METAGENOMIC ANALYSIS OF SPRING AND STREAM WATERS IN THE CHESAPEAKE AND OHIO CANAL NATIONAL HISTORICAL PARK

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
MASTER OF SCIENCE IN ENVIRONMENTAL ENGINEERING

---

by
Asad Ullah Khan
July 2015

Examining Committee Members:

Dr. Benoit Van Aken, Advisory Chair, Civil and Environmental Engineering, Temple University
Dr. Rouzbeh Tehrani, Civil and Environmental Engineering, Temple University
Dr. Evelyn Walters, Civil and Environmental Engineering, Temple University
Dr. Hui (Lisa) Yu, Civil and Environmental Engineering, Temple University

ABSTRACT

In the current century, the most critical crises faced by human kind will likely be climate change, shortage of energy supplies, and pollution of the environment. A large variety of contaminants are susceptible to be released in the environment from households and from agricultural and industrial activities. During the last decades, physical, chemical, and biological technologies have been developed for pollution remediation and for assessing the extent of environmental contamination in water resources. Because of the large diversity of contaminants, the systematic and comprehensive analysis of elemental and compound pollutants cannot practically be conducted over an extensive network of water bodies. As a consequence, large-scale surface water monitoring programs frequently rely on biological assessment protocols based on macroinvertebrates, microalgae, or fishes, allowing to integrate the impact of many potential contaminants into single indices that are easy to interpret.

However, standard bioassessment protocols are currently based on the morphological identification of representative sets of indicator organisms, which requires extensive stream sampling and laboratory observation in the laboratory and taxonomic identification. These operations are time- and personnel-consuming and require a great deal of experience. In this project, we have developed and validated an innovative water quality bioindicator based on the metagenomic analysis of the total prokaryotic microbial community in the water.

Microorganisms are essential components of the aquatic ecosystem and their diversity, nature, and distribution typically reflect variations of the environmental

conditions and water quality parameters. Although conventional, cultivation-based methods for microbial characterization are important in investigating the microbial communities, they are time and resources consuming. New polymerase chain reaction (PCR)-based molecular methods, such as metagenomic pyrosequencing, have the potential to quickly provide the detailed information on the microbial communities present in any environment. Advanced bioinformatics computing in connection with the resources of extensive genomic databases allow providing the detailed distribution of the microbial species present in the samples, which, in this project, was used as a fingerprint of water quality.

The proposed research has been conducted using water samples collected from the Chesapeake and Ohio Canal National Historical Park (CHOH) in Maryland. Comprehensive characterization of the aquatic bacterial communities has been performed using metagenomic pyrosequencing. In parallel, a suite of relevant water quality parameters were monitored in the samples using standard methods. Using redundancy analyses (RDA), meaningful relationships were established between water characteristics and the metagenomic biomarker, showing its potential utilization as a general water quality indicator.

This study provides the basis for the development of an innovative method for the fast and cost-effective assessment of water quality based on the aquatic prokaryotic microbiome. Phylogenetic analyses conducted on the metagenomic data revealed that the dominant prokaryotic phyla detected in the 19 samples are similar to the ones typically detected in freshwater environments. Microbial diversity indices showed that all 2012 samples were characterized by a low biodiversity, while 2013 samples were characterized

by a higher diversity, which is likely the result of different meteorological conditions in 2012 and 2013.

Clustering analysis and principal component analysis (PCA) were conducted to investigate the relationships between the relative abundance of the prokaryotic phyla and water quality parameters. The results showed that the samples collected from the same sites in different years cluster well together when compared based on the water quality parameters. On the contrary, the samples collected in 2012 made a separate group of cluster and same is true for 2013 samples when compared based on the prokaryotic phyla. These observations suggest a larger temporal variation of the microbial communities than the physico-chemical parameters of the water.

PCA focusing on prokaryotic communities showed that *Proteobacteria* and *Bacteroides* phyla, including aerobic heterotrophic, fast growing bacteria – referred to as copiotrophic or 'r-type' organisms --, cluster together. On the other hand, the other phyla, including mostly anaerobic and/or autotrophic, slow growing bacteria – referred to as oligotropic or 'K-type' organisms --, form a rather distinct cluster.

The dependence of the prokaryotic relative abundance on the water quality parameters for the 19 samples was then interrogated using RDA. As showed by PCA investigations, the r-type phyla cluster together and correlate with high alkalinity and conductivity. On the contrary, the K-type phyla cluster together and correlate collectively with sulfate and nitrate. As expected, the copiotrophic, fast-growing, r-type phyla also correlate with the stream samples, while the oligotrophic, slow-growing, K-type phyla correlate better with spring, cave, and mine samples.

This study provides the basis for the development of an innovative method for the fast and cost-effective assessment of water quality based on the prokaryotic microbiome.

DEDICATION

Dedicated to my late father Mr. Zar Khalil Khan who instilled in me the attitude of perseverance and the aspiration of aiming high in achieving my goals. Unfortunately, he did not stay in this world long enough to congratulate me on this great achievement in my life.

ACKNOWLEDGEMENTS

I express my deepest sense of gratitude to my advisor Dr. Benoit Van Aken for the opportunities he provided me during my master degree program at Temple University. Dr.Van Aken expertly guided me through my graduation and his unwavering enthusiasm kept me constantly engaged with my research. Dr. Van Aken helped me immensely in statistical analysis of my research work and guided me in honing my writing skills. I am honored to have been chosen as his student and got the opportunity of conducting research in his lab. I have been extremely fortunate to have an advisor whose patience and support helped me to finish my master thesis.

I would like to thank my committee members for their advice throughout my research work. It has been an honor to have such a diverse and knowledgeable committee

I would also thank to my colleagues and friends for their cooperation and support. I greatly value their friendship and deeply appreciate their belief in me. I take this opportunity to thank Civil and Environmental Engineering department at Temple University for their trust on my abilities and sponsored my studies at Temple University. I would thank to my family for their constant encouragement attention and support. I am also thankful to Dr. Vesper and her research group members for their contribution in the completion of the thesis. Finally, I am thankful to NRPP-Natural Resource Management for sponsoring this project

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AMP	Adenosine monophosphate
APS	Adenosine 5-phosphosulfate
ATP	Adenosine triphosphate
BAM	Bioaccumulation monitoring
BMPs	Best management practices
BOD	Biochemical oxygen demand
bp	Base pair
CA	Correspondence analysis
CCA	Canonical correspondence
CHOH	Chesapeake and Ohio Canal National Historical Park
COD	Chemical oxygen demand
DCA	Detrended correspondence analysis
DGGE	Denaturing gradient gel electrophoresis
eDNA	Environmental DNA
EM	Ecosystem monitoring
eRNA	Environmental RNA
MPN	Most probable number
NCBI	National Center for Biotechnology Information
NMDS	Non-metric multidimensional scaling
NPDES	National pollutant discharge elimination system
NPS	Non-point source
NPS	National Park Services
OCPs	Organochloride pesticides
OTU	Operational taxonomic unit
PAHs	Polycyclic aromatic hydrocarbons
PCA	Principal component analysis
PCBs	Polychlorinated biphenyls
PCR	Polymerase chain reaction
PPi	Pyrophosphate
QIIME	Quantitative Insight into Microbial Ecology
RBP	Rapid bioassessment protocol

RDA	Redundancy analyses
RDP	Ribosomal database project
RTE	Rare, threatened, and endangered
TMDLs	Total maximum daily loads
TOC	Total organic carbon
T-RFLP	Terminal-restriction fragment length polymorphism
USEPA	United States Environmental Protection Agency
WQIs	Water quality indices

CHAPTER 1

PROBLEM STATEMENT

1.1. Introduction

Increasing agricultural activities, urban communities, and industries release large amounts of toxic chemicals (xenobiotic) in the environment. These chemicals can be inorganics, such as heavy metals (cadmium, lead, arsenic, mercury) and cyanide, or organic, such as polychlorinated biphenyls (PCBs), organochloride pesticides (OCPs), polycyclic aromatic hydrocarbons (PAHs), polycyclic dibenzofurans (PCDFs), and dibenzodioxins (PCDDs). For several decades, the potential risk imposed by these pollutants to the aquatic and terrestrial ecosystem has been recognized. In particular, the aquatic environment frequently constitutes the first repository of most of these contaminants, due to both direct discharges and/or hydrological and atmospheric deposition processes [1].

Assessing water quality is therefore essential for the protection of surface water resources in every state in the US. In order to protect aquatic resources, most states have developed extensive monitoring programs of water quality in streams and rivers. Monitoring water quality is also an essential tool for the identification of non-point source (NPS) pollution loads, assessing best management practices (BMPs) to control NPS pollution, and determining total maximum daily loads (TMDLs) used for permitting purposes, e.g., United States Environmental Protection Agency (USEPA) National Pollutant Discharge Elimination System (NPDES) [2].

Because it is virtually impossible to monitor specifically every single potential contaminants that may be present in an extensive network of water bodies, surface water monitoring programs frequently rely on biological assessment protocols, such as the EPA's Rapid Bioassessment Protocol (RBP) [3]. Bioassessment methods are based on the identification, taxonomical classification, and abundance of a suite of model indicator organisms, including macroinvertebrates, microalgae, and fishes. These operations are currently performed using tedious and time-consuming conventional morphological observations. First, the sampling of the under assessment water body is time- and personnel-consuming. Then, collected organisms are identified and counted using visual and/or microscopic observation in the laboratory, which is also time-consuming and requires a high degree of expertise from the personnel. In this project, we have developed and validated an innovative water quality bioindicator based on the metagenomic analysis of total prokaryotic microbial community in the water. Indeed, the recent development of metagenomic pyrosequencing techniques has allowed obtaining rapidly and at relatively low cost the detailed profile of complex microbial communities in various environments. Metagenomic profiles are then expected to inform on prevailing environmental conditions and they can be used as powerful indicators of water quality.

1.2. Hypothesis

Microbial communities play an important role in nutrients cycling and attenuation of pollutants in the environment [4]. Microorganisms adapt quickly both physiologically and evolutionary to changes in ecological conditions and environmental stresses [5, 6]. Freshwater systems are one of the major habitats of microorganisms. Changes in water

characteristics, including temperature, pH, and salinity have been resulted into microbial composition changes in the aquatic environments [7]. Similarly, microbial communities in the aquatic environment have been shown to be highly responsive to low concentration of pollutants and nutrients [8]. Even though the mechanisms by which bacterial communities respond to changes in environmental factors and pollutants need to be further investigated, several studies have reported that the microbial communities are responsive to environmental conditions, including climate change and chemical pollutions [9]. For example, a finding of the study endorsed that there is a significant difference in microbial community structure of a ground water plume contaminated by iron leachate and unpolluted ground water [10].

The hypothesis underlying the proposed research is that the diversity, composition, and/or structure of the microbial communities will reflect the overall quality of the aquatic environment and could therefore be used as an indicator of water quality. High-resolution characterization of the microbial communities in the aquatic environment is nowadays possible using advanced molecular methods, such as metagenomic pyrosequencing [11]. The comprehensive profile of the microbial community in the water derived from the metagenome is expected to provide a very detailed fingerprint specific to water quality parameters.

1.3. Overall Objective of the Proposed Research

The proposed research has been conducted using water samples collected from the Chesapeake and Ohio Canal National Historical Park (CHOH) in Maryland, in which my thesis advisor, Dr. Benoit Van Aken, is currently conducting a funded study focusing

on the impact of water quality on rare, threatened, and endangered (RTE) species in spring and cave water (National Park Service (NPS) – Natural Resource Management (NRPP) – Cooperative Agreement #P11AC60552: "Assessing the Vulnerability of Sensitive Karst Habitats Containing RTE Species in CHOH". In fact, the entire aquatic ecosystem in the CHOH has been recognized as suffering from increasing anthropogenic intervention, including agricultural runoff and residential septic tanks. The study presented here relates specifically to the characterization of the quality of water bodies in the CHOH, which is suspected to impact the distribution and abundance of RTE species along the park.

The overall objective of the proposed research was to perform the comprehensive characterization of the bacterial communities in a suite of spring and stream waters collected in several sites along the CHOH and determine their potentiality as general and specific water quality indicators.

1.4. Specific Objectives

The proposed research has been conducted through the completion of the following technical objectives:

a) To characterize the microbial community structure in selected water samples collected from the CHOH using metagenomic pyrosequencing method.

One-liter aliquots of water were collected from selected water bodies across the CHOH and filtered on site through 0.45 μm filter membrane. The total deoxyribonucleic acid (DNA) was extracted from the filters and used for PCR amplification of 16S rDNA fragments using universal primers. The resulting clone libraries were sequenced on a

Roche 454 pyrosequencing system. The pyrosequencing data were filtered and processed to determine the relative abundance of major taxonomic groups and biodiversity indices using the Quantitative Insight into Microbial Ecology (QIIME) software package.

b) To monitor relevant water quality parameters in the water samples from the CHOH using standard methods. Temperature, pH, conductivity, and alkalinity were measured in the field. Other parameters, including inorganic C, inorganic cations (Ca, K, Mg Na), nutrients (NO_3^- and PO_4^{3-}), sulfate (SO_4^{2-}) and selected metals were measured in the laboratory.

c) To establish meaningful relationship between metagenomic biomarkers and water quality parameters. Relationship between microbial community structure (biomarkers) and water quality parameters (environmental variables) were determined by RDA using the software package CANOCO. CANOCO is a popular software for multivariate statistical analysis using ordination methods in the field of ecology. Multivariate analysis are used to relate the relative abundances of bacterial species (dependent variables) to water quality parameters (independent variables) [12].

1.5. Expected Outcomes and Significance

This study is first expected to provide a new method for the fast and cost-effective assessment of water quality based on the structure of the aquatic prokaryotic community. In addition, results from this research could bring new insights on the microbial ecology and water quality vulnerability in water bodies in the CHOH, which may help protect RTE species.

This research is significant because a fast and accurate method for the overall assessment of the quality of surface water is essential for the efficient management and protection of water resources in US streams and rivers.

CHAPTER 2

LITERATURE REVIEW

2.1. Overview

More than 3.5 million miles of streams and rivers exist in the US. These rivers and streams have many uses and functions, such as providing habitats for fishes and other aquatic species, mitigating flood damages and pollution, transportation, providing sources for irrigation and drinking water, and recreational activities. Unfortunately, human institutions, such as cities, town, farmland, mines, factories, and industries, have had increasingly larger impacts on the condition of watercourses. Restoring US water bodies to their pristine condition and/or preventing further pollution is therefore of paramount importance for the environment, including wildlife and human health and well-being [13].

Water quality management and protection require extensive monitoring of water quality in streams, lake, and river, which is today accomplished through a variety of different approaches summarized below.

2.2. Water Quality Assessment Methods

To determine the quality of water in rivers, streams, lakes and reservoir, water quality indices (WQIs) have been established by different regulatory agencies. WQIs are based on comparison between regulatory standards (i.e., the target) and actual concentration of water quality parameters that are recorded in the water [14]. Currently, a

variety of physical and chemical parameters, such as pH, temperature, dissolved oxygen, hardness, conductivity, phosphorus, nitrogen, toxic metals, and pesticides are commonly monitored and used to compute WQIs. In addition, specific compounds of concern may be monitored, such as chlorinated solvents, PAHs, or PCBs. Biological parameters include biochemical oxygen demand (BOD) and bacteriological indicators, such as total and fecal coliforms [15]. However, it is impractical to monitor such a large number of parameters on a large scale with a regular frequency. Consequently, water quality monitoring frequently relies on the assessment of indicator organisms present in the water, an approach referred to as biomonitoring.

2.3. Aquatic Pollution Monitoring

Utilizing living organisms in routine and systematic investigations to assess changes in environmental conditions of water quality is referred to as biomonitoring. The hypothesis behind biomonitoring is that different aquatic organisms exhibit different sensitivities to water quality in general, and environmental contamination in particular. Poor quality and contaminated waters typically house relatively more tolerant species -- and less sensitive ones -- and exhibit lower biodiversity than pristine water. The relative abundance of tolerant/sensitive species and biodiversity indices can therefore be used as indicators of water quality. Although time and personnel intensive, these methods are still preferred than the systematic analysis of a long suite of chemical/physical/bacteriological parameters, which may not capture the major concern in terms of water quality.

One may recognize four basic types of environmental monitoring methods to evaluate the risks of contaminants for aquatic life [16].

I. ***Bioaccumulation monitoring (BAM)*:** In BAM, contaminants levels are measured in aquatic organisms to determine the level of pollutant. This measurement inform on the level of contamination of the water as well as the level of exposure of wildlife. For example a study found that bald eagles, a higher trophic level organism, are at risk of accumulating PCBs in Hudson river [17].

II. ***Biological effect monitoring (BEM)*:** BEM identifies endpoints that reflect early detrimental effects in aquatic organisms that may or may not be reversible following contaminants exposure. This measurement informs more specifically on the detrimental effects of contaminants in the aquatic wildlife. A study by Hayes et al, 2002 concluded that atrazine (herbicide), an endocrine disruptor, could impaired sexual development in amphibian species [18].

III. ***Health monitoring (HM)*:** HM investigates the occurrence of permanent damage to the cellular structures due to contaminant exposure.

IV. ***Ecosystem monitoring (EM)*:** Ecosystem monitoring can be understood as the collection, analysis, and interpretation of data on changes in the natural environment that occur in a given ecosystem. EM attempts to observe living and non-living aspects of the biosphere, identify the response of the environment to human interventions, and predict the actual or likely impact on the environment.

2.4. Why Biological Monitoring?

Although the source of water contamination could be investigated using conventional methods, such as physical and chemical analyses, they provide only limited and indirect information on the health of the aquatic ecosystem and the biological

response to pollution. Directly monitoring the organisms – or the ecosystem – instead of environmental quality parameters helps capture more precisely the adverse effect on the organism's health in an ecosystem.

Due to the high variability in chemical concentration of water bodies, it is almost impossible to determine the overall effect of contaminants on ecosystem health. These variations are related to fluctuating point source discharges, inconstant water flow patterns, and/or precipitation events.

Furthermore, chemical analyses are unable to capture biological threats to aquatic organisms and ecosystem health, such as the presence of invasive species or pathogenic organisms. On the other hand, living organisms and biological communities are exposed – and are susceptible to be responsive – to all environmental stresses caused by human intervention and natural events over a longer period of time. Consequently, the physiological conditions and diversity of organisms in a water body are considered to be the representative of the overall quality of their surroundings and environment [19].

