills, South Dakota. *Geological Society of America Bulletin*, *105*(2), 241-250.

Forde, B. M., & O'Toole, P. W. (2013). Next-generation sequencing technologies and their impact on microbial genomics. *Brief Funct Genomics*, *12*(5), 440-453. https://doi.org/10.1093/bfgp/els062

Francis, C. A., Co, E.-M., & Tebo, B. M. (2001). Enzymatic Manganese(II) Oxidation by a Marine α-Proteobacterium. *Applied and Environmental*

Microbiology, *67*(9), 4024-4029. https://doi.org/10.1128/AEM.67.9.4024-4029.2001

Francis, C. A., Roberts, K. J., Beman, J. M., Santoro, A. E., & Oakley, B. B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. *Proc Natl Acad Sci U S A*, *102*(41), 14683-14688. https://doi.org/10.1073/pnas.0506625102

Freese, H. M., Karsten, U., & Schumann, R. (2006). Bacterial abundance, activity, and viability in the eutrophic River Warnow, northeast Germany. *Microb Ecol*, *51*(1), 117-127. https://doi.org/10.1007/s00248-005-0091-5

Fuerst, J. A., & Sagulenko, E. (2011). Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. *Nature Reviews Microbiology*, *9*(6), 403-413. https://www.nature.com/articles/nrmicro2578

Fujitani, H., Ushiki, N., Tsuneda, S., & Aoi, Y. (2014). Isolation of sublineage I Nitrospira by a novel cultivation strategy. *Environ Microbiol*, *16*(10), 3030-3040. https://doi.org/10.1111/1462-2920.12248

Garber, A. (2020). MagicLamp: toolkit for annotation of 'omics datasets using curated HMM sets. In. GitHub repository: https://github.com/Arkadiy-Garber/MagicLamp.

Garcia-Solsona, E., Garcia-Orellana, J., Masqué, P., Rodellas, V., Mejías, M., Ballesteros, B., & Domínguez, J. (2010). Groundwater and nutrient discharge through karstic coastal springs (Castelló, Spain). *Biogeosciences*, *7*(9), 2625-2638.

Garrido-Cardenas, J. A., & Manzano-Agugliaro, F. (2017). The metagenomics worldwide research. *Curr Genet*, *63*(5), 819-829. https://doi.org/10.1007/s00294-017-0693-8

Garrity, G. M., Bell, J. A., & Lilburn, T. (2005). Acidithiobacillales ord. nov. In *Bergey’s Manual® of Systematic Bacteriology* (pp. 60-63). Springer.

Geszvain, K., Butterfield, C. N., Davis, R. E., Madison, A. S., Lee, S.-W., Parker, D. L., Soldatova, A., Spiro, T. G., Luther, G. W., III, & Tebo, B. M. (2012). The molecular biogeochemistry of manganese(II) oxidation. *Biochemical Society Transactions*, *40*(6), 1244-1248. https://doi.org/10.1042/BST20120229

Ghai, R., Mizuno, C. M., Picazo, A., Camacho, A., & Rodriguez-Valera, F. (2013). Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. *Sci Rep*, *3*, 2471. https://doi.org/10.1038/srep02471

Ghaly, T. M., Geoghegan, J. L., Tetu, S. G., & Gillings, M. R. (2020). The peril and promise of integrons: beyond antibiotic resistance. *Trends in microbiology*, *28*(6), 455-464.

Gillings, M. R. (2014). Integrons: past, present, and future. *Microbiol Mol Biol Rev*, *78*(2), 257-277. https://doi.org/10.1128/MMBR.00056-13

Glöckner, J., Kube, M., Shrestha, P. M., Weber, M., Glöckner, F. O., Reinhardt, R., & Liesack, W. (2010). Phylogenetic diversity and metagenomics of candidate division OP3. *Environmental Microbiology*, *12*(5), 1218-1229. https://sfamjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1111/j.1462-2920.2010.02164.x?download=true

Gloor, G. (2015). ALDEx2: ANOVA-Like Differential Expression tool for compositional data. *ALDEX manual modular*, *20*, 1-11.

Gloor, G. (2016). Compositional data analysis for high throughput sequencing: an example from 16S rRNA gene sequencing. In.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol*, *8*, 2224. https://doi.org/10.3389/fmicb.2017.02224

Glud, R. N., Wenzhöfer, F., Middelboe, M., Oguri, K., Turnewitsch, R., Canfield, D. E., & Kitazato, H. (2013). High rates of microbial carbon turnover in sediments in the deepest oceanic trench on Earth. *Nature Geoscience*, *6*(4), 284-288.

Goldscheider, N., Hunkeler, D., & Rossi, P. (2006). Review: Microbial biocenoses in pristine aquifers and an assessment of investigative methods. *Hydrogeology Journal*, *14*(6), 926-941. https://doi.org/10.1007/s10040-005-0009-9

Gonzalez, I., Laiz, L., Hermosin, B., Caballero, B., Incerti, C., & Saiz-Jimenez, C. (1999). Bacteria isolated from rock art paintings: the case of Atlanterra shelter (south Spain). *Journal of Microbiological Methods*, *36*(1–2), 123-127. https://doi.org/http://dx.doi.org/10.1016/S0167-7012(99)00017-2

Gray, C. J., & Engel, A. S. (2013). Microbial diversity and impact on carbonate geochemistry across a changing geochemical gradient in a karst aquifer. *ISME J*, *7*(2), 325-337. https://doi.org/10.1038/ismej.2012.105

Gray, N. D., & Head, I. M. (2001). Linking genetic identity and function in communities of uncultured bacteria. *Environmental Microbiology*, *3*(8), 481-492. https://doi.org/DOI 10.1046/j.1462-2920.2001.00214.x

Greene, E. A. (1993). *Hydraulic properties of the Madison aquifer system in the western Rapid City area, South Dakota* (Vol. 93). US Department of the Interior, US Geological Survey.

Griebler, C., & Lueders, T. (2009). Microbial biodiversity in groundwater ecosystems. *Freshwater Biology*, *54*(4), 649-677. https://doi.org/10.1111/j.1365-2427.2008.02013.x

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *68*(4), 669-685. https://doi.org/10.1128/mmbr.68.4.669-685.2004

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, *5*(10), R245-R249.

Harf, C., & Monteil, H. (1988). Interactions between free-living amoebae and Legionella in the environment. *Water Science and Technology*, *20*(11-12), 235-239.

