12-2015

# Bacterial diversity and function within an epigenic cave system and implications for other limestone cave systems

Kathleen Merritt Brannen-Donnelly
*University of Tennessee - Knoxville*, kbrannen@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Kathleen Merritt Brannen-Donnelly entitled "Bacterial diversity and function within an epigenic cave system and implications for other limestone cave systems." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Geology.

Annette S. Engel, Major Professor

We have read this dissertation and recommend its acceptance:

Terry Hazen, Steven Wilhelm, Andrew Steen, Larry McKay

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

# Bacterial diversity and function within an epigenic cave system and implications for other limestone cave systems

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Kathleen Merritt Brannen-Donnelly
December 2015

**ACKNOWLEDGEMENTS**

I have had advice, assistance, and encouragement from many people during the preparation of this dissertation. I must first thank my family for insisting that a bachelor's degree was the only path out of high school, and for all of their continued support during graduate school. I would also like to thank my husband, Brendan, for his support, because he supported my ideals of independence and equality long before we were married. Brendan knew that graduate school was the only path for me, and together we have traveled down the long and winding path that is graduate school. I would also like to thank all of the graduate students in my research group: Sarah Keenan, Brendan Headd, Terri Brown, Chanda Drennan, Caroline Dietz, Walt Doty, and Aaron Geomann. This group of students has taught me so much about science, politics, and group dynamics. They also made research more exciting, and provided many moments that I will remember forever. Thanks also to my lab manager Audrey Paterson who helped with organization of sample trips as well as assistance sampling. Besides my research group, there are many people that assisted me with cave sampling, which is not everyone's idea of vacation time: A.J. Campion, Kevin Kissell, Jared Apuuli, and Scott Engel. Thanks to Coy Aiken and Sam Plummer, rangers at the Carter Caves State Resort Park, for access to Cascade Cave System. Also thanks to Sam Plummer for providing precipitation data. I would also like to thank several people who helped me with coding in various languages: Steve Techtmann, Hannah Woo, Junqi Yin, Drew Steen, and Mike Cheng. I would like to thank my committee members, Larry McKay, Terry Hazen, Steven Wilhelm, and Drew Steen, who helped shape my research ideas into obtainable goals. Lastly, I would like to thank my advisor Annette Summers Engel for her support. Annette inspired me to be a scientist during my undergraduate. I have had many great experiences and opportunities because of Annette's support during both my undergraduate and graduate career.

**ABSTRACT**

There are approximately 48,000 known cave systems in the United States of America, with caves formed in carbonate karst terrains being the most common. Epigenic systems develop from the downward flow of meteoric water through carbonate bedrock and the solutional enlargement of interconnected subsurface conduits. Despite carbonate karst aquifers being globally extensive and important drinking water sources, microbial diversity and function are poorly understood compared to other Earth environments. After several decades of research, studies have shown that microorganisms in caves affect water quality, rates of carbonate dissolution and precipitation, and ecosystem nutrition through organic matter cycling. However, limited prior knowledge exists for the most common system, epigenic caves, regarding microbial taxonomic diversity, their metabolic capabilities, and how community function changes during and following environmental disturbances. To evaluate community development and succession, as well as potential roles in organic matter cycling, bacteria from the Cascade Cave System (CCS) in Kentucky were investigated. From geochemical and metagenomic data collected during a five-month colonization experiment, taxonomically distinct planktonic and sediment-attached bacterial communities formed along the epigenic cave stream. This represents one of the largest metagenomic studies done from any cave. Betaproteobacteria, Gammaproteobacteria, Alphaproteobacteria, and Opitutae were the most abundant groups. Planktonic bacteria pioneered sediment-attached communities, likely attributed to functional differences related to cell motility and attachment. Organic matter cycling affected exogenous heterotrophic community composition and function downstream

because of diminished organic matter quality over time. This was reflected in significantly different abundances of genes encoding for carbohydrate and lignin degradation between habitats and depending on cave location. The ubiquity of environmental controls on bacteria functional diversity in karst is unknown because these environments have generally been left out of microbial biogeography research. In spatial meta-analyses of bacterial diversity data from global cave systems, the ubiquity of some bacteria in karst is evident. Despite evidence for undersampling and difficulties comparing sequencing technologies and strategies, some caves appear to have novel lineages while other caves have taxonomically similar communities despite being 1000s of kilometers apart. The implications are that microbes in karst (i.e., carbonate) caves around the world are functionally comparable.

**TABLE OF CONTENTS**

**LIST OF TABLES**

**LIST OF FIGURES**

# Chapter 1 INTRODUCTION

Karst landscapes comprise up to 20% of the Earth's dry land surface, which largely coincides with the distribution of carbonate sedimentary rocks (Ford and Williams, 2013). Karst is distinguished from other bedrock terrane because of the movement of meteoric water from the surface into the subsurface through self-evolving, diffuse or conduit flow systems that develop from the dissolution of soluble rock by slightly acidic, usually $CO_2$-charged solutions (Palmer, 2007). Because of the ability to store and transport vast quantities of water, an estimated 25% of global drinking water is sourced from karst aquifers (Ford and Williams, 2013). In the United States of America, there are approximately 48,000 known cave systems (Culver et al., 1999; Palmer, 2007). Globally, the number of caves can easily be estimated to reach over a million.

Hydrological connectivity between the surface and subsurface means that allochthonous (i.e., surface-derived, terrigenous) headwater streams deliver water, nutrients, and organic matter (OM) into the subsurface (Brooks et al., 1999; Simon et al., 2007). The absence of sunlight excludes photosynthetic primary production in caves, and although some systems have multi-trophic level ecosystems supported by chemolithoautotrophy (Sarbu et al., 1996; Engel et al., 2004; Chen et al., 2009), the majority of cave ecosystems have been shown to rely on allochthonous OM transported from the surface (Simon et al., 2003; Simon et al., 2007; Lee et al., 2012). Depending on the hydrological connectivity of a cave to the surface, the types and abundance of allochthonous nutrients can vary over time within the same cave, as well as between caves. This variation in allochthonous input has been hypothesized to influence cave species biomass and diversity (Simon and Benfield, 2001; Cooney and Simon, 2009;

Huntsman et al., 2011; Venarsky et al., 2012). However, these previous studies did not statistically evaluate organic matter abundance and breakdown rates among cave streams, which suggests that there is a more complex relationship between the influx of organic matter and the processes and rates at which the organic matter is broken down.

A review of microbial diversity in shallow groundwater systems (Griebler and Lueders, 2009) states that subsurface microbial communities are distinct from those found in soil and surface waters, not because of the presence of endemic groundwater microbial species, but because of the "specific phylogenetic composition of groundwater microbial communities and by their special physiological capabilities." If shallow groundwater environments are subject to influxes of both allochthonous microbial communities and allochthonous inorganic and organic nutrients, then the differences between these connected surface and subsurface environments are (1) extended periods of darkness in the subsurface depending on the length of the subsurface flowpath, (2) a cutoff point for allochthonous inputs (depending on the hydrologic connectivity of the subsurface to the surface), and (3) the possible presence of endemic subsurface macrofauna.

Despite general microbial ecology studies in groundwater and karst environments (Griebler and Lueders, 2009; Engel, 2010; Lee et al., 2012; Engel, 2015), and total organic carbon (TOC), dissolved (DOM), and particulate organic matter (POM) assessments from karst systems (Graening and Brown, 2003; Simon et al., 2003; Farnleitner et al., 2005; Simon et al., 2007; Birdwell and Engel, 2010; Simon et al., 2010), there has been limited knowledge regarding the controls that diverse microbial groups have on the nature of OM, carbon, and

nutrients in groundwater-dependent ecosystems. In contrast, the marine microbial diversity and the oceanic carbon cycle are far better understood (Azam et al., 1994), likely due in part to the ocean being a large accessible reservoir, and the ocean's role in climate change because of its capacity to uptake anthropogenic carbon (Jiao et al., 2010). Although microbial communities in cave streams rely on the input of allochthonous inorganic and organic nutrients, the dogma has been that pristine cave systems support stable subsurface ecosystems (Goldscheider et al., 2006; Griebler and Lueders, 2009). Less attention has been given to understanding which microorganisms are living in cave systems that rely on allochthonous OM, what these microorganisms are doing, and how microbial diversity and function change over time (Lee et al., 2012; Engel, 2015).

The goals of this dissertation were: (1) Survey the known cave bacterial diversity using 16S rRNA genes obtained from an open access databases to understand trends in bacterial diversity in cave systems around the world, as well as to place the bacterial diversity of the study cave–Cascade Cave System, Kentucky–in biological and ecological context with other cave systems; (2) Survey the bacterial diversity of the Cascade Cave System over time, as well as changes in aqueous geochemistry, flood disturbances, and sediment mobilization, to understand how the bacterial community responded to environmental disturbances, specifically cave flooding, and how community sucession may be initated following disturbances; (3) Assess functional differences between planktonic and sediment-attached microbes based on habitat types inside the cave and evaluate functional capabilities in response to environmental disturbances. Details about the geology and hydrology of the

Cascade Cave System are in Chapter 3. Overall, as one of the largest meta-analyses of 16S rRNA gene data from caves, and the largest metagenomic study to date, this dissertation adds key knowledge about the distribution and function of bacteria in epigenic caves.

**REFERENCES**


Azam, F., Smith, D., Steward, G., and Hagström, Å. (1994). Bacteria-organic matter coupling and its significance for oceanic carbon cycling. *Microbial Ecology* 28**,** 167-179.

Birdwell, J.E., and Engel, A.S. (2010). Characterization of dissolved organic matter in cave and spring waters using UV–Vis absorbance and fluorescence spectroscopy. *Organic Geochemistry* 41**,** 270-280. doi: 10.1016/j.orggeochem.2009.11.002.

Brooks, P.D., Mcknight, D.M., and Bencala, K.E. (1999). The relationship between soil heterotrophic activity, soil dissolved organic carbon (DOC) leachate, and catchment-scale DOC export in headwater catchments. *Water Resources Research* 35**,** 1895-1902. doi: 10.1029/1998wr900125.

Cooney, T.J., and K.S. Simon. (2009). Influence of dissolved organic matter and invertebrates on the function of microbial films in groundwater*. Microbiol Ecology* 58(3): 599–610.

Chen, Y., Wu, L., Boden, R., Hillebrand, A., Kumaresan, D., Moussard, H., Baciu, M., Lu, Y., and Colin Murrell, J. (2009). Life without light: microbial diversity and evidence of sulfur- and ammonium-based chemolithotrophy in Movile Cave. *ISME J* 3**,** 1093-1104. doi: 10.1038/ismej.2009.57.

Culver, D. C., Hobbs III, H. H., Christman, M. C., & Master, L. L. (1999). Distribution map of caves and cave animals in the United States. *Journal of Cave and Karst Studies*, 61(3), 139-140.

Engel, A.S. (2010). Microbial diversity of cave ecosystems, in Barton, L., Mandl, M., and Loy, A. (eds.), *Geomicrobiology: Molecular & Environmental Perspectives*, Springer. p. 219-238. DOI:10.1007/978-90-481-9204-5_10.

Engel, A.S. (2015). Bringing microbes into focus for cave science: An introduction, in A.S. Engel (ed.), *Microbial Life of Cave Systems,* De Gruyter. p. 1-22.

Engel, A.S., Stern, L.A., and Bennett, P.C. (2004). Microbial contributions to cave formation: New insights into sulfuric acid speleogenesis. *Geology* 32**,** 369. doi: 10.1130/g20288.1.

Farnleitner, A.H., Wilhartitz, I., Ryzinska, G., Kirschner, A.K., Stadler, H., Burtscher, M.M., Hornek, R., Szewzyk, U., Herndl, G., and 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**,** 1248-1259. doi: 10.1111/j.1462-2920.2005.00810.x.

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

Goldscheider, N., Hunkeler, D., and Rossi, P. (2006). Review: Microbial biocenoses in pristine aquifers and an assessment of investigative methods. *Hydrogeology Journal* 14**,** 926-941.

Graening, G. O., and Brown, A. V. Ecosystem dynamics and pollution effects in an Ozark cave stream. *Journal of the American Water Resources Association* 39, 1497–1507. doi: 10.1111/j.1752-1688.2003.tb04434.x

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

Huntsman, B.M., M.P. Venarsky, and J.P. Benstead. (2011). Relating carrion breakdown rates to ambient resource level and community structure in four cave stream ecosystems. *Journal North American Benthological Society* 30(4): 882–892.

Jiao, N., Herndl, G.J., Hansell, D.A., Benner, R., Kattner, G., Wilhelm, S.W., Kirchman, D.L., Weinbauer, M.G., Luo, T., Chen, F., and Azam, F. (2010). Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Reviews Microbiology* 8**,** 593-599. doi: 10.1038/nrmicro2386.

Lee, N.M., Meiginger, D.B., Aubrecht, R., Kovacic, L., Saiz-Jimenez, C. Baskar, S.R., Schleifer, K.-H., Liebl, W., Porter, M.L., Engel, A.S. (2012) *Caves and karst environments*, in Bell, E. (ed.) Life at Extremes: Environments, Organisms and Strategies for Survival. CAB Interantioanl Publishing. p. 320-344.

Palmer, A.N. (2007). *Cave Geology.* Cave books Dayton.

Sarbu, S.M., Kane, T.C., and Kinkle, B.K. (1996). A chemoautotrophically based cave ecosystem. *Science* 272**,** 1953.

Simon, K., Benfield, E., and Macko, S. (2003). Food web structure and the role of epilithic biofilms in cave streams. *Ecology* 84**,** 2395-2406.

Simon, K.S., Pipan, T., and Culver, D.C. (2007). A conceptual model of the flow and distribution of organic carbon in caves. *Journal of Cave and Karst Studies* 69**,** 279-284.

Simon, K.S., Pipan, T., Ohno, T., and Culver, D.C. (2010). Spatial and temporal patterns in abundance and character of dissolved organic matter in two karst aquifers. *Fundamental and Applied Limnology / Archiv für Hydrobiologie* 177**,** 81-92. doi: 10.1127/1863-9135/2010/0177-0081.

Venarsky, M.P., J.P. Benstead, and A.D. Huryn. 2012. Effects of organic matter and season on leaf litter colonisation and breakdown in cave streams*. Freshwater Biol*ogy 57(4): 773–786.

# Chapter 2 KNOWN BACTERIAL COMMUNITY DIVERSITY OF CAVES

Some of the results in this chapter will be submitted for review and potential publication. KBD and ASE designed the study, and KBD collected, analyzed, and interpreted the data. KBD and ASE wrote the manuscript.

## INTRODUCTION

Subsurface habitats have abundant microorganisms, with abundances and biomass that likely exceeds values estimated for other Earth environments (McMahon and Parnell, 2014). Karst landscapes formed in carbonate bedrock are one type of subsurface habitat that are characterized by the rapid transfer of surface water into the subsurface through hydraulic flow systems consisting of sinkholes, caves, and springs (Ford & Williams, 2013). These surface-to-subsurface flow systems develop from the dissolution of soluble rock, such as limestone, by slightly acidic water. The hydrological connectivity between the surface and subsurface allows headwater streams to deliver important allochthonous (i.e., surface-derived) nutrients and organic matter (OM) to the subsurface (e.g., Brooks et al., 1999; Jardine et al., 2006). Karst systems can vary from a few square meters to hundreds of square kilometers in aerial extent, and from a single, small cave system <10 m long to regional 100s km of extensive, complex passageways forming a karst aquifer. In general, karst landscapes and processes are geologically, hydrologically, and geochemically homogeneous, and cover up to 20% of the Earth's terrestrial, ice-free surface (Ford & Williams 2013).

Overall, there is less information about the microbial ecology of karst groundwater compared to other types of groundwater-dependent ecosystems (Griebler and Lueders, 2009).

From genetics and microbial ecology studies, some karst cave systems with slow water flow and long retention times host stable autochthonous microbial communities (Farnleitner et al., 2005; Pronk et al., 2008; Wilhartitz et al., 2009), while other karst cave systems do not host a stable microbial community due to rapid water flow and environmental disturbances (Brannen-Donnelly and Engel, 2015). There are also cave ecosystems that can support multiple trophic levels via primary production by chemolithoautotrophic microorganisms (Sarbu et al., 1996; Macalady et al., 2006; Engel et al., 2010). The most common type of cave ecosystems are dependent upon surficial influx of dissolved organic matter, and are considered energy limited (Simon et al., 2007; Simon et al., 2010; Zhou et al., 2012; Venarsky et al., 2014). In general, although microbial cycling of solutes is known to be a key component of biogeochemical cycling that is responsible for water quality in the subsurface (Spizzico et al., 2005), metabolic processes and turnover rates of nutrients by microorganisms are still largely unknown at the scale of a karst aquifer (Simon et al., 2007). Moreover, there is no consensus or model of the flow and distribution of carbon or other nutrients in karst cave systems (Simon et al., 2003; Simon et al., 2007; Hallbeck and Pedersen, 2008; Griebler and Lueders, 2009).

One of the reasons why karst environments are not as well-studied as other subsurface systems is that the vast majority of cave systems worldwide have not been found (Lee et al., 2012). Few of the known cave systems have characterized microbial diversity or function (Lee et al., 2012). These environments are physically difficult to access. Karst aquifers have limited accessibility though wells and boreholes, and samples may be contaminated by the drilling process (Engel and Northup, 2008). Alternatively, using caves as access points into subsurface

karst environments allows for less risk of contamination, but also allows for different types of cave habitats to be explored and sampled (e.g., planktonic versus attached microbial communities), and also allows for *in situ* experimentation (Engel and Northup, 2008).

Caves are defined as any subsurface void that a human is physically able to enter, where at least some part of it is completely dark (Palmer, 2007). Caves are a dominant subsurface feature in karst environments, and there are at least 48,000 known caves in the United States of America (Culver et al., 1999; Palmer, 2007). Even in well explored areas in North America and Europe, it is estimated that only 50% of known caves have been studied and only 10% of caves around the world have been found (Lee et al., 2012).

GenBank, that the first open access repository for DNA sequences by the National Center for Biotechnology Information (NCBI) in the U.S., was started in 1982 There have been 197 published studies (some not formerly published, but deposited in GenBank) using molecular genetics methods to describe microbial (mostly bacterial) diversity from 16S rRNA genes retrieved from caves and karst settings since the first published study by Vlasceanu et al. (1997) (Engel, 2015). The rate of publication using newer, culture-independent has not increased in comparison to the rate of publications of culture-dependent methods (Engel, 2015). This is because the focus of many cave diversity studies was to evaluate the roles of microorganisms on either passive or active precipitation and dissolution of carbonate minerals (Lee et al., 2012). Moreover, despite the number of studies available via open access databases, the number of studies that have comparable alpha- and beta- diversity is relatively low (Engel et al., 2010; Lee et al., 2012), mostly because methods used to obtain gene sequences, and also

gene sequencing technologies, have changed over time. But, comparing microbial diversity across different cave environments, such as from carbonate caves versus basaltic lava tubes, is essential to understand key relationships between microorganisms and their environment underground (Griebler and Lueders, 2009).

The purpose of this chapter was to review the cave and karst 16S rRNA gene sequence information from the NCBI GenBank database and to conduct a meta-analysis of the genetic data to improve upon our understanding of bacterial diversity in caves and karst. Thus far, and to my knowledge, there have been no published meta-analyses of available 16S rRNA gene sequences from cave and karst systems. The previously published and available 16S rRNA gene sequences were clustered and compared with newly obtained sequences from the Cascade Cave System, Kentucky (details provided in Chapter 3), to evaluate whether carbonate caves have unique bacterial community compositions and to better understand the distribution of bacteria in carbonate caves compared to other subsurface environments. This was also done to understand how comparable the bacterial communities from Cascade Cave were in the context of other studied limestone caves worldwide.

## METHODS

### Collection and analysis of open access sequences

Sequence libraries from the NCBI GenBank database (Benson et al., 2013) were downloaded[1]. Sequences from the NCBI GenBank database were trimmed of their barcodes and primers, if they were present, using QIIME and cutadapt (Caporaso et al., 2010; Martin, 2011). All libraries were quality screened (Caporaso et al., 2010). Genbank sequences were combined with sequences from Cascade Cave System (see Chapter 3) and clustered into operational taxonomic units (OTUs) based on 97% sequence similarity (Kunin et al., 2010) using uclust through QIIME (Crawford et al., 2009; Caporaso et al., 2010; Edgar et al., 2011). Analyses in QIIME were performed on "Blacklight," an SGI Altix UV1000 shared-memory machine with Intel Xeon 7500 processers, through the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE) partnership (Towns et al., 2014).

From the 23,658 OTUs generated for the full dataset (501,926 16S rRNA gene reads; Table A2-1, all tables located in Appendix I), representative sequences were chosen for classification by the Ribosomal Database Project (RDP) Classifier at 80% confidence intervals using QIIME (Wang et al., 2007). The majority of all OTUs were classified to the class level (88%). Following sequence alignment with pyNast, a phylogenetic tree was generated in QIIME using the maximum likelihood fasttree method (Price et al., 2009; Caporaso et al., 2010). The

[1] based on the search function "bacteria AND 16S AND (cave OR karst OR aquifer OR groundwater OR mine OR lava) AND 300 : 1000[SLEN] NOT soil NOT river NOT coal NOT tailings NOT potassium NOT drainage NOT landfill NOT (whole AND genome) NOT Animal NOT Fungi NOT Archaea"

relative abundances of sequences affiliated with major phyla and candidate divisions were calculated, and reported as percentages, based on incidence (as presence/absence) compared to the total number of sequences obtained per library. Statistical analyses were based on the UniFrac distance metric (Lozupone & Knight, 2005) that used the phylogenetic tree, which only represented 21,795 OTUs after alignment. UniFrac is more powerful than the nonphylogenetic distance measures because it uses different degrees of similarity between sequences (Lozupone and Knight, 2005).

**Statistical analyses**

Shannon diversity (H') and Chao1 indices were calculated for each of the cave libraries obtained using the package phyloseq (version 1.10.0) in the computer program R (McMurdie and Holmes, 2013). Higher numbers for both indices indicate greater OTU-level richness and evenness. To determine if the distribution of OTUs between samples could be attributed to the sample environment type, a permutational multivariate analysis of variance was performed using the adonis function with 9999 permutations in the vegan package in R (Oksanen et al., 2013). To determine if proteobacterial classes varied by environment type, an analysis of variance (ANOVA) was performed on the abundance of sequences per proteobacterial class. A principal coordinates analysis (PCoA) was done on a weighted UniFrac distance metric to represent the similarity of libraries based on their distribution of OTUs, using the phyloseq package in R (Lozupone and Knight, 2005; McMurdie and Holmes, 2013). Cluster analyses can be used to find groups within data without the help of a response variable (Tibshirani et al.,

2001). A cluster analysis was used on the phylogenetic tree because no common metadata were available for all libraries. The goodness of clustering measure for the number of clusters in the phylogenetic tree was calculated using the R package cluster (Tibshirani et al., 2001). Code and details regarding data processing are provided in Code I, located in Appendix IV.

A distance plot was created using a Jaccard distance metric with the phylogeo package in R to calculate the ecological and geographic distances between libraries and to analyze the spatial relationships between libraries (Charlop-Powers and Brady, 2015). To interpret the distribution of estimated clusters across spatial distance, k-means clustering was performed and individually mapped to show the location and abundance of the sequences in each cluster using phylogeo in R implemented using Blacklight (Towns et al., 2014; Charlop-Powers and Brady, 2015).

## RESULTS

Only 0.05% of the over 9.6 million 16S rRNA bacterial gene sequences in the NCBI GenBank database could be attributed to cave and karst settings. As a comparison, the number of cave and karst sequences from the European Nucleotide Archive (ENA) database was 0.08% of all sequences. The number of gene sequences from any terrestrial subsurface environment was also low, with GenBank at 0.08% and ENA at 0.23%. For the meta-analysis, a total of 98,715 sequences after quality screening were obtained from 93 sample libraries, with each library being grouped into 11 cave and karst environment types based on bedrock lithology or dominant aqueous chemistry (Table A2-1). At least 45 caves were represented, as well as karst

springs and samples from karst aquifer wells. Metadata were incomplete, with 67 libraries having published information about the site or sampling and sequencing methods. Differentiated environment types included: caves formed in dolomite, “DS_cave”; caves formed in ice, “Ice_cave”; general subsurface karst environments, “karst”; karst springs, “karst_spring”; basaltic lava tubes, “lava_tube”; caves formed in limestone “LS_cave”; caves with the presence of gypsum or the aqueous presence of any reduced sulfur species (such as $H_2S$), “sulfur_cave”; karst with the presence of gypsum or presence of reduced sulfur species (such as $H_2S$), “sulfur_karst”; spring samples of unknown lithology, “surface_spring”; and samples whose geologic location is unknown, “unknown.”

All but one of the sample libraries were generated from one sample collected from a single cave and 16S rRNA genes were sequenced following PCR amplification, shot-gun cloning, and Sanger chain-termination sequencing genes from isolated clones with correctly sized rRNA gene fragments (see Table A2-1 for list of references for each sample). Only 9 libraries had more than 100 sequences. GenBank sequences were clustered with 403,211 pyrosequences obtained from 48 separate samples from the Cascade Cave System (Brannen-Donnelly & Engel, 2015). After clustering, there were 23,658 OTUs, but only 20,136 were successfully aligned and used to make a phylogenetic tree (Table A2-1). Compared to massively parallel, high-throughput, next-generation sequencing technologies using Illumina or Roche 454 platforms, the total diversity from cloned samples was low and generally considered to underrepresent the diversity of the original material (Nikolaki & Tsiamis, 2013). Specifically, OTU richness measurements like Chao1 indicated that 454 tag pyrosequencing studies from Cascade Cave

and Lava Beds National Monument, New Mexico, USA (Northup et al., 2012) had higher species richness, likely due to the fact that these studies obtained pyrosequences and had more data than the clone-based studies (Fig. A2-1; figures located in Appendix I). Sample evenness from Shannon indices indicated that most libraries had evenly distributed OTU diversity (Fig. A2-1). Because OTU richness was low in most libraries, calculations of library similarity/difference based on the variation of OTUs in each library was completed with an unweighted UniFrac distance metric utilizing a phylogenetic tree (Lozupone and Knight, 2005)

A total of 57 phyla were represented by the full dataset (Table A2-2), including 19 putative candidate divisions. Proteobacteria was the most abundant phylum retrieved for all cave types except ice caves (Table A2-2). Within the proteobacterial classes, the relative abundances of each class significantly varied by environment type (ANOVA F-value = 5.35, p-value = 0.002). The PCoA of UniFrac distance metrics explained ~31% of the variation in OTU distribution between libraries (Fig. A2-2), but there was no identifiable trend in library composition similarity based on environment type. However, the adonis results that assessed OUT distribution between samples based on grouping by environment type (F-value = 1.13, p-value = 0.04) were statistically significant, but the effect of the environment type is not the only factor effecting the distribution of OTUs between samples

From the cluster analysis, the gap statistic indicated that there were three optimal clusters for the phylogenetic tree (Fig. A2-3). The three clusters were plotted spatially (Fig. A2-4), with the major taxonomic groups within each cluster as follows: cluster 1, all phyla except Bacteroidetes; cluster 2, all phyla except Acidobacteria, Bacteroidetes, Elusimicrobia, Candidate

Division OD1, and Planctomycetes; cluster 3, Acidobacteria, Bacteroidetes, Elusimicrobia, Candidate Division OD1, Planctomycetes, and Proteobacteria. Where more than one phylum was represented in a cluster, there was a split between OTUs within that phyla, and some of the OTUs had a smaller pairwiase distance to OTUs from a different phylum in the same cluster than to OTUs in the same phylum. Libraries comprising each cluster were obtained from several caves, from different cave types, and from different continents. Moreover, the spatial distances among libraries compared to the Jaccard Index values indicated that there was a bimodal distribution of distances and OTU similarity between libraries, with some libraries having OTUs comprised of sequences obtained from geographically proximal locations while other libraries had OTUs comprised of sequences obtained from caves separated by 1000s of km (Fig. A2-5).

## DISCUSSION

Of the thousands to tens of thousands, to even millions, of different, extant bacterial and archaeal species predicted (Curtis et al., 2002; Achtman and Wagner 2008; Yarza et al., 2014), only 11,000 species have been classified, predominately from using culture-based methods (Yarza et al., 2014). By 2017, the rate of discovery of new bacterial and archaeal species is expected to decline, despite the exponential increase in publically available sequences from next-generation sequencing efforts (Yarza et al., 2014). However, this rate of discovery assumes that all types of environments have already been thoroughly sampled. For caves and karst habitats, it is clear that they have been undersampled and available genetic data are much less than those from other terrestrial and marine environments on Earth (Mora

et al., 2011; Henschel et al., 2015 ), despite recent efforts of some countries to better understand subsurface microbial populations and ecosytsems (Griebler et al., 2010; Navarro-Ortega et al., 2015).

From the publically available datasets, some cave systems have more than one entry, representing multiple publications. Since 1997, only an estimated 135 studies have been published describing the microbiology of caves worldwide using culture-independent methods (Engel, 2015). The more well-studied caves tended to have some archeological value, such as those with Paleolithic cave paintings (Lascaux Cave, Altamira Cave, and Doña Trinidad cave; see Table A2-1 for references). Other systems offered ecological and potential astrobiological insight because of extreme environmental conditions, like low pH, or having chemolithoautotrophy at the base of their cave ecosystem (e.g., Movile Cave, Romania; Lower Kane Cave, Wyoming; Frasassi caves, Italy; see Table A2-1 for references). In contrast, the microbiology of cave systems like Cascade Cave, where the microbial community relies on allochthonous nutrients, are not well represented in the databases, even though epigenic caves are the most common types of caves in terms of speleogenesis and ecosystem classification (Palmer, 2007). Moreover, from publication meta-analysis, many studies did not have similar purposes, sample measurements (besides sequence data), or reporting strategies. For instance, sample metadata, including geochemical analyses, were not required when sequences were provided to GenBank several years ago. Newer databases, such as the NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/Traces/sra), require metadata, including minimal geochemical data like temperature and pH, for sequence data.

For this study, cave type was broadly simplified by lithology or dominant chemistry, which was done because many studies lacked geochemical metadata and because many cave and karst environments are homogeneous at the level of rock type or aqueous geochemistry (Ford and Williams, 2013). OTU distribution from the libraries significantly grouped by cave type or cave environment, confirming what earlier studies of cave microbial diversity found (Engel, 2010; Lee et al., 2012; Paterson & Engel, 2015). Microbial diversity in caves is likely related to cave type, which results in specific geochemistry processes and specific nutrients which influence the metabolisms of specific bacterial groups.

The cluster analysis also indicated that taxonomic similarity among caves does not correspond to geographic distance, as some OTUs formed with sequences from caves 1,000s of km away from one another, whereby other OTUs formed with sequences from only one cave or caves that were geographically proximal to each other. The meta-analysis is the first spatial evaluation of the taxonomic distribution of bacteria in multiple types of cave systems. Recent publications discussing microbial biogeography and biogeographic effects on microbial ecology do not mention cave and karst environments (Fierer and Jackson, 2006; Martiny et al., 2006; Nemergut et al., 2011; Hanson et al., 2012). But, microbial biogeography of cave and karst environments is intriguing because caves form over thousands to millions of years. Unless strongly influenced by surface processes (e.g., flooding, dripwater, animal migration, etc.), microbes that get transported into caves and colonize surfaces or populate isolated water bodies may remain underground for long periods of time. The potential for adaptation to the cave environment, as well as speciation, is high.

In conclusion, despite evidence for undersampling and difficulties comparing sequencing technologies and strategies, bacterial diversity data from globally distributed cave systems indicate that some bacteria are ubiquitous. Specifically, some caves appear to have novel taxonomic OTU lineages while other caves have taxonomically similar OTUs even though there are 1000s of kilometers separating them. The implications are that it may be possible to compare microbes in karst (i.e., carbonate) caves around the world and to consider that these communities are functionally comparable. As such, the research findings place bacterial taxonomic and functional diversity from Cascade Cave, a system formed in limestone and supplied with surface-derived organic matter in context, and provide avenues for comparison. Lastly, the microbial diversity in cave and karst systems is underrepresented by comparison to other environment types like the oceans or soils, but the potential for novel diversity is high and cave habitats remain a place to uncover unique microbial species in the future.

**REFERENCES**


Achtman, M. & Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. *Nature Rev. Microbio*l. 6, 431–440.

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2013). GenBank. *Nucleic Acids Research* 41**,** D36-42. doi: 10.1093/nar/gks1195.

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

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, 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., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7**,** 335-336.

Charlop-Powers, Z., and Brady, S.F. (2015). phylogeo: an R package for geographic analysis and visualization of microbiome data. *Bioinformatics* 31**,** 2909-2911. doi: 10.1093/bioinformatics/btv269.

Crawford, P.A., Crowley, J.R., Sambandam, N., Muegge, B.D., Costello, E.K., Hamady, M., Knight, R., and Gordon, J.I. (2009). Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. *Proceedings of the National Academy of Science USA* 106**,** 11276-11281. doi: 10.1073/pnas.0902366106.

Culver, D.C., Hobbs III, H.H., Christman, M.C., and Master, L.L. (1999). Distribution map of caves and cave animals in the United States. *Journal of Cave and Karst Studies* 61**,** 139-140.

Curtis, T. P., Sloan, W. T., & Scannell, J. W. (2002). Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci*. 99, 10494–10499.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27**,** 2194-2200. doi: 10.1093/bioinformatics/btr381.

Engel, A.S. (2015). Bringing microbes into focus for cave science: An introduction, in A.S. Engel (ed.), *Microbial Life of Cave Systems,* De Gruyter. p. 1-22.

Engel, A.S., and Engel, S.A. (2009). A field guide for the karst of Carter Caves State Resort Park and the surrounding area, northeastern Kentucky, in *Select Field Guides to Cave and Karst Lands of the United States*, Karst Waters Institute Special Publication 15**,** 87-106.

Engel, A.S. (2010). Microbial diversity of cave ecosystems, in Barton, L., Mandl, M., and Loy, A. (eds.), *Geomicrobiology: Molecular & Environmental Perspectives*, Springer. p. 219-238. DOI:10.1007/978-90-481-9204-5_10.

Engel, A.S., Meisinger, D.B., Porter, M.L., Payn, R.A., Schmid, M., Stern, L.A., Schleifer, K.H., and Lee, N.M. (2010). Linking phylogenetic and functional diversity to nutrient spiraling in microbial mats from Lower Kane Cave (USA). *ISME J* 4**,** 98-110. doi: 10.1038/ismej.2009.91.

Engel, A.S., and Northup, D.E. (2008). Caves and karst as model systems for advancing the microbial sciences. *Frontiers of Karst Research*: Karst Waters Institute Special Publication 13, Leesburg, Virginia: 37-48.

Farnleitner, A.H., Wilhartitz, I., Ryzinska, G., Kirschner, A.K., Stadler, H., Burtscher, M.M., Hornek, R., Szewzyk, U., Herndl, G., and 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**,** 1248-1259. doi: 10.1111/j.1462-2920.2005.00810.x.

Fierer, N., and Jackson, R.B. (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Science USA* 103**,** 626-631. doi: 10.1073/pnas.0507535103.

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

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

Griebler, C., Stein, H., Kellermann, C., Berkhoff, S., Brielmann, H., Schmidt, S., Selesi, D., Steube, C., Fuchs, A., and Hahn, H.J. (2010). Ecological assessment of groundwater ecosystems – Vision or illusion? *Ecological Engineering* 36**,** 1174-1190. doi: 10.1016/j.ecoleng.2010.01.010.

Hallbeck, L., and Pedersen, K. (2008). Characterization of microbial processes in deep aquifers of the Fennoscandian Shield. *Applied Geochemistry* 23**,** 1796-1819. doi: 10.1016/j.apgeochem.2008.02.012.

Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C., and Martiny, J.B. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology* 10**,** 497-506. doi: 10.1038/nrmicro2795.

Henschel A., Anwar M. Z., and Manohar, V. (2015). Comprehensive Meta-analysis of Ontology Annotated 16S rRNA Profiles Identifies Beta Diversity Clusters of Environmental Bacterial Communities. *PLoS Comput Biol.* 11(10): e1004468. doi:10.1371/journal.pcbi.1004468

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* 12**,** 118-123. doi: 10.1111/j.1462-2920.2009.02051.x.

Lee, N.M., Meiginger, D.B., Aubrecht, R., Kovacic, L., Saiz-Jimenez, C. Baskar, S.R., Schleifer, K.-H., Liebl, W., Porter, M.L., Engel, A.S. (2012) *Caves and karst environments*, in Bell, E. (ed.) Life at Extremes: Environments, Organisms and Strategies for Survival. CAB Interantioanl Publishing. p. 320-344.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* 71**,** 8228-8235. doi: 10.1128/AEM.71.12.8228-8235.2005.

Macalady, J.L., Lyon, E.H., Koffman, B., Albertson, L.K., Meyer, K., Galdenzi, S., and Mariani, S. (2006). Dominant microbial populations in limestone-corroding stream biofilms, Frasassi

cave system, Italy. *Applied and Environmental Microbiology* 72**,** 5596-5609. doi: 10.1128/AEM.00715-06.

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

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

Mcmurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8**,** e61217. doi: 10.1371/journal.pone.0061217.

Mora, C., Tittensor, D. P., Adl, S., Simpson, S. G. B. and Worm, B. (2011). How many species are on Earth and in the ocean. *PLoS Biol.* 9, e1001127.

Navarro-Ortega, A., Acuna, V., Bellin, A., Burek, P., Cassiani, G., Choukr-Allah, R., Doledec, S., Elosegi, A., Ferrari, F., Ginebreda, A., Grathwohl, P., Jones, C., Rault, P.K., Kok, K., Koundouri, P., Ludwig, R.P., Merz, R., Milacic, R., Munoz, I., Nikulin, G., Paniconi, C., Paunovic, M., Petrovic, M., Sabater, L., Sabaterb, S., Skoulikidis, N.T., Slob, A., Teutsch, G., Voulvoulis, N., and Barcelo, D. (2015). Managing the effects of multiple stressors on aquatic ecosystems under water scarcity. The GLOBAQUA project. *Science Total Environment* 503-504**,** 3-9. doi: 10.1016/j.scitotenv.2014.06.081.

Nemergut, D.R., Costello, E.K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S.K., Fierer, N., Townsend, A.R., Cleveland, C.C., Stanish, L., and Knight, R. (2011). Global patterns in the biogeography of bacterial taxa. *Environmental Microbiology* 13**,** 135-144. doi: 10.1111/j.1462-2920.2010.02315.x.

Nikolaki, Sofia, and George Tsiamis. (2013). Microbial diversity in the era of omic technologies. *BioMed Research International 2013*. Volume 2013, Article ID 958719.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., O'hara, R., Simpson, G., Solymos, P., Stevens, M., and Wagner, H. (2013). vegan: Community Ecology Package.

Palmer, A.N. (2007). *Cave Geology.* Cave books Dayton.

Paterson, A.T., and Engel, A.S. (2015). Predicting bacterial diversity in caves associated with sulfuric acid speleogenesis, in *Microbial Life of Cave Systems,* ed. A.S. Engel. (Berlin: De Gruyter), 193-214.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology Evolution* 26**,** 1641-1650. doi: 10.1093/molbev/msp077.

Pronk, M., Goldscheider, N., and Zopfi, J. (2008). Microbial communities in karst groundwater and their potential use for biomonitoring. *Hydrogeology Journal* 17**,** 37-48. doi: 10.1007/s10040-008-0350-x.

Sarbu, S.M., Kane, T.C., and Kinkle, B.K. (1996). A chemoautotrophically based cave ecosystem. *Science* 272**,** 1953.

Simon, K., Benfield, E., and Macko, S. (2003). Food web structure and the role of epilithic biofilms in cave streams. *Ecology* 84**,** 2395-2406.

Simon, K.S., Pipan, T., and Culver, D.C. (2007). A conceptual model of the flow and distribution of organic carbon in caves. *Journal of Cave and Karst Studies* 69**,** 279-284.

Simon, K.S., Pipan, T., Ohno, T., and Culver, D.C. (2010). Spatial and temporal patterns in abundance and character of dissolved organic matter in two karst aquifers. *Fundamental and Applied Limnology / Archiv für Hydrobiologie* 177**,** 81-92. doi: 10.1127/1863-9135/2010/0177-0081.

Spizzico, M., Lopez, N., and Sciannamblo, D. (2005). Analysis of the potential contamination risk of groundwater resources circulating in areas with anthropogenic activities. *Natural Hazards and Earth System Science* 5**,** 109-116.

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63**,** 411-423.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., and Peterson, G.D. (2014). XSEDE: accelerating scientific discovery. *Computing in Science & Engineering* 16**,** 62-74.

Venarsky, M.P., Huntsman, B.M., Huryn, A.D., Benstead, J.P., and Kuhajda, B.R. (2014). Quantitative food web analysis supports the energy-limitation hypothesis in cave stream ecosystems. *Oecologia* 176**,** 859-869. doi: 10.1007/s00442-014-3042-3.

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

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73**,** 5261-5267. doi: 10.1128/AEM.00062-07.

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

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K. H., Whitman, W. B., Euzéby, J., Amann, R. and Rosselló-Móra, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology* 12, 635–645. doi:10.1038/nrmicro3330

Zhou, Y., Kellermann, C., and Griebler, C. (2012). Spatio-temporal patterns of microbial communities in a hydrologically dynamic pristine aquifer. *FEMS Microbiology Ecology* 81**,** 230-242. doi: 10.1111/j.1574-6941.2012.01371.x.

**APPENDIX I**

**FIGURES**

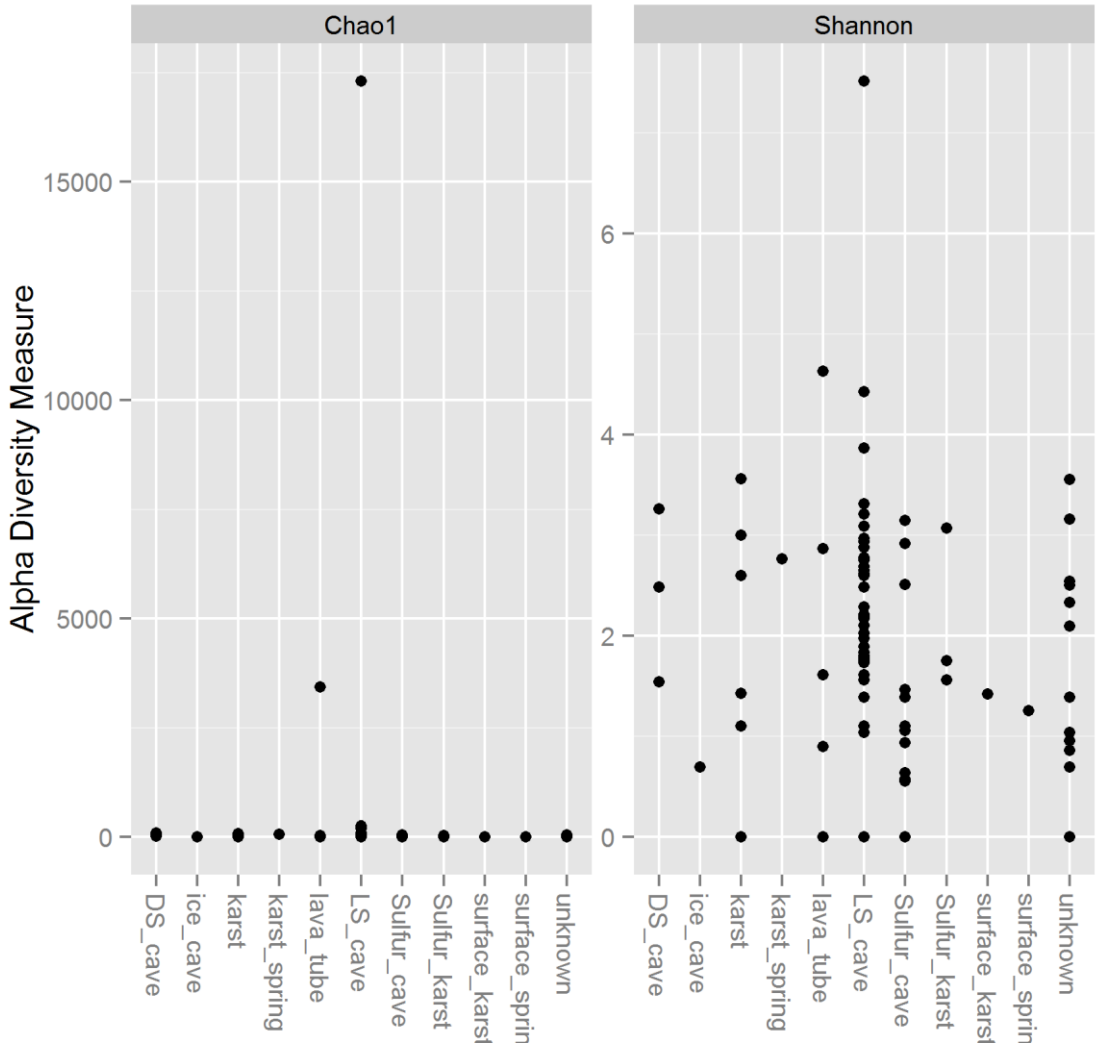



Figure A 2-1: OTU richness and evenness for all samples by environment type. Chao1 Index measure OTU richness. Shannon diversity Index measure OTU richness and evenness.