Most commonly, organisms from various trophic levels are simultaneously used as biological indicators to determine water quality. These includes bacteria, algae, macroinvertebrates, plants, and fishes.

The main reason for using aquatic macroinvertebrates as a water quality indicators is that some species are more sensitive to pollution than others, so that the ratios of the abundance of sensitive species to tolerant ones can be used as an indicator of water quality [20]. Another reason for using macroinvertebrates is the relative ease of

sampling and the existence of well-established sampling and identification procedures [19].

Another lower-trophic group of organisms commonly used to monitor water quality is algae. The short life span of algae make them attractive as a water quality indicator especially to evaluate the effect of pollutants in the short term. Algae are known to be very sensitive to most chemical and physical factors affecting water quality. They are important water quality indicators because some pollutants which accumulate more readily than others, have deleterious effect on algal assemblages. In addition, algae's sensitive metabolic system indicates variability of natural and environmental disorder. Algae is preferred because of their inexpensive and easy cultivation and sampling technique in the laboratory [21].

Fish could also be used as a water quality indicator [22]. One of the reasons, using fish is because of their acceptability to general public as a water quality indicator. Fish are preferable water quality indicator due to their integration of pollutants effects over longer time periods and over extensive area of study.

In comparison, there are a limited number of studies available that focus on microbial community used as a water quality indicators [23]. Most of these studies focused on the monitoring of indicator organisms, such as *Escherichia coli* or fecal bacteria, in water bodies to assess contamination by sewage, septic tanks, or other source of fecal material from human or animal origin [24]. Recent advancement in molecular microbiological techniques has made possible to characterize the complete prokaryotic community in any environment, including the aquatic ecosystem. These techniques allow using the microbial community as a fingerprint of a specific environment. These

techniques enable to study microbial community structures of many samples simultaneously in speedy and cost-effective manners.

2.5. Methods for the Characterization of the Microbial Communities

2.5.1 Culture or Cultivation Dependent Methods

Conventionally, culture-dependent methods, such as viable plate counts and most probable number (MPN), were used for the quantification of microbial communities in a variety of environments, including water, wastewater, soil, sediments, activated sludge, and enteric microflora of animals and humans [25]. These methods were used for more than a century to detect and quantify bacteria in the environment [26]. In culture–dependent methods, microbes in environmental samples are cultivated on a laboratory medium. The selection of the medium for cultivation depends on the goal of the study. Complex, unspecific media (e.g., nutrient broth) are typically used to quantitatively characterize the overall microbial community – or large groups of organisms (e.g., total heterotrophic counts). On the contrary, selective media and cultivation conditions can be used to selectivity identify and quantify target microbes (e.g., specific media for fecal coliform growth) [27]. Although commonly used, culture-dependent methods has many shortcomings especially for the accurate quantification of the bacterial abundance [28].

Quantitatively, culture-based methods are not reliable because they systematically underestimate the actual numbers of microorganisms in the environment by one to three order of magnitude, as it was recently shown by the use of methods based on DNA (this paradox has been referred to as the "great plate count anomaly", by analogy to the UV anomaly in quantum physics). It is believed that this situation originates from the

difficulty to reproduce *in vitro* the metabolic and physiological conditions favorable for bacterial growth. The high diversity of bacteria in most environmental samples makes it difficult to create conditions allowing the majority of them growing on culture plates media [29]. That is why it is usually considered that only less than 10% – and often less than 1% – of total bacteria in environmental samples could be cultivated in the laboratory [30]. Furthermore, the relative abundance of specific bacterial phyla could be too low to be detected by conventional culture-dependent techniques, even though their numbers could be large enough to cause significant effect on the environment, e.g., pathogens [31].

2.5.2 Culture-Independent Methods

The limitations of culture-dependent methods to characterize environmental microbial communities have been overcome by the emergence of culture-independent methods, also known as genetic or molecular biology methods. These methods become popular in 1980s due to the invention of the PCR, the emergence of sequencing techniques, and the subsequent development of molecular ecological methods [32]. In most of these methods, DNA and/or RNA – although other biomolecules, including proteins and lipids, have also been used for microbial characterization – is extracted directly from environmental samples (environmental DNA (eDNA) or environmental RNA (eRNA)) without any cultivation steps and analyzed by various molecular techniques, including electrophoresis, PCR, or sequencing [33].

The application of culture-independent techniques has dramatically overcome the traditional limitations faced by cultivation methods and it has recently revealed a new and improved view of the microbial world.

Culture-independent techniques can be based on the analysis of the whole genomes or selected genes, such as 16S rDNA and 18S rDNA for prokaryotes and eukaryotes respectively. Due to the advancement in molecular techniques, the characterization of the phylogeneticity and functional diversity of microorganisms can be performed with greater depth, allowing microbial ecology to progress rapidly.

To analyze the 16S rRNA/rDNA, culture independent methods such as Denaturing Gradient Gel Electrophoresis (DGGE), Terminal-Restriction Fragment Length Polymorphism (T-RFLP) and clone libraries have been widely used over the past two decades [34, 35]. In this study we adopted new sequencing methodology called pyrosequencing to analyze microbial community structure. The technique is widely applicable technology for in depth characterization of nucleic acids. Before presenting the pyrosequencing technique with greater detail, the phylogenetic DNA marker, primer design and PCR method will be explained further.

2.5.2.1 Phylogenetic DNA Marker and Primer Design

Although several markers have been proposed over the years for species identification (e.g., plant barcoding project, metagenomics analyses), 16S rDNA has emerged the marker of the choice for identification and phylogenetic characterization of microbial communities. The requirements for suitable marker include universality (i.e., to be present in all species), conservation (i.e., to have similar regions to allow the design of

primers susceptible to anneal to sequences in all species), and enough variation in the region between the primer (i.e., to allow univocal identification at the species level). Among other requirements, the region between the primers should be long enough, again, for the univocal identification of species. The conserved gene coding for the small ribosomal-subunit RNA, 16S rRNA/16S rDNA, satisfies best these criteria and it has become the most-commonly used biomarker for prokaryotic phylogenetic analysis. The ribosomal DNA include highly-conserved regions suitable for the design of universal primers susceptible to amplify 90 – 99% of all prokaryotic species [36]. Another advantage of the 16S rDNA is the presence of several copies per genome, making it a strong marker for detection of less represented species. On the other hand, even though the complete 16S rDNA sequence is about 1,500-base pair (bp) long, routine PCR-based methods and DNA-pyrosequencing result in generation of relatively short sequence which requires focusing on hypervariable regions within the 16S rRNA gene. Sequences of hypervariable regions allow identifying bacteria to the genus or species level [37, 38].

2.5.2.2 PCR Technique

In PCR techniques, total DNA/RNA extracted from environmental sample is used as a template for the amplification of a specific marker and characterization of the microorganisms. Using 'universal primers', a large majority of marker genes can be amplified, generating a mixture of microbial genes that are signatures of organisms (prokaryotes/eukaryotes) present in the environmental sample. It is utmost important to select a marker conserved enough for the annealing of 'universal primers' to conserved domains – therefore capturing a large majority of the organisms present -- while neighboring regions flanked by the primers with enough variability to allow

discrimination to the genus or species levels. Although a variety of markers have been proposed over the years, the genes coding for the large subunit of the ribosome, 16S ribosomal (rDNA) for prokaryotes and 18S rDNA for eukaryotes have today emerged as the gold standard for phylogenetic and metagenomic analyses. Extensive public databases, such as the Ribosomal Database Project (RDP) and the National Center for Biotechnology Information (NCBI) contain hundreds of thousands of 16S rDNA sequences that can be interrogated by advanced bioinformatic applications for obtaining extremely detailed picture of the microbial communities present in various environments [39].

2.6. Pyrosequencing Technique

Recently, the DNA extracted from environmental samples – as it is, or after PCR amplification of a specific marker -- has been analyzed through the pyrosequencing technology which is a next-generation sequencing technique. This technique allows sequencing the total eDNA or an entire clone library representative of the microbial community present in a sample in a rapid and comprehensive manner [40].

In 1977, the DNA sequencing technology was developed by two groups, Maxam & Gilbert [41] and Sanger & his colleagues [42]. These early methods were based on the dideoxy chain termination technique. Early sequencing was a tiresome and time-consuming job. For example, it could take a week for sequencing only few thousands bases. Automated sequencing resolved this issue and allowed sequencing larger DNA segments in shorter periods of time. With the introduction of capillary DNA sequencers, the quality of sequencing was improved and sequencing of longer read length was

possible in only a few hours. One of the major issues of Sanger sequencing with respect to metagenomic analyses is that it requires a unique pool of similar DNA sequences, such as obtained from the PCR amplification of a single, univoqual gene sequence. 16S rDNA amplification using eDNA as a template results in the generation of hundreds of thousands of dissimilar sequences that must be separated before sequencing (e.g., by cloning and amplification in a bacterial vector), which is time- and personnel-intensive.

Today, advances in so-called next-generation sequencing have offered the possibility to sequence large pool of different DNA sequences, such as generated in metagenomic libraries. Sequencing by hybridization, parallel signature sequencing based on ligation and cleavage, and pyrosequencing are three most recognized second-generation sequencing technique frequently used in the field of metagenomics [43].

In 1996, a new sequencing technology named pyrosequencing was developed by Mostafa Ronaghi and Pål Nyrén at Royal Institute of Technology in Stockholm. This technique allows sequencing of eDNA or an entire clone library representative of the microbial community in a rapid manner [40]. Besides identification of microorganisms potentially to the species level, the technique has the advantage over other molecular techniques of providing quantitative estimation of microbial community [44]. Because of the very large number of sequences obtained (hundreds of thousands), statistical bioinformatics methods allow obtaining a precise quantification of the relative abundance of the taxa present in the samples. Pyrosequencing is a DNA sequencing technique that is based on the detection of released inorganic pyrophosphate (PPi) during DNA synthesis. The light generated after a chain of enzymatic reactions is proportional to the incorporated nucleotides. The process starts when the DNA sequence incorporates added

nucleotide and release PPi. PPi converts to Adenosine triphosphate (ATP) by enzyme ATP sulfurylase. The ATP provides energy to luciferase which oxidize luciferin and consequently light will be generated and could be seen as a peak in a pyrogram™ [43]. The Remaining nucleotides are degraded by enzyme apyrase. As the process continues, the complementary DNA strand is developed and the nucleotide sequence is determined from the peak in the pyrogram™. Following equations by Ronaghi et al. [43] show the general principle behind different pyrosequencing reaction systems.

$$(NA)_n + \text{Nucleotide} \xrightarrow{Polymerase} (NA)_{n+1} + \text{PPi}$$

$$PPi + \text{APS} \xrightarrow{ATP\ sulfurylase} \text{ATP} + SO_4^{-2}$$

$$\text{ATP} + \text{Luciferin} + O_2 \xrightarrow{Luciferase} \text{AMP} + \text{PPi} + \text{Ocyluciferin} + CO_2 + \text{Light}$$

Figure 2.1 further explain the pyrosequencing process.

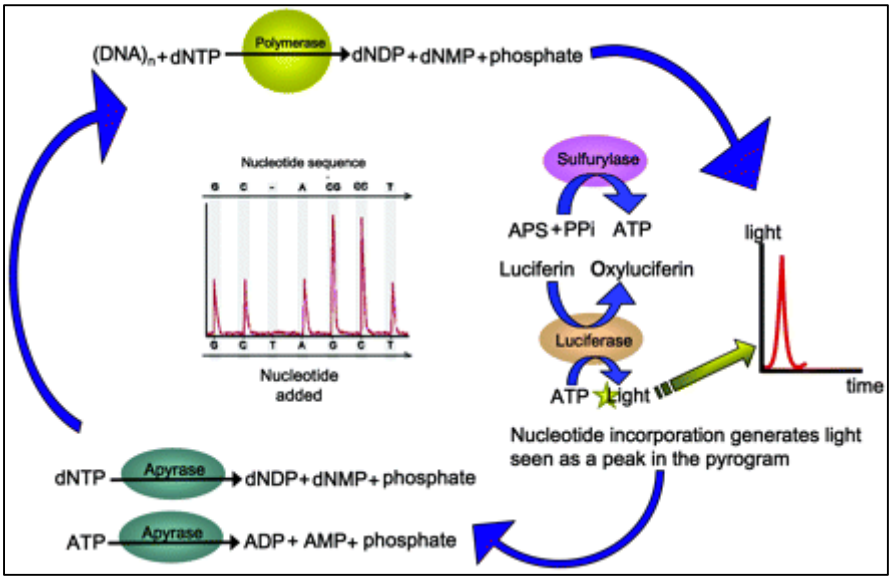

Figure 2.1 General principle behind different pyrosequencing reaction systems: First a matching nucleotide is incorporated into the DNA sequence by the enzyme DNA polymerase, with the release of a PPi. PPi is then converted to ATP by the enzyme ATP sulfurylase. Finally, the ATP formed leads to production of light through the oxidation of luciferin by the enzyme luciferase. [45].

Different platforms are available today for high throughput sequencing such as the Roche 454 pyrosequencing system (Life Science), the Illumina/Solexa platform, and the SOLiD™ platform (Applied Biosystems) [46]. In the research presented in this thesis, we used 454 Roche pyrosequencing (Life Sciences, subsidiary of Roche Diagnostic) primarily because the possibility to generate directly 500 – 600 (bp) sequences without the need for further *in silico* assembly. Such fragment length allows identification of most bacteria to the genus or species level. Other platforms currently only allow sequencing of shorter fragments, which requires additional processing susceptible to introduce errors.

2.7. Processing Pyrosequencing Data

The pyrosequencing process generates hundreds of thousands of sequences so called reads, which must be processed using advanced bioinformatic and statistical computational methods. Processing of our pyrosequencing data has been performed in two major steps of phylogenetic analysis and multivariate analysis.

2.7.1 Phylogenetic Analysis

The phylogenetic analysis includes filtering out short and poor quality reads, pooling sequences belonging to each sample together (based on their respective barcode), pooling similar sequences in each sample together, and assigning phylogenetic taxon to sequences. The outcome of this process will provide information about the identification of microbial phyla present in each sample, the relative abundance of the identified phyla in each sample, and biodiversity of the microbial community in each samples. In the present study, phylogenetic analyses have been performed using the microbial ecology software package QIIME [47].

2.7.1.1 Filtering, Sorting, and Pooling

In brief, pyrosequencing data are first filtered to remove short reads (based on the fragment length) and poor fidelity sequences (based on the comparison with the known region of the reads: primers, adaptors, and barcodes). Filtered reads belonging to each sample are then pooled together based on the specific barcode sequence. Within each samples, similar sequences (e.g., sharing more than 97% base similarity) are pooled, each pool representing a single group of microorganisms or phylotype. Sequences representing each phylotype are then aligned and a consensus sequence representative of the phylotype will be generated.

2.7.1.2 Phylogenetic Identification and Relative Abundance

Phylotypes are identified by comparison of their consensus sequence with sequences in public databases, such as the NCBI database [48] and the RDP [49]. Each phylotype is then assigned a phylogenetic identification or operational taxonomic unit (OTU). OTUs can be defined up to the species level, but most commonly to the genus or high phylogenetic level (e.g., family, class, etc.). Based on the number of sequences corresponding to the phylotype, the relative abundance of each phylogenetic group – OTUs – in the samples are calculated. Biases can be present due to the differences in the PCR and sequencing efficiency of different base sequences.

2.7.1.3 Biodiversity and Rarefaction Analysis

Based on the number and relative abundance of OTUs, the diversity within each sample – alpha-diversity – is calculated using various diversity indices (e.g., Shannon,

Chao1). The diversity and richness in OTUs in each samples was examined further by using rarefaction analysis (rarefaction curves).

2.7.2 Multivariate Analysis

The phylogenetic analysis generates 'stand-alone' data about the structure of the microbial community in each samples and the relationships between communities across samples. In order to establish correlation between the structure of the microbial communities and environmental conditions, in the case of this study, water quality parameters – one must use other techniques collectively known as multivariate statistical methods. Multivariate statistical methods have been used for many years by ecologists to process multidimensional data on community composition, properties of populations, and properties of the environment. These data need to be analyzed in a multidimensional perspective as observing the properties of each variable separately does provide useful information in most cases [50]. In the present study, the extent to which bacterial community structure relates – and can be used to predict – environmental variables (i.e., water quality parameters) has been performed using a kind of multivariate method called RDA using CANOCO 5 software (Biometris, Wageningen, The Netherlands).

Ecologist and environmental microbiologists have often observed that the environmental conditions affect biological or microbiological composition. The gradual change in species composition or species composition gradient can therefore be related to quantitative environmental parameters. Although these relationships can be analyzed through a wide range of statistical methods, the most popular method for analyzing the relationships between metagenomic data and environmental parameters are called

ordination methods. As in Principal Component Analysis (PCA), ordination methods represent relationships between species composition (or response or independent variable), environmental parameters (explanatory or environmental variable), and/or samples with two-dimensional diagrams in which proximity implies similarity. In the case of multivariate analysis when no environmental variables are accessible, the approach to be chosen in ordination methods, includes indirect gradient analysis (e.g. PCA, correspondence analysis (CA), detrended correspondence analysis (DCA), and non-metric multidimensional scaling (NMDS) or cluster analysis. If environmental variables are available for a set of response variables, the methods to be chosen is the direct gradient analysis (e.g. – RDA and canonical correspondence – (CCA).

Table 2.1 summarizes the type of statistical models that can be used to process metagenomics data.

Table 2.1 The type of multivariate statistical models adapted from [50] CVA, canonical variate analysis; DCA, detrended correspondence analysis; NMDS, non-metric multidimensional scaling

Response Variable(s)	**Predictor(s)**	
	Absent	**Present**
One	• Distribution summary	• Regression models *sensu lato*
Many	• Indirect gradient analysis • (PCA,DCA,NMDS) • Cluster analysis	• Direct gradient analysis • Discriminate analysis (CVA)

2.7.2.1 Gradient Analysis Methods

The regression analysis aims to quantify the dependence of a single dependent variable – univariate response –, such as the abundance of a single species -- to environmental – explanatory variables. The dependence of a set of dependent variables –

multivariate response such as the abundance of multiple species to environmental variables can be analyzed using constrained ordination methods**.** Ordination methods – constrained or unconstrained – are used to select axes of the greatest variability in the dependent variable(s) for a set of samples and visualize in a single diagram – ordination diagram – samples and species distances. In the case where no quantifiable explanatory variables are available, the analysis is called unconstrained ordination, e.g., PCA. In the case where the ordination axes also coincide with quantifiable environmental variables, they can be correlated with the ordination axes. In this case, the analysis is called constrained ordination. The constrained ordination analysis allows determining the dependence of multivariate dependent variable, such as the species composition, on multiple environmental variables, e.g. RDA.

