A DRINKING WATER MICROBIOME FROM SOURCE TO TAP: COMMUNITY DIVERSITY, FUNCTIONALITY, AND MICROBIAL INTERACTION

BY

YA ZHANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Environmental Science in Civil Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Wen-Tso Liu, Chair
Dr. Mark W. LeChevallier, American Water
Professor Benito J. Mariñas
Professor Thanh H. Nguyen

ABSTRACT

Despite the long history of water research, understanding drinking water microbiome continuum spanning from source water, treatment in the production process, distribution network, and up to the point where water enters a building is rather challenging owing to the complexity in community assembly, water matrices, physical structure, and chemical gradients from source to tap. Previous studies on drinking water microbiomes have primarily investigated “who are there?” and “how do they change over time and across space?” in selected stages of drinking water systems. However, it is important to ask additional questions that include but are not limited to “what are they doing?”, “why are they there?” and more critically “who is doing what?”, and “what are the interrelationships among them, and between them and their environment?”. To answer these questions, it requires not only the advent of new methods, but also the transformation of drinking water microbiology from a descriptive discipline to a hypothesis-driven science that attempts to elucidate mechanisms with the intention to predict and shape the microbiome continuum.

The studies included in this dissertation resolved the ecological patterns of a groundwater-sourced drinking water microbiome at different scales. At the community level, the treatment process could be viewed as ecological disturbances on the drinking water microbiome continuum over space in the system by combining 16S rRNA gene amplicon sequencing and metagenomics. Abstraction caused a substantial decrease in both the abundance and number of functional genes related to methanogenesis and syntrophs in raw water. The softening process reduced microbial diversity and selected an *Exiguobacterium*-related population, which was attributed to its ability to use the phosphotransferase system (PTS) as regulatory machinery to control the energy conditions of the cell. After disinfection and entering the distribution system, microbial populations and their functions remained relatively stable. Predation by eukaryotic populations could be another disturbance to the bacterial microbiome, which could further drive the diversification of the bacterial community.

At the population level, nine draft genomes of pathogen-related species from the genera *Legionella*, *Mycobacterium*, *Parachlamydia,* and *Leptospira* were constructed and characterized in relation to their abundance, diversity, potential pathogenicity, genetic exchange, and distribution across the groundwater-sourced drinking water system. The presence/absence of specific virulence machinery could be effectively used to determine the pathogenicity potential of these genomes. Clustered regularly interspaced short palindromic repeats-CRISPR-associated proteins (CRISPR-Cas) genetic signatures were identified as a potential biomarker in the monitoring of *Legionella* related strains across different drinking water systems.

At the multi-species level, methano-/methylo-trophs were investigated, which were overlooked populations dominant and prevalent in drinking water microbiomes of groundwater systems. Using genome-resolved metagenomics, 34 methylotroph-related draft genomes were recovered together with another 133 draft genomes belonging to a variety of taxa. Both Type I and Type II methanotrophs dominated the finished water and distribution system. They mostly possessed methylotrophy pathways involving many enzymes rather than single enzyme systems. Network analysis determined potential species interaction between methanotrophs and a number of non-methanotrophic methylotrophs and other heterotrophs. The latter two groups had the capability to supply essential metabolites to methanotrophs as indicated by metabolic interdependency analysis.

This series of studies established a framework to understand the drinking water microbiome continuum through the inference of evolutionary and ecological processes that shape the microbiome from genomic/metagenomic data. They also offered new perspectives to some questions waiting to be answered by future studies, including “How to define a ‘healthy’ microbiome and microbial indicators?”, “How to effectively monitor opportunistic pathogens in drinking water microbiomes?”, and “Can drinking water microbiomes be predict and intentionally shaped?”.

ACKNOWLEDGMENTS

I would like to express my profound gratitude to my advisor, Professor Wen-Tso Liu, for his patient guidance and thorough support toward my dissertation work along the way. I am inspired by his high standards for research, insights on the research area, and charismatic leadership. I am also impressed by his character traits: intelligent but humble, considerate, and passionate. We spent nine years together, which is even longer than the time that I have had with my father so far and means a lot both to my research and personal life. I still remember that he wrote the four Chinese characters corresponding to “opening, developing, changing, and concluding” on a piece of paper to teach me how to write good research articles because we were both Chinese-speaking and those four characters were the four steps of Chinese classic writing. I also remember the “gift” he brought to me after he visited my undergraduate college, which was a picture of me at the age of 18 with short hair looking like a boy. It is because of all these small but memorable moments that make my Ph.D. life a unique experience.

I greatly appreciate the guidance from, and interaction with my dissertation committee members. Dr. Mark LeChevallier has over 30-year experience working with drinking water microbiomes and water quality engineering. His previous studies and methodologies provided crucial guidance for my current study. He gave me many names and publications related to my studies when we met accidentally at the Willard Airport of the University of Illinois. I was so impressed by his devotion to science when he told me that he read papers while waiting at airports. I feel grateful that he is willing to come onsite to my defense after his retirement from American Water. Every time I met Professor Benito Mariñas, I learned something. I would never forget his class on transport phenomena and modelling, though I was so nervous during every class because he thought and spoke so fast. His positive attitude towards fundraising activities for the department deeply touched me. When I asked for recommendation letters from him, he gave me a quick test on faculty position interview and a very useful tip. He has so many attributes that I wish I could have. In particular, I want to thank Professor Helen Nguyen for her guidance on research and support about situations women in engineering are facing. She has become a pattern for me on teaching undergraduates effectively and guiding young researchers with passion and empathy.

The research work in this dissertation has received much support from Illinois American Water. Ms. Elizabeth Doellman provided valuable assistance in sampling and data collection. Mr. Charles Andrew McCarrey helped me collect pipe sections. Many researchers provided meaningful discussion with their expertise, including Seungdae Oh from University of Illinois at Urbana-Champaign, Masaaki Kitajima from Hokkaido University (Japan), Stan J.J, Brouns and Gang Liu from Delft University of Technology (the Netherlands), and Paul van der Wielen from KWR Watercycle Research Institute (the Netherlands).

I want to thank several researchers for their guidance and enlightenment on scientific research and systematic thinking when I was a beginner. They are Yoichi Kamagata and Hideyuki Tamaki from National Institute of Advanced Industrial Science and Technology (Japan).

Many thanks to Dr. Shaoying Qi for always being helpful for experiments, equipment, and lab safety and management. He has also shared with me his knowledge and personal thoughts about Environmental Engineering research and education.

Thanks also go to students and researchers in Professor Wen-Tso Liu’s research group. Many of the students and postdoctoral researchers have given valuable feedback on my work. I want to thank Fangqiong Ling for her previous study on the same system.

I would like to thank Professor Vernon L. Snoeyink for hosting the Cultural Awareness and Speech Enhancement (CASE) class and talking with me frequently, from which I benefited tremendously. Members like Nanxi Lu and Jinyong Liu provided valuable guidance on culture difference and presentation and communication skills. I benefited tremendously from talking with Professor Snoeyink on drinking water treatment.

Lastly, I would like to give special thanks to my parents, for being incredibly supportive for my long-time stay in the United States. Many thanks to you two!

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

1.1 The history of drinking water treatment

The history of drinking water treatment can be traced back to 4000 BC, when methods including filtering through charcoal, exposing to sunlight, boiling, and straining were recorded in ancient Sanskrit and Greek writings. It was reported that in 1500 BC, the Egyptians used chemical alum to facilitate the settling of suspended particles (Halliday, 2004). The knowledge was forgotten in medieval Europe, and waste and wastewater were discharged in cities and a series of waterborne disease outbreaks including cholera and typhoid fever occurred (USEPA, 2000). It was not until the 1700s that Europe re-established the regular use of drinking water treatment filtration technologies as an effective means of removing particles from water (USEPA, 2000). The effectiveness of these water treatment technologies could not be successfully measured to understand why water treatment using both filtration and disinfection were essential until after the invention of cultivation technique and the germ theory of disease between 1850s and 1890s (Koch and Duncan, 1894). Currently, drinking water treatment and distribution are practiced in two ways based on the consideration of chlorination for biological safety and the exposure to harmful substances from disinfection by-products (DBPs). One proposes to maintain a minimal level of residual disinfectants to suppress microbial regrowth in distribution system (DS) and the other (the Netherlands, Switzerland, Austria, and Germany) tries to substantially reduce available carbon for regrowth during distribution instead of adding disinfectant residuals. Clearly both approaches recognize the importance of microbes in drinking water systems. Concurrently, regulatory agencies have implemented rules and regulations for drinking water quality to protect public health. However, the microbiological standard of drinking water still relies on heterotrophic plate count and indicator microorganisms (*Escherichia coli* and total coliforms) proposed more than 100 years ago to determine the adequacy of water treatment and the integrity of the DS.

1.2 The development of microbial ecology of drinking water systems

Accurate description of microbes in drinking water systems (DWSs) from source water, treatment in the production process, distribution network, and up to the point where water enters a building is rather challenging owing to its complexity in water matrices, physical structure and chemical gradients from source to tap. In the United States (US) alone, 34 billion gallons of water is produced daily by > 151,000 public water systems and passes through almost 1 million miles of pipelines to individual households (National Research Council, 2006; USEPA, 2017). A typical surface water system usually consists of abstraction, coagulation, sedimentation, filtration, disinfection, and distribution. More complex systems include ozonation, (biological) activated carbon, membrane separation, and ultraviolet disinfection (Rosario-Ortiz et al., 2016). A municipal water DS further consists of pipes, pumps, valves, storage tanks, reservoirs, meters, fittings, and other hydraulic accessories made of many different materials (National Research Council, 2006).

As early as 1945, Wilson described the concept of microbial ecology in drinking water DS (Wilson, 1945), and suggested that the ecological niches inside a DS could be determined by knowing the type and number of bacteria developed. After more than 70 years, we are beginning to gain a glimpse of the microbial ecology of drinking water systems and appreciate the complexity of such ecosystems with the invention of new detection methods. As Koch said, “As soon as the right method was found, discoveries came as easily as ripe apples from a tree.”. The inventions of Leeuwenhoek’s simple spherical lenses, the electron microscope, pure culture technique, and recombinant DNA all support such a paradigm shift, and elevate our understandings of the microbial world. Over the past 20 years, cultivation-independent approaches that primarily base on the use of rRNA gene sequences have transformed our understanding of the microbial world from deep sea, to human microbiome, and now to drinking water microbiome in a remarkable way.

1.3 Organization of this dissertation

This dissertation describes the ecological patterns of a groundwater-sourced drinking water microbiome at different scales.

- Chapter 1 provides background including the history of drinking water treatment and the development of microbial ecology of drinking water systems.
- Chapter 2 reviews our current understandings on drinking water microbiome, which is closely related to the advent of new methods for monitoring and characterizing microbial communities. The potential of using omics technologies to study drinking water microbiomes is also discussed.
- Chapter 3 resolves the ecological patterns observed at the community level, including structural diversity and metabolic potential of drinking water microbiome continuum under disturbances from the treatment process and predation from eukaryotes.
- Chapter 4 concentrates on pathogen-related species at the population level with regard to their abundance, diversity, potential pathogenicity, genetic exchange, and distribution across the studied drinking water system.
- Chapter 5 expands to the multi-species level, investigating the compositional and functional diversity, interspecies relationship, and metabolic interdependency of methano-/methylo-trophic bacteria.
- Chapter 6 finishes with the conclusions from this research and identifies challenges we are facing and suggestions for further work.

1.4 References

National Research Council (2006) *Drinking water distribution systems: assessing and reducing risks*: National Academies Press.

Halliday, S. (2004) *Water: a turbulent history*: Sutton.

Koch, R., and Duncan, G. (1894) *Professor Koch on the Bacteriological Diagnosis of Cholera, Water-filtration and Cholera, and the Cholera in Germany During the Winter of 1892-93. Translated by G. Duncan. Etc. [Reprinted from the "Scotsman."]*: Edinburgh.

Rosario-Ortiz, F., Rose, J., Speight, V., von Gunten, U., and Schnoor, J. (2016) How do you like your tap water? *Science* **351**: 912-914.

USEPA (2000) *The history of drinking water treatment*: Office of Water.

USEPA (2017). Information about public water systems. URL https://www.epa.gov/dwreginfo/information-about-public-water-systems

Wilson, C. (1945) Bacteriology of water pipes. *J Am Water Works Assoc* **37**: 52-58.

CHAPTER 2 CURRENT UNDERSTANDING ON DRINKING WATER MICROBIOMES

2.1 Abstract

To advance the understandings on drinking water microbiomes, this review concentrates on i) what cultivation-independent tools have been developed for the analyses of drinking water microbiomes; ii) what knowledge we have gained so far on the population dynamics and microbial ecology of drinking water microbiomes; and iii) how the next-generation techniques are able to provide new perspectives on the complexity of the microbial community within drinking water ecosystem. The goal is to guide operational practices in water utilities to create and maintain a “healthy” microbiome in the distribution system (DS), and to enlighten future research on drinking water microbiomes.