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: overviews and challenges. *Biotechniques*, *56*(2), 61-64, 66, 68, passim. https://doi.org/10.2144/000114133

Hedlund, B. P., Dodsworth, J. A., Murugapiran, S. K., Rinke, C., & Woyke, T. (2014). Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". *Extremophiles*, *18*(5), 865-875. https://doi.org/10.1007/s00792-014-0664-7

Heo, Y., Wu, X.-L., Chen, D., Ma, J., & Hwu, W.-M. (2014). BLESS: bloom filter-based error correction solution for high-throughput sequencing reads. *Bioinformatics*, *30*(10), 1354-1362.

Herrmann, M., Opitz, S., Harzer, R., Totsche, K. U., & Kusel, K. (2017). Attached and Suspended Denitrifier Communities in Pristine Limestone Aquifers Harbor High Fractions of Potential Autotrophs Oxidizing Reduced Iron and Sulfur Compounds. *Microb Ecol*, *74*(2), 264-277. https://doi.org/10.1007/s00248-017-0950-x

Hershey, O. S., & Barton, H. A. (2018). The Microbial Diversity of Caves. In *Cave Ecology* (pp. 69-90). https://doi.org/10.1007/978-3-319-98852-8_5

Hershey, O. S., Kallmeyer, J., & Barton, H. A. (2019). A Practical Guide to Studying the Microbiology of Karst Aquifers. In *Karst Water Environment* (pp. 191-207). https://doi.org/10.1007/978-3-319-77368-1_7

Hershey, O. S., Kallmeyer, J., Wallace, A., Barton, M. D., & Barton, H. A. (2018). High Microbial Diversity Despite Extremely Low Biomass in a Deep Karst Aquifer [10.3389/fmicb.2018.02823]. *Frontiers in microbiology*, *9*, 2823. https://www.frontiersin.org/article/10.3389/fmicb.2018.02823

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

Hiraoka, S., Yang, C. C., & Iwasaki, W. (2016). Metagenomics and Bioinformatics in Microbial Ecology: Current Status and Beyond. *Microbes Environ*, *31*(3), 204-212. https://doi.org/10.1264/jsme2.ME16024

Høeg, O. A. (1946). Cyanophyceae and bacteria in calcareous sediments in the interior of limestone caves in Nord-Rana, Norway. *Nytt Magazin for Naturvidenskapene*, *85*, 99-104.

Hose, G., & Lategan, M. (2012). Sampling strategies for biological assessment of groundwater ecosystems. *CRC CARE Technical Report no 21*.

Hudak, P. (2000). Regional trends in nitrate content of Texas groundwater. *Journal of Hydrology*, *228*(1-2), 37-47.

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. *ISME J*, *9*(8), 1846-1856. https://doi.org/10.1038/ismej.2015.2

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, *180*(18), 4765-4774.

Hugenholtz, P., & Kyrpides, N. C. (2009). A changing of the guard. *Environ Microbiol*, *11*(3), 551-553. https://doi.org/10.1111/j.1462-2920.2009.01888.x

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., & Bohannan, B. J. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, *67*(10), 4399-4406.

Hugler, M., & Sievert, S. M. (2011). Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. *Ann Rev Mar Sci*, *3*, 261-289. https://doi.org/10.1146/annurev-marine-120709-142712

Hutchinson, G. E. (1961). The paradox of the plankton. *The American Naturalist*, *95*(882), 137-145.

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*. https://doi.org/Artn 11910.1186/1471-2105-11-119

Iker, B. C., Kambesis, P., Oehrle, S. A., Groves, C., & Barton, H. A. (2010). Microbial atrazine breakdown in a karst groundwater system and its effect on ecosystem energetics. *J Environ Qual*, *39*(2), 509-518. https://doi.org/10.2134/jeq2009.0048

Interlandi, S. J., & Kilham, S. S. (2001). Limiting resources and the regulation of diversity in phytoplankton communities. *Ecology*, *82*(5), 1270-1282.

Jeraldo, P., Chia, N., & Goldenfeld, N. (2011). On the suitability of short reads of 16S rRNA for phylogeny-based analyses in environmental surveys. *Environ Microbiol*, *13*(11), 3000-3009. https://doi.org/10.1111/j.1462-2920.2011.02577.x

John, D. E., & Rose, J. B. (2005). Review of factors affecting microbial survival in groundwater. *Environmental science & technology*, *39*(19), 7345-7356.

Johnson, J. E., Webb, S. M., Ma, C., & Fischer, W. W. (2016). Manganese mineralogy and diagenesis in the sedimentary rock record. *Geochimica et Cosmochimica Acta*, *173*, 210-231. https://doi.org/https://doi.org/10.1016/j.gca.2015.10.027

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., & D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci U S A*, *109*(40), 16213-16216. https://doi.org/10.1073/pnas.1203849109

Kallmeyer, J., Smith, D. C., Spivack, A. J., & D'Hondt, S. (2008). New cell extraction procedure applied to deep subsurface sediments. *Limnology and Oceanography: Methods*, *6*(6), 236-245.

Karl, D., Bird, D. F., Björkman, K., Houlihan, T., Shackelford, R., & Tupas, L. (1999). Microorganisms in the accreted ice of Lake Vostok, Antarctica. *Science*, *286*(5447), 2144-2147.

Kato, K., Nagaosa, K., Kimura, H., Katsuyama, C., Hama, K., Kunimaru, T., Tsunogai, U., & Aoki, K. (2009). Unique distribution of deep groundwater bacteria constrained by geological setting. *Environ Microbiol Rep*, *1*(6), 569-574. https://doi.org/10.1111/j.1758-2229.2009.00087.x

Kembel, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Comput Biol*, *8*(10), e1002743. https://doi.org/10.1371/journal.pcbi.1002743

Kennedy, K., Hall, M. W., Lynch, M. D., Moreno-Hagelsieb, G., & Neufeld, J. D. (2014). Evaluating bias of illumina-based bacterial 16S rRNA gene profiles. *Appl Environ Microbiol*, *80*(18), 5717-5722. https://doi.org/10.1128/AEM.01451-14

Kerr, B., Riley, M. A., Feldman, M. W., & Bohannan, B. J. (2002). Local dispersal promotes biodiversity in a real-life game of rock–paper–scissors. *Nature*, *418*(6894), 171-174. https://www.nature.com/articles/nature00823.pdf

Klimchouk, A., Palmer, A. N., De Waele, J., Auler, A. S., & Audra, P. (2017). *Hypogene karst regions and caves of the world*. Springer.