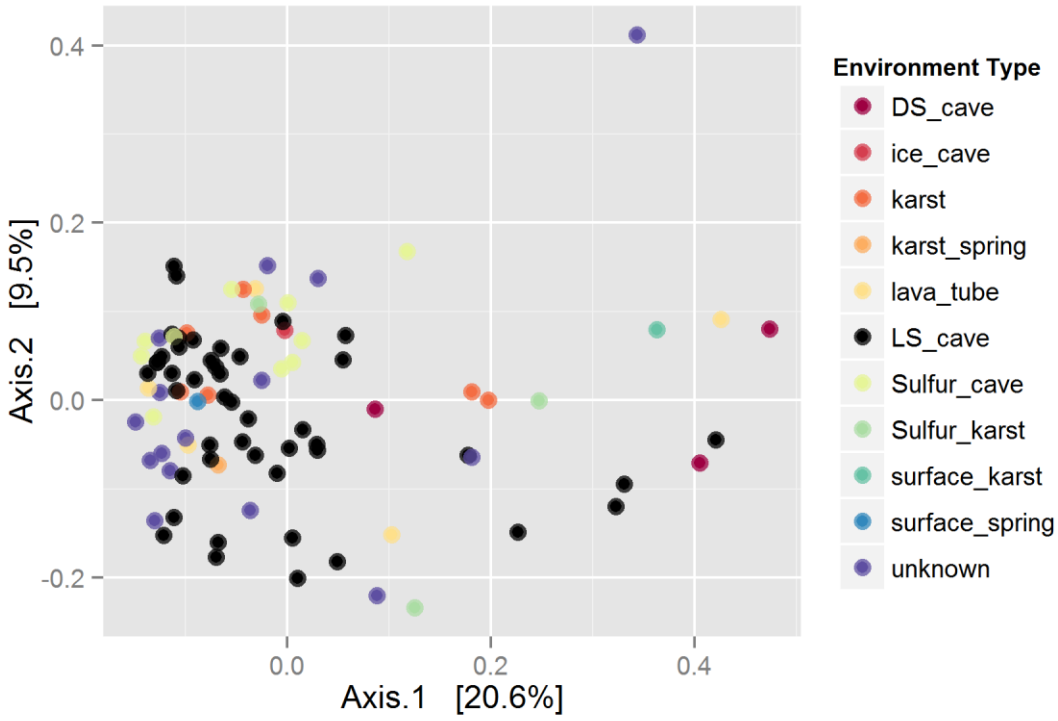



Figure A 2-2: PCoA of all samples based on a weighted UniFrac distance metric. The first two axes explain 28% of the variance of OTUs among samples. “DS” = dolostone, “LS” = limestone, representing caves with carboante bedrock as the dominant geologic feature. “Sulfur” indicates the presence of gypsum or the aqueous presence of any sulfur species (such as $H_2S$ and $SO_4^{2-}$), even though these caves formed in carbonate bedrock.

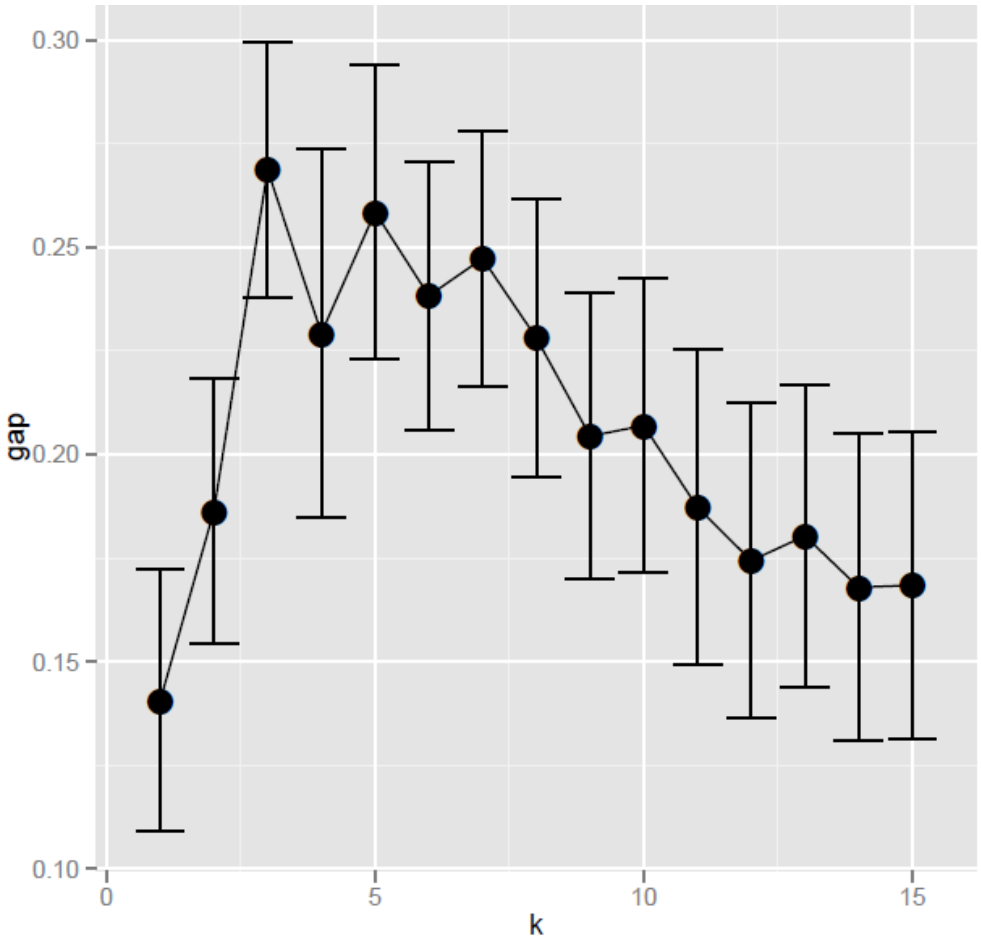
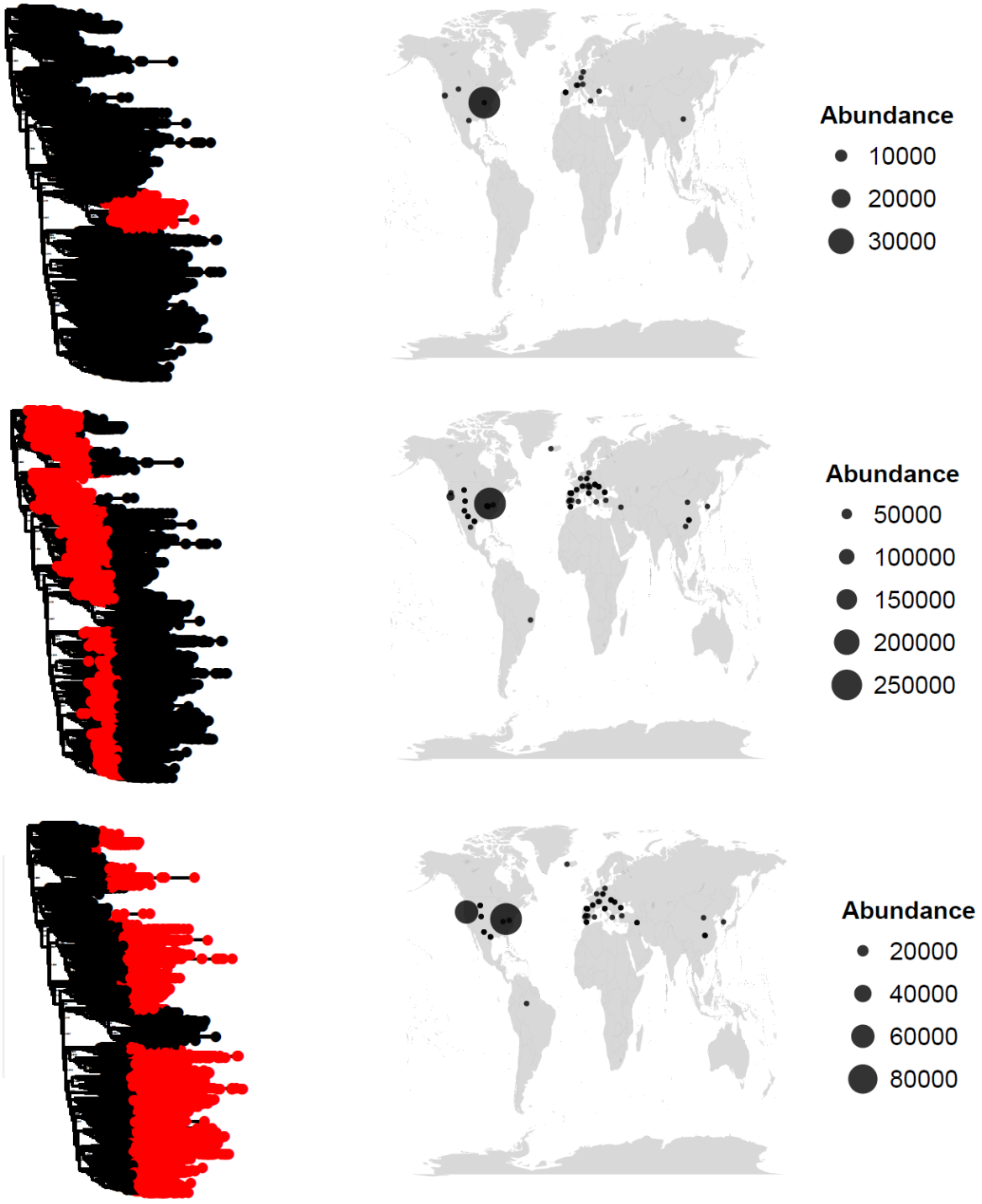



Figure A 2-3: The goodness of clustering measure or the “gap” statistic. For each number of clusters k, the clustering fit compares log(W(k)) with E*[log(W(k))], where the latter is defined by simulations from a reference distribution. The R code used to create this graph is available in the Appendix IV.

Figure A 2-4: The location of sequence subgroups as identified by 3 k-means clusters. There are regions from many of the same phyla of the phylogenetic tree in each cluster. The major taxonomic groups within each cluster were as follows: cluster 1, all phyla except Bacteroidetes; cluster 2, all phyla except Acidobacteria, Bacteroidetes, Elusimicrobia, OD1, and Planctomycetes; cluster 3, Acidobacteria, Bacteroidetes, Elusimicrobia, OD1, Planctomycetes, and Proteobacteria.


Abundance
10000
20000
30000
Abundance
50000
100000
150000
200000
250000
Abundance
20000
40000
60000
80000

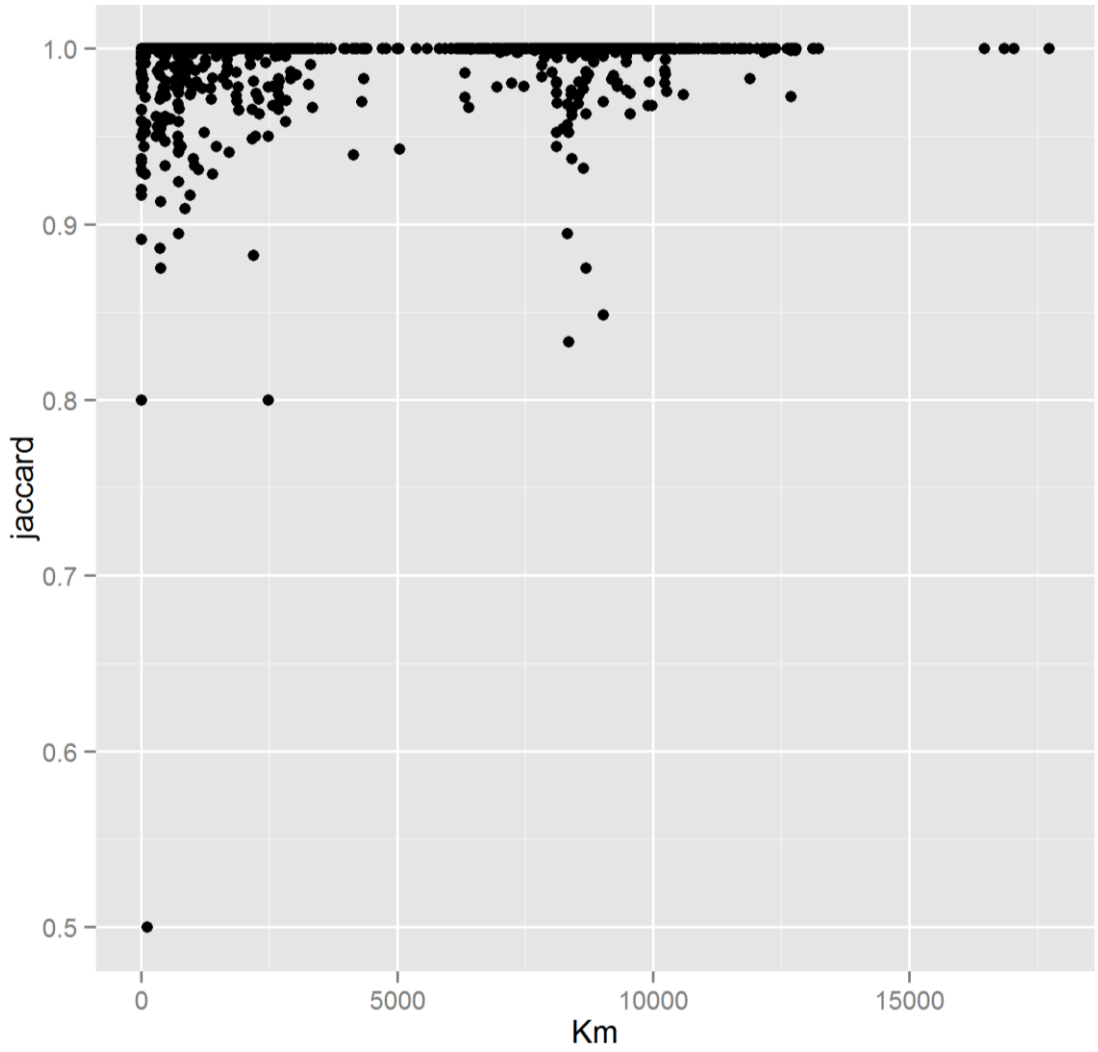



Figure A 2-5: Pairwise sample ecological distance, of geographic distance (kilometers, Km) versus Jaccard Distance calculated between every set of samples. A value of 1 indicates that there are no overlapping OTUs among samples.

**TABLES**

Table A 2-1: Number of sequences and OTUs from each GenBank sample analyzed in this study. The environment type, title of sequence, latitude, longitude, journal citation, cave name, region and country information for each sample are provided, if available.

| Environment Type | latitude | longitude | Cave Name(s) | Country | # OTUs | # sequences | Author, journal |
|---|---|---|---|---|---|---|---|
| karst | 47.17382 | 12.6822 | | Austria | 3 | 3 | Ryzinska et al., Environmental Microbiology 7.8 (2005): 1248-1259. |
| LS_cave | 50.4864 | 5.026386 | Scladina | Belguim | 20 | 35 | Orlando et al., Unpublished |
| LS_cave | 46.5314 | 22.59371 | Magura Cave | Bulgaria | 6 | 6 | Tomova, Journal of Cave and Karst Studies 75.3 (2013): 218. |
| DS_cave | 30.45 | 110.4145 | Heshang Cave | China | 95 | 915 | Li et al. Applied Geochemistry 26.3 (2011): 341-347. |
| DS_cave | 30.45 | 110.4145 | Heshang Cave | China | 7 | 27 | Liu et al., Journal of Earth Science 21 (2010): 325-328. |
| DS_cave | 30.45 | 110.4145 | Heshang Cave | China | 14 | 14 | Liu et al., Organic Geochemistry 42.1 (2011): 108-115. |
| karst | 27.35949 | 107.2034 | | China | 1 | 1 | Tang and Lian, Unpublished |
| karst | 39 | 109 | | China | 9 | 12 | Zhang and Wang, Int. Biodeterior. Biodegradation 76, 92-97 (2013) |
| LS_cave | 45.05583 | 1.167308 | Lascaux cave | France | 6 | 7 | Martin-Sanchez, Instituto de Recursos Naturales y Agrobiología |
| LS_cave | 45.05472 | 1.167651 | Lascaux cave | France | 8 | 11 | Martin-Sanchez, The Conservation of Subterranean Cultural Heritage – Saiz-Jimenez (Ed) © 2014 Taylor & Francis Group, London, ISBN 978-1-138-02694-0 |
| karst | 53.17251 | 13.10961 | | Germany | 74 | 481 | Cousin, International Microbiology 11.2 (2009): 91-100. |
| LS_cave | 50.42806 | 11.01861 | Herrenberg Cave | Germany | 23 | 23 | Rusznyak, Applied and environmental microbiology 78.4 (2012): 1157-1167. |
| LS_cave | 50.42806 | 11.01861 | Herrenberg Cave | Germany | 6 | 6 | Rusznyak, Applied and environmental microbiology 78.4 (2012): 1157-1167. |
| LS_cave | 39.17727 | 20.20728 | Blue Pot Cave | Greece | 24 | 69 | Gruenke, ISME Jounral 4 (8), 1031-1043 (2010) |
| karst | 47.52937 | 18.95932 | | Hungary | 5 | 23 | Anda et al., Extremophiles (2015): 1-11. |
| karst | 47.51841 | 19.03601 | | Hungary | 21 | 25 | Borsodi et al. Geomicrobiology Journal 29.7 (2012): 611-627. |
| karst_spring | 47.51841 | 19.03601 | | Hungary | 22 | 36 | Borsodi, Acta Microbiologica et Immunologica Hungarica, 61 (3), pp. 329–346 (2014) |
| LS_cave | 47.51841 | 19.03601 | | Hungary | 6 | 8 | Borsodi, Geomicrobiology Journal 29.7 (2012): 611-627. |

Table A2-1 Continued

| Environment Type | latitude | longitude | Cave Name(s) | Country | # OTUs | # sequences | Author, journal |
|---|---|---|---|---|---|---|---|
| LS_cave | 47.51841 | 19.03601 | Molnár János and Rudas-Török | Hungary | 8 | 10 | Borsodi, Geomicrobiology Journal 29.7 (2012): 611-627. |
| lava_tube | 64.74786 | -23.8179 | Vatnshellir Cave | Iceland | 19 | 26 | Unpublished |
| unknown | 36.62698 | 44.29664 | | Iraq | 13 | 22 | Unpublished |
| unknown | 36.62698 | 44.29664 | | Iraq | 5 | 29 | Unpublished |
| Sulfur_cave | 46.19691 | 12.86587 | Frasassi caves | Italy | 1 | 1 | Engel et al., International Journal of Speleology (2013) |
| Sulfur_cave | 43.4012 | 12.97611 | Frasassi caves | Italy | 2 | 31 | Macalady et al., Environmental Microbiology 9 (6), 1402-1414 (2007) |
| Sulfur_cave | 43.4012 | 12.97611 | Frasassi caves | Italy | 4 | 45 | Macalady et al., The ISME Journal (2008) 2, 590–601 |
| Sulfur_cave | 43.4012 | 12.97611 | Frasassi caves | Italy | 3 | 3 | Schaperdoth et al., Frontiers in microbiology 2 (2011). |
| karst | 26.98672 | -102.08 | | Mexico | 3 | 3 | Souza et al., PNAS U.S.A. 103 (17), 6565-6570 (2006) |
| ice_cave | 46.48986 | 22.80968 | Scarisoara Ice Cave | Romania | 2 | 4 | Pascu et al., Acta Carsologica (2014) |
| LS_cave | 46.55402 | 22.56947 | Ursilor Cave Cave and Cave Pestera cu Apa din Valea Lesului | Romania | 9 | 17 | Epur et al., Geomicrobiology Journal 31.2 (2014): 116-127. |
| Sulfur_cave | 43.82568 | 28.56103 | Movile Cave | Romania | 36 | 115 | Chen et al., ISME Journal 3 (2009): 1093-1104. |
| Sulfur_cave | 43.82568 | 28.56103 | Movile Cave | Romania | 4 | 4 | Hutchens et al., Environmental Microbiology 6.2 (2004): 111-120. |
| Sulfur_cave | 43.82568 | 28.56103 | Movile Cave | Romania | 14 | 23 | Wischer et al. The ISME journal 9.1 (2015): 195-206. |
| surface_karst | 46.36038 | 14.0278 | | Slovenia | 5 | 10 | Cankar et al., FEMS Microbiology Letters (2005) 244 (2), 341-345 |
| LS_cave | 36.98889 | 128.3817 | Gosu | South Korea | 21 | 30 | Chang et al., Chemical Geology (2010) 276, Issues 3–4 |
| LS_cave | 39.36561 | 2.852976 | Pas de Vallgornera cave | Spain | 36 | 89 | Bisbal, International Journal of Speleology 43.2 (2014): 8. |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 7 | 9 | Gonzalez et al., Journal Applied Microbiology 104 (3), 681-691 (2008) |

Table A2-1 Continued

| Environment Type | latitude | longitude | Cave Name(s) | Country | # OTUs | # sequences | Author, journal |
|---|---|---|---|---|---|---|---|
| LS_cave | 40.09892 | -5.44154 | Castañar de Ibor | Spain | 12 | 12 | Jurado, Atmospheric Environment Volume 40, Issue 38, December 2006, |
| LS_cave | 36.87806 | -4.84562 | Doña Trinidad cave | Spain | 5 | 5 | Jurado, Environmental Science and Pollution Research 18.6 (2011): 1037-1045. |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 4 | 4 | Jurado, Environmental Science and Pollution Research 21.1 (2014): 473-484 |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 15 | 16 | Jurado, FEMS Microbiology Ecolgy 81 (1), 281-290 (2012) |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 27 | 34 | Jurado, FEMS Microbiology Ecology 81 (1), 281-290 (2012) |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 9 | 9 | Jurado, Naturwissenschaften 96.9 (2009): 1027-1034. |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 9 | 13 | Portillo and Gonzalez, Unpublished |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 119 | 232 | Portillo et al., Research Microbiolgy 160 (1), 41-47 (2009) |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 6 | 6 | Portillo et al., Unpublished |
| LS_cave | 43.29842 | -3.97024 | Covalanas Cave | Spain | 75 | 225 | Rivalta et al., Unpublished |
| LS_cave | 43.29842 | -3.97024 | Covalanas Cave | Spain | 1 | 2 | Rivalta et al., Unpublished |
| LS_cave | 43.29842 | -3.97024 | Monedas Cave | Spain | 20 | 53 | Sarro et al., Unpublished |
| LS_cave | 43.45953 | -5.06916 | Tito Bustillo | Spain | 16 | 21 | Schabereiter-Gurtner, Environmental microbiology 4.7 (2002): 392-400. |
| LS_cave | 43.33528 | -4.65236 | Llonin and La Garma Caves | Spain | 17 | 24 | Schabereiter-Gurtner, FEMS Microbiology Ecology 47 (2), 235-247 (2004) |
| LS_cave | 43.45953 | -5.06916 | Tito Bustillo | Spain | 5 | 6 | Schabereiter-Gurtner, Journal Microbiology Methods 45 (2), 77-87 (2001) |
| LS_cave | 43.37767 | -4.12228 | Altamira Cave | Spain | 7 | 9 | Schabereiter-Gurtner, Unpublished |
| LS_cave | 36.87806 | -4.84562 | Doña Trinidad cave | Spain | 18 | 55 | Stomeo et al., Coalition 14:24–27 |

Table A2-1 Continued

| Environment Type | latitude | longitude | Cave Name(s) | Country | # OTUs | # sequences | Author, journal |
|---|---|---|---|---|---|---|---|
| LS_cave | 39.88346 | -3.48534 | Cave of Doña Trinidad and Santimamiñe Cave | Spain | 10 | 15 | Stomeo et al., International Biodeterioration & Biodegradation 62 (4), 483-486 (2008) |
| LS_cave | 36.87806 | -4.84562 | Doña Trinidad cave | Spain | 21 | 37 | Stomeo et al., International Biodeterioration & Biodegradation 62 (4), 483-486 (2008) |
| LS_cave | 39.4614 | -6.3723 | Maltravieso Rock Cave | Spain | 14 | 42 | Unpublished |
| LS_cave | | | Caves: Altamira, Tito Bustillo, Candamo, and Grotta dei Cervi; Catacombs: Domitilla and Saint Callixtus | Spain and Italy | 3 | 4 | Sanchez-Moral, Sergio, et al. Geomicrobiology Journal 20.5 (2003): 491-500. |
| LS_cave | 46.75373 | 7.724263 | Bärenschacht cave | Switzerland | 48 | 486 | Shabarova, Environmental microbiology 15.9 (2013): 2476-2488. |
| LS_cave | 39.92836 | 29.58543 | Oylat Cave | Turkey | 7 | 10 | Gulecal, FEMS microbiology ecology 86.1 (2013): 101-113. |
| lava_tube | 34.892 | -107.932 | Carlsbad Cavern, Lechuguilla Cave | USA | 1 | 1 | Northup et al, Astrobiology 11.7 (2011): 601-618. |
| lava_tube | 34.892 | -107.932 | | USA | 4 | 9 | Northup et al., Astrobiology 11.7 (2011): 601-618. |
| lava_tube | 41.73424 | -121.516 | | USA | 3922 | 94574 | Northup et al., Submitted (27-AUG-2012) |
| LS_cave | 37.29581 | -85.5293 | Unnamed Cave | USA | 1 | 1 | Banks et al., Geomicrobiology Journal 27.5 (2010): 444-454. |
| LS_cave | 37.12331 | -86.1307 | Parker Cave | USA | 7 | 11 | Barton et al., Unpublished |
| LS_cave | 38.34953 | -83.1068 | Cascade Cave | USA | 18481 | 403211 | Brannen-Donnelly and Engel, Frontiers in microbiology (2015) 6:729 |
| LS_cave | 38.34953 | -83.1068 | Carter Salt Peter Cave | USA | 3 | 8 | Carmichael and Brauer, Journal of Cave and Karst Studies 75(3): 189-204 |

Table A2-1 Continued

| Environment Type | latitude | longitude | Cave Name(s) | Country | # OTUs | # sequences | Author, journal |
|---|---|---|---|---|---|---|---|
| LS_cave | 37.72435 | -80.0565 | Cesspool Cave | USA | 3 | 3 | Engel, Geomicrobiology Journal 18, 259-274 (2001) |
| LS_cave | 37.18707 | -86.1 | Mammoth Cave | USA | 1 | 1 | Fowler et al., American Mineralogist 83 (1998): 1583-1592. |
| LS_cave | 36.82435 | -84.8751 | | USA | 1 | 2 | Fowler et al., Diss. University of Louisville, 2005. |
| LS_cave | 39.55316 | -107.32 | | USA | 18 | 171 | Spear et al., Applied and environmental microbiology 73.19 (2007): 6172-6180. |
| Sulfur_cave | 32.14812 | -104.557 | Lechuguilla Cave | USA | 4 | 6 | Barton et al., Unpublished |
| Sulfur_cave | 32.14812 | -104.557 | Lechuguilla Cave | USA | 5 | 9 | Dichosa et al., Geomicrobiology Journal (2005) 22, Issue 3-4 |
| Sulfur_cave | 44.86486 | -108.257 | Lower Kane Cave | USA | 2 | 3 | Engel et al., FEMS Microbiology Ecology (2004) 51 (1), 31-53 |
| Sulfur_cave | 32.14812 | -104.557 | Lechuguilla and Spider Caves | USA | 3 | 3 | Northup et al., Environmental Microbiology 5 (11), 1071-1086 (2003) |
| Sulfur_cave | 44.86486 | -108.257 | Lower Kane Cave | USA | 21 | 44 | Porter and Engel, International Journal of Speleology 38.1 (2009): 4. |
| Sulfur_cave | 44.86486 | -108.257 | Lower Kane Cave | USA | 3 | 18 | Rossmassler et al., FEMS microbiology ecology 79.2 (2012): 421-432. |
| Sulfur_karst | 29.75277 | -98.1731 | Edwards Aquifer well water | USA | 5 | 6 | Bates et al., The Southwestern Naturalist 51.3 (2006): 299-309. |
| Sulfur_karst | 29.85644 | -98.4327 | Edwards Aquifer well water | USA | 9 | 47 | Engel and Randall, Karst Waters Institute Special Publication 14 (2008): 52-56. |
| Sulfur_karst | 29.75277 | -98.1731 | Edwards Aquifer well water | USA | 24 | 48 | Gray and Engel, ISME Journal 7 (2), 325-337 (2013) |
| surface_spring | 39.54969 | -107.323 | Glenwood Springs | USA | 4 | 8 | Barton and Luiszer, Journal of Cave and Karst Studies 67.1 (2005): 28-38. |
| karst | 46.74667 | 6.541676 | | Switzerland | 15 | 17 | Pronk, Hydrogeology Journal 17.1 (2009): 37-48. |
| lava_tube | 43.58331 | -121.077 | | USA | 5 | 10 | Popa, Astrobiology 12.1 (2012): 9-18. |
| unknown | | | | | 1 | 1 | Ellersdorfer, Unpublished |
| unknown | | | | | 12 | 34 | Engel, Submitted (12-JUL-2007) |

Table A2-1 Continued

| Environment Type | latitude | longitude | Cave Name(s) | Country | # OTUs | # sequences | Author, journal |
|---|---|---|---|---|---|---|---|
| unknown | | | | | 2 | 2 | Joshi and Banerjee, Unpublished |
| unknown | | | | | 38 | 80 | Koren and Rosenberg, Submitted (2008) |
| unknown | | | | | 1 | 1 | Kumar et al., Submitted (22-SEP-2010) |
| Unknown | | | | | 2 | 2 | Onal and Rodrigues, Submitted (14-JUN-2014) |
| unknown | | | | | 13 | 14 | Taylor and Barton, Unpublished |
| unknown | -17.6655 | -43.6836 | | Brazil | 1 | 1 | Unpublished |
| unknown | | | | | 4 | 4 | Unpublished |
| unknown | | | | | 3 | 15 | Unpublished |
| unknown | | | | | 29 | 68 | Yasir and Ullah, Submitted (12-OCT-2013) |

Table A 2-2: Relative abundances of sequences in each phyla by environment type. Proteobacteria are listed at the class-level.

| Phyla | DS cave | Ice cave | karst | Karst spring | Lava tube | LS cave | Sulfur cave | Sulfur karst | Surface karst | Surface spring |
|---|---|---|---|---|---|---|---|---|---|---|
| Alphaproteobacteria | 6.74 | 0.36 | 8.57 | 6.41 | 14.37 | 8.36 | 48.51 | 0.00 | 0.00 | 16.03 |
| Betaproteobacteria | 11.37 | 8.00 | 28.57 | 4.04 | 34.00 | 26.83 | 10.89 | 0.00 | 0.00 | 1.39 |
| Deltaproteobacteria | 0.21 | 0.91 | 5.71 | 1.80 | 1.87 | 1.39 | 0.00 | 0.00 | 0.00 | 0.00 |
| Epsilonproteobacteria | 0.00 | 0.00 | 0.00 | 0.00 | 0.34 | 12.89 | 0.00 | 0.00 | 37.50 | 5.57 |
| Gammaproteobacteria | 65.89 | 40.00 | 5.71 | 6.95 | 15.84 | 42.51 | 35.64 | 100.00 | 12.50 | 29.97 |
| Bacteroidetes | 3.68 | 39.82 | 0.00 | 1.15 | 10.61 | 4.53 | 1.98 | 0.00 | 0.00 | 2.79 |
| Verrucomicrobia | 0.21 | 0.00 | 0.00 | 0.64 | 3.80 | 0.70 | 0.00 | 0.00 | 0.00 | 0.00 |
| Actinobacteria | 5.26 | 1.64 | 2.86 | 52.27 | 3.52 | 0.00 | 0.99 | 0.00 | 37.50 | 7.32 |
| Acidobacteria | 0.32 | 0.91 | 2.86 | 10.93 | 3.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Nitrospirae | 0.32 | 1.64 | 8.57 | 7.99 | 2.76 | 0.70 | 0.00 | 0.00 | 0.00 | 0.70 |
| Planctomycetes | 0.32 | 0.00 | 0.00 | 1.15 | 1.85 | 1.05 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cyanobacteria | 0.00 | 0.00 | 0.00 | 0.21 | 1.09 | 0.00 | 0.00 | 0.00 | 0.00 | 11.85 |
| Chloroflexi | 0.00 | 0.00 | 0.00 | 1.18 | 0.98 | 0.00 | 0.00 | 0.00 | 12.50 | 3.83 |
| Firmicutes | 4.74 | 1.82 | 31.43 | 0.23 | 0.50 | 0.00 | 1.98 | 0.00 | 0.00 | 19.16 |
| Elusimicrobia | 0.00 | 0.00 | 0.00 | 0.10 | 0.41 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Gemmatimonadetes | 0.00 | 0.55 | 0.00 | 1.12 | 0.38 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 |
| Fibrobacteres | 0.42 | 0.00 | 0.00 | 0.00 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Armatimonadetes | 0.00 | 0.00 | 0.00 | 0.07 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Chlorobi | 0.00 | 0.00 | 0.00 | 0.43 | 0.07 | 0.70 | 0.00 | 0.00 | 0.00 | 0.00 |
| Spirochaetes | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Fusobacteria | 0.00 | 0.00 | 0.00 | 0.01 | 0.03 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Caldithrix | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Aquificae | 0.00 | 2.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Lentisphaerae | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Tenericutes | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table A2-2 Continued

| Phyla | DS cave | Ice cave | karst | Karst spring | Lava tube | LS cave | Sulfur cave | Sulfur karst | Surface karst | Surface spring |
|---|---|---|---|---|---|---|---|---|---|---|
| Thermi | 0.00 | 0.73 | 0.00 | 0.00 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Candidate Division OP3 | 0.21 | 0.18 | 0.00 | 0.08 | 0.61 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division OD1 | 0.00 | 0.00 | 0.00 | 0.01 | 0.44 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division TM6 | 0.00 | 0.00 | 0.00 | 0.01 | 0.27 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division SR1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division TM7 | 0.11 | 0.18 | 0.00 | 0.04 | 0.15 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division GN02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division WS3 | 0.00 | 0.00 | 5.71 | 0.31 | 0.12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Candidate Division FBP | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division WS2 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division OP11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division NKB19 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division NC10 | 0.00 | 0.18 | 0.00 | 0.11 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division SBR1093 | 0.00 | 0.00 | 0.00 | 0.59 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.70 |
| Cand. Division PAUC34f | 0.00 | 0.18 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division GN04 | 0.00 | 0.00 | 0.00 | 0.11 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division GAL15 | 0.00 | 0.18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division AD3 | 0.00 | 0.00 | 0.00 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division BRC1 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Cand. Division WPS-2 | 0.00 | 0.00 | 0.00 | 0.35 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Unclassified bacteria | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Other | 0.00 | 0.55 | 0.00 | 1.58 | 2.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.70 |

# Chapter 3 BACTERIAL DIVERSITY DIFFERENCES ALONG AN EPIGENIC CAVE STREAM REVEAL EVIDENCE OF COMMUNITY DYNAMICS, SUCCESSION, AND STABILITY

This chapter has been published: Brannen-Donnelly, K., and Engel, A.S. (2015). Bacterial diversity differences along an epigenic cave stream reveal evidence of community dynamics, succession, and stability. *Frontiers in Microbiology* 6. doi: 10.3389/fmicb.2015.00729.

KBD and ASE designed the study and sampling protocol. KBD collected and analyzed the data, and KBD and ASE interpreted the data and wrote the manuscript. ASE provided funding from the Jones Endowment. Additional acknowledgements are included at the end of the chapter.

## ABSTRACT

Unchanging physicochemical conditions and nutrient sources over long periods of time in cave and karst subsurface habitats, particularly aquifers, can support stable ecosystems, termed autochthonous microbial endokarst communities (AMEC). AMEC existence is unknown for other karst settings, such as epigenic cave streams. Conceptually, AMEC should not form in streams due to faster turnover rates and seasonal disturbances that have the capacity to transport large quantities of water and sediment and to change allochthonous nutrient and organic matter sources. Our goal was to investigate whether AMEC could form and persist in hydrologically active, epigenic cave streams. We analyzed bacterial diversity from cave water, sediments, and artificial substrates (Bio-Traps®) placed in the cave at upstream and downstream locations. Distinct communities existed for the water, sediments, and Bio-Trap® samplers. Throughout the study period, a subset of community members persisted in the water, regardless of hydrological disturbances. Stable habitat conditions based on flow regimes resulted in more than one contemporaneous, stable community throughout the epigenic cave

stream. However, evidence for AMEC was insufficient for the cave water or sediments. Community succession, specifically as predictable exogenous heterotrophic microbial community succession, was evident from decreases in community richness from the Bio-Traps®, a peak in Bio-Trap® community biomass, and from changes in the composition of Bio-Trap® communities. The planktonic community was compositionally similar to Bio-Trap® initial colonizers, but the downstream Bio-Trap® community became more similar to the sediment community at the same location. These results can help in understanding the diversity of planktonic and attached microbial communities from karst, as well as microbial community dynamics, stability, and succession during disturbance or contamination responses over time.

## INTRODUCTION

Caves are diagnostic dissolutional features in karst landscapes underlain by soluble rock (e.g., limestone or dolomite) where surface water sinks into the subsurface and flows in a network of self-evolving underground stream passages (Ford and Williams, 2013). Although hydrological flow regimes, watershed geometry, aqueous geochemistry, and bedrock geology differ between karst systems (Nico Goldscheider et al., 2006; Simon et al., 2007; Bonacci et al., 2008), many have similar, stable environmental conditions and components that contribute to habitability and ecosystem development (Hahn and Fuchs, 2009; Griebler et al., 2010). Microbes are important components of all subterranean ecosystems (Chapelle, 2000) and of every type of karst habitat (Griebler and Lueders, 2009). Although the compositions of microbial communities (from the aspect of alpha-diversity) have been widely evaluated from

karst (Griebler and Lueders, 2009), much still remains to be explored, including microbial diversity trends over time (Engel, 2010). Microbes regulate chemical reactions that cause mineral dissolution and precipitation (Engel et al., 2004; Engel and Randall, 2011; Lian et al., 2011) and affect contaminant remediation (Thomas and Ward, 1992). As such, interest in microbial communities in various karst settings has increased (Griebler et al., 2010), and attempts have been made to understand whether microbial diversity differs throughout distinct types of karst systems and what ecosystem conditions control or regulate community composition. For instance, in karst aquifers and cave pools where water residence times are exceedingly long, from months to years, autochthonous microbial endokarst communities (AMEC) develop (Farnleitner et al., 2005; Pronk et al., 2008). Understanding AMEC is important to groundwater ecology, biogeochemistry of karst aquifers, and water resource management and conservation (Farnleitner et al., 2005; Pronk et al., 2008; Griebler and Lueders, 2009; Zhou et al., 2012).

Previously described AMEC have been sampled as planktonic phenomena from annual and monthly sample events of karst springs (Farnleitner et al., 2005; Pronk et al., 2008). A uniform definition for AMEC has not been applied, despite other types of groundwater systems having taxonomic and functionally distinct attached and planktonic communities (Hazen et al., 1991; Alfreider et al., 1997; Lehman, 2007; Flynn et al., 2008; Zhou et al., 2012). Conceptually AMEC should be homogenized communities of planktonic and attached microbial cells from within a karst aquifer setting. Under elevated flow conditions during recharge events, high flow velocities would mobilize sediment (Dogwiler and Wicks, 2004) and cause high shear stress on

sediment-attached cells (Rehmann and Soupir, 2009; Ghimire and Deng, 2013). Biofilm development on sediments and aquifer surfaces would be limited and attached cells would become entrained into the water column and become part of the planktonic community (Rehmann and Soupir, 2009). A prescribed minimum time limit for AMEC formation in karst has not been described, but this is not surprising because the stability and potential AMEC successional patterns over time in most groundwater systems are also not well understood (Farnleitner et al., 2005). Typical AMEC bacterial compositions are apparently comprised of Acidobacteria, Nitrospira, Gammaproteobacteria, and Deltaproteobacteria, and AMEC comprise the majority of the overall community abundances (Farnleitner et al., 2005; Pronk et al., 2008). The major taxonomic groups in AMEC are phylogenetically related to surface-derived groups, but not identical, thereby highlighting the importance of being sourced from within a subsurface system. Although no truly endemic karst microorganisms have been identified (Griebler and Lueders, 2009), arguably enhanced genetic divergences between surface communities and AMEC could result from long flow path travel distances and longer periods of isolation between the surface and subsurface.

As such, it is unclear whether AMEC are present or can persist in systems where turnover rates are expected to be high, such as in cave streams. Cave streams are dynamic, usually turbulent underground features that form from sinking surface water. Water is sourced from the surface and may reemerge from a conduit some distance later as a spring. Cave stream habitats that become established based on prevailing physicochemical gradients may only last for hours to weeks, according to the hydrological connection (i.e., continuous, flashy,

etc.) with the surface. Sediment suspension and deposition events caused by recharge flooding or flushing of the system could compositionally homogenize water and sediment microbial communities (at the level of beta-diversity), which would hamper the ability to detect AMEC from transported allochthonous communities. In this study, we investigated the diversity and prevalence of microorganisms from 16S rRNA gene sequences in stream water and cave sediments along a continuously flowing cave stream of fixed length but having different flow rates due to storm events over a six month period. In addition to documenting novel bacterial diversity for an epigenic cave stream, we compared water-transported (i.e. planktonic) and sediment (i.e. attached) bacterial diversity to test the hypothesis that an AMEC exists, despite storm water and sediment disturbances and differential contribution of surface-derived bacterial groups into the subsurface. We expected water and sediment communities to be similar to each other after high flow events, but that sediment communities would represent AMEC in between high flow events that would resuspend some or all of the cave sediments.

We also hypothesized that disturbance events reveal successional patterns between upstream and downstream communities. Studying microbial community successional patterns has proven difficult in many ecological systems (Shade et al., 2013). For this study, we used the definition of succession from Fierer et al. (2010), as the "orderly and predictable manner by which communities change over time following the colonization of a new environment." During four months, we seeded bacterial communities on artificial substrates (Bio-Trap® samplers) that were fixed in one upstream and one downstream location in the cave system. The Bio-Traps® were subsampled every month so that only a portion of material was removed and the rest

remained in a sampler. This experimentally contrasted cave stream sediment samples, which had the potential to be redistributed and mobilized during the study. The newly formed Bio-Trap® communities every month were compared to preexisting water and sediment communities to test the hypothesis that Bio-Trap® communities would resemble sediment communities over time, despite being colonized initially by planktonic microbes. Combined, these results provide evidence for cave stream community assembly and community succession. Underlying drivers that could explain spatial and temporal changes in bacterial diversity were statistically evaluated against stream discharge, rainfall, and geochemistry, including fluorescence spectral data for chromophoric dissolved organic matter (CDOM) that highlighted organic matter seasonal changes.

## MATERIALS AND METHODS

### Site characterization

We conducted the study from July through December, 2013, in the Cascade Cave system within Carter Caves State Resort Park (CCSRP) in Carter County, Kentucky (Fig. A3-1). The system is comprised of at least three surveyed caves that formed within the carbonate Slade Formation (Mississippian) (Engel and Engel, 2009). The caves are situated in the James Branch stream watershed, which flows into Tygart's Creek at local base-level (Dougherty, 1985; Engel and Engel, 2009). The entire watershed is approximately 4 $km^2$. The surface stream flows over Pennsylvanian and Mississippian interbedded sandstone and shale units before it sinks underground at a waterfall called Fort Falls (herein referred to as the surface sampling

location). The cave system has flowing water year-round. Jones Cave is the first access point to the cave stream (herein referred to as the upstream sampling location). There is a karst window 500 m downstream from Jones Cave where surface water enters the subsurface from a small surface stream. The entrance to another cave, Sandy Cave, is located at the window. Cascade Cave has several entrances, and one is reached downstream of the karst window and Sandy Cave. Where the cave stream emerges at the surface as a karst spring and another entrance to Cascade Cave, we sampled at the Lake Room (herein referred to as the downstream sampling location). The total estimated distance of the underground cave stream from the top of the water fall to resurgence is approximately 1.5 km. Preliminary (i.e. unpublished) tracer tests from Fort Falls to the Lake Room indicate a base flow travel time of about 12 hours. All of the sampling was done in less than 3 hours to evaluate contemporaneous microbial communities that could be present or established at each location, specifically planktonic communities from water, attached communities from sediment, and newly formed communities from the Bio-Trap® devices.

At each sample location and time, water flow rates were calculated by an average of three flow readings using a Geopacks Basic Flowmeter. Passage or channel cross-sectional area and water depth were measured to calculate discharge (Q) as the product of velocity, depth of the water, and channel width. Sediment particle transport was calculated by comparing stream velocity to the Stokes Settling Velocity for all the grain sizes present in the sediment samples (methods describes below), according to Ferguson and Church (2004). With no automated meteorological station data from CCSRP, daily precipitation data are measured and recorded at

the Fort Falls location by a citizen scientist who works in CCSRP (Fig. A3-2).

**Water and sediment sampling and analyses**

At each sampling location, physicochemical properties were measured using standard electrode methods (American Public Health et al., 2005), including pH, temperature, dissolved oxygen, total dissolved solids, and conductivity. At least 500 mL of cave stream water were manually filtered through duplicate 0.22 □m Sterivex$^{TM}$ (PVDF, EMD Millipore) filters. Filters were frozen at -20$^{o}$ C until use. The filtered water was collected for anion (using clean HDPE bottles), cation (using acid-washed HDPE bottles), and total organic carbon (TOC) and total nitrogen (TN) analyses (using baked glass VOA vials). Cations were preserved with trace metal grade nitric acid. Samples were put on ice for transport and stored at 4° C until analysis. Alkalinity, representing bicarbonate concentration, was measured from 0.2 μm-filtered water in the field by manual titration to an end-point of pH 4.3 with 0.1 N $H_2SO_4$ (American Public Health et al., 2005). Major dissolved ions were measured on a Dionex ICS-2000 ion chromatograph, with standards checks accurate within two standard deviations. Total inorganic carbon (TIC) and dissolved organic carbon (DOC) concentrations were analyzed for filtered water with a Shimadzu Model TOC-V Total Carbon Analyzer. DOC was reported as the difference between dissolved nonpurgable organic carbon and TIC (American Public Health et al., 2005). The standard used for minimum detection limit was $C_8H_5KO_4$, and the precision between replicate sample injections was 2% of the relative percent difference (RPD) for DOC >4 mg/L and 5% RPD for DOC <4 mg/L. TN content were measured by a high temperature catalytic oxidation with

chemiluminescence minimum detection level of 0.01 mg/L (ASTM, 2008).

During some sampling times, only bare carbonate rock was exposed at a sample location in the cave where sediment had been present previously. If sediments were available to collect at a sampling location, then at least 25 g were aseptically collected from 0-2 cm deep and placed into sterile Falcon tubes; as such, any one particle had to be <20 mm to fit in the tube. Sediment was stored at -20$^{o}$ C until use. Sediment grain size was analyzed in triplicate for each sample from sieving air-dried material through sieves for >2 mm, 1 mm, 500 µm, 250 µm, 150 µm, and <150 µm. Weights of each sieved aliquot were measured to ±0.0001 g at least three times.

**Fluorescence spectroscopy**

Qualitative information about organic matter sources, composition, bioavailability, and the differences between allochthonous and autochthonous DOM can be determined from the natural concentration of CDOM (Coble, 1996; McKnight et al., 2001). The relative contributions of different CDOM sources in the filtered stream water were evaluated from excitation emission matrices (EEMs) produced by a Horiba Scientific Fluoromax4 spectrofluorometer with a Xenon lamp. A total of 43 emission scans were completed for each sample with setting of λEM = 250–550-nm, 2.5-nm steps; λEX = 240–550-nm, and 5-nm steps. Instrument settings were PMT voltage 800V, EX/EM slits 5-nm each, and an integration time 0.1 sec. Spectral corrections for primary and secondary inner filter effects of EEMs were made using absorbance spectra collected using a Thermo Scientific Evolution 200 series spectrophotometer in a 1-cm

cuvette over the 200-700 nm wavelength range with pyrogen-free deionized (DI) (>18.1 MΩ) water as the reference. Raman scattering was removed from EEMs by subtracting a DI water blank spectrum collected from each sample spectrum. Rayleigh scattering effects were edited from each spectrum, following correction and blank subtraction (Lakowicz, 2007).