When both dependent and environmental variables are available, both unconstrained ordination and constrained ordination should be conducted. The unconstrained ordination will provide more complete information of the variability of the dependent variable, even though not explained by any of the explanatory variables (in other words, the constrained ordination may miss trends in the dependent variables if not explained by explanatory variables). On the contrary, the constrained ordination will inform on variability of the dependent variable that is explained by the environmental variables. The two analyses are therefore complementary.

2.7.2.2 Ordination Diagram

The outcome of ordination analysis is displayed as ordination diagrams, in which the samples are represented by symbols and both dependent variables – species and

quantitative environmental variables are represented by arrows in linear methods such as RDA (and by symbols in weighted averaging methods). The values of variables increase in the direction of the arrows.

The statistical significance of a relationship is tested by the estimation of the probability of obtaining results different from those expected under the null hypotheses, which requires knowing the distribution of the test statistics. Often the distribution of the test statistics under the null hypothesis is not known, as it depends on the environmental variables, their correlation structure, and the distribution of the dependent variable. In this case, the distribution can be simulated by Monte Carlo permutation test. In short, if the null hypothesis – the dependent variables are independent of the environmental variables is true, one assumes that the values of the environmental variables are randomly assigned to the individual samples. This can be tested by performing ordination analysis with a permuted or shuffled – data set and computing the value of the test statistics for each permutation. The significance level of the test is based on the proportion of the permutations returning test statistics no lower than in the experimental data set, which is represented by the *p*-value. In other words, the level of significance decreases with the number of random permutations showing test statistics comparable to the one observed with the real data set. The test has the advantage to be independent of knowledge or assumptions about the distribution of the dependent variable. Under some assumptions, such as the normality of the data, the distribution of the F-values under the null hypothesis of independence is known as F-distribution. If this distribution is not known, it can be simulated using random permutations of the data set and calculating the corresponding F-values.

2.7.2.3 Interpretation of the Ordination Diagram (bi-Plot)

The objective of the PCA is to minimize the number of dimensions of the variability by selecting a few principal components (2 or 3 components) which are responsible for the most of the variation in data and to obtain a representation of the data without losing information. In PCA, the variation in a set of correlated variables is transformed into a small set of uncorrelated components. These uncorrelated components are called principal components (PCs) and consist in a linear function of the initial variables. First few PCs accounted for most of the variation in the original dataset. PCs are calculated from the eigenvectors of the covariance or from the correlation matrix of the initial variables. It is easy for a PCA analyst to analyze the first few PCs with respect to original variables and get maximum understanding of the data. If a large proportion of variability achieved by first few PCs, then the objective of reducing the dimensionality is achieved. PCA provides a better representation of the data variation than simple or weighted averages of the initial variables. The selection of PCs can be made using the original datasets or its covariance or correlation matrix in the case the original dataset is not available.

RDA apply a similar reductive approach to make more meaningful multivariate regression analysis, i.e., the dependence of a dependent variable to a set of independent – or explanatory variables. In a similar fashion, RDA expresses the variation of a dependent variable in response to correlated explanatory multi-variables as a function of a set of uncorrelated components, each of which being a combination of the initial independent – explanatory – variables [51].

Because of the dimension reduction, both PCA and RDA results would be easily visualized with a bi-plot. The bi-plot is a two dimensional plot where the observations (in our case, the sampling locations) and variables (in our case the microbial phyla or water quality parameters) are plotted with scattered points and arrows respectively. In bi-plot, the X and Y axes are the two PCs. The X-axis represents the first PC, while Y-axis represents the second PC. In our analysis, the circles on the bi-plot represent the samples (i.e., the water sampling locations), while the vectors or arrows represent the environmental variables (i.e., the microbial phyla and water quality parameters). Both the direction and the length of the vectors have a specific meaning. Vectors originate from the origin of bi-plot and are oriented in any direction. The direction of the arrow represents the direction of steepest rise of the plane -- or gradient --, i.e., the direction in which the value of the environmental variables (phyla or water quality parameters) increases the most. The length of the arrow equals the rate of change of the dependent variable in that particular direction. Furthermore, the bi-plot also represents the relationship between any two arrows (environmental variables) by the cosine of the angles between the arrows (vectors). If the angle between two vectors (environmental variables) is 180 degree, the correlation between the variables is -1, which represents a negative correlation. On the contrary, two arrows pointing in the same direction signify a positive relationship. If the angle between arrows is 90 degree, there are no correlations between the two corresponding variables. The location of the observations (i.e., water sampling locations) on the bi-plot can also be interpreted. Observations that are close together have similar scores on the PCs displayed in the plot. In other word, Euclidean distance between plots represents the similarity or dissimilarity between the observations.

CHAPTER 3

PROPOSED EXPERIMENTAL APPROACH

The proposed research was conducted using spring and stream water samples collected in the Chesapeake and Ohio Canal National Historical Park (CHOH). This site is selected because my thesis advisor, Dr. Van Aken, has current funding from the National Park Service (NPS), National Resource Management (NRM), to study the microbial communities in surface and spring waters in CHOH. The primary goal of the project is to assess the vulnerability of sensitive karst habitats hosting RTE aquatic species in CHOH. Protection of aquatic RTE species in CHOH requires an integrated approach of physical, chemical, geological, and biological sciences. Dr. Van Aken's primary task on the project is to use advanced molecular biology tools to characterize the microbial community structure in various aquifers of the CHOH, determine how different aquifers are related to each other based on the microbial profile, and to define meaningful relationships between metagenomic biomarkers and water quality parameters. The overall project is conducted by a team of geologist, chemists, stream biologists, and microbiologists.

This chapter outlines the methods that are being used for the conduction of the project, which include (1) characterizing the aquatic microbial community structure using metagenomic pyrosequencing methods, (2) monitoring various water quality parameters in the water samples from the CHOH using standard methods, and (3) establishing meaningful relationships between metagenomic biomarkers and water quality parameters.

3.1. Site Description

The Chesapeake and Ohio Canal (C&O) extends 296.9 km (184.5 miles) along the Potomac River in Maryland and District of Columbia (through pastoral farm country and forest to Cumberland, Maryland). Groundwater discharge locations within CHOH--springs, cave pools, and streams are important habitat for globally rare aquatic invertebrates, such as *Stygobromus gracilipes, Stygobromus biggersi, Caecidotea pricei* etc.

The quality of streams and groundwater is affected by a variety of natural and human processes, such as agricultural and residential land use. Several major types of contaminating chemicals found in water in the watershed include nutrients, trace elements, pesticides, chlorinated industrial compounds, and volatile organic compounds. Geologically, the CHOH landscape consists mainly in karstic formations. In karstic landscape, water from surface sources, transmits rapidly into subsurface environment, making underground water more vulnerable to pollution from above. Water that enters the CHOH is therefore highly dependent upon land use and activities on areas adjacent to the park, which are primarily located on private lands. Today, the park is considered a "living laboratory" that helps a better understanding of how these environmental factors have shaped park landscapes and ecosystems. Further details about the water sample locations are given in Table 3.1 and Figure 3.1.

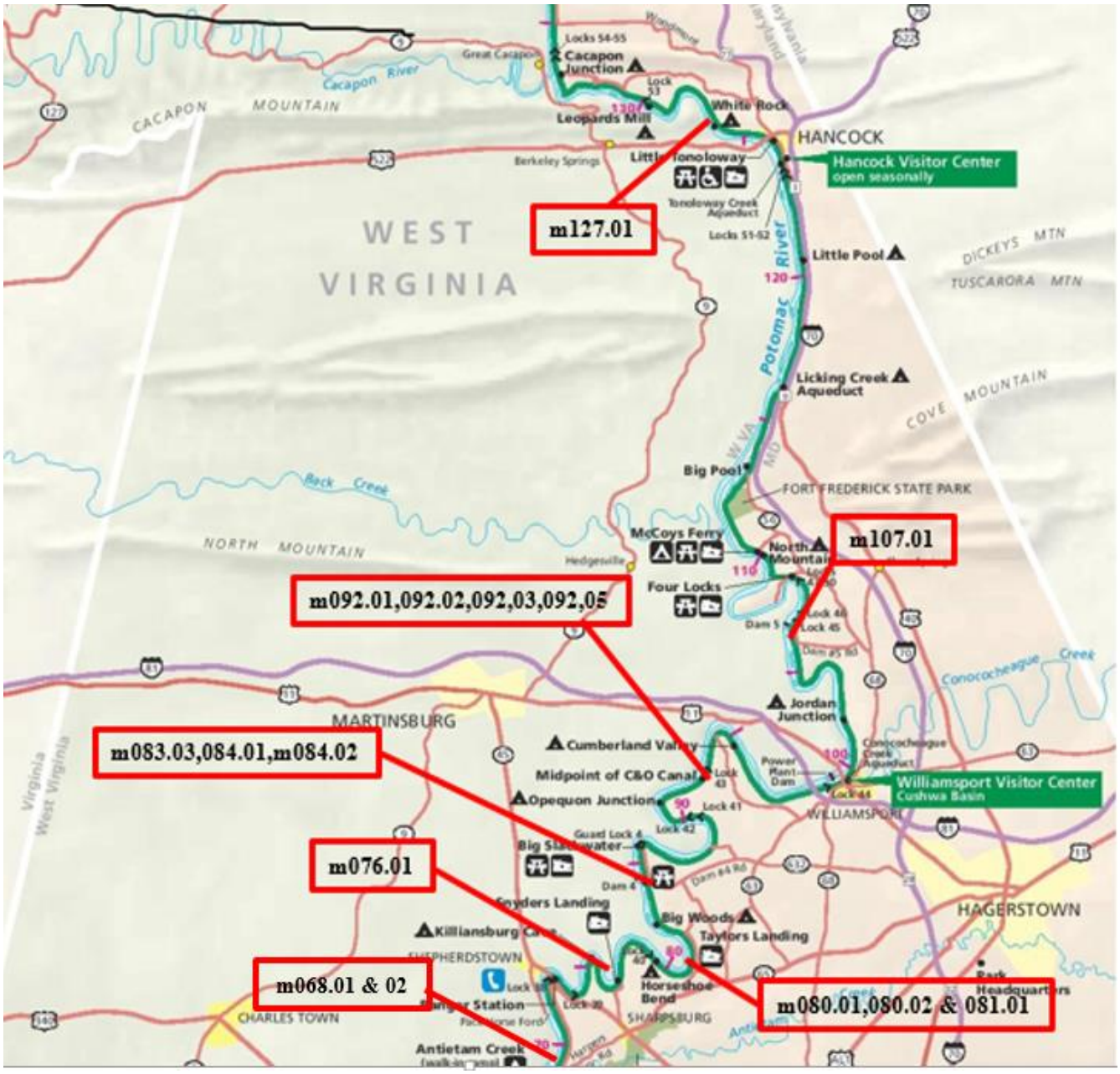

Figure 3.1 Water sampling sites along the Potomac River. "Chesapeake and Ohio Canal National Historical Park.". Maryland: National Park Services.

Table 3.1 Description of water samples, including type, locations and sampling year. Water samples were collected from springs, streams, caves and a mine along the Potomac River. Geological formation of the sampling location are divided into carbonate or non-carbonate (clastic)

Sample Location Code*	**Type (Waterbody)**	**Geological Formation**	**Sampling Year**
*m068.01**/13 & m068.02**/13	Spring	Non Carbonate	2013
m076.01/12	Spring	Carbonate	2012
m080.01/12/13 & m080.02/13	Stream	Carbonate	2012-13
m081.01/13	Spring	Carbonate	2013
m083.03/12	Cave	Carbonate	2012
m084.01/12/13 & m084.02/12	Spring	Carbonate	2012-13
m092.01/12/13, m092.02/12, m093.03/12/13 & m092.05/13	Spring	Carbonate	2012-13
M107.01/13	Spring	Carbonate	2012
M127.01/13	Mine	Non Carbonate	2013

* "m" represent miles, numbers represents actual distance from bench mark and sampling year.
**Two or more samples from same locations are represented by numbers in consecutive manners.

3.2. Methods and Techniques

3.2.1 Sample Collection

National Park Services and the Chesapeake and Ohio Canal (C & O) administration have collaborated and divided the 184.5 miles of the CHOH into 40 tourist attractions. The Canal starts from Tide Lock on the Potomac River at the mouth of the Rock Creek and is referred to as “Mile Post 0” on the CHOH map. Cumberland Visitor Centre is the end of the CHOH at the distance of 184.5 from Tide Lock and referred to as “Mile Post 184.5” in the CHOH map. We collected our water samples between these two extremes and chose Tide Lock as the bench mark. We used simple codes to differentiate water sample locations. For coding purposes we used alphabet “m” for miles and then numbers to represent actual distance from the bench mark (Tide Lock) and sampling year, for example, sample “m068.01/13” is 68 miles away from Tide Lock and was collected in 2013. If there were two or more samples collected from same location then they are represented by numbers in consecutive manner, such as sample m068.01 and sample m068.02. For quality assurance/quality control (QA/QC), water samples were collected in triplicates from selected surface water, spring water, and cave water for microbial community analyses. A total of 19 water samples collected in both years of 2012 (August) and 2013 (July). Ten water samples were collected in 2012 and nine water samples in 2013. A description of the water sample type and location is given in Table 3.1. For biomass collection, one-liter samples were filtered in the field through sterile, hydrophobic membrane bacteriological filters (0.45-µm pore size, 25-mm diameter, Millipore, USA) using a portable vacuum system (G 180 Portable Aspirator, Comco).

The filter samples were placed into microtubes and kept on ice until their storage in the lab at -80 °C prior for DNA extraction. Sample collection and handling were performed using sterile techniques.

3.2.2 Physical and Chemical Parameters Measurement

Various geochemical water quality parameters were being determined, including temperature, pH, electronic conductivity, alkalinity, inorganic C, Ca, K, Mg Na, Ni, P, Si, F^- Cl^-, NO_3^-, and SO_4^{2-}.

Temperature, pH, and electrical conductivity were measured in the field using a portable multiparameter meter equipped with various probes. A range of other water parameters were determined in the laboratory. For the measurement of all other parameters, the samples were frozen at -20°C until processing. Other parameters were measured in the laboratory using standard methods . Water quality analyses were the responsibility of the group of Dr. Vesper at West Virginia University (WVU). Part of the analyses were performed 'in house' (at WVU) and others, such as ICP-MS analyses were performed by commercial certified laboratories. In consequence, the analytical methods have not been further described in this thesis.

3.2.3 DNA Extraction, PCR, Pyrosequencing and Microbial Community's Analysis

3.2.3.1 DNA Extraction

The total genomic DNA was extracted from biomass-containing filters using the MOBIO PowerSoil DNA Extraction Kit (MOBIO Laboratories, Carlsbad, CA), following

the instruction of the manufacturer with some modifications. In particular, the filters were cut into about 1×1-mm^2 pieces before to be placed into the PowerBead tubes containing 1-mm (diameter) glass beads. After adding Solution C1 (lysis buffer), the PowerBead tubes were vigorously agitated on a Mini-BeadBeater-1 (Biospec product, Bartlesville, OK) at 4,200 rpm for 40 sec for complete homogenization and cell lysis. Then, solutions C2 and C3, which are patented inhibitor removal solutions, are used to remove organic and inorganic material. Next, solution C4, a concentrated salt solution, is used for binding DNA to the silica membrane. After DNA binding on the silica column, solution C5 is added, which is an ethanol-based solution for removal of residual salt, humic acids, and other contaminants, while allowing the DNA to stay bound to the silica membrane. Finally, the extracted and purified DNA was eluted with 100 μL of Solution C6 (sterile elution buffer) and stored at -80°C until further processing [52].

3.2.3.2 PCR Amplification of Bacterial, Archaeal, and Eukaryotic rRNA Genes

The extracted DNA was amplified by PCR for 16S rDNA gene using combined barcoded universal primers containing Roche 454 adapters. The following universal primers were designed to amplify hyper-variable regions of the ribosomal DNA: U-519F (5'-Barcode-CAGCMGCCGCGGTAATWC) and U-1086R (5'-Barcode-CTGACGRCRGCCATGC) for bacterial and archaeal 16s rDNA gene, A-519F (5'-Barcode-CAGCMGCCGCGGTAA) and A-1058R (5'-Barcode-GGCCATGCACCWCCTCTC) for archaeal 16s rDNA gene; E-343F (5'-Barcode-TACGGRAGGCAGCAG) and E-806R (5'-Barcode-GGACTACCAGGGTATCTAAT) eubacterial 16s rDNA gene. The PCR amplification was carried out on a StepOnePlus™

Real-Time PCR System (Applied Biosystems, Carlsbad, CA) using the HotStarTaq® Master Mix Kit (Qiagen, Foster City, CA). Cycling conditions were as follows: initial denaturation 10 min at 94°C; 40 cycles: denaturation 30 seconds at 94°C, annealing 30 seconds at 52°C, extension 45 seconds at 72°C; final extension 10 min at 72°C. After visualization by agarose gel electrophoresis, the PCR products flanked with adaptors were purified using the Agencourt Ampure PCR Purification system (Beckman Coulter, Beverly, MA) and quantified using a Nanodrop-2000. DNA extraction and PCR amplification were conducted in triplicate for each sample. A metagenomic library was then constructed by pooling equimolar concentrations of the 16S DNA products. The library was sequenced using a 454 GS FLX+ Pyrosequencing System (Roche Diagnostics, Indianapolis, IN).

3.2.3.3 Sequence Processing, Taxonomic Identity, Abundance, and Phylogeny

Pyrosequencing data was filtered and analyzed to determine the microbial diversity and the distribution of major taxonomic groups using the microbial ecology software package QIIME [47]. For each sample, the software identified the DNA sequences by comparison with sequences in the NCBI database [48], and the RDP [49]. The software calculated the diversity indices and the relative abundance of phylogenetic groups at various levels (i.e., species and genus).