Koch, H., Lucker, S., Albertsen, M., Kitzinger, K., Herbold, C., Spieck, E., Nielsen, P. H., Wagner, M., & Daims, H. (2015). Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus Nitrospira. *Proc Natl Acad Sci U S A*, *112*(36), 11371-11376. https://doi.org/10.1073/pnas.1506533112

Kolinko, S., Jogler, C., Katzmann, E., Wanner, G., Peplies, J., & Schüler, D. (2012). Single-cell analysis reveals a novel uncultivated magnetotactic bacterium within the candidate division OP3. *Environmental Microbiology*, *14*(7), 1709-1721.

Köllner, K. E., Carstens, D., Schubert, C. J., Zeyer, J., & Bürgmann, H. (2013). Impact of particulate organic matter composition and degradation state on the vertical structure of particle-associated and planktonic lacustrine bacteria. *Aquatic Microbial Ecology*, *69*(1), 81-92. https://doi.org/10.3354/ame01623

Korbel, K., Chariton, A., Stephenson, S., Greenfield, P., & Hose, G. C. (2017). Wells provide a distorted view of life in the aquifer: implications for sampling, monitoring and assessment of groundwater ecosystems. *Sci Rep*, *7*, 40702. https://doi.org/10.1038/srep40702

Kováč, Ľ. (2018). Caves as Oligotrophic Ecosystems. In *Cave Ecology* (pp. 297-307). https://doi.org/10.1007/978-3-319-98852-8_13

Kozich, J. J., Westcott, S. L., Baxter, N. T., Highlander, S. K., & Schloss, P. D. (2013). Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol*, *79*(17), 5112-5120. https://doi.org/10.1128/AEM.01043-13

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Sci Rep*, *7*(1), 17668. https://doi.org/10.1038/s41598-017-17333-x

Kuhn, E., Ichimura, A. S., Peng, V., Fritsen, C. H., Trubl, G., Doran, P. T., & Murray, A. E. (2014). Brine assemblages of ultrasmall microbial cells within the ice cover of Lake Vida, Antarctica. *Appl Environ Microbiol*, *80*(12), 3687-3698. https://doi.org/10.1128/AEM.00276-14

Kwon, E. Y., Kim, G., Primeau, F., Moore, W. S., Cho, H. M., DeVries, T., Sarmiento, J. L., Charette, M. A., & Cho, Y. K. (2014). Global estimate of submarine groundwater discharge based on an observationally constrained radium isotope model. *Geophysical Research Letters*, *41*(23), 8438-8444.

Labbate, M., Case, R. J., & Stokes, H. W. (2009). The integron/gene cassette system: an active player in bacterial adaptation. *Horizontal gene transfer*, 103-125.

Lander, E. S. (2011). Initial impact of the sequencing of the human genome [10.1038/nature09792]. *Nature*, *470*(7333), 187-197. https://doi.org/http://www.nature.com/nature/journal/v470/n7333/abs/10.1038-nature09792-unlocked.html#supplementary-information

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol*, *31*(9), 814-821. https://doi.org/10.1038/nbt.2676

LaRock, E. J., & Cunningham, K. I. (1995). Helictite bush formation and aquifer cooling in Wind Cave, Wind Cave National Park, South Dakota. *Journal of Cave and Karst Studies*, *57*(1), 43-51.

Learman, D. R., Voelker, B. M., Vazquez-Rodriguez, A. I., & Hansel, C. M. (2011). Formation of manganese oxides by bacterially generated superoxide. *Nature Geoscience*, *4*(2), 95-98. https://doi.org/10.1038/ngeo1055

Lefèvre, C. T. (2016). Genomic insights into the early-diverging magnetotactic bacteria. *Environmental Microbiology*, *18*(1), 1-3. https://sfamjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1111/1462-2920.12989?download=true

Lefèvre, C. T., Trubitsyn, D., Abreu, F., Kolinko, S., de Almeida, L. G. P., de Vasconcelos, A. T. R., Lins, U., Schüler, D., Ginet, N., & Pignol, D. (2013). Monophyletic origin of magnetotaxis and the first magnetosomes. *Environmental Microbiology*, *15*(8), 2267-2274.

Lehman, R. M. (2007). Understanding of Aquifer Microbiology is Tightly Linked to Sampling Approaches. *Geomicrobiology Journal*, *24*(3-4), 331-341. https://doi.org/10.1080/01490450701456941

Li, X., & Li, J. (2015). Dead-End Filtration. In *Encyclopedia of Membranes* (pp. 1-3). https://doi.org/10.1007/978-3-642-40872-4_2196-1

Lin, W., Pan, Y., & Bazylinski, D. A. (2017). Diversity and ecology of and biomineralization by magnetotactic bacteria. *Environmental microbiology reports*, *9*(4), 345-356. https://sfamjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1111/1758-2229.12550?download=true

Long, A. J., Ohms, M. J., & McKaskey, J. D. (2012). *Groundwater flow, quality (2007-10), and mixing in the Wind Cave National Park area, South Dakota*. US Department of the Interior, US Geological Survey.

Long, A. J., & Valder, J. F. (2011). Multivariate analyses with end-member mixing to characterize groundwater flow: Wind Cave and associated aquifers. *Journal of Hydrology*, *409*(1-2), 315-327. https://doi.org/10.1016/j.jhydrol.2011.08.028

López-García, P., Rodriguez-Valera, F., Pedrós-Alió, C., & Moreira, D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, *409*(6820), 603-607. https://www.nature.com/articles/35054537.pdf

Loy, A., Maixner, F., Wagner, M., & Horn, M. (2007). probeBase—an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic acids research*, *35*(suppl_1), D800-D804.

Lozupone, C., & Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol*, *71*(12), 8228-8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Lucker, S., Wagner, M., Maixner, F., Pelletier, E., Koch, H., Vacherie, B., Rattei, T., Damste, J. S., Spieck, E., Le Paslier, D., & Daims, H. (2010). A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. *Proc Natl Acad Sci U S A*, *107*(30), 13479-13484. https://doi.org/10.1073/pnas.1003860107

Ludwig, W., Strunk, O., Klugbauer, S., Klugbauer, N., Weizenegger, M., Neumaier, J., Bachleitner, M., & Schleifer, K. H. (1998). Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis*, *19*(4), 554-568.