2.2 Current methods for drinking water microbiome monitoring

To routinely monitor and characterize drinking water microbiomes, a suite of methods has been developed (Figure 2.1). These methods are mainly used to measure microbial density, microbial composition, and microbial activities, and detailed description has been reported previously (Liu et al., 2013a; Douterelo et al., 2014b). It should be noted that no single method can provide all the information, including the presence/absence and concentrations of specific opportunistic pathogens, microbial density, community composition and structure, and spatial arrangement. The current strategy is to combine several methods to improve the view of the microbiome in the studied system. Also, all the methods have known biases associated with and should be used with caution.

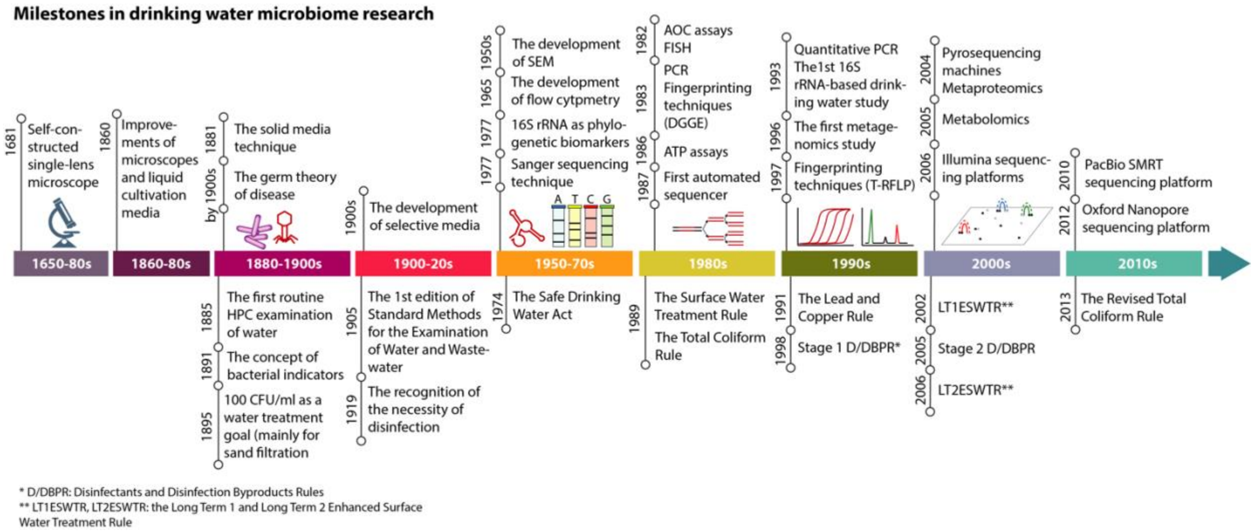

Milestones in drinking water microbiome research
1681 Self-constructed single-lens microscope
1860 Improvements of microscopes and liquid cultivation media
1881 The solid media technique
by 1900s The germ theory of disease
1900s The development of selective media
1950s The development of SEM
1965 The development of flow cytpmetry
1977 16S rRNA as phylogenetic biomarkers
1977 Sanger sequencing technique
A T C G
1982 AOC assays FISH
1983 PCR Fingerprinting techniques (DGGE)
1986 ATP assays
1987 First automated sequencer
1993 Quantitative PCR The1st 16S rRNA-based drinking water study
1996 The first metagenomics study
1997 Fingerprinting techniques (T-RFLP)
2004 Pyrosequencing machines Metaproteomics
2005 Metabolomics
2006 Illumina sequencing platforms
2010 PacBio SMRT sequencing platform
2012 Oxford Nanopore sequencing platform
1650-80s
1860-80s
1880-1900s
1900-20s
1950-70s
1980s
1990s
2000s
2010s
1885 The first routine HPC examination of water
1891 The concept of bacterial indicators
1895 100 CFU/ml as a water treatment goal (mainly for sand filtration
1905 The 1st edition of Standard Methods for the Examination of Water and Wastewater
1919 The recognition of the necessity of disinfection
1974 The Safe Drinking Water Act
1989 The Surface Water Treatment Rule
The Total Coliform Rule
1991 The Lead and Copper Rule
1998 Stage 1 D/DBPR*
2002 LT1ESWTR**
2005 Stage 2 D/DBPR
2006 LT2ESWTR**
2013 The Revised Total Coliform Rule
* D/DBPR: Disinfectants and Disinfection Byproducts Rules
** LT1ESWTR, LT2ESWTR: the Long Term 1 and Long Term 2 Enhanced Surface Water Treatment Rule

Figure 2.1 History on the development of major microbiological methods (upper half), and the introduction of key regulations and rules in drinking water production, primarily based on the US system (bottom half).

Table 2.1 Summary of methods used to measure microbial density, microbial composition, and microbial activities in drinking water microbiome studies.

Microbial density	Microbial composition	Microbial activities
Cultivation methods (HPC, selective and differential media)	**Community fingerprint** (DGGE, T-RFLP, PCR-ALH, SSCP)	ATP assay
Cell counting (microscopic counts, FCM)	**16S rRNA gene amplicon analysis** (clone library and Sanger sequencing, NGS)	Enzymatic activity tests
Molecular methods (qPCR, viable qPCR, ddPCR)	16S rRNA gene hybridization (DNA microarray)	Assimilable organic carbon (AOC)
	Spatial distribution (FISH, SEM)	

Methods for measuring microbial density Cultivation is still the most widely-used technique to quantify microbial density in drinking water sector since the late 18th century. Heterotrophic plate count (HPC) and bacterial indicators (i.e., total coliforms and *E. coli*) were introduced to determine the adequacy of water treatment and the integrity of the DS (Frankland, 1894)(Koch and Duncan, 1894)(Payment et al., 2003)(US

EPA 1984)(Bartram et al., 2004). Isolating and enumerating disease-causing microorganisms is always a priority in drinking water studies. Thus, selective and differential media techniques have been developed to cultivate many known pathogens. For example, the buffered charcoal yeast extract (BCYE) agar is a selective medium developed to isolate the once difficult-to-culture *L. pneumophila* in 1980s that caused the outbreak of Legionnaires' disease in 1976 (Pasculle et al., 1980; Edelstein, 1981).

An alternative and most direct way to quantify microbial density is cell counting. Cells in water samples can be directly counted using a microscopy, or stained with fluorescent dyes and counted under an epifluorescence microscope or flow cytometry (FCM) as total cell count (TCC). The usage of FCM in DSs and premise plumbing is still limited to systems without residual disinfectants. For drinking water containing residual disinfectants, pretreatment using membrane filtration to concentrate bacteria at an appropriate density is required due to low cell number and the interference of bacteria-like particles (Lautenschlager et al., 2013; Besmer et al., 2016; Van Nevel et al., 2017), which is time-consuming and subjective (Santic et al., 2007).

Since the mid-1980s, various forms of molecular methods (e.g., quantitative PCR (qPCR), viable qPCR and digital droplet PCR (ddPCR) have been developed to provide qualitative and quantitative information related to total or specific bacterial cells, and to the ratio of live and dead cells (Chen and Chang, 2010; Yanez et al., 2011; Lee et al., 2015). These molecular tools are mostly based on the use of a key biomarker known as 16S rRNA gene (Woese 1987). The 16S rRNA gene sequence is conserved among the domains *Bacteria* and *Archaea* and composed of regions that are variable enough to be used for phylogenetic classification at different levels of specificities (i.e., species, genera … phyla and domains) (Liu and Stahl, 2007).

Methods measuring microbial composition Characterizing and monitoring microbial populations is an important first step to elucidate the complexity of microbial ecology in DWSs. Two major types of PCR amplification methods can be carried out. The first type of PCR-based methods is generally termed “community fingerprint". It analyzes the amplified 16S rRNA genes and generates a pattern-based profile of community structure, most commonly represented by a banding pattern of nucleic acid fragments resolved by gel electrophoresis. In general, these community fingerprinting methods allow one to rapidly examine the microbial diversity within a microbial ecosystem, or compare the differences and similarities on the microbial community structure among different ecosystems. Many studies have also demonstrated the capabilities of these molecular tools to provide more rapid and better insights into microbial diversity than cultivation methods in different natural and engineered environments (Liu and Stahl, 2007). In the last decade, many microbial fingerprinting methods have been developed to facilitate the determination of microbial diversity in various ecosystems. Commonly used methods including DGGE (denaturing gradient gel electrophoresis), T-RFLP (terminal restriction fragment length polymorphism), PCR-ALH (amplicon length heterogeneity), and SSCP (single-strand-conformation polymorphism) were detailed previously (Liu and Stahl, 2007).

The second type of molecular methods is to obtain a dataset of 16S rRNA gene sequences from the extracted community genomic DNA. Initially, this was achieved through clone library construction of 16S rRNA gene sequences. In recent years, the composition of 16S rRNA gene sequences in a microbial sample can be obtained using the next-generation sequencing (NGS) technology. Both approaches describe the microbial composition based on the number of unique 16S rRNA sequences and the abundance of individual 16S rRNA sequences. The 16S rRNA sequences can be further compared with all 16S rRNA sequences stored in a public database. This allows one to infer the phylogeny affiliation of individual 16S rRNA sequences and determine whether the sequences are novel or related to known organisms based on the similarity of sequence homology (e.g., <97% for defining a new species). Furthermore, based on the sequence information, one can design oligonucleotide probes specific for a target organism or a group of organisms, and then apply them in whole-cell hybridization or membrane hybridization for confirming the presence of the targeted organisms or for quantitative measurement of those targeted organisms in the environment. Likewise, the active members within the microbial community can be determined using the corresponding RNA-based analysis instead of DNA-based methods.

DNA microarray technology has emerged as a high-throughput platform for nucleic acid analysis in environmental microbiology studies (Zhou, 2003; Bodrossy and Sessitsch, 2004; Li and Liu, 2004). For microbial identification and community analysis, this platform generally uses rRNA genes as the phylogenetic marker (Guschin et al., 1997; Liu et al., 2001; Small et al., 2001; Loy et al., 2002; Wilson et al., 2002; Peplies et al., 2003). Initially, hundreds up to thousands of rRNA-based oligonucleotide probes with a length between 15 and 25 nt are designed to target the rRNA gene sequences of interested microorganisms at different levels of specificities (i.e., species, genera … phyla and domains), and then spotted or *in-situ* synthesized onto a microarray substrate (Loy et al., 2002; Wilson et al., 2002). Followed by hybridization with fluorescently labeled rRNA/rDNA targets and washing at optimal conditions, signals are measured and statistically analyzed to infer microbial community structures in complex environmental samples (Small et al., 2001; El Fantroussi et al., 2003; Peplies et al., 2004).

Finally, the microbial composition can also be determined by examining the spatial distribution of microorganisms *in-situ*. Fluorescence *in-situ* hybridization (FISH) and scanning electron microscope (SEM) are the most commonly used methods to show the spatial structure and arrangement of microbial communities. The development of SEM in the 1950s (Fischer et al., 2005) has enabled the visualization of 3-dimensional topography of biofilm in water meters (Hong et al., 2010). Using 16S rRNA gene as a biomarker, FISH technique has been introduced in the late 1980s to identify and quantify microbial populations at different phylogenetic levels, and in combination with other techniques to determine microbial functions in their natural positions (Delong et al., 1989; Amann et al., 1990).

Methods measuring microbial activities To measure microbial activity, the currently available methods included ATP assay, enzymatic activity tests, and assimilable organic carbon (AOC) tests (Vanderkooij et al., 1982; Stutz et al., 1986; Manz et al., 1993; Henne et al., 2012; Lautenschlager et al., 2014). ATP assay determines all biologically active microorganisms based on the total amount of ATP measured through bioluminescence assay (Stutz et al., 1986). Enzymatic activity tests quantify specific enzymes by monitoring the increase of fluorescence intensities or absorbance owing to the degradation of substrates by specific enzymatic activities such as polysaccharide-degrading enzymes (α- and β-glucosidase, cellobiohydrolase, xylosidase, chitinase) as a function of time (Roskoski, 2007; Lautenschlager et al., 2014). AOC concentration represents the fraction of dissolved organic carbon that may readily support microbial growth and is determined by measuring the maximum level of growth of two bacterial isolates (*Pseudomonas fluorescens* P-17 and *Spirillum* sp. strain NOX) in a water sample, which usually takes 5-7 days (Vanderkooij et al., 1982). However, activity measurements have not been widely incorporated in drinking water microbiome studies because they are time-consuming and labor-intensive.