Sequencing were assigned to samples by the 10-bp barcodes and grouped into phylotypes (≥ 97 sequence similarities). Operational taxonomic units (OTUs) were picked by Uclust and Cd-hit procedures. Representative set of sequences from each OTU were aligned with PyNAST application using Green Genes pre-aligned core set as a

template. Representative sequences were assigned phylogenetic taxa using RDP reference database for bacteria and archaea.

The diversity and richness of OTUs were further examined using rarefaction analysis. Rarefaction curves and different alpha-diversity indices (Shannon, Chao1, PD_whole_tree, and equitability) were calculated using QIIME.

3.2.3.4 Relationships between Metagenomic Biomarkers and Environmental Variables

An agglomerative hierarchical clustering method was used to analyze the relationships between environmental variables and metagenomics biomarkers. Hierarchical clustering groups objects in a hierarchy based on similarity between them. The output of the analysis is a dendrogram, a tree like structure which is a graphical representation of the resulting hierarchy. XLSTAT, a statistical analysis software was used for hierarchical clustering analysis .The XLSTAT software generated dendrograms after importing the relative abundance of the major prokaryotic phyla data and environmental variable data from Excel files.

PCA was conducted using the multivariate statistical analysis software package CANOCO 5. Both dependent variables; the relative abundance of major prokaryotic phyla (16) and environmental variables; water quality parameters (18) for the 19 samples were first imported in the software from Excel files. Unconstrained ordination analysis PCA was then conducted with focus on the relative abundance of prokaryotic phyla.

To directly examine how environmental variables (i.e., water quality parameters) related to bacterial community structure, RDA was applied using CANOCO 5 software

(Biometris, Wageningen, The Netherlands). All water geochemical parameters, nutrients, and metal concentrations data were firstly screened by calculating their parametric Pearson's correlations coefficients (see Table 4.4) and the significantly co-correlated variables ($|r| > 0.9$) were removed. Because CANOCO 5 software requests a set of explanatory variables that is significantly smaller than the number of samples in order to estimate the significance of the RDA regression – Monto Carlo permutation test (see Table 4.5) –, the remaining variables were then further reduced based on their microbial and water quality significance. The resulting filtered environmental variables were considered the most explanatory variables and they were used to examine their relationships with the microbial community composition and diversity through RDA. The unrestricted Monte Carlo permutation test (499 permutations) was used to test the statistical significance of the regression analysis. The significance of each individual environmental – explanatory -- variable that best described the most influential gradients was estimated by using the 'forward selection' option of the CANOCO 5 RDA procedure.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter presents and discusses the results obtained while conducting this metagenomic study. According to the specific objective of these, the first section covers the characterization of the microbial communities in the water samples collected from the CHOH using metagenomic pyrosequencing methods. Results include the relative abundance of major taxonomic groups observed in the samples and corresponding biodiversity indices, as well as further analysis of microbial community structures using clustering analysis and PCA.

The second section present results from the water quality analysis conducted on the water samples from the CHOH using standard methods: temperature, pH, conductivity, alkalinity, inorganic C, inorganic cations (Ca, K, Mg and Na), nutrients (NO_3- and PO_4^{3-}), and sulfate (SO_4^{2-}). These data are further discussed using clustering analysis and PCA.

The third section tries to establish meaningful relationships between metagenomic biomarkers (i.e., microbial community structure) and the water quality parameters, which were investigated by multivariate statistical analysis using RDA.

4.1. Metagenomic Analysis of the CHOH Water Samples

One-liter aliquots of water were collected from selected water bodies across the CHOH and filtered on site. The total DNA was extracted from the filters and used for PCR amplification of 16S rDNA fragments using universal primers. The resulting clone

library were sequenced on a Roche 454 pyrosequencing system. Pyrosequencing data were filtered and processed to determine the relative abundance of major taxonomic groups and biodiversity indices of prokaryotic bacteria in water samples by using the QIIME software package. Relationships and distances between the water sample locations with respect to the microbial community structure were investigated using clustering analysis (dendrogram) by using XLSTAT, a statistical software suite for Microsoft excel and by PCA using the CANOCO software package.

4.1.1 Relative Abundance of Bacterial Phyla in CHOH Water Samples

Sequences from pyrosequencing were assigned taxonomic affiliations using BLAST and the most abundant prokaryotic phyla that represents 1% or more of the total sequences detected in the water samples collected during the campaigns of years 2012 and 2013 are showed in Figure 4.1 and Figure 4.2, respectively. Other phyla with relative abundance less than 1% are classified as minor. The bacteria communities across all samples comprised about 60 phyla (99.5% of total sequences). In addition, 0.1% of total sequences were unclassified, which may belong to novel or not-well characterized phylotypes. The predominant bacteria phylum in all sites was *Proteobacteria,* which accounts for 55.4% of total classified sequences. Other dominant bacteria phyla in all sites were *Bacteroidetes* and *Actinobacteria*, comprising 12.4% and 10.6% of total classified sequences, respectively. As reported in many of previous studies, these bacteria phyla are also the dominant groups in fresh water and river sediments, in which they contribute significantly to essential biochemical processes [53, 54]. More detailed descriptions of the dominant phyla are in section 4.1.1.1.

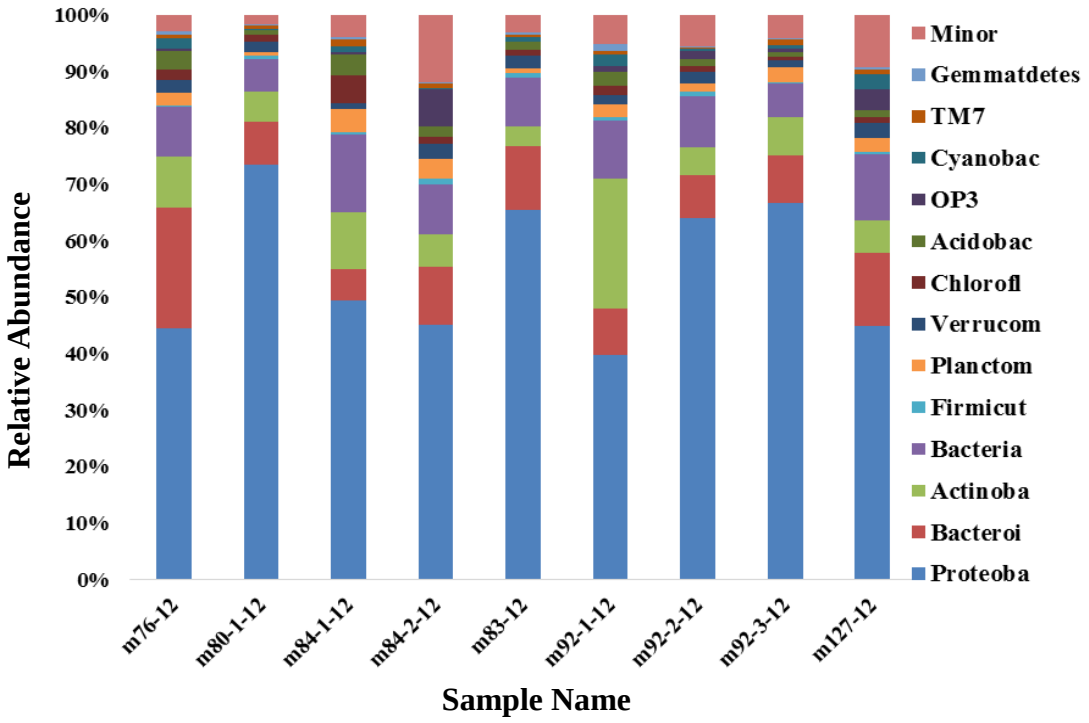

Figure 4.1 Prokaryotic phyla representing ≥ 1% of total sequence detected in all collected samples in year 2012.

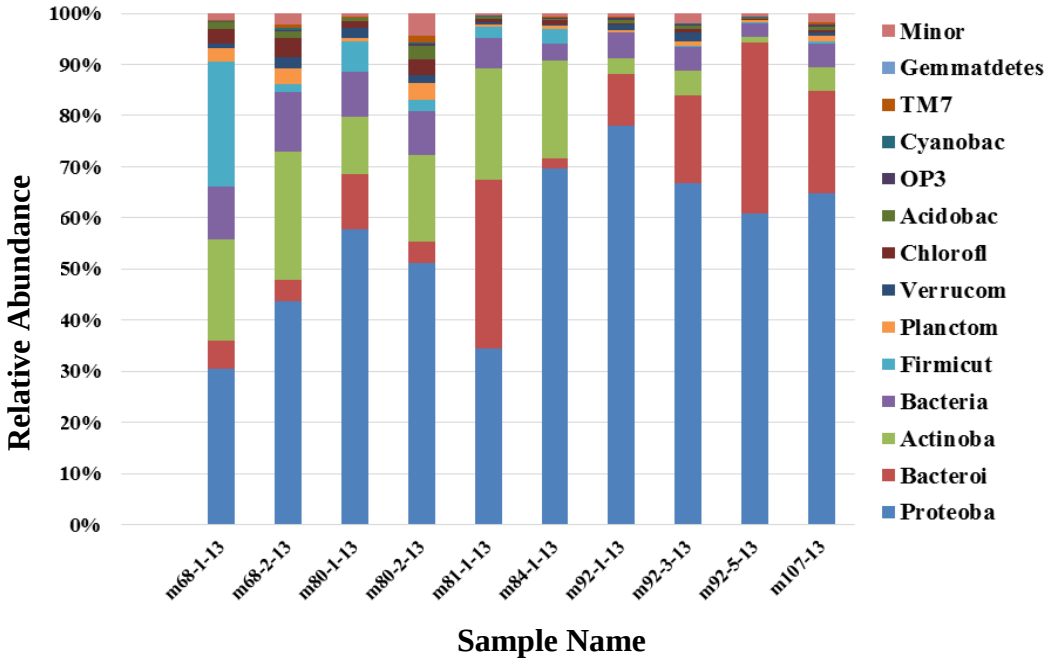

Figure 4.2 Prokaryotic phyla representing ≥ 1% of total sequence detected in all collected samples in year 2013.

4.1.1.1 Brief description of major prokaryotic phylum identified

Acidobacteria. Bacteria in the phylum *Acidobacteria* exist in a wide variety of environments. The phylum *Acidobacteria* is metabolically and genetically a diverse group. Although some studies suggested that oligotrophic environment is not favorable to *Acidobacteria* members, it seems copiotrophic environment may favor members of the phylum *Proteobacteria*, which is therefore detrimental to the phylum *Acidobacteria* [55]. As the name suggests, *Acidobacteria* include many acidophilic microorganisms.

Actinobacteria. Members of the phylum *Actinobacteria* are abundant in soil and freshwater. Although many soil *Actinobacteria* are filamentous, freshwater *Actinobacteria* are free-living, typically small bacteria, with a rod, coccus, or selenoid shape. Freshwater lake *Actinobacteria* are phototrophic or heterotrophic organism. *Actinobacteria* produce enzymes that help degrade organic plant material, lignin, chitin, and other recalcitrant compounds. Many are facultative anaerobes and grow best under anaerobic conditions [56].

Bacteroidetes. The phylum *Bacteroidetes* shows extensive phenotypic and metabolic diversity and are found in abundance in soil, in the aquatic environment, and as symbionts of plants, animals, and humans. Although, most *Bacteroidetes* are chemoheterotrophs, only few phototrophic have been described. Unlike other freshwater bacteria, *Bacteroidetes* do not show seasonal fluctuations, which are likely due to their dependency on organic matter or algal blooms [57] .

Chloroflexi. The members of the phylum *Chloroflexi* also have low trophic requirements as they are filamentous green non-sulfur bacteria, which perform

photosynthesis without producing oxygen – phototrophy. Isolated representatives of the *Choroflexi* display a broad range of phenotypes [58] *Chloroflexi* found in diverse habitats including geothermal springs [59], hypersaline mats [60], the deep subsurface [61], and in anaerobic environment [62]. *Chloroflexi* prefer photoorganoheterotrophic environments [63] and metabolize halogenated hydrocarbons by using them as electron acceptor for anaerobic respiration, and can therefore be indicative of contamination by chlorinated compounds [64]. *Chloroflexi* are more common in fresh water habitat [65].

Crenarchaeota: Members of the Crenachaeota phylum that were first described as sulfur-dependent extremophiles, although they have been shown to be abundant in the marine environment. Crenarchaea are today thought to be very abundant in soil and freshwater environments and to be a major contributor to carbon fixation. Members of Crenachaeota are autotrophic aerobic sulfur-oxidized organisms or heterotrophic anaerobic sulfate-reducers [66].

Cyanobacteria. *Cyanobacteria* (blue-green algae) are photosynthetic bacteria, populating the soil and aquatic environment. *Cyanobacteria* perform oxygenic photosynthesis using chlorophyll and phycobilisomes. *Cyanobacteria* perform nitrogen fixation sometimes involving specialized cell called heterocysts, and play a key role in nutrient cycling in freshwater. *Cyanobacteria* are susceptible to release toxins in water, especially in the case of algal blooms induced by excess of nutrients [66].

Firmicutes. *Firmicutes* include large numbers of anaerobic and photosynthetic bacteria, as well as important pathogens. The abundance of *Firmicutes* may indicate contamination with fecal matter as the group includes many enteric bacteria.

Gemmatimonadetes. The members of the *Gemmatimonadetes* phylum have been found in soil, freshwater, and sediments. A wide range of environments where *Gemmatimonadetes* have been found suggests an adaptation to low soil moisture. The phylum *Gemmatimonadetes* is similar to but distinct from the phylum *Cyanobacteria*. The metabolic pathways and enzymes of known members of the phylum are unique and allow it to grow under both aerobic and anaerobic conditions. Studies show that *Gemmatimonadetes* are found infrequently in freshwater body [67].

Planctomyces. *Planctomycetes* are characterized by a peptidoglycan-less cell wall and budding mode of reproduction. *Planctomyces* are abundant in freshwater, marine, and soil environment and in association with invertebrate animals. *Planctomyces* are chemoheterotrophs or chemoautotrophs. Heterotrophic members can use diverse source of carbon substrates, such as N-acetylglucosamine. Autotrophic 'anammox' *planctomycetes* are capable of performing anaerobic oxidation of ammonium. Their metabolic diversity allow them to populate diverse habitats [68].

Nitrospira. *Nitrospira* bacteria have been isolated from ocean water, freshwater, sediments, and soils. Members of the *Nitrospira* phylum oxidize ammonia into nitrate also known as nitrification, which is an important process of global nitrogen cycle. In freshwater bodies, *Nitrospira* are predominantly responsible for nitrification [69]. Because of their autotrophic character, *Nitrospira* are slow-growing organisms and difficult to maintain [70]. Their generation times could reach up to 90 h for *Nitrospira* in marine [71]. *Nitrospira* are also believed to play an important role in carbon fixation [72].

Proteobacteria. The phylum *Proteobacteria* includes Gram-negative bacteria, the most studied microorganisms, which are useful in agricultural, industrial, and medical

fields. The phylum includes six classes, among which α-, β-, γ-*Proteobacteria* are the major ones [57]. The class α-*Proteobacteria* is found in a wide diversity of habitats and exhibits a wide variety of life-styles. Many members of *Proteobacteria* play an important role in global nitrogen cycle, because symbionts performs atmospheric nitrogen fixation in plants (e.g., Rhizobiales). α-*Proteobacteria* are often dominant in the marine ecosystem. Freshwater α-*Proteobacteria* are also widely distributed and they are known to be resistant to grazing and compete well under low-nutrient conditions. Some species are capable of degrading complex compounds. β-*Proteobacteria* are often dominant in freshwater lakes. Members of the phylum are fast growing under high nutrient conditions. Their abundance depends on the water depth, pH, and availability of carbon substrate.

Verrucomicrobia. The phylum *Verrucomicrobia* is ubiquitous in soil and it has been recently considered as one of the most abundant bacterial phyla in the environment. Most *Verrucomicrobia* are free-living, mesophilic, facultative obligate anaerobic, and oligotrophic. *Verrucomicrobia* are commonly found in extreme conditions such as hot springs, fumaroles, and in Antarctica [73]. Other than the *Proteobacteria*, *verrucomicrobia* is the only phylum contains aerobic methonotrophs [73, 74]. The phylum also has an important role of nitrogen fixation in soil [75]. Members of Verrucomicrobia have frequently been isolated form acidophilic and thermophilic environments (e.g., hot springs). Members of the phylum can oxidize methane.

OP3. Many members of the Obsidian Pool phylum (OP) are found in different extremophilic environments, such as geothermal and mineral rich soils, deep subsurface aquifers, and hydrothermal vents, although they have also been observed in other

environmental niches, including mesophilic lake water and flooded paddy soils. In particular, members of OP3 phylum are found in many anoxic environments including marine sediments, hypersaline deep sea, fresh water lakes, aquifers, cave seepage waters, flooded paddy environments and methanogenic reactors [76].

TM7. Candidate phylum TM7 is a newly described bacterial group exclusively characterized by environmental DNA sequences, which indicates that TM7 members are found in diverse terrestrial, aquatic, and clinical environments [39].

4.1.2 Diversity of Microbial Communities in CHOH Samples

The pyrosequencing process generated over 240,000 reads in total of 19 samples for both 2012 and 2013 sampling campaigns. After filtration, 201,746 reads longer than 200 bp and shorter than 1,000 bp – based on 97% similarities – were identified, with an average reads length of 526 bp. The number of reads per sample ranged from 1,409 to 9,989. The Good's coverage – which is used as an estimator for sequencing completeness, varied from 40% to 82% with an average of 34%.

4.1.2.1 Alpha Diversity

The alpha diversity is defined as the diversity of organisms in each sample. To determine the alpha diversity, various estimators have been suggested for decades [77, 78]. The alpha diversity is commonly measured by three methods: rarefaction curve, species richness estimators – e.g., Chao1 index–, which are often used together with rarefaction curves, and community diversity indices, such as Shannon index. Several diversity indices covering these three methods are available in QIIME to determine the alpha diversity. For species richness, the estimator Chao I is one of the most commonly used indices .Chao1

index relies on the number of rare OTUs found in a sample [79]. The formula used in QIIME pipeline for Chao 1 estimation is:

$$S_{Chao1} = S_{obs} + \frac{{n_1}^2}{2n_2}$$

Where S_{chao1} is the estimated number of species, S_{obs} is the observed number of species, n_1 is the number of singleton taxa (taxa represented by a single read in that community), and n_2 is the number of doubleton taxa. Samples containing large number of singletons favor the greater possibility of more undetected OTUs and thus Chao1 Index will estimate greater species richness than the sample without rare OTUs. Chao 1 value in individual samples indicated that m068-01 has the highest richness while site m080-01 has the lowest richness. Furthermore the observed species (represented by OTUs) in sample m068-01 is highest in number supporting the Chao1 value for the sample.