Luef, B., Frischkorn, K. R., Wrighton, K. C., Holman, H. Y., Birarda, G., Thomas, B. C., Singh, A., Williams, K. H., Siegerist, C. E., Tringe, S. G., Downing, K. H., Comolli, L. R., & Banfield, J. F. (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. *Nat Commun*, *6*, 6372. https://doi.org/10.1038/ncomms7372

Lynch, M. D., & Neufeld, J. D. (2015). Ecology and exploration of the rare biosphere. *Nat Rev Microbiol*, *13*(4), 217-229. https://doi.org/10.1038/nrmicro3400

Maclay, R. W. (1995). *Geology and hydrology of the Edwards Aquifer in the San Antonio area, Texas* (Vol. 95). US Geological Survey.

Maniloff, J. (1997). Nannobacteria: Size limits and evidence. *Science*, *276*(5320), 1776-1776. <Go to ISI>://WOS:A1997XF10300006

Marcy, Y., Ouverney, C., Bik, E. M., Losekann, T., Ivanova, N., Martin, H. G., Szeto, E., Platt, D., Hugenholtz, P., Relman, D. A., & Quake, S. R. (2007). Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A*, *104*(29), 11889-11894. https://doi.org/10.1073/pnas.0704662104

Marti, R., Becouze-Lareure, C., Ribun, S., Marjolet, L., Bernardin Souibgui, C., Aubin, J. B., Lipeme Kouyi, G., Wiest, L., Blaha, D., & Cournoyer, B. (2017). Bacteriome genetic structures of urban deposits are indicative of their origin and impacted by chemical pollutants. *Sci Rep*, *7*(1), 13219. https://doi.org/10.1038/s41598-017-13594-8

Mazel, D. (2006). Integrons: agents of bacterial evolution. *Nat Rev Microbiol*, *4*(8), 608-620. https://doi.org/10.1038/nrmicro1462

McCormack, T., Gill, L., Naughton, O., & Johnston, P. (2014). Quantification of submarine/intertidal groundwater discharge and nutrient loading from a lowland karst catchment. *Journal of Hydrology*, *519*, 2318-2330.

McKinney, W. (2010). Data structures for statistical computing in python. Proceedings of the 9th Python in Science Conference,

McMahon, S., & Parnell, J. (2014). Weighing the deep continental biosphere. *FEMS Microbiol Ecol*, *87*(1), 113-120. https://doi.org/10.1111/1574-6941.12196

McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, *8*(4), e61217. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3632530/pdf/pone.0061217.pdf

Mi, S., Song, J., Lin, J., Che, Y., Zheng, H., & Lin, J. (2011). Complete genome of Leptospirillum ferriphilum ML-04 provides insight into its physiology and environmental adaptation. *The Journal of Microbiology*, *49*(6), 890-901.

Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W., & Banfield, J. F. (2011). EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology*, *12*(5), 1-14. https://doi.org/10.1186/gb-2011-12-5-r44

Mirete, S., Morgante, V., & Gonzalez-Pastor, J. E. (2016). Functional metagenomics of extreme environments. *Curr Opin Biotechnol*, *38*, 143-149. https://doi.org/10.1016/j.copbio.2016.01.017

Miteva, V. I., & Brenchley, J. E. (2005). Detection and isolation of ultrasmall microorganisms from a 120,000-year-old Greenland glacier ice core. *Appl Environ Microbiol*, *71*(12), 7806-7818. https://doi.org/10.1128/AEM.71.12.7806-7818.2005

Miyoshi, T., Iwatsuki, T., & Naganuma, T. (2005). Phylogenetic characterization of 16S rRNA gene clones from deep-groundwater microorganisms that pass through 0.2-micrometer-pore-size filters. *Appl Environ Microbiol*, *71*(2), 1084-1088. https://doi.org/10.1128/AEM.71.2.1084-1088.2005

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. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5607374/pdf/ismej201794a.pdf

Muyzer, G., De Waal, E. C., & Uitterlinden, A. G. (1993). Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Applied and Environmental Microbiology*, *59*(3), 695-700.

Myrttinen, A., Becker, V., van Geldern, R., Würdemann, H., Morozova, D., Zimmer, M., Taubald, H., Blum, P., & Barth, J. A. C. (2010). Carbon and oxygen isotope indications for CO2 behaviour after injection: First results from the Ketzin site (Germany). *International Journal of Greenhouse Gas Control*, *4*(6), 1000-1006. https://doi.org/10.1016/j.ijggc.2010.02.005

Naus, C. A., Driscoll, D. G., & Carter, J. M. (2001). *Geochemistry of the Madison and Minnelusa aquifers in the Black Hills area, South Dakota*. US Department of the Interior, US Geological Survey.

Nayfach, S., & Pollard, K. S. (2016). Toward Accurate and Quantitative Comparative Metagenomics. *Cell*, *166*(5), 1103-1116. https://doi.org/10.1016/j.cell.2016.08.007

Nealson, K. H., Tebo, B. M., & Rosson, R. A. (1988). Occurrence and Mechanisms of Microbial Oxidation of Manganese. In A. I. Laskin (Ed.), *Advances in Applied Microbiology* (Vol. 33, pp. 279-318). Academic Press. https://doi.org/https://doi.org/10.1016/S0065-2164(08)70209-0

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., & Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev*, *75*(1), 14-49. https://doi.org/10.1128/MMBR.00028-10

Newton, R. J., & McMahon, K. D. (2011). Seasonal differences in bacterial community composition following nutrient additions in a eutrophic lake. *Environ Microbiol*, *13*(4), 887-899. https://doi.org/10.1111/j.1462-2920.2010.02387.x

North, N. N., Dollhopf, S. L., Petrie, L., Istok, J. D., Balkwill, D. L., & Kostka, J. E. (2004). Change in bacterial community structure during in situ biostimulation of subsurface sediment cocontaminated with uranium and nitrate. *Appl Environ Microbiol*, *70*(8), 4911-4920. https://doi.org/10.1128/AEM.70.8.4911-4920.2004

Ohrel Jr, R., & Register, K. (2006). *Volunteer Estuary Monitoring: A methods Manual . Washington, USEPA, Techn. Doc*.

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. *ISME J*, *8*(2), 478-491. https://doi.org/10.1038/ismej.2013.159

Ortiz, M., Neilson, J. W., Nelson, W. M., Legatzki, A., Byrne, A., Yu, Y., Wing, R. A., Soderlund, C. A., Pryor, B. M., & Pierson III, L. S. (2013). Profiling bacterial diversity and taxonomic composition on speleothem surfaces in Kartchner Caverns, AZ. *Microbial Ecology*, *65*(2), 371-383.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., & Edwards, R. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic acids research*, *33*(17), 5691-5702.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, *276*, 734-740.