Fluorescence data were interpreted from index analyses from individual emission scans or extracted from EEMs using methods previously described (Birdwell and Engel, 2010). We used the Fluorescence Index (FI) to assess terrestrial and microbial contributions to CDOM fluorescence (McKnight et al, 2001), the Humification Index (HIX) to estimate the degree of DOM humification (Ohno, 2002), and the Biological or Freshness Index (BIX) to evaluate the contribution of biological or microbial processes to CDOM fluorescence (Huguet et al., 2009).

**Microbial succession experiment**

Standard Bio-Trap® samplers baited with 30 g of 2-mm diameter Bio-Sep® beads made of Nomex® composite and powdered activated carbon were obtained from Microbial Insights, Inc. (Knoxville, TN, USA) (www.microbe.com). Slits on the samplers were 0.4 mm wide, and the inside of the samplers were wrapped with 0.011 mm mesh screen to reduce sediment and macrofauna intrusion. Bio-Traps® were suspended in triplicate (overall weight 1.3 kg) via ropes attached to the cave wall by using nondestructive, spring-loaded camming devices at Jones Cave (upstream location) and in the Lake Room (downstream location) (Fig. A3-3). At base-flow (i.e., low flow) conditions, Bio-Traps® were in contact with sediment or bare rock at the bottom of the stream channel, but were not buried in the sediment. The samplers were also weighted

by using 0.2 kg weights so that they would become suspended in the water column only during exceptionally high flow events (i.e., in excess of 0.5 m/s). The Bio-Traps® were sampled every month for four months. From each Bio-Trap®, 2.5 g beads were separated out and frozen until extraction. During the study period, no fine-grained or sand particles were observed in the Bio-Traps®. At the time of deployment (August 2013), the water column and sediment microbial communities were sampled at Fort Falls (surface location) and at both Bio-Trap® sample locations. Over the next four months at both Bio-Trap® locations, water column, surface sediment, and Bio-Trap® microbial communities were sampled. Only the water column and sediment microbial communities were sampled at the Fort Falls location during those time points.

**DNA extraction and pyrosequencing**

DNA was extracted from two Sterivex™ filters collected at each sampling location using a method modified from Riemann et al. (2008). Briefly, sucrose lysis buffer (0.75 M sucrose, 0.5 M Tris-HCl, 0.4 M EDTA) and 5 mg/mL lysozyme (Fisher BioReagents) were added to each filter prior to incubation at 37˚C for 30 minutes. Proteinase K (100-µg/mL final concentration; Fisher BioReagents) and 10% SDS were added, and digestion continued at 55˚C overnight. The lysate was drawn from the filter and combined with a 1X TE buffer wash of the filter prior to adding 0.3 M sodium acetate and molecular grade 100% isopropanol. Lysates were centrifuged and pellets were separated from the supernatants and resuspended in TE buffer. Nucleic acids were precipitated from the suspensions using 25:24:1 phenol:chloroform:isoamyl alcohol (pH 8)

twice, and 24:1 chloroform:isoamyl alcohol once, prior to pelleting by centrifugation. Pellets were washed with 100% molecular grade ethanol twice and then resuspended in 1X TE buffer.

MoBio PowerSoil® Extraction kits, following manufacturer instructions (MoBio Laboratories, Inc., Carlsbad, CA, USA), were used to extract total nucleic acids from 0.25 g of beads collected from each Bio-Trap® and separately from 0.25 g of sediments at each sampling location. Extractions for each sample type per sample period and location were done in triplicate.

The quality and quantity of extracted DNA were verified by examining products on TBE agarose gels with ethidium bromide staining after electrophoresis and by measuring the ratio of absorbance maxima at 260 and 280 nm, and 260 and 230 nm, with a Thermo Scientific Nanodrop 2000c Spectrophotometer. Duplicate (for water) or triplicate (for Bio-sep® beads or sediment) extractions at a sampling location and month were homogenized prior to purification, barcoding, and amplicon pyrosequencing using a Roche 454 FLX Titanium instrument and reagents, as described in Dowd et al. (2008), at the Molecular Research LP (MrDNA) laboratory (www.mrdnalab.com; Shallowater, Texas, USA). The V1-V3 region of 16S rRNA genes was amplified using 27F-534R primers (Dowd et al., 2008).

**qPCR analyses**

Bacterial biomass was estimated for all samples by quantitative PCR (qPCR) using a CFX96 Real-Time PCR System (Bio-Rad Laboratories, Hercules, CA, USA), according to the approach described by Ortiz et al. (2014). Briefly, for a 10-ml qPCR reaction with a 2x

SensiFAST™ SYBR® No-ROX Kit (Bioline Meridian Life Science Company, Tauton, MA), 400 mg/mL bovine serum albumin solution, 400 nM of each primer, and 400 pg DNA extract were used. Primers used for bacterial 16S rRNA amplification were 338F and 518R (Ortiz et al., 2014). A standard curve was used to calculate the number of 16S rRNA amplicons (Zhu et al., 2005):

$$N = [(A/B) \times d] \times (V/C)$$

N is the total number of cells in the initial sample; *A* is the number of 16S rRNA amplicons per PCR tube, as calculated from the standard curve; *B* is the number of µL of cell lysate in the PCR tube, and *d* the lysate dilution factor; V is the initial lysate volume expressed in µL, and C is the average number of 16S rRNA copies per bacterial cell. Based on the retrieved bacterial diversity from our samples, and specifically of the predominance of Proteobacteria, we used the value 4.2 based on the genome assessment work of Vetrovsky and Baldrian (2013). N was divided by the amount of water filtered for each sample, or the amount of sediment or Bio-Trap® beads used during the extractions, to find the number of cells per mL of water, or the number of cells per gram of sediment or Bio-Traps®, respectively.

**Sequence analyses**

Amplicon sequence data were quality screened and chimera checked prior to clustering into operational taxonomic units (OTUs) based on 95% sequence similarity using QIIME (Crawford et al., 2009; Caporaso et al., 2010; Edgar et al., 2011). A 95% cut-off was used to cluster OTUs at the genus level because of the short length of the pyrosequences (Kunin et al., 2010). The greengenes 13_8 database (DeSantis et al., 2006) was used as the reference for the

usearch61 method for chimera checking (Edgar et al., 2011) and for picking OTUs using the open reference method (DeSantis et al., 2006). From the 18,177 OTUs generated for the full dataset (397,144 amplicons; Supplemental Table 1), representative sequences were chosen for classification by the RDP Classifier at 80% confidence intervals using QIIME (Wang et al., 2007). Alpha-diversity was calculated in QIIME to generate rarefaction curves (Fig. A3-4) (Crawford et al., 2009; Caporaso et al., 2010) and Shannon diversity (H') and Chao1 indices were calculated in the computer program R using the package phyloseq (version 1.10.0) (McMurdie and Holmes, 2013). Higher numbers for both indices indicate greater OTU-level richness. All OTUs shared between samples were compared for presence/absence. Details regarding data processing are provided in the Supplemental Materials and Methods in R markdown format.

All raw amplicons obtained from this study were submitted to the NCBI Sequence Read Archive (SRA) under the Bioproject PRJNA283038, with the accession numbers SAMN03451533 - SAMN03451581 (http://www.ncbi.nlm.nih.gov). Summaries for the amplicon data, including SRA Accession Numbers for each sample, are included in Table A3-1 (all tables located in Apendix II).

**Statistical analyses**

The significance of changes in geochemical variables over time and between sampling months, as well as in microbial diversity data, were analyzed statistically using several approaches. Analysis of variance (ANOVA), reported as the F-test value with significance evaluated from a p-value of $<0.05$, was done with geochemical data between month, season,

and location using the car package in R (Fox and Weisberg, 2011). Summary of code completed in R is located in Appendix IV Code II. Sediment grain size comparisons were done using the G2Sd package for sediment size analysis (Gallon and Fournier, 2013). Permutational multivariate analysis of variance (PERMANOVA), calculated with the adonis function in the vegan package for R, was used to detect similarities in the means of multivariate groups described by material type (i.e., water, sediment, Bio-Trap®), location (i.e., surface, upstream, and downstream), and month, such that community OTU representation would be equivalent for all groups. PERMANOVA was also used to detect similarities in the composition and/or relative abundances of different OTUs based on geochemical variable (i.e., Cl, Ca, HIX, etc.). PERMANOVA was performed with the Adonis function from the vegan package for community ecology on a Bray-Curtis dissimilarity matrix and significance was assessed with 9999 permutations (Oksanen et al., 2013). Non-metric multidimensional scaling (NMDS) was used on a Bray-Curtis dissimilarity matrix to represent the pairwise dissimilarity graphically between OTUs in each sample. Statistically significant environmental variables (p-value <0.05) were plotted as vectors representing the average of factor levels using envfit, from the vegan package (Oksanen et al., 2013).

To investigate any linear relationships between the distribution of OTUs between samples and any redundant geochemical gradients, a redundancy analysis (RDA) was performed. The significance of RDA axes was calculated by the PCAsignificance function in the BiodiversityR package for R (Kindt, 2014). To evaluate the relationship between OTU distribution among sediment samples and sediment size, another RDA was performed on only

sediment samples. RDAs were performed with the rda function from the vegan package on a Bray-Curtis dissimilarity matrix (Oksanen et al., 2013), which was produced using a culled dataset of only OTUs present more than three times in at least 20% of the samples (McMurdie and Holmes, 2013). Only 313 of the original 18,177 OTUs remained and application of a 2.0 CV cutoff resulted in 178 OTUs.

## RESULTS

### Stream dynamics, sediment characteristics, and aqueous geochemistry

Several major rainfall events occurred during the study within the watershed. Stream discharge fluctuated from below detection limit by flow meter to as high as 1.36 $m^3/s$ at the downstream location (Table A3-2). At these flow rates during the study period, sediment particles up to 2 mm in diameter may have been mobilized during four different precipitation events based on Stokes calculations. Excluding the largest particles (i.e. cobbles), coarse to very coarse sand (0.5 - 2 mm diameter) was sampled from the upstream location at Jones Cave. The average particle sizes downstream in the Lake Room were fine-medium sand (0.125 - 0.5 mm) (Fig. A3-5). There was <1% contribution of silt- or clay-sized particles at both sampling locations. After a large storm event in December (Fig. A3-2), sediment remobilization and redistribution was evident and finer particles were deposited at the downstream location (Fig. A3-2).

Geochemical parameters for all of the stream water pH, ranging from 7.1 to 7.8, at each location significantly varied by month (ANOVA F-test = 33; p-value <0.001), as did stream temperature, ranging from 21°C (July) to 4°C (December) (p-value <0.001) (Table A3-2). Other

geochemical parameters, including alkalinity, also significantly differed by month (ANOVA F-test = 8.7; p-values for all analyses <0.05). The amount of DOC (ranging from 0.27 – 6.6 mg/L) and total dissolved N (ranging from 0.33 – 1 mg/L) did not significantly differ for any analysis by month or between locations. However, the quality of the carbon, as assessed by using fluorescence indices FI and HIX, did significantly differ by month (Table A3-3). In July and August, CDOM fluorescence was dominated by humic acids derived from terrigenous material and less proteinaceous CDOM than later months in the Fall and Winter seasons.

**Controls on bacterial biomass and diversity**

The number of 16S rRNA gene copies qPCR reaction ranged from $1 \times 10^5$ to $1 \times 10^2$ copies/sample, which was used to calculate biomass per gram of sediment or Bio-Trap® beads, or per mL water. Bio-Trap® samples had higher biomass (up to $2.6 \times 10^6$ cells/gram) than the other sample types; water had the least biomass at only $1 \times 10^4$ cells/mL (Fig. A3-6). Sediment biomass was greatest in August and decreased through the winter months, but biomass in the cave stream was relatively stable throughout the study period. Biomass in the Bio-Trap® samplers for both sampling locations were nearly the same, with the least biomass at the beginning of the experiment and the highest biomass in November.

The 18,177 OTUs were affiliated with 402 classified genera. The most abundant classes for all the OTUs included Betaproteobacteria (35% of all sequences), Gammaproteobacteria (16% of all sequences), Alphaproteobacteria (15% of all sequences), and Opitutae (4% of all sequences). The planktonic community throughout the cave stream was dominated by

Betaproteobacteria (48%), Alphaproteobacteria (8%), and Opitutae (6%). The sediment samples throughout the cave were dominated by Gammaproteobacteria (34%), followed by Alphaproteobacteria (16%) and Betaproteobacteria (12%). The Bio-Trap® communities from both locations had nearly equal distributions of Betaproteobacteria (26%), Alphaproteobacteria (24%), and Gammaproteobacteria (23%). Over time, observed Bio-trap® community OTU abundances decreased (Fig. A3-7A), but calculated richness and evenness were unchanged (according to H' and Chao1, Figs. A3-7B and A3-7C, respectively).

Prior to testing hypotheses related to AMEC existence and community succession, changes in bacterial diversity based on environmental gradients over time were evaluated. Each sample's taxonomic profile was compared temporally and spatially. Overall OTU taxonomic distribution between locations was significantly distinct from each other (i.e., upstream versus downstream) (PERMANOVA p-value <0.05, $r^2$ = 6%), and taxonomy differed significantly by month (PERMANOVA p-value <0.001, $r^2$ = 18%). OTU taxonomy clustered significantly by sample type (i.e., water, sediment, Bio-Traps®), according to both ordination in NMDS space (Fig. A3-8) and a RDA (Fig. A3-9) that tested potential multidimensional and linear relationships among environment gradients and taxonomy, respectively. Changes in seasonal CDOM quality from FI and HIX fluorescence indices accounted for observed bacterial diversity variation for water and Bio-Trap® samples, but not the sediment samples (RDA axis 2, 14.9%; Fig. A3-9). Instead, diversity from the sediment samples clustered by location and according to sediment size (Fig. A3-10), which also differed over time.

**Shared community membership and potential succession**

The number of shared OTUs were evaluated based on sample location, type (sediment, water, Bio-Traps®), and month to assess community stability, which could potentially provide evidence for AMEC. A shared OTU was identified if amplicons from more than one sample type, location, or month were present. Overall, the number of shared OTUs for any location or sample month was low, between 0.1 - 4% (Table A3-4), in contrast to the total number of OTUs retrieved during the six months. No OTUs comprised amplicons from all sediment, water, and Bio-Trap® samples from any location and any month (Table A3-4). But, there were shared OTUs from the sediments, water, and Bio-Traps® at each location over the six month study period (Table A3-4; Fig. A3-11), although the total number of shared OTUs was different for each material. Specifically, shared OTUs for sediment samples were comparatively lower (0.01 - 4% of the total) than the water and Bio-Trap® samples, which shared 20 - 65% of the OTUs when binned by sample type. The shared and prevalent OTUs over time showed sequence abundance changes (Fig. A3-11). Some of the most prevalent OTUs had a similar trend over time in both upstream and downstream locations (Fig. A3-11).

To assess community succession, comparisons among shared OTUs from sediment, water, and Bio-Traps® were made. Evidence for community succession was indicated if OTUs were comprised of amplicons from Bio-Traps® and either water or sediment over time. Upstream and downstream Bio-Trap® samples had more shared OTUs with water (20 OTUs upstream and 13 downstream) than with sediment (0 OTUs upstream and 1 downstream). Downstream, the number of shared OTUs between Bio-Traps® and sediments increased by the

end of the study (Table A3-5). This trend was not observed upstream, as the number of shared OTUs between Bio-Traps® and sediments remained low (Table A3-5).

## DISCUSSION

Originally described from karst spring water, AMEC represent stable communities that develop over months to years and that form from a mix of planktonic and biofilm (i.e. attached) communities within a karst aquifer (Farnleitner et al., 2005; Pronk et al., 2008). Karst aquifers have interconnected networks of solutionally-enlarged conduits and voids, solutionally-enlarged fractures and bedding partings, and bedrock matrix. Each component has its own flow regime, ranging from fast and potentially turbulent flow in conduits to Darcian or diffusive flow in fractures and the matrix (Ford and Williams, 2013). AMEC have previously been found within saturated conduits and voids and along fractures in the subsurface, where flow may be fast but residence times are long so environmental conditions remain stable, particularly pH and temperature (Farnleitner et al., 2005; Pronk et al., 2008). When AMEC were originally described, attached communities were not analyzed, presumably due to difficulties sampling karst bedrock surfaces from wells (Engel and Northup, 2008). Well boreholes completed in karst aquifers usually intercept fractures, conduits, and voids, and nothing but water can usually be physically sampled when voids are encountered. Moreover, these zones are cased off during well completion and inhibit future access to aquifer bedrock surfaces. Karst well construction and sampling contrasts other groundwater systems, such as porous sand and gravel aquifers, because aquifer sediment and/or rock material can be physically sampled from

cores during well construction. From these other types of groundwater systems, planktonic and attached microbial communities can be distinct based on taxonomic (Hazen et al., 1991; Alfreider et al., 1997; Lehman, 2007; Flynn et al., 2008; Zhou et al., 2012) and functional diversity (Wilhartitz et al., 2009). Moreover, planktonic microbial communities in porous sand and gravel aquifers can be seasonally dynamic while sediment-attached communities are unchanging (Zhou et al., 2012). Understanding how AMEC form and evolve is important because karst systems are highly susceptible to contamination (Vesper et al., 2001) and AMEC may play an important role the stability of microbial communities during ecosystem biogeochemical cycling or contaminant response.

Caves allow for direct entry into karst aquifer systems (Yagi et al., 2010; Morasch, 2013). Prior to this study, knowledge about cave stream bacterial diversity was limited and understanding how environmental parameters impact cave stream bacteria was poor (Engel, 2010). The hydrology of cave streams is different from that of the original AMEC habitats because residence times can be much shorter, on the order of hours to days, and environmental conditions can vary daily (Farnleitner et al., 2005; Pronk et al., 2008). Cave streams are hydrologically comparable to surface streams, and stable communities comparable to AMEC have not yet been identified from surface streams (Lyautey et al., 2005; Besemer et al., 2007; Lear et al., 2008; Besemer et al., 2012; Wey et al., 2012). However, in surface streams, sediment-attached microbial communities have been shown to express seasonal diversity trends (Feris et al., 2003; Hullar et al., 2006; Wey et al., 2012) and the distribution of planktonic bacteria and bacteria attached to fine benthic organic matter can also correlates to surface

stream pH (Fierer et al., 2007). As such, because cave streams are hydrologically connected to the surface, seasonal trends linked to physicochemistry may be observed from cave stream microbial communities. We found that, although there were significant differences for some environmental parameters over time, there were no significant differences in bacterial diversity over time at any one location along the cave stream. The duration of study may have been too short to observe potential lasting effects of seasonality on community assembly.

Conceptually, there is a low probability of AMEC development in cave streams because of more rapid removal or redistribution of material of all sizes (from clay particles to large logs), including microbial communities. In contrast to the original AMEC studies of planktonic communities (Farnleitner et al., 2005; Pronk et al., 2008), we hypothesized that sediment communities would be compositionally stable over time and provide evidence for AMEC formation because planktonic communities would likely be dominated by transient populations from the surface and stream water residence times would be too short for autochthonous communities to develop, in contrast to cave pools (Shabarova and Pernthaler, 2010; Shabarova et al., 2013; Shabarova et al., 2014). There were shared OTUs among the water samples throughout the entire study (Table A3-4), and the shared OTUs between the surface water and cave water indicated that some of the planktonic bacteria were ubiquitously distributed throughout the cave system (Table A3-4). This may be due to their survival throughout the duration of the flowpath, not that they are AMEC. In prior studies, to indicate a unique habitat consistent with microorganisms sourced autochthonously from within a system, >30 % of total sequences should be considered unclassified (<50 % sequence similarity) past the domain level

(Farnleitner et al., 2005; Pronk et al., 2008). From alpine systems, AMEC consist of Acidobacteria, Nitrospira, Gammaproteobacteria, and Deltaproteobacteria (Farnleitner et al., 2005; Pronk et al., 2008). In our study, the diversity of shared OTUs from the cave stream was different than previously described AMEC. Compared to the full bacterial diversity, the shared communities represented very little of the total diversity retrieved for all sample types (< 4%; Table A3-4). Consequently, we do not believe there is sufficient evidence that AMEC developed in the cave stream water. Also, as was originally described, AMEC should represent common bacterial groups that occur both in the water and from attached biofilms on sediments and aquifer surfaces (Farnleitner et al., 2005; Pronk et al., 2008). Sediment remobilization would cause similarity in planktonic and sediment-attached communities. However, our results do not support this because there were few OTUs shared between water and sediment communities over time. But, as separate habitats, water and sediments each shared OTUs throughout the entire study period (Table A3-4). Sediments at each location had distinct bacterial community compositions (Fig. A3-7) that correlated to sediment size. At the upstream location, only two OTUs were shared (representing 0.1 % of the overall community) over time, perhaps because sediment upstream may be more transient than downstream. Downstream, 16 OTUs, or 0.9 % of the total diversity were shared over time, and were comprised of Alpha- and Betaproteobacteria, Chloroflexi, Actinobacteria, and Acidobacteria. There were no OTUs shared between the surface sediments and upstream cave sediments, but four OTUs were shared between the surface sediments and downstream cave sediments. This may provide evidence that the cave sediment communities are not endemic to the karst system, but more

work needs to be done in the future and over longer periods of time to verify this result.

One reason why there is limited evidence for AMEC in the cave stream may be linked to the frequency of flooding. Significant rainfall events have the capacity to mobilize sediments of certain sizes. Based on calculated volume estimates for the different areas of the cave, flooding frequency, and particle size distribution, the smaller sediment upstream in the cave were probably only in place at most eight weeks during the study period. For AMEC to form in cave sediments, we would expect that the sediments should remain in place, or that attached communities are able to colonize newly (re)deposited sediments after an extended period of time. This would also increase the ability to readily distinguish AMEC from transient microbial communities. The monthly sampling intervals during the study period may have been too long to capture a stable community in the sediments because AMEC diversity was not easily distinguished from the sediments. Collectively from these results, it is unclear that AMEC, as defined originally as being autochthonous communities within a karst system (Farnleitner et al., 2005), formed in the cave stream sediments that were sampled in this study. We should point out that our sampling was biased towards smaller sediment sizes, and AMEC may develop on larger cobbles and boulders that are not mobilized as frequently as the smaller sediment sizes. Future work should sample the large sediment particles and the submerged cave wall and stream bottom surfaces because it may be possible that AMEC are present on more stable surfaces in the stream.

Lastly, we examined the potential for successional patterns in cave stream communities by using artificial substrates (i.e., Bio-Trap® samplers). Knowledge about community succession

and AMEC development in cave stream systems has been completely lacking. We hypothesized that Bio-Trap® communities would resemble sediment communities over time and we compared the community compositions among the planktonic and sediment-attached communities with those of the Bio-Traps®. Initially, even though the upstream and downstream planktonic communities differed, the Bio-Traps® at the upstream and downstream locations were dominated by OTUs shared with water at each location. Differences between the upstream and downstream communities were likely due to stochastic effects and dispersal potential (Fierer et al., 2010), but it is clear from the data that the planktonic microorganisms were the pioneering community for the Bio-Traps®. From a succession perspective, the downstream Bio-Traps® had more OTUs comprised of sediment amplicons at the end of the study (Table A3-5), but the upstream Bio-Traps® had the same small number of sediment-shared amplicons throughout the study. These results imply that the rate at which sediment-attached microorganisms colonize new surfaces differs depending on the location along the cave stream flowpath. At the end of the study, the relative abundances of several shared OTUs decreased at the upstream location but increased at the downstream location, suggesting that distinct Bio-Trap® communities formed according to the environmental conditions at each location (Fierer et al., 2010).

Variance among Bio-Trap® and water bacterial community compositions was positively correlated with CDOM quality along the cave stream flowpath, but CDOM quality did not correlate to sediment microbial community diversity. Bio-Trap® communities were likely utilizing CDOM in the water and not the sediments. This distinction is consistent with surface

stream studies (Hullar et al., 2006) as well as karst aquifers (Simon et al., 2010), and the differences may be due to organic matter in the streambed being partitioned differently from the water column (Simon et al., 2010). Although the effects of temperature on the nature of CDOM in surface streams has been shown to play an important role in planktonic bacterial community structure and function (Van der Gucht et al., 2005; Hullar et al., 2006), it is still unclear how environmental conditions affect CDOM in the cave streams and subsequent microbial community composition and assembly. Cave streams lack CDOM photodegradation, as well as the active photosynthesis that occurs in surface streams, which means that CDOM transported into the cave from the surface has the potential to retain its original properties. But, as CDOM is cycled along the flowpath, upstream CDOM is transformed and transported downstream or into the sediments for additional processing. The potential for CDOM quality to diminish with increasing travel time downstream may impact the composition and assembly of heterotrophic communities along the flowpath. The type of heterotrophic community that developed in the cave stream over time is consistent with exogenous (versus endogenous) communities because these commonly form aquatic biofilms under reduced light conditions and reach a diversity plateau with only small shifts in biomass once the community reaches the plateau phase (Fierer et al., 2010). Bio-Trap® samplers had a biomass peak in November (Fig. A3-3), and the overall trend in biomass and diversity suggests an exogenous heterotrophic community (Fierer et al., 2010). Future research should address if specific differences exist regarding the nature and behavior of water versus sediment organic matter and how those changes affect exogenous community composition and assembly over time.

In conclusion, microbes are essential for organic carbon and nutrient cycling in karst systems (Gibert et al., 1994; Simon et al., 2007). We found several distinct shared planktonic and attached bacterial communities in the cave stream, which is a novel outcome. However, although we found shared OTUs that were stable for the duration of our study, there were no OTUs shared between the planktonic and attached microbial communities. Therefore, we have limited evidence for an AMEC in this cave stream. Nevertheless, the definition of AMEC should be updated, as we struggled during our data analysis to find a set of ubiquitous requirements that could be used for comparison. The Bio-Trap® bacterial communities that stabilized over time in both upstream and downstream locations along a cave stream provide evidence that succession following a large-scale (perhaps sterilizing) environmental disturbance does occur in cave streams (Fierer et al., 2010). Despite the many flooding events during this study period, the community richness trend was predictable over time for all the Bio-Trap® samples, even though the pioneering microbial community was not the same. Sediment size and mobilization play a key role in the sediment-attached karst microbial community structure, and organic carbon quality governs the planktonic karst microbial community structure in a cave stream. These findings also indicate that cave stream communities with short water residence times can follow successional patterns in response to disturbances, like flooding or contamination events, although community stability only exists for short periods of time between disturbances.

**ACKNOWLEDGEMENTS**

Funding for this research was provided in part by a National Science Foundation Graduate Research Fellowship and the Cave Conservancy Foundation Karst Studies Scholarship to K.M.B.-D., and the Jones Endowment for Aqueous Geochemistry at the University of Tennessee. T. Hazen provided Bio-Traps® for the research, and B. Donnelly, C. Dietz, and A. Campion assisted in field sample collection. Thanks to S. Plummer for the rainfall data, as well as the staff at Carter Cave State Resort Park for granting permission to conduct the study. Thanks to H. H. Hobbs, III, for providing the cave map. A. Steen and J. Price provided R assistance. We also thank two anonymous reviewers for improving the manuscript.

## REFERENCES


Alfreider, A., Krossbacher, M., and Psenner, R. (1997). Groundwater samples do not reflect bacterial densities and activity in subsurface systems. *Water Research* 31**,** 832-840.

American Public Health, A., Eaton, A.D., American Water Works, A., and Water Environment, F. (2005). *Standard Methods for the Examination of Water and Wastewater.* Washington, D.C.: APHA-AWWA-WEF.

Astm (2008). D5373-08. Standard Test Methods for Instrumental Determination of Carbon, Hydrogen, and Nitrogen in Laboratory Samples of Coal. *Annual Book of ASTM Standards***,** 19428-12959.

Besemer, K., Peter, H., Logue, J.B., Langenheder, S., Lindstrom, E.S., Tranvik, L.J., and Battin, T.J. (2012). Unraveling assembly of stream biofilm communities. *ISME J* 6**,** 1459-1468. doi: 10.1038/ismej.2011.205.

Besemer, K., Singer, G., Limberger, R., Chlup, A.K., Hochedlinger, G., Hodl, I., Baranyi, C., and Battin, T.J. (2007). Biophysical controls on community succession in stream biofilms. *Appl Environ Microbiol* 73**,** 4966-4974. doi: 10.1128/AEM.00588-07.

Birdwell, J.E., and Engel, A.S. (2010). Characterization of dissolved organic matter in cave and spring waters using UV–Vis absorbance and fluorescence spectroscopy. *Organic Geochemistry* 41**,** 270-280. doi: 10.1016/j.orggeochem.2009.11.002.

Bonacci, O., Pipan, T., and Culver, D.C. (2008). A framework for karst ecohydrology. *Environmental Geology* 56**,** 891-900. doi: 10.1007/s00254-008-1189-0.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, 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., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7**,** 335-336.

Chapelle, F.H. (2000). The significance of microbial processes in hydrogeology and geochemistry. *Hydrogeology Journal* 8**,** 41-46.

Coble, P. (1996 ). Characterization of marine and terrestrial DOM in seawater using excitation-emission spectroscopy. *Marine Chemistry* 51**,** 325-346.

Crawford, P.A., Crowley, J.R., Sambandam, N., Muegge, B.D., Costello, E.K., Hamady, M., Knight, R., and Gordon, J.I. (2009). Regulation of myocardial ketone body metabolism by the gut microbiota during nutrient deprivation. *Proc Natl Acad Sci U S A* 106**,** 11276-11281. doi: 10.1073/pnas.0902366106.

Desantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G.L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72**,** 5069-5072. doi: 10.1128/AEM.03006-05.

Dogwiler, T., and Wicks, C.M. (2004). Sediment entrainment and transport in fluviokarst systems. *Journal of Hydrology* 295**,** 163-172. doi: 10.1016/j.jhydrol.2004.03.002.

Daugherty, P.H. (1985). An Overview of the Geology and Physical Geography of Kentucky. pp. 5-17, in *Caves and Karst of Kentucky: Special Publication 12*. Kentucky Geological Survey

Dowd, S.E., Callaway, T.R., Wolcott, R.D., Sun, Y., Mckeehan, T., Hagevoort, R.G., and Edrington, T.S. (2008). Evaluation of the bacterial diversity in the feces of cattle using 16S rDNA bacterial tag-encoded FLX amplicon pyrosequencing (bTEFAP). *BMC Microbiol* 8**,** 125. doi: 10.1186/1471-2180-8-125.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27**,** 2194-2200. doi: 10.1093/bioinformatics/btr381.

Engel, A.S. (2010). Microbial diversity of cave ecosystems, in Barton, L., Mandl, M., and Loy, A. (eds.), *Geomicrobiology: Molecular & Environmental Perspectives*, Springer. p. 219-238. DOI:10.1007/978-90-481-9204-5_10.

Engel, A.S., and Engel, S.A. (2009). A field guide for the karst of Carter Caves State Resort Park and the surrounding area, northeastern Kentucky, in *Select Field Guides to Cave and Karst Lands of the United States*, Karst Waters Institute Special Publication 15**,** 87-106.

Engel, A.S., and Northup, D.E. (2008). Caves and karst as model systems for advancing the microbial lsciences,in *Frontiers of Karst Research*, eds J.B.Martin and W. W. White (Leesburg: Karst Waters Institute Special Publication), 37–48.

Engel, A.S., and 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**,** 313-327. doi: 10.1080/01490451.2010.500197.

Engel, A.S., Stern, L.A., and Bennett, P.C. (2004). Microbial contributions to cave formation: New insights into sulfuric acid speleogenesis. *Geology* 32**,** 369. doi: 10.1130/g20288.1.

Farnleitner, A.H., Wilhartitz, I., Ryzinska, G., Kirschner, A.K., Stadler, H., Burtscher, M.M., Hornek, R., Szewzyk, U., Herndl, G., and Mach, R.L. (2005). Bacterial dynamics in spring water of alpine karst aquifers indicates the presence of stable autochthonous microbial endokarst communities. *Environ Microbiol* 7**,** 1248-1259. doi: 10.1111/j.1462-2920.2005.00810.x.

Ferguson, R.I., and Church, M. (2004). A simple universal equation for grain settling velocity *Journal of Sedimentary Research* 74**,** 933-937.

Feris, K., Ramsey, P., Frazar, C., Rillig, M., Gannon, J., and Holben, W. (2003). Structure and seasonal dynamics of hyporheic zone microbial communities in free-stone rivers of the estern United States. *Microbial Ecology* 46**,** 200-215.

Fierer, N., Morse, J.L., Berthrong, S.T., Bernhardt, E.S., and Jackson, R.B. (2007). Environmental controls on the landscape-scale biogeography of stream bacterial communities. *Ecology* 88**,** 2162-2173.

Fierer, N., Nemergut, D., Knight, R., and Craine, J.M. (2010). Changes through time: integrating microorganisms into the study of succession. *Res Microbiol* 161**,** 635-642. doi: 10.1016/j.resmic.2010.06.002.

Flynn, T.M., Sanford, R.A., and Bethke, C.M. (2008). Attached and suspended microbial communities in a pristine confined aquifer. *Water Resources Research* 44**,** W07425-W07432. doi: 10.1029/2007wr006633.

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

Fox, J., and Weisberg, S. (2011). " An {R} Companion to Applied Regression". (Thousand Oaks, CA: Sage).

Gallon, R.K., and Fournier, J. (2013). "G2Sd : Grain-size Statistics and Description of Sediment". 2.0 ed.).

Ghimire, B., and Deng, Z. (2013). Hydrograph-based approach to modeling bacterial fate and transport in rivers. *Water Res* 47**,** 1329-1343. doi: 10.1016/j.watres.2012.11.051.

Gibert, J., Danielopol, D., and Stanford, J.A. (1994). *Groundwater ecology.* Academic Press.

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

Griebler, C., Stein, H., Kellermann, C., Berkhoff, S., Brielmann, H., Schmidt, S., Selesi, D., Steube, C., Fuchs, A., and Hahn, H.J. (2010). Ecological assessment of groundwater ecosystems – Vision or illusion? *Ecological Engineering* 36**,** 1174-1190. doi: 10.1016/j.ecoleng.2010.01.010.

Hahn, H.J., and Fuchs, A. (2009). Distribution patterns of groundwater communities across aquifer types in south-western Germany. *Freshwater Biology* 54**,** 848-860. doi: 10.1111/j.1365-2427.2008.02132.x.

Hazen, T.C., Jimenes, L., Lopez De Victoria, G., and Fliermans, C.B. (1991). Comparison of Bacteria from Deep Subsurface Sediment and Adjacent Groundwater. *Microbial Ecology* 22**,** 293-304.

Huguet, A., Vacher, L., Relexans, S., Saubusse, S., Froidefond, J.M., and Parlanti, E. (2009). Properties of fluorescent dissolved organic matter in the Gironde Estuary. *Organic Geochemistry* 40**,** 706-719. doi: 10.1016/j.orggeochem.2009.03.002.

Hullar, M.A., Kaplan, L.A., and Stahl, D.A. (2006). Recurring seasonal dynamics of microbial communities in stream habitats. *Appl Environ Microbiol* 72**,** 713-722. doi: 10.1128/AEM.72.1.713-722.2006.

Kindt, R. (2014). BiodiversityR: GUI for biodiversity, suitability and community ecology analysis.

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 12**,** 118-123. doi: 10.1111/j.1462-2920.2009.02051.x.

Lakowicz, J.R. (2007). *Principles of fluorescence spectroscopy.* Springer Science & Business Media.

Lear, G., Anderson, M.J., Smith, J.P., Boxen, K., and Lewis, G.D. (2008). Spatial and temporal heterogeneity of the bacterial communities in stream epilithic biofilms. *FEMS Microbiol Ecol* 65**,** 463-473. doi: 10.1111/j.1574-6941.2008.00548.x.

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

Lian, B., Yuan, D., and Liu, Z. (2011). Effect of microbes on karstification in karst ecosystems. *Chinese Science Bulletin* 56**,** 3743-3747. doi: 10.1007/s11434-011-4648-z.

Lyautey, E., Jackson, C.R., Cayrou, J., Rols, J.L., and Garabetian, F. (2005). Bacterial community succession in natural river biofilm assemblages. *Microb Ecol* 50**,** 589-601. doi: 10.1007/s00248-005-5032-9.

Mcknight, D.M., Boyer, E.W., Westerhoff, P.K., Doran, P.T., Kulbe, T., and Anderson, D.T. (2001). Spectrofluorometric characterization of dissolved organic matter for indication of precursor organic material and aromaticity. *Limnology & Oceanography* 46**,** 38-48.

McMahon, S., & Parnell, J. (2014). Weighing the deep continental biosphere. *FEMS microbiology ecology*, 87(1), 113-120.

Mcmurdie, P.J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8**,** e61217. doi: 10.1371/journal.pone.0061217.

Morasch, B. (2013). Occurrence and dynamics of micropollutants in a karst aquifer. *Environ Pollut* 173**,** 133-137. doi: 10.1016/j.envpol.2012.10.014.

Nico Goldscheider, Daniel Hunkeler, and Rossi, P. (2006). Review: Microbial biocenoses in pristine aquifers and an assessment of investigative methods. *Hydrogeology Journal* 14**,** 926-941.

Ohno, T. (2002). Fluorescence Inner-Filtering Correction for Determining the Humification Index of Dissolved Organic matter. *Envionmental Science and Technology* 38**,** 742-746.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., O'hara, R., Simpson, G., Solymos, P., Stevens, M., and Wagner, H. (2013). "vegan: Community Ecology Package".).

Ortiz, M., Legatzki, A., Neilson, J.W., Fryslie, B., Nelson, W.M., Wing, R.A., Soderlund, C.A., Pryor, B.M., and 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**,** 478-491. doi: 10.1038/ismej.2013.159.

Pronk, M., Goldscheider, N., and Zopfi, J. (2008). Microbial communities in karst groundwater and their potential use for biomonitoring. *Hydrogeology Journal* 17**,** 37-48. doi: 10.1007/s10040-008-0350-x.

Rehmann, C.R., and Soupir, M.L. (2009). Importance of interactions between the water column and the sediment for microbial concentrations in streams. *Water Res* 43**,** 4579-4589. doi: 10.1016/j.watres.2009.06.049.

Riemann, L., Leitet, C., Pommier, T., Simu, K., Holmfeldt, K., Larsson, U., and Hagstrom, A. (2008). The native bacterioplankton community in the central baltic sea is influenced by freshwater bacterial species. *Appl Environ Microbiol* 74**,** 503-515. doi: 10.1128/AEM.01983-07.

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

Shabarova, T., Villiger, J., Morenkov, O., Niggemann, J., Dittmar, T., and Pernthaler, J. (2014). Bacterial community structure and dissolved organic matter in repeatedly flooded subsurface karst water pools. *FEMS Microbiol Ecol* 89**,** 111-126. doi: 10.1111/1574-6941.12339.

Shabarova, T., Widmer, F., and Pernthaler, J. (2013). Mass effects meet species sorting: transformations of microbial assemblages in epiphreatic subsurface karst water pools. *Environ Microbiol* 15**,** 2476-2488. doi: 10.1111/1462-2920.12124.

Shade, A., Caporaso, J.G., Handelsman, J., Knight, R., and Fierer, N. (2013). A meta-analysis of changes in bacterial and archaeal communities with time. *ISME J* 7**,** 1493-1506. doi: 10.1038/ismej.2013.54.

Simon, K.S., Pipan, T., and Culver, D.C. (2007). A conceptual model of the flow and distribution of organic carbon in caves. *Journal of Cave and Karst Studies* 69**,** 279-284.

Simon, K.S., Pipan, T., Ohno, T., and Culver, D.C. (2010). Spatial and temporal patterns in abundance and character of dissolved organic matter in two karst aquifers. *Fundamental and Applied Limnology / Archiv für Hydrobiologie* 177**,** 81-92. doi: 10.1127/1863-9135/2010/0177-0081.

Thomas, J.M., and Ward, C.H. (1992). Subsurface microbial ecology and bioremediation. *Journal of Hazardous Materials* 32**,** 179-194.

Van Der Gucht, K., Vandekerckhove, T., Vloemans, N., Cousin, S., Muylaert, K., Sabbe, K., Gillis, M., Declerk, S., De Meester, L., and Vyverman, W. (2005). Characterization of bacterial communities in four freshwater lakes differing in nutrient load and food web structure. *FEMS Microbiol Ecol* 53**,** 205-220. doi: 10.1016/j.femsec.2004.12.006.

Vesper, D.J., Loop, C.M., and White, W.B. (2001). Contaminant transport in karst aquifers. *Theoretical and Applied Karstology* 13**,** 101-111. doi: 10.1007/s10040-003-0299-8.

Vetrovsky, T., and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* 8**,** e57923. doi: 10.1371/journal.pone.0057923.

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

Wey, J.K., Jurgens, K., and Weitere, M. (2012). Seasonal and successional influences on bacterial community composition exceed that of protozoan grazing in river biofilms. *Appl Environ Microbiol* 78**,** 2013-2024. doi: 10.1128/AEM.06517-11.

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

Yagi, J.M., Neuhauser, E.F., Ripp, J.A., Mauro, D.M., and Madsen, E.L. (2010). Subsurface ecosystem resilience: long-term attenuation of subsurface contaminants supports a dynamic microbial community. *ISME J* 4**,** 131-143. doi: 10.1038/ismej.2009.101.

Zhou, Y., Kellermann, C., and Griebler, C. (2012). Spatio-temporal patterns of microbial communities in a hydrologically dynamic pristine aquifer. *FEMS Microbiol Ecol* 81**,** 230-242. doi: 10.1111/j.1574-6941.2012.01371.x.

Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* 52**,** 79-92. doi: 10.1016/j.femsec.2004.10.006.

## APPENDIX II

**FIGURES**

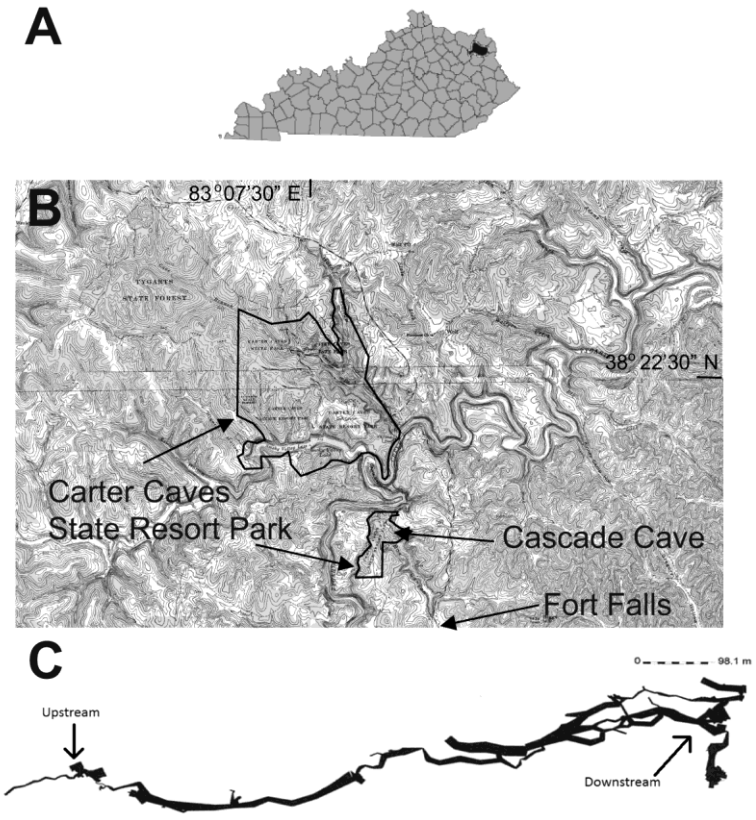



Figure A 3-1: (A) Black area denotes Carter County, Kentucky. (B) Spliced topographic maps from the United States Geological Survey showing the location of Carter Caves State Resort Park boundaries, relative location of Cascade Cave and Fort Falls. Specific location details are withheld at the request of the park. (C) A generalized line-plot map of the Cascade Cave system, including Cascade Cave (downstream), Sandy Cave, and Jones Cave (upstream). Map provided by Dr. Horton H. Hobbs, III, and the Wittenberg University Speleological Society, Springfield, Ohio (USA).

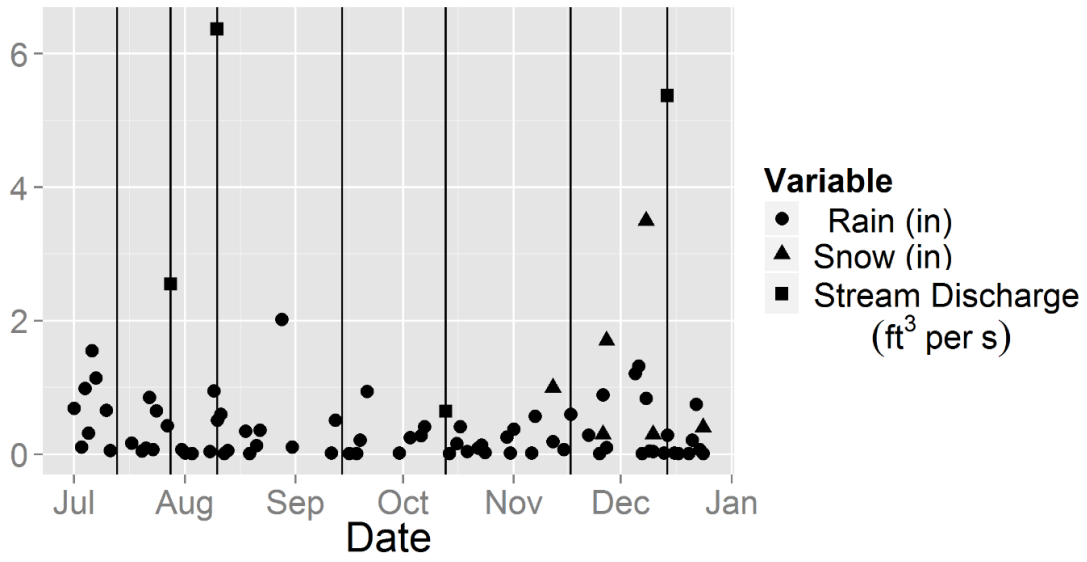



Figure A 3-2: Precipitation events over a 5 month period from the Olive Hill citizen scientist. Each vertical line represents the date of a sample event. Stream discharge measured at the downstream location of the Cascade Cave system. All non-precipitation events have been removed. Stream discharge rates below detection limit are not plotted.

Figure A 3-3: (Top) Bio-Traps® deployed at the upstream location in Cascade Cave System. (Bottom) Bio-Traps® deployed at the downstream location in Cascade Cave System.