Limitations of current methods Current methods for the monitoring of drinking water microbiomes do have limitations. Cultivation-based methods are known to be time-consuming, low in sensitivity, and ineffective in recovering most organisms (Staley and Konopka, 1985; Berney et al., 2008; Hammes et al., 2008). Thus, current studies mostly rely on molecular tools to gain insights into drinking water microbiomes. We systematically summarize biases that can occur or be associated with key steps from sampling to data interpretation in those molecular-based methods (Table 2.2). Firstly, experimental design in technical and biological replicates is critical to statistically determine variations between samples, but is often not considered due to time, manpower, and cost (Brooks et al., 2015). For example, sample-to-sample heterogeneity often occur during biofilm samplings. Sampling biofilm in replicates in full-scale systems can be difficult because of limited access, high cost, and high chances of contamination

(Gomez-Smith et al., 2015; Ling et al., 2016). The sample-to-sample heterogeneity can become more significant when analyses of molecules at different metabolic levels (i.e., DNA, RNA, proteins, and metabolites) are integrated (Muller et al., 2013). Sample volume and concentration methods also play a crucial role in determining whether enough biomass can be successfully obtained for downstream molecular analyses. At present, no standard practice has been established for the minimal sampling volume required and the concentration method used. Sample volume ranging from 100 mL to 2000 L is necessary and often dependent on downstream analyses and research objectives. The extraction efficiency of DNA, RNA, and protein from collected samples can vary from 35 to 85%, and can sometimes influence the taxonomic outcomes of microbiota assessments (Hwang et al., 2012b; Guo and Zhang, 2013; Henderson et al., 2013; Moran et al., 2013; Stark et al., 2014; Tsementzi et al., 2014; Brooks et al., 2015). Biases can also occur during PCR amplification (30-50%) (Brooks et al., 2015), due to differences in GC content of microbes (Duhaime et al., 2012), rRNA gene copy number (Klappenbach et al., 2001), and primer annealing efficiency (Wu et al., 2009). All these factors can lead up to 10% variation in estimating the relative abundance of specific microbial groups (Angly et al., 2014). During DNA sequencing, various degrees of error can occur depending on the sequencing platforms with error rate ranging from <1% up to 14% (Roberts et al., 2013; Ross et al., 2013; Feng et al., 2015; Jain et al., 2016). Last but not the least, 16S rRNA is the most commonly used biomarker for identifying microbial populations. As short sequence reads are often used, the classification of microbial populations is only accurate to the genus or family level (Schloss, 2010). Also, 16S rRNA gene sequence cannot accurately provide the physiological function unlike whole genome-based methods (Jain et al., 2017). This can be a significant problem in distinguishing pathogenic strains from commensals (Steele and Streit, 2005; Edberg, 2009).

Table 2.2 Limitations of current methods as shown by systematic errors reported in the literature.

Experimental/analytical steps	Range	Systematic error	References
Sample number (technical and biological replicates)	2-15	5%	(Prosser, 2010; Brooks et al., 2015; Bautista-de los Santos et al., 2016b)
Sample volume and concentration methods	100 ml - 2000 L	N/A	(APHA et al., 1998; Chao et al., 2013; Liu et al., 2013b; Wang et al., 2017; Zhang et al., 2017)
DNA extraction efficiency	1 – 100%	35-85%	(Hwang et al., 2012b; Guo and Zhang, 2013; Henderson et al., 2013; Brooks et al., 2015)
RNA extraction efficiency	1 – 100%	50%-80%	(Moran et al., 2013; Stark et al., 2014; Tsementzi et al., 2014)
Protein extraction efficiency	1 – 100%	50%-60%	(Keiblinger et al., 2012; Hansen et al., 2014)
PCR biases	$1.4^N – 2^N$	30%-50%	(Duhaime et al., 2012; Brooks et al., 2015)
rRNA gene copy number	~ 1 – 20	~10% in abundance estimate	(Klappenbach et al., 2001; Vetrovsky and Baldrian, 2013; Angly et al., 2014)
DNA sequencing	90 – 100%	0.1%-10%	(Lee et al., 2012; Roberts et al., 2013; Ross et al., 2013; Feng et al., 2015; Jain et al., 2016)
16S rRNA resolution (length and targeted variable regions)	genus - family	12-15% in underestimating diversity	(Brooks et al., 2015; Jain et al., 2017)

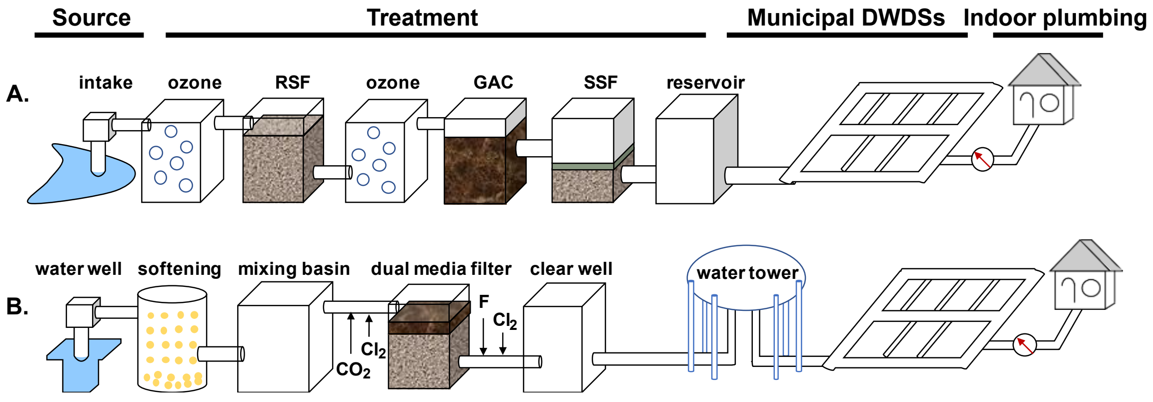


Figure 2.2 The configuration of typical DWSs. A) A complex system in Switzerland, B) a simple system in a mid-size town in the US.

2.3 Current understandings on drinking water microbiomes

2.3.1 Changes in drinking water microbiomes due to environmental factors

Current drinking water production plants apply physical and chemical means to remove unwanted chemicals and microorganisms, and some use microbiological processes such as sand filters and granular activated carbon filters to biologically remove excessive nutrients and break down soluble organic matters (Figure 2.2). The produced water with or without residual disinfectants is then delivered to consumers through complex distribution network. Thus, it is important to understand how drinking water microbiomes can be influenced by different environmental factors, including source water matrices, treatments within drinking water production plant, DSs prior to entering indoor plumbing, and indoor plumbing (Table 2.3). Due to the substantial difference between the water system before water meter and premise plumbing, this review focuses mostly to the former. Readers can refer to the following reviews published dedicatedly to premise plumbing (Wang et al., 2013; Dai et al., 2017; Wang et al., 2017).

Effect of source water matrices Surface water and groundwater are the two types of source waters most commonly used to produce drinking water, and can contain distinct microbial populations owing to differences in physical and chemical gradients. These microbial populations present can 'seed' downstream microbiota in the treatment train, the DS, and indoor plumbing. Groundwater systems are mostly known to be anoxic or anaerobic with relatively high concentrations of compounds (i.e., iron, manganese, ammonia, sulfur compounds, methane, and dissolved organic carbon) supporting the growth of anaerobic communities (van der Wielen et al., 2009; Holinger et al., 2014; Albers et al., 2015; Ling et al., 2016; Bruno et al., 2017; Zhang et al., 2017). Due to aeration on abstraction, most of these anaerobes cannot survive under the oxidative stress and are generally not detected downstream (Roeselers et al., 2015; Zhang et al., 2017).

Table 2.3 Treatment effects on drinking water microbiomes from the literature.

Treatment stages	Purpose	Impacts on drinking water microbiome	References
Source water	Water supply	- Serve as inoculum to downstream microbiota. - Seeding effect is more significant in surface water systems than groundwater systems. - Seasonal variations in surface water influences changes in downstream microbiota. - Biofilm community in surface water systems have slightly higher diversity than ground water systems.	(Pinto et al., 2012; Pinto et al., 2014; Sun et al., 2014; Gomez-Alvarez et al., 2015; Roeselers et al., 2015; Revetta et al., 2016; Douterelo et al., 2017; Zhang et al., 2017)
Softening	Hardness removal	- Softening effluent is dominated by a single bacterial population.	(Zhang et al., 2017)
Coagulation and sedimentation	Removal of turbidity, infectious agents, and DBP precursors	- Their influence on microbial community is relatively small.	(Eichler et al., 2006; Poitelon et al., 2010; Zeng et al., 2013; Lin et al., 2014)
Ozonation	Disinfection and removal of NOM	- It has significant impacts on total cell counts and community diversity. - The resulting AOC supports a different community in downstream biofilters compared with incoming water.	(Vaz-Moreira et al., 2013; Zeng et al., 2013; Lautenschlager et al., 2014)
Filtration	Removal of turbidity, pathogens, and various contaminants	- Filtration is the key step shaping downstream microbiota through removal of incoming particles and seeding of outflow owing to the sloughing of biofilms on filter media. - Various biological processes can occur in filters. - Some groups have greater potential to seed downstream. - Eukaryotes play an important role in bacterial dynamics in filters.	(Cerrato et al., 2010; Kasuga et al., 2010a, b; White et al., 2012; Zearley and Summers, 2012; Feng et al., 2013; Liao et al., 2013; Holinger et al., 2014; Lautenschlager et al., 2014; Albers et al., 2015; Gulay et al., 2016; Marcus et al., 2017)
Disinfection	Pathogen inactivation	- Disinfection treatments can lead to decreases in microbial diversity, in particular, on the active members. - Chlorination and chloramination might select for different bacterial populations. - The molecular mechanism of chlorine resistance is attributed to glutathione synthesis.	(Eichler et al., 2006; Gomez-Alvarez et al., 2012; Hwang et al., 2012a; Chao et al., 2013; Chiao et al., 2014; Sun et al., 2014; Wang et al., 2014b; Bautista-de los Santos et al., 2016a)
Distribution	Water transportation	- Drinking water microbiota are involved in many important processes in DSs, i.e., the formation of biofilms and loose-deposits on pipe walls, harboring pathogens in biofilm, nitrification, manganese oxidation, methane oxidation, and inducing pipe corrosion. - Many factors have been determined to play a role in microbial regrowth, such as temperature; the amount of usable carbon; flow regime (hydrodynamics); water residence time; pipe materials; and the presence of corrosion products.	(LeChevallier et al., 1996; Camper et al., 1999; Berry et al., 2006; Lautenschlager et al., 2010; Liu et al., 2013a; Ji et al., 2015; Proctor et al., 2015; Bautista-de los Santos et al., 2016a; Dai et al., 2017)

In comparison, the seeding effect, or treatment breakthrough, has been reported in treatment processes using surface water as source water (Pinto et al., 2012; Gomez-Alvarez et al., 2015).Seasonal variations in bulk water community are also commonly observed with seasonal changes in surface water (Pinto et al., 2014; Douterelo et al., 2017).

Softening Softening is used to remove calcium, magnesium, and certain other metal cations that contribute to hardness, with concurrent benefit to the removal of heavy metals, natural organic matter, turbidity, and pathogens (Peters, 2011). During the softening process, lime and soda ash are added to raise pH rapidly to 10.3 for calcium precipitation or 11.0 for magnesium precipitation. It is anticipated that not many microbes can survive under this drastic change in pH, as shown by a recent study reporting a substantial reduction in community diversity before and after the softening process (Zhang et al., 2017). In comparison, a wide variety of different microorganisms were observed to colonize the calcite pellets in a full-scale pellet softening reactor preceded by ozonation. They proliferated as soon as the pH in the water was neutralized due to calcite crystallization in the presence of highly biodegradable nutrients. Biomass could reach as high as 220 mg ATP/m^3 of the reactor and were responsible for the removal of 60% of AOC from the water (Hammes et al., 2011).

Coagulation, flocculation, and sedimentation These processes are used to remove suspended solids including small particulars and colloids (0.001 – 1.0 µm), improve water turbidity color, and reduce the level of microbial pathogens and DBP precursors (Peters, 2011). Coagulation and flocculation turn small particles present in source water into larger particles called ‘flocs’, which are then removed during sedimentation and filtration. These processes are reported to have no observable effects on microbial community structure (Eichler et al., 2006; Poitelon et al., 2010; Lin et al., 2014). However, one study observed a significant community shift during sedimentation (Zeng et al., 2013).

Filtration Filtration separates suspended or colloidal impurities from water by passing it through a porous medium (e.g., a bed of sand, coal, activated carbon, or garnet) to improve turbidity, and remove pathogens and many organic and inorganic contaminants. Depending on the source water quality, one or a series of filters are used at a DWS, including rapid sand filters (RSFs), granular activated carbon (GAC) filters, and slow sand filters (SSFs). Filtration is a key step in shaping DS microbiota by removing incoming microbes as a form of particles through mechanical screening and by seeding outflow with microbes as planktonic cells or aggregates detached from filter media (Peters, 2011). Microbial biomass can be enriched on the filters and reach up to a density of 10^9 copies of 16S rRNA gene per g-filter material or 10^{15}-10^{16} cells per m^3 filter material. Depending on water quality, these microbes have various metabolic functions, including oxidation of ammonia, iron, and manganese, metabolism of sulfur compounds, and degradation of dissolved organic carbon and trace organic micropollutants (Magic-Knezev and van der Kooij, 2004; Velten et al., 2007; de Vet et al., 2009; van der Wielen et al., 2009; Cerrato et al., 2010; Kasuga et al., 2010a, b; White et al., 2012; Zearley and Summers, 2012; Feng et al., 2013; Liao et al., 2013; Holinger et al., 2014; Lautenschlager et al., 2014; Albers et al., 2015; Gulay et al., 2016; Marcus et al., 2017). Dense bacterial cells and protozoa are frequently observed at the top layer of slow sand filters (i.e., *Schmutzdecke*), and eukaryotic predation has been shown to play a critical role in the dynamics of the bacterial community in the filters (Lautenschlager et al., 2014; Haig et al., 2015).