The alpha diversity could also be measured by community diversity indices which combine species richness and abundance into one value. Shannon-Weiner index is one of the commonly used community diversity indices, which was coined by Shannon and Weiner (also known as Shannon's diversity index) [80]. The Shannon index is calculated as follows:

$$H = -\sum_{b=1}^{A}(c_i * lnc_i)$$

Where H is the Shannon diversity index, c_i is the fraction of the entire population made up of the b^{th} species, and A is the total number of species in the sample [81].

The H value represents the uncertainty of the species composition (entropy or degree of surprise). Two samples with overall the same abundance value and equal numbers of representative species, but in different proportion, would have different H values. The sample with evenly distributed species would have high H value, while sample with uneven species distribution would have low H value. So the H value indicates not only the number of species but how the abundance of the species is distributed among all the species in the community. Shannon index integrates species richness and evenness into a single parameter [82]. The metric PD_Whole Tree is the Faith's Phylogenetic Diversity and it is based on the phylogenetic tree. Basically, it adds up all the branch lengths and measure the diversity. Closely related OTUs will bring a small increase in diversity. On the other hand, unalike OTUs contribute large increase in diversity [52].

Table 4.1 shows the alpha diversity estimation of all water samples based on different diversity indices. Shannon index value of the site m068-02 is the highest (10.69), showing the highest diversity and evenness of species, while site m080 has the lowest value (5.218), showing the lowest richness and uneven distribution of species. Site m068.01/13 shows the highest value of PD whole tree (344) and thus the highest diversity, while sample m080.1 has the lowest PD whole (71) and thus lowest alpha diversity. Equitability of all samples ranges from 0.60 to 0.95, indicating high evenness of species in these samples.

Table 4.1 Alpha diversity indices explaining richness and evenness with in samples

Site	Shannon	Chao 1	PD Whole Tree	Observed Species	Equitability	Good Coverage %
m080.01/12	5.21	1430.03	71	1525	0.933	88.93
m092.01/13	6.25	5284.489	119.6	1356	0.601	82.8
m083.02/12	6.66	2862.903	119	1694	0.932	67.92
m076.01/12	6.72	3077.143	140	2500	0.911	75.58
m083.03/12	6.83	3041.519	134	2437	0.927	72.46
m083.01/12	6.94	3361.504	136	2289	0.947	68.97
m092.01/12	6.94	3380.203	153.5	2621	0.933	73.2
m092.02/12	6.95	3807.503	166	2708	0.926	70.24
m092.03/12	7.12	6588.515	194	2955	0.933	67.71
m127.01/12	7.15	5160.003	198.6	3366	0.926	70.946
m084.01/13	8.76	12151.58	238.1	3279	0.75	72.33
m092.05/13	9.36	11860.62	207.5	2602	0.825	60.15
m080.01/13	9.66	4771.311	172	2089	0.876	80.21
m107.01/13	9.98	10188.69	191	2347	0.891	56.52
m081.01/13	10.04	9827.872	185.4	2162	0.906	56.8
m080.02/13	10.32	9538.695	189.5	1843	0.951	46.05
m068.01/13	10.67	20518.07	344.5	4866	0.871	58.35
m068.02/13	10.69	11598.56	226	2425	0.951	45.53
m092.03/13	10.8	12519.1	233.7	2404	0.961	42.07

4.1.2.2 Rarefaction Curve

Rarefaction curves (Figure 4.3 and Figure 4.4) are also used to estimate the alpha diversity. A rarefaction curve plots the number of species as a function of the number of individuals sampled. At start, curve begins with a steep slope, which at some point flatten as fewer species are being discovered in each additional sample. Samples with gentler slope of rarefaction curves, contributes less to the total number of OTUs [83].

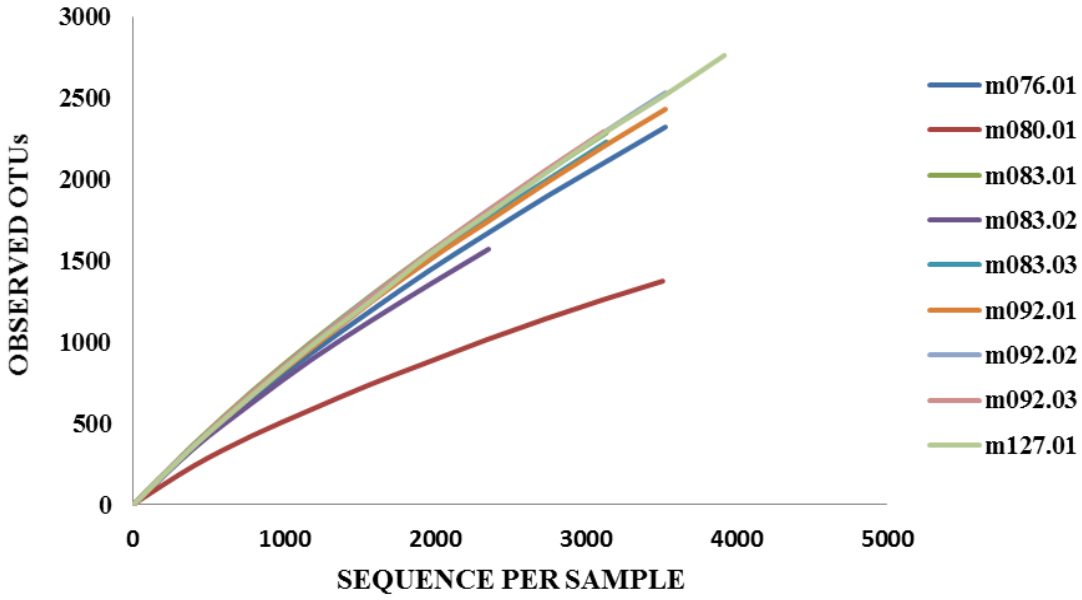

Figure 4.3 Rarefaction curve at 97% sequence similarlity indicating the observed OTUs in 2012 samples.

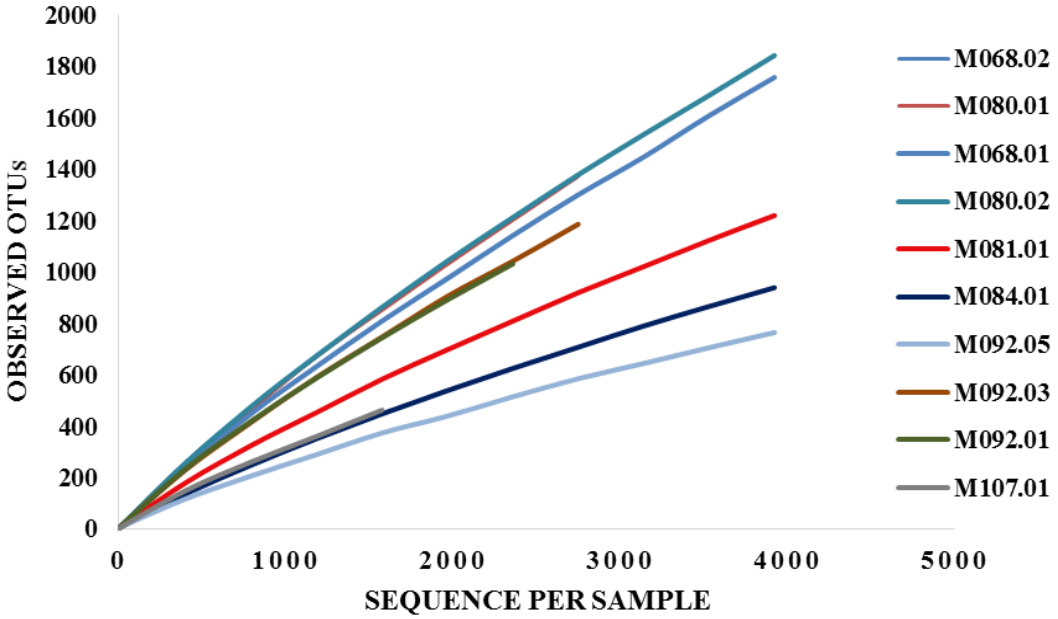

Figure 4.4 Rarefaction curve at 97% sequence similarity indicating the observed OTUs in 2013 samples.

QIIME software can perform many rarefaction analyses to estimate the alpha diversity systematically. QIIME performs many rarefactions at multiple depths and repeat many times at each depths. For this purpose, QIIME picks the OTU table and generates a new folder, which contains many OTU tables, all of which are repeats of rarefactions at specific depths. These OTU tables used by QIIME to generate rarefaction curves. QIIME pipeline generated the rarefaction curves for all the alpha diversity indices (Shannon, Chao1, Observed OTUs, etc.). In Figure 4.4, the gentle rarefaction curves of samples in year 2013 show that the species (OTUs) in the sample are easily determined. On the other hand, steeper slopes of rarefaction curves in show that in 2012 samples estimation of number of species (OTUs) are difficult. In other words for rarefaction curves at 97% of similarity phylotype level for most samples in 2012 were not saturated, indicating that the bacteria diversities in water body were probably higher than those found in 2012 water

samples. In such cases, large number of sequences per samples will be needed to complete a diversity survey of water sample.

Microbial diversity indices show that all 2012 samples are characterized by a low biodiversity, while 2013 samples are characterized by a high diversity, which is likely the result of different meteorological conditions in 2012 and 2013. One exception is sample m092-01-2013 which is one of the two stream samples collected (m092-01 and m080-01).

Surprisingly, across the two sampling campaigns, the diversity was generally higher in spring samples than in stream, cave, and mine samples. The stream and cave samples, m092-01, m080-01, m083-03, m092-02, are characterized by a low diversity (below 7), while the mine sample, m127-01, is slightly higher (7.15). One exception is the stream sample m080-01-2013 which shows a higher diversity (9.66). This observation suggests that microorganisms collected from spring water may have developed under specific conditions (e.g., anaerobic, autotrophic, extremophilic) typically expected in the deep subsurface – k-type microorganisms –, while microorganisms collected from surface water (streams, caves, and mine pools) are more heterotrophic aerobic organisms performing aerobic oxidation of organic carbon – r-type microorganisms.

4.1.3 Cluster Analysis

To group the water samples according to their similarity in relative abundance of phyla, cluster analysis method was used. Cluster analysis is a multivariate method in which objects/samples are subdivided into groups (clusters) such that the more similar

objects make individual groups. There are two main clustering methods: (1) Hierarchical methods and (2) k-means clustering methods.

Hierarchical clustering methods are further subdivided into agglomerative methods and divisive methods. In agglomerative hierarchical methods, similar objects make a clusters. Next, closest clusters join to each other to make a bigger cluster, engulfing samples of both adjoining clusters. This process is repeated until all clusters join together to make one cluster.

Figure 4.5 shows a dendrogram which is generated by XLSTAT software package by using Agglomerative Hierarchical Clustering (AHC) command after importing relative abundance data of each sample into the software. The dendrogram shows three groups of cluster, differentiated by different line colors. Group 1 (Top left) clustered four sampling site .Sampling site m068-2-13 and m068-1-13 are at same location and thus clustered together. However, sampling sites in group 3 (bottom left) are heterogeneous with respect to geological formation, but represent more homogeneity in term of bacteria phylum then the other groups because it is flatter on the dendrogram. Although there are some variations, nearly all the samples collected in each year cluster together separately, when compared based on the prokaryotic phyla, suggesting a relative stability of water quality. The relationship is further elaborated in PCA section.

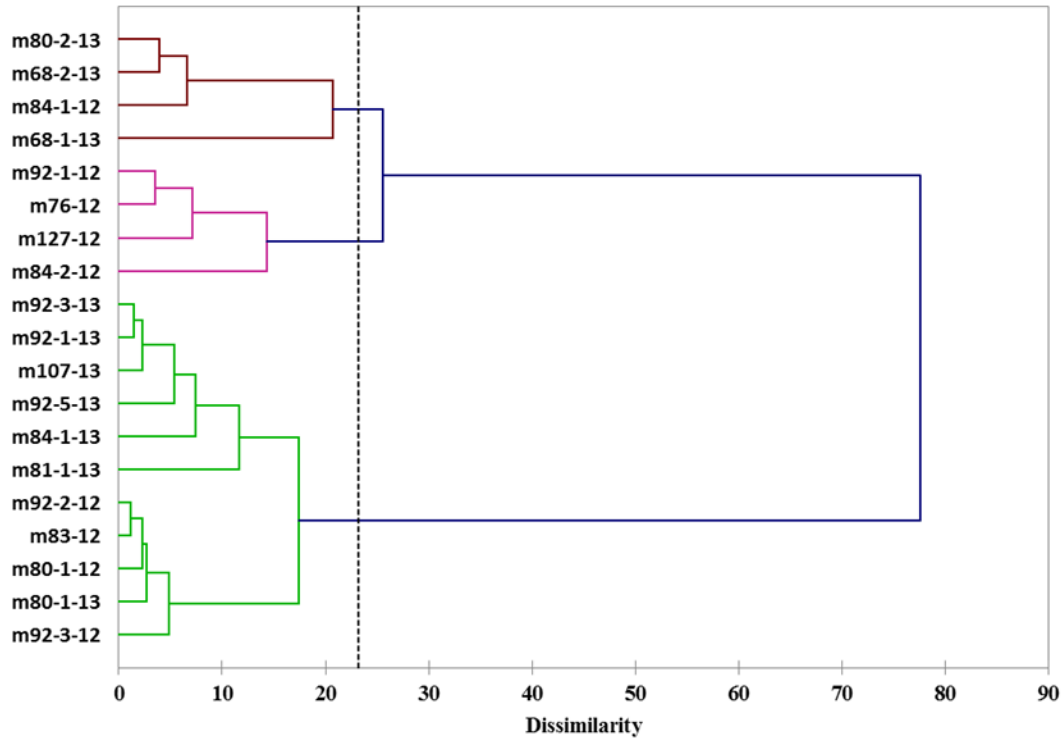

Figure 4.5 Dendrogram showing relationship between microbial community structures and water sample locations**.**

4.1.4 Principal Component Analysis (PCA)

PCA was conducted using the multivariate statistical analysis software package CANOCO 5 In short, both dependent variables; the relative abundance of major prokaryotic phyla (16) and environmental variables; water quality parameters (18) for the 19 samples were first imported in the software from Excel files. Unconstrained ordination analysis PCA was then conducted with focus on the relative abundance of prokaryotic phyla. Figure 4.6 shows the major microbial groups identified in the samples (i.e., prokaryotic phyla that represent 1% or more of the total sequences detected in all samples).

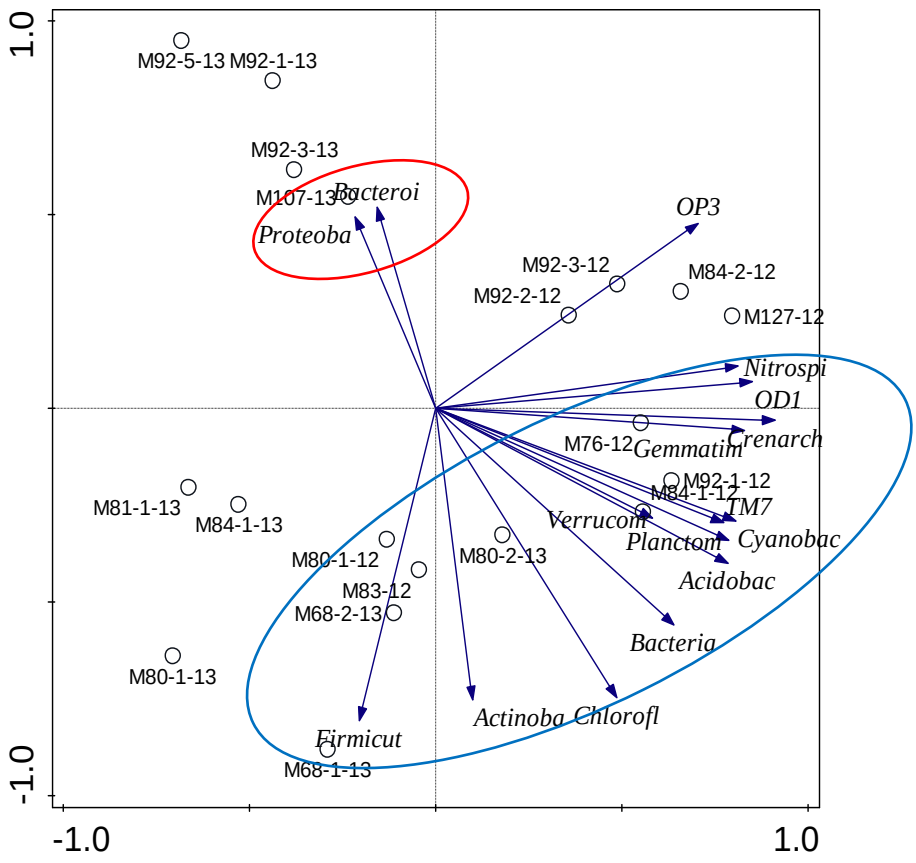

Figure 4.6 Ordination diagram (PCA) with focus on environmental variables showing the relationships between 16 major prokaryotic phyla and 19 sampling sites.

Dominant phyla in most natural environments are *Acidobacteria, Actinobacteria, Bacteroidetes, Cyanobacteria*, *Firmicutes*, *Proteobacteria*, and *Verrucomicrobia* [57, 84]. Newton et al [56] reported that major bacterial groups observed in freshwater lakes are commonly, *Proteobacteria* – especially β-*Proteobacteria* --, *Actinobacteria, Bacteroidetes, Cyanobacteria*, and *Verrucomicrobia*, which are well represented in our samples. Minor phyla include, among others, *Acidobacteria, Chloroflexi*, *Firmicutes, Gemmatimonadetes,* and *Nitrospira*, which were also observed in our samples.

Traditionally, trophic parameters are the best predictors of microbial abundance in the environment, which is well reflected on the ordination diagram (Figure 4.6).