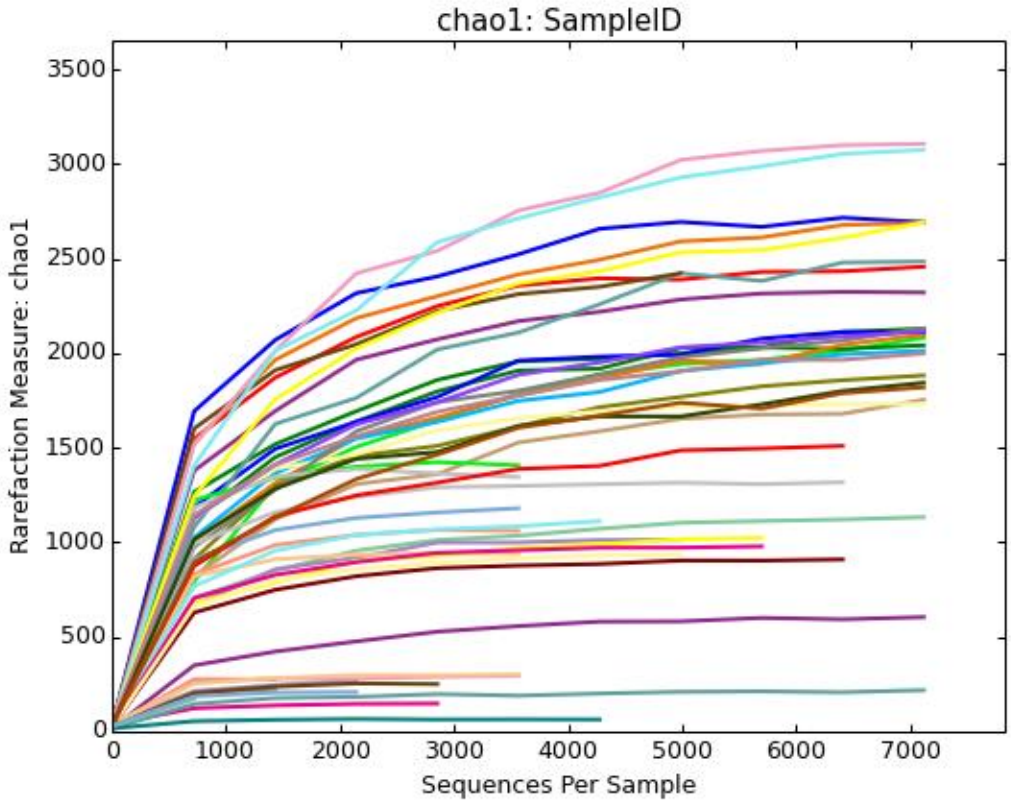



Figure A 3-4: Rarefaction curves generated by QIIME using the Chao1 diversity metric. The calculation is cut after 7000 sequences. The samples with the lowest diversity are the sediment samples.

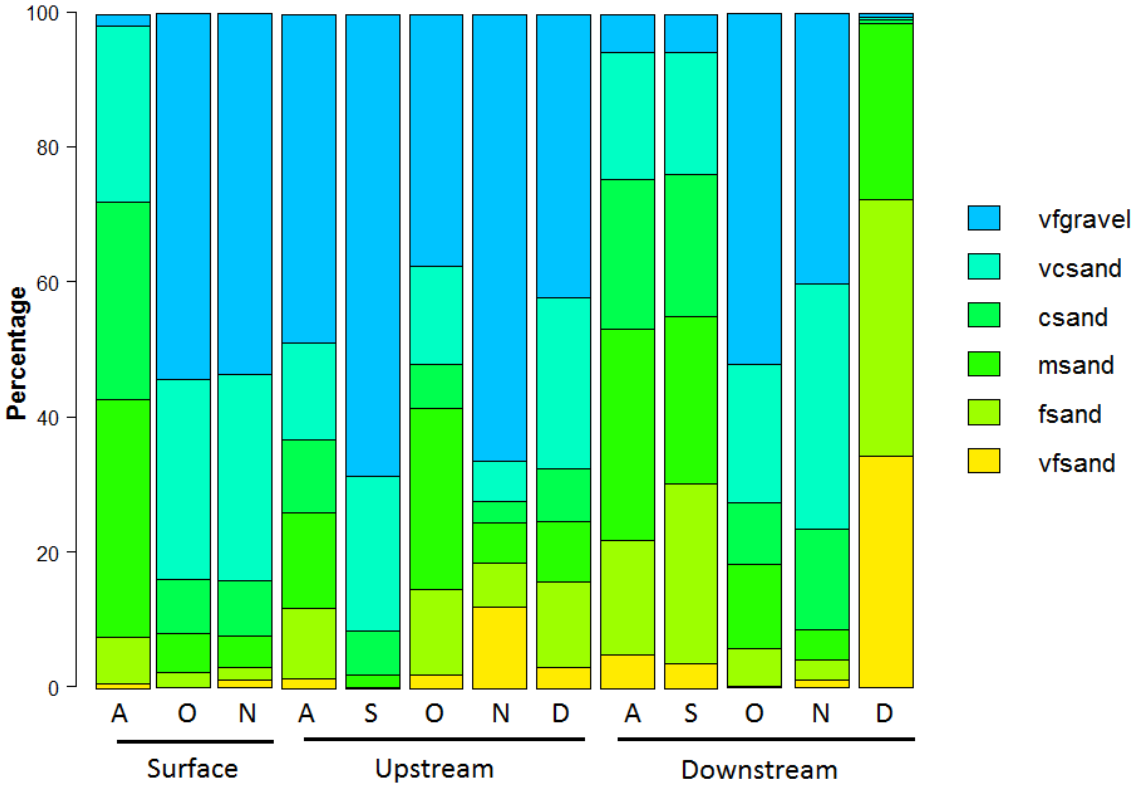



Figure A 3-5: Percent grain size distribution of all sediment samples. Produced with the G2SD package for R.

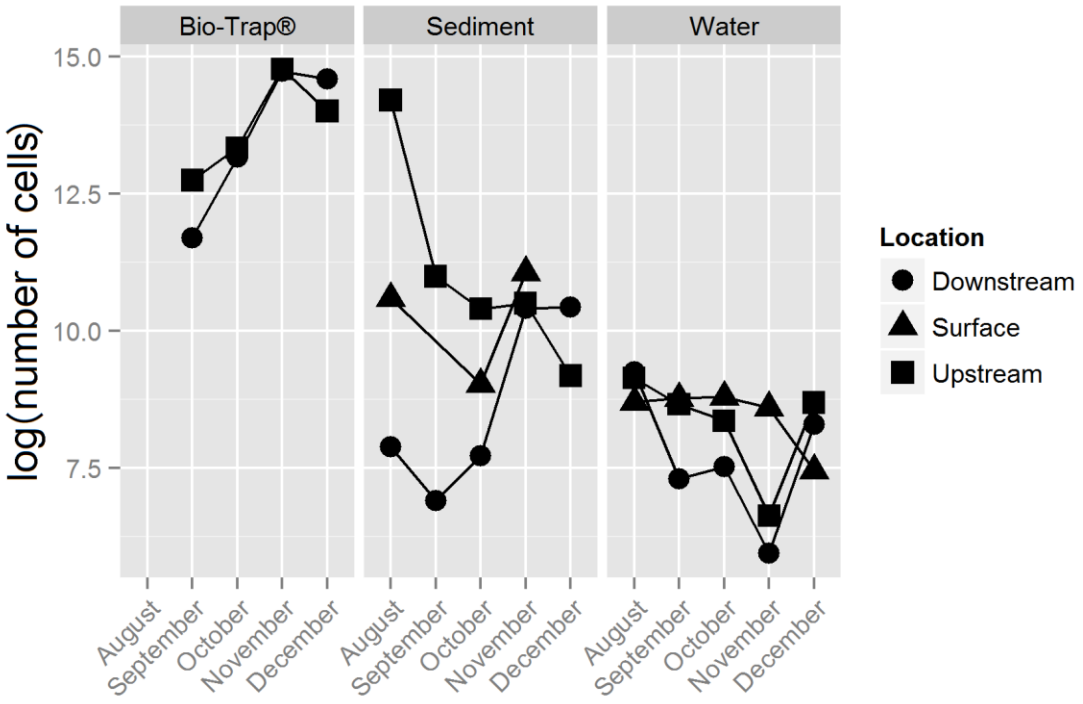
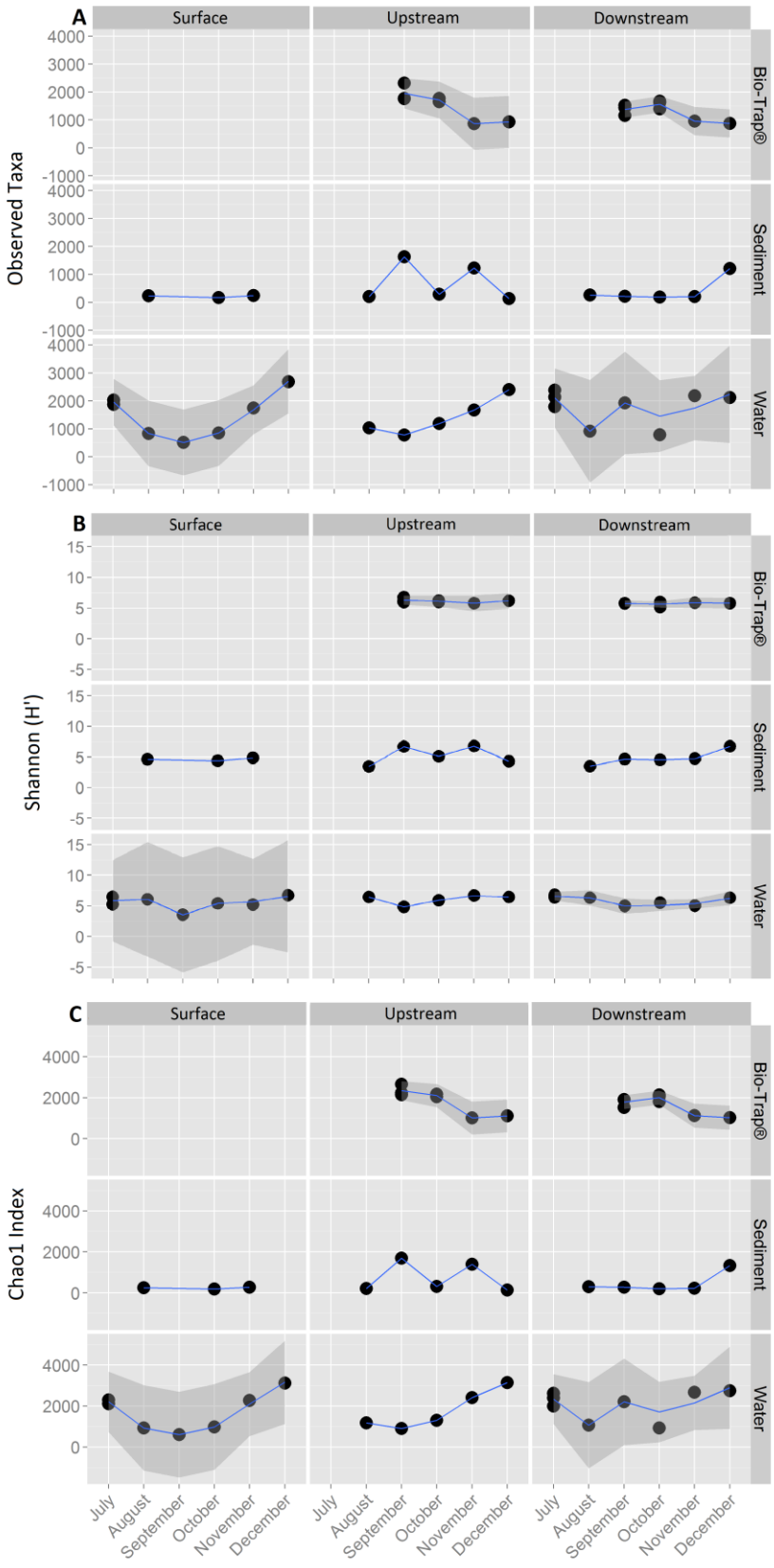



Figure A 3-6: Bio-Trap®, sediment, and water biomass estimates from qPCR results, displayed as log(number of cells) over time for each type of sample at the surface, upstream, and downstream locations.

Figure A 3-7: Alpha-diversity richness and evenness indices of (A) Observed, (B) Shannon, and (C) Chao1, by sample type and location over a six month period.


A
Surface
Upstream
Downstream
Observed Taxa
Bio-Trap®
Sediment
Water
B
Shannon (H')
C
Chao1 Index
July
August
September
October
November
December

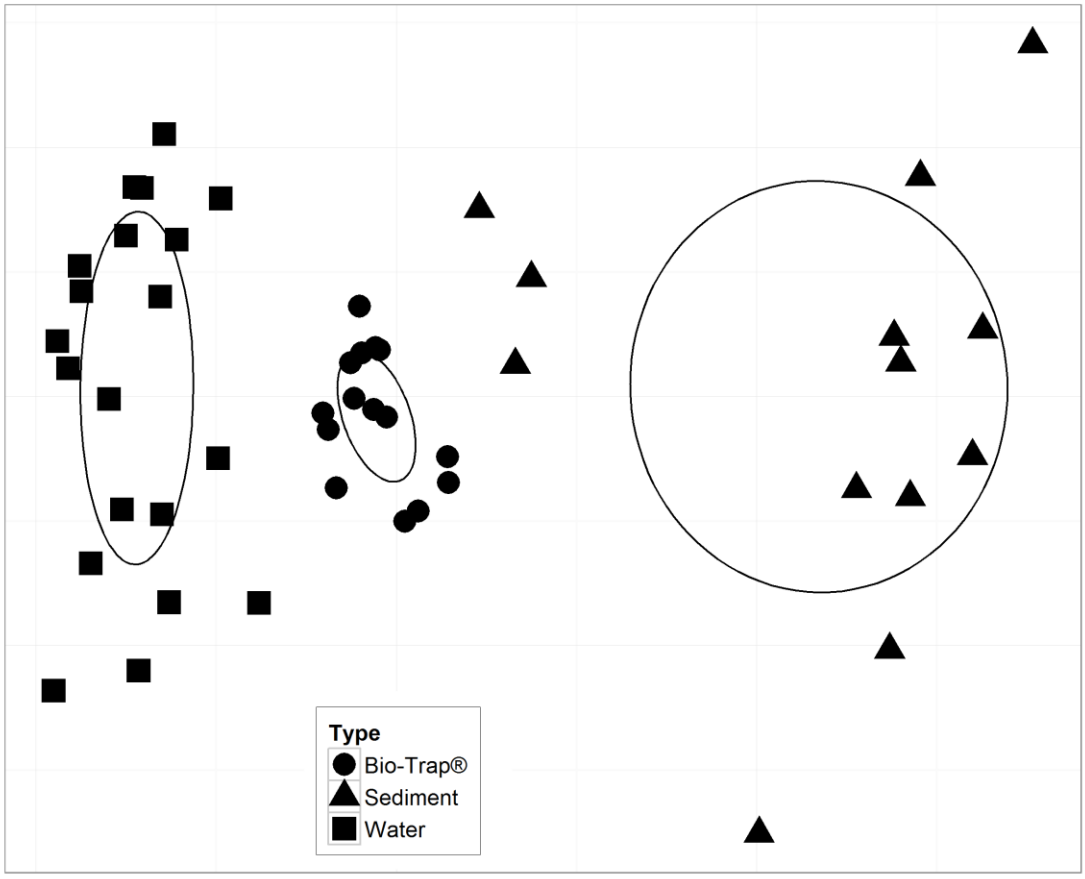



Figure A 3-8: Nonmetric multidimensional scaling (NMDS) plot based on a Bray-Curtis dissimilarity matrix; stress = 0.082. Ellipses represent the standard error of the weighted average of scores of samples, and the direction of the principal axis of the ellipse is defined by the weighted correlation of samples. There were no statistically significant environmental vectors (p-value<0.05).

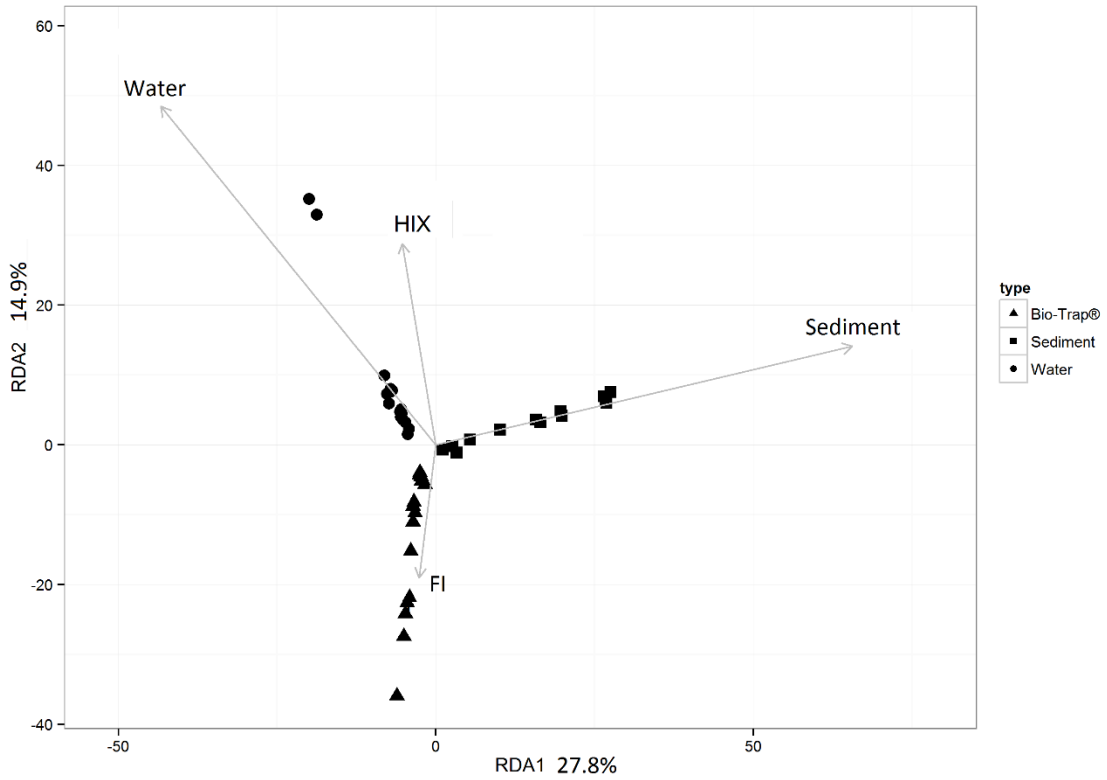



Figure A 3-9: Redundancy analysis (RDA) of the culled OTU dataset as a function of the fluorescence indices HIX and FI. Significance of each RDA axis was calculated with the RDAsignificance function from the BiodiversityR package for R (Kindt, 2014).

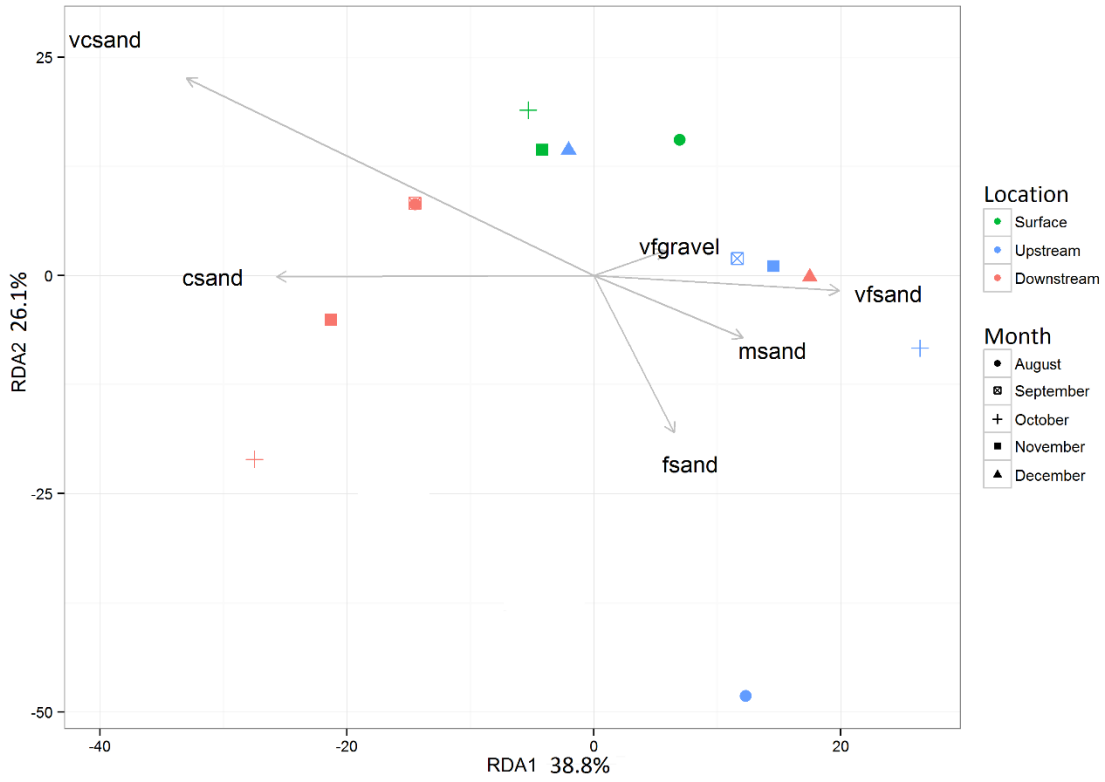
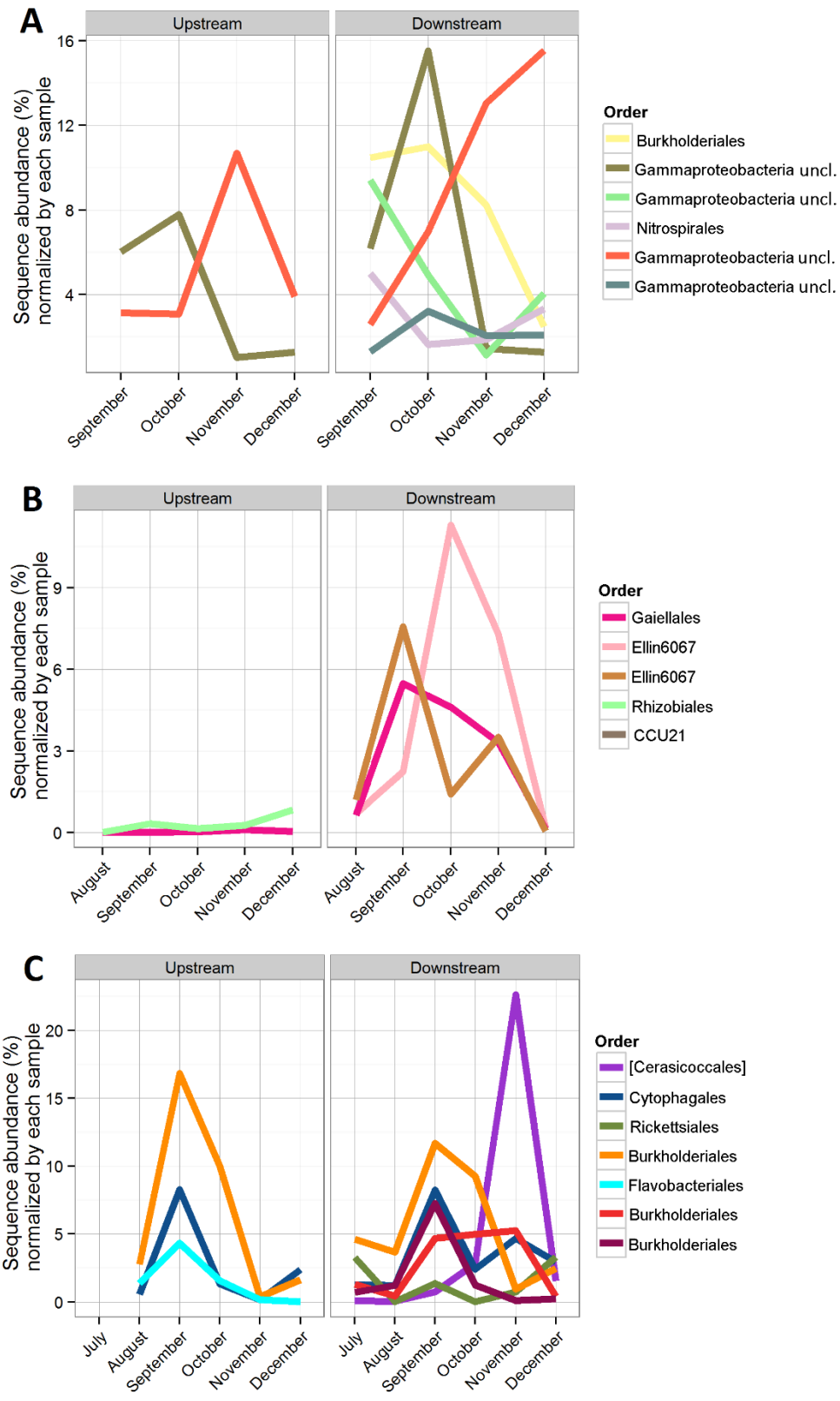



Figure A 3-10: Redundancy analysis (RDA) of the culled OTU dataset as a function of the grain size analysis from the G2SD package gran_stat function output (Gallon and Fournier, 2013). Significance of each RDA axis was calculated with the RDAsignificance function from the BiodiversityR package for R (Kindt, 2014).

Figure A 3-11: Sequence abundance of OTUs present for the duration of the study, normalized by the total abundance of sequences in the sample. Each OTU is colored by its taxonomic order, and the same color represents the same OTU across locations. (A) Bio-Trap® samples, triplicates were averaged for the sequence abundances; (B) Sediment samples; (C) Water samples.


A
Upstream
Downstream
Sequence abundance (%) normalized by each sample
16
12
8
4
September
October
November
December
Order
Burkholderiales
Gammaproteobacteria uncl.
Gammaproteobacteria uncl.
Nitrospirales
Gammaproteobacteria uncl.
Gammaproteobacteria uncl.
B
Upstream
Downstream
Sequence abundance (%) normalized by each sample
9
6
3
0
August
September
October
November
December
Order
Gaiellales
Ellin6067
Ellin6067
Rhizobiales
CCU21
C
Upstream
Downstream
Sequence abundance (%) normalized by each sample
20
15
10
5
0
July
August
September
October
November
December
Order
[Cerasicoccales]
Cytophagales
Rickettsiales
Burkholderiales
Flavobacteriales
Burkholderiales
Burkholderiales

**TABLES**

Table A 3-1: Summary of pyrosequencing data for each of the samples used in this study. Average seq. length after trimming and % Chimeric were calculated in five batches.

| Sample Name | Small Read Archive Run Accession Number | Number of Seqs. (raw) | Number of Seqs. After Trimming | Average Seq. Length After Trimming | % Chimeric | OTUs (95% sequence identity) | Shannon Index | Chao1 Index |
|---|---|---|---|---|---|---|---|---|
| CCRB.13d1 | SAMN03451539 | 7969 | 7191 | 492.8 | 25.1 | 875 | 5.77 | 1023.23 |
| CCRB.13d2 | SAMN03451542 | 6438 | 5652 | 492.8 | 25.1 | 932 | 6.19 | 1113.24 |
| CCRB.13N1 | SAMN03451551 | 8940 | 8061 | 492.8 | 25.1 | 962 | 5.84 | 1124.91 |
| CCRB.13N2 | SAMN03451554 | 7002 | 6276 | 492.8 | 25.1 | 867 | 5.78 | 1011.06 |
| CCRB.13O1A | SAMN03451561 | 15907 | 13349 | 471.1 | 38.1 | 1668 | 5.97 | 2130.62 |
| CCRB.13O1B | SAMN03451562 | 14262 | 11779 | 471.1 | 38.1 | 1392 | 5.14 | 1813.07 |
| CCRB.13O1C | SAMN03451563 | 17481 | 14329 | 471.1 | 38.1 | 1625 | 5.70 | 2063.34 |
| CCRB.13O2A | SAMN03451566 | 14414 | 12146 | 471.1 | 38.1 | 1772 | 6.21 | 2164.13 |
| CCRB.13O2B | SAMN03451567 | 13333 | 11369 | 471.1 | 38.1 | 1654 | 5.99 | 2046.55 |
| CCRB.13O2C | SAMN03451568 | 5310 | 4624 | 471.1 | 38.1 | 62 | 2.62 | 62.50 |
| CCRB.13S2A | SAMN03451573 | 12848 | 10622 | 471.1 | 38.1 | 1441 | 5.75 | 1913.68 |
| CCRB.13S2B | SAMN03451574 | 16369 | 13327 | 471.1 | 38.1 | 1523 | 5.74 | 1912.62 |
| CCRB.13S2C | SAMN03451575 | 9586 | 7755 | 471.1 | 38.1 | 1159 | 5.78 | 1520.17 |
| CCRB.13S3A | SAMN03451577 | 13941 | 11884 | 471.1 | 38.1 | 1758 | 6.13 | 2147.56 |
| CCRB.13S3B | SAMN03451578 | 16328 | 13889 | 471.1 | 38.1 | 2322 | 6.78 | 2661.20 |
| CCRB.13S3C | SAMN03451579 | 14418 | 12319 | 471.1 | 38.1 | 1767 | 5.94 | 2228.02 |
| CCRS.13A1 | SAMN03451533 | 4104 | 3731 | 469.9 | 34.6 | 258 | 3.44 | 294.85 |
| CCRS.13A2 | SAMN03451535 | 13143 | 11881 | 469.9 | 34.6 | 207 | 4.60 | 212.27 |
| CCRS.13A3 | SAMN03451537 | 3460 | 3099 | 469.9 | 34.6 | 230 | 6.72 | 248.00 |
| CCRS.13d1 | SAMN03451540 | 6022 | 5350 | 469.9 | 34.6 | 1214 | 4.27 | 1331.75 |
| CCRS.13d2 | SAMN03451543 | 3663 | 3176 | 469.9 | 34.6 | 131 | 4.71 | 142.40 |
| CCRS.13N1 | SAMN03451552 | 1736 | 1497 | 469.9 | 34.6 | 203 | 6.76 | 225.67 |
| CCRS.13N2 | SAMN03451555 | 6021 | 1693 | 469.9 | 34.6 | 1226 | 4.85 | 1389.20 |
| CCRS.13N3 | SAMN03451557 | 1905 | 5182 | 469.9 | 34.6 | 246 | 4.51 | 270.05 |
| CCRS.13O1 | SAMN03451559 | 2581 | 2182 | 469.9 | 34.6 | 186 | 5.09 | 205.33 |

Table A3-1 Continued

| Sample Name | Small Read Archive Run Accession Number | Number of Seqs. (raw) | Number of Seqs. After Trimming | Average Seq. Length After Trimming | % Chimeric | OTUs (95% sequence identity) | Shannon Index | Chao1 Index |
|---|---|---|---|---|---|---|---|---|
| CCRS.13O2 | SAMN03451564 | 4805 | 4243 | 469.9 | 34.6 | 289 | 4.35 | 306.00 |
| CCRS.13O3 | SAMN03451569 | 1561 | 1359 | 469.9 | 34.6 | 166 | 4.65 | 190.80 |
| CCRS.13S1 | SAMN03451571 | 2520 | 2222 | 469.9 | 34.6 | 213 | 6.69 | 273.05 |
| CCRS.13S2 | SAMN03451572 | 14811 | 13077 | 469.9 | 34.6 | 1633 | 5.78 | 1698.35 |
| CCRW.13A1 | SAMN03451534 | 5120 | 4521 | 492.8 | 25.1 | 918 | 6.30 | 1063.75 |
| CCRW.13A2 | SAMN03451536 | 5718 | 5093 | 492.8 | 25.1 | 1030 | 6.42 | 1179.08 |
| CCRW.13A3 | SAMN03451538 | 4896 | 4261 | 492.8 | 25.1 | 836 | 6.08 | 935.23 |
| CCRW.13D1 | SAMN03451541 | 15327 | 13513 | 495.6 | 29.2 | 2117 | 6.33 | 2748.12 |
| CCRW.13D2 | SAMN03451544 | 17377 | 15459 | 495.6 | 29.2 | 2399 | 6.44 | 3139.65 |
| CCRW.13D3 | SAMN03451545 | 19492 | 17461 | 495.6 | 29.2 | 2684 | 6.70 | 3117.50 |
| CCRW.13JA1 | SAMN03451546 | 14537 | 12877 | 492.9 | 25.1 | 2139 | 6.61 | 2393.11 |
| CCRW.13JA3 | SAMN03451547 | 14016 | 12026 | 492.9 | 25.1 | 1787 | 6.37 | 1999.21 |
| CCRW.13JB1 | SAMN03451548 | 16038 | 14340 | 492.9 | 25.1 | 2388 | 6.78 | 2623.98 |
| CCRW.13JB3 | SAMN03451549 | 15484 | 13593 | 492.9 | 25.1 | 2032 | 6.48 | 2303.13 |
| CCRW.13JC1 | SAMN03451550 | 22010 | 19737 | 492.9 | 25.1 | 1876 | 5.21 | 2106.68 |
| CCRW.13N1 | SAMN03451553 | 22412 | 19960 | 495.6 | 29.2 | 2184 | 5.03 | 2672.43 |
| CCRW.13N2 | SAMN03451556 | 6823 | 6012 | 495.6 | 29.2 | 1676 | 6.66 | 2410.48 |
| CCRW.13N3 | SAMN03451558 | 17428 | 15493 | 495.6 | 29.2 | 1746 | 5.18 | 2272.56 |
| CCRW.13O1 | SAMN03451560 | 6546 | 5881 | 492.8 | 25.1 | 785 | 5.50 | 924.38 |
| CCRW.13O2 | SAMN03451565 | 8481 | 7597 | 492.8 | 25.1 | 1188 | 5.91 | 1299.04 |
| CCRW.13O3 | SAMN03451570 | 8331 | 7014 | 492.8 | 25.1 | 851 | 5.38 | 979.13 |
| CCRW.13S3 | SAMN03451576 | 8333 | 7513 | 492.8 | 25.1 | 781 | 4.81 | 905.05 |
| CCRW.13SF3 | SAMN03451580 | 10862 | 9876 | 492.8 | 25.1 | 513 | 3.53 | 606.31 |
| CCRW.13SL3 | SAMN03451581 | 26016 | 23437 | 495.6 | 29.2 | 1924 | 4.98 | 2204.79 |
| CCRB.13d1 | SAMN03451539 | 7969 | 7191 | 492.8 | 25.1 | 875 | 5.77 | 1023.23 |

Table A 3-2: Geochemical and hydrological data from each sample. NM = not measured. DO = dissolved organic carbon measured as the difference between dissolved non-purgeable organic carbon and total inorganic carbon. Total N = total dissolved nitrogen measured as all N compounds present in a sample, including N in DOM. Water flow rate, BDL = below detection limit for the flow measurements, NC = not calculated because velocity measurements were below detection. FI = Fluorescence index, see text for description. HIX = Humification index, see text for description. BIX = Biological index, see text for description.

| Sample Month | Sample Location | Sample Name | temp °C | pH | Alkalinity mg/L | DOC mg/L | Total N mg/L | Cl mg/L | $SO_4^{2-}$ mg/L | $NO_3^-$ mg/L | Na mg/L | Mg mg/L | Ca mg/L | Flow rate m/s | Discharge m3/s | FI | HIX | BIX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| July | Surface | CCRW.13JB1 | 17.1 | 7.1 | 56.85 | 3.08 | 0.46 | 10.1 | 30.1 | BDL | 5.99 | 2.71 | 9.23 | BDL | BDL | 1.97 | 6.97 | 0.63 |
| July | Downstream | CCRW.13JA1 | 21.5 | 7.4 | 58.07 | 6.62 | 0.46 | 5.61 | 19.09 | BDL | 5.23 | 4.81 | 17.45 | BDL | BDL | 2.0 | 10.33 | 0.58 |
| July | Surface | CCRW.13JB3 | 21.5 | 7.5 | 84.66 | NM | NM | 25.58 | 33.34 | BDL | 12.24 | 10.25 | 8.34 | 0.78 | 0.21 | 2.01 | 11.08 | 0.61 |
| July | Downstream | CCRW.13JA3 | 17.3 | 7.3 | 84.42 | NM | NM | 18.68 | 27.05 | 1.09 | 8.93 | 7.66 | 20.67 | 2.23 | 0.6 | 2.0 | 10.33 | 0.58 |
| August | Surface | CCRW.13A3 | 24 | 7.2 | 84.42 | 3.0 | 0.77 | 17.84 | 31.03 | 1.41 | 8.11 | 8.50 | 10.22 | 5.25 | 0.82 | 2.01 | 8.51 | 0.65 |
| August | Upstream | CCRW.13A2 | 20.3 | 7.3 | 93.6 | 5.2 | 0.99 | 11.6 | 24.31 | 1.28 | 7.74 | 6.89 | 14.78 | 4.4 | 0.69 | 1.99 | 8.29 | 0.64 |
| August | Downstream | CCRW.13A1 | 18.3 | 7.3 | 87.84 | NM | NM | 10.59 | 17.01 | 1.03 | 6.69 | 5.56 | 15.3 | 6.23 | 1.68 | 1.96 | 8.88 | 0.64 |
| September | Surface | CCRW.13SF3 | 22 | 7.4 | 107.36 | 3.0 | 0.31 | 24.23 | 28.62 | 0.27 | 12.53 | 10.46 | 41.48 | BDL | BDL | 2.1 | 0.93 | 0.68 |
| September | Upstream | CCRW.13S3 | 17.9 | 7.5 | 110.28 | 2.85 | 0.55 | 34.66 | 32.61 | BDL | 10.4 | 8.93 | 40.79 | BDL | BDL | 2.14 | 0.93 | 0.66 |
| September | Downstream | CCRW.13SL3 | 18 | 7.4 | 127.36 | 1.79 | 0.65 | 19.07 | 21.69 | 1.55 | 9.52 | 7.39 | 47.09 | BDL | BDL | 2.11 | 0.91 | 0.66 |
| October | Surface | CCRW.13O3 | 16.9 | 7.7 | 125.41 | 0.27 | 0.28 | 30.12 | 29.37 | BDL | 11.34 | 10.24 | 38.62 | BDL | BDL | 2.11 | 0.92 | 0.69 |
| October | Upstream | CCRW.13O2 | 15.2 | 7.7 | 152.01 | 2.54 | 0.33 | 23.78 | 30.94 | BDL | 10.22 | 10.19 | 41.5 | BDL | BDL | 2.12 | 0.93 | 0.69 |
| October | Downstream | CCRW.13O1 | 13.7 | 7.8 | 133.95 | 3.81 | 0.57 | 22.59 | 20.37 | 1.05 | 7.64 | 7.05 | 46.21 | BDL | BDL | 2.12 | 0.93 | 0.69 |
| November | Surface | CCRW.13N3 | 10.1 | 7.5 | 105.65 | 1.48 | 0.32 | 23.49 | 38.44 | BDL | 11.73 | 9.93 | 32.02 | 0.76 | 0.2 | 2.13 | 0.91 | 0.67 |
| November | Upstream | CCRW.13N2 | 8.7 | 7.5 | 88.57 | 2.48 | 0.40 | 20.09 | 34.61 | BDL | 9.68 | 9.16 | 35.66 | 3.2 | 0.05 | 2.19 | 0.91 | 0.69 |
| November | Downstream | CCRW.13N1 | 10 | 7.5 | 130.05 | 3.99 | 0.42 | 22.28 | 29.44 | 0.82 | 9.64 | 7.73 | 40.22 | 1.67 | 0.14 | 2.16 | 0.9 | 0.7 |
| December | Surface | CCRW.13D3 | 4.3 | 7.1 | 93.2 | 1.7 | 0.93 | 17.9 | 35.23 | BDL | 9.44 | 6.09 | 9.00 | 3.7 | 10.18 | 2.04 | 0.9 | 0.59 |
| December | Upstream | CCRW.13D2 | 3.8 | 7.3 | 62.46 | 1.67 | 0.98 | 19.45 | 33.77 | BDL | 14.59 | 7.35 | 13.16 | 4.95 | 0.7 | 2.12 | 0.9 | 0.66 |
| December | Downstream | CCRW.13D1 | 4.9 | 7.5 | 93.2 | 1.69 | 0.56 | 18.33 | 28.32 | BDL | 9.29 | 6.52 | 15.63 | 2.95 | 1.32 | 2.2 | 0.9 | 0.66 |

Table A 3-3: Results from ANOVA of CDOM fluorescence index by month and by location.

| | Bio-Trap® | | | Bio-Trap® | |
|---|---|---|---|---|---|
| | Upstream August | Upstream December | | Downstream August | Downstream December |
| Water | 153 | 216 | Water | 110 | 199 |
| Sediment | 12 | 26 | Sediment | 13 | 212 |

Table A 3-4: Number of shared OTUs by taxonomic Phylum and Class. Numbers in parentheses represent the percent abundance (sequences in shared OTUs normalized by the shared group total).

| Phylum | Class | Bio-Trap® | | | Sediment | | | | Water | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Up-stream | Down-stream | Shared | Surface | Up-stream | Down-stream | Shared | Surface | Up-stream | Down-stream | Shared |
| Acidobacteria | Acidobacteria-6 | 8 (1.3) | 4 (1.1) | 4 (1.3) | 1 (4.2) | 1 (0.2) | 1 (0.4) | - | - | - | - | - |
| Acidobacteria | Chloracidobacteria | 3 (0.2) | 1 (0) | 1 (0.1) | - | - | - | - | - | - | - | - |
| Actinobacteria | Acidimicrobiia | - | - | - | 2 (2.8) | - | - | - | - | - | - | - |
| Actinobacteria | Thermoleophilia | - | - | - | 5 (3.8) | - | 3 (4.1) | - | - | - | - | - |
| Actinobacteria | Actinobacteria | - | - | - | - | - | - | - | 1 (0.1) | 3 (1.1) | 4 (1.6) | 2 (0.4) |
| Chloroflexi | Ellin6529 | - | - | - | 1 (1.3) | - | 1 (0.2) | - | - | - | - | - |
| Chloroflexi | P2-11E | - | - | - | 2 (0.5) | - | - | - | - | - | - | - |
| Chloroflexi | Anaerolineae | 3 (0.3) | - | - | - | - | - | - | - | - | - | - |
| Chloroflexi | Chloroflexi | 1 (0) | - | - | - | - | - | - | - | - | - | - |
| Proteobacteria | Alphaproteobacteria | 88 (14.9) | 61 (12.8) | 48 (12.0) | 3 (3.1) | - | 8 (5) | - | - | 9 (1.3) | 2 (2) | 2 (1.7) |
| Proteobacteria | Betaproteobacteria | 52 (12.8) | 55 (14.2) | 38 (13.3) | 4 (1.6) | 1 (0) | 3 (2.6) | - | 53 (20.2) | 39 (14.6) | 41 (18.2) | 19 (13.1) |
| Proteobacteria | Deltaproteobacteria | 1 (0) | 1 (0) | 11 (17.3) | - | - | - | - | - | - | - | - |
| Proteobacteria | Gammaproteobacteria | 18 (16.6) | 24 (18.6) | - | - | - | - | - | - | 3 (0.3) | 2 (0) | - |
| Proteobacteria | Epsilonproteobacteria | - | - | - | - | - | - | - | 1 (0.2) | - | - | - |
| Proteobacteria | NA | - | 2 (0.8) | - | - | - | - | - | - | - | - | - |
| Thermi | Deinococci | 1 (0) | - | - | - | - | - | - | - | - | - | - |
| Bacteroidetes | Saprospirae | 3 (0.6) | 5 (0.7) | 1 (0.5) | - | - | - | - | - | - | 1 (0.1) | - |
| Bacteroidetes | Cytophagia | 5 (0.9) | 7 (1.5) | 5 (1.3) | - | - | - | - | 1 (1.3) | 4 (3.2) | 2 (4) | 1 (3.6) |
| Bacteroidetes | Sphingobacteriia | 1 (0) | 4 (0.4) | - | - | - | - | - | - | - | - | 1 (1.0) |
| Bacteroidetes | Flavobacteriia | - | - | - | - | - | - | - | 1 (0.7) | 4 (1.5) | 1 (0.8) | - |
| Gemmatimonadetes | Gemmatimonadetes | 1 (0) | 2 (0.1) | - | - | - | - | - | - | - | - | - |
| Nitrospirae | Nitrospira | 5 (2.6) | 6 (8.6) | 5 (6.4) | - | - | - | - | - | 1 (0.1) | - | - |
| Planctomycetes | OM190 | 1 (0) | 1 (0) | - | - | - | - | - | - | - | - | - |
| Planctomycetes | Planctomycetia | 9 (0.9) | 6 (1.1) | 1 (0.7) | - | - | - | - | - | - | - | - |
| Planctomycetes | vadinHA49 | - | - | - | - | - | - | - | 1 (0) | - | - | - |

Table A 3-5: Number of OTUs shared between Bio-Trap® samples and Water/Sediment environment types in both August and December for both locations inside the cave.

| | Bio-Trap® | | | Bio-Trap® | |
|---|---|---|---|---|---|
| | Upstream August | Upstream December | | Downstream August | Downstream December |
| Water | 153 | 216 | Water | 110 | 199 |
| Sediment | 12 | 26 | Sediment | 13 | 212 |

# Chapter 4
# METAGENOMIC VIEW OF ECOLOGICAL DIVERSITY AMONG MICROBIAL COMMUNITIES IN AN EPIGENIC CAVE SYSTEM

This chapter has been formatted as a manuscript to be submitted to the ISME Journal. KBD and ASE designed the study and sampling protocol. KBD collected and analyzed the data, and KBD and ASE interpreted the data and wrote the manuscript. ASE provided funding from the Jones Endowment. KJ Price provided statistical analyses using the software SAS and provided the methods section for those analyses. He will be a co-author when the manuscript is submitted for publication. Additional acknowledgements are included at the end of the chapter.


**ABSTRACT**

In many epigenic cave systems, the main source of energy and nutrients are products of plant litter decomposition transported from the surface into the subsurface. Although it is well accepted that microorganisms are responsible for mediating the distribution of energy and nutrients in these ecosystems, microbial taxonomic and functional diversity associated with most biogeochemical cycles in different cave habitats, such as planktonic versus sediment-attached, are poorly understood. Here, we examined 31 metagenomes obtained from planktonic communities, communities attached to sediments, and communities attached to *in situ* artificial substrates to search for ecological functions associated with the degradation of plant litter, which is one of the most commonly hypothesized biogeochemical processes occurring in epigenic caves that affect subsurface ecosystems. Our analyses of functional gene structure revealed unprecedented information about microbial communities separated by water, sediment, or artificial habitat during a five-month period. Also, the genes encoding for the degradation of certain components of plant litter were present despite environmental

disturbances and the lack of a stable microbial community in any one of the sampled environments. This study provides new insight into functional capabilities of planktonic versus attached microorganisms in subsurface aquatic systems, as well as evidence for some degradation pathways that occur throughout the cave ecosystem. These potential pathways may play a role in the heterotrophic production of $CO_2$ in the system that could affect karst development through carbonate dissolution.