Ozonation Ozonation is used as a disinfection and oxidation process to enhance microbial removal, control taste and odor, and eliminate micropollutants from water (von Gunten, 2003). The strong oxidative stress imposed by ozonation causes a significant reduction of the total cell counts and community diversity (Vaz-Moreira et al., 2013; Zeng et al., 2013; Lautenschlager et al., 2014). It also oxidizes natural organic matter (NOM) into low-molecular-weight and possibly biodegradable AOC, which is then removed by biological filters. A distinct microbial community in biofilters following an ozonation step was identified in comparison to the incoming raw water (Lautenschlager et al., 2014).

Disinfection Free chlorination and chloramination are two major types of disinfection treatments used to inactivate pathogens during drinking water production and transportation processes. Disinfection treatments can lead to decreases in microbial diversity for systems maintaining a disinfectant residual (Gomez-Alvarez et al., 2012; Chao et al., 2013; Sun et al., 2014; Bautista-de los Santos et al., 2016a). Due to the difference in the inactivation mechanisms, chlorination and chloramination were reported to select for different bacterial populations in a DWS with alternating disinfection treatment between chlorination and chloramination (Hwang et al., 2012a; Wang et al., 2014b). However, when more systems were incorporated and compared, this trend was not observed (Bautista-de los Santos et al., 2016a).

Distribution network DS pipes carry drinking water from a centralized treatment plant or well supplies to consumers' taps, providing the required water quantity and quality at a suitable pressure. Managing the network is a primary challenge from both an operational and public health standpoint due to the expansive physical infrastructure (Snoeyink et al., 2006). Microbial regrowth with spatiotemporal variation is the major concern in distribution as the physicochemical and nutritional conditions provided by pipe walls are very different from those found during treatment. Recent studies were able to identify the microbial community and dominant species associated with many important processes in DSs. These processes included the formation of biofilms and loose-deposits on pipe walls (Kelly et al., 2014; Liu et al., 2014; Wang et al., 2014a), harboring pathogens in biofilm (Wang et al., 2013; Ling et al., 2016), nitrification (Regan et al., 2003; Zhang et al., 2008; van der Wielen et al., 2009; Wang et al., 2014b), oxidation of manganese (Marcus et al., 2017) and methane (Kelly et al., 2014; Ling et al., 2016), and inducing pipe corrosion (Beech and Sunner, 2004; Zhang et al., 2008; Li et al., 2010; Chen et al., 2013; Jin et al., 2015; Li et al., 2015b). Many factors have been determined to play a role in the microbial regrowth in DSs, including temperature, especially warm water conditions, the amount of usable carbon, flow regime (hydrodynamics), water residence time, pipe materials, and the presence of corrosion products (LeChevallier et al., 1996; Camper et al., 1999; Berry et al., 2006; Liu et al., 2013a; Proctor et al., 2015; Bautista-de los Santos et al., 2016a; Dai et al., 2017). For premise plumbing, pipe diameter is reported as a key critical factor (Lautenschlager et al., 2010; Ji et al., 2015).

2.3.2 Spatiotemporal shifts in microbiomes as a continuum throughout a DWS

Drinking water microbiomes can be viewed as a continuum of microbial communities that exhibit spatiotemporal dynamics. They can shift dramatically over distance as described in the previous section and over time at short- or long-term intervals. For long-term changes, seasonal variations have been observed in the bulk water phase, the biofilm phase, and the cold and hot waterlines in DSs, and were found to be correlated with disinfection treatment and seasonal temperature change (Henne et al., 2013; Pinto et al., 2014; Ling et al., 2016; Prest et al., 2016). Short-term fluctuations occur on a scale of hours-to-weeks, and are difficult to capture due to low frequency in sampling. These spatiotemporal changes can be further exemplified by studies about four DWSs in Europe and the US.

The city of Braunschweig, Germany, uses two surface water reservoirs as raw water to produce drinking water through two systems with coagulation-flocculation, sand barriers, and chlorination (0.2-0.7 mg/L). Containing no residual chlorine, the treated water from the two sources is transported to a storage container and mixed at a constant ratio (Lesnik et al., 2016). The drinking water microbiome was analyzed at different stages of the system, including the production system, the distribution network, and cold and hot waterlines of premise plumbing, together with the monitoring of *Legionella* from source to cold and hot waterlines (Eichler et al., 2006; Henne et al., 2012; Henne et al., 2013; Lesnik et al., 2016). The findings suggested that microflora in the DS were influenced by both source water and chlorination with chlorination having a more profound impact on the active community than the overall microbial community (Eichler et al., 2006). The biofilm community and the bulk water community did not share any core microbial population. The bulk water community was observed to have a high number of the low-abundance bacterial populations, and the biofilm community had a reduced diversity. It was hypothesized that low-abundance bacterial populations in the bulk water could function as an inoculum to seed the biofilm community (Henne et al., 2012). Seasonal dynamics observed in drinking water microbiome were highly influenced by source water (Henne et al., 2013). It was further observed that treatment processes had apparent impact on *Legionella* species. The types of *Legionella* species in cold waterlines and hot waterlines were different with substantially more *Legionella pneumophila* in hot waterlines (Lesnik et al., 2016).

The city of Zurich, Switzerland, uses surface water (Lake Zürich) to produce 50% of daily drinking water demand through sequential ozonation and filtration steps (RSF, GAC, and SSF), together with untreated groundwater (49%) and spring water (1%). The produced water contains no residual disinfectants in the distribution network. The drinking water microbiome was analyzed using online flow cytometry and HPC for cell counts, ATP for microbial activities, and 16S rRNA-based molecular methods (Lautenschlager et al., 2010; Lautenschlager et al., 2013; Lautenschlager et al., 2014). Results based on phylogenetic, enzymatic, and metabolic analyses indicated that microbial communities in the water phase shifted from source water, to RSF effluent, GAC effluent, and then SSF effluent. Filter microbial communities in RSF, GAC, and SSF differed among each other, and from those observed in the effluent of individual filters. The microbial community of SSF filter remained unchanged in the subsequent reservoir during a two-year consecutive sampling (Lautenschlager et al., 2014). Within the DS, the microbiome composition in bulk water sampled at different locations throughout the distribution network remained remarkably stable during the two-year sampling period. High abundances of candidate phyla were detected compared with systems with residual disinfectants (Lautenschlager et al., 2013). In premise plumbing, cell concentration increased in the first liter of tap water after overnight stagnation, followed by step-wise decrease when the water was flushed (the first 2 liters). This increase in cell concentration could only be partially explained by the growth due to available AOC (Lautenschlager et al., 2010).

At the city of Ann Arbor, Michigan, USA, the DWS uses surface water (Huron River) and local wells (groundwater) as raw water at approximately 2:1 in the winter and 8:1 in the summer. The water is treated through a process including lime softening, coagulation, flocculation, sedimentation, ozonation, dual media filtration, and chloramination. The finished water contains approximately 3 mg Cl_2/L chloramine as the residual disinfectants with a pH between 9.1 and 9.3. Microbial community analysis together with modelling and network analyses were used to describe and generalize the trend observed with drinking water microbiome. The results indicated that filtration by dual media sand filters played a primary role in shaping the microbial community in the DS, and bacterial taxa that colonized the filter and sloughed off in the filter effluent persisted in the DS (Pinto et al., 2012). The drinking water microbiome in the distribution network exhibited a strong temporal trend of seasonal cycling correlating with temperature and source water usage patterns, and weaker spatial dynamics. The the relative abundance of a taxon and the frequency of its detection were positively correlated (Pinto et al., 2014). The findings could be further used to develop a predictive framework for microbial management.

The cities of Champaign and Urbana, Illinois, USA use groundwater from the Mahomet aquifer containing dissolved methane as the source water to produce drinking water through two-stage lime softening, recarbonation, chlorination and filtration (Gunsalus et al., 1972; Flynn et al., 2013) (Hwang et al., 2012a). Chloramine was used as the residual disinfectant for many years but was switched to free chlorine in 2012, and the finished water maintains a residual disinfectant concentration at approximately 2.5 mg Cl_2/L with a pH of 8.8. Abstraction and softening processes were shown to cause major microbial community shifts throughout the system. After an extensive period of monitoring of the microbiome in bulk water and water meter biofilms (Hong et al., 2010; Ling et al., 2016), the shifts in microbial communities were shown to be correlated with disinfectant types and sampling time for the water-phase samples but not for the biofilm-phase samples from water meters. Between bulk water and water meter biofilms, the shared core microbiome contained a high abundance of populations related to methano- and methylotrophs and exhibited seasonal variations (Ling et al., 2016). A variety of eukaryotic groups were detected throughout the system, indicating that predation could be another factor driving the diversification of the bacterial community.

2.4 Microbial ecology of drinking water microbiome

Typically, DWSs are viewed as complex microbial ecosystems that create various ecological niches to support the growth of microbes and microbial community in individual niches is shaped by a variety of deterministic factors, as described in section 3. Recently, neutral processes including random death, dispersal and speciation are considered to play a crucial role in community assembly in the distribution system and premise plumbing. Additional efforts are being made to transform our understanding of drinking water microbiomes from descriptive processes to hypothesis-driven ecological processes (Horner-Devine et al., 2004).

Core microbiome   Identifying a core microbiome is an important step for gaining insights into the microbial function associated with an ecosystem, and is often used to provide guidance on how to manipulate microbial communities to achieve desired outcomes. A core microbiome is typically defined as the suite of members shared among microbial consortia from similar habitats (Shade and Handelsman, 2012). As the water-phase microbial communities within a given system are relatively stable, irrespective of the sampling locations over short time-scales (Lautenschlager et al., 2013; Pinto et al., 2014; Roeselers et al., 2015), studies have attempted to define core microbiome (i.e., shared microbial taxa) in the bulk water phase after the production process across different DWSs. At high taxonomical levels, such as phyla, classes, and families, the core microbiome is made up primarily of *Alpha- and Beta-proteobacteria,* and to a lesser extent of *Gammaproteobacteria, Nitrospirae, Planctomycetes, Acidobacteria, Bacteroidetes* and *Chloroflexi* (Eichler et al., 2006; Pinto et al., 2012; Vaz-Moreira et al., 2013; Zeng et al., 2013; Lautenschlager et al., 2014; Lin et al., 2014; Bautista-de los Santos et al., 2016a). Families of *Burkholderiaceae, Methylophilaceae, Comamonadaceae,* and *Rhodocyclaceae* were abundant among *Betaproteobacteria*, whereas *Sphingomonadaceae*, *Caulobacteraceae,* and *Methylobacteriaceae* were dominant in *Alphaproteobacteria* (Eichler et al., 2006; Pinto et al., 2012; Vaz-Moreira et al., 2013; Zeng et al., 2013; Douterelo et al., 2014b). It is however difficult to define core microbiome down to genus or species levels across systems using different disinfectants (chlorine, chloramine, and without disinfectants), likely due to the substantial differences in selection pressures from theses disinfectants (Bautista-de los Santos et al., 2016a). Few studies have attempted to define the active core microbiome. Furthermore, it might be impossible to define core microbiome in the biofilm phase of DSs owing to the numerous but spatially heterogeneous ecological niches and continual ecological succession (Ling et al., 2016). When using 16S rRNA gene as the biomarker to define core microbiome, caution should be taken in relation to its limitations in differentiating closely-related populations at lower phylogenetic levels (e.g., genus or species), and in its ability to predict microbial functions associated with a given microbial community. Often different species within a genus do not carry out the same microbial function.

Microbial interactions Microbes in a DWS are present in the bulk water phase or the biofilm phase, and exchange of microorganisms between these two phases are anticipated. Unlike water-phase populations, which have a short transit time within a DWS, biofilm-phase microbes can be viewed as the indigenous populations in a DWS. They are organized in highly-structured habitats and exhibit considerable structural, chemical and biological heterogeneity (Allen et al., 1980; Ridgway and Olson, 1981; Wimpenny et al., 2000; Stewart and Franklin, 2008). In the past studies, biofilms were taken as composite samples along the pipe wall for microbial community analysis.