Proteobacteria and *Bacteroides* cluster together, which is consistent with their copiotrophic character – fast heterotrophic metabolism consuming labile organic matter. *Proteobacteria* and *Bacteroidetes* are traditionally considered to be the dominant bacterial phyla in aquatic ecosystem with β-*Proteobacteria* usually predominant in lakes and rivers and α- and γ-*Proteobacteria* dominant in marine systems [85]. These two groups demark clearly from other groups shown on Figure 4.6, which include other major bacterial phyla observed on the environment, i.e., *Acidobacteria, Firmicutes*, and *Actinobacteria*, which is characterized as oligotrophs – slow growers using less biodegradable carbon sources [84]. Fierer et al [84] proposed that bacterial phyla can be classified into ecological categories on the basis of r-type or K-type character: r-type bacteria are growing rapidly and develop where resources are abundant, K-types bacteria grow at lower rate but have higher substrate affinities, allowing them to be competitive in anaerobic and/or oligotrophic environments. *Planctomyces* are non-extreme bacteria similar to *Actinobacteria* and, to which they appear close in Figure 4.6.

Unlike comparison based on water quality parameters, comparison based on microbial community structures shows less consistency between samples collected from the same site during the two sampling campaigns, i.e., m080-01, m084-01, m092-01, and m092-03. Interestingly, all sites north of site m084-1/2 (mileage 84 and higher), with the exception of m076, cluster together and exhibit higher abundance of *Proteobacteria* and *Bacteroides*, indicating higher abundance of easily metabolizable organic matter.

The ordination diagram shows high dependence of microbial communities to the general chemistry water quality parameters, pH, temperature, conductivity, alkalinity, and inorganic carbon. One may observe a negative correlation between most microbial groups

and the conductivity, reflecting the freshwater character of microbial species. Strong positive correlations were observed between *Proteobacteria* and pH and temperature, which is consistent with the neutrophile and mesophile character of the group.

4.2. Analysis of Water Quality Parameters in the CHOH Water Samples

Temperature, pH, conductivity, and alkalinity were measured in the field. Other parameters, including inorganic C, inorganic cations (e.g., Ca, K, Mg Na), nutrients (e.g., NO_3^- and PO_4^{3-}), sulfate (e.g., SO_4^{2-}), and selected metals (e.g., Fe, Cu, Mg) were measured in the laboratory. Water quality analyses were performed by the group of Dr. Vesper at West Virginia University. Relationships and distances between the water samples with respect to water quality parameters were investigated using clustering analysis (denderogram) using XLSTAT and by PCA using the CANOCO software package. Water quality parameters values are given in Table 4.2 and Table 4.3.

Table 4.2 Water quality parameters (geophysical and nutrient)

Sampling Sites	pH	Temp (°C)	EC (mS/cm)	Alk (meq/L)	Inorg C (ppm C)	NO_3^- (ppm)	PO_4^{3-} (ppm)	SO_4^{2-} (ppm)
m76-12	6.50	12.4	0.59	5.60	56.41	7.77	0.00	16.83
m80-1-12	7.38	17.8	0.70	4.88	64.13	5.84	0.00	30.74
m84-1-12	6.51	13.0	0.70	6.18	71.00	7.89	0.00	16.63
m84-2-12	6.56	12.6	0.68	5.96	70.99	6.96	0.00	16.71
m83-12	7.97	25.41	0.49	3.76	57.74	10.06	2.49	26.33
m92-1-12	7.62	21.86	0.55	5.52	59.88	2.32	0.00	20.09
m92-2-12	6.93	17.59	0.72	6.84	68.77	1.86	0.00	20.57
m92-3-12	5.74	12.07	0.71	6.58	86.00	2.32	0.00	19.58
m127-12	5.66	11.01	0.63	4.96	74.53	0.00	0.00	93.76
m68-1-13	6.85	12.66	0.66	5.88	69.31	4.91	0.12	14.43
m68-2-13	6.58	12.24	0.61	5.20	65.62	4.87	0.08	13.65
m80-1-13	7.67	18.52	0.67	5.84	69.09	0.00	0.05	32.61
m80-2-13	6.33	12.09	0.78	6.48	72.56	3.20	0.10	27.33
m81-1-13	6.81	12.64	0.64	5.42	66.81	4.80	0.10	22.64
m84-1-13	6.67	12.91	0.78	6.14	70.72	0.00	0.12	17.88
m92-1-13	7.73	20.90	0.66	5.08	64.97	2.74	0.00	20.90
m92-3-13	6.28	11.90	0.83	7.08	75.82	0.00	0.16	20.18
m92-5-13	6.75	13.36	0.88	7.10	75.93	4.11	0.09	18.98
m107-13	6.91	12.21	0.64	4.50	61.82	0.00	0.09	15.86

Table 4.3 Water quality parameters (Metals)

Sampling Site	µg/L								ppm	
	Ba	**Ca**	**K**	**Mg**	**Na**	**Si**	**Sr**	**Zn**	**F**	**Cl**
m76-12	112.00	80108	3283	23279	5442	7726	295	4.00	0.29	13.96
m80-1-12	65.00	84900	3534	21248	15644	6755	235	0.00	0.28	29.82
m84-1-12	85.00	123318	3208	11531	8950	8256	1015	10.00	0.21	19.68
m84-2-12	84.00	123994	3251	11556	8920	8296	1017	7.00	0.21	19.82
m83-12	27.00	50914	5642	21957	8749	6104	240	18.00	0.09	12.84
m92-1-12	109.50	91302	4433	12170	5163	8077	351	9.00	0.12	13.39
m92-2-12	132.00	134085	2378	12714	4503	6376	382	13.00	0.12	13.39
m92-3-12	105.00	133234	2359	12638	4428	6357	379	11.0	0.14	13.83
m127-12	98.00	77975	2844	32121	5165	5119	6420	12.00	0.18	3.77
m68-1-13	37.00	62609	3214	29575	2973	3816	48	7.00	0.11	6.60
m68-2-13	41.00	62386	3133	29438	2976	3797	48	12.00	0.05	6.00
m80-1-13	65.00	86220	3549	18195	14304	5437	221	10.00	0.27	33.30
m80-2-13	101.00	94205	2148	19365	9734	5592	211	14.00	0.20	20.60
m81-1-13	130.00	94013	2039	9459	7012	5403	1200	32.00	0.10	18.30
m84-1-13	116.00	112369	3129	10163	8397	6700	955	26.00	0.17	21.10
m92-1-13	126.00	86195	4399	11009	4382	6217	341	122.0	0.08	12.00
m92-3-13	101.00	12188	2092	11163	4276	5058	359	12.00	0.06	13.40
m92-5-13	132.00	126696	5048	11091	4845	5833	391	24.00	0.07	9.50
m107-13	87.00	79157	2004	9508	8050	5643	290	18.00	0.13	22.90

4.2.1 Cluster Analysis

Figure 4.7 shows a dendrogram which is generated by XLSTAT software package by using Agglomerative Hierarchical Clustering (AHC) command after importing 18 water quality parameters of each water sample into the software .The dendrogram shows that the first four sites sampled during both sampling campaigns 2012 and 2013, i.e., m080-01, m084-01, m092-01, and m092-03 cluster closely together, indicating the relative temporal stability of water quality over the years. Similarly, most of the sites located physically close to each other also show similar water quality, i.e., m068-01 and m068-02, m084-01 and m084-02, and m092-02, m092-03, and m092-5 (although not m092-01).Furthermore, water samples m068-01, m068-02 and m127-01 are all clustered together (except m083-01) in one group (top left) in dendrogram. The geological formation of these three water samples are clastic (non-carbonate). Thus having lower concentration of calcium (Table 4.3).Except few, the results show that the samples collected from the same sites in different years cluster together when compared based on the water quality parameters.

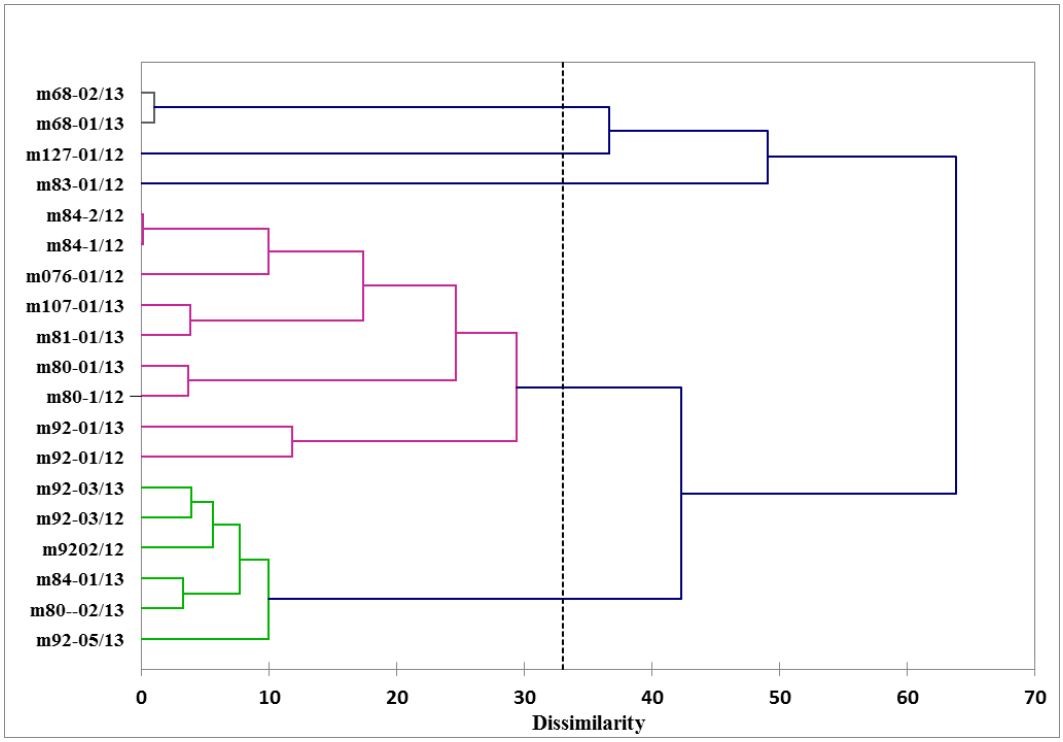

Figure 4.7 Dendrogram of water quality parameters of all samples.

4.2.2 Principal Component Analysis (PCA)

As seen in clustering analysis, ordination diagram (PCA) in Figure 4.8 shows that the four sites sampled during both sampling campaigns 2012 and 2013, i.e., m080-01, m084-01, m092-01, and m092-03 cluster closely together, indicating the relative temporal stability of water quality over the years. Similarly, most of the sites located physically close to each other also show similar water quality, i.e., m068-01 and m068-02, m084-01 and m084-02, and m092-02, m092-03, and m092-5 (although not m092-01).

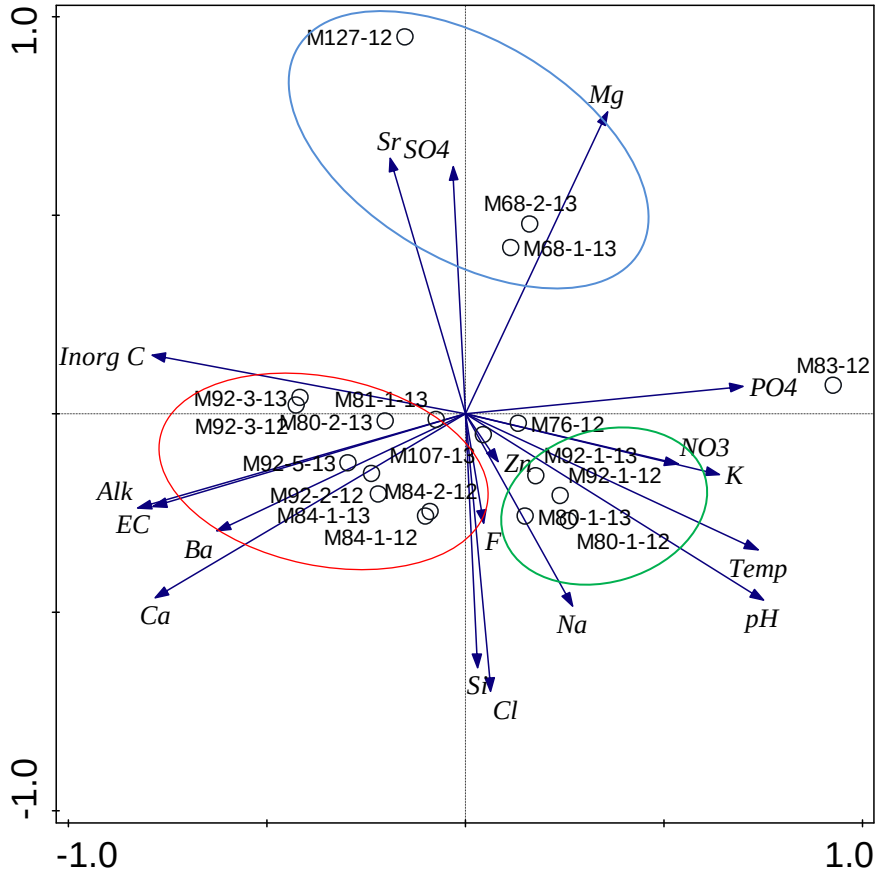

Figure 4.8 Ordination diagram (PCA) with focus on environmental variables showing the relationships between 18 water quality parameters and the 19 sampling sites.

The relationships between environmental variables are as expected showing positive correlations between inorganic carbon, alkalinity, conductivity, and calcium, between phosphate, nitrate, and pH, and between chloride and sodium. Negative correlations were observed between alkalinity and pH. Although most samples consisted in spring water, a few four were collected from stream and caves. One sample, m127, was collected from a stagnant pool in an abandoned mine and demarks clearly from the rest of the samples. Similarly, m083, was collected from a stagnant pool in a cave and differs significantly from other samples. On the other hand, two other samples, collected

from caves, m092-1 and m092-2, do not show much differences by comparison with other spring samples. Geological data may also explain the relationships between water quality and sampling sites. Sites m076 to m107 (mileage 76 to 107) are located in carbonate formations, while sites 068-1/2 and m127 (mileage 068 and 127) are located in clastic formations (Figure 4.9). In fact, sites 068-1/2 are located at the edge of carbonate and clastic formations.

Indeed, many but not all samples in the carbonate formation, from sites m076 to m107, are associated with high alkalinity, inorganic carbon, and calcium, as expected in association with carbonate minerals. Particularly, spring and cave samples located in the carbonate formation, m80-2, m81-1, m84-1/2, m107, and m92-2/3/5 (red envelop) correlate well with alkalinity, inorganic carbon, and calcium. Two other sites located in the carbonate formations, m80-1 and m92-1 (green envelop) are stream samples, which correlate with high nutrients, potassium, nitrate, and phosphate, pH, and temperature. On the contrary, sites m068 and m127 (blue envelop) which are outside or at the very edge of the carbonate formation correlate with sulfate and magnesium.

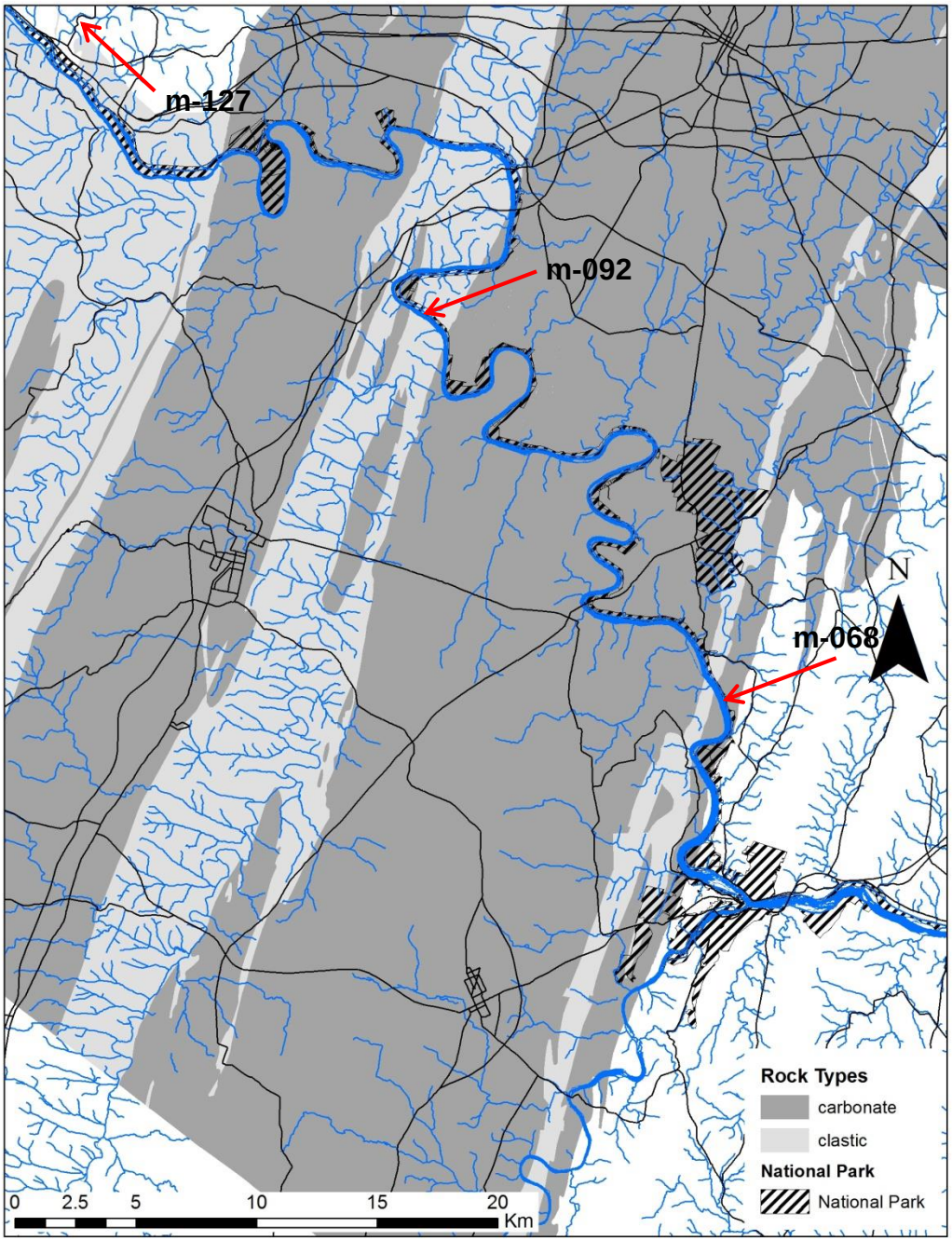

Figure 4.9 Overview map of the study area showing bedrock types: clastic – non-carbonate (light grey) and carbonate (dark grey).