## INTRODUCTION

Microorganisms are responsible for the biogeochemical cycling of elements that affect the quality of water (Spizzico et al., 2005). Compared to research done to understand microbial diversity and function in groundwater (Hazen et al., 1991; Farnleitner et al., 2005; Pronk et al., 2008; Griebler and Lueders, 2009; Hahn and Fuchs, 2009; Shabarova and Pernthaler, 2010; Lin et al., 2012; Zhou et al., 2012; Shabarova et al., 2014; Hug et al., 2015; Smith et al., 2015), particularly for contaminated environments (Dussart-Baptista et al., 2003; Hemme et al., 2010; Yagi et al., 2010), our knowledge of microbial functional diversity in epigenic cave habitats, including water moving through a system or sediments stored within conduits, have received less attention (Goldscheider et al., 2006; Griebler and Lueders, 2009; Morasch, 2013;Byl et al., 2014). To date, there have been four metagenomics studies from caves that focus on microbial communities from sulfuric acid systems with acidic conditions or from nitrogen-dominated groundwater systems (Bhullar et al., 2012; Jones et al., 2012; Tetu et al., 2013; Ortiz et al., 2014). However, these systems represent about 10% of known caves worldwide (Palmer, 2007).

In contrast, epigenic cave systems are more common, forming from carbonic acid dissolution in the shallow subsurface of a karst landscape as interconnected, self-evolving hydrological passageways and conduits associated with sinking streams (Palmer, 2007; Ford and Williams, 2013). Only one metagenomics study has been completed from this type of karst system, although the study focused on oligotrophic and chemolithoautotrophic microbial communities associated with speleothems (Ortiz et al., 2014). No metagenomics research has been done from flowing karst cave streams, which are among the most common types of cave habitats encountered underground (Palmer, 2007; Culver and Pipan, 2014).

Because of the absence of photosynthesis in dark habitats, subsurface ecosystems, including caves, depend on energy and nutrition sourced from plant litter decomposition. Plant litter decomposition is one of the biosphere's most complex ecological processes (Sinsabaugh et al., 2002). Primary and secondary substrates and metabolites released from soils, exuded by plant roots, and resulting from the degradation of plant litter are transported into the subsurface as allochthonous particulate and dissolved organic matter (OM) (Simon et al., 2003; Simon et al., 2007; Cooney and Simon, 2009; Simon et al., 2010; Venarsky et al., 2012). The amount and quality of allochthonous OM that reaches the subsurface is controlled by hydrological connectivity to the surface (Simon et al., 2007; Ford and Williams, 2013). In the case of karstic cave systems, headwater streams containing terrigenous OM directly sink or become pirated into the subsurface network of passages and conduits (e.g., Brooks et al., 1999; Jardine et al., 2006). Cave ecosystems dependent upon allochthonous OM are generally characterized as energy-limited (Venarsky et al., 2014) because allochthonous inputs into cave

systems are less than those of surface streams due to a lower number of direct riparian inputs (Graening and Brown, 2003). Also, the allochthonous OM inputs are generally of lesser quality in subsurface environments due to enhanced OM biological processing in surface and soil habitats (Graening and Brown, 2003; Engel, 2010). Gradients of resource availability along the surface-to-cave stream flowpaths are possible and often impact species biomass, diversity, and function within caves (Venarsky et al., 2012).

The aim of this study was to assess potential metabolisms associated with plant litter degradation in the Cascade Cave stream ecosystem, in northeastern Kentucky, from analyzing functional and taxonomic profiles from multiple metagenomes obtained over time from flowing water, sediments, and biofilms formed on artificial substrates (Bio-Trap® samplers). Our previous research demonstrated that distinct microbial communities persisted in the stream and sediments, regardless of hydrological disturbance (Brannen-Donnelly and Engel, 2015). Community succession was also evident based on changes in community composition over time on Bio-Traps® (Brannen-Donnelly and Engel, 2015). In general, because sediment- or rock-attached microbial communities are thought to be responsible for most activities in aquatic subsurface systems (Hazen et al., 1991; Flynn et al., 2008; Wilhartitz et al., 2009), and the roles of planktonic microorganisms in groundwater systems have been largely unknown to date, we were motivated to test the hypothesis that Bio-Trap® communities would be functionally similar to planktonic communities initially, but transition to being more functionally similar to attached sediment communities over time. Moreover, metabolic strategies and potential capabilities associated with plant litter degradation would be more prevalent for sediment-

attached communities but would differ along the cave stream as OM quality and quantity changed.

## METHODS

### Sampling, DNA extraction, sequencing, and annotation

The study occurred from August to December 2013 in the Cascade Cave system within Carter Caves State Resort Park in Carter County, Kentucky. A general description of the cave system is described in Engel and Engel (2009) and Brannen-Donnelly and Engel (2015). Briefly, three surveyed caves comprise a system within the James Branch stream watershed. The surface sampling location was at Fort Falls, where surface water sinks into the subsurface karst before discharging approximately 1.5 km later into Tygart's Creek at local base-level (Brannen-Donnelly and Engel, 2015). The upstream sampling location was at Jones Cave and the downstream sampling location was at the Lake Room in Cascade Cave. In addition to monthly water and sediment sampling, as described in Brannen-Donnelly and Engel (2015), unbaited Bio-Trap® samplers (Microbial Insights, Knoxville, TN, USA) were deployed in triplicate in August at the cave upstream and downstream sampling locations. Subsamples were collected from the Bio-Traps® monthly from September-December. Methods and results for sediment characterization, major dissolved ion geochemistry, and other analyses, including dissolved organic carbon (DOC) concentration, chromophoric dissolved OM (CDOM) assessment based on fluorescence spectroscopy, and calculations of qualitative and comparative indices, are described in Brannen-Donnelly and Engel (2015).

Total environmental nucleic acids extraction and quality screening for water, sediment, and Bio-Trap® samples are described in Brannen-Donnelly and Engel (2015). Approximately 50 ng of DNA from each of the 31 samples was prepared using Nextera DNA Sample Preparation Kits (Illumina, San Diego, CA, USA), following the manufacturer's instructions, at the Molecular Research LP laboratory (www.mrdnalab.com; Shallowater, TX, USA). Shotgun metagenome library concentrations were measured using the Qubit® dsDNA HS Assay Kit (Life Technologies). Reads were sequenced using an Illumina HiSeq 2500 system. All paired-end reads were submitted to the Metagenomics Analysis Server (MG-RAST) v.3.6 (http://metagenomics.anl.gov/) for pipeline analysis of trimming, dereplication, DRISEE (Keegan et al., 2012), screening, gene calling (Rho et al., 2010), and annotation using default settings (Meyer et al., 2008). All metagenomes generated for this study are publicly available through MG-RAST under their MG-RAST ID numbers (Table A4-1; tables located in Appendix III).

**Metagenome examination**

Several different standard methods can be used to analyze taxonomic and functional classifications of metagenomic data. Taxonomy of unassembled metagenomic reads from MG-RAST data was assessed from the representative hit classification of the SEED annotation source, using the computational defaults (1e-5 e-value, 60% minimum identity; Meyer *et al.*, 2008). Unassembled read assignments can offer an acceptable and comprehensive view of the functional capabilities of the microbial community (Delmont et al., 2012; Montana et al., 2012; Cavalcanti et al., 2014; Nyyssonen et al., 2014; Yergeau et al., 2014; Hu et al., 2015). To test if

the functional gene taxonomy was significantly different among samples grouped by sample type (i.e., water, sediment, Bio-Trap® sampler), location, or month, the SEED taxonomy abundance data were binned by phyla and an analysis of variance on a Bray-Curtis dissimilarity metric was performed using the adonis function within the vegan package in R (Oksanen et al., 2013).

Functional reads matching to the Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2014) Orthology with default settings from MG-RAST (1e-5 e-value, 60% minimum identity; Meyer et al., 2008) were used for a broad assessment of metabolic types in planktonic and attached samples. A sample type pair-wise comparison of functional read abundances matching KEGG level 1 categories was done using the Wald test and Benjamini-Hochberg p-value adjustment with the DESeq2 package in R (Love et al., 2014). The KEGG Orthology level 2 categories were used to estimate the relative abundances of functional reads matching genes for different metabolism types using the DESeq2 package in R (Love et al., 2014). The DESeq2 package uses a negative binomial frequency distribution model from an originally non-normalized abundance profile. This method estimates significant differential expression of functional genes for samples have different abundances of functional reads across samples and widely different type of functional reads (Anders and Huber, 2014; Love et al., 2014). The p-value was assessed as significant if $< 0.05$, and was corrected for multiple tests using the Benjamini-Hochberg adjustment in DESeq2. A principal component plot (PCA) of all samples was performed using a Euclidean distance metric on the regularized log transformed data from the negative binomial likelihood ratio test from the DESeq2 package in R.

The Subsystems functional hierarchical classification system (Aziz et al., 2008) with default settings from MG-RAST (1e-5 e-value, 60% minimum identity; Meyer et al., 2008) was used for the following functional read analysis. The Subsystems functional hierarchical classification system was chosen because it assigned functions to the largest number of reads for each sample, compared to other annotation databases. Hierarchical average neighbor clustering was used to assess discrete groups with varying degrees of (dis)similarity using the program Statistical Analysis of Metagenomic Profiles program (STAMP; Parks et al. 2014). Clustering was performed with unweighted pair-group method using arithmetic averages (UPGMA) of unassembled raw functional read abundances at a dendrogram threshold of 0.75 (Parks et al., 2014). Jaccard Indices were calculated across each sample to assess Bio-Trap® sample similarity to other sample types, using the application of the vegdist function on the unassembled raw functional read abundances of all Subsystems function level categories (7756 categories in total) in R (Oksanen et al., 2013).

Two different classes of enzymes involved in plant litter degradation were considered as defined by the Carbohydrate-Active enZYmes Database (CAZy; Lombard *et al.* 2014). The glycoside hydrolases (EC 3.2.1.-) are a large group of enzymes that hydrolyse the glycosidic bond between at least one carbohydrate compound. The auxiliary activities family classification is a group of enzymes that cleave complex carbohydrates, including ligninolytic enzymes and lytic polysaccharide mono-oxygenases. To evaluate if the unassembled raw functional read abundances of reads matching CAZy enzyme classes involved in plant litter degradation were significantly different between sample type, location, or month, a Randomized Block Design

(RBD) split-split plot repeated measures ANOVA was performed in SAS (version 9.4; SAS Institute Inc.). A randomized block design controls for any variability induced by different enzyme types while being able to measure differences more explicitly in average abundance values between location, sample type, and month (Saxton, 2002). The fixed conditions of the RBD split-split plot analysis were as follows: the whole-plot factor was the location within the cave (upstream or downstream), the split-plot factor was the month the measurement was taken (August through December), and the split-split plot factor was the type of sample (water, sediment, or Bio-Trap®). The type of sample was considered a repeated measure over space to account for any dependencies that may exist from measuring sediment, water, and Bio-Trap® sample types simultaneously at each individual cave location. The blocking factor was the enzyme type (glycoside hydrolases, or ligninolytic enzymes and lytic polysaccharide mono-oxygenases). The formula used for the RBD split-split plot analysis was:

$$y_{ijkl} = \mu + B_i + T_j + B * T_{ij} + F_k + T * F_{jk} + B * F * T_{ijk} + G_l + G * T_{jl} + G * F_{kl} + G * T * F_{jkl} + B * G * F * T_{ijkl}$$

where B = block, T = location, F = month, and G = sample type. The response variable was defined as the average abundance of all reads taken at each combination of factor levels (n = 46). The model information is as follows: the response distribution was Gaussian, the link function was identity, and the estimation technique was restricted maximum likelihood. An unstructured variance co-variance matrix was used to account for unequal variances between different levels of the fixed factors. Tukey–Kramer HSD p-value adjustments were

implemented when performing paired Least Squares (LS) Mean Differences comparisons. Codes for statistical analyses performed in R (version 3.2.1) are provided in Appendix IV Code III.

## RESULTS

### Habitat and CDOM variability

During the study period, the cave stream continuously flowed through the three caves, although flow varied from below detection limit to 1.36 $m^3/s$ due to surface precipitation events. Brannen-Donnelly and Engel (2015) describe four flood events that remobilized sediment particles < 2 mm in size, which corresponded to the largest particles (except cobbles) collected upstream but was larger than the average particles downstream. Because the study period spanned seasonal changes from July to December, stream water pH, temperature, and alkalinity differed at each sampling location, but other geochemical parameters, such as major cation and anion concentrations, did not (Brannen-Donnelly and Engel, 2015). The DOC concentration upstream and downstream did not significantly vary for any one month, although concentrations decreased from the summer to the winter, for instance ranging from 5.2 mg/L in August to 1.67 mg/L in December at the upstream location (Brannen-Donnelly and Engel, 2015). CDOM fluorescence significantly differed by month and by location. Upstream water in July and August was dominated by chromophores resulting from humified terrigenous material, with increasing contributions of humified CDOM downstream compared to upstream. In December, contributions from chromophores resulting from proteinaceous material increased, which suggested less input of CDOM from terrigenous material in the winter, as would be

expected in an area dominated by deciduous vegetation (Brannen-Donnelly and Engel, 2015).

**Metagenome overview and taxonomic composition**

The 31 Cascade Cave metagenomes ranged in size from 33 thousand to 3 million reads. Analyzed using MG-RAST, these samples resulted in 15,677,399 reads after trimming, and 40% of the reads had predicted known functions (Table A4-1). There were three sediment samples deemed to be of poor quality based on the dominance of poor quality reads, and were not used in this study (Table A4-1). Taxonomic analysis of the assembled reads produced classifications for 53.5% of the reads (Table A4-1). Of those, Bacteria dominated (95-98% of the classified reads), followed by Unassigned (1-3%), and Archaea (1%). Within the bacterial domain, 99% of the reads could be further assigned to a phylum, and most were affiliated with the Proteobacteria (69% of the reads), followed by Bacteroidetes (7%), Actinobacteria (5%), and Firmicutes (3%). This bacterial taxonomic distribution was similar to the 454 tag pyrosequencing analyses completed by Brannen-Donnelly and Engel (2015) that also found Proteobacteria and Actinobacteria were the dominant phyla (Table A4-2).

The raw taxonomic abundances of the metagenomic samples were significantly different between water, sediment, and Bio-Trap® sample types (adonis p-value = 0.01), but not by sample location or month. The taxonomy of the CAZy reads was also representative of the general bacterial diversity from all samples (Table A4-2). Based on the KEGG hierarchical classification system, the most abundant level 2 categories in all samples (Table A4-3) were Amino Acid Metabolism (21.5% of all functional reads), Biosynthesis of Other Secondary

Metabolites (12.9% of all functional reads), and Carbohydrate Metabolism (10.9% of all functional reads). The PCA results indicated a grouping of samples based on environment type (i.e., water, sediment, Bio-Trap®) for the first PC axis, which explained 44% of the variance in functional read abundance between KEGG categories (Fig. A4-1; figures located in Appendix III). Average genome sizes significantly differed (p-value <0.005) by sample type, but not by location or month, with sediment metagenomes having the largest average genome sizes and water samples having the smallest (Fig. A4-2).

**Comparative functional metagenomic analyses**

The log2 fold change analyses allow for pair-wise comparison of significant differences in abundance of functional reads matching different KEGG categories among sample types (Fig. A4-3). Reads matching the cell motility KEGG level 2 category were enriched in both water and Bio-Trap samples, as compared to sediment samples. Read matching the Cell Communication category was significantly enriched in both the water and sediment samples, as compared to the Bio-Trap samples. Both Bio-Trap and sediment samples were significantly enriched with reads matching the Xenobiotics Biodegradation and Metabolism category, as compared to the water samples. There was also a significant enrichment of reads matching the KEGG level 3 category for Methane Metabolism in the sediment samples, as compared to the water samples (Fig. A4-4). The Bio-Trap samples had a significant enrichment in the KEGG level 3 category for Nitrogen Metabolism, and the water samples had a significant enrichment the KEGG level 3 category for Oxidative Phosphorylation when compared with each other (Fig. A4-4). The Bio-

Trap and sediment samples did not have any significant enrichment in the abundances of functional read matching to any KEGG level 3 category pertaining to metabolism.

All three classes of CAZy enzymes involved in plant litter degradation were present in all sample types and locations during the study period (Table A4-2). The ligninolytic enzymes and lytic polysaccharide mono-oxygenases had the highest abundance of the three enzyme classes for all samples, and the glycoside hydrolases had the second highest abundances. Because abundances of reads encoding for polysaccharide lyases were significantly less than the abundances for the other two CAZy classes, the polysaccharide lyases were not included in further analyses. The sediment samples had the highest abundance for all enzyme classes compared to all other sample types. The average read values of the two CAZy enzyme classes, as assessed from the RBD split-split plot repeated measures ANOVA Type III test of fixed effects, were significantly different based on sample type and month, but not by location (Table A4-4). All interactions between sample type, location, and month were significant (Table A4-4), suggesting that the abundance of CAZy genes varied by type and also over time, but also by time within each type of sample. These abundance changes could have been impacted by several types of disturbances during the study period because cave sediment was mobilized by four flooding events, although Bio-Trap® samplers were bolted in place and beads could not move. Even though abundances of the two CAZy gene classes significantly varied by sample type and over time, it is important to note that these genes were present for the entire study. Moreover, the combination of both taxonomic and functional analyses indicated that the two CAZy enzyme classes within the cave systems were not restricted to a particular taxonomic

group, or a particular time period, and that different microbial groups would potentially be able to degrade the same types of compounds throughout the ecosystem despite the environmental disturbances.

Lastly, we expected that community functional capabilities associated with the Bio-Trap® samplers would change as the communities matured, rather than be distinct from both water and sediment samples throughout the study period. In general, the prevalent functional Subsystems categories from the Bio-Trap® samplers were the same during the first few months of the study and during the last few months. If the planktonic microorganisms in the cave system played a role as the first colonizers in the cave system (Brannen-Donnelly and Engel, 2015), then we hypothesized that the Bio-Trap® communities would be more similar to the planktonic communities in diversity and function at the earlier stages of succession, but more similar to sediment communities in diversity and function at later stages of succession. However, the Bio-Trap® communities were more similar to other Bio-Trap® communities, according to the Jaccard Index values and the PCA (Table A4-5 and Fig. A4-1). They also had the smallest mean Jaccard Index values compared with the other sample types (Table A4-5).

## DISCUSSION

The functional and taxonomic profiles from metagenomes obtained over time from Cascade Cave stream water, sediments, and biofilms formed on Bio-Trap® samplers provide new information about the metabolic potential of microbes flowing through and colonizing solid surfaces (albeit, also potentially movable) inside caves, as well as about microbial

community functional and taxonomic succession following disturbances. Because these dark ecosystems rely on allochthonous material for energy and nutrients, metagenomic data revealed changes in overall ecological function associated with carbon degradation pathways, in particular those associated with plant litter degradation. However, even though the carbon degradation pathways change over time and across sample types, these pathways are prevalent in the cave ecosystem. Furthermore, these pathways provide a level of ecosystem stability, which is surprising because microbial communities change over time, likely due to the environmental disturbances (e.g., flooding) during this study.

Differences between both structure and function of planktonic and attached microorganisms have been documented in some subsurface environments (Hazen et al., 1991; Alfreider et al., 1997; Lehman, 2007; Flynn et al., 2008; Zhou et al., 2012), as well as surface streams (Araya et al., 2003; Besemer et al., 2012). The log2 fold change analyses between the water and sediment samples provided information that the abundance of functional reads did not significantly differ between metabolic gene KEGG level 2 categories between these two habitats; however, there was a significant enrichment in functional genes matching methane metabolism in the cave sediments as compared to the planktonic community. The Bio-Trap® samples had functional gene abundances that were more similar to the sediment samples than the water column, even though taxonomically there were more OTUs shared between the water column and the Bio-Trap® samplers at most time points (Brannen-Donnelly and Engel, 2015).

Functional read analyses provide evidence that nitrogen and methane metabolic cycling were more likely to occur in the Bio-Trap® and sediment samples, respectively, compared to the water column. From the broad analyses, there were also some functional capabilities indicating that the mobility and communication of microorganisms may not be occurring in the same way in the different habitats. There were significant differences in functional gene abundances between 28% of Subsystems functional categories. A previous comparison of functional genes from a cave speleothem to other environment types (i.e., soil, ocean, or rhizosphere; Ortiz *et al.* 2014) found that 50% of the functional genes were significantly different, as assessed by the COG database. Different functional niches for two *Prochlorococcus* spp. were assigned after 25% of their functional genes were found to be distinct (Rocap et al., 2003). Although niche separation has not been strictly defined in terms of the percent similarity of functional gene similarity (Rocap et al., 2003; Lennon et al., 2012), the functional differences between the cave environment types do not provide enough evidence to define distinct functional niches within the cave system, when compared to functional differences from other environments.

Burke et al. (2011) suggest that the functional similarities may be more important than taxonomy in order to understand bacterial succession and diversity in the environment. After an environmental disturbance, the microorganisms that will play a role in the new environment will be the ones that arrive there first, and colonization of space is random from within a functionally equivalent group of microorganisms. The log2 fold change analyses indicate that the Bio-Trap® samples were more similar to the sediment samples than the water samples for

the duration of the study. The functional gene analyses contrast with the 16S rRNA taxonomy analyses from Brannen-Donnelly and Engel (2015), where both the upstream and downstream Bio-Traps® had more OTUs in common with planktonic microorganisms in the earlier months, and the downstream Bio-Traps® had an increase in OTUs shared between the attached microorganisms over time.

We also searched for plant litter degradation functions that were common in all samples and that may indicate resiliency from environmental disturbances. The sediment samples did not have higher abundances for all CAZy enzyme classes compared to all of the other sample types, even though the sediment samples had the largest average genome size. Larger microbial genome sizes also generally include more reads for secondary metabolism and energy conversion, explaining broader metabolic diversity in the sediment microorganisms (Konstantinidis and Tiedje, 2004). The continuous presence of all three investigated classes of CAZy enzymes involved in plant litter degradation suggested that the degradation of these compound classes could occur despite the environmental disturbances in type and amount of DOM as well as sediment remobilization. However, the RBD split-split plot repeated measures ANOVA Type III test for fixed effects results indicate that the abundances of CAZy enzyme classes were significantly different based on environment type, as well as by time and location within each environment type. The glycoside hydrolase family of enzymes includes 258,218 different enzyme modules, and the auxiliary enzyme family of enzymes include 10,526 different enzyme modules, all of which are not contained in the samples from this study (Lombard et al. 2014). The source of variation in abundance of reads for these enzyme classes between our

samples may be due to the very large number of enzymes within these classes. While carbohydrates and some parts of lignin compounds were present, each environment type may be producing a different type of enzyme to degrade these molecules. These results are the first indication of a degradation of the same compound class in an epigenic cave system that is regularly flooded, and has changes in the quantity and quality of DOM. The variation of the abundance of reads matching the CAZy enzyme classes may help the cave microorganisms to degrade the changing quality of carbon.

Studies that have assessed large-scale ecosystem functions have found that larger scale processes can be independent of changes in microbial community diversity (Marschner, 2003; Langenheder et al., 2005; Frossard et al., 2012; Purahong et al., 2014). We believe the functions that glycoside hydrolases and ligninolytic enzymes provide could potentially be considered large-scale ecosystem functions in this cave system due to their continued presence and high relative abundance. While some functions may be more sensitive to a change in microbial community diversity or environmental perturbation, large-scale functions carried out by multiple types of microorganisms are not (Langenheder et al., 2006). One of the reasons why large-scale ecosystem functions might not be related to changes in microbial community diversity could be that the microbial community diversity includes generalist species capable of surviving in a wide range of environmental conditions (Rosenfeld, 2002; Langenheder et al., 2006; Frossard et al., 2012). However, we know that the microbial community structure changed in all cave environment types over the study period (Brannen-Donnelly and Engel 2015), so the same species of generalists must not exist in this cave system. Another reason for

the persistence of the CAZy enzyme classes could be that functional redundancy existed within the diversity of the cave environment (Rosenfeld, 2002; Langenheder et al., 2006; Allison and Martiny, 2008). It has been previously shown that functional redundancy for cellulose degradation across a high species richness supported a greater number of individuals and subsequently greater rates of total cellulose decomposition (Wohl et al., 2004).

## CONCLUSIONS

Allochthonous OM is an important source of energy for many cave ecosystems, including those with and without stable microbial communities. The reads involved in the degradation of certain compounds from allochthonous OM in the Cascade Cave System are present regardless of environment type, even though the microbial community diversity in each environment type changes over time. The genes are also present despite changes in amount and quality of DOM that is transported though the cave system, and despite mobilization of the sediment habitat for the microorganisms. Although the taxonomic diversity of bacteria within the cave changed over time, these data provide evidence that the bacteria may have stable functional ability to degrade specific classes of DOM, despite all of the environmental changes during the study period. This is a novel discovery for microbial processes occurring in the terrestrial subsurface. The functional succession only captured functional similarity to attached microorganisms, which is in contrast to the taxonomic succession. Many of the functional genes that were shared between all of the environment types were not restricted to a particular taxonomic group, which means that different species of microorganisms are able to provide the

same functionality to its cave ecosystem niche. This cave system may fundamentally differ from a groundwater system that has a stable microbial community in many ways, but the possibility of stable large-scale ecosystem functions may not be one of them.

**ACKNOWLEDGEMENTS:** K.B.-D. was supported by a National Science Foundation graduate research fellowship and a Cave Conservancy Foundation graduate fellowship. The Jones Endowment for Aqueous Geochemistry at the University of Tennessee provided additional funding for the research. T. Hazen provided the Bio-Trap® samplers, and S.A. Engel, B. Donnelly, C. Dietz, and A. Campion assisted in field deployment of the Bio-Traps® and sample collection. The cave research was permitted by the staff at Carter Cave State Resort Park in Kentucky. H. Woo is also thanked for her assistance with DESeq2.

## REFERENCES


Alfreider, A., Krossbacher, M., and Psenner, R. (1997). Groundwater samples do not reflect bacterial densities and activity in subsurface systems. *Water Research* 31**,** 832-840.

Allison, S.D., and Martiny, J.B. (2008). Colloquium paper: resistance, resilience, and redundancy in microbial communities. *Proc Natl Acad Sci U S A* 105 Suppl 1**,** 11512-11519.

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biol*, 11(10), R106.

Araya, R., Tani, K., Takagi, T., Yamaguchi, N., and Nasu, M. (2003). Bacterial activity and community composition in stream water and biofilm from an urban river determined by fluorescent in situ hybridization and DGGE analysis. *FEMS Microbiology Ecology* 43**,** 111-119.

Aziz, R.K., Bartels, D., Best, A.A., Dejongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., Mcneil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9**,** 75.

Besemer, K., Peter, H., Logue, J.B., Langenheder, S., Lindstrom, E.S., Tranvik, L.J., and Battin, T.J. (2012). Unraveling assembly of stream biofilm communities. *ISME J* 6**,** 1459-1468.

Bhullar, K., Waglechner, N., Pawlowski, A., Koteva, K., Banks, E.D., Johnston, M.D., Barton, H.A., and Wright, G.D. (2012). Antibiotic resistance is prevalent in an isolated cave microbiome. *PLoS One* 7**,** e34953.

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

Byl, T.D., Metge, D.W., Agymang, D.T., Bradley, M., Hileman, G., and Harvey, R.W. (2014). Adaptations of indigenous bacteria to fuel contamination in karst aquifers in South-Central Kentucky. *Journal of Cave and Karst Studies* 76**,** 104-113.

Cavalcanti, G.S., Gregoracci, G.B., Dos Santos, E.O., Silveira, C.B., Meirelles, P.M., Longo, L., Gotoh, K., Nakamura, S., Iida, T., Sawabe, T., Rezende, C.E., Francini-Filho, R.B., Moura, R.L., Amado-Filho, G.M., and Thompson, F.L. (2014). Physiologic and metagenomic attributes of the rhodoliths forming the largest $CaCO_3$ bed in the South Atlantic Ocean. *ISME J* 8**,** 52-62.

Cooney, T.J., and Simon, K.S. (2009). Influence of dissolved organic matter and invertebrates on the function of microbial films in groundwater. *Microb Ecol* 58**,** 599-610.

Culver, D.C., and Pipan, T. (2014). *Shallow Subterranean Habitats: Ecology, Evolution, and Conservation.* Oxford University Press.

Delmont, T.O., Prestat, E., Keegan, K.P., Faubladier, M., Robe, P., Clark, I.M., Pelletier, E., Hirsch, P.R., Meyer, F., Gilbert, J.A., Le Paslier, D., Simonet, P., and Vogel, T.M. (2012). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J* 6**,** 1677-1687.

Dussart-Baptista, L., Massei, N., Dupont, J.P., and Jouenne, T. (2003). Transfer of bacteria-contaminated particles in a karst aquifer: evolution of contaminated materials from a sinkhole to a spring. *Journal of Hydrology* 284**,** 285-295.

Engel, A.S. (2010). Microbial diversity of cave ecosystems, in Barton, L., Mandl, M., and Loy, A. (eds.), *Geomicrobiology: Molecular & Environmental Perspectives*, Springer. p. 219-238. DOI:10.1007/978-90-481-9204-5_10.

A. S. Engel and S. A. Engel. (2009). "A field guide for the karst of CarterCaves State Resort Park and the surrounding area, Northeastern Kentucky," in *Field Guide to Cave and Karst Lands of the United States*, A. S. Engel and S. A. Engel, Eds., Karst Waters Institute Special Publication 15, pp. 154–171, Karst Waters Institute, Leesburg, Va, USA.

Farnleitner, A.H., Wilhartitz, I., Ryzinska, G., Kirschner, A.K., Stadler, H., Burtscher, M.M., Hornek, R., Szewzyk, U., Herndl, G., and 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**,** 1248-1259.

Flynn, T.M., Sanford, R.A., and Bethke, C.M. (2008). Attached and suspended microbial communities in a pristine confined aquifer. *Water Resources Research* 44**,** W07425-W07432.

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

Frossard, A., Gerull, L., Mutz, M., and Gessner, M.O. (2012). Disconnect of microbial structure and function: enzyme activities and bacterial communities in nascent stream corridors. *ISME J* 6**,** 680-691.

Goldscheider, N., Hunkeler, D., and Rossi, P. (2006). Review: Microbial biocenoses in pristine aquifers and an assessment of investigative methods. *Hydrogeology Journal* 14**,** 926-941.

Graening, G. O., and Brown, A. V. Ecosystem dynamics and pollution effects in an Ozark cave stream. *Journal of the American Water Resources Association* 39, 1497–1507. doi: 10.1111/j.1752-1688.2003.tb04434.x

Griebler, C., and Lueders, T. (2009). Microbial biodiversity in groundwater ecosystems. *Freshwater Biology* 54**,** 649-677.

Hahn, H.J., and Fuchs, A. (2009). Distribution patterns of groundwater communities across aquifer types in south-western Germany. *Freshwater Biology* 54**,** 848-860.

Hazen, T.C., Jimenes, L., Lopez De Victoria, G., and Fliermans, C.B. (1991). Comparison of Bacteria from Deep Subsurface Sediment and Adjacent Groundwater. *Microbial Ecology* 22**,** 293-304.

Hemme, C.L., Deng, Y., Gentry, T.J., Fields, M.W., Wu, L., Barua, S., Barry, K., Tringe, S.G., Watson, D.B., He, Z., Hazen, T.C., Tiedje, J.M., Rubin, E.M., and Zhou, J. (2010). Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *ISME J* 4**,** 660-672.

Henderson, I.R., Navarro-Garcia, F., Desvaux, M., Fernandez, R.C., and Ala'aldeen, D. (2004). Type V protein secretion pathway: the autotransporter story. *Microbiology Molecular Biology Reviews* 68**,** 692-744.

Hu, Q., Guo, X., Liang, Y., Hao, X., Ma, L., Yin, H., and Liu, X. (2015). Comparative metagenomics reveals microbial community differentiation in a biological heap leaching system. *Research Microbiology* 166**,** 525-534.

Hug, L.A., Thomas, B.C., Brown, C.T., Frischkorn, K.R., Williams, K.H., Tringe, S.G., and Banfield, J.F. (2015). Aquifer environment selects for microbial species cohorts in sediment and groundwater. *ISME J* 9**,** 1846-1856.

Jones, D.S., Albrecht, H.L., Dawson, K.S., Schaperdoth, I., Freeman, K.H., Pi, Y., Pearson, A., and Macalady, J.L. (2012). Community genomic analysis of an extremely acidophilic sulfur-oxidizing biofilm. *ISME J* 6**,** 158-170.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, gkr988.

Keegan, K.P., Trimble, W.L., Wilkening, J., Wilke, A., Harrison, T., D'souza, M., and Meyer, F. (2012). A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. *PLoS Computional Biology* 8**,** e1002541.

Konstantinidis, K.T., and Tiedje, J.M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci U S A* 101**,** 3160-3165.

Langenheder, S., Lindstrom, E.S., and Tranvik, L.J. (2006). Structure and function of bacterial communities emerging from different sources under identical conditions. *Applied and Environmental Microbiology* 72**,** 212-220.

Langenheder, S., Lindström, E.S., and Tranvik, L.J. (2005). Weak coupling between community composition and functioning of aquatic bacteria. *Limnology and Oceanography* 50**,** 957-967.

Lehman, R.M. (2007). Understanding of Aquifer Microbiology is Tightly Linked to Sampling Approaches. *Geomicrobiology Journal* 24**,** 331-341.

Lennon, J.T., Aanderud, Z.T., Lehmkuhl, B., and Schoolmaster Jr, D.R. (2012). Mapping the niche space of soil microorganisms using taxonomy and traits. *Ecology* 93**,** 1867-1879.

Lin, X., Mckinley, J., Resch, C.T., Kaluzny, R., Lauber, C.L., Fredrickson, J., Knight, R., and Konopka, A. (2012). Spatial and temporal dynamics of the microbial community in the Hanford unconfined aquifer. *ISME Journal* 6**,** 1665-1676.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42**,** D490-495.

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

Marschner, P. (2003). Structure and function of the soil microbial community in a long-term fertilizer experiment. *Soil Biology and Biochemistry* 35**,** 453-461.

Meyer, F., Paarmann, D., D'souza, M., Olson, R., Glass, E.M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R.A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9**,** 386.

Montana, J.S., Jimenez, D.J., Hernandez, M., Angel, T., and Baena, S. (2012). Taxonomic and functional assignment of cloned sequences from high Andean forest soil metagenome. *Antonie Van Leeuwenhoek* 101**,** 205-215.

Moore, B.S., Hertweck, C., Hopke, J.N., Izumikawa, M., Kalaitzis, J.A., Nilsen, G., O'hare, T., Piel, J., Shipley, P.R., and Xiang, L. (2002). Plant-like biosynthetic pathways in bacteria: from benzoic acid to chalcone. *Journal of natural products* 65**,** 1956-1962.

Morasch, B. (2013). Occurrence and dynamics of micropollutants in a karst aquifer. *Environmental Pollution* 173**,** 133-137.

Nyyssonen, M., Hultman, J., Ahonen, L., Kukkonen, I., Paulin, L., Laine, P., Itavaara, M., and Auvinen, P. (2014). Taxonomically and functionally diverse microbial communities in deep crystalline rocks of the Fennoscandian shield. *ISME Journal* 8**,** 126-138.

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., O'hara, R., Simpson, G., Solymos, P., Stevens, M., and Wagner, H. (2013). "vegan: Community Ecology Package".).

Ortiz, M., Legatzki, A., Neilson, J.W., Fryslie, B., Nelson, W.M., Wing, R.A., Soderlund, C.A., Pryor, B.M., and Maier, R.M. (2014). Making a living while starving in the dark: metagenomic insights into the energy dynamics of a carbonate cave. *ISME Journal* 8**,** 478-491.

Palmer, A.N. (2007). *Cave Geology.* Cave books, Dayton.

Parks, D.H., Tyson, G.W., Hugenholtz, P., and Beiko, R.G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30**,** 3123-3124.

Pronk, M., Goldscheider, N., and Zopfi, J. (2008). Microbial communities in karst groundwater and their potential use for biomonitoring. *Hydrogeology Journal* 17**,** 37-48.

Purahong, W., Schloter, M., Pecyna, M.J., Kapturska, D., Daumlich, V., Mital, S., Buscot, F., Hofrichter, M., Gutknecht, J.L., and Kruger, D. (2014). Uncoupling of microbial community structure and function in decomposing litter across beech forest ecosystems in Central Europe. *Sci Rep* 4**,** 7014.

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 38**,** e191.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., Hess, W.R., Johnson, Z.I., Land, M., Lindell, D., Post, A.F., Regala, W., Shah, M., Shaw, S.L., Steglich, C., Sullivan, M.B., Ting, C.S., Tolonen, A., Webb, E.A., Zinser, E.R., and Chisholm, S.W. (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* 424**,** 1042-1047.

Rosenfeld, J.S. (2002). Functional redundancy in ecology and conservation. *Oikos* 98**,** 156-162.

Saxton, M.A. (2002). *Design and Analysis for Biological Research With SAS Software*. Department of Animal Sciences, University of Tennessee, Knoxville, Tennessee.

Shabarova, T., and Pernthaler, J. (2010). Karst pools in subsurface environments: collectors of microbial diversity or temporary residence between habitat types. *Environ Microbiol* 12**,** 1061-1074.

Shabarova, T., Villiger, J., Morenkov, O., Niggemann, J., Dittmar, T., and Pernthaler, J. (2014). Bacterial community structure and dissolved organic matter in repeatedly flooded subsurface karst water pools. *FEMS Microbiol Ecol* 89**,** 111-126.

Simon, K., Benfield, E., and Macko, S. (2003). Food web structure and the role of epilithic biofilms in cave streams. *Ecology* 84**,** 2395-2406.

Simon, K.S., Pipan, T., and Culver, D.C. (2007). A conceptual model of the flow and distribution of organic carbon in caves. *Journal of Cave and Karst Studies* 69**,** 279-284.

Simon, K.S., Pipan, T., Ohno, T., and Culver, D.C. (2010). Spatial and temporal patterns in abundance and character of dissolved organic matter in two karst aquifers. *Fundamental and Applied Limnology / Archiv für Hydrobiologie* 177**,** 81-92.

Sinsabaugh, R.L., Carreiro, M.M., and Alvarez, S. (2002). Enzyme and microbial dynamics of litter decomposition. *Enzymes in the Environment, Activity, Ecology, and Applications.* Marcel Dekker, New York, Basel**,** 249-265.

Smith, R.J., Paterson, J.S., Sibley, C.A., Hutson, J.L., and Mitchell, J.G. (2015). Putative effect of aquifer recharge on the abundance and taxonomic composition of endemic microbial communities. *PLoS One* 10**,** e0129004.

Spizzico, M., Lopez, N., and Sciannamblo, D. (2005). Analysis of the potential contamination risk of groundwater resources circulating in areas with anthropogenic activities. *Natural Hazards and Earth System Science* 5**,** 109-116.

Tetu, S.G., Breakwell, K., Elbourne, L.D., Holmes, A.J., Gillings, M.R., and Paulsen, I.T. (2013). Life in the dark: metagenomic evidence that a microbial slime community is driven by inorganic nitrogen metabolism. *ISME J* 7**,** 1227-1236.

Venarsky, M.P., Benstead, J.P., and Huryn, A.D. (2012). Effects of organic matter and season on leaf litter colonisation and breakdown in cave streams. *Freshwater Biology* 57**,** 773-786.

Venarsky, M.P., Huntsman, B.M., Huryn, A.D., Benstead, J.P., and Kuhajda, B.R. (2014). Quantitative food web analysis supports the energy-limitation hypothesis in cave stream ecosystems. *Oecologia* 176**,** 859-869.

Wilhartitz, I.C., Kirschner, A.K., Stadler, H., Herndl, G.J., Dietzel, M., Latal, C., Mach, R.L., and Farnleitner, A.H. (2009). Heterotrophic prokaryotic production in ultraoligotrophic alpine karst aquifers and ecological implications. *FEMS Microbiol Ecol* 68**,** 287-299.

Yagi, J.M., Neuhauser, E.F., Ripp, J.A., Mauro, D.M., and Madsen, E.L. (2010). Subsurface ecosystem resilience: long-term attenuation of subsurface contaminants supports a dynamic microbial community. *ISME J* 4**,** 131-143.

Yergeau, E., Sanschagrin, S., Maynard, C., St-Arnaud, M., and Greer, C.W. (2014). Microbial expression profiles in the rhizosphere of willows depend on soil contamination. *ISME J* 8**,** 344-358.

Zhou, Y., Kellermann, C., and Griebler, C. (2012). Spatio-temporal patterns of microbial communities in a hydrologically dynamic pristine aquifer. *FEMS Microbiol Ecol* 81**,** 230-242.

## APPENDIX III

**FIGURES**

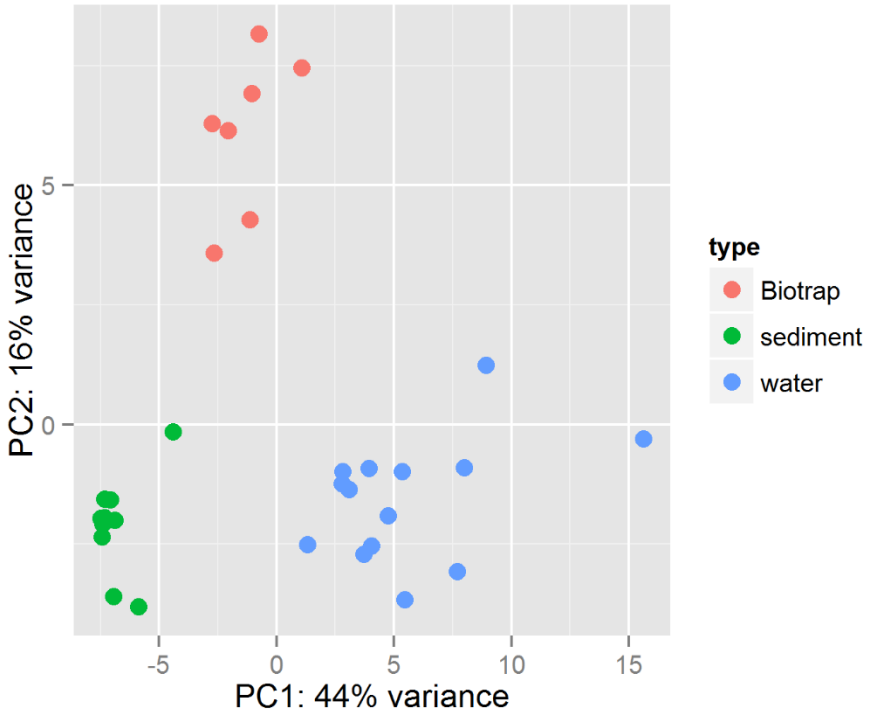



Figure A 4-1: Principal coordinate analysis showing clusters of sample type based on the Euclidean distance metric constructed from regularized log transformed gene abundance data from the negative binomial likelihood ratio test.

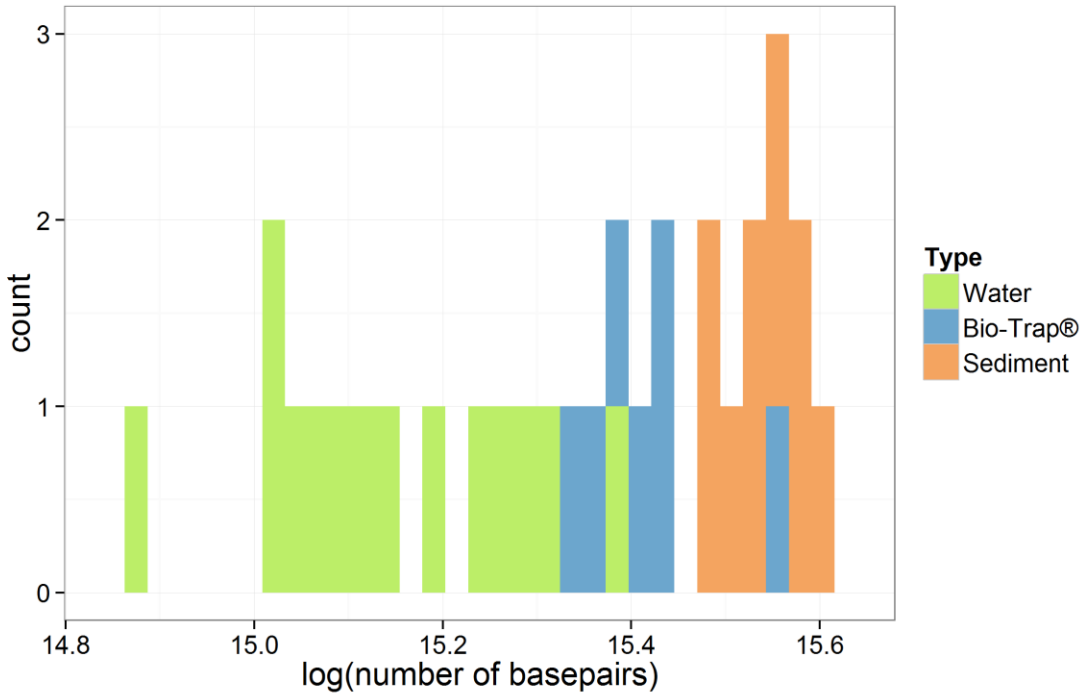



Figure A 4-2: Histogram of the average genome size of each sample by log number of basepairs. Color indicates the sample environment type.