However, biofilm likely exhibited heterogeneity along the radial direction in the pipe wall. Contrary to intuition, the findings by Ridgway and Olson (Liu et al., 2017b) revealed that biofilm located in the middle part of pipe walls possessed the highest diversity and harbored the highest abundance of possible pathogens. Studies have shown that biofilm assemblages could influence the bulk water communities in the DS, and the effect was dependent on the level of biofilm sloughing from the pipe surface (Henne et al., 2012; Douterelo et al., 2013; Roeselers et al., 2015; Ling et al., 2016; Douterelo et al., 2017). A recent study estimated that the sloughing of 20% biofilm from PVC pipes or 10% biofilm from HDPE pipes would significantly alter the bulk water community (Liu et al., 2017a). In an extreme situation, when flushing is practiced by water utilities to remove any loosely adhered material, significant differences of the mobilized material between plastic and cast iron pipe sections could be observed (Douterelo et al., 2014a).

Another form of microbial interaction in a DWS is between eukaryotes and bacterial populations. Eukaryotes are an important component of the microbial assemblages in DWSs, many of which are resistant to disinfection processes and can feed on bacteria by phagocytosis, creating disturbances to the bacterial community (Griffiths et al., 1999; Ronn et al., 2002; De Mesel et al., 2004; Pernthaler, 2005; Bell et al., 2010; Jousset, 2012; Haig et al., 2015). Reported eukaryotic groups in DWSs included amoebae, nematodes, fungi, flagellates, segmented worms, arthropods, and flat worms (Delafont et al., 2016; Zhang et al., 2017; Oh et al., 2018). Some bacteria, including many opportunistic pathogens and closely-related species, have developed resistance mechanisms against phagocytosis and can use eukaryotic cells as hosts to protect themselves against stresses in DWSs (Cervero-Arago et al., 2015; Delafont et al., 2016; Buse et al., 2017). Moreover, many endosymbionts of free-living amoebae, including *Chlamydiae*, are potential pathogens and found to be prevalent in drinking water environments by a recent study (Zhang et al., 2017).

Biogeography Biogeography refers to the geographical distributions of organisms over the Earth in both space and time (Beijerinck, 1913; Horner-Devine et al., 2004).

Geographical differences in microbiomes have been observed for waste-treating ecosystems like anaerobic digester sludge (Mei et al., 2017). However, few studies have attempted to verify the existence of geographic difference with drinking water microbiomes (Roeselers et al., 2015) or understand which environmental factors exert the strongest influences (Horner-Devine et al., 2004; Bautista-de los Santos et al., 2016a). A survey of drinking water microbiomes along the Mississippi River found that the drinking water microbiota of New Orleans, LA differed from other communities surveyed with high relative abundances of phylotypes, indicative of fresh and saltwater infiltration (e.g., *Planctomycetes* and *Bacteroidetes*) and potential opportunistic pathogens (e.g., *Legionella* and *Mycobacterium* spp.) (Holinger et al., 2014; Hull et al., 2017). This survey further observed that the abundant taxa were generally shared among all systems and system-specific taxa were not abundant (Holinger et al., 2014). Similar findings were reported among systems across a restricted area. Roeselers et al. (Roeselers et al., 2015) surveyed 32 drinking water distribution networks in the Netherlands, all using groundwater from (un)confined sandy aquifers as the source water and no disinfectant residual in the networks, and observed network-specific taxa, which were of low abundances. However, these studies investigated diversity only through 16S rRNA gene amplicon analyses. Future studies revealing diversity at different scales or levels of resolution are needed for accurate description of the biogeography of drinking water microbiomes.

2.5 Potential of using meta-omics techniques to study drinking water microbiomes

Current studies in drinking water microbiomes primarily investigate “who is there under what conditions?”. It is important to ask additional questions that include but are not limited to “what are they doing?”, “why are they there?” and more critically “who is doing what?”, and “what are the interrelationships among them, and between them and their environment?" (Rittmann et al., 2006). These critical questions can be systematically addressed using NGS and meta-omics technologies (Figure 2.3).

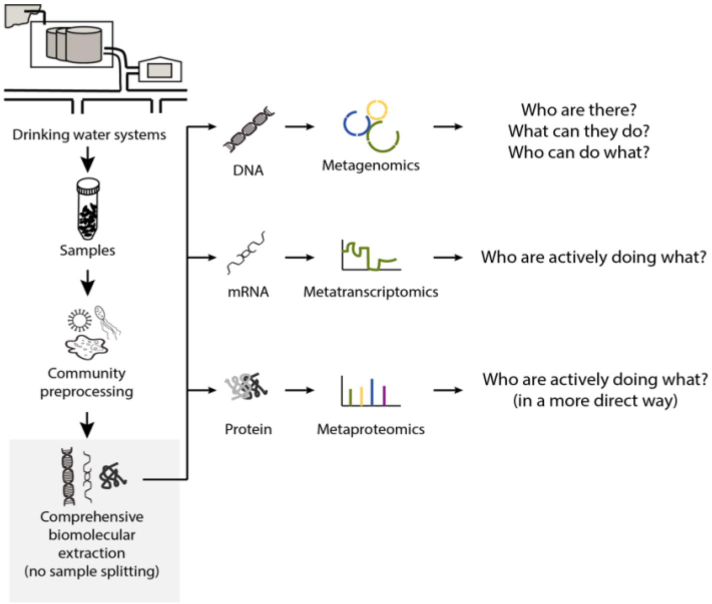


Figure 2.3 Types of meta-omics and the key questions that can be addressed by the meta-omics tools in the study of drinking water microbiomes.

NGS and meta-omics technologies The advance in NGS technology in the last decade serves as the pivotal force to the development of omics tools. The Sanger method is considered as a "first-generation" technology, and newer methods using new sequencing chemistry are referred to as NGS, which includes "454", "Illumina", PacBio SMRT and the Oxford Nanopore MinION. "454" and Illumina platforms can produce a large number of sequences that are low in cost, high in throughput and accuracy, and short in run times (Maxam and Gilbert, 1977). They produce short reads (Ross et al., 2013), which makes downstream bioinformatics analyses difficult. For this reason, technologies such as PacBio SMRT and the Oxford Nanopore MinION were developed to produce long reads (> 5 kb) but at a higher error rate (12-14% for PacBio and 8% for Nanopore) (Roberts et al., 2013; Feng et al., 2015; Jain et al., 2016). Using these NGS technologies alone or in combination has enabled the development of metagenomics and metatranscriptomics to study microbial functions and activities in various ecosystems (Beja et al., 2000; Tyson et al., 2004; Venter et al., 2004; Urich et al., 2008; Giannoukos et al., 2012; Xiong et al., 2012).

Metagenomics or environmental genomics is the genomic analysis of microorganisms in a microbial community that can provide insights into community physiology (Handelsman, 2005; Sharpton, 2014), and enables the discovery of new microbial taxa and genes without cultivation. The procedure involves extracting DNA from all cells in a community, shearing DNA into fragments, sequencing fragmented DNA (Handelsman, 2004; Tyson et al., 2004; Venter et al., 2004), assembling all sequences into an ecosystem genome comprised of many genomes of the innate microbial populations ("metagenome") (Handelsman, 2004), and phylogenetically classifying the genomic fragments to specific microorganisms ("binning") (McHardy et al., 2007). This approach has been greatly improved by using novel assemblers (*e.g.*, metaSPAdes and MEGAHIT) (Namiki et al., 2012; Peng et al., 2012; Li et al., 2015a; Nurk et al., 2016) and binning methods (McHardy et al., 2007; Pati et al., 2011; Patil et al., 2012; Wu et al., 2014) together with software that integrate information from essential single-copy genes (*e.g.*, MaxBin) and multiple metagenomes of related samples (*e.g.*, MetaBAT and GroopM) (Albertsen et al., 2013; Imelfort et al., 2014; Kang et al., 2015; Wu et al., 2016). Researchers can now determine individual bins' phylogeny ("phylogenomics") using software such as PhyloPhlAn (Segata et al., 2013) and genome completeness/contamination with marker genes using software such as CheckM (Parks et al., 2015). These advancements enable accurate metagenomic assembly, binning, and recovery of genomes for phylogenetically novel organisms without cultivating them (Wrighton et al., 2012; Wu et al., 2014; Kang et al., 2015).

Metatranscriptomics is based on sequencing the total message RNA (mRNA) in a microbial community to identify genes or pathways that are actively expressed. This process involves extracting total RNA from microbial communities, removing ribosome RNA (rRNA) to obtain high levels of mRNA transcripts, reverse transcribing mRNA into cDNAs, ligating to adapters, and then sequencing using NGS (He et al., 2010b; Sorek and Cossart, 2010). This method is often used together with metagenomics to provide insight into microbial community functions and activities. Metatranscriptomics has been widely used in a variety of environments, including soil (Urich et al., 2008), sediment (Dumont et al., 2013), gut microbiomes (Giannoukos et al., 2012; Xiong et al., 2012), and activated sludge (He et al., 2010a; Yu and Zhang, 2012). It is also a powerful tool to identify novel pathways in uncultured microorganisms (Haroon et al., 2013).

Metaproteomics is developed to evaluate microbial activity within an ecosystem at a specific time based on protein expression (Wilmes and Bond, 2004; Zampieri et al., 2016). Unlike metagenomics and metratranscriptomics that use NGS technologies, metaproteomics uses liquid chromatography tandem mass spectrometry (LC-MS/MS). The process starts with extracting protein, followed by LC-MS/MS to generate MS spectra, and then comparing spectra with peptides from thousands of proteins of diverse taxonomic groups. These comparisons can be achieved in two ways: through searching against existing protein/peptide databases or by matching to theoretical peptide spectra generated *in silico* from metagenomes of the same sample or of similar environments (Zampieri et al., 2016; Timmins-Schiffman et al., 2017). Metaproteomics is a powerful tool to unravel the active metabolic processes in different environments in a more direct way than metagenomics or metatranscriptomics. This approach has been applied to complicated environments, including soils (Benndorf et al., 2007; Williams et al., 2010; Wang et al., 2011), sediments (Benndorf et al., 2009; Bruneel et al., 2011), marine habitats (Morris et al., 2010; Sowell et al., 2011), freshwater systems (Ng et al., 2010; Habicht et al., 2011; Lauro et al., 2011), and activated sludge (Wilmes et al., 2008).

Applications of meta-omics in drinking water microbiome studies While many studies have applied omics tools in various microbial ecosystems, only a limited number of studies have been applied to study drinking water microbiomes. Some of these studies are based on the use of cosmid library construction that is low in sequence throughput or early NGS technologies that cannot derive long assembled contigs to provide correct linkage between microbes and functionalities (Schmeisser et al., 2003; Chistoserdova, 2014). Using NGS techniques, two studies (Gomez-Alvarez et al., 2012; Chao et al., 2013) investigated the impact of water treatment on drinking water microbiome. Their results revealed that chlorine and chloramine treatments caused differences in community structures, disinfectant mechanisms, and virulence genes (Gomez-Alvarez et al., 2012). Changes in protective functions (i.e., glutathione synthesis) were observed in treated water compared with raw water (Chao et al., 2013). Oh et al. (Oh et al., 2018) applied metagenomics to understand how microorganisms inhabiting filtration media could be beneficial to water production in a full-scale treatment plant configured with a RSF, GAC, SSF, and the top layer of SSF known as *Schmutzdecke* that is biologically active. The findings revealed that the filter bacterial communities significantly differed from those in the source water and final effluent communities, respectively. *Bradyrhizobiaceae* were abundant in GAC, whereas *Nitrospira* were enriched in the sand-associated filters (RSF, SCM, and SSF). The GAC community was enriched with functions associated with aromatics degradation, many of which were encoded by *Rhizobiales* (~ 30% of the total GAC community). Findings further suggested that the GAC community potentially selected fast-growers among the four filter communities, consistent with the highest dissolved organic matter removal rate observed with GAC.

Limitations of the meta-omics technologies Applying meta-omics in drinking water microbiome studies can face several challenges. The first one can be related to sample preparation, as a large quantity of genomic DNA, RNA and protein is required for downstream sequencing and LC-MS/MS analyses. For drinking water microbiomes in the water phase, sampling a large volume of water (i.e., over 1000 L) is often required for systems containing residual disinfectant. As conventional concentration devices are not suitable for this purpose (Chao et al., 2013; Zhang et al., 2017), studies have used point-of-use water purifiers (Chao et al., 2013; Zhang et al., 2017), which involve more than one mechanism to concentrate cells and the biases remain unknown. Thus, there is a need to standardize a device for concentrating large volumes of drinking water. Likewise, studying biofilm-phase drinking water microbiome in full-scale DWSs can be challenging (Gomez-Smith et al., 2015; Ling et al., 2016). Currently, two approaches are used: one is to cut pipes and the other is to insert coupons into pipes and retrieve them after biofilms develop (Douterelo et al., 2014b). The former is labor-intensive, expensive, and prone to contamination from surrounding environments, and the latter can distort hydraulic conditions in pipes and cause deviations from real pipes. An alternative solution is to sample biofilms from the inner surface of water meters (Hong et al., 2010). For metatranscriptomics and metaproteomics studies, sampling preparation needs to be carefully evaluated (Hansen et al., 2014) because mRNA is liable to degradation by RNases that are ubiquitously present in the environment. Proper stabilization and storage procedures are critical to obtain sufficient quantities of high-quality mRNA. Lastly, meta-omics studies can generate huge datasets that require vigorous bioinformatics analyses and computing capacity (Thomas et al., 2012). However, errors associated with bioinformatics can be problematic and substantially influence the final interpretation (Kunin et al., 2008; Timmins-Schiffman et al., 2017). Most studies are required to establish curated databases of their interests, partially because of the scattered data submitted and stored in various databases. At present, several web-based pipelines are available, including MG-RAST (Meyer et al., 2008), KBase (Arkin et al., 2016), CyVerse/iPlant Discovery Environment (Goff et al., 2011), and IMG-ER (Markowitz et al., 2012). Future studies will require advanced tools to simultaneously interact with multiple databases for microbial genomes, metagenomes, protein, antibiotic resistance genes, and viral genomes. The next step is to develop a systematic data management framework by leveraging current and future plans for expanding our understanding of drinking water microbiomes globally.