Pace, N. R., Stahl, D. A., Lane, D. J., & Olsen, G. J. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Advances Microbial Ecology*, *9*, 1-55.

Palmer, A., & Palmer, M. (2000). Speleogenesis of the Black Hills maze caves, South Dakota, USA. *Speleogenesis. Evolution of karst aquifers. Huntsville, National Speleological Society*, 274-281.

Palmer, A. N. (1981). *The Geology of Wind Cave.* Wind Cave Natural History Association.

Palmer, A. N. (1990). Groundwater processes in karst terranes. In *Groundwater Geomorphology; The Role of Subsurface Water in Earth-Surface Processes and Landforms* (pp. 177-210). https://doi.org/10.1130/SPE252-p177

Palmer, A. N. (2011). Distinction between epigenic and hypogenic maze caves. *Geomorphology*, *134*(1-2), 9-22. https://doi.org/10.1016/j.geomorph.2011.03.014

Palmer, A. N. (2017). *Cave Geology*. Cave Books.

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*(5), 1403-1414.

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 Res*, *25*(7), 1043-1055. https://doi.org/10.1101/gr.186072.114

Parro, V., & Moreno-Paz, M. (2003). Gene function analysis in environmental isolates: The *nif* regulon of the strict iron oxidizing bacterium *Leptospirillum ferrooxidans. Proceedings of the National Academy of Sciences*, *100*(13), 7883. https://doi.org/10.1073/pnas.1230487100

Pedersen, K., Arlinger, J., Hallbeck, L., & Pettersson, C. (1996). Diversity and distribution of subterranean bacteria in groundwater at Oklo in Gabon, Africa, as determined by 16S rRNA gene sequencing. *Molecular Ecology*, *5*(3), 427-436. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-294X.1996.d01-320.x?sid=nlm%3Apubmed

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Petrusevski, B., Bolier, G., Van Breemen, A., & Alaerts, G. (1995). Tangential flow filtration: a method to concentrate freshwater algae. *Water Research*, *29*(5), 1419-1424.

Pinowska, A., Stevenson, R., Sickman, J., Albertin, A., & Anderson, M. (2007). Integrated interpretation of survey for determining nutrient thresholds for macroalgae in Florida springs. *Florida Department of Environmental Protection, Tallahassee, Florida, USA*.

Pinto, A. J., & Raskin, L. (2012). PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLoS One*, *7*(8), e43093. https://doi.org/10.1371/journal.pone.0043093

Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product rations in multitemplate PCR. *Applied Environmental Microbiology.*, *64*, 3724-3730.

Prakash, O., Shouche, Y., Jangid, K., & Kostka, J. E. (2013). Microbial cultivation and the role of microbial resource centers in the omics era. *Appl Microbiol Biotechnol*, *97*(1), 51-62. https://doi.org/10.1007/s00253-012-4533-y

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, *5*(3), e9490. https://escholarship.org/content/qt1z34b690/qt1z34b690.pdf?t=qae5az

Pronk, M., Goldscheider, N., & Zopfi, J. (2005). Dynamics and interaction of organic carbon, turbidity and bacteria in a karst aquifer system. *Hydrogeology Journal*, *14*(4), 473-484. https://doi.org/10.1007/s10040-005-0454-5

Pronk, M., Goldscheider, N., & Zopfi, J. (2009). Microbial communities in karst groundwater and their potential use for biomonitoring. *Hydrogeology Journal*, *17*(1), 37-48. https://doi.org/10.1007/s10040-008-0350-x

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic acids research*, *35*(21), 7188-7196.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glockner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*, *41*(Database issue), D590-596. https://doi.org/10.1093/nar/gks1219

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.

Raz, Y., & Tannenbaum, E. (2010). The Influence of Horizontal Gene Transfer on the Mean Fitness of Unicellular Populations in Static Environments. *Genetics*, *185*(1), 327-337. https://doi.org/10.1534/genetics.109.113613

Reeder, J., & Knight, R. (2010). Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nature methods*, *7*(9), 668-669. https://escholarship.org/content/qt1vv43478/qt1vv43478.pdf?t=phrsve

Rinke, C., Low, S., Woodcroft, B. J., Raina, J. B., Skarshewski, A., Le, X. H., Butler, M. K., Stocker, R., Seymour, J., Tyson, G. W., & Hugenholtz, P. (2016). Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ*, *4*, e2486. https://doi.org/10.7717/peerj.2486

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W. T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P., & Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431-437. https://doi.org/10.1038/nature12352

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. *BioRxiv*.

Rodrigue, S., Malmstrom, R. R., Berlin, A. M., Birren, B. W., Henn, M. R., & Chisholm, S. W. (2009). Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One*, *4*(9), e6864. https://doi.org/10.1371/journal.pone.0006864

Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A Review of Bioinformatics Tools for Bio-Prospecting from

Metagenomic Sequence Data. *Front Genet*, *8*, 23. https://doi.org/10.3389/fgene.2017.00023

Rowe, J. M., DeBruyn, J. M., Poorvin, L., LeCleir, G. R., Johnson, Z. I., Zinser, E. R., & Wilhelm, S. W. (2012). Viral and bacterial abundance and production in the Western Pacific Ocean and the relation to other oceanic realms. *FEMS microbiology ecology*, *79*(2), 359-370. https://doi.org/10.1111/j.1574-6941.2011.01223.x

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1977). Nucleotide sequence of bacteriophage [phi]X174 DNA [10.1038/265687a0]. *Nature*, *265*(5596), 687-695. http://dx.doi.org/10.1038/265687a0

Schauer, M., Massana, R., & Pedros-Alio, C. (2000). Spatial differences in bacterioplankton composition along the Catalan coast (NW Mediterranean) assessed by molecular fingerprinting. *FEMS microbiology ecology*, *33*(1), 51-59. https://doi.org/Doi 10.1016/S0168-6496(00)00043-X

Schütz, K., Kandeler, E., Nagel, P., Scheu, S., & Ruess, L. (2010). Functional microbial community response to nutrient pulses by artificial groundwater recharge practice in surface soils and subsoils. *FEMS microbiology ecology*, *72*(3), 445-455.

Scott, W. (1909). An ecological study of the plankton of Shawnee Cave, with notes on the cave environment. *Biological Bulletin*, *17*, 386-407.