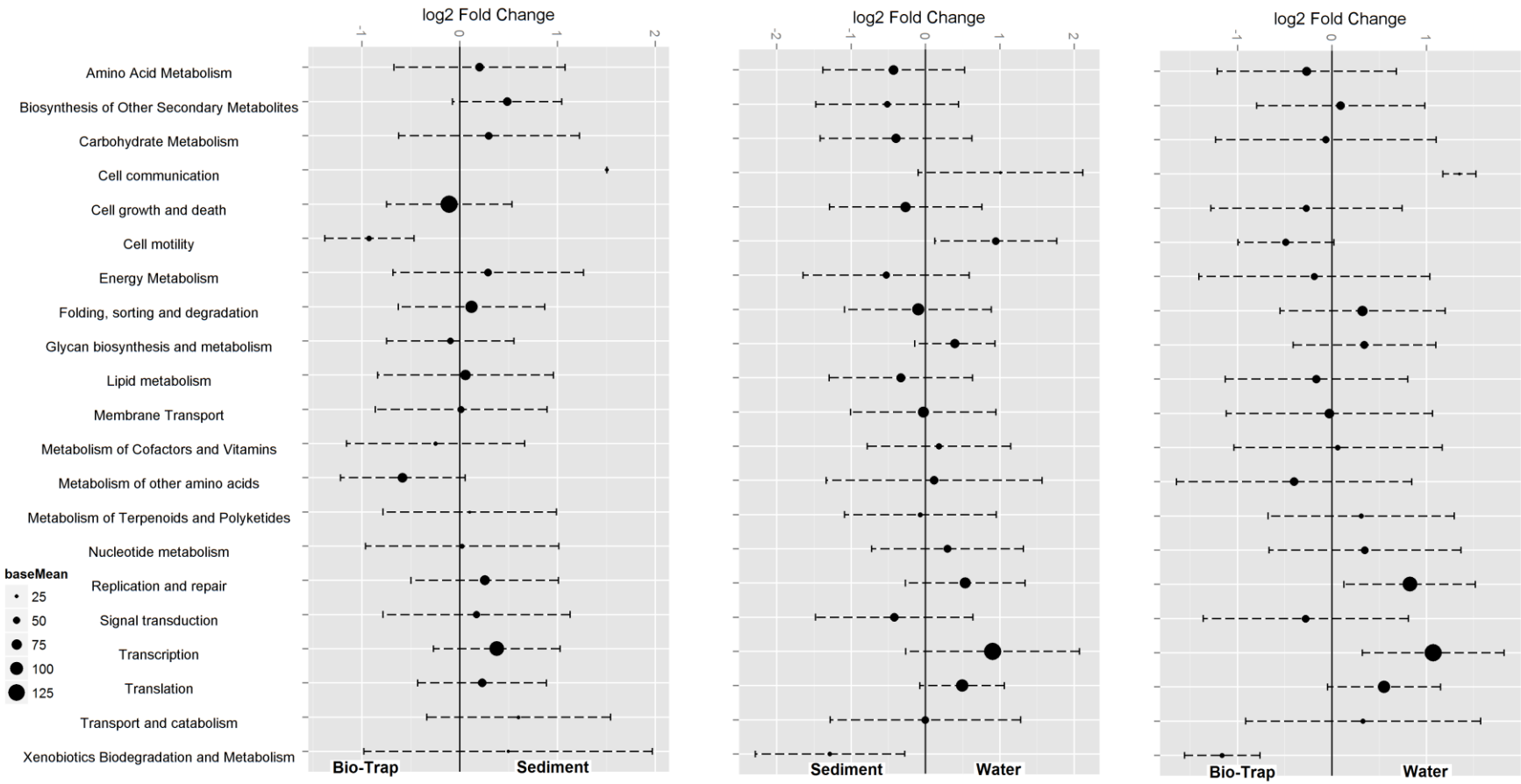



Figure A 4-3: Log2 fold changes in read abundances compare each sample type for all KEGG level 2 hierarchical categories.

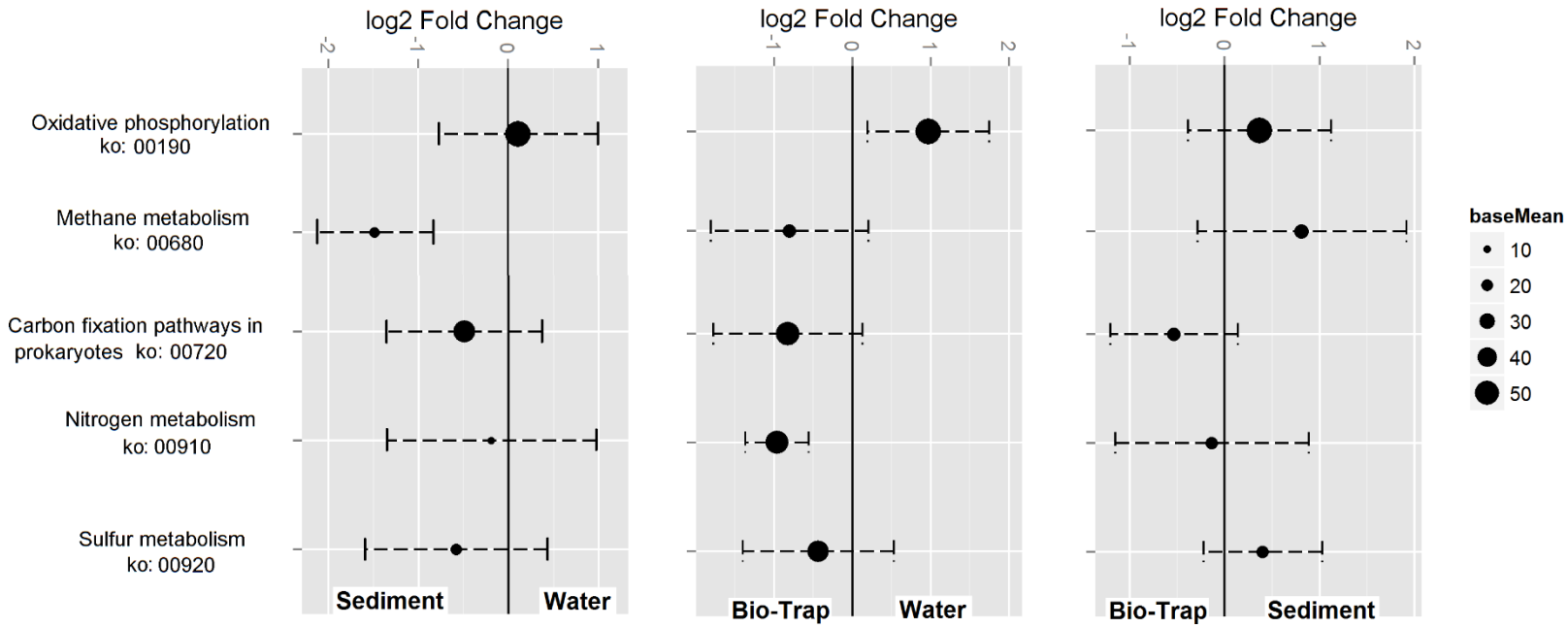



Figure A 4-4: Log2 fold changes in read abundances for all KEGG level 3 hierarchical categories to compare bacterial and archaeal metabolisms. KEGG ontology numbers are listed under the category name

**TABLES**

Table A 4-1: Summaries of the metagenome data, including MG-RAST ID numbers. Numbers of base pairs (bp), sequences, mean sequence length, mean GC content, and gene copies are from the raw data. Reads after quality filter and number predicted protein features are from the quality analyses through MG-RAST. Number annotated reads, SEED Subsystems predicted functions, and abundance from SEED taxonomy are from the annotated read through MG-RAST. * denotes the samples not used in this study due to their poor quality. Month is denoted by 1-12 calendar months.

| MG-RAST ID | name | Location | Month | Type | bp | # of seq. | Mean seq. length | Mean GC content | 16S rRNA copies MG-RAST | Number predicted protein features | Number of identified protein features | Number annotated reads | Subsyst. predicted functions | Abundance from SEED taxonomy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4577492.3 | CCRB-13d1 | Lake Room | 12 | Bio-Trap® | 102744317 | 452810 | 226 | 60.9 | 1398 | 430438 | 220358 | 182148 | 251656 | 412106 |
| 4577493.3 | CCRB-13d2 | Jones | 12 | Bio-Trap® | 85815054 | 378491 | 226 | 61 | 763 | 365251 | 173608 | 140852 | 185290 | 309392 |
| 4577494.3 | CCRB-13N1 | Lake Room | 11 | Bio-Trap® | 101507965 | 449606 | 225 | 61 | 1319 | 426409 | 209929 | 171521 | 235511 | 392191 |
| 4577495.3 | CCRB-13N2 | Jones | 10 | Bio-Trap® | 172515507 | 748024 | 230 | 61 | 2897 | 701417 | 341039 | 279178 | 402289 | 651596 |
| 4577496.3 | CCRB-13O1 | Lake Room | 10 | Bio-Trap® | 202553737 | 901205 | 224 | 60 | 6141 | 806218 | 406344 | 334378 | 520334 | 834944 |
| 4577497.3 | CCRB-13S2 | Lake Room | 9 | Bio-Trap® | 63254656 | 270454 | 233 | 60 | 902 | 257334 | 38409 | 115614 | 155657 | 257981 |
| 4577498.3 | CCRB-13S3 | Jones | 9 | Bio-Trap® | 50856755 | 219348 | 231 | 61 | 582 | 210595 | 104040 | 85566 | 110370 | 185330 |
| 4577499.3 | CCRF-13A1 | Lake Room | 8 | Water | 44329393 | 195397 | 226 | 57 | 523 | 190400 | 83138 | 67302 | 85628 | 144713 |
| 4577500.3 | CCRF-13A2 | Jones | 8 | Water | 53438692 | 231083 | 231 | 56 | 985 | 222411 | 95379 | 77381 | 100052 | 163559 |
| 4579010.3 | CCRF-13A3 | Fort Falls | 8 | Water | 44204417 | 191779 | 230 | 53 | 3660 | 181036 | 81774 | 66862 | 86337 | 142670 |
| 4579011.3 | CCRF-13d1 | Lake Room | 12 | Water | 7582203 | 33135 | 228 | 51 | 426 | 30594 | 11936 | 9575 | 12491 | 20223 |
| 4579012.3 | CCRF-13d2 | Jones | 12 | Water | 48307027 | 210831 | 229 | 51 | 1019 | 200117 | 84643 | 69267 | 90893 | 144958 |

Table A4-1 Continued

| MG-RAST ID | name | Location | Month | Type | bps | # of seq. | Mean sequence length | Mean GC content | 16S rRNA copies MG-RAST | Number predicted protein features | Number of identified protein features | Number annotated reads | Subsys. predicted functions | Abundance from SEED taxonomy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4579013.3 | CCRF-13d3 | Fort Falls | 12 | Water | 84034190 | 371055 | 226 | 52 | 1414 | 352011 | 139898 | 109973 | 136861 | 240552 |
| 4579014.3 | CCRF-13N1 | Lake Room | 11 | Water | 88839532 | 386988 | 229 | 51 | 788 | 355987 | 153162 | 123192 | 176291 | 279129 |
| 4579015.3 | CCRF-13N2 | Jones | 11 | Water | 27195066 | 117733 | 230 | 52 | 1085 | 108989 | 40369 | 31515 | 39715 | 67570 |
| 4579016.3 | CCRF-13N3 | Fort Falls | 11 | Water | 57049312 | 247812 | 230 | 47 | 1819 | 213482 | 90193 | 72009 | 95481 | 157525 |
| 4579017.3 | CCRF-13O1 | Lake Room | 10 | Water | 72701088 | 315101 | 230 | 56 | 614 | 301900 | 135460 | 108946 | 143408 | 239960 |
| 4577665.3 | CCRF-13O2 | Jones | 10 | Water | 327064058 | 1465184 | 223 | 54 | 4626 | 1377893 | 574775 | 467523 | 739159 | 1080566 |
| 4577501.3 | CCRF-13O3 | Fort Falls | 10 | Water | 77344721 | 335447 | 230 | 54 | 2009 | 315896 | 153302 | 128465 | 178603 | 275167 |
| 4577502.3 | CCRF-13S3 | Jones | 9 | Water | 48127053 | 206918 | 232 | 51 | 984 | 197402 | 88821 | 73441 | 103514 | 158314 |
| 4577666.3 | CCRF-13SF3 | Fort Falls | 9 | Water | 116304134 | 508964 | 228 | 53 | 3817 | 417300 | 220370 | 189919 | 367672 | 529846 |
| 4585047.3 | CCRS-13A2 | Jones | 8 | Sediment | 119853679 | 728065 | 164 | 63 | 350 | 591621 | 220831 | 176900 | 226645 | 202143 |
| 4585048.3 | CCRS-13A3 | Fort Falls | 8 | Sediment | 103386734 | 657317 | 157 | 61 | 513 | 516172 | 182190 | 145429 | 187963 | 163109 |
| 4577667.3 | CCRS-13d1 | Lake Room | 12 | Sediment | 125457517 | 554815 | 226 | 61 | 309 | 537363 | 228593 | 180595 | 237974 | 202734 |
| 4585049.3 | CCRS-13d2 | Jones | 12 | Sediment | 60879213 | 388662 | 156 | 63 | 2777 | 291177 | 118953 | 97042 | 118462 | 107863 |
| 4577503.3 | CCRS-13N2 | Jones | 11 | Sediment | 86451307 | 384860 | 224 | 62 | 560 | 372406 | 155854 | 123099 | 156543 | 133752 |
| 4585051.3 | CCRS-13N3 | Fort Falls | 11 | Sediment | 111150170 | 686408 | 161 | 62 | 318 | 541749 | 201883 | 163972 | 212586 | 184875 |
| 4585052.3 | CCRS-13O1 | Lake Room | 10 | Sediment | 102971176 | 712132 | 145 | 62 | 292 | 503487 | 181146 | 145854 | 189579 | 163910 |
| 4585053.3 | CCRS-13O2 | Jones | 10 | Sediment | 66786944 | 428265 | 156 | 62 | 496 | 324665 | 120435 | 96786 | 123744 | 105425 |
| 4585054.3 | CCRS-13O3 | Fort Falls | 10 | Sediment | 142154862 | 902606 | 157 | 62 | 4179 | 701119 | 259058 | 209618 | 274328 | 241312 |

Table A4-1 Continued

| MG-RAST ID | name | Location | Month | Type | bps | # of seq. | Mean sequence length | Mean GC content | 16S rRNA copies MG-RAST | Number predicted protein features | Number of identified protein features | Number annotated reads | Subsyst. predicted functions | Abundance from SEED taxonomy |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4585056.3 | CCRS-13S3 | Jones | 9 | Sediment | 112193503 | 668444 | 167 | 63 | 359 | 547966 | 209320 | 167862 | 219067 | 189980 |
| 4585046.3* | CCRS-13A1 | Lake Room | 8 | Sediment | 358,284,791 | 2168303 | 165 | 51 | 2,830 | 179064 | 27486 | 19889 | 58347 | * |
| 4585055.3* | CCRS-13S2 | Jones | 9 | Sediment | 438,778,323 | 2086251 | 154 | 47 | 4,179 | 135881 | 21296 | 15316 | 39331 | * |
| 4585050.3* | CCRS-13N1 | Lake Room | 11 | Sediment | 541,634,070 | 3326064 | 162 | 50 | 2,777 | 133307 | 17803 | 12804 | 46128 | * |

Table A 4-2: Number of functional sequences that matched phyla from the SEED Taxonomic database. Samples are split into groups based on environment type, with number of sequences (n) and percent abundance (%) of the group. Totals for Proteobacteria and proteobacterial classes are included. The number and percent abundance of functional reads from the CAZy gene classes matching different phyla from the SEED Taxonomic database are also listed.

| Phylum or class | Water | | | | Sediment | | | | Bio-Trap | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | n | % | CAZy n | CAZy % | n | % | CAZy n | CAZy % | n | % | CAZy n | CAZy % |
| Acidobacteria | 52340 | 1.47 | 1318 | 2.14 | 76635 | 4.81 | 3116 | 4.90 | 59861 | 2.00 | 1491 | 2.59 |
| Actinobacteria | 171040 | 4.82 | 3277 | 5.32 | 138124 | 8.67 | 5369 | 8.44 | 121929 | 4.08 | 2353 | 4.08 |
| Aquificae | 7890 | 0.22 | 119 | 0.19 | 2738 | 0.17 | 145 | 0.23 | 3712 | 0.12 | 62 | 0.11 |
| Bacteroidetes | 408458 | 11.50 | 7125 | 11.56 | 66897 | 4.20 | 2580 | 4.05 | 134436 | 4.50 | 2681 | 4.65 |
| Chlamydiae | 11492 | 0.32 | 160 | 0.26 | 1255 | 0.08 | 29 | 0.05 | 1987 | 0.07 | 26 | 0.05 |
| Chlorobi | 21821 | 0.61 | 392 | 0.64 | 12695 | 0.80 | 314 | 0.49 | 13792 | 0.46 | 226 | 0.39 |
| Chloroflexi | 35420 | 1.00 | 772 | 1.25 | 39695 | 2.49 | 1855 | 2.91 | 29737 | 1.00 | 666 | 1.15 |
| Cyanobacteria | 55304 | 1.56 | 730 | 1.18 | 34082 | 2.14 | 1313 | 2.06 | 46090 | 1.54 | 548 | 0.95 |
| Deferribacteres | 3589 | 0.10 | 22 | 0.04 | 971 | 0.06 | 25 | 0.04 | 1572 | 0.05 | 15 | 0.03 |
| Deinococcus-Thermus | 13790 | 0.39 | 280 | 0.45 | 11890 | 0.75 | 706 | 1.11 | 11836 | 0.40 | 415 | 0.72 |
| Dictyoglomi | 1884 | 0.05 | 74 | 0.12 | 1481 | 0.09 | 94 | 0.15 | 664 | 0.02 | 25 | 0.04 |
| Elusimicrobia | 2774 | 0.08 | 33 | 0.05 | 519 | 0.03 | 26 | 0.04 | 801 | 0.03 | 14 | 0.02 |
| Firmicutes | 142907 | 4.02 | 2183 | 3.54 | 52779 | 3.31 | 2330 | 3.66 | 59602 | 2.00 | 1046 | 1.81 |
| Fusobacteria | 4242 | 0.12 | 43 | 0.07 | 1002 | 0.06 | 34 | 0.05 | 0 | 0.00 | 19 | 0.03 |
| Nitrospirae | 15803 | 0.44 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 43429 | 1.45 | 0 | 0.00 |
| Planctomycetes | 88246 | 2.48 | 723 | 1.17 | 53148 | 3.33 | 1295 | 2.03 | 98219 | 3.29 | 902 | 1.56 |
| Proteobacteria (total) | 2357863 | 66.40 | 42728 | 69.33 | 1067341 | 66.96 | 43113 | 67.74 | 2272227 | 76.07 | 46452 | 80.54 |
| Alphaproteobacteria class | 518974 | 22.01 | 9299 | 21.76 | 387859 | 36.34 | 15508 | 35.97 | 121929 | 5.37 | 17368 | 37.39 |
| Betaproteobacteria class | 1226630 | 52.02 | 21825 | 51.08 | 357848 | 33.53 | 14647 | 33.97 | 792456 | 34.88 | 15642 | 33.67 |
| Deltaproteobacteria class | 195790 | 8.30 | 3801 | 8.90 | 171244 | 16.04 | 6489 | 15.05 | 133967 | 5.90 | 3041 | 6.55 |
| Epsilonproteobacteria class | 21515 | 0.91 | 250 | 0.59 | 3290 | 0.31 | 114 | 0.26 | 5018 | 0.22 | 75 | 0.16 |

Table A4-2 Continued

| Phylum | Water | | | | Sediment | | | | Bio-Trap | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | n | % | CAZy n | CAZy % | n | % | CAZy n | CAZy % | n | % | CAZy n | CAZy % |
| Gammaproteobacteria class | 389862 | 16.53 | 7465 | 17.47 | 144226 | 13.51 | 6244 | 14.48 | 477715 | 21.02 | 10248 | 22.06 |
| Zetaproteobacteria class | 1117 | 0.05 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| unclassified (derived from Proteobacteria) | 3975 | 0.17 | 88 | 0.21 | 2874 | 0.27 | 111 | 0.26 | 3882 | 0.17 | 78 | 0.17 |
| Spirochaetes | 12264 | 0.35 | 42 | 0.07 | 2614 | 0.16 | 33 | 0.05 | 4955 | 0.17 | 19 | 0.03 |
| Tenericutes | 1864 | 0.05 | 16 | 0.03 | 381 | 0.02 | 8 | 0.01 | 369 | 0.01 | 15 | 0.03 |
| Thermotogae | 6498 | 0.18 | 179 | 0.29 | 3086 | 0.19 | 189 | 0.30 | 2492 | 0.08 | 84 | 0.15 |
| unclassified (derived from Bacteria) | 4630 | 0.13 | 69 | 0.11 | 4830 | 0.30 | 238 | 0.37 | 2389 | 0.08 | 74 | 0.13 |
| Verrucomicrobia | 112663 | 3.17 | 1312 | 2.13 | 20377 | 1.28 | 762 | 1.20 | 55717 | 1.87 | 514 | 0.89 |
| Other | 14980 | 0.42 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 | 19440 | 0.65 | 0 | 0.00 |

Table A 4-3: Number of functional reads matching KEGG level 2 functional categories listed by environment type and for all samples. Percent abundance is listed based on the group of samples.

| KEGG Category Level 2 | Bio-Trap | | Water | | Sediment | | All Samples | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| Amino Acid Metabolism | 173038 | 22.02 | 181854 | 22.01 | 205566 | 20.65 | 560458 | 21.49 |
| Biosynthesis of Other Secondary Metabolites | 8505 | 1.08 | 10002 | 1.21 | 10996 | 1.1 | 29503 | 1.13 |
| Carbohydrate Metabolism | 100688 | 12.81 | 111891 | 13.54 | 124945 | 12.55 | 337524 | 12.94 |
| Cell communication | 177 | 0.02 | 459 | 0.06 | 541 | 0.05 | 1177 | 0.05 |
| Cell growth and death | 14459 | 1.84 | 17034 | 2.06 | 19000 | 1.91 | 50493 | 1.94 |
| Cell motility | 19719 | 2.51 | 11969 | 1.45 | 18688 | 1.88 | 50376 | 1.93 |
| Energy Metabolism | 46055 | 5.86 | 52611 | 6.37 | 56541 | 5.68 | 155207 | 5.95 |
| Folding, sorting and degradation | 22275 | 2.83 | 23938 | 2.9 | 31299 | 3.14 | 77512 | 2.97 |
| Glycan biosynthesis and metabolism | 18351 | 2.33 | 16696 | 2.02 | 26416 | 2.65 | 61463 | 2.36 |
| Lipid metabolism | 22507 | 2.86 | 26225 | 3.17 | 26593 | 2.67 | 75325 | 2.89 |
| Membrane Transport | 89422 | 11.38 | 93680 | 11.34 | 100992 | 10.15 | 284094 | 10.9 |
| Metabolism of Cofactors and Vitamins | 44016 | 5.6 | 41679 | 5.04 | 58222 | 5.85 | 143917 | 5.52 |
| Metabolism of other amino acids | 10270 | 1.31 | 9341 | 1.13 | 11108 | 1.12 | 30719 | 1.18 |
| Metabolism of Terpenoids and Polyketides | 11439 | 1.46 | 11660 | 1.41 | 15188 | 1.53 | 38287 | 1.47 |
| Nucleotide metabolism | 31745 | 4.04 | 31919 | 3.86 | 45909 | 4.61 | 109573 | 4.2 |
| Replication and repair | 35240 | 4.48 | 37248 | 4.51 | 56112 | 5.64 | 128600 | 4.93 |
| Signal transduction | 48542 | 6.18 | 54300 | 6.57 | 56344 | 5.66 | 159186 | 6.1 |
| Signaling molecules and interaction | 12 | 0.00 | 33 | 0.00 | 20 | 0.00 | 65 | 0.00 |
| Transcription | 14212 | 1.81 | 15542 | 1.88 | 21476 | 2.16 | 51230 | 1.96 |
| Translation | 52471 | 6.68 | 53517 | 6.48 | 84392 | 8.48 | 190380 | 7.3 |
| Transport and catabolism | 9478 | 1.21 | 10362 | 1.25 | 12227 | 1.23 | 32067 | 1.23 |
| Xenobiotics Biodegradation and Metabolism | 13312 | 1.69 | 14200 | 1.72 | 12789 | 1.28 | 40301 | 1.55 |

Table A 4-4: RBD split-split plot repeated measures ANOVA Type III test of fixed effects. "*" implies an interaction between effects.

| Effect | Num DF | Den DF | F value | Pr>F |
|---|---|---|---|---|
| Location | 1 | 14.800 | 2.68 | 0.1226 |
| Month | 4 | 10.920 | 12.71 | 0.0004 |
| Location*Month | 4 | 9.752 | 5.16 | 0.0168 |
| Type | 2 | 8.955 | 8.59 | 0.0083 |
| Location*Type | 2 | 6.856 | 15.52 | 0.0028 |
| Type*Month | 7 | 8.525 | 10.14 | 0.0015 |
| Location*Type*Month | 2 | 6.865 | 5.72 | 0.0345 |

Table A 4-5: Jaccard Index values for each sample compared to its nearest neighbor, second nearest neighbor, and third nearest neighbor

| Sample | | | Jaccard Index | Nearest Neighbor | | | Jaccard Index | Second Nearest Neighbor | | | Jaccard Index | Third Nearest Neighbor | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Location | Month | Type | | Location | Month | Type | | Location | Month | Type | | Location | Month | Type |
| Jones | 9 | Bio-Trap® | 0.18 | Lake Room | 9 | Bio-Trap® | 0.19 | Lake Room | 12 | Bio-Trap® | 0.19 | Lake Room | 11 | Bio-Trap® |
| Jones | 10 | Bio-Trap® | 0.15 | Lake Room | 10 | Bio-Trap® | 0.17 | Lake Room | 12 | Bio-Trap® | 0.17 | Lake Room | 11 | Bio-Trap® |
| Jones | 12 | Bio-Trap® | 0.17 | Lake Room | 12 | Bio-Trap® | 0.18 | Lake Room | 11 | Bio-Trap® | 0.18 | Jones | 10 | Bio-Trap® |
| Lake Room | 9 | Bio-Trap® | 0.17 | Lake Room | 12 | Bio-Trap® | 0.18 | Lake Room | 11 | Bio-Trap® | 0.18 | Jones | 9 | Bio-Trap® |
| Lake Room | 10 | Bio-Trap® | 0.15 | Jones | 10 | Bio-Trap® | 0.16 | Lake Room | 12 | Bio-Trap® | 0.17 | Lake Room | 11 | Bio-Trap® |
| Lake Room | 11 | Bio-Trap® | 0.16 | Lake Room | 12 | Bio-Trap® | 0.17 | Lake Room | 10 | Bio-Trap® | 0.17 | Jones | 10 | Bio-Trap® |
| Lake Room | 12 | Bio-Trap® | 0.16 | Lake Room | 11 | Bio-Trap® | 0.16 | Lake Room | 10 | Bio-Trap® | 0.17 | Jones | 10 | Bio-Trap® |
| Fort Falls | 8 | Sediment | 0.18 | Jones | 8 | Sediment | 0.18 | Jones | 9 | Sediment | 0.18 | Fort Falls | 10 | Sediment |
| Fort Falls | 10 | Sediment | 0.18 | Fort Falls | 11 | Sediment | 0.18 | Jones | 8 | Sediment | 0.18 | Lake Room | 10 | Bio-Trap® |
| Fort Falls | 11 | Sediment | 0.18 | Fort Falls | 10 | Sediment | 0.18 | Jones | 8 | Sediment | 0.19 | Fort Falls | 8 | Sediment |
| Jones | 8 | Sediment | 0.18 | Lake Room | 10 | Bio-Trap® | 0.18 | Fort Falls | 10 | Sediment | 0.18 | Lake Room | 10 | Sediment |
| Jones | 9 | Sediment | 0.18 | Lake Room | 12 | Sediment | 0.18 | Jones | 8 | Sediment | 0.18 | Fort Falls | 8 | Sediment |
| Jones | 10 | Sediment | 0.20 | Lake Room | 10 | Sediment | 0.20 | Jones | 11 | Sediment | 0.20 | Jones | 9 | Sediment |
| Jones | 11 | Sediment | 0.19 | Jones | 9 | Sediment | 0.19 | Lake Room | 10 | Sediment | 0.19 | Jones | 8 | Sediment |
| Jones | 12 | Sediment | 0.20 | Jones | 8 | Sediment | 0.21 | Lake Room | 10 | Sediment | 0.21 | Jones | 9 | Sediment |
| Lake Room | 8 | Sediment | 0.37 | Jones | 11 | Water | 0.37 | Lake Room | 9 | Sediment | 0.40 | Jones | 9 | Bio-Trap® |
| Lake Room | 9 | Sediment | 0.37 | Lake Room | 8 | Sediment | 0.39 | Lake Room | 11 | Sediment | 0.41 | Jones | 11 | Water |
| Lake Room | 10 | Sediment | 0.18 | Jones | 8 | Sediment | 0.18 | Jones | 9 | Sediment | 0.18 | Fort Falls | 8 | Sediment |
| Lake Room | 11 | Sediment | 0.39 | Lake Room | 9 | Sediment | 0.40 | Lake Room | 8 | Sediment | 0.42 | Jones | 11 | Water |
| Lake Room | 12 | Sediment | 0.18 | Jones | 9 | Sediment | 0.18 | Jones | 10 | Bio-Trap® | 0.18 | Fort Falls | 8 | Sediment |
| Fort Falls | 8 | Water | 0.23 | Lake Room | 10 | Water | 0.23 | Fort Falls | 9 | Water | 0.24 | Lake Room | 9 | Bio-Trap® |
| Fort Falls | 9 | Water | 0.21 | Jones | 10 | Bio-Trap® | 0.21 | Fort Falls | 10 | Water | 0.21 | Lake Room | 12 | Bio-Trap® |
| Fort Falls | 10 | Water | 0.21 | Fort Falls | 9 | Water | 0.21 | Lake Room | 10 | Bio-Trap® | 0.21 | Jones | 10 | Bio-Trap® |
| Fort Falls | 11 | Water | 0.24 | Fort Falls | 10 | Water | 0.24 | Fort Falls | 9 | Water | 0.25 | Fort Falls | 8 | Water |

Table A4-5 Continued

| Sample | | | Jaccard Index | Nearest Neighbor | | | Jaccard Index | Second Nearest Neighbor | | | Jaccard Index | Third Nearest Neighbor | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Location | Month | Type | | Location | Month | Type | | Location | Month | Type | | Location | Month | Type |
| Fort Falls | 12 | Water | 0.22 | Fort Falls | 10 | Water | 0.22 | Jones | 10 | Water | 0.22 | Lake Room | 10 | Bio-Trap® |
| Jones | 8 | Water | 0.23 | Lake Room | 9 | Bio-Trap® | 0.23 | Jones | 12 | Bio-Trap® | 0.23 | Lake Room | 11 | Bio-Trap® |
| Jones | 9 | Water | 0.23 | Fort Falls | 9 | Water | 0.23 | Lake Room | 10 | Water | 0.23 | Lake Room | 9 | Bio-Trap® |
| Jones | 10 | Water | 0.18 | Lake Room | 10 | Bio-Trap® | 0.19 | Jones | 10 | Bio-Trap® | 0.20 | Fort Falls | 10 | Sediment |
| Jones | 11 | Water | 0.28 | Jones | 9 | Water | 0.29 | Jones | 9 | Bio-Trap® | 0.29 | Lake Room | 9 | Bio-Trap® |
| Jones | 12 | Water | 0.23 | Lake Room | 11 | Water | 0.24 | Lake Room | 10 | Water | 0.24 | Jones | 8 | Water |
| Lake Room | 8 | Water | 0.23 | Lake Room | 9 | Bio-Trap® | 0.23 | Lake Room | 12 | Bio-Trap® | 0.23 | Jones | 9 | Bio-Trap® |
| Lake Room | 10 | Water | 0.20 | Jones | 12 | Bio-Trap® | 0.20 | Lake Room | 12 | Bio-Trap® | 0.21 | Lake Room | 9 | Bio-Trap® |
| Lake Room | 11 | Water | 0.21 | Fort Falls | 9 | Water | 0.22 | Lake Room | 12 | Sediment | 0.22 | Jones | 12 | Bio-Trap® |
| Lake Room | 12 | Water | 0.40 | Lake Room | 8 | Sediment | 0.40 | Jones | 11 | Water | 0.43 | Lake Room | 11 | Sediment |

# Chapter 5 CONCLUSIONS

The overall aim of this dissertation was to fill in the large knowledge gap regarding the controls that diverse microbial groups have on the nature of OM, carbon, and nutrients in the most common type of cave system over time. An epigenic cave system was chosen because of its ease of access for sampling, its close physical location to the surface, and environmental disturbances that include floods and changes in nutrients over time. The first objective in this dissertation was to survey and compare known bacterial diversity from all publically available cave and karst NCBI GenBank sequences. Unfortunately, the quantity of geological and geochemical metadata associated with these sequences only allowed for a broad generalizations about the bacterial diversity from cave and karst environment types. Also, there are still many caves and cave system types that have not had their microbial diversity thoroughly assessed. Cave microbial diversity is highly underrepresented compared to other terrestrial habitats on Earth. Nonetheless, some OTUs were found from caves separated by 1000s of kilometers, suggesting that some bacterial groups in caves may be globally distributed and likely reflect the geological and environmental conditions of the cave habitat and not biogeographic barriers to distribution. It is possible that there are broad-scale geochemical and ecological processes that affect the distribution of microbial communitites in cave systems.

These results provide evidence that bacterial community diversity of cave systems is not unique to each cave system. Because of their isolation, limited energy, and limited hydrologic connectivity to the surface, and a small number of studies in general, it has been assumed that microbial diversity in cave systems is uncomparable. If cave bacterial community compositions are not unique, then it is possible to compare findings from one cave system to another.

Consequently, the bacterial diversity and functional processes occurring in Cascade Cave System, Kentucky, can be correlated to other non-sulfidic limestone caves. This outlook will change the way scientists can and will study the microbiology and gemicrobiology of cave systems in the future.

The second objective in this dissertation was to survey the bacterial diversity of the Cascade Cave System over time, as well as evaluate changes in aqueous geochemistry, flood disturbances, and sediment mobilization in the cave and compare those features to potential diversity changes. In general, there are very few temporal studies of microbial diversity correlated to changes in environmental conditions (Griebler & Lueders, 2009; Engel, 2010; Engel, 2015), and this dissertation increased the collective knowledge about the effects of disturbance in the most common type of cave system (Palmer, 2007). There were several environmental disturbances that occurred in the cave system during the study period, sediment remobilization and DOM quality and quantity changes. Several distinct, shared planktonic and attached bacterial communities were observed from the cave stream. However, although we found shared OTUs that were stable for the duration of our study, there were no OTUs shared between the planktonic and attached microbial communities. Therefore, there was no evidence for a shared or stable microbial community across all environment types in the cave system. The bacterial succession in Cascade Cave System stabilized over time in both locations along the stream flowpath, providing evidence that succession following large-scale environmental disturbances does occur in cave streams. Also, the planktonic microorganisms in Cascade Cave system were the pioneering community, and there were differences in the abundance of

shared planktonic and attached communities at the end of the study in the two cave locations. We also found that sediment size and mobilization play a key role in the sediment-attached karst microbial community structure, and organic carbon quality governs the planktonic karst microbial community structure in a cave streams. Following the results about microbial diversity from other limestone cave systems around the world being similar to Cascade Cave System, it is possible that some of the same relationships between microbial diversity and environmental disturbances could be occurring.

The last goal of this dissertation was to assess functional capabilities of microorganisms in Cascade Cave System, Kentucky in order to understand functional changes in the cave system over time, as well as functional differences between the planktonic and attached environment types inside the cave system. Most of the function-level abundance of reads were not significantly different between environment types, however the types of genes that were significantly different based on environment type were related to differences of habitat type (attachment, mobility, and secondary metabolism). These results are the first known genetic functional differences between environment types in a cave and karst environment type. The reads matching genes involved in the degradation of carbohydrates and lignin are present regardless of environment type, even though the microbial community diversity in each environment type changes over time. The genes are also present despite environmental disturbances, such as the quantity and quality of DOM that is transported though the cave system, and sediment mobilization. The functional succession only captured functional similarity to attached microorganisms, which is in contrast to the taxonomic succession. Also,

because many of the functional genes that were shared between all of the environment types were from many different taxonomic groups, different species of microorganisms are able to provide similar functions to their cave ecosystem habitat type. The microorganisms are constantly chaning over time in the cave system, but the functional redundancy of cellulose and lignin bacterial degradation is able to provide the cave ecosystem with the degradation products regardless of environmental disturbances. These results mean that functional redundancy between bacteria is an important ecosystem factor for the assessment of the cave ecosystem's resiliency to environmental disturbances.

Finally, all of the code in this dissertation can enable anyone to reproduce the results from this dissertation (provided the data), or analyze different data in the same manner. There is not a single package of code in R that includes functions or graphics for all of the analyses in this dissertation. The code provides the cutting-edge methods for microbial sequence normalization, analyses, and graphics, as well as geochemical data analyses. As data sets become larger, it will soon not be possible to open or visualize a data set in some basic programs like Text Editors and Microsoft Excel. The code also provides some summary functions to summarize large data sets and results for data sets that are too large to open in basic programs.

## REFERENCES

Engel, A.S. (2010). Microbial diversity of cave ecosystems, in Barton, L., Mandl, M., and Loy, A. (eds.), *Geomicrobiology: Molecular & Environmental Perspectives*, Springer. p. 219-238. DOI:10.1007/978-90-481-9204-5_10.

Engel, A.S. (2015). Bringing microbes into focus for cave science: An introduction, in A.S. Engel (ed.), *Microbial Life of Cave Systems,* De Gruyter. p. 1-22.

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

Palmer, A.N. (2007). *Cave Geology.* Cave books, Dayton.

**APPENDIX IV**

**Code I**

```
####################################Parsing GenBank files for information

library(ape)
library(plyr)
library(reshape2)
library(dplyr)
library(rlist)
## read.GenBank.R (2012-02-17)

##   Read DNA Sequences from GenBank via Internet

## Copyright 2002-2012 Emmanuel Paradis

## This file is part of the R-package `ape'.
## See the file ../COPYING for licensing issues.

####################function pulled from Brian O'Meara's github #page
#https://github.com/bomeara/genbankcredit/blob/master/notes.md

read.GenBank <-
  function(access.nb, seq.names = access.nb, species.names = TRUE,
           gene.names = FALSE, as.character = FALSE, pubmed = TRUE)
  {
    N <- length(access.nb)
    ## If there are more than 400 sequences, we need to break down the
    ## requests, otherwise there is a segmentation fault.
    nrequest <- N %/% 400 + as.logical(N %% 400)
    X <- character(0)
    for (i in 1:nrequest) {
      a <- (i - 1) * 400 + 1
      b <- 400 * i
      if (i == nrequest) b <- N
      URL <- paste("http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nucleotide&id=",
                   paste(access.nb[a:b], collapse = ","),
                   "&rettype=gb&retmode=text", sep = "")
      X <- c(X, scan(file = URL, what = "", sep = "\n", quiet = TRUE))
    }
    FI <- grep("^ {0,}ORIGIN", X) + 1
    LA <- which(X == "//") - 1
    obj <- vector("list", N)
    for (i in 1:N) {
      ## remove all spaces and digits
      tmp <- gsub("[[:digit:]]", "", X[FI[i]:LA[i]])
      obj[[i]] <- unlist(strsplit(tmp, NULL))
```

```
  }
  names(obj) <- seq.names
  if (!as.character) obj <- as.DNAbin(obj)
  if (species.names) {
   tmp <- character(N)
   sp <- grep("ORGANISM", X)
   for (i in 1:N)
    tmp[i] <- unlist(strsplit(X[sp[i]], " +ORGANISM +"))[2]
   attr(obj, "species") <- gsub(" ", "_", tmp)
  }
  if (gene.names) {
   tmp <- character(N)
   sp <- grep(" +gene +<", X)
   for (i in 1:N)
    tmp[i] <- unlist(strsplit(X[sp[i + 1L]], " +/gene=\""))[2]
   attr(obj, "gene") <- gsub("\"$", "", tmp)
  }
  if (pubmed) {
   tmp <- vector("list", N)
   endPub <- grep("//", X)
   refs <- grep("^REFERENCE", X)
   pub <- grep("^\\s+PUBMED", X)
   auth <- grep("^\\s+AUTHORS", X)
   title <- grep("^\\s+TITLE", X)
   journal <- grep("^\\s+JOURNAL", X)
   feat <- grep("^FEATURES", X)
   for (i in 1:N) {
    begPub <- ifelse(i == 1, 1, endPub[i-1])
    nRefs <- refs[refs > begPub & refs < endPub[i]]
    refLst <- vector("list", length(nRefs))
    for (j in 1:length(nRefs)) {
     rgRef <- c( nRefs[j], ifelse(j == length(nRefs), feat[i], nRefs[j+1]))
     tmpRes <- vector("list", 4)
     names(tmpRes) <- c("pubmedid", "authors", "title", "journal")
     tmpRes$pubmedid <- gsub("^\\s+PUBMED\\s+(\\d+)", "\\1", X[pub[pub > rgRef[1] & pub <
rgRef[2]]])
     tmpRes$authors <- paste0(X[auth[j]:(title[j]-1)], collapse=" ")
     tmpRes$title <- paste0(X[title[j]:(journal[j]-1)], collapse=" ")
     tmpRes$journal <- paste0(X[journal[j]], collapse=" ") # JOURNAL always 1 line?
     tmpRes <- lapply(tmpRes, function(x) { gsub("\\s{2,}", " ", gsub("\\s+[A-Z]+\\s+", "", x)) })
     refLst[[j]] <- tmpRes
    }
    tmp[[i]] <- refLst
   }
   names(tmp) <- access.nb
   attr(obj, "references") <- tmp
```

```
  }
  obj
}

############get Accession list from NCBI GenBank of your seqs
gi_numbers<-read.csv("sequence.gi2.csv",header=FALSE)
str(gi_numbers)

#let's seperate out large data set into 10000 increments
gi_sample<-as.data.frame(gi_numbers[1:10000,])
gi_sample2<-as.data.frame(gi_numbers[10001:20000,])
gi_sample3<-as.data.frame(gi_numbers[20001:30000,])
gi_sample4<-as.data.frame(gi_numbers[30001:40000,])
gi_sample5<-as.data.frame(gi_numbers[40001:50000,])
gi_sample6<-as.data.frame(gi_numbers[50001:60000,])
gi_sample7<-as.data.frame(gi_numbers[60001:70000,])
gi_sample8<-as.data.frame(gi_numbers[70001:80000,])
gi_sample9<-as.data.frame(gi_numbers[80001:90000,])
gi_sample10<-as.data.frame(gi_numbers[90001:104551,])



#########now get GenBank info going through a loop
#may have to do this a few times because it is dependant on the internet

my_output1 <- apply(gi_sample, 1, function(x) read.GenBank(x))
my_output3 <- apply(gi_sample3, 1, function(x) read.GenBank(x))
my_output4 <- apply(gi_sample4, 1, function(x) read.GenBank(x))
my_output8 <- apply(gi_sample8, 1, function(x) read.GenBank(x))
my_output2 <- apply(gi_sample2, 1, function(x) read.GenBank(x))
my_output5 <- apply(gi_sample5, 1, function(x) read.GenBank(x))
my_output6 <- apply(gi_sample6, 1, function(x) read.GenBank(x))
my_output7 <- apply(gi_sample7, 1, function(x) read.GenBank(x))
my_output9 <- apply(gi_sample9, 1, function(x) read.GenBank(x))
my_output10 <- apply(gi_sample10, 1, function(x) read.GenBank(x))



# Write a function to extract the title of the first reference in each element, Thanks to Drew Steen for a
little help subsetting

get_info <- function(x) {
  ti <- attr(x, "references")[[1]][[1]]$title
  nm <- names(attr(x, "references"))[1]
  jour<-attr(x, "references")[[1]][[2]]$journal
```

```
  # Return a named vector
  c(ti=ti, nm=nm,jour=jour)
}

# GEt the names and titles out
seq_info1 <- ldply(my_output1, get_info)
seq_info2 <- ldply(my_output2, get_info)
seq_info3 <- ldply(my_output3, get_info)
seq_info4 <- ldply(my_output4, get_info)
seq_info5 <- ldply(my_output5, get_info)
seq_info6 <- ldply(my_output6, get_info)
seq_info7 <- ldply(my_output7, get_info)
seq_info8 <- ldply(my_output8, get_info)
seq_info9 <- ldply(my_output9, get_info)
seq_info10 <- ldply(my_output10, get_info)


###put all that info together
all_seq_info<-
rbind(seq_info1,seq_info2,seq_info3,seq_info4,seq_info5,seq_info6,seq_info7,seq_info8,seq_info9,seq
_info10)


head(all_seq_info)
###srite to file, and now you can look up the references
write.csv(all_seq_info,"all_seq_info.csv")


##############################################analysis of sequence data

library(ape)
library(vegan)
library(phyloseq)
require(ggplot2)
library(biom)
library(rlist)
library(plyr)
library(reshape2)
library(phylogeo)
library(cluster)
library(dplyr)


################################################
```

```
##import_biom does not work for this biom file sadly. if it works for you, skip all this and just do
that
##import using biom package
x<-read_biom("otu_table_by_sample.biom")
x2<-as(biom_data(x), "matrix")
#make an otu matrix
otu<-otu_table(x2,taxa_are_rows=TRUE)


#taxonomy is a bit more dificult
taxa<-observation_metadata(x)
taxa_format <- do.call(rbind, lapply(lapply(taxa, unlist), "[",
                       unique(unlist(c(sapply(taxa,names))))))
tax_table1<-tax_table(taxa_format)

#sample metadata
meta<-read.csv("phyloseq_sample_meta3.csv",header=TRUE)
row.names(meta)<-meta$sample_name
str(meta)
meta_ready<-sample_data(meta)
str(meta_ready)

#now create the phyloseq object
finally<- phyloseq(otu,tax_table1,meta_ready)
finally

#import a tree for our phyloseq object
tree<-read.tree("rep_set_aligned.tre")

#now merge it all together
physeq1 = merge_phyloseq(finally, tree)
physeq1
str(sample_data(physeq1))
#How many unique Genera are there?
taxa<-as.data.frame(tax_table(physeq1))
#How many classified OTUs do we have?
apply(taxa, 2, function(x) length(which(!is.na(x))))
#How many unclassified OTUs do we have?
apply(taxa, 2, function(x) length(which(is.na(x))))

#How many sequences do we have?
sum(otu_table(physeq1))
```

```
#####get rid of host objects
physeq3<-subset_samples(physeq1, !env_type %in%
c("beetle_host","bat_host","amphipod_host","sandfly_host"))
physeq4<-subset_samples(physeq1, !env_type %in%
c("beetle_host","bat_host","amphipod_host","sandfly_host","ice_cave"))

#########how many phyla do we have?
tax<-tax_table(physeq3)
str(tax)

length(unique(tax_table(physeq3)[,2]))


##############################################################################
##
#phew all that just to get it imported. Now let's do the work


#########################################################diversity analyses

richness<- plot_richness(physeq3, x = "env_type",measures = c("Chao1", "Shannon"))
richness
ggsave("richness.tiff",richness,height=6,width=6,units="in",dpi=300)



#########################################################let's do a PCoA
ordu = ordinate(physeq3, "PCoA", "unifrac",weighted=TRUE)
pcoa1<-plot_ordination(physeq1, ordu, color = "env_type") +
  scale_colour_manual(values =  c("#9e0142","#d53e4f","#f46d43","#fdae61","#fee08b",
                                  "#000000","#e6f598","#abdda4","#66c2a5","#3288bd","#5e4fa2"))     +
  geom_point(size = 3, alpha = 0.75)
pcoa1

ggsave("pcoa1.tiff",pcoa1,height=4,width=6,units="in",dpi=300)


##########ADONIS
dist<-distance(physeq3,"unifrac")
```

```
meta_physeq3<-as.data.frame(sample_data(physeq3))
adonis(dist~meta_physeq3$env_type, permutations=9999)

#####################################################gap stat and clusters
##########first do an ordination
exord = ordinate(physeq1, method = "MDS", distance = "unifrac")

########################################code from phyloseq tutorial
#########this code gets ready to do a gap statistic analysis
###########got through and change K.max if you want to change the number of ####clusters
you test for.
pam1 = function(x, k) {
  list(cluster = pam(x, k, cluster.only = TRUE))
}
x = phyloseq:::scores.pcoa(exord, display = "sites")
# gskmn = clusGap(x[, 1:2], FUN=kmeans, nstart=20, K.max = 9, B = 500)
gskmn = clusGap(x[, 1:2], FUN = pam1, K.max = 9, B = 50)
gskmn
gap_statistic_ordination = function(ord, FUNcluster, type = "sites", K.max = 9,
                                    axes = c(1:2), B = 500, verbose = interactive(), ...) {
  require("cluster")
  # If 'pam1' was chosen, use this internally defined call to pam
  if (FUNcluster == "pam1") {
    FUNcluster = function(x, k) list(cluster = pam(x, k, cluster.only = TRUE))
  }
  # Use the scores function to get the ordination coordinates
  x = phyloseq:::scores.pcoa(ord, display = type)
  # If axes not explicitly defined (NULL), then use all of them
  if (is.null(axes)) {
    axes = 1:ncol(x)
  }
  # Finally, perform, and return, the gap statistic calculation using
  # cluster::clusGap
  clusGap(x[, axes], FUN = FUNcluster, K.max = K.max, B = B, verbose = verbose,
          ...)
}

plot_clusgap = function(clusgap, title = "Gap Statistic calculation results") {
  require("ggplot2")
  gstab = data.frame(clusgap$Tab, k = 1:nrow(clusgap$Tab))
  p = ggplot(gstab, aes(k, gap)) + geom_line() + geom_point(size = 5)
  p = p + geom_errorbar(aes(ymax = gap + SE.sim, ymin = gap - SE.sim))
```

```
  p = p + ggtitle(title)
  return(p)
}

gs = gap_statistic_ordination(exord, "pam1", B = 50, verbose = FALSE)
print(gs, method = "Tibs2001SEmax")
###################################
#how many clusters should we have?
clustgap<- plot_clusgap(gs)


#now let's actually calculate the clusters. This is memory and time intensive, especially if you
use your whole tree
cluster_map<- map_clusters(physeq2, clusternum=3)
cluster_map

ggsave("global_clust2.tiff",cluster_map,height=15,width=6,units="in",dpi=300)
```