2.6 References

Albers, C.N., Ellegaard-Jensen, L., Harder, C.B., Rosendahl, S., Knudsen, B.E., Ekelund, F., and Aamand, J. (2015) Groundwater chemistry determines the prokaryotic community structure of waterworks sand filters. *Environ Sci Technol* **49**: 839-846.

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

Allen, M.J., Taylor, R.H., and Geldreich, E.E. (1980) The occurrence of microorganisms in water main encrustations. *J Am Water Works Assoc* **72**: 614-625.

Amann, R.I., Krumholz, L., and Stahl, D.A. (1990) Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental-studies in microbiology. *J Bacteriol* **172**: 762-770.

Angly, F.E., Dennis, P.G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P., and Tyson, G.W. (2014) CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* **2**.

APHA, AWWA, and WEF (1998) *Standard Methods for the Examination of Water and Wastewater*. Washington, DC: American Publica Health Association.

Arkin, A.P., Stevens, R.L., Cottingham, R.W., Maslov, S., Henry, C.S., Dehal, P. et al. (2016) The DOE systems biology knowledgebase (KBase). *bioRxiv*.

Bartram, J., Cotruvo, J., Exner, M., Fricker, C., and Glasmacher, A. (2004) Heterotrophic plate count measurement in drinking water safety management - Report of an Expert Meeting Geneva, 24-25 April 2002. *Int J Food Microbiol* **92**: 241-247.

Bautista-de los Santos, Q.M., Schroeder, J.L., Sevillano-Rivera, M.C., Sungthong, R., Ijaz, U.Z., Sloan, W.T., and Pinto, A.J. (2016a) Emerging investigators series: microbial communities in full-scale drinking water distribution systems - a meta-analysis. *Environ Sci: Water Res Technol* **2**: 631-644.

Bautista-de los Santos, Q.M., Schroeder, J.L., Blakemore, O., Moses, J., Haffey, M., Sloan, W., and Pinto, A.J. (2016b) The impact of sampling, PCR, and sequencing replication on discerning changes in drinking water bacterial community over diurnal time-scales. *Water Res* **90**: 216-224.

Beech, W.B., and Sunner, J. (2004) Biocorrosion: towards understanding interactions between biofilms and metals. *Curr Opin Biotechnol* **15**: 181-186.

Beijerinck, M. (1913) *De Infusies en de Ontdekking der Backterien, Jaarboek van de Koninklijke Akademie v. Wetenschappen*. Muller, Amsterdam, the Netherlands.

Beja, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. et al. (2000) Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1906.

Bell, T., Bonsall, M.B., Buckling, A., Whiteley, A.S., Goodall, T., and Griffiths, R.I. (2010) Protists have divergent effects on bacterial diversity along a productivity gradient. *Biol Lett* **6**: 639-642.

Benndorf, D., Balcke, G.U., Harms, H., and von Bergen, M. (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME J* **1**: 224-234.

Benndorf, D., Vogt, C., Jehmlich, N., Schmidt, Y., Thomas, H., Woffendin, G. et al. (2009) Improving protein extraction and separation methods for investigating the metaproteome of anaerobic benzene communities within sediments. *Biodegradation* **20**: 737-750.

Berney, M., Vital, M., Hulshoff, I., Weilenmann, H.U., Egli, T., and Hammes, F. (2008) Rapid, cultivation-independent assessment of microbial viability in drinking water. *Water Res* **42**: 4010-4018.

Berry, D., Xi, C., and Raskin, L. (2006) Microbial ecology of drinking water distribution systems. *Curr Opin Biotechnol* **17**: 297-302.

Besmer, M.D., Epting, J., Page, R.M., Sigrist, J.A., Huggenberger, P., and Hammes, F. (2016) Online flow cytometry reveals microbial dynamics influenced by concurrent natural and operational events in groundwater used for drinking water treatment. *Sci Rep* **6**.

Bodrossy, L., and Sessitsch, A. (2004) Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol* **7**: 245-254.

Brooks, J.P., Edwards, D.J., Harwich, M.D., Rivera, M.C., Fettweis, J.M., Serrano, M.G. et al. (2015) The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol* **15**.

Bruneel, O., Volant, A., Gallien, S., Chaumande, B., Casiot, C., Carapito, C. et al. (2011) Characterization of the active bacterial community involved in natural attenuation processes in arsenic-rich creek sediments. *Microb Ecol* **61**: 793-810.

Bruno, A., Sandionigi, A., Rizzi, E., Bernasconi, M., Vicario, S., Galimberti, A. et al. (2017) Exploring the under-investigated "microbial dark matter" of drinking water treatment plants. *Sci Rep* **7**.

Buse, H.Y., Ji, P., Gomez-Alvarez, V., Pruden, A., Edwards, M.A., and Ashbolt, N.J. (2017) Effect of temperature and colonization of Legionella pneumophila and Vermamoeba vermiformis on bacterial community composition of copper drinking water biofilms. *Microb Biotechnol* **10**: 773-788.

Camper, A., Burr, M., Ellis, B., Butterfield, P., and Abernathy, C. (1999) Development and structure of drinking water biofilms and techniques for their study. *J Appl Microbiol* **85**: 1s-12s.

Cerrato, J.M., Falkinham, J.O., Dietrich, A.M., Knocke, W.R., McKinney, C.W., and Pruden, A. (2010) Manganese-oxidizing and -reducing microorganisms isolated from biofilms in chlorinated drinking water systems. *Water Res* **44**: 3935-3945.

Cervero-Arago, S., Rodriguez-Martinez, S., Puertas-Bennasar, A., and Araujo, R.M. (2015) Effect of common drinking water disinfectants, chlorine and heat, on free Legionella and Amoebae-associated Legionella. *PLoS One* **10**.

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X.X., Wu, W.M., and Zhang, T. (2013) Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Sci Rep* **3**: 3550.

Chen, L., Jia, R.B., and Li, L. (2013) Bacterial community of iron tubercles from a drinking water distribution system and its occurrence in stagnant tap water. *Environ Sci Process Impacts* **15**: 1332-1340.

Chen, N.T., and Chang, C.W. (2010) Rapid quantification of viable legionellae in water and biofilm using ethidium monoazide coupled with real-time quantitative PCR. *J Appl Microbiol* **109**: 623-634.

Chiao, T.H., Clancy, T.M., Pinto, A., Xi, C.W., and Raskin, L. (2014) Differential resistance of drinking water bacterial populations to monochloramine disinfection. *Environ Sci Technol* **48**: 4038-4047.

Chistoserdova, L. (2014) Is metagenomics resolving identification of functions in microbial communities? *Microb Biotechnol* **7**: 1-4.

Dai, D., Prussin, A.J., 2nd, Marr, L.C., Vikesland, P.J., Edwards, M.A., and Pruden, A. (2017) Factors shaping the human exposome in the built environment: opportunities for engineering control. *Environ Sci Technol* **51**: 7759-7774.

De Mesel, I., Derycke, S., Moens, T., Van der Gucht, K., Vincx, M., and Swings, J. (2004) Top-down impact of bacterivorous nematodes on the bacterial community structure: a microcosm study. *Environ Microbiol* **6**: 733-744.

de Vet, W.W.J.M., Dinkla, I.J.T., Muyzer, G., Rietveld, L.C., and van Loosdrecht, M.C.M. (2009) Molecular characterization of microbial populations in groundwater sources and sand filters for drinking water production. *Water Res* **43**: 182-194.

Delafont, V., Bouchon, D., Hechard, Y., and Moulin, L. (2016) Environmental factors shaping cultured free-living amoebae and their associated bacterial community within drinking water network. *Water Res* **100**: 382-392.

Delong, E.F., Wickham, G.S., and Pace, N.R. (1989) Phylogenetic stains - ribosomal RNA-based probes for the identification of single cells. *Science* **243**: 1360-1363.

Douterelo, I., Sharpe, R.L., and Boxall, J.B. (2013) Influence of hydraulic regimes on bacterial community structure and composition in an experimental drinking water distribution system. *Water Res* **47**: 503-516.

Douterelo, I., Husband, S., and Boxall, J.B. (2014a) The bacteriological composition of biomass recovered by flushing an operational drinking water distribution system. *Water Res* **54**: 100-114.

Douterelo, I., Jackson, M., Solomon, C., and Boxall, J. (2017) Spatial and temporal analogies in microbial communities in natural drinking water biofilms. *Sci Total Environ* **581**: 277-288.

Douterelo, I., Boxall, J.B., Deines, P., Sekar, R., Fish, K.E., and Biggs, C.A. (2014b) Methodological approaches for studying the microbial ecology of drinking water distribution systems. *Water Res* **65**: 134-156.

Duhaime, M.B., Deng, L., Poulos, B.T., and Sullivan, M.B. (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ Microbiol* **14**: 2526-2537.

Dumont, M.G., Pommerenke, B., and Casper, P. (2013) Using stable isotope probing to obtain a targeted metatranscriptome of aerobic methanotrophs in lake sediment. *Environ Microbiol Rep* **5**: 757-764.

Edberg, S.C. (2009) Does the possession of virulence factor genes mean that those genes will be active? *J Water Health* **7**: S19-S28.

Edelstein, P.H. (1981) Improved semiselective medium for isolation of *Legionella pneumophila* from contaminated clinical and environmental specimens. *J Clin Microbiol* **14**: 298-303.

Eichler, S., Christen, R., Holtje, C., Westphal, P., Botel, J., Brettar, I. et al. (2006) Composition and dynamics of bacterial communities of a drinking water supply system as assessed by RNA- and DNA-based 16S rRNA gene fingerprinting. *Appl Environ Microbiol* **72**: 1858-1872.

El Fantroussi, S., Urakawa, H., Bernhard, A.E., Kelly, J.J., Noble, P.A., Smidt, H. et al. (2003) Direct profiling of environmental microbial populations by thermal dissociation analysis of native rRNAs hybridized to oligonucleotide microarrays. *Appl Environ Microbiol* **69**: 2377-2382.

Feng, S., Chen, C., Wang, Q.F., Zhang, X.J., Yang, Z.Y., and Xie, S.G. (2013) Characterization of microbial communities in a granular activated carbon-sand dual media filter for drinking water treatment. *Int J Environ Sci Technol* **10**: 917-922.

Feng, Y.X., Zhang, Y.C., Ying, C.F., Wang, D.Q., and Du, C.L. (2015) Nanopore-based fourth-generation DNA Sequencing technology. *Genomics Proteomics Bioinformatics* **13**: 200-201.

Fischer, E.R., Hansen, B.T., Nair, V., Hoyt, F.H., and Dorward, D.W. (2005) Scanning electron microscopy. In *Curr Protoc Microbiol*: John Wiley & Sons, Inc.

Flynn, T.M., Sanford, R.A., Ryu, H., Bethke, C.M., Levine, A.D., Ashbolt, N.J., and Domingo, J.W.S. (2013) Functional microbial diversity explains groundwater chemistry in a pristine aquifer. *BMC Microbiol* **13**.

Frankland, P. (1894) *Micro-organisms in water: Their significance, identification and removal, together with an account of the bacteriological methods employed in their investigaion, specially designed for the use of those connected with the sanitary aspects of water-supply*: Longmans, Green.

Giannoukos, G., Ciulla, D.M., Huang, K., Haas, B.J., Izard, J., Levin, J.Z. et al. (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol* **13**.

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E., Gessler, D. et al. (2011) The iPlant collaborative: cyberinfrastructure for plant biology. *Front Plant Sci* **2**.

Gomez-Alvarez, V., Revetta, R.P., and Domingo, J.W.S. (2012) Metagenomic analyses of drinking water receiving different disinfection treatments. *Appl Environ Microbiol* **78**: 6095-6102.

Gomez-Alvarez, V., Humrighouse, B.W., Revetta, R.P., and Domingo, J.W.S. (2015) Bacterial composition in a metropolitan drinking water distribution system utilizing different source waters. *J Water Health* **13**: 140-151.

Gomez-Smith, C.K., LaPara, T.M., and Hozalski, R.M. (2015) Sulfate reducing bacteria and mycobacteria dominate the biofilm communities in a chloraminated drinking water distribution system. *Environ Sci Technol* **49**: 8432-8440.