.

4.3. Relationships between Microbial Community Structure and Water Quality Parameters in the CHOH Water Samples

The dependence of the prokaryotic relative abundance on the water quality parameters for all the 19 samples was interrogated using RDA using the multivariate statistical analysis software package CANOCO 5 [50]. Both dependent variables – relative abundance of major prokaryotic phyla (16) – and environmental variables – water quality parameters (18) – for the 19 samples were imported in the software from Excel files. Constrained ordination analysis RDA was then conducted with focus on metagenomic data.

Significant RDA requests reducing the number of environmental – explanatory variables –, which was performed first by removing collinear variables. Collinear variables were identified by generating a Pearson's correlation matrix (Prism 6.0, GraphPad): Two collinear variables – with $r \geq 0.9$ –, chloride and strontium, were removed (see Table 4.5). Then significant explanatory variables were further selected based on the microbial significance and the maximum number of variables allowed when conducting RDA analysis, i.e., 7 explanatory variables for 19 samples. (In fact, CANOCO 5 software could not estimate the significance of the RDA regression – Monto Carlo permutation test – with more than 7 explanatory variables for 19 samples). Several combinations were manually tested and we selected the most meaningful one based on the microbial significance and statistical significance of the test (based on Monte Carlo permutations). The ordination diagram presented in Figure 4.10 includes general chemistry parameters (pH, temperature, conductivity, and alkalinity) and nutrient

parameters (nitrate, phosphate, and sulfate) most susceptible to affect the microbial communities. The RDA significance test on all axes gave an F-value of 2.2 and a *p*-value of 0.004. The significance of the individual environmental – explanatory – variables was estimated by using the 'forward selection' option of the CANOCO 5 RDA procedure (Table 4.4)

Table 4.4 Summary analysis of the RDA conducted on 7 selected explanatory variables using the option 'forward selection'. Adjusted P-values were calculated using the 'false discovery rate' method

Simple Term Effects					Conditional Term Effects				
variable	**Explains %**	**pseudo-F**	**P**	**P (adj)**	**variable**	**Explains %**	**pseudo-F**	**P**	**P (adj)**
pH	12.0	2.30	0.066	0.406	**pH**	12.0	2.30	0.068	0.159
EC	10.9	2.10	0.116	0.406	**Temp**	19.1	4.40	0.008	0.056
Temp	6.7	1.20	0.294	0.525	**EC**	11.7	3.10	0.020	0.070
NO_3^-	6.6	1.20	0.300	0.525	**PO_4**	5.30	1.40	0.234	0.328
SO_4^{2-}	5.1	0.90	0.514	0.664	**NO_3**	5.50	1.50	0.146	0.256
PO_4^{3-}	4.60	0.80	0.664	0.664	**Alk**	2.80	0.80	0.588	0.686
Alk	3.50	0.60	0.610	0.664	**SO_4**	2.20	0.60	0.738	0.738

Summary Table

Statistic	Axis1	Axis2	Axis3	Axis4
Eigen value	0.399	0.079	0.074	0.019
Explained variation (cumulative)	39.91	47.85	55.26	57.17
Psuedo-canonical correlation	0.891	0.708	0.776	0.698
Explained filled variation (cumulative)	68.17	81.37	94.39	97.650

Analysis: Constrained-Method: RDA
Total Variation is 355.17109,
Explanatory variables account for 58.5%
(Adjusted explained variation is 32.2%)

Table 4.5 Pearson correlation matrix by Prism 6.0 (GraphPad) identified two collinear explanatory variables

	pH	Temp	EC	Alk	Inorg C	Ba	Ca	K	Mg	Na	Si	Sr	Zn	F	Cl	NO_3	PO_4
Temp	0.88																
EC	-0.40	-0.49															
Alk	-0.46	-0.45	0.81														
Inorg C	-0.67	-0.53	0.69	0.68													
Ba	-0.25	-0.20	0.45	0.51	0.25												
Ca	-0.40	-0.30	0.70	0.81	0.65	0.65											
K	0.63	0.71	-0.29	-0.32	-0.37	-0.24	-0.29										
Mg	-0.14	-0.07	-0.38	-0.36	-0.16	-0.62	-0.68	0.10									
Na	0.36	0.24	0.01	-0.24	-0.17	-0.24	-0.06	0.06	-0.09								
Si	0.12	0.23	-0.10	0.06	-0.19	0.33	0.43	0.22	-0.48	0.29							
Sr	-0.48	-0.28	-0.09	-0.17	0.24	0.14	-0.05	-0.14	0.35	-0.09	-0.06						
Zn	0.33	0.31	0.01	-0.16	-0.08	0.35	-0.05	0.25	-0.30	-0.23	-0.05	-0.05					
F	-0.05	-0.08	-0.06	-0.06	-0.13	-0.06	0.02	-0.11	0.16	0.70	0.48	0.12	-0.37				
Cl	0.34	0.15	0.13	-0.04	-0.12	-0.04	0.14	-0.15	-0.39	0.90	0.34	-0.33	-0.15	0.63			
NO_3	0.22	0.23	-0.39	-0.30	-0.40	-0.39	-0.24	0.44	0.20	0.10	0.33	-0.25	-0.13	0.14	-0.08		
PO_4	0.43	0.58	-0.45	-0.51	-0.35	-0.50	-0.44	0.53	0.16	0.10	-0.07	-0.11	-0.02	-0.24	-0.10	0.47	
SO_4	-0.29	-0.07	-0.12	-0.25	0.18	0.01	-0.20	-0.04	0.48	0.10	-0.18	0.91	-0.07	0.21	-0.19	-0.29	0.00

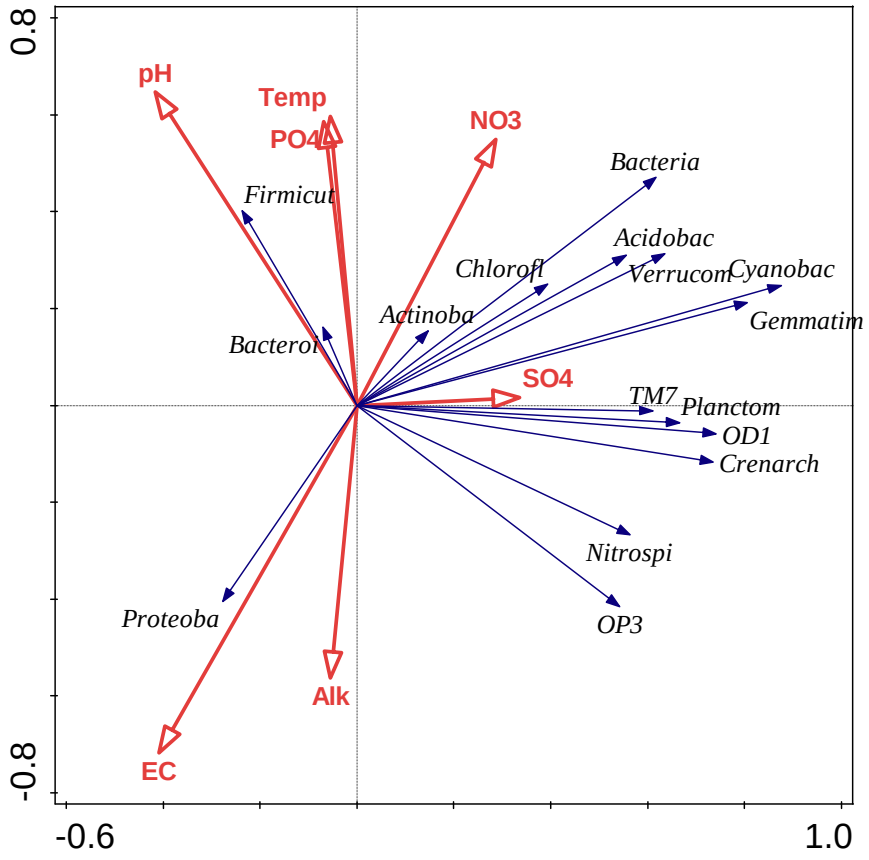

Figure 4.10 Ordination diagram generated from RDA showing the dependence of the relative abundance of prokaryotic communities (16 phyla) on water quality parameters (7 parameters) for the 19 sampling sites.

As seen in prior PCA investigations, Figure 4.8 shows that the r-type phyla, *Proteobacteri*a and *Bacteroides*, cluster together. The *Proteobacteria* phylum correlates with high alkalinity and conductivity, while the *Bacteroides* phylum (and *Firmicutes* phylum) correlates with high pH, temperature, phosphate, and nitrate. Altogether, these parameters suggest a copiotrophic environment, which is consistent with the dominance of *Proteobacteria* and *Bacteroides* [56]. The phylum *Firmicutes* include mostly aerobic or anaerobic (denitrifying) microorganisms, which may consistent with its association with the two other r-type phyla, as well as with nitrate. On the contrary, K-type phyla,

including *Acidobacteria, Actinobacteria, Chloflexi, Cyanobacteria, Crenarchaeota, Gemmatimonadetes, Planctomyces, Nitrospira, Verrucomicrobia, OD1, OP3, and TM7,* cluster together and correlate collectively with sulfate and nitrate, which are major electron-acceptor for anaerobes. Sulfate and Nitrate are important product which are generated during autotrophic processes in the subsurface where nitrate is generated by using ammonia and sulfate by reduced sulfur as electron donors. Members of *Actinobacteria* are facultative anaerobes using nitrate and sulfate as electron-acceptor, which is consistent with the correlation observed on Figure 4.10. The *Acidobacteria* phylum includes bacteria adapted to low nutrient conditions and capable to perform nitrate reduction, which may also explain the correlation with nitrate [86]. Moreover, *Acidobacteria* are genetically similar to *Cyanobacteria*, which can be observed in Figure 4.10. Similarly, the phylum *Gemmatimonadetes* include nitrate-reducers and is similar to the phylum *Cyanobacteria*. Members of the phylum *Verrucomicrobia* have frequently been isolated from acidic and thermophilic environments from the subsurface. The phylum also shares similarities with the phylum *Planctomycetes. Cyanobacteria* and *Chloroflexi* phyla include autotrophic photosynthetic bacteria – therefore with low nutrient requirements, which is consistent with the oligotrophic character of the cluster.

In addition, the two phyla correlate with nitrate and sulfate as the growth of these organisms are typically regulated by the abundance of nitrogen and phosphate. Many members of the *Planctomycetes* phylum are capable to use highly diversified carbon sources, which allow them to have a competitive advantage in oligotrophic environments – K-type organisms. Members of *Planctomyces* are also known to perform anaerobic oxidation of ammonia using nitrate as electron-acceptors ('anammox' metabolism). The

full sequencing of a *Planctomyces* isolate has revealed the presence of multiple sulfatase genes, indicating that members of the phylum have the capability to reduce sulfate. This may explain the strong correlation between the phylum and sulfate. Members of the *Crenachaeota* phylum include sulfur-dependent extremophiles that are either autotrophic aerobic sulfur-oxidizers or heterotrophic anaerobic sulfate-reducers, therefore explaining its strong correlation with sulfate. Members of the TM7 have been detected in sulfidic marine environment where oxidation of organic matter occurs through sulfate reduction (the Black Sea), potentially explaining the association between this phylum and sulfate [87]). Similarly, members of the OD1 phylum include sulfate reducers, which likely explain their correlation with sulfate. Members of the *Nitrospira* phylum are known to oxidize ammonia into nitrate. However, members of the phylum are also known to use sulfate as electron acceptor, therefore explaining the correlation with sulfate. Members of the *OP3* phylum have been isolated from various anoxic environments and are known to perform sulfate oxidation, which is consistent with the correlation of the phylum with sulfate. For instance, *OP3* bacteria have been isolated from deep-sea sulfate reducing communities involved in alkane oxidation in marine environment [13].

In addition to the relationships presented in ordination diagram in Figure 4.11 also shows the relationships between sampling sites with respect to prokaryotic phyla and water quality parameters.

As expected, the copiotrophic, fast-growing, r-type phyla, *Proteobacteria*, *Bacteroides*, and *Firmicutes*, correlate with the stream samples, m80-1 and m92-1 (upper-right quadrant). On the contrary, oligotrophic, slow-growing, K-type phyla, correlate better with spring, cave, and mine samples, m76, m84-1/2, m92-2/3, and m127

(left side). The presence of many autotrophic and anaerobic extremophiles in these phyla suggests that the water flows were originating from the deep subsurface characterized by high temperatures, low pH, low oxygen, and low organic carbon. On the other hand, five spring samples, m80-2, m84-1, m92-3/5, and m107, located in the carbonate formation, correlate strongly with *Proteobacteria*, conductivity, and alkalinity (lower-right quadrant).

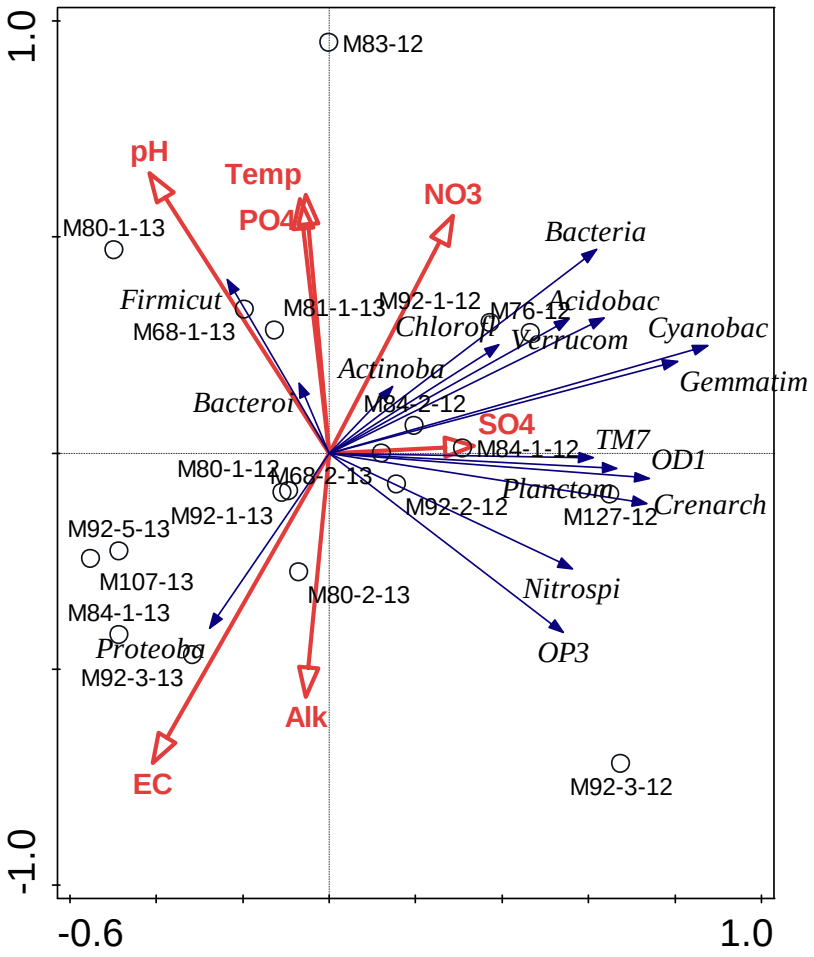

Figure 4.11 Ordination diagram generated from RDA showing the dependence of the relative abundance of prokaryotic communities (16 phyla) on water quality parameters (7 parameters) and sampling sites (19 samples).

CHAPTER 5

CONCLUSIONS

Phylogenetic analyses conducted on the metagenomic data show that the dominant prokaryotic phyla detected in the 19 samples are similar the ones typically detected in freshwater environments*: Proteobacteria, Actinobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia, Acidobacteria, Chloroflexi, Firmicutes, Gemmatimonadetes,* and *Nitrospira*. Microbial diversity indices show that all 2012 samples are characterized by a low biodiversity, while 2013 samples are characterized by a high diversity, which is likely the result of different meteorological conditions in years 2012 and 2013. Across the two sampling campaigns, the diversity was generally higher in the spring samples than in the stream, cave, and mine samples, suggesting that microorganisms collected from spring water may have developed under specific conditions in the deep subsurface, while microorganisms collected from surface water are more heterotrophic aerobic organisms performing aerobic.

Clustering analysis and PCA were conducted to investigate the relationships between the relative abundance of the 16 major identified prokaryotic phyla and the 18 water quality parameters monitored in 19 water samples collected in the summer 2012 and 2013. The results show that the samples collected from the same sites in different years cluster together when compared based on the water quality parameters. On the contrary, the samples collected in each year cluster together, respectively, when compared based on the prokaryotic phyla, suggesting again a relative stability of water

quality parameters over the years, although the microbial communities are subjected to a larger variability.

PCA focusing on prokaryotic phyla shows that *Proteobacteria* and *Bacteroides* phyla cluster together, which is consistent with their copiotrophic character – r-type bacteria. On the other hand, the other phyla (e.g., *Acidobacteria, Actinobacteria, Chloflexi, Cyanobacteria, Crenarchaeota,* etc.) form a distinct cluster, which is consistent with their oligotrophic character – r-type microorganisms, including autotrophs, anaerobes, and extremophiles. High dependence of the microbial communities to general chemistry water quality parameters (e.g., pH, temperature, conductivity, alkalinity, and inorganic carbon) was observed.

The relationships between environmental variables show positive correlations between inorganic carbon, alkalinity, conductivity, and calcium, between phosphate, nitrate, and pH, and between chloride and sodium. The type of water sampled (i.e., spring, cave, stream, mine) and geological formation (e.g., carbonate or clastic formations) seems to explain the relationships between water quality and sampling sites. Indeed, stream, cave, and mine samples cluster separately from other spring water samples. Similarly, most spring and cave samples in the carbonate formation are associated with high alkalinity, inorganic carbon, and calcium, as expected in association with carbonate minerals. Stream samples located in the carbonate formations correlate with high nutrients, potassium, nitrate, and phosphate, pH, and temperature. Samples collected from the clastic formation correlate sulfate and magnesium.