Sedlar, K., Kupkova, K., & Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput Struct Biotechnol J*, *15*, 48-55. https://doi.org/10.1016/j.csbj.2016.11.005

Shabarova, T., & Pernthaler, J. (2010). Karst pools in subsurface environments: collectors of microbial diversity or temporary residence between habitat types. *Environ Microbiol*, *12*(4), 1061-1074. https://doi.org/10.1111/j.1462-2920.2009.02151.x

Shen, H., Rogelj, S., & Kieft, T. L. (2006). Sensitive, real-time PCR detects low-levels of contamination by Legionella pneumophila in commercial reagents. *Mol Cell Probes*, *20*(3-4), 147-153. https://doi.org/10.1016/j.mcp.2005.09.007

Shimizu, S., Akiyama, M., Ishijima, Y., Hama, K., Kunimaru, T., & Naganuma, T. (2006). Molecular characterization of microbial communities in fault-bordered aquifers in the Miocene formation of northernmost Japan. *Geobiology*, *4*(3), 203-213.

Silva, G. G., Green, K. T., Dutilh, B. E., & Edwards, R. A. (2016). SUPER-FOCUS: a tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics*, *32*(3), 354-361. https://doi.org/10.1093/bioinformatics/btv584

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 688-688.

Singh, V. K., Singh, K., & Baum, K. (2018). The Role of Methionine Sulfoxide Reductases in Oxidative Stress Tolerance and Virulence of Staphylococcus aureus and Other Bacteria. *Antioxidants (Basel)*, *7*(10). https://doi.org/10.3390/antiox7100128

Sinreich, M., Pronk, M., & Kozel, R. (2014). Microbiological monitoring and classification of karst springs. *Environmental Earth Sciences*, *71*(2), 563-572. https://doi.org/10.1007/s12665-013-2508-7

Smith, R. J., Jeffries, T. C., Roudnew, B., Fitch, A. J., Seymour, J. R., Delpin, M. W., Newton, K., Brown, M. H., & Mitchell, J. G. (2012). Metagenomic comparison of microbial communities inhabiting confined and unconfined aquifer ecosystems. *Environ Microbiol*, *14*(1), 240-253. https://doi.org/10.1111/j.1462-2920.2011.02614.x

Snyder, L. A., Loman, N., Pallen, M. J., & Penn, C. W. (2009). Next-generation sequencing--the promise and perils of charting the great microbial unknown. *Microb Ecol*, *57*(1), 1-3. https://doi.org/10.1007/s00248-008-9465-9

So, A., Pel, J., Rajan, S., & Marziali, A. (2010). Efficient genomic DNA extraction from low target concentration bacterial cultures using SCODA DNA extraction technology. *Cold Spring Harb Protoc*, *2010*(10), pdb prot5506. https://doi.org/10.1101/pdb.prot5506

Soergel, D. A., Dey, N., Knight, R., & Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME J*, *6*(7), 1440-1444. https://doi.org/10.1038/ismej.2011.208

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., & Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci U S A*, *103*(32), 12115-12120.

Stahl, D. A., Lane, D. J., Olsen, G. J., & Pace, N. R. (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science*, *224*, 409-411.

Starikova, I., Harms, K., Haugen, P., Lunde, T. T., Primicerio, R., Samuelsen, O., Nielsen, K. M., & Johnsen, P. J. (2012). A trade-off between the fitness cost of functional integrases and long-term stability of integrons. *PLoS Pathog*, *8*(11), e1003043. https://doi.org/10.1371/journal.ppat.1003043

Stokes, H. W. T., & Hall, R. M. (1989). A novel family of potentially mobile DNA elements encoding site-specific gene-integration functions: integrons. *Molecular microbiology*, *3*(12), 1669-1683.

Sujith, P. P., & Bharathi, P. A. (2011). Manganese oxidation by bacteria: biogeochemical aspects. *Prog Mol Subcell Biol*, *52*, 49-76. https://doi.org/10.1007/978-3-642-21230-7_3

Sunda, W. G., & Huntsman, S. A. (1988). Effect of sunlight on redox cycles of manganese in the southwestern Sargasso Sea. *Deep Sea Research Part A. Oceanographic Research Papers*, *35*(8), 1297-1317. https://doi.org/https://doi.org/10.1016/0198-0149(88)90084-2

Suzuki, M. T., & Giovannoni, S. J. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, *62*(2), 625-630. https://doi.org/Doi 10.1128/Aem.62.2.625-630.1996

Takai, K., Mormile, M. R., McKinley, J. P., Brockman, F. J., Holben, W. E., Kovacik Jr, W. P., & Fredrickson, J. K. (2003). Shifts in archaeal communities associated with lithological and geochemical variations in subsurface Cretaceous rock. *Environmental Microbiology*, *5*(4), 309-320. https://sfamjournals.onlinelibrary.wiley.com/doi/pdfdirect/10.1046/j.1462-2920.2003.00421.x?download=true

Taniguchi, M., Burnett, W. C., Cable, J. E., & Turner, J. V. (2002). Investigation of submarine groundwater discharge. *Hydrological Processes*, *16*(11), 2115-2129.

Team, R. C. (2020). *R: A language and environment for statistical computing*. In https://www.R-project.org/

Tebo, B. M., Johnson, H. A., McCarthy, J. K., & Templeton, A. S. (2005). Geomicrobiology of manganese(II) oxidation. *Trends Microbiol*, *13*(9), 421-428. https://doi.org/10.1016/j.tim.2005.07.009

Teeling, H., & Glockner, F. O. (2012). Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. *Brief Bioinform*, *13*(6), 728-742. https://doi.org/10.1093/bib/bbs039

Thomas, T., Gilbert, J., & Meyer, F. (2012). Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp*, *2*(1), 3. https://doi.org/10.1186/2042-5783-2-3

Torsvik, V., Øvreås, L., & Thingstad, T. F. (2002). Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science*, *296*(5570), 1064-1066. https://science.sciencemag.org/content/sci/296/5570/1064.full.pdf

Tringe, S. G., Von Mering, C., Kobayashi, A., Salamov, A. A., Chen, K., Chang, H. W., Podar, M., Short, J. M., Mathur, E. J., & Detter, J. C. (2005). Comparative metagenomics of microbial communities. *Science*, *308*(5721), 554-557.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., & Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, *428*(6978), 37-43. https://doi.org/10.1038/nature02340

Urbach, E., Vergin, K. L., Young, L., Morse, A., Larson, G. L., & Giovannoni, S. J. (2001). Unusual bacterioplankton community structure in ultra-oligotrophic Crater Lake. *Limnology and Oceanography*, *46*(3), 557-572.