Griffiths, B.S., Bonkowski, M., Dobson, G., and Caul, S. (1999) Changes in soil microbial community structure in the presence of microbial-feeding nematodes and protozoa. *Pedobiologia* **43**: 297-304.

Gulay, A., Musovic, S., Albrechtsen, H.J., Abu Al-Soud, W., Sorensen, S.J., and Smets, B.F. (2016) Ecological patterns, diversity and core taxa of microbial communities in groundwater-fed rapid gravity filters. *ISME J* **10**: 2209-2222.

Gunsalus, R.P., Zeikus, J.G., and Wolfe, R.S. (1972) Microbial modification of ground water.

Guo, F., and Zhang, T. (2013) Biases during DNA extraction of activated sludge samples revealed by high throughput sequencing. *Appl Microbiol Biotechnol* **97**: 4607-4616.

Guschin, D.Y., Mobarry, B.K., Proudnikov, D., Stahl, D.A., Rittmann, B.E., and Mirzabekov, A.D. (1997) Oligonucleotide microchips as genosensors for determinative and environmental studies in microbiology. *Appl Environ Microbiol* **63**: 2397-2402.

Habicht, K.S., Miller, M., Cox, R.P., Frigaard, N.U., Tonolla, M., Peduzzi, S. et al. (2011) Comparative proteomics and activity of a green sulfur bacterium through the water column of Lake Cadagno, Switzerland. *Environl Microbiol* **13**: 203-215.

Haig, S.J., Schirmer, M., D'Amore, R., Gibbs, J., Davies, R.L., Collins, G., and Quince, C. (2015) Stable-isotope probing and metagenomics reveal predation by protozoa drives E-coli removal in slow sand filters. *Isme Journal* **9**: 797-808.

Hammes, F., Berney, M., Wang, Y.Y., Vital, M., Koster, O., and Egli, T. (2008) Flow-cytometric total bacterial cell counts as a descriptive microbiological parameter for drinking water treatment processes. *Water Res* **42**: 269-277.

Hammes, F., Boon, N., Vital, M., Ross, P., Magic-Knezev, A., and Dignum, M. (2011) Bacterial colonization of pellet softening reactors used during drinking water treatment. *Appl Environ Microbiol* **77**: 1041-1048.

Handelsman, J. (2004) Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**: 669-+.

Handelsman, J. (2005) Metagenomics: Application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **69**: 195-195.

Hansen, S.H., Stensballe, A., Nielsen, P.H., and Herbst, F.A. (2014) Metaproteomics: Evaluation of protein extraction from activated sludge. *Proteomics* **14**: 2535-2539.

Haroon, M.F., Hu, S.H., Shi, Y., Imelfort, M., Keller, J., Hugenholtz, P. et al. (2013) Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* **500**: 567.

He, S.M., Kunin, V., Haynes, M., Martin, H.G., Ivanova, N., Rohwer, F. et al. (2010a) Metatranscriptomic array analysis of 'Candidatus Accumulibacter phosphatis'-enriched enhanced biological phosphorus removal sludge. *Environ Microbiol* **12**: 1205-1217.

He, S.M., Wurtzel, O., Singh, K., Froula, J.L., Yilmaz, S., Tringe, S.G. et al. (2010b) Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods* **7**: 807-U858.

Henderson, G., Cox, F., Kittelmann, S., Miri, V.H., Zethof, M., Noel, S.J. et al. (2013) Effect of DNA extraction methods and sampling techniques on the apparent structure of cow and sheep rumen microbial communities. *PLoS One* **8**.

Henne, K., Kahlisch, L., Brettar, I., and Hofle, M.G. (2012) Analysis of structure and composition of bacterial core communities in mature drinking water biofilms and bulk water of a citywide network in germany. *Appl Environ Microbiol* **78**: 3530-3538.

Henne, K., Kahlisch, L., Hofle, M.G., and Brettar, I. (2013) Seasonal dynamics of bacterial community structure and composition in cold and hot drinking water derived from surface water reservoirs. *Water Res* **47**: 5614-5630.

Holinger, E.P., Ross, K.A., Robertson, C.E., Stevens, M.J., Harris, J.K., and Pace, N.R. (2014) Molecular analysis of point-of-use municipal drinking water microbiology. *Water Res* **49**: 225-235.

Hong, P.Y., Hwang, C.C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2010) Pyrosequencing analysis of bacterial biofilm communities in water meters of a drinking water distribution system. *Appl Environ Microbiol* **76**: 5631-5635.

Horner-Devine, M.C., Carney, K.M., and Bohannan, B.J.M. (2004) An ecological perspective on bacterial biodiversity. *Proc R Soc Lond* **271**: 113-122.

Hull, N.M., Holinger, E.P., Ross, K.A., Robertson, C.E., Harris, J.K., Stevens, M.J., and Pace, N.R. (2017) Longitudinal and Source-to-Tap New Orleans, LA, USA Drinking Water Microbiology. *Environmental Science & Technology* **51**: 4220-4229.

Hwang, C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2012a) Microbial community dynamics of an urban drinking water distribution system subjected to phases of chloramination and chlorination treatments. *Appl Environ Microbiol* **78**: 7856-7865.

Hwang, C.C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2012b) Evaluation of methods for the extraction of DNA from drinking water distribution system biofilms. *Microbes Environ* **27**: 9-18.

Imelfort, M., Parks, D., Woodcroft, B.J., Dennis, P., Hugenholtz, P., and Tyson, G.W. (2014) GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**: e603.

Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T., and Aluru, S. (2017) High-throughput ANI Analysis of 90K prokaryotic genomes reveals clear species boundaries. *bioRxiv*.

Jain, M., Olsen, H.E., Paten, B., and Akeson, M. (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community (vol 17, 239, 2016). *Genome Biol* **17**.

Ji, P., Parks, J., Edwards, M.A., and Pruden, A. (2015) Impact of Water Chemistry, Pipe Material and Stagnation on the Building Plumbing Microbiome. *PLoS One* **10**: e0141087.

Jin, J., Wu, G., and Guan, Y. (2015) Effect of bacterial communities on the formation of cast iron corrosion tubercles in reclaimed water. *Water Res* **71**: 207-218.

Jousset, A. (2012) Ecological and evolutive implications of bacterial defences against predators. *Environmental Microbiology* **14**: 1830-1843.

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

Kasuga, I., Nakagaki, H., Kurisu, F., and Furumai, H. (2010a) Predominance of ammonia-oxidizing archaea on granular activated carbon used in a full-scale advanced drinking water treatment plant. *Water Res* **44**: 5039-5049.

Kasuga, I., Nakagaki, H., Kurisu, F., and Furumai, H. (2010b) Abundance and diversity of ammonia-oxidizing archaea and bacteria on biological activated carbon in a pilot-scale drinking water treatment plant with different treatment processes. *Water Sci Technol* **61**: 3070-3077.

Keiblinger, K.M., Wilhartitz, I.C., Schneider, T., Roschitzki, B., Schmid, E., Eberl, L. et al. (2012) Soil metaproteomics - Comparative evaluation of protein extraction protocols. *Soil Biol Biochem* **54**: 14-24.

Kelly, J.J., Minalt, N., Culotti, A., Pryor, M., and Packman, A. (2014) Temporal variations in the abundance and composition of biofilm communities colonizing drinking water distribution pipes. *PLoS ONE* **9**: e98542.

Klappenbach, J.A., Saxman, P.R., Cole, J.R., and Schmidt, T.M. (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res* **29**: 181-184.

Koch, R., and Duncan, G. (1894) *Professor Koch on the Bacteriological Diagnosis of Cholera, Water-filtration and Cholera, and the Cholera in Germany During the Winter of 1892-93. Translated by G. Duncan, Etc. [Reprinted from the "Scotsman."]*: Edinburgh.

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008) A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* **72**: 557-578, Table of Contents.

Lauro, F.M., DeMaere, M.Z., Yau, S., Brown, M.V., Ng, C., Wilkins, D. et al. (2011) An integrative study of a meromictic lake ecosystem in Antarctica. *ISME J* **5**: 879-895.

Lautenschlager, K., Boon, N., Wang, Y., Egli, T., and Hammes, F. (2010) Overnight stagnation of drinking water in household taps induces microbial growth and changes in community composition. *Water Res* **44**: 4868-4877.

Lautenschlager, K., Hwang, C., Liu, W.T., Boon, N., Koster, O., Vrouwenvelder, H. et al. (2013) A microbiology-based multi-parametric approach towards assessing biological stability in drinking water distribution networks. *Water Res* **47**: 3015-3025.

Lautenschlager, K., Hwang, C., Ling, F., Liu, W.T., Boon, N., Koster, O. et al. (2014) Abundance and composition of indigenous bacterial communities in a multi-step biofiltration-based drinking water treatment plant. *Water Res* **62**: 40-52.

LeChevallier, M.W., Welch, N.J., and Smith, D.B. (1996) Full-scale studies of factors related to coliform regrowth in drinking water. *Appl Environ Microbiol* **62**: 2201-2211.

Lee, C.K., Herbold, C.W., Polson, S.W., Wommack, K.E., Williamson, S.J., McDonald, I.R., and Cary, S.C. (2012) Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS One* **7**.

Lee, E.S., Lee, M.H., and Kim, B.S. (2015) Evaluation of propidium monoazide-quantitative PCR to detect viable Mycobacterium fortuitum after chlorine, ozone, and ultraviolet disinfection. *Int J Food Microbiol* **210**: 143-148.

Lesnik, R., Brettar, I., and Hofle, M.G. (2016) *Legionella* species diversity and dynamics from surface reservoir to tap water: from cold adaptation to thermophily. *ISME J* **10**: 1064-1080.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015a) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*: btv033.

Li, D., Li, Z., Yu, J., Cao, N., Liu, R., and Yang, M. (2010) Characterization of bacterial community structure in a drinking water distribution system during an occurrence of red water. *Appl Environ Microbiol* **76**: 7171-7180.

Li, E.S.Y., and Liu, W.T. (2004) DNA microarray technology in microbial ecology studies- principle, application and current limitations. *Microbes Environ* **18**: 175-187.

Li, X.X., Wang, H.B., Hu, C., Yang, M., Hu, H.Y., and Niu, J.F. (2015b) Characteristics of biofilms and iron corrosion scales with ground and surface waters in drinking water distribution systems. *Corros Sci* **90**: 331-339.

Liao, X.B., Chen, C., Wang, Z., Wan, R., Chang, C.H., Zhang, X.J., and Xie, S.G. (2013) Pyrosequencing analysis of bacterial communities in drinking water biofilters receiving influents of different types. *Process Biochem* **48**: 703-707.

Lin, W.F., Yu, Z.S., Zhang, H.X., and Thompson, I.P. (2014) Diversity and dynamics of microbial communities at each step of treatment plant for potable water generation. *Water Res* **52**: 218-230.

Ling, F., Hwang, C., LeChevallier, M.W., Andersen, G.L., and Liu, W.T. (2016) Core-satellite populations and seasonality of water meter biofilms in a metropolitan drinking water distribution system. *ISME J* **10**: 582-595.

Liu, G., Verberk, J.Q., and Van Dijk, J.C. (2013a) Bacteriology of drinking water distribution systems: an integral and multidimensional review. *Appl Microbiol Biotechnol* **97**: 9265-9276.

Liu, G., Ling, F.Q., Magic-Knezev, A., Liu, W.T., Verberk, J.Q.J.C., and Van Dijk, J.C. (2013b) Quantification and identification of particle-associated bacteria in unchlorinated drinking water from three treatment plants by cultivation-independent methods. *Water Res* **47**: 3523-3533.

Liu, G., Bakker, G.L., Li, S., Vreeburg, J.H.G., Verberk, J.Q.J.C., Medema, G.J. et al. (2014) Pyrosequencing reveals bacterial communities in unchlorinated drinking water distribution system: An integral study of bulk water, suspended solids, loose deposits, and pipe wall biofilm. *Environ Sci Technol* **48**: 5467-5476.

Liu, G., Tao, Y., Zhang, Y., Lut, M., Knibbe, W.J., van der Wielen, P. et al. (2017a) Hotspots for selected metal elements and microbes accumulation and the corresponding water quality deterioration potential in an unchlorinated drinking water distribution system. *Water Res* **124**: 435-445.

Liu, J.Q., Ren, H.X., Ye, X.B., Wang, W., Liu, Y., Lou, L.P. et al. (2017b) Bacterial community radial-spatial distribution in biofilms along pipe wall in chlorinated drinking water distribution system of East China. *Appl Microbiol Biotechnol* **101**: 749-759.

Liu, W.T., and Stahl, D.A. (2007) Molecular approaches for the measurement of density, diversity and phylogeny. In *Manual of Environmental Microbiology, Third Edition*. Washington DC: American Society for Microbiology, pp. 139-156.

Liu, W.T., Mirzabekov, A.D., and Stahl, D.A. (2001) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ Microbiol* **3**: 619-629.

Loy, A., Lehner, A., Lee, N., Adamczyk, J., Meier, H., Ernst, J. et al. (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Appl Environ Microbiol* **68**: 5064-5081.