The dependence of the prokaryotic relative abundance on the water quality parameters for all the 19 samples was interrogated using RDA. As showed by PCA

investigations, the r-type phyla, *Proteobacteria*, *Bacteroides*, and *Firmicutes* cluster together and correlates with high alkalinity and conductivity (*Proteobacteria*) and with high pH, temperature, phosphate, and nitrate (*Bacteroides* and *Firmicutes*), suggesting that members of these phyla are more abundant copiotrophic environments. On the contrary, K-type phyla (e.g., *Acidobacteria*, *Actinobacteria*, *Chloflexi, Cyanobacteria*, etc.) cluster together and correlate collectively with sulfate and nitrate, which are major electron-acceptor for anaerobes and important product generated by autotrophic processes in the subsurface using ammonia and reduced sulfur as electron donors. Members of the *Actinobacteria, Acidobacteria,* and *Gemmatimonadetes* phyla frequently include nitrate-reducer, which likely explain the correlation with nitrate. *Cyanobacteria* and *Chloroflexi* phyla include autotrophic photosynthetic bacteria, which is consistent with the oligotrophic character of the cluster and explain the correlation with nitrate and sulfate. Members of the *Planctomycetes* phylum are capable to use highly diversified carbon sources, which allow them to have a competitive advantage in oligotrophic environments – K-type organisms. Members of the *Crenachaeota, Nitrospira, Planctomyces*, *TM7, OD1, OP3* phyla are known to perform sulfate oxidation, which is consistent with the correlation with sulfate. As expected, the copiotrophic, fast-growing, r-type phyla, *Proteobacteria, Bacteroides,* and *Firmicutes,* correlate with the stream samples, while the oligotrophic, slow-growing, K-type phyla, correlate better with spring, cave, and mine samples. The presence of many autotrophic and anaerobic extremophiles in these phyla suggest that the water flows were originating from the deep subsurface characterized by high temperatures, low pH, low oxygen, and low organic carbon.

This research constitutes one of the first efforts to use molecular microbiology techniques and multivariate analyses to find out meaningful relationships between microbial community structure and water quality parameters in surface water and to investigate whether one can use the bacterial community structure as a water quality indicator. The result showed that bacterial communities have considerable potential to be used as a sensitive indicator of water quality. The study revealed that the diversity of the microbial communities depends on the type of water bodies, suggesting the strong impact of environmental conditions on the bacterial community structure. The multivariate analyses (PCA and RDA) further confirm our hypothesis because certain types of bacteria (k type, r type) showed a strong correlation with the general water quality parameters (pH, NO_3^-, alkalinity) as shown in the bi-plots. Although, DNA pyrosequencing is an expensive method, it would still be cost effective as compared to physico-chemical methods to analyze the quality of water as it allows to integrate the major water quality parameters into one single index, i.e., the bacterial community structure.

REFERENCES

1. Hahn, M.E. and J.J. Stegeman. *Molecular Biology and Biotechnology in Marine Toxicology*. in *Opportunities for Environmental Applications of Marine Biotechnology:: Proceedings of the October 5-6, 1999, Workshop*. 2000. National Academies Press.

2. DEP, P., *Pennsylvania's Nonpoint Source Management Program Update*. 2008. p. 87.

3. Barbour, M.T., et al., *Rapid bioassessment protocols for use in streams and wadeable rivers.* USEPA, Washington, 1999.

4. Rodrigues, D.F., D.P. Jaisi, and M. Elimelech, *Toxicity of functionalized single-walled carbon nanotubes on soil microbial communities: implications for nutrient cycling in soil.* Environmental science & technology, 2012. **47**(1): p. 625-633.

5. Baltar, F., et al., *Microbial functioning and community structure variability in the mesopelagic and epipelagic waters of the subtropical northeast atlantic ocean.* Applied and environmental microbiology, 2012. **78**(9): p. 3309-3316.

6. Wang, Q., et al., *Using microbial community functioning as the complementary environmental condition indicator: A case study of an iron deposit tailing area.* European Journal of Soil Biology, 2012. **51**: p. 22-29.

7. Aguilera, A., et al., *Eukaryotic community distribution and its relationship to water physicochemical parameters in an extreme acidic environment, Rio Tinto (southwestern Spain).* Applied and environmental microbiology, 2006. **72**(8): p. 5325-5330.

8. Sun, M.Y., et al., *Bacterial communities are sensitive indicators of contaminant stress.* Marine pollution bulletin, 2012. **64**(5): p. 1029-1038.

9. First, M.R. and J.T. Hollibaugh, *Environmental factors shaping microbial community structure in salt marsh sediments.* Marine Ecology Progress Series, 2010. **399**: p. 15-26.

10. Röling, W.F., et al., *Relationships between microbial community structure and hydrochemistry in a landfill leachate-polluted aquifer.* Applied and Environmental Microbiology, 2001. **67**(10): p. 4619-4629.

11. Liu, Z., et al., *Phylogenetic diversity, composition and distribution of bacterioplankton community in the Dongjiang River, China.* FEMS microbiology ecology, 2012. **80**(1): p. 30-44.

12. Gauch, H.G., *Multivariate analysis in community ecology*. 1982: Cambridge University Press.

13. Kleindienst, S., et al., *Diverse sulfate-reducing bacteria of the Desulfosarcina/Desulfococcus clade are the key alkane degraders at marine seeps.* The ISME journal, 2014. **8**(10): p. 2029-2044.

14. Khan, F., T. Husain, and A. Lumb, *Water quality evaluation and trend analysis in selected watersheds of the Atlantic region of Canada.* Environmental Monitoring and assessment, 2003. **88**(1-3): p. 221-248.

15. Abbasi, T. and S. Abbasi, *Water quality indices based on bioassessment: The biotic indices.* Journal of water and health, 2011. **9**(2): p. 330-348.

16. de Zwart, D., *Monitoring water quality in the future, Volume 3: Biomonitoring.* 1995.

17. Custer, C.M., T.W. Custer, and P.M. Dummer, *Patterns of organic contaminants in eggs of an insectivorous, an omnivorous, and a piscivorous bird nesting on the Hudson River, New York, USA.* Environmental Toxicology and Chemistry, 2010. **29**(10): p. 2286-2296.

18. Hayes, T.B., et al., *Hermaphroditic, demasculinized frogs after exposure to the herbicide atrazine at low ecologically relevant doses.* Proceedings of the National Academy of Sciences, 2002. **99**(8): p. 5476-5480.

19. Maine Department of Environmental Protection. *Why Biological Monitoring*. 2013 [cited 2015 March-20]; Available from: https://www1.maine.gov/dep/water/monitoring/biomonitoring/why.htm.

20. ribosomal Database Project. *Seqmatch-start*. 1992 [cited 2015; Available from: http://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp.

21. Joubert, G., *A bioassay application for quantitative toxicity measurements, using the green algae Selenastrum capricornutum.* Water Research, 1980. **14**(12): p. 1759-1763.

22. Parks, J., et al., *Young northern pike, yellow perch and crayfish as bioindicators in a mercury contaminated watercourse.* Environmental monitoring and assessment, 1991. **16**(1): p. 39-73.

23. Córdova-Kreylos, A.L., et al., *Diversity, composition, and geographical distribution of microbial communities in California salt marsh sediments.* Applied and Environmental Microbiology, 2006. **72**(5): p. 3357-3366.

24. Gauthier, F. and F. Archibald, *The ecology of "fecal indicator" bacteria commonly found in pulp and paper mill water systems.* Water Research, 2001. **35**(9): p. 2207-2218.

25. Al-Awadhi, H., et al., *Bias problems in culture-independent analysis of environmental bacterial communities: a representative study on hydrocarbonoclastic bacteria.* SpringerPlus, 2013. **2**(1): p. 369.

26. Vaz-Moreira, I., et al., *Culture-dependent and culture-independent diversity surveys target different bacteria: a case study in a freshwater sample.* Antonie Van Leeuwenhoek, 2011. **100**(2): p. 245-257.

27. RK, R. and B. CA, *Encyclopedia of food microbiology*. 2nd ed. 1999: Academic Press.

28. Cocolin, L., et al., *The late blowing in cheese: a new molecular approach based on PCR and DGGE to study the microbial ecology of the alteration process.* International journal of food microbiology, 2004. **90**(1): p. 83-91.

29. Andreote, F.D., J.L. Azevedo, and W.L. Araújo, *Assessing the diversity of bacterial communities associated with plants.* Brazilian Journal of Microbiology, 2009. **40**(3): p. 417-432.

30. Hugenholtz, P. and N.R. Pace, *Identifying microbial diversity in the natural environment: a molecular phylogenetic approach.* Trends in biotechnology, 1996. **14**(6): p. 190-197.

31. Girones, R., et al., *Molecular detection of pathogens in water–the pros and cons of molecular techniques.* Water research, 2010. **44**(15): p. 4325-4339.

32. Tringe, S.G., et al., *The airborne metagenome in an indoor urban environment.* PloS one, 2008. **3**(4): p. e1862.

33. Cocolin, L., Dolci Paola, Rantsiou, Kalliopi, *Molecular methos dor the identification of microorganisms in traditional meat products*, in *Meat Biotechnology*. 2008, Springer Science & Business Media. p. 93.

34. Muyzer, G., *DGGE/TGGE a method for identifying genes from natural ecosystems.* Current opinion in microbiology, 1999. **2**(3): p. 317-322.

35. Nocker, A., M. Burr, and A.K. Camper, *Genotypic microbial community profiling: a critical technical review.* Microbial ecology, 2007. **54**(2): p. 276-289.

36. Hou, W., et al., *A comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing*. PloS one, 2013. **8**(1): p. e53350.

37. Tärnberg, M., et al., *Identification of randomly selected colonies of lactobacilli from normal vaginal fluid by pyrosequencing of the 16S rDNA variable V1 and V3 regions.* Apmis, 2002. **110**(11): p. 802-810.

38. Hans-Ju«rg Monstein , S.N.-B.a., Jon Jonasson a, *Rapid molecular identi¢cation and subtyping of Helicobacter pylori by pyrosequencing of the 16S rDNA variable V1 and V3 regions.* FEMS Microbiology Letters, 2001.

39. Hugenholtz, P., *Exploring prokaryotic diversity in the genomic era.* Genome Biol, 2002. **3**(2): p. 1-0003.8.

40. Kröber, M., et al., *Phylogenetic characterization of a biogas plant microbial community integrating clone library 16S-rDNA sequences and metagenome sequence data obtained by 454-pyrosequencing.* Journal of Biotechnology, 2009. **142**(1): p. 38-49.

41. Maxam, A.M. and W. Gilbert, *A new method for sequencing DNA.* Proceedings of the National Academy of Sciences, 1977. **74**(2): p. 560-564.

42. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proceedings of the National Academy of Sciences, 1977. **74**(12): p. 5463-5467.

43. Ronaghi, M., *Pyrosequencing sheds light on DNA sequencing.* Genome research, 2001. **11**(1): p. 3-11.

44. Ranasinghe, P.D., et al., *Revealing microbial community structures in large-and small-scale activated sludge systems by barcoded pyrosequencing of 16S rRNA gene.* Water Science & Technology, 2012. **66**(10): p. 2155-2161.

45. Ronaghi, M., et al., *Real-time DNA sequencing using detection of pyrophosphate release.* Analytical biochemistry, 1996. **242**(1): p. 84-89.

46. Petrosino, J.F., et al., *Metagenomic pyrosequencing and microbial identification.* Clinical chemistry, 2009. **55**(5): p. 856-866.

47. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nature methods, 2010. **7**(5): p. 335-336.

48. National Center of Biotechnology Information. *Analyze*. 2015 [cited 2015 01-12]; Available from: http://www.ncbi.nlm.nih.gov/home/analyze.shtml.

49. Ribosomal Database Project. *Seqmatch-start*. 1992 [cited 2015 01-12]; Available from: http://rdp.cme.msu.edu/seqmatch/seqmatch_intro.jsp.

50. Lepš, J. and P. Šmilauer, *Multivariate analysis of ecological data using CANOCO*. 2003: Cambridge university press.

51. Adathakula, S., D, Venkata.K, *Anomaly Detection Using Principal Component Analysis.* International Journal of Computer Science and Telecommunications 2014. **5**(4).

52. BIO, M. *Power Soil DNA Isolation Kit*. 2010 [cited 2015 03-01]; Available from: http://www.mobio.com/.

53. Araya, R., et al., *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, 2003. **43**(1): p. 111-119.

54. Yergeau, E., et al., *Next-Generation Sequencing of Microbial Communities in the Athabasca River and Its Tributaries in Relation to Oil Sands Mining Activities.* Applied and Environmental Microbiology, 2012. **78**(21): p. 7626-7637.

55. Lee, S.-H., J.-O. Ka, and J.-C. Cho, *Members of the phylum Acidobacteria are dominant and metabolically active in rhizosphere soil.* FEMS microbiology letters, 2008. **285**(2): p. 263-269.

56. Newton, R.J., et al., *Phylogenetic ecology of the freshwater Actinobacteria acI lineage.* Applied and environmental microbiology, 2007. **73**(22): p. 7169-7176.

57. Newton, R.J., et al., *A guide to the natural history of freshwater lake bacteria.* Microbiology and Molecular Biology Reviews, 2011. **75**(1): p. 14-49.

58. Björnsson, L., et al., *Filamentous Chloroflexi (green non-sulfur bacteria) are abundant in wastewater treatment processes with biological nutrient removal.* Microbiology, 2002. **148**(8): p. 2309-2318.

59. Boomer, S.M., et al., *Molecular characterization of novel red green nonsulfur bacteria from five distinct hot spring communities in Yellowstone National Park.* Applied and environmental microbiology, 2002. **68**(1): p. 346-355.

60. Nübel, U., et al., *Diversity and Distribution in Hypersaline Microbial Mats of Bacteria Related to Chloroflexusspp.* Applied and environmental microbiology, 2001. **67**(9): p. 4365-4371.

61. Chandler, D., et al., *Phylogenetic diversity of archaea and bacteria in a deep subsurface paleosol.* Microbial Ecology, 1998. **36**(1): p. 37-50.

62. Sekiguchi, Y., et al., *In situ detection, isolation, and physiological properties of a thin filamentous microorganism abundant in methanogenic granular sludges: a novel isolate affiliated with a clone cluster, the green non-sulfur bacteria,*

subdivision I. Applied and environmental microbiology, 2001. **67**(12): p. 5740-5749.

63. Overmann, J., *Green nonsulfur bacteria.* eLS, 2008.

64. Zinder, S., *Anaerobic utilization of halohydrocarbons*, in *Handbook of Hydrocarbon and Lipid Microbiology*. 2010, Springer. p. 2049-2064.

65. Gich, F., J. Garcia-Gil, and J. Overmann, *Previously unknown and phylogenetically diverse members of the green nonsulfur bacteria are indigenous to freshwater lakes.* Archives of microbiology, 2001. **177**(1): p. 1-10.

66. Madigan, M.T., et al., *Brock Biology of Microorganisms 13th edition*. 2010: Benjamin Cummings.

67. DeBruyn, J.M., et al., *Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil.* Applied and environmental microbiology, 2011. **77**(17): p. 6295-6300.

68. Fuerst, J., *Planctomycetes: a phylum of emerging interest for microbial evolution and ecology.* World Federation for Culture Collections Newsletter, 2004(38): p. 1-11.

69. Hovanec, T.A., et al., *Nitrospira-like bacteria associated with nitrite oxidation in freshwater aquaria.* Applied and environmental microbiology, 1998. **64**(1): p. 258-264.

70. Off, S., M. Alawi, and E. Spieck, *Enrichment and physiological characterization of a novel Nitrospira-like bacterium obtained from a marine sponge.* Applied and environmental microbiology, 2010. **76**(14): p. 4640-4646.

71. Watson, S.W., et al., *Nitrospira marina gen. nov. sp. nov.: a chemolithotrophic nitrite-oxidizing bacterium.* Archives of Microbiology, 1986. **144**(1): p. 1-7.

72. Daims, H., et al., *In Situ Characterization ofNitrospira-Like Nitrite-Oxidizing Bacteria Active in Wastewater Treatment Plants.* Applied and environmental microbiology, 2001. **67**(11): p. 5273-5284.

73. Islam, T., et al., *Methane oxidation at 55 C and pH 2 by a thermoacidophilic bacterium belonging to the Verrucomicrobia phylum.* Proceedings of the National Academy of Sciences, 2008. **105**(1): p. 300-304.

74. Dunfield, P.F., et al., *Methane oxidation by an extremely acidophilic bacterium of the phylum Verrucomicrobia.* Nature, 2007. **450**(7171): p. 879-882.

75. Khadem, A.F., et al., *Nitrogen fixation by the verrucomicrobial methanotroph 'Methylacidiphilum fumariolicum'SolV.* Microbiology, 2010. **156**(4): p. 1052-1059.

76. Kumar, M.R. and V. Saravanan, *Candidate OP phyla: importance, ecology and cultivation prospects.* Indian journal of microbiology, 2010. **50**(4): p. 474-477.

77. Whittaker, R.H., *Evolution and measurement of species diversity.* Taxon, 1972: p. 213-251.

78. Magurran, A.E., *Measuring biological diversity.* African Journal of Aquatic Science, 2004. **29**(2): p. 285-286.

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

80. Shannon, C.E., *A mathematical theory of communication.* ACM SIGMOBILE Mobile Computing and Communications Review, 2001. **5**(1): p. 3-55.

81. Magurran, A.E., *Why diversity?*, in *Ecological Diversity and Its Measurement*. 1988, Springer. p. 1-5.

82. Magurran, A.E., *Ecological diversity and its measurement*. 2013: Springer Science & Business Media.

83. Wooley, J.C., A. Godzik, and I. Friedberg, *A primer on metagenomics.* PLoS Comput Biol, 2010. **6**(2): p. e1000667.

84. Fierer, N., M.A. Bradford, and R.B. Jackson, *Toward an ecological classification of soil bacteria.* Ecology, 2007. **88**(6): p. 1354-1364.

85. Stevens, H., et al., *Phylogeny of Proteobacteria and Bacteroidetes from oxic habitats of a tidal flat ecosystem.* FEMS microbiology ecology, 2005. **54**(3): p. 351-365.

86. Ward, N.L., et al., *Three genomes from the phylum Acidobacteria provide insight into the lifestyles of these microorganisms in soils.* Applied and Environmental Microbiology, 2009. **75**(7): p. 2046-2056.

87. Leloup, J., et al., *Diversity and abundance of sulfate-reducing microorganisms in the sulfate and methane zones of a marine sediment, Black Sea.* Environmental Microbiology, 2007. **9**(1): p. 131-142.