Venables, W. N., & Ripley, B. D. (2013). *Modern applied statistics with S-PLUS*. Springer Science & Business Media.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., & Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, *304*(5667), 66-74. https://doi.org/10.1126/science.1093857

Vigneron, A., Cruaud, P., Langlois, V., Lovejoy, C., Culley, A. I., & Vincent, W. F. (2020). Ultra-small and abundant: Candidate phyla radiation bacteria are potential catalysts of carbon transformation in a thermokarst lake ecosystem. *Limnology and Oceanography Letters*, *5*(2), 212-220.

Vit, C., Richard, E., Fournes, F., Whiteway, C., Eyer, X., Lapaillerie, D., Parissi, V., Mazel, D., & Loot, C. (2021). Cassette recruitment in the chromosomal Integron of Vibrio cholerae. *Nucleic Acids Res*, *49*(10), 5654-5670. https://doi.org/10.1093/nar/gkab412

Vlasceanu, L., Popa, R., & Kinkle, B. K. (1997). Characterization of Thiobacillus thioparus LV43 and its distribution in a chemoautotrophically based groundwater ecosystem. *Appl Environ Microbiol*, *63*(8), 3123-3127.

Walters, W. A., Caporaso, J. G., Lauber, C. L., Berg-Lyons, D., Fierer, N., & Knight, R. (2011). PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics*, *27*(8), 1159-1161. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3072552/pdf/btr087.pdf

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*, *73*(16), 5261-5267. https://doi.org/10.1128/AEM.00062-07

Ward, M. H., deKok, T. M., Levallois, P., Brender, J., Gulis, G., Nolan, B. T., VanDerslice, J., & International Society for Environmental, E. (2005). Workgroup report: Drinking-water nitrate and health--recent findings and research needs. *Environ Health Perspect*, *113*(11), 1607-1614. https://doi.org/10.1289/ehp.8043

Watanabe, K., Kodama, Y., & Harayama, S. (2001). Design and evaluation of PCR primers to amplify bacterial 16S ribosomal DNA fragments used for community fingerprinting. *Journal of Microbiological Methods*, *44*(3), 253-262.

Water Protection Ordinance, Swiss Federal Law (1998).

Weinstein, Y., Yechieli, Y., Shalem, Y., Burnett, W. C., Swarzenski, P. W., & Herut, B. (2011). What is the role of fresh groundwater and recirculated seawater in conveying nutrients to the coastal ocean? *Environmental science & technology*, *45*(12), 5195-5200.

White, W. B. (1988). *Geomorphology and hydrology of karst terrains*.

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

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *Proceedings of the National Academy of Sciences*, *95*(12), 6578-6583. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC33863/pdf/pq006578.pdf

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of statistical software*, *21*(12), 1-20.

Wickham, H. (2009). *Ggplot2 : elegant graphics for data analysis*. Springer. Table of contents http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&doc_number=017387312&line_number=0001&func_code=DB_RECORDS&service_type=MEDIA

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of statistical software*, *40*(1), 1-29.

Wilhartitz, I. C., Kirschner, A. K., Stadler, H., Herndl, G. J., Dietzel, M., Latal, C., Mach, R. L., & Farnleitner, A. H. (2009). Heterotrophic prokaryotic production in ultraoligotrophic alpine karst aquifers and ecological implications. *FEMS Microbiol Ecol*, *68*(3), 287-299. https://doi.org/10.1111/j.1574-6941.2009.00679.x

Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K. P., Paczian, T., Trimble, W. L., Bagchi, S., Grama, A., Chaterji, S., & Meyer, F. (2016). The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res*, *44*(D1), D590-594. https://doi.org/10.1093/nar/gkv1322

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

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

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput Biol*, *6*(2), e1000667. https://doi.org/10.1371/journal.pcbi.1000667

Wu, S., Zhu, Z., Fu, L., Niu, B., & Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics*, *12*(1), 1-9.

Yanagawa, K., Nunoura, T., McAllister, S., Hirai, M., Breuker, A., Brandt, L., House, C., Moyer, C. L., Birrien, J.-L., & Aoike, K. (2013). The first microbiological contamination assessment by deep-sea drilling and coring by the D/V Chikyu at the Iheya North hydrothermal field in the Mid-Okinawa Trough (IODP Expedition 331). *Frontiers in microbiology*, *4*, 327. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3820981/pdf/fmicb-04-00327.pdf

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., & Glöckner, F. O. (2014). The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic acids research*, *42*(D1), D643-D648.

Yu, H., & Leadbetter, J. R. (2020). Bacterial chemolithoautotrophy via manganese oxidation. *Nature*, *583*(7816), 453-458. https://doi.org/10.1038/s41586-020-2468-5

APPENDIX

Table A.1. Estimated cell counts in other aquatic environments, Chapter 3

Aquatic Environment	**Depth (m below surface)**	**Typical bacterial abundance (cells/mL)**	**Reference**
Open Ocean	Various depths	$1.0 \times 10^4 – 1.0 \times 10^7$	Whitman *et al*., 1998
Screened well, San Juan Basin, New Mexico	182 -190	2×10^6	Takai *et al*., 2003
CO_2 Sink Reservoir, Ketzin, Germany	647	$2 - 6 \times 10^6$	Myrttinen *et al*., 2010
Eutrophic River Warnow	0	2.4×10^6	Freese *et al*., 2006
North Atlantic	5 -200	$8.2 \times 10^5 – 2.4 \times 10^6$	Rowe *et al*., 2012
West Pacific	5 - 200	$2.9 \times 10^5 -1.2 \times 10^6$	Rowe *et al*. 2012
Crater Lake	0 - 200	$2.0 \times 10^5 – 1 \times 10^6$	Urbach *et al*., 2001
Sedimentary Rock Borehole, Hokkaido, Japan	0 - 482	$4.61 \times 10^4 - 5.06 \times 10^6$	Kato *et al*., 2009
Sargasso Sea	5 - 200	$4.6 - 8.8 \times 10^5$	Rowe *et al*. 2012
Bangomb site, Gabon, Africa	5 - 105	$4.5 \times 10^4 - 5.8 \times 10^5$	Pedersen *et al*., 1996
Artesian Well, Paris, France	800	$1.0 \times 10^4 - 2.5 \times 10^5$	Basso *et al*., 2005
Swiss Cave Pool	950	5.2×10^5	Shabarova and Pernthaler, 2010
Limestone Karst Aquifer Spring	0	6.8×10^4	Farnleitner *et al*. 2005
Dolomite Karst Aquifer Spring	0	1.5×10^4	Farnleitner *et al*. 2005
Fault-bordered aquifer, Northern Japan	550	3×10^3	Shimizu *et al*., 2006
WICA lakes	200	2.3×10^3	This study
Lake Vostok	1500 - 2750	$2.0 \times 10^2 -1.0 \times 10^3$	Karl *et al*. 1999
Dolomite, igneous rock, South African Mines	1700 - 3600	$2.0 \times 10^2 - 3.4 \times 10^3$	Borgonie *et al*., 2011