**Code II**

R Scripts

```
require(grid) #this is only required for the envfit arrows.
## Loading required package: grid
require(plyr) #make sure plyr is loaded b4 dplyr
require(dplyr) #data formatting and more
require(phyloseq) #OTU analysis
require(reshape2) #to get data into long format
require(ggplot2) #plotting
require(RColorBrewer) #ColorBrewer palettes
require(vegan) #Community Ecology Package: Ordination, Diversity and Dissimilarities
require(rmarkdown) #to make this script document
require(knitr) #to make this script document
require(BiodiversityR)
Load in data to a phyloseq object
#import qiime biom with taxonomy and tre file
taxvec1 = c("k__Bacteria", "p__Firmicutes", "c__Bacilli", "o__Bacillales",
"f__Staphylococcaceae")
mydata<-
import_biom("otu_table_mc2_w_tax.biom","rep_set.tre",parseFunction=parse_taxonomy_gre
engenes)
## Warning in parseFunction(i$metadata$taxonomy): No greengenes prefixes were found.
## Consider using parse_taxonomy_default() instead if true for all OTUs.
## Dummy ranks may be included among taxonomic ranks now.
parse_taxonomy_greengenes(taxvec1)
#import mapping file
qiimedata <- import_qiime_sample_data("metadata_mapping.txt")
#merge them together
data <- merge_phyloseq(mydata, qiimedata)
Run numbers for general stats about our phyloseq object and full data set
#How many unique Genera are there?
taxa<-as.data.frame(tax_table(data))
#How many classified OTUs do we have?
apply(taxa, 2, function(x) length(which(!is.na(x))))
#How many unclassified OTUs do we have?
apply(taxa, 2, function(x) length(which(is.na(x))))

#How many sequences do we have?
sum(otu_table(data))
```

```
#How many OTUs do we have in each sample?
otus<-as.data.frame(otu_table(data))
head(otus)
num_otus<-as.data.frame(colSums(otus != 0))

#Let's see which are the most abundant Classes
class.sum <- tapply(taxa_sums(data), tax_table(data)[, "Class"], sum, na.rm = TRUE)
class.sum.table<-as.data.frame(class.sum)
class.sum.table$frac.abund<-(class.sum.table$class.sum/sum(class.sum.table$class.sum))*100
class.sum.table<-sort(class.sum.table$frac.abund, TRUE)
class.sum.table<-as.data.frame(class.sum.table)
OTU similarity table analyses
###############water
water <- subset_samples(data, Type == "Water")

#####fort falls
ff_w <- subset_samples(water, Location == "Surface")
shared_w_ff2<-filter_taxa(ff_w, function(x) (sum(x) !=0 ), TRUE)
shared_w_ff<-filter_taxa(ff_w, function(x) all(x !=0 ), TRUE)

##OTU and tax shared tables
shared_w_ff_1<-
cbind(as.data.frame(otu_table(shared_w_ff)),as.data.frame(tax_table(shared_w_ff)))
shared_w_ff_1$otu<-rownames(shared_w_ff_1)
head(shared_w_ff_1)
shared_w_ff_1_melt<-melt(shared_w_ff_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_w_ff_1_melt)

#total number of seqs
num_seqs_w_ff<-sum(as.data.frame(otu_table(ff_w)))

#Add our other vaiables, but really we just need month
melt_vars1<-as.data.frame(cbind("sample"=as.character(qiimedata$Sample.ID),
                        "type"=as.character(qiimedata$Type),
                        "location"=as.character(qiimedata$Location),
                        "month"=as.character(qiimedata$Month)))

rownames(melt_vars1)<-melt_vars1$sample
head(melt_vars1)
```

```
#use merge to add metadata to melted similarity data frame
shared_w_ff_1_melt2<-merge(shared_w_ff_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_w_ff_1_melt2)

w_ff_melt<-shared_w_ff_1_melt2[,c(1,3:4,10:11,13:14)]
head(w_ff_melt)

w_ff_melt2<-w_ff_melt %>%
  group_by(Phylum,Class,type,location) %>%
  summarise(abundance=sum(value)/num_seqs_w_ff*100,
            OTUs=n()/7)

#####Upstream
j_w <- subset_samples(water, Location == "Upstream")
shared_w_j<-filter_taxa(j_w, function(x) all(x !=0 ), TRUE)
shared_w_j2<-filter_taxa(j_w, function(x) (sum(x) !=0 ), TRUE)

##OTU and tax shared tables
shared_w_j_1<-
cbind(as.data.frame(otu_table(shared_w_j)),as.data.frame(tax_table(shared_w_j)))
shared_w_j_1$otu<-rownames(shared_w_j_1)
head(shared_w_j_1)
shared_w_j_1_melt<-melt(shared_w_j_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_w_j_1_melt)

#total number of seqs
num_seqs_w_j<-sum(as.data.frame(otu_table(j_w)))

#use merge to add metadata to melted similarity data frame
shared_w_j_1_melt2<-merge(shared_w_j_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_w_j_1_melt2)

w_j_melt<-shared_w_j_1_melt2[,c(1,3:4,10:11,13:14)]
head(w_j_melt)

w_j_melt2<-w_j_melt %>%
  group_by(Phylum,Class,type,location) %>%
  summarise(abundance=sum(value)/num_seqs_w_j*100,
            OTUs=n()/5)
```

```
#####lake room
lr_w <- subset_samples(water, Location == "Downstream")
shared_w_lr<-filter_taxa(lr_w, function(x) all(x !=0 ), TRUE)
shared_w_lr2<-filter_taxa(lr_w, function(x) (sum(x) !=0 ), TRUE)

##OTU and tax shared tables
shared_w_lr_1<-
cbind(as.data.frame(otu_table(shared_w_lr)),as.data.frame(tax_table(shared_w_lr)))
shared_w_lr_1$otu<-rownames(shared_w_lr_1)
head(shared_w_lr_1)
shared_w_lr_1_melt<-melt(shared_w_lr_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_w_lr_1_melt)

#total number of seqs
num_seqs_w_lr<-sum(as.data.frame(otu_table(lr_w)))


#use merge to add metadata to melted similarity data frame
shared_w_lr_1_melt2<-merge(shared_w_lr_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_w_lr_1_melt2)

w_lr_melt<-shared_w_lr_1_melt2[,c(1,3:4,10:11,13:14)]
head(w_lr_melt)

w_lr_melt2<-w_lr_melt %>%
  group_by(Phylum,Class,type,location) %>%
  summarise(abundance=sum(value)/num_seqs_w_lr*100,
            OTUs=n()/8)



################biotrap
biotrap <- subset_samples(data, Type== "Bio-Trap¨")

#####Upstream
j_bio <- subset_samples(biotrap, Location == "Upstream")
shared_b_j<-filter_taxa(j_bio, function(x) all(x !=0 ), TRUE)
shared_b_j2<-filter_taxa(j_bio, function(x) (sum(x) !=0 ), TRUE)

##OTU and tax shared tables
```

```
shared_b_j_1<-
cbind(as.data.frame(otu_table(shared_b_j)),as.data.frame(tax_table(shared_b_j)))
shared_b_j_1$otu<-rownames(shared_b_j_1)
head(shared_b_j_1)
shared_b_j_1_melt<-melt(shared_b_j_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_b_j_1_melt)

#total number of seqs
num_seqs_b_j<-sum(as.data.frame(otu_table(j_bio)))

#use merge to add metadata to melted similarity data frame
shared_b_j_1_melt2<-merge(shared_b_j_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_b_j_1_melt2)

b_j_melt<-shared_b_j_1_melt2[,c(1,3:4,10:11,13:14)]
head(b_j_melt)

b_j_melt2<-b_j_melt %>%
 group_by(Phylum,Class,type,location) %>%
 summarise(abundance=sum(value)/num_seqs_b_j*100,
        OTUs=n()/7)


#####lake room
lr_bio <- subset_samples(biotrap, Location == "Downstream")
shared_b_lr<-filter_taxa(lr_bio, function(x) all(x !=0 ), TRUE)
shared_b_lr2<-filter_taxa(lr_bio, function(x) (sum(x) !=0 ), TRUE)

##OTU and tax shared tables
shared_b_lr_1<-
cbind(as.data.frame(otu_table(shared_b_lr)),as.data.frame(tax_table(shared_b_lr)))
shared_b_lr_1$otu<-rownames(shared_b_lr_1)
head(shared_b_lr_1)
shared_b_lr_1_melt<-melt(shared_b_lr_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_b_lr_1_melt)

#total number of seqs
num_seqs_b_lr<-sum(as.data.frame(otu_table(lr_bio)))

#use merge to add metadata to melted similarity data frame
```

```
shared_b_lr_1_melt2<-merge(shared_b_lr_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_b_lr_1_melt2)

b_lr_melt<-shared_b_lr_1_melt2[,c(1,3:4,10:11,13:14)]
head(b_lr_melt)

 b_lr_melt2<-b_lr_melt %>%
 group_by(Phylum,Class,type,location) %>%
 summarise(abundance=sum(value)/num_seqs_b_lr*100,
        OTUs=n()/8)


###############sed
sediment <- subset_samples(data, Type == "Sediment")

#####fort falls
ff_s <- subset_samples(sediment, Location == "Surface")
shared_s_ff<-filter_taxa(ff_s, function(x) all(x !=0 ), TRUE)
shared_s_ff2<-filter_taxa(ff_s, function(x) (sum(x) !=0 ), TRUE)

##OTU and tax shared tables
shared_s_ff_1<-
cbind(as.data.frame(otu_table(shared_s_ff)),as.data.frame(tax_table(shared_s_ff)))
shared_s_ff_1$otu<-rownames(shared_s_ff_1)
head(shared_s_ff_1)
shared_s_ff_1_melt<-melt(shared_s_ff_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_s_ff_1_melt)

#total number of seqs
num_seqs_s_ff<-sum(as.data.frame(otu_table(ff_s)))


#use merge to add metadata to melted similarity data frame
shared_s_ff_1_melt2<-merge(shared_s_ff_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_s_ff_1_melt2)

s_ff_melt<-shared_s_ff_1_melt2[,c(1,3:4,10:11,13:14)]
head(s_ff_melt)

s_ff_melt2<-s_ff_melt %>%
 group_by(Phylum,Class,type,location) %>%
```

```
 summarise(abundance=sum(value)/num_seqs_s_ff*100,
        OTUs=n()/3)


#####Upstream
j_s <- subset_samples(sediment, Location == "Upstream")
shared_s_j<-filter_taxa(j_s, function(x) all(x !=0 ), TRUE)
shared_s_j2<-filter_taxa(j_s, function(x) (sum(x) !=0 ), TRUE)

##OTU and tax shared tables
shared_s_j_1<-
cbind(as.data.frame(otu_table(shared_s_j)),as.data.frame(tax_table(shared_s_j)))
shared_s_j_1$otu<-rownames(shared_s_j_1)
head(shared_s_j_1)
shared_s_j_1_melt<-melt(shared_s_j_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_s_j_1_melt)

#total number of seqs
num_seqs_s_j<-sum(as.data.frame(otu_table(j_s)))


#use merge to add metadata to melted similarity data frame
shared_s_j_1_melt2<-merge(shared_s_j_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_s_j_1_melt2)

s_j_melt<-shared_s_j_1_melt2[,c(1,3:4,10:11,13:14)]
head(s_j_melt)

s_j_melt2<-s_j_melt %>%
 group_by(Phylum,Class,type,location) %>%
 summarise(abundance=sum(value)/num_seqs_s_j*100,
        OTUs=n()/5)

#####lake room
lr_s <- subset_samples(sediment, Location == "Downstream")
shared_s_lr<-filter_taxa(lr_s, function(x) all(x !=0 ), TRUE)
shared_s_lr2<-filter_taxa(lr_s, function(x) (sum(x) !=0 ), TRUE)

##OTU and tax shared tables
shared_s_lr_1<-
cbind(as.data.frame(otu_table(shared_s_lr)),as.data.frame(tax_table(shared_s_lr)))
```

```
shared_s_lr_1$otu<-rownames(shared_s_lr_1)
head(shared_s_lr_1)
shared_s_lr_1_melt<-melt(shared_s_lr_1)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_s_lr_1_melt)

#total number of seqs
num_seqs_s_lr<-sum(as.data.frame(otu_table(lr_s)))


#use merge to add metadata to melted similarity data frame
shared_s_lr_1_melt2<-merge(shared_s_lr_1_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
str(shared_s_lr_1_melt2)

s_lr_melt<-shared_s_lr_1_melt2[,c(1,3:4,10:11,13:14)]
head(s_lr_melt)

s_lr_melt2<-s_lr_melt %>%
  group_by(Phylum,Class,type,location) %>%
  summarise(abundance=sum(value)/num_seqs_s_lr*100,
            OTUs=n()/5)


###################make our final table!
shared_table<-
as.data.frame(rbind(s_ff_melt2,s_j_melt2,s_lr_melt2,b_j_melt2,b_lr_melt2,w_ff_melt2,w_j_m
elt2,w_lr_melt2))
shared_table_otus<-dcast(shared_table,Phylum+Class~type+location,value.var="OTUs")
shared_table_abund<-
dcast(shared_table,Phylum+Class~type+location,value.var="abundance")
Anosim and Adonis analyses
#dissimilarity matrix
bray<-ordinate(data,distance="bray",method="NMDS")

#Adonis
bray_dist<-distance(data,"bray")

type<- adonis(bray_dist~sample_data(data)$Type,permutations=9999)
month<- adonis(bray_dist~sample_data(data)$Month,permutations=9999)
location<- adonis(bray_dist~sample_data(data)$Location,permutations=9999)
densityplot(permustats(location))
```

```
fi<- adonis(bray_dist~sample_data(data)$FI,permutations=9999)
hix<- adonis(bray_dist~sample_data(data)$HIX,permutations=9999)
NMDS script
Data processing
data_filter= filter_taxa(data, function(x) sum(x > 3) > (0.2*length(x)), TRUE)
ntaxa(data)
ntaxa(data_filter)

#Standardize abundances to the median sequencing depth
total = median(sample_sums(data_filter))
standf = function(x, t=total) round(t * (x / sum(x)))
data_trans = transform_sample_counts(data_filter, standf)

#Filter the taxa using a cutoff of 2.0 for the Coefficient of Variation
data_trans_cv = filter_taxa(data_trans, function(x) sd(x)/mean(x) > 2.0, TRUE)
ntaxa(data_trans_cv)
NMDS plot
taken from the excellent stackoverflow Q+A:
http://stackoverflow.com/questions/13794419/plotting-ordiellipse-function-from-vegan-package-onto-nmds-plot-created-in-ggplo
#Ordinate NMDS using Bray, use stress plot to find the best dissimilarit/distance
nmds.bray <- ordinate(data,"NMDS","bray")
stressplot(nmds.bray)

as.data.frame(nmds.bray$points)

environ<-data.frame(sample_data(data)[,2:16])
str(environ)
##site data
sites <- data.frame(scores(nmds.bray, choices=c(1,2),display = c("sites"))) #dataframe of species scoes for plotting
head(sites)
sites$Location <- sample_data(data)$Location #otherwise factor doesn't drop unused levels and it will throw an error
sites$Type <- sample_data(data)$Type
sites$Month <- sample_data(data)$Month

#envfit
nmds.bray.envfit <- envfit(as.data.frame(nmds.bray$points),
                    env = environ,na.rm=TRUE, perm = 999) #standard envfit
plot(nmds.bray)
plot(nmds.bray.envfit,p=0.05)
```

```
#ellipses
# function for ellipsess - just run this, is used later
veganCovEllipse <- function (cov, center = c(0, 0), scale = 1, npoints = 100)
{
  theta <- (0:npoints) * 2 * pi/npoints
  Circle <- cbind(cos(theta), sin(theta))
  t(center + scale * t(Circle %*% chol(cov)))
}

#data for ellipse, in this case using the management factor
df_ell.dune.management <- data.frame() #sets up a data frame before running the function.
for(g in levels(sites$Type)){
  df_ell.dune.management <- rbind(df_ell.dune.management, cbind(as.data.frame(with(sites
[sites$Type==g,],

veganCovEllipse(cov.wt(cbind(NMDS1,NMDS2),wt=rep(1/length(NMDS1),length(NMDS1)))$co
v,center=c(mean(NMDS1),mean(NMDS2)))))
                                          ,Type=g))
}

# data for labelling the ellipse
NMDS.mean.biotrap=aggregate(sites[ ,c("NMDS1", "NMDS2")],
                    list(group = sites$Type), mean)

# data for labelling the ellipse
NMDS.mean=aggregate(sites[,c("NMDS1", "NMDS2")],
                list(group = sites$Type), mean)


## finally plotting.
getPalette = colorRampPalette(brewer.pal(8, "Accent"))
#display.brewer.pal(8,"Dark2")

nmds2 <- ggplot(data = sites, aes(x = NMDS1, y = NMDS2))+ #sets up the plot. brackets around
the entire thing to make it draw automatically
  geom_point(aes(x = NMDS1, y = NMDS2, shape=Type),size = 6) + #puts the site points in from
the ordination, shape determined by site, size refers to size of point
  geom_path(data = df_ell.dune.management, aes(x = NMDS1, y = NMDS2, group = Type)) +
#this is the ellipse, seperate ones by type. If you didn't change the "alpha" (the shade) then you
need to keep the "group  annotate("text",x = -0.70,y = 0.5,label="Filter",size=6) + #labels for the
centroids - I haven't used this since we have a legend. but you could also dithc the legend, but
```

```
plot will get v messy
  theme_bw(base_size = 15) +
  theme(axis.text.x=element_blank(),
                    axis.text.y=element_blank(),axis.ticks=element_blank(),
                    axis.title.x=element_blank(), axis.title.y=element_blank()
  )

ggsave(file="NMDS_plot.tiff",plot=nmds2,width=10,height=7,units="in",dpi=300)
PCA/RDA plot
with help from https://oliviarata.wordpress.com/2014/09/06/rda-in-ggplot2/
require(grid) #this is only required for the envfit arrows.
require(plyr) #make sure plyr is loaded b4 dplyr
require(dplyr) #data formatting and more
require(phyloseq) #OTU analysis
require(reshape2) #to get data into long format
require(ggplot2) #plotting
require(RColorBrewer) #ColorBrewer palettes
require(vegan) #Community Ecology Package: Ordination, Diversity and Dissimilarities
require(rmarkdown) #to make this script document
require(knitr) #to make this script document
#create a veganified phyloseq object
veganotu <- function(physeq) {
  OTU <- otu_table(physeq)
  if (taxa_are_rows(OTU)) {
    OTU <- t(OTU)
  }
  OTU <- as(OTU, "matrix")
  return(OTU)
}


vegan_otu<-as.data.frame(veganotu(data_trans_cv))
sample.data<-data.frame(sample_data(data_trans_cv))

all.rda<- rda(vegan_otu~Type+HIX+FI, data=sample.data)
all.rda
PCAsignificance(all.rda,axes=4)
plot(all.rda)

scor = scores(all.rda, display=c("sp", "cn", "bp"), scaling=2)

# type centroids
```

```
type_numeric_centroids <- data.frame(scor$centroids)
type_numeric_centroids
type_numeric_centroids$type <- rownames(type_numeric_centroids)
type_numeric_centroids

#sites
site_scores <- data.frame(scores(all.rda)$sites)
site_scores$type <-sample_data(data_trans_cv)$Type
str(site_scores)

# arrows
type_continuous_arrows <- data.frame(scor$biplot)
type_continuous_arrows
type_continuous_arrows$type_class <- rownames(type_continuous_arrows) #turning
rownames into a variable
type_continuous_arrows

mult <- attributes(scores(all.rda))$const # scaling for the arrows

RDA_plot <- ggplot(site_scores, aes(x = RDA1, y = RDA2))+
  theme_bw() +
  geom_point(aes(size = 3,shape=type)) +
  scale_shape_manual(values = c('Bio-Trap®' = 17, 'Water' = 16,
'Sediment'=15),guide="legend") +
  geom_segment(data = type_continuous_arrows,
           aes(x = 0, xend = mult * RDA1,
               y = 0, yend = mult * RDA2),
           arrow = arrow(length = unit(0.25, "cm")), colour = "grey") + #grid is required for arrow
to work.
  geom_text(data = type_continuous_arrows,
         aes(x= (mult + mult/5) * RDA1, y = (mult + mult/5) * RDA2,
             label = type_class),
         size = 5,
         hjust = 0.5)


ggsave(plot = RDA_plot, file = "RDA.tiff",width=10,height=7,units="in",dpi=300)
## Warning: Removed 16 rows containing missing values (geom_point).
Richness and Evenness plots
##reorder months
sample_data(data)$Month <- factor(sample_data(data)$Month, levels = c("July",
"August","September","October", "November","December"),ordered=TRUE)
```

```
levels(sample_data(data)$Month)

richness_all<- plot_richness(data,measures =rbind("Observed","Chao1",
"Shannon"),color="Month", shape="Location")
str(richness_all)
head(richness_all$data$variable)
richness<-as.data.frame(richness_all$data)

write.csv(richness,"richness_data.csv")

richness_table<-
as.data.frame(cbind("Month"=as.character(richness_all$data$Month),"Location"=as.character
(richness_all$data$Location),"Type"=as.character(richness_all$data$Type),"Index"=as.characte
r(richness_all$data$variable),"value"=as.numeric(richness_all$data$value)))
str(richness_table)
richness_table$value<-as.numeric(as.character(richness_table$value))
richness_table$Month = with(richness_table, factor(Month, levels =
c("July","August","September","October", "November","December"),ordered=TRUE))

Chao1 <- subset(richness_table, Index== "Chao1")
Shannon <- subset(richness_table, Index== "Shannon")
Observed<- subset(richness_table, Index== "Observed")

Chao1_plot<-ggplot(data=Chao1,aes(x=Month,y=value))+
  theme_grey(base_size = 20)+
  geom_point(size=6) +
  stat_smooth(method="loess",aes(group=1)) +
  facet_grid(Type~Location,scales="fixed") +
  theme(axis.text.x = element_text(angle =
45,vjust=1,hjust=1),axis.title.x=element_blank(),axis.title.y=element_blank()) +
  labs(title = "Chao1")
ggsave(file="chao1.tiff",plot=Chao1_plot,width=10,height=8,units="in",dpi=300)
Shannon_plot<-ggplot(data=Shannon,aes(x=Month,y=value))+
  theme_grey(base_size = 20)+
  geom_point(size=6) +
  stat_smooth(method="loess",aes(group=1)) +
  facet_grid(Type~Location,scales="fixed") +
  theme(axis.text.x = element_text(angle =
45,vjust=1,hjust=1),axis.title.x=element_blank(),axis.title.y=element_blank()) +
  labs(title = "Shannon")
ggsave(file="shannon.tiff",plot=Shannon_plot,width=10,height=8,units="in",dpi=300)
```

```
Observed_plot<-ggplot(data=Observed,aes(x=Month,y=value))+
  theme_grey(base_size = 20)+
  geom_point(size=6) +
  stat_smooth(method="loess",aes(group=1)) +
  facet_grid(Type~Location,scales="fixed") +
  theme(axis.text.x = element_text(angle =
45,vjust=1,hjust=1),axis.title.x=element_blank(),axis.title.y=element_blank()) +
  labs(title = "Observed")
ggsave(file="Observed_all.tiff",plot=Observed_plot,width=10,height=8,units="in",dpi=300)
OTU tracking over time graphs

###################################water
water <- subset_samples(data, Type == "Water")
shared_w<-filter_taxa(water, function(x) all(x !=0 ), TRUE)
#16

#### water by location
#fort falls
ff_w <- subset_samples(water, Location == "Surface")
w_ff<-filter_taxa(ff_w, function(x) all(x !=0 ), TRUE)
shared_w_ff<-as.data.frame(tax_table(w_ff))
shared_w_ff_otus<-as.data.frame(otu_table(w_ff))
shared_w_ff$abundance<-rowSums(shared_w_ff_otus)
shared_w_ff$type<-"Water"
shared_w_ff$loc1<-"Surface"
shared_w_ff$loc2<-"Surface"
head(shared_w_ff)

#Upstream
j_w <- subset_samples(water, Location == "Upstream")
shared_w_j<-filter_taxa(j_w, function(x) all(x !=0 ), TRUE)

#How many sequences do we have?
sum(as.data.frame(otu_table(j_w)))
#create new normalized abundance. Normalized to the total number of sequences in each
biotrap samples
j_w_trans<-transform_sample_counts(j_w, function(OTU) OTU/sum(OTU)*100)
shared_w_j<-filter_taxa(j_w_trans, function(x) all(x !=0 ), TRUE)

#for a plot, we need to melt this data to a long format
shared_w_j_otu_tax<-
cbind(as.data.frame(otu_table(shared_w_j)),as.data.frame(tax_table(shared_w_j)))
```

```
shared_w_j_otu_tax$otu<-rownames(shared_w_j_otu_tax)
head(shared_w_j_otu_tax)
shared_wj_melt<-melt(shared_w_j_otu_tax)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_wj_melt)

#use merge to add metadata to melted similarity data frame
melt_shared_wj<-merge(shared_wj_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
head(melt_shared_wj)


#drop unsused levels
melt_shared_wj<-droplevels(melt_shared_wj)
#order months by time
months<-c("July","August","September","October","November","December")
melt_shared_wj$month<-factor(melt_shared_wj$month,levels=months)

#use only the most abundant OTUs
greater_than_four_wj<-subset(melt_shared_wj,subset=value>4,select=otu)
greater_than_four_wj_percent<-subset(melt_shared_wj, otu %in% greater_than_four_wj$otu)

#Average the replicates
#no replicates here
#create percent column as duplicate of value so we can combine datasets later
greater_than_four_wj_percent$percent<-greater_than_four_wj_percent$value

#lake room
lr_w <- subset_samples(water, Location == "Downstream")
w_lr<-filter_taxa(lr_w, function(x) all(x !=0 ), TRUE)

#How many sequences do we have?
sum(as.data.frame(otu_table(lr_w)))
#create new normalized abundance. Normalized to the total number of sequences in each
biotrap samples
lr_w_trans<-transform_sample_counts(lr_w, function(OTU) OTU/sum(OTU)*100)
shared_w_lr<-filter_taxa(lr_w_trans, function(x) all(x !=0 ), TRUE)

#for a plot, we need to melt this data to a long format
shared_w_lr_otu_tax<-
cbind(as.data.frame(otu_table(shared_w_lr)),as.data.frame(tax_table(shared_w_lr)))
shared_w_lr_otu_tax$otu<-rownames(shared_w_lr_otu_tax)
```

```
head(shared_w_lr_otu_tax)
shared_wlr_melt<-melt(shared_w_lr_otu_tax)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_wlr_melt)

#use merge to add metadata to melted similarity data frame
melt_shared_wlr<-merge(shared_wlr_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
head(melt_shared_wlr$month)


#drop unsused levels
#melt_shared_wlr<-droplevels(melt_shared_wlr)
#order months by time
months<-c("July","August","September","October","November","December")
melt_shared_wlr$month<-factor(melt_shared_wlr$month,levels=months)

#use only the most abundant OTUs
greater_than_four_wlr<-subset(melt_shared_wlr,subset=value>4,select=otu)
greater_than_four_wlr_percent<-subset(melt_shared_wlr, otu %in%
greater_than_four_wlr$otu)

#Average the replicates
greater_than_four_wlr_percent2<-ddply(greater_than_four_wlr_percent, .(otu,month),
transform,  percent=mean(value),.drop=FALSE)


#Now let's put our locations together
water_greater_than_four<-
rbind(greater_than_four_wlr_percent2,greater_than_four_wj_percent)

janky_comb<-ggplot(water_greater_than_four,aes(x=month,y=percent,group=otu,color=otu))
+
  geom_line(size=2) +
  facet_wrap(~location) +
  theme_bw(base_size=12) +
  theme(axis.title.x=element_blank(),panel.grid.major=element_line(colour = "darkgrey")) +
  ylab(("Sequence abundance (%) normalized by each sample")) +

scale_color_manual(values=c("darkorchid3","dodgerblue4","darkolivegreen4","darkorange","c
yan","firebrick2","deeppink4"))

ggsave("water_comb_order.tiff",janky_comb,width=10,height=5,units="in",dpi=300)
```

```
################################## biotrap by location
#Upstream
j_b <- subset_samples(biotrap, Location == "Upstream")
#How many sequences do we have?
sum(as.data.frame(otu_table(j_b)))
#create new normalized abundance. Normalized to the total number of sequences in each
biotrap samples
j_b_trans<-transform_sample_counts(j_b, function(OTU) OTU/sum(OTU)*100)
shared_b_j<-filter_taxa(j_b_trans, function(x) all(x !=0 ), TRUE)

#for a plot, we need to melt this data to a long format
shared_b_j_otu_tax<-
cbind(as.data.frame(otu_table(shared_b_j)),as.data.frame(tax_table(shared_b_j)))
shared_b_j_otu_tax$otu<-rownames(shared_b_j_otu_tax)
head(shared_b_j_otu_tax)
shared_bj_melt<-melt(shared_b_j_otu_tax)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_bj_melt)

#use merge to add metadata to melted similarity data frame
melt_shared_bj<-merge(shared_bj_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
head(melt_shared_bj)

#drop unsused levels
melt_shared_bj<-droplevels(melt_shared_bj)
#order months by time
months<-c("September","October","November","December")
melt_shared_bj$month<-factor(melt_shared_bj$month,levels=months)

#use only the most abundant OTUs
greater_than_four_otu<-subset(melt_shared_bj,subset=value>4,select=otu)
greater_than_four_percent<-subset(melt_shared_bj, otu %in% greater_than_four_otu$otu)

#Average the replicates
greater_than_four_percent2<-ddply(greater_than_four_percent, .(otu,month), transform,
percent=mean(value),.drop=FALSE)

#lake room
lr_b <- subset_samples(biotrap, Location == "Downstream")
shared_b_lr<-filter_taxa(lr_b, function(x) all(x !=0 ), TRUE)
#How many sequences do we have?
```

```
sum(as.data.frame(otu_table(lr_b)))
#create new normalized abundance. Normalized to the total number of sequences in each
biotrap samples
lr_b_trans<-transform_sample_counts(shared_b_lr, function(OTU) OTU/sum(OTU)*100)
#for a plot, we need to melt this data to a long format
shared_b_lr_otu_tax<-
cbind(as.data.frame(otu_table(lr_b_trans)),as.data.frame(tax_table(lr_b_trans)))
shared_b_lr_otu_tax$otu<-rownames(shared_b_lr_otu_tax)
head(shared_b_lr_otu_tax)
shared_blr_melt<-melt(shared_b_lr_otu_tax)
## Using Kingdom, Phylum, Class, Order, Family, Genus, Species, Rank1, otu as id variables
head(shared_blr_melt)

#use merge to add metadata to melted similarity data frame
melt_shared_blr<-merge(shared_blr_melt,melt_vars1, by.x = 10, by.y = 0, all.x= TRUE)
head(melt_shared_blr)


#drop unsused levels
melt_shared_blr<-droplevels(melt_shared_blr)
#order months by time
months<-c("September","October","November","December")
melt_shared_blr$month<-factor(melt_shared_blr$month,levels=months)

#Pull out OTUs that are greater than 4% abundance
greater_than_four_otu_lr<-subset(melt_shared_blr,subset=value>4,select=otu)
greater_than_four_percent_lr<-subset(melt_shared_blr, otu %in%
greater_than_four_otu_lr$otu)
greater_than_four_percent2_lr<-ddply(greater_than_four_percent_lr, .(otu,month), transform,
percent=mean(value),.drop=FALSE)


#####Now let's combine them together
biotrap_shared<-rbind(greater_than_four_percent2_lr,greater_than_four_percent2)

jamky_bio<-ggplot(biotrap_shared,aes(x=month,y=percent,group=otu,color=otu)) +
 geom_line(size=2) +
 theme_bw(base_size=12) +
 facet_wrap(~location) +
 theme(axis.title.x=element_blank(),panel.grid.major=element_line(colour = "darkgrey")) +
 ylab(("Sequence abundance (%) normalized by each sample")) +
 scale_color_manual(values=c("khaki1","khaki4","lightgreen","thistle",
```

```
                    "tomato","paleturquoise4"))
ggsave("biotrap_succession_class.tiff",jamky_bio,width=10,height=5,units="in",dpi=300)
```

**Code III**

Metagenomics taxonomy

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
#################################################################
########import and summarize data

library(plyr)
library(dplyr)
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(ggplot2)
library(vegan)
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.3-0
######import all of the data
######These are data from the MG-RAST workbench.
######All samples are not possible to upload at once so I separated by sample type
######These are all annotations for each sample from the SEED taxonomy best hit
classifications
bio<-read.csv("bio.csv",header=TRUE)
#####if you want to see how the data is structured take out the "#" below
```

```
#str(bio)
unique(bio$metagenome)
## [1] 4577496 4577492 4577495 4577493 4577494 4577497 4577498
sed<-read.csv("sed.csv",header=TRUE)
sed1<-sed[,c(1,3:11)]
#####if you want to see how the data is structured take out the "#" below
#str(sed1)
unique(sed1$metagenome)
##  [1] 4585049 4585051 4585053 4585048 4585056 4577503 4577667 4585054
##  [9] 4585052 4585047
water<-read.csv("water.csv",header=TRUE)
######if you want to see how the data is structured take out the "#" below
#str(water)
unique(water$metagenome)
##  [1] 4579011 4579015 4579010 4577500 4579014 4577502 4577665 4579016
##  [9] 4579012 4579017 4579013 4577666 4577499 4577501
meta<-read.csv("metadata_w_16s.csv")
unique(meta$metagenome)
##  [1] 4577492 4577493 4577494 4577495 4577496 4577497 4577498 4577499
##  [9] 4577500 4577501 4577502 4577503 4577665 4577666 4577667 4579010
## [17] 4579011 4579012 4579013 4579014 4579015 4579016 4579017 4585047
## [25] 4585048 4585049 4585051 4585052 4585053 4585054 4585056
####combine the data sets together
seed_tax1<-rbind(bio,sed1,water)
seed_tax<-merge(seed_tax1,meta,by="metagenome")
#####if you want to see how the data is structured take out the "#" below
#str(seed_tax)
seed_tax$metagenome<-as.factor(seed_tax$metagenome)
write.csv(seed_tax,"seed_tax.csv")
sum(seed_tax$abundance)
## [1] 8383395
##now let's summarize taxonomy info with category percentages
#domain
domain1 <- seed_tax %>%
  group_by(metagenome) %>%
  mutate(countT= sum(abundance)) %>%
  group_by(domain, add=TRUE) %>%
  mutate(percent=(100*abundance/countT))
domain<-as.data.frame(domain1)
####if you need to see what we did take out the "#" below
#str(domain)
```

```
#domain plot
taxa_plot3<- ggplot(domain, aes(x=metagenome,y=percent,fill=domain)) +
  geom_bar(aes(order = desc(domain)),stat="identity")  +
  facet_grid(~Type,scales="free") +
  theme(axis.ticks = element_blank(), axis.text.x = element_blank()) +
  guides(fill=guide_legend(ncol=2))
taxa_plot3

ggsave("domain1.tiff",taxa_plot3,height=7,width=10,units="in",dpi=300)
####less than helpful plot

#phylum
phylum1 <- seed_tax %>%
  group_by(metagenome) %>%
  mutate(countT= sum(abundance)) %>%
  group_by(phylum, add=TRUE) %>%
  mutate(percent=(100*abundance/countT))
phylum<-as.data.frame(phylum1)
####if you need to see what we did take out the "#" below
#str(phylum)

#let's just get the most abundant phylum info
phylum2 <- phylum %>%
  group_by(phylum) %>%
  summarise(max=sum(abundance)/8383395*100)
phylum2<-as.data.frame(phylum2)
phylum_sorted<-phylum2[with(phylum2, order(max)), ]

#phylum plot
taxa_plot1<- ggplot(phylum, aes(x=metagenome,y=percent,fill=phylum)) +
  geom_bar(aes(order = desc(phylum)),stat="identity")  +
  facet_grid(~Type,scales="free") +
  theme(axis.ticks = element_blank(), axis.text.x = element_blank()) +
  guides(fill=guide_legend(ncol=2))
taxa_plot1

ggsave("phylum1.tiff",taxa_plot1,height=7,width=10,units="in",dpi=300)
####also less than helpful plot but there you go
###but since proteobacteria are abundant, let's take a closer look at them only

#proteos only
proteo<-subset(seed_tax,seed_tax$phylum %in% "Proteobacteria")
```

```
class1 <- proteo %>%
  group_by(metagenome) %>%
  mutate(countT= sum(abundance)) %>%
  group_by(class, add=TRUE) %>%
  mutate(percent=(100*abundance/countT))
class<-as.data.frame(class1)
#####if you want to see how the data is structured take out the "#" below
#str(class)

#proteo plot
taxa_plot2<- ggplot(class, aes(x=metagenome,y=percent,fill=class)) +
  geom_bar(aes(order = desc(class)),stat="identity")  +
  facet_grid(~Type,scales="free") +
  theme(axis.ticks = element_blank(), axis.text.x = element_blank())
taxa_plot2

ggsave("proteo1.tiff",taxa_plot2,height=6,width=6,units="in",dpi=300)
####another bar plot
Great, we have some of the taxonomic summary information. Let's do a statistically significant
test to see if the taxonomy varies by environment type.
##############################################################################
############ADONIS

library(plyr)
library(dplyr)
library(ggplot2)
library(vegan)
library(reshape2)

##using the seed_tax data set, place into wide format so we can perform other analyses
otu_ish<-dcast(seed_tax,metagenome~phylum, value.var="abundance")
## Aggregation function missing: defaulting to length
#must have row.names
row.names(otu_ish)<-otu_ish$metagenome
#########if you want to see how the data is structured take out the "#" below
#str(otu_ish)

#get rid of that first pesky column
otu<-otu_ish[,2:55]
#########if you want to see how the data is structured take out the "#" below
#str(otu)
```

```
#perform distance metric
dist<-vegdist(otu,method="bray")
#do the adonis
adonis(otu ~ Type, data=meta, permutations=99)
##
## Call:
## adonis(formula = otu ~ Type, data = meta, permutations = 99)
##
## Permutation: free
## Number of permutations: 99
##
## Terms added sequentially (first to last)
##
##          Df SumsOfSqs  MeanSqs F.Model      R2 Pr(>F)
## Type      2  0.290720 0.145360   325.1 0.95871   0.01 **
## Residuals 28  0.012519 0.000447         0.04129
## Total     30  0.303240                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Average genome size

Kathleen Brannen-Donnelly

Tuesday, August 18, 2015

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

```
#####need the following packages
library(dplyr)
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
library(reshape2)
library(ggplot2)
library(vegan)
```

```
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.3-0
library(car)


####import the average genome size data from the
####AverageCensus program from https://github.com/snayfach/MicrobeCensus
avg_sizes<-read.csv("metagenome_avg_size.csv",header=TRUE)
#####a little data manipulation follows
avg_sizes$metagenome<-as.character(avg_sizes$metagenome)
#get rid of poor quality samples
avg_sizes1<-subset(avg_sizes, !(metagenome %in% c("4585055.3","4585050.3","4585046.3")))

####import the metadata
metadata<-read.csv("metadata.csv",header=TRUE)
#####a little data manipulation follows
metadata$metagenome<-as.character(metadata$metagenome)
#get rid of poor quality samples
metadata1<-subset(metadata, !(metagenome %in% c("4585055.3","4585050.3","4585046.3")))

####merge metadata and avg genome size
avg_size_meta<-
merge(metadata1,avg_sizes1,by.x="metagenome",by.y="metagenome",all=TRUE)

##### Is the average genome size similar among types of samples?
##anova
mod1<-lm(log(avg_size_meta$avg_size)~Type+Location+Month,avg_size_meta)
Anova(mod1)
## Anova Table (Type II tests)
##
## Response: log(avg_size_meta$avg_size)
##           Sum Sq Df F value    Pr(>F)
## Type     0.90617  2 51.0448 5.44e-09 ***
## Location 0.02819  2  1.5882    0.2268
## Month    0.06859  4  1.9317    0.1408
## Residuals 0.19528 22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
######plot a histogram of avg genome size
#let's order the sample types
material.name<-c("Water","Bio-Trap®","Sediment")
```

```
#now let's just make sure our months are in actual sequential order
avg_size_meta$Type <- factor(avg_size_meta$Type, levels = material.name)
levels(avg_size_filt$Type)
## [1] "Water"     "Bio-Trap®" "Sediment"
###plot a histogram of sizes
hist<-ggplot(avg_size_meta,aes(x=log(avg_size),fill=Type)) +
 geom_histogram(stat="bin") +
 scale_fill_manual(values=c("darkolivegreen2", "skyblue3","sandybrown")) +
 xlab("log(number of basepairs)") +
 theme_bw(base_size=16)
hist
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.

ggsave("hist_genome_size.tiff",hist,width=8,height=5,units="in",dpi=300)
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
########what's the average size by type
average<- avg_size_meta %>%
 group_by(Type) %>%
 summarise(typee=log(mean(avg_size)))
```