Table A.2. Wind Cave metagenomic annotation statistics

Sample	IMG Genome ID	Genome Size (Mb)	Gene Count	GC %	CDS Count	CDS %	COG Count	COG %	Seed Count	Seed %
Wind Cave, Large fraction, 2017	*	321.73	494833	54.89	489019	98.83	253883	51.92	65207	13.33
Wind Cave, unfractionated, 2017 (JGI)	3300021357	320.06	1216485	45.79	1199541	98.61	544905	45.43	158892	13.25
Wind Cave, Large fraction, 2018	*	293.88	440063	57.85	434709	98.78	263806	60.69	55350	12.73
Wind Cave, unfractionated, 2018	*	87.00	135897	55.42	134051	98.64	79985	59.67	18844	14.06

*Will be made available at Barton Lab GitHub

Table A.3. Metagenomic annotation statistics for other Cave biomes examined

Sample	IMG Genome ID	Genome Size (Mb)	Gene Count	GC %	CDS Count	CDS %	COG Count	COG %	Seed Count	Seed %
Carbonate Cave, Grottes des Collemboles, Begium*	3300005781	121.59	8556468	62.64	8515791	99.52	3565507	41.67	63807	0.75*
Iron Ore cave, brazil	3300021476	564.26	1969152	60.41	1960276	99.55	929499	47.20	268697	13.71
Wishing Well Cave, Virginia	3300031576	2021.13	4912229	59.30	4877777	99.30	2385555	48.56	454679	9.32
Kartchner Caverns	2209111007	195.24	555572	65.80	552953	99.53	252936	45.53	101339	18.33
	2199352033	138.88	384714	62.44	382582	99.45	170993	44.45	58573	15.31
	2199352032	168.97	488539	61.76	485931	99.47	218215	44.67	81063	16.68
	2199352031	181.90	529910	62.44	527210	99.49	228371	43.10	80194	15.21
	2189573024	138.50	365407	58.98	363306	99.43	170302	46.61	70017	19.27
	3300004454	136.51	424662	58.99	422609	99.52	199055	46.87	68224	16.14

*unassembled

Table A.4. Metagenomic annotation statistics for Groundwater biomes examined

Sample	IMG Genome ID	Genome Size (Mb)	Gene Count	GC %	CDS Count	CDS %	COG Count	COG %	Seed Count	Seed %
Cold Creek Source, Nevada	3300025123	607.60	1070400	52.48	1064648	99.46	422441	39.47	176499	16.58
Devils Hole, Nevada	3300025765	23.83	101703	58.79	100941	99.25	47343	46.55	16477	16.32
Ash Meadows Crystal Spring, Nevada	3300025130	642.08	1179482	59.01	1171040	99.28	633860	53.74	237649	20.29
Horonobe Underground Research Laboratory, Japan	3300029821	268.48	352146	49.47	348013	98.83	222594	63.21	30201	8.68
Sanford Underground Research Facility (SURF), South Dakota	3300007352	513.21	815222	54.17	807594	99.06	389944	47.83	85567	10.60
	3300007354	279.82	478086	53.08	473855	99.12	234652	49.08	53520	11.29

Table A.5. Metagenomic annotation statistics for Freshwater biomes examined

Sample	IMG Genome ID	Genome Size (Mb)	Gene Count	GC %	CDS Count	CDS %	COG Count	COG %	Seed Count	Seed %
Glacier valley Borup Fiord, Nunavut, Canada	3300025140	734.59	1024020	52.41	1017046	99.32	533667	52.11	160697	15.80
Lake La Cruz, Castile-La Mancha, Spain	3300028569	760.84	1458059	50.21	1445315	99.13	648956	44.51	150483	10.41
Oligotrophic Sparkling Lake, Wisconsin, USA	3300018815	374.32	1160925	57.04	1152371	99.26	708308	61.01	333551	28.94
Lake Erie, Canada	3300027870	1226.94	3394051	59.44	3375379	99.45	1966471	57.94	860231	25.49
Lake Tanganyika, Tanzania	3300020083	1362.07	3751598	55.93	3722470	99.22	1805666	48.13	429729	11.54
	3300020200	1080.30	2656613	53.82	2635641	99.21	1063291	40.02	244452	9.27

Table A.6. Metagenomic annotation statistics for Deep Ocean Subsurface biomes examined

Sample	IMG Genome ID	Genome Size (Mb)	Gene Count	GC %	CDS Count	CDS %	COG Count	COG %	Seed Count	Seed %
South Atlantic Ocean, Benguela	3300024263	695.06	2254741	46.28	2237440	99.23	1200063	53.22	223792	10.00
Anholt, Denmark	3300024353	801.52	1233515	49.82	1222993	99.15	655552	53.15	112130	9.17
Black Sea	3300024433	1387.49	2170781	46.71	2155214	99.28	830338	38.25	141341	6.56
South Pacific Ocean, Chile	3300024429	1282.10	2084361	45.30	2067776	99.20	1080099	51.82	220933	10.68
Indian Ocean, Sumatra	3300024432	663.81	2422292	42.98	2405592	99.31	906837	37.44	170329	7.08
	3300024265	1672.78	3241927	44.81	3216648	99.22	1231436	37.98	266352	8.28

Table A.7. Metagenomic annotation statistics for Deep Oceanic Trench biomes examined

Sample	IMG Genome ID	Genome Size (Mb)	Gene Count	GC %	CDS Count	CDS %	COG Count	COG %	Seed Count	Seed %
Kermadec Trench	3300024060	846.63	2663714	55.58	2649608	99.47	1307406	49.08	333281	12.58
Mariana Trench	3300024058	1619.15	2560110	54.55	2547840	99.52	1248475	48.77	336976	13.23
	3300024516	1425.61	2618447	54.37	2605477	99.50	1195719	45.67	330022	12.67