CAZy plot

Kathleen Brannen-Donnelly

Tuesday, August 18, 2015

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

First we will load in the data, and sumamrize some key points

```
########packages required
require(Rcpp)
## Loading required package: Rcpp
require(ggplot2)
## Loading required package: ggplot2
require(reshape2)
## Loading required package: reshape2
require(plyr)
## Loading required package: plyr
library(car)
require(dplyr)
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
```

```
##
##    arrange, count, desc, failwith, id, mutate, rename, summarise,
##    summarize
##
## The following objects are masked from 'package:stats':
##
##    filter, lag
##
## The following objects are masked from 'package:base':
##
##    intersect, setdiff, setequal, union
library(vegan)
## Loading required package: permute
## Loading required package: lattice
## This is vegan 2.3-0
########read in data
data<-read.csv("subsystems1.csv",header=TRUE)
str(data)
## 'data.frame':    236755 obs. of  10 variables:
##  $ metagenome    : num  4579015 4585052 4585056 4579016 4585055 ...
##  $ level.1       : Factor w/ 28 levels "Amino Acids and Derivatives",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ level.2       : Factor w/ 168 levels "-","ABC transporters",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ level.3       : Factor w/ 1134 levels "(GlcNAc)2 Catabolic Operon",..: 354 354 354 354 354
354 354 354 354 354 ...
##  $ function.level: Factor w/ 7756 levels "(3R)-hydroxymyristoyl-[ACP] dehydratase (EC 4.2.1.-
)",..: 1604 1604 1604 1604 1604 1604 1604 1604 1604 1604 ...
##  $ abundance     : int  1 6 6 3 4 2 6 13 13 5 ...
##  $ avg.eValue    : num  -18 -19 -23.4 -29.7 -21.7 ...
##  $ avg...ident   : num  61.1 76.5 82.3 83.2 75.9 ...
##  $ avg.align.len : num  72 57.2 61.6 75 59.3 ...
##  $ X..hits       : int  1 5 4 3 3 2 4 5 7 4 ...
head(data)
##   metagenome                level.1 level.2
## 1    4579015 Amino Acids and Derivatives       -
## 2    4585052 Amino Acids and Derivatives       -
## 3    4585056 Amino Acids and Derivatives       -
## 4    4579016 Amino Acids and Derivatives       -
## 5    4585055 Amino Acids and Derivatives       -
## 6    4579013 Amino Acids and Derivatives       -
##                          level.3         function.level abundance
## 1 Creatine and Creatinine Degradation Creatinase (EC 3.5.3.3)         1
## 2 Creatine and Creatinine Degradation Creatinase (EC 3.5.3.3)         6
```

```
## 3 Creatine and Creatinine Degradation Creatinase (EC 3.5.3.3)       6
## 4 Creatine and Creatinine Degradation Creatinase (EC 3.5.3.3)       3
## 5 Creatine and Creatinine Degradation Creatinase (EC 3.5.3.3)       4
## 6 Creatine and Creatinine Degradation Creatinase (EC 3.5.3.3)       2
##   avg.eValue avg...ident avg.align.len X..hits
## 1    -18.00      61.11       72.00      1
## 2    -19.00      76.47       57.20      5
## 3    -23.40      82.33       61.60      4
## 4    -29.67      83.22       75.00      3
## 5    -21.67      75.94       59.33      3
## 6    -23.50      86.30       58.50      2
unique(data$metagenome)
##  [1] 4579015 4585052 4585056 4579016 4585055 4579013 4585051 4585048
##  [9] 4585054 4585053 4579010 4579017 4585047 4585049 4577667 4579014
## [17] 4579012 4579011 4585046 4585050 4577496 4577492 4577502 4577495
## [25] 4577494 4577665 4577503 4577497 4577666 4577493 4577499 4577498
## [33] 4577500 4577501
# data is already in long form, thanks mgrast
#####read in metadata
metadata<-read.csv("metadata_w_16s2.csv",header=TRUE)
head(metadata)
##   metagenome  Location     Month      Type sample.name  reads X16s.copies
## 1    4577492 Lake Room  December Bio-Trap®   CCRB-13d1 452810       1398
## 2    4577493     Jones  December Bio-Trap®   CCRB-13d2 378491        763
## 3    4577494 Lake Room  November Bio-Trap®   CCRB-13N1 449606       1319
## 4    4577495     Jones   October Bio-Trap®   CCRB-13N2 748024       2897
## 5    4577496 Lake Room   October Bio-Trap®   CCRB-13O1 901205       6141
## 6    4577497 Lake Room September Bio-Trap®   CCRB-13S2 270454        902
##   Estimate.of.cells
## 1          332.8571
## 2          181.6667
## 3          314.0476
## 4          689.7619
## 5         1462.1429
## 6          214.7619
unique(metadata$metagenome)
##  [1] 4577492 4577493 4577494 4577495 4577496 4577497 4577498 4577499
##  [9] 4577500 4577501 4577502 4577503 4577665 4577666 4577667 4579010
## [17] 4579011 4579012 4579013 4579014 4579015 4579016 4579017 4585046
## [25] 4585047 4585048 4585049 4585050 4585051 4585052 4585053 4585054
## [33] 4585055 4585056
```

```
# Add metadata for samples to data
#use merge to add metadata to melted similarity data frame
data_merge<-merge(data, metadata, by.x = 1, by.y = 1, all.x= TRUE)
str(data_merge)
## 'data.frame':    236755 obs. of  17 variables:
##  $ metagenome      : num  4577492 4577492 4577492 4577492 4577492 ...
##  $ level.1         : Factor w/ 28 levels "Amino Acids and Derivatives",..: 22 11 5 2 9 15 23 2 7 15
...
##  $ level.2         : Factor w/ 168 levels "-","ABC transporters",..: 1 168 1 30 54 1 1 154 47 1 ...
##  $ level.3         : Factor w/ 1134 levels "(GlcNAc)2 Catabolic Operon",..: 1132 666 214 755 449
381 835 785 407 47 ...
##  $ function.level   : Factor w/ 7756 levels "(3R)-hydroxymyristoyl-[ACP] dehydratase (EC
4.2.1.-)",..: 5106 4374 2415 689 241 4591 6100 5018 863 2961 ...
##  $ abundance       : int  36 64 56 102 1 19 162 12 7 6 ...
##  $ avg.eValue      : num  -18.2 -13.6 -18.2 -16.8 -13 ...
##  $ avg...ident     : num  75.4 70 71.3 72.4 82.6 ...
##  $ avg.align.len    : num  55.9 54 62.1 57.7 46 ...
##  $ X..hits         : int  29 34 40 61 1 8 101 12 7 6 ...
##  $ Location        : Factor w/ 3 levels "Fort Falls","Jones",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Month           : Factor w/ 5 levels "August","December",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Type            : Factor w/ 3 levels "Bio-Trap®","Sediment",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sample.name     : Factor w/ 34 levels "CCRB-13d1","CCRB-13d2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ reads           : int  452810 452810 452810 452810 452810 452810 452810 452810 452810
452810 ...
##  $ X16s.copies     : int  1398 1398 1398 1398 1398 1398 1398 1398 1398 1398 ...
##  $ Estimate.of.cells: num  333 333 333 333 333 ...
data_merge$metagenome<-as.factor(data_merge$metagenome)

#ok, double checked that by hand in the .csv and the metadata matches with the correct
metagenome

#####we are getting rid of 3 poor quality samples
data_merge1<-subset(data_merge, !(metagenome %in%
c("4585055.3","4585050.3","4585046.3")))
unique(data_merge1$metagenome)
##  [1] 4577492.3 4577493.3 4577494.3 4577495.3 4577496.3 4577497.3 4577498.3
##  [8] 4577499.3 4577500.3 4577501.3 4577502.3 4577503.3 4577665.3 4577666.3
## [15] 4577667.3 4579010.3 4579011.3 4579012.3 4579013.3 4579014.3 4579015.3
## [22] 4579016.3 4579017.3 4585047.3 4585048.3 4585049.3 4585051.3 4585052.3
## [29] 4585053.3 4585054.3 4585056.3
## 34 Levels: 4577492.3 4577493.3 4577494.3 4577495.3 4577496.3 ... 4585056.3
```

```
###now we need to make a new column with the proportion of sequences
#make metagenome a factor
data_merge1$metagenome<-as.factor(data_merge1$metagenome)
str(data_merge1)
## 'data.frame':    224061 obs. of  17 variables:
##  $ metagenome      : Factor w/ 34 levels "4577492.3","4577493.3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ level.1         : Factor w/ 28 levels "Amino Acids and Derivatives",..: 22 11 5 2 9 15 23 2 7 15
...
##  $ level.2         : Factor w/ 168 levels "-","ABC transporters",..: 1 168 1 30 54 1 1 154 47 1 ...
##  $ level.3         : Factor w/ 1134 levels "(GlcNAc)2 Catabolic Operon",..: 1132 666 214 755 449
381 835 785 407 47 ...
##  $ function.level   : Factor w/ 7756 levels "(3R)-hydroxymyristoyl-[ACP] dehydratase (EC
4.2.1.-)",..: 5106 4374 2415 689 241 4591 6100 5018 863 2961 ...
##  $ abundance       : int  36 64 56 102 1 19 162 12 7 6 ...
##  $ avg.eValue      : num  -18.2 -13.6 -18.2 -16.8 -13 ...
##  $ avg...ident     : num  75.4 70 71.3 72.4 82.6 ...
##  $ avg.align.len    : num  55.9 54 62.1 57.7 46 ...
##  $ X..hits         : int  29 34 40 61 1 8 101 12 7 6 ...
##  $ Location        : Factor w/ 3 levels "Fort Falls","Jones",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Month           : Factor w/ 5 levels "August","December",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Type            : Factor w/ 3 levels "Bio-Trap®","Sediment",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sample.name     : Factor w/ 34 levels "CCRB-13d1","CCRB-13d2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ reads           : int  452810 452810 452810 452810 452810 452810 452810 452810 452810
452810 ...
##  $ X16s.copies     : int  1398 1398 1398 1398 1398 1398 1398 1398 1398 1398 ...
##  $ Estimate.of.cells: num  333 333 333 333 333 ...
data_merge2<-ddply(data_merge1, c("metagenome"), transform,
            frac.abund = (abundance / sum(abundance))*100,
            norm.copies = (abundance / Estimate.of.cells),
            abund.16s = (abundance / X16s.copies))
head(data_merge2)
##   metagenome                    level.1
## 1  4577492.3       Regulation and Cell signaling
## 2  4577492.3              Membrane Transport
## 3  4577492.3        Clustering-based subsystems
## 4  4577492.3                  Carbohydrates
## 5  4577492.3 Fatty Acids, Lipids, and Isoprenoids
## 6  4577492.3              Nitrogen Metabolism
##                level.2
## 1                    -
## 2 Uni- Sym- and Antiporters
## 3                    -
```

```
## 4           CO2 fixation
## 5            Fatty acids
## 6                      -
##                                level.3
## 1                    Zinc regulated enzymes
## 2 NhaA, NhaD and Sodium-dependent phosphate transporters
## 3                     CBSS-281090.3.peg.464
## 4           Photorespiration (oxidative C2 cycle)
## 5                Fatty Acid Biosynthesis FASI
## 6                     Denitrification
##                                                    function.level
## 1                          Phosphoribosyl-AMP cyclohydrolase (EC 3.5.4.19)
## 2                                  Na+/H+ antiporter NhaA type
## 3                                  FIG004453: protein YceG like
## 4              Aminomethyltransferase (glycine cleavage system T protein) (EC 2.1.2.10)
## 5                       3-oxoacyl-coenzyme A reductase of elongase (EC 1.1.1.62)
## 6 NnrU family protein, required for expression of nitric oxide and nitrite reductases (Nir and
Nor)
##   abundance avg.eValue avg...ident avg.align.len X..hits  Location
## 1       36    -18.19      75.41       55.92     29 Lake Room
## 2       64    -13.65      70.04       54.05     34 Lake Room
## 3       56    -18.20      71.27       62.09     40 Lake Room
## 4      102     -16.75      72.39       57.74     61 Lake Room
## 5        1    -13.00      82.61       46.00      1 Lake Room
## 6       19    -15.50      70.41       55.00      8 Lake Room
##     Month      Type sample.name  reads X16s.copies Estimate.of.cells
## 1 December Bio-Trap®   CCRB-13d1 452810       1398        332.8571
## 2 December Bio-Trap®   CCRB-13d1 452810       1398        332.8571
## 3 December Bio-Trap®   CCRB-13d1 452810       1398        332.8571
## 4 December Bio-Trap®   CCRB-13d1 452810       1398        332.8571
## 5 December Bio-Trap®   CCRB-13d1 452810       1398        332.8571
## 6 December Bio-Trap®   CCRB-13d1 452810       1398        332.8571
##    frac.abund norm.copies    abund.16s
## 1 0.0107705748 0.108154506 0.0257510730
## 2 0.0191476885 0.192274678 0.0457796853
## 3 0.0167542275 0.168240343 0.0400572246
## 4 0.0305166286 0.306437768 0.0729613734
## 5 0.0002991826 0.003004292 0.0007153076
## 6 0.0056844700 0.057081545 0.0135908441
str(data_merge2)
## 'data.frame':    224061 obs. of  20 variables:
##  $ metagenome      : Factor w/ 34 levels "4577492.3","4577493.3",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ level.1        : Factor w/ 28 levels "Amino Acids and Derivatives",..: 22 11 5 2 9 15 23 2 7 15
...
##  $ level.2        : Factor w/ 168 levels "-","ABC transporters",..: 1 168 1 30 54 1 1 154 47 1 ...
##  $ level.3        : Factor w/ 1134 levels "(GlcNAc)2 Catabolic Operon",..: 1132 666 214 755 449
381 835 785 407 47 ...
##  $ function.level   : Factor w/ 7756 levels "(3R)-hydroxymyristoyl-[ACP] dehydratase (EC
4.2.1.-)",..: 5106 4374 2415 689 241 4591 6100 5018 863 2961 ...
##  $ abundance       : int  36 64 56 102 1 19 162 12 7 6 ...
##  $ avg.eValue      : num  -18.2 -13.6 -18.2 -16.8 -13 ...
##  $ avg...ident     : num  75.4 70 71.3 72.4 82.6 ...
##  $ avg.align.len   : num  55.9 54 62.1 57.7 46 ...
##  $ X..hits         : int  29 34 40 61 1 8 101 12 7 6 ...
##  $ Location        : Factor w/ 3 levels "Fort Falls","Jones",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Month           : Factor w/ 5 levels "August","December",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Type            : Factor w/ 3 levels "Bio-Trap®","Sediment",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sample.name     : Factor w/ 34 levels "CCRB-13d1","CCRB-13d2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ reads           : int  452810 452810 452810 452810 452810 452810 452810 452810 452810
452810 ...
##  $ X16s.copies     : int  1398 1398 1398 1398 1398 1398 1398 1398 1398 1398 ...
##  $ Estimate.of.cells: num  333 333 333 333 333 ...
##  $ frac.abund      : num  0.010771 0.019148 0.016754 0.030517 0.000299 ...
##  $ norm.copies     : num  0.108 0.192 0.168 0.306 0.003 ...
##  $ abund.16s       : num  0.025751 0.04578 0.040057 0.072961 0.000715 ...
write.csv(data_merge2,"subsystems_data_merged.csv")

###what is the max metabolism category?
##we need to sum them up by their level.1 to get max
met<- data_merge2 %>%
  group_by(level.1) %>%
  summarise(category.sums=sum(abundance),perc.abund=sum(abundance)/8422107*100)

###what is the max metabolism category by type of sample?
##we need to sum them up by their level.1 to get max
met2<- data_merge2 %>%
  group_by(level.1,Type) %>%
  summarise(category.sums=sum(abundance),perc.abund=sum(abundance)/8422107*100)
met_sort<- with(met2, order(Type, perc.abund))
met_sorted<-met2[met_sort,]


##Now we will look for three CAZy enzyme classifications
```

```
##this uses data object data_merge2 from subsystems
#Let's seperate out the genes we want the glycoside hydrolases from www.cazy.org
#use"\\s" to denote that there will be a space before the ec number; can't be "S"
#we want the space so that the function doesn't find that combo with another numebr in front
bc that would be wrong
glycoside_hydrolases1<-filter(data_merge2, grepl("\\s3.2.1", function.level))
glycoside_hydrolases2<-filter(data_merge2, grepl("\\s2.4.1", function.level))

#now bind these two together
glyc_hydro<-rbind(glycoside_hydrolases1,glycoside_hydrolases2)
str(glyc_hydro)
## 'data.frame':    5105 obs. of  20 variables:
##  $ metagenome      : Factor w/ 34 levels "4577492.3","4577493.3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ level.1         : Factor w/ 28 levels "Amino Acids and Derivatives",..: 2 2 2 2 2 2 2 2 2 4 ...
##  $ level.2         : Factor w/ 168 levels "-","ABC transporters",..: 1 83 43 43 1 43 43 1 83 1 ...
##  $ level.3         : Factor w/ 1134 levels "(GlcNAc)2 Catabolic Operon",..: 969 568 470 470 969
129 470 969 608 650 ...
##  $ function.level   : Factor w/ 7756 levels "(3R)-hydroxymyristoyl-[ACP] dehydratase (EC
4.2.1.-)",..: 1017 653 640 6770 656 1017 656 1024 656 4071 ...
##  $ abundance       : int  51 16 32 17 5 51 5 23 5 4 ...
##  $ avg.eValue      : num  -15.5 -19.5 -22.8 -17 -10 ...
##  $ avg...ident     : num  68.4 68.6 76.6 69.9 61.3 ...
##  $ avg.align.len   : num  57 60.2 65.4 55.2 49.8 ...
##  $ X..hits         : int  37 13 17 9 5 37 5 23 5 4 ...
##  $ Location        : Factor w/ 3 levels "Fort Falls","Jones",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Month           : Factor w/ 5 levels "August","December",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Type            : Factor w/ 3 levels "Bio-Trap®","Sediment",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sample.name     : Factor w/ 34 levels "CCRB-13d1","CCRB-13d2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ reads           : int  452810 452810 452810 452810 452810 452810 452810 452810 452810
452810 ...
##  $ X16s.copies     : int  1398 1398 1398 1398 1398 1398 1398 1398 1398 1398 ...
##  $ Estimate.of.cells: num  333 333 333 333 333 ...
##  $ frac.abund      : num  0.01526 0.00479 0.00957 0.00509 0.0015 ...
##  $ norm.copies     : num  0.1532 0.0481 0.0961 0.0511 0.015 ...
##  $ abund.16s       : num  0.03648 0.01144 0.02289 0.01216 0.00358 ...
glyc_hydro$enz.type<-"glycoside hydrolases"

#plot it with each locatin in a grid
glyc_hydro_plot<-ggplot(glyc_hydro,aes(x=Month,y=abund.16s,fill=Type)) +
 theme_bw(base_size=16) +
 geom_bar(stat="identity") +
 scale_fill_manual(values=c("darkolivegreen2", "sandybrown","skyblue3")) +
```

```
  facet_grid(Location~Type) +
  ylab("functional gene abundance/16s gene abundance") +
  theme(axis.text.x = element_text(angle = 45,vjust=1,hjust=1),axis.title.x=element_blank()) +
  theme(panel.grid.major = element_line(colour = "grey70"),
        panel.grid.minor = element_blank()) +
  ggtitle("SEED Subsystems glycoside hydrolases") +
  theme(plot.title = element_text(lineheight=.8, face="bold"))
ggsave("endo_gluc_plot_grid.tiff",endo_gluc_plot,width=7,height=6,units="in",dpi=300)
```

*#Polysaccharide Lyases*

*#Let's seperate out the genes we want the Polysaccharide Lyases from www.cazy.org*
*#use"\\s" to denote that there will be a space before the ec number; can't be "S"*
*#we want the space so that the function doesn't find that combo with another numebr in front bc that would be wrong*

```
polysaccharide_lyases<-filter(data_merge2, grepl("\\s4.2.2", function.level))
str(polysaccharide_lyases)
## 'data.frame':    107 obs. of  20 variables:
##  $ metagenome      : Factor w/ 34 levels "4577492.3","4577493.3",..: 1 1 1 1 1 2 2 2 3 3 ...
##  $ level.1         : Factor w/ 28 levels "Amino Acids and Derivatives",..: 2 4 4 2 2 4 4 2 2 2 ...
##  $ level.2         : Factor w/ 168 levels "-","ABC transporters",..: 83 19 19 1 83 19 19 83 1 83 ...
##  $ level.3         : Factor w/ 1134 levels "(GlcNAc)2 Catabolic Operon",..: 371 42 42 969 371 42
42 371 969 371 ...
##  $ function.level   : Factor w/ 7756 levels "(3R)-hydroxymyristoyl-[ACP] dehydratase (EC
4.2.1.-)",..: 4793 593 5205 4793 4654 593 5205 4654 4793 4793 ...
##  $ abundance       : int  1 2 1 1 4 1 1 3 3 3 ...
##  $ avg.eValue      : num  -7 -11 -11 -7 -11.8 ...
##  $ avg...ident     : num  64.9 62.7 60.8 64.9 74 ...
##  $ avg.align.len   : num  37 52.5 51 37 45 ...
##  $ X..hits         : int  1 2 1 1 4 1 1 1 3 3 ...
##  $ Location        : Factor w/ 3 levels "Fort Falls","Jones",..: 3 3 3 3 3 2 2 2 3 3 ...
##  $ Month           : Factor w/ 5 levels "August","December",..: 2 2 2 2 2 2 2 2 3 3 ...
##  $ Type            : Factor w/ 3 levels "Bio-Trap®","Sediment",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sample.name     : Factor w/ 34 levels "CCRB-13d1","CCRB-13d2",..: 1 1 1 1 1 2 2 2 3 3 ...
##  $ reads           : int  452810 452810 452810 452810 452810 378491 378491 378491 449606
449606 ...
##  $ X16s.copies     : int  1398 1398 1398 1398 1398 763 763 763 1319 1319 ...
##  $ Estimate.of.cells: num  333 333 333 333 333 ...
##  $ frac.abund      : num  0.000299 0.000598 0.000299 0.000299 0.001197 ...
```

```
##  $ norm.copies     : num  0.003 0.00601 0.003 0.003 0.01202 ...
##  $ abund.16s       : num  0.000715 0.001431 0.000715 0.000715 0.002861 ...
polysaccharide_lyases$enz.type<-"polysaccharide lyases"

#plot it with each locatin in a grid
polysac_lyase_plot<-ggplot(polysaccharide_lyases,aes(x=Month,y=abund.16s,fill=Type)) +
  theme_bw(base_size=16) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=c("darkolivegreen2", "sandybrown","skyblue3")) +
  facet_grid(Location~Type) +
  ylab("functional gene abundance/16s gene abundance") +
  theme(axis.text.x = element_text(angle = 45,vjust=1,hjust=1),axis.title.x=element_blank()) +
  theme(panel.grid.major = element_line(colour = "grey70"),
        panel.grid.minor = element_blank()) +
  ggtitle("SEED Subsystems polysaccharide lyases") +
  theme(plot.title = element_text(lineheight=.8, face="bold"))
polysac_lyase_plot

ggsave("polysac_lyase_plot.tiff",polysac_lyase_plot,width=7,height=6,units="in",dpi=300)


#Auxiliary Activities family classification
#AKA peroxidases
#assigned by Cazy.org

#There are a few here, we will have to go thru one at a time.
#use"\\s" to denote that there will be a space before the ec number; can't be "S"
#we want the space so that the function doesn't find that combo with another numebr in front
bc that would be wrong

all_auxes<-c("\\s1.1.1",  "\\s1.1.3.",  "\\s1.1.3.4",  "\\s1.1.3.7",    "\\s1.1.3.9",   "\\s1.1.3.10",
         "\\s1.1.3.13", "\\s1.1.3.16",  "\\s1.1.3.38",  "\\s1.1.99.18", "\\s1.1.99.29", "\\s1.3.3.5",
         "\\s1.6.5.6",  "\\s1.10.3.",   "\\s1.10.3.2",  "\\s1.11.1.",   "\\s1.11.1.5",  "\\s1.11.1.11",
         "\\s1.11.1.13",    "\\s1.11.1.14", "\\s1.11.1.16", "\\s3.2.1.4",   "\\s3.2.1.78")

#Let's seperate out the genes we want the Polysaccharide Lyases from www.cazy.org
aux1<-filter(data_merge2, grepl(all_auxes[1], function.level))
aux2<-filter(data_merge2, grepl(all_auxes[2], function.level))
aux3<-filter(data_merge2, grepl(all_auxes[3], function.level))
aux4<-filter(data_merge2, grepl(all_auxes[4], function.level))
aux5<-filter(data_merge2, grepl(all_auxes[5], function.level))
aux6<-filter(data_merge2, grepl(all_auxes[6], function.level))
```

```
aux7<-filter(data_merge2, grepl(all_auxes[7], function.level))
aux8<-filter(data_merge2, grepl(all_auxes[8], function.level))
aux9<-filter(data_merge2, grepl(all_auxes[9], function.level))
aux10<-filter(data_merge2, grepl(all_auxes[10], function.level))
aux11<-filter(data_merge2, grepl(all_auxes[11], function.level))
aux12<-filter(data_merge2, grepl(all_auxes[12], function.level))
aux13<-filter(data_merge2, grepl(all_auxes[13], function.level))
aux14<-filter(data_merge2, grepl(all_auxes[14], function.level))
aux15<-filter(data_merge2, grepl(all_auxes[15], function.level))
aux16<-filter(data_merge2, grepl(all_auxes[16], function.level))
aux17<-filter(data_merge2, grepl(all_auxes[17], function.level))
aux18<-filter(data_merge2, grepl(all_auxes[18], function.level))
aux19<-filter(data_merge2, grepl(all_auxes[19], function.level))
aux20<-filter(data_merge2, grepl(all_auxes[20], function.level))
aux21<-filter(data_merge2, grepl(all_auxes[21], function.level))
aux22<-filter(data_merge2, grepl(all_auxes[22], function.level))
aux23<-filter(data_merge2, grepl(all_auxes[23], function.level))


all_auxes2<-
rbind(aux1,aux2,aux3,aux4,aux5,aux6,aux7,aux8,aux9,aux10,aux11,aux13,aux14,aux15,aux16,
            aux17,aux18,aux19,aux20,aux21,aux22,aux23)

all_auxes2$enz.type<-"ligninolytic enzymes/lytic polysaccharide mono-oxygenases"



#plot it with each locatin in a grid
auxes_plot<-ggplot(all_auxes2,aes(x=Month,y=abund.16s,fill=Type)) +
 theme_bw(base_size=16) +
 geom_bar(stat="identity") +
 scale_fill_manual(values=c("darkolivegreen2", "sandybrown","skyblue3")) +
 facet_grid(Location~Type) +
 ylab("functional gene abundance/16s gene abundance") +
 theme(axis.text.x = element_text(angle = 45,vjust=1,hjust=1),axis.title.x=element_blank()) +
 theme(panel.grid.major = element_line(colour = "grey70"),
     panel.grid.minor = element_blank()) +
 ggtitle("SEED Subsystems peroxidases") +
 theme(plot.title = element_text(lineheight=.8, face="bold"))

ggsave("auxes_plot.tiff",auxes_plot,width=7,height=6,units="in",dpi=300)
```

```
############let's see if we can combine them all.

enzymes<-rbind(glyc_hydro,polysaccharide_lyases,all_auxes2)
head(enzymes)
enzymes<-droplevels(enzymes)

##we need to sum them up by their enz. type for each category we want to bin
enzymes2<- enzymes %>%
 group_by(Location,Month,Type,enz.type) %>%
 summarise(enz.totals=sum(abund.16s))

#let's order the sample types
enz.name<-c("glycoside hydrolases","polysaccharide lyases","ligninolytic enzymes/lytic
polysaccharide mono-oxygenases")

#now let's just make sure our months are in actual sequential order
enzymes2$enz.type <- factor(enzymes2$enz.type , levels = enz.name)
levels(enzymes2$Location) <-c("Surface","Upstream","Downstream")

enz_plot<-ggplot(enzymes2,aes(x=Month,y=enz.totals,color=enz.type)) +
 theme_bw(base_size=16) +
 geom_point(size=4) +
 scale_color_manual(values=c("gray", "black","indianred1")) +
 facet_grid(Location~Type) +
 ylab("functional gene abundance/16s gene abundance") +
 theme(axis.text.x = element_text(angle = 45,vjust=1,hjust=1),axis.title.x=element_blank()) +
 theme(panel.grid.major = element_line(colour = "grey70"),
       panel.grid.minor = element_blank()) +
 ggtitle("Subsystems") +
 theme(plot.title = element_text(lineheight=.8, face="bold"))
enz_plot

ggsave("enz_plot.tiff",enz_plot,width=11,height=6,units="in",dpi=300)

##Now we will do some statistics
require(Rcpp)
require(ggplot2)
require(reshape2)
require(plyr)
library(car)
require(dplyr)
```

```
library(vegan)

##this uses data object data_merge from subsystems.R
##this also uses the object enzymes2 from subsystems_cazy_plot
enzy<-as.data.frame(enzymes2)
str(enzy)
## 'data.frame':    93 obs. of  5 variables:
##  $ Location  : Factor w/ 3 levels "Surface","Upstream",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Month     : Factor w/ 5 levels "August","December",..: 1 1 1 1 1 1 2 2 2 3 ...
##  $ Type      : Factor w/ 3 levels "Bio-Trap®","Sediment",..: 2 2 2 3 3 3 3 3 3 2 ...
##  $ enz.type  : Factor w/ 3 levels "glycoside hydrolases",..: 1 3 2 1 3 2 1 3 2 1 ...
##  $ enz.totals: num  10.193 23.6355 0.0136 0.7224 1.3145 ...
####Are any of the enzy classes significiantly different from each other?
mod<-lm(enzy$enz.totals~enzy$Type)
anova1<-Anova(mod)
anova1
## Anova Table (Type II tests)
##
## Response: enzy$enz.totals
##           Sum Sq Df F value    Pr(>F)
## enzy$Type 2055.6  2  14.629 3.154e-06 ***
## Residuals 6323.2 90
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
###Are any of the enzy classes significantly diff by typ, location, or month?
glyc<-filter(enzy, grepl("glycoside hydrolases",enz.type))
mod2<-lm(glyc$enz.totals~glyc$Type+glyc$Location+glyc$Month)
anova2<-Anova(mod2)
anova2
## Anova Table (Type II tests)
##
## Response: glyc$enz.totals
##              Sum Sq Df F value    Pr(>F)
## glyc$Type     495.62  2 30.0751 5.081e-07 ***
## glyc$Location   3.43  2  0.2079    0.8139
## glyc$Month     24.85  4  0.7540    0.5661
## Residuals     181.27 22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
poly<-filter(enzy, grepl("polysaccharide lyases",enz.type))
mod3<-lm(poly$enz.totals~poly$Type+poly$Location+poly$Month)
```

```
anova3<-Anova(mod3)
anova3
## Anova Table (Type II tests)
##
## Response: poly$enz.totals
##                Sum Sq Df F value   Pr(>F)
## poly$Type     0.00093005  2  9.0636 0.001345 **
## poly$Location 0.00003438  2  0.3351 0.718867
## poly$Month    0.00009778  4  0.4764 0.752600
## Residuals     0.00112875 22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
lig<-filter(enzy, grepl("ligninolytic enzymes/lytic polysaccharide mono-oxygenases",enz.type))
mod4<-lm(lig$enz.totals~lig$Type+lig$Location+lig$Month)
anova4<-Anova(mod4)
anova4
## Anova Table (Type II tests)
##
## Response: lig$enz.totals
##             Sum Sq Df F value    Pr(>F)
## lig$Type     2968.74  2 30.2914 4.796e-07 ***
## lig$Location   33.85  2  0.3454    0.7117
## lig$Month     215.84  4  1.1012    0.3808
## Residuals    1078.06 22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##########################################################DESeq2

library(ggplot2)
library(DESeq2) #this is the latest version of DESeq
library(plyr) #the next 3 packages of for manipulating data frames structures
library(reshape2)
library(tidyr)
set.seed("123")

####Subsystems Function level analysis ####
#any(data_deseq_countdata<0) #double check if the conversion from data frame to matrix
worked
#properly
```

```
data_deseq_countdata2<-
data.frame(function.level=factor(data_merge1$function.level),abundance=as.numeric(data_m
erge1$abundance),
                        sample.name=factor(data_merge1$sample.name))
head(data_deseq_countdata2)
str(data_deseq_countdata2)
###place into wide format
data_deseq_countdata2_wide<-dcast(data_deseq_countdata2, function.level~sample.name,
value.var="abundance",fun.aggregate = sum)
head(data_deseq_countdata2_wide)
str(data_deseq_countdata2_wide)

#now get the data in the proper format and order
rownames(data_deseq_countdata2_wide)<-data_deseq_countdata2_wide[,1]
data_deseq_countdata2_wide$function.level<-NULL
data_deseq_countdata2_wide<-
data_deseq_countdata2_wide[,order(names(data_deseq_countdata2_wide))]
head(data_deseq_countdata2_wide)

#read in metadata about the samples.
metadata<-read.csv("metadata_w_16s2.csv",header=TRUE)
head(metadata)
metadata2<-subset(metadata, !(metagenome %in% c("4585055.3","4585050.3","4585046.3")))
data_deseq_metadata2<-
data.frame(sample.name=factor(metadata2$sample.name),location=factor(metadata2$Locatio
n),
                      type=factor(metadata2$Type),month=factor(metadata2$Month))
str(data_deseq_metadata2)
rownames(data_deseq_metadata2)<-data_deseq_metadata2[,1]
data_deseq_metadata2<-
data_deseq_metadata2[order(data_deseq_metadata2$sample.name),]


################Let's compare water vs biotrap
wat_bio<-data_deseq_countdata2_wide[, -grep("CCRS-",
colnames(data_deseq_countdata2_wide))]
str(wat_bio)

wat_bio_meta<-data_deseq_metadata2[-grep("CCRS-",
data_deseq_metadata2$sample.name),]
str(wat_bio_meta)
wat_bio_meta<-droplevels(wat_bio_meta)
```

```
#start deseq, design is using metadata criteria.  in this case we are  using substrate.
#When doing the DESeq analysis we have specified a Wald test.
#The alternative is LRT and is better for more than two class problems.
dds_meta1<-DESeqDataSetFromMatrix(countData=wat_bio,
                        colData= wat_bio_meta,design= ~type)
dds_meta1<-DESeq(dds_meta1,test="Wald",fitType = "local")
head(dds_meta1)
#convert deseq object into results
res_meta1<-results(dds_meta1,cooksCutoff = TRUE)
#Cookscutoff is not accurate without 3 reps
#summary(res_meta)
plotMA(res_meta1, main="DESeq2",ylim=c(-5,5))
#plot showing abundance vs log2fold change, could be supp fig.
#this shows that the functions with low expression do not have high log fold change.
#this is the intention of the deseq2 package to avoid overexaggerated log fold changes for
#low expressed functions

#filter the results by alpha value or other criteria
alpha=0.05
sigtab_meta1 = res_meta1[which(res_meta1$padj < alpha), ]
#sigtab_meta = res_meta[which(res_meta$padj < alpha & res_meta$baseMean >1000), ]
#we could filter results by both alpha value and the normalized average.

mcols(sigtab_meta1)$description
#this gives you the description about each column in the results table.
#"water vs. biotrap" means functions with positive log2 fold change are more
#abundant in the water.
sigtab_meta1<-sigtab_meta1[order(sigtab_meta1$log2FoldChange),]

#put the results in order of log2fold change

#convert results into a regular data frame for ggplot2 to use
sigtab_meta_dataframe1<-as(sigtab_meta1, "data.frame")
head(sigtab_meta_dataframe1)
#make the function as factor for plotting
sigtab_meta_dataframe1$function.<-rownames(sigtab_meta_dataframe1)
sigtab_meta_dataframe1$function.<-as.factor(sigtab_meta_dataframe1$function.)

#this step can take a long time, maybe even crash R, if you don't subset the
#columns you really need from the annotation hierarch
data_genes1<-data_gene_hier[,c("level1","function.")]
done1<-merge(sigtab_meta_dataframe1, data_genes1, by="function.")
```

```
#add the annotation hierarch to the results data frame

# plot results
ggplot(done1, aes(level1,log2FoldChange))+geom_point(aes(size=baseMean)) +
  theme(axis.text.x=element_text(angle=-90, size=8, colour="black")) +
  geom_hline(yintercept=0)

###a simpler looking plot.
#to calculate the average and std dev of the log2 fold change and basemean per level1
done_summary1<-aggregate(log2FoldChange~level1, data=done1, mean)
done_summary2<-aggregate(log2FoldChange~level1, data=done1, sd)
done_summary3<-aggregate(baseMean~level1, data=done1, mean)
done_sum2<-merge(done_summary1,done_summary2, by="level1")
done_sum<-merge(done_sum2,done_summary3, by="level1")
#done_sum has all the needed info for plotting
#done_sum<-done_sum[order(done_sum$log2FoldChange.x),]
#make the order by log2foldchange
#plot average and std dev.
limits<- aes(ymax = log2FoldChange.x + log2FoldChange.y, ymin=log2FoldChange.x -
log2FoldChange.y)
#to put the error bars
bubbleplot_siglevel1=ggplot(done_sum, aes(level1,log2FoldChange.x))+
  geom_point(aes(size=baseMean))+
  theme(axis.text.x=element_text(angle=90, size=10, colour="black"))+
  geom_hline(yintercept=0)+
  geom_errorbar(limits, width=0.2, linetype=5)+
  scale_y_continuous(name="log2 Fold Change",breaks=seq(-6,6,1))
bubbleplot_siglevel1

ggsave("water_vs_bio.tiff",bubbleplot_siglevel1,height=7,width=9,units="in",dpi=300)


################Let's compare sed vs biotrap
sed_bio<-data_deseq_countdata2_wide[, -grep("CCRF-",
colnames(data_deseq_countdata2_wide))]
str(sed_bio)

sed_bio_meta<-data_deseq_metadata2[-grep("CCRF-",
data_deseq_metadata2$sample.name),]
str(sed_bio_meta)
sed_bio_meta<-droplevels(sed_bio_meta)
#start deseq, design is using metadata criteria.  in this case we are  using substrate.
```

```
#When doing the DESeq analysis we have specified a Wald test.
#The alternative is LRT and is better for more than two class problems.
dds_meta2<-DESeqDataSetFromMatrix(countData=sed_bio,
                          colData= sed_bio_meta,design= ~type)
dds_meta2<-DESeq(dds_meta2,test="Wald",fitType = "local")
head(dds_meta2)
#convert deseq object into results
res_meta2<-results(dds_meta2,cooksCutoff = TRUE)
#Cookscutoff is not accurate without 3 reps
#summary(res_meta)
plotMA(res_meta2, main="DESeq2",ylim=c(-5,5))
#plot showing abundance vs log2fold change, could be supp fig.
#this shows that the functions with low expression do not have high log fold change.
#this is the intention of the deseq2 package to avoid overexaggerated log fold changes for
#low expressed functions

#filter the results by alpha value or other criteria
alpha=0.05
sigtab_meta2 = res_meta2[which(res_meta2$padj < alpha), ]
#sigtab_meta = res_meta[which(res_meta$padj < alpha & res_meta$baseMean >1000), ]
#we could filter results by both alpha value and the normalized average.

mcols(sigtab_meta2)$description
#this gives you the description about each column in the results table.
#"sed vs. biotrap" means functions with positive log2 fold change are more
#abundant in the sed.
sigtab_meta2<-sigtab_meta2[order(sigtab_meta2$log2FoldChange),]

#put the results in order of log2fold change

#convert results into a regular data frame for ggplot2 to use
sigtab_meta_dataframe2<-as(sigtab_meta2, "data.frame")
head(sigtab_meta_dataframe2)
#make the function as factor for plotting
sigtab_meta_dataframe2$function.<-rownames(sigtab_meta_dataframe2)
sigtab_meta_dataframe2$function.<-as.factor(sigtab_meta_dataframe2$function.)

#this step can take a long time, maybe even crash R, if you don't subset the
#columns you really need from the annotation hierarch
data_genes2<-data_gene_hier[,c("level1","function.")]
done2<-merge(sigtab_meta_dataframe2, data_genes2, by="function.")
#add the annotation hierarch to the results data frame
```

```
# plot results
ggplot(done2, aes(level1,log2FoldChange))+geom_point(aes(size=baseMean)) +
  theme(axis.text.x=element_text(angle=-90, size=8, colour="black")) +
  geom_hline(yintercept=0)

###a simpler looking plot.
#to calculate the average and std dev of the log2 fold change and basemean per level1
done_summary1<-aggregate(log2FoldChange~level1, data=done2, mean)
done_summary2<-aggregate(log2FoldChange~level1, data=done2, sd)
done_summary3<-aggregate(baseMean~level1, data=done2, mean)
done_sum2<-merge(done_summary1,done_summary2, by="level1")
done_sum<-merge(done_sum2,done_summary3, by="level1")
#done_sum has all the needed info for plotting
#done_sum<-done_sum[order(done_sum$log2FoldChange.x),]
#make the order by log2foldchange
#done_sum$level1_order<-factor(done_sum$level1, as.character(done_sum$level1))
#plot average and std dev.
limits<- aes(ymax = log2FoldChange.x + log2FoldChange.y, ymin=log2FoldChange.x -
log2FoldChange.y)
#to put the error bars
bubbleplot_siglevel2=ggplot(done_sum, aes(level1,log2FoldChange.x))+
  geom_point(aes(size=baseMean))+
  theme(axis.text.x=element_text(angle=90, size=10, colour="black"))+
  geom_hline(yintercept=0)+
  geom_errorbar(limits, width=0.2, linetype=5)+
  scale_y_continuous(name="log2 Fold Change",breaks=seq(-6,6,1))
bubbleplot_siglevel2

ggsave("sed_vs_bio.tiff",bubbleplot_siglevel2,height=7,width=9,units="in",dpi=300)

################Let's compare sed vs wat
sed_wat<-data_deseq_countdata2_wide[, -grep("CCRB-",
colnames(data_deseq_countdata2_wide))]
str(sed_wat)

sed_wat_meta<-data_deseq_metadata2[-grep("CCRB-",
data_deseq_metadata2$sample.name),]
str(sed_wat_meta)
sed_wat_meta<-droplevels(sed_wat_meta)
#start deseq, design is using metadata criteria.  in this case we are  using substrate.
#When doing the DESeq analysis we have specified a Wald test.
```

```
#The alternative is LRT and is better for more than two class problems.
dds_meta3<-DESeqDataSetFromMatrix(countData=sed_wat,
                          colData= sed_wat_meta,design= ~type)
dds_meta3<-DESeq(dds_meta3,test="Wald",fitType = "local")
head(dds_meta3)
#convert deseq object into results
res_meta3<-results(dds_meta3,cooksCutoff = TRUE)
resultsNames(dds_meta3)
#Cookscutoff is not accurate without 3 reps
#summary(res_meta)
plotMA(res_meta3, main="DESeq2",ylim=c(-5,5))
#plot showing abundance vs log2fold change, could be supp fig.
#this shows that the functions with low expression do not have high log fold change.
#this is the intention of the deseq2 package to avoid overexaggerated log fold changes for
#low expressed functions

#filter the results by alpha value or other criteria
alpha=0.05
sigtab_meta3 = res_meta3[which(res_meta3$padj < alpha), ]
#sigtab_meta = res_meta[which(res_meta$padj < alpha & res_meta$baseMean >1000), ]
#we could filter results by both alpha value and the normalized average.

mcols(sigtab_meta3)$description
#this gives you the description about each column in the results table.
#"substrate xylan vs. amended" means functions with positive log2 fold change are more
abundant in the xylan.
#"substrate water vs. sed" means functions with positive log2 fold change are more abundant
in the water.


sigtab_meta3<-sigtab_meta3[order(sigtab_meta3$log2FoldChange),]

#put the results in order of log2fold change

#convert results into a regular data frame for ggplot2 to use
sigtab_meta_dataframe3<-as(sigtab_meta3, "data.frame")
head(sigtab_meta_dataframe3)
#make the function as factor for plotting
sigtab_meta_dataframe3$function.<-rownames(sigtab_meta_dataframe3)
sigtab_meta_dataframe3$function.<-as.factor(sigtab_meta_dataframe3$function.)

#this step can take a long time, maybe even crash R, if you don't subset the
```

```
#columns you really need from the annotation hierarch
data_genes3<-data_gene_hier[,c("level1","function.")]
done3<-merge(sigtab_meta_dataframe3, data_genes3, by="function.")
#add the annotation hierarch to the results data frame

# plot results
ggplot(done3, aes(level1,log2FoldChange))+geom_point(aes(size=baseMean)) +
  theme(axis.text.x=element_text(angle=-90, size=8, colour="black")) +
  geom_hline(yintercept=0)



###a simpler looking plot.
#to calculate the average and std dev of the log2 fold change and basemean per level1
done_summary1<-aggregate(log2FoldChange~level1, data=done3, mean)
done_summary2<-aggregate(log2FoldChange~level1, data=done3, sd)
done_summary3<-aggregate(baseMean~level1, data=done3, mean)
done_sum2<-merge(done_summary1,done_summary2, by="level1")
done_sum<-merge(done_sum2,done_summary3, by="level1")
#done_sum has all the needed info for plotting
#done_sum<-done_sum[order(done_sum$log2FoldChange.x),]
#make the order by log2foldchange

#plot average and std dev.
limits<- aes(ymax = log2FoldChange.x + log2FoldChange.y, ymin=log2FoldChange.x -
log2FoldChange.y)
#to put the error bars
bubbleplot_siglevel3=ggplot(done_sum, aes(level1,log2FoldChange.x))+
  geom_point(aes(size=baseMean))+
  theme(axis.text.x=element_text(angle=90, size=10, colour="black"))+
  geom_hline(yintercept=0)+
  geom_errorbar(limits, width=0.2, linetype=5)+
  scale_y_continuous(name="log2 Fold Change",breaks=seq(-6,6,1))
  #scale_x_discrete(labels = add_newlines(done_sum$level1, 23), name = "")
bubbleplot_siglevel3

ggsave("sed_vs_wat.tiff",bubbleplot_siglevel3,height=7,width=9,units="in",dpi=300)

############################go down a level

data_genes5<-data_gene_hier[,c("level2","function.")]
done5<-merge(sigtab_meta_dataframe3, data_genes5, by="function.")
```

```
#add the annotation hierarch to the results data frame

# plot results
ggplot(done5, aes(level2,log2FoldChange))+geom_point(aes(size=baseMean)) +
  theme(axis.text.x=element_text(angle=-90, size=8, colour="black")) +
  geom_hline(yintercept=0)


###a simpler looking plot.
#to calculate the average and std dev of the log2 fold change and basemean per level1
done_summary1<-aggregate(log2FoldChange~level2, data=done5, mean)
done_summary2<-aggregate(log2FoldChange~level2, data=done5, sd)
done_summary3<-aggregate(baseMean~level2, data=done5, mean)
done_sum2<-merge(done_summary1,done_summary2, by="level2")
done_sum<-merge(done_sum2,done_summary3, by="level2")
#done_sum has all the needed info for plotting
done_sum<-done_sum[order(done_sum$log2FoldChange.x),] #make the order by
log2foldchange
#done_sum$level2_order<-factor(done_sum$level2, as.character(done_sum$level2))
#plot average and std dev.
limits<- aes(ymax = log2FoldChange.x + log2FoldChange.y, ymin=log2FoldChange.x -
log2FoldChange.y)
#to put the error bars
bubbleplot_siglevel5=ggplot(done_sum, aes(level2,log2FoldChange.x))+
  geom_point(aes(size=baseMean))+
  theme(axis.text.x=element_text(angle=90, size=10, colour="black"))+
  geom_hline(yintercept=0)+
  geom_errorbar(limits, width=0.2, linetype=5)+
  scale_y_continuous(name="log2 Fold Change",breaks=seq(-6,6,1))
  #scale_x_discrete(labels = add_newlines(done_sum$level2, 23), name = "")
bubbleplot_siglevel5

write.csv(done_sum,"water_vs_sed_level2.csv")

ggsave("sed_vs_wat2.tiff",bubbleplot_siglevel5,height=15,width=25,units="in",dpi=300)
```

## VITA

Kathleen Merritt Brannen-Donnelly was born in Pahokee, Florida. During her childhood, Kathleen lived in several states and had various educational experiences in private, public, and magnet schools. Because Kathleen's father was hired to work at a sugar refinery in Louisiana, Kathleen graduated from high school in Louisiana the year that Hurricane Katrina flooded New Orleans. Her plans until that point had been to attend Tulane University; however, Katrina decimated the region, and Tulane did not have many science major positions available. Kathleen decided to apply for scholarships to Louisiana State University. Kathleen was offered a scholarship to go to freshman geology field camp, and saw mountains for the first time. Kathleen decided to major in both geology and biology, because it was hard to choose between her two academic interests. Kathleen worked in Annette Summers Engel's lab as a lab manager and student researcher where she completed an undergraduate thesis. During her tenure at LSU, Kathleen had many wonderful opportunities to travel around the world for conferences and research, including Antarctica, Slovenia, Spain, and Turkey. After touring several schools, Kathleen decided to take her National Science Foundation Graduate Research Fellowship to the University of Tennessee. During her time at the University of Tennessee, Kathleen married her fellow student Brendan Donnelly, and also adopted a dog named Pipistrelle (Pippa for short). There have also been many wonderful opportunities for Kathleen to travel to conferences and present research in Mexico, Slovenia, and Italy.