Magic-Knezev, A., and van der Kooij, D. (2004) Optimisation and significance of ATP analysis for measuring active biomass in granular activated carbon filters used in water treatment. *Water Res* **38**: 3971-3979.

Manz, W., Szewzyk, U., Ericsson, P., Amann, R., Schleifer, K.H., and Stenstrom, T.A. (1993) In situ identification of bacteria in drinking water and adjoining biofilms by hybridization with 16S and 23S rRNA-directed fluorescent oligonucleotide probes. *Appl Environ Microbiol* **59**: 2293-2298.

Marcus, D.N., Pinto, A., Anantharaman, K., Ruberg, S.A., Kramer, E.L., Raskin, L., and Dick, G.J. (2017) Diverse manganese(II)-oxidizing bacteria are prevalent in drinking water systems. *Environ Microbiol Rep* **9**: 120-128.

Markowitz, V.M., Chen, I.M., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y. et al. (2012) IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* **40**: D115-122.

Maxam, A.M., and Gilbert, W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci USA* **74**: 560-564.

McHardy, A.C., Martin, H.G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods* **4**: 63-72.

Mei, R., Nobu, M.K., Narihiro, T., Kuroda, K., Sierra, J.M., Wu, Z.Y. et al. (2017) Operation-driven heterogeneity and overlooked feed-associated populations in global anaerobic digester microbiome. *Water Res* **124**: 77-84.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M. et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**.

Moran, M.A., Satinsky, B., Gifford, S.M., Luo, H.W., Rivers, A., Chan, L.K. et al. (2013) Sizing up metatranscriptomics. *ISME J* **7**: 237-243.

Morris, R.M., Nunn, B.L., Frazar, C., Goodlett, D.R., Ting, Y.S., and Rocap, G. (2010) Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J* **4**: 673-685.

Muller, E.E.L., Glaab, E., May, P., Vlassis, N., and Wilmes, P. (2013) Condensing the omics fog of microbial communities. *Trends Microbiol* **21**: 325-333.

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* **40**: e155.

Ng, C., DeMaere, M.Z., Williams, T.J., Lauro, F.M., Raftery, M., Gibson, J.A.E. et al. (2010) Metaproteogenomic analysis of a dominant green sulfur bacterium from Ace Lake, Antarctica. *ISME J* **4**: 1002-1019.

Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. (2016) metaSPAdes: a new versatile de novo metagenomics assembler. *ArXiv e-prints* **1604**: arXiv:1604.03071.

Oh, S., Hammes, F., and Liu, W.T. (2018) Metagenomic characterization of biofilter microbial communities in a full-scale drinking water treatment plant. *Water Res* **128**: 278-285.

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

Pasculle, A.W., Feeley, J.C., Gibson, R.J., Cordes, L.G., Myerowitz, R.L., Patton, C.M. et al. (1980) Pittsburgh pneumonia agent - direct isolation from human-lung tissue. *J Infect Dis* **141**: 727-732.

Pati, A., Heath, L.S., Kyrpides, N.C., and Ivanova, N. (2011) ClaMS: A Classifier for Metagenomic Sequences. *Stand Genomic Sci* **5**: 248-253.

Patil, K.R., Roune, L., and McHardy, A.C. (2012) The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS One* **7**: e38581.

Payment, P., Sartory, D., and Reasoner, D. (2003) The history and use of HPC in drinking-water quality management. In *Heterotrophic Plate Counts and Drinking-water Safety*. Bartram, J., Cotruvo, J., Exner, M., Fricker, C., and Glasmacher, A. (eds). Alliance House, UK: IWA Publishing, pp. 20-48.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420-1428.

Peplies, J., Glockner, F.O., and Amann, R. (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. *Appl Environ Microbiol* **69**: 1397-1407.

Peplies, J., Lau, S.C., Pernthaler, J., Amann, R., and Glockner, F.O. (2004) Application and validation of DNA microarrays for the 16S rRNA-based analysis of marine bacterioplankton. *Environ Microbiol* **6**: 638-645.

Pernthaler, J. (2005) Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* **3**: 537-546.

Peters, R.W. (2011) Water and wastewater engineering: Design principles and practice. In: Wiley Online Library.

Pinto, A.J., Xi, C.W., and Raskin, L. (2012) Bacterial community structure in the drinking water microbiome is governed by filtration processes. *Environ Sci Technol* **46**: 8851-8859.

Pinto, A.J., Schroeder, J., Lunn, M., Sloan, W., and Raskin, L. (2014) Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. *MBio* **5**: e01135-01114.

Poitelon, J.B., Joyeux, M., Welte, B., Duguet, J.P., Prestel, E., and DuBow, M.S. (2010) Variations of bacterial 16S rDNA phylotypes prior to and after chlorination for drinking water production from two surface water treatment plants. *J Ind Microbiol Biotechnol* **37**: 117-128.

Prest, E.I., Weissbrodt, D.G., Hammes, F., van Loosdrecht, M.C.M., and Vrouwenvelder, J.S. (2016) Long-term bacterial dynamics in a full-scale drinking water distribution system. *PLoS One* **11**.

Proctor, C.R., and Hammes, F. (2015) Drinking water microbiology-from measurement to management. *Curr Opin Biotechnol* **33C**: 87-94.

Proctor, C.R., Edwards, M.A., and Pruden, A. (2015) Microbial composition of purified waters and implications for regrowth control in municipal water systems. *Environmental Science-Water Research & Technology* **1**: 882-892.

Prosser, J.I. (2010) Replicate or lie. *Environ Microbiol* **12**: 1806-1810.

Regan, J.M., Harrington, G.W., Baribeau, H., De Leon, R., and Noguera, D.R. (2003) Diversity of nitrifying bacteria in full-scale chloraminated distribution systems. *Water Res* **37**: 197-205.

Revetta, R.P., Gomez-Alvarez, V., Gerke, T.L., Domingo, J.W., and Ashbolt, N.J. (2016) Changes in bacterial composition of biofilm in a metropolitan drinking water distribution system. *J Appl Microbiol* **121**: 294-305.

Ridgway, H.F., and Olson, B.H. (1981) Scanning electron microscope evidence for bacterial colonization of a drinking water distribution system. *Appl Environ Microbiol* **41**: 274-287.

Rittmann, B.E., Hausner, M., Loffler, F., Love, N.G., Muyzer, G., Okabe, S. et al. (2006) A vista for microbial ecology and environmental biotechnology. *Environ Sci Technol* **40**: 1096-1103.

Roberts, R.J., Carneiro, M.O., and Schatz, M.C. (2013) The advantages of SMRT sequencing. *Genome Biol* **14**: 405.

Roeselers, G., Coolen, J., van der Wielen, P.W.J.J., Jaspers, M.C., Atsma, A., de Graaf, B., and Schuren, F. (2015) Microbial biogeography of drinking water: patterns in phylogenetic diversity across space and time. *Environ Microbiol* **17**: 2505-2514.

Ronn, R., McCaig, A.E., Griffiths, B.S., and Prosser, J.I. (2002) Impact of protozoan grazing on bacterial community structure in soil microcosms. *Appl Environ Microbiol* **68**: 6094-6105.

Roskoski, R. (2007) Enzyme Assays. In *xPharm: The Comprehensive Pharmacology Reference*. New York: Elsevier, pp. 1-7.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R. et al. (2013) Characterizing and measuring bias in sequence data. *Genome Biol* **14**.

Santic, D., Krstulovic, N., and Solic, M. (2007) Comparison of flow cytometric and epifluorescent counting methods for marine heterotrophic bacteria. *Acta Adriatica* **48**: 107-114.

Schloss, P.D. (2010) The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* **6**.

Schmeisser, C., Stockigt, C., Raasch, C., Wingender, J., Timmis, K.N., Wenderoth, D.F. et al. (2003) Metagenome survey of biofilms in drinking-water networks. *Applied and Environmental Microbiology* **69**: 7298-7309.

Segata, N., Bornigen, D., Morgan, X.C., and Huttenhower, C. (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* **4**: 2304.

Shade, A., and Handelsman, J. (2012) Beyond the Venn diagram: the hunt for a core microbiome. *Environ Microbiol* **14**: 4-12.

Sharpton, T.J. (2014) An introduction to the analysis of shotgun metagenomic data. *Front Plant Sci* **5**.

Small, J., Call, D.R., Brockman, F.J., Straub, T.M., and Chandler, D.P. (2001) Direct detection of 16S rRNA in soil extracts by using oligonucleotide microarrays. *Appl Environ Microbiol* **67**: 4708-4716.

Snoeyink, V., Hass, C., Boulos, P., Burlinghame, G., Camper, A., Clark, R. et al. (2006) *Drinking water distribution systems: Assessing and reducing risks*: National Academics Press: Washington, DC.

Sorek, R., and Cossart, P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet* **11**: 9-16.

Sowell, S.M., Abraham, P.E., Shah, M., Verberkmoes, N.C., Smith, D.P., Barofsky, D.F., and Giovannoni, S.J. (2011) Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *ISME J* **5**: 856-865.

Staley, J.T., and Konopka, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* **39**: 321-346.

Stark, L., Giersch, T., and Wunschiers, R. (2014) Efficiency of RNA extraction from selected bacteria in the context of biogas production and metatranscriptomics. *Anaerobe* **29**: 85-90.

Steele, H.L., and Streit, W.R. (2005) Metagenomics: Advances in ecology and biotechnology. *FEMS Microbiol Lett* **247**: 105-111.

Stewart, P.S., and Franklin, M.J. (2008) Physiological heterogeneity in biofilms. *Nat Rev Microbiol* **6**: 199-210.

Stutz, W., Leki, G., and Lopez Pila, J.M. (1986) The ATP concentration of drinking water compared to the colony count. *Zentralbl Bakteriol Mikrobiol Hyg B* **182**: 421-429.

Sun, H., Shi, B., Bai, Y., and Wang, D. (2014) Bacterial community of biofilms developed under different water supply conditions in a distribution system. *Sci Total Environ* **472**: 99-107.

Thomas, T., Gilbert, J., and Meyer, F. (2012) Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp* **2**: 3.

Timmins-Schiffman, E., May, D.H., Mikan, M., Riffle, M., Frazar, C., Harvey, H.R. et al. (2017) Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J* **11**: 309-314.

Tsementzi, D., Poretsky, R., Rodriguez, L.M., Luo, C.W., and Konstantinidis, K.T. (2014) Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environ Microbiol Rep* **6**: 640-655.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

Urich, T., Lanzen, A., Qi, J., Huson, D.H., Schleper, C., and Schuster, S.C. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One* **3**.

van der Wielen, P.W.J.J., Voost, S., and van der Kooij, D. (2009) Ammonia-oxidizing bacteria and Archaea in groundwater treatment and drinking water distribution systems. *Appl Environ Microbiol* **75**: 4687-4695.

Van Nevel, S., Koetzsch, S., Proctor, C.R., Besmer, M.D., Prest, E.I., Vrouwenvelder, J.S. et al. (2017) Flow cytometric bacterial cell counts challenge conventional heterotrophic plate counts for routine microbiological drinking water monitoring. *Water Res* **113**: 191-206.

Vanderkooij, D., Visser, A., and Hijnen, W.A.M. (1982) Determining the concentration of easily assimilable organic carbon in drinking water. *J Am Water Works Assoc* **74**: 540-545.

Vaz-Moreira, I., Egas, C., Nunes, O.C., and Manaia, C.M. (2013) Bacterial diversity from the source to the tap: a comparative study based on 16S rRNA gene-DGGE and culture-dependent methods. *FEMS Microbiol Ecol* **83**: 361-374.

Velten, S., Hammes, F., Boller, M., and Egli, T. (2007) Rapid and direct estimation of active biomass on granular activated carbon through adenosine tri-phosphate (ATP) determination. *Water Res* **41**: 1973-1983.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Vetrovsky, T., and Baldrian, P. (2013) The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS One* **8**.

von Gunten, U. (2003) Ozonation of drinking water: Part I. Oxidation kinetics and product formation. *Water Res* **37**: 1443-1467.

Wang, H., Edwards, M.A., Falkinham, J.O., and Pruden, A. (2013) Probiotic approach to pathogen control in premise plumbing systems? A review. *Environ Sci Technol* **47**: 10117-10128.

Wang, H., Masters, S., Edwards, M.A., Falkinham, J.O., and Pruden, A. (2014a) Effect of disinfectant, water age, and pipe materials on bacterial and eukaryotic community structure in drinking water biofilm. *Environ Sci Technol* **48**: 1426-1435.

Wang, H., Bedard, E., Prevost, M., Camper, A.K., Hill, V.R., and Pruden, A. (2017) Methodological approaches for monitoring opportunistic pathogens in premise plumbing: A review. *Water Res* **117**: 68-86.

Wang, H., Proctor, C.R., Edwards, M.A., Pryor, M., Domingo, J.W.S., Ryu, H. et al. (2014b) Microbial community response to chlorine conversion in a chloraminated drinking water distribution system. *Environ Sci Technol* **48**: 10624-10633.

Wang, H.B., Zhang, Z.X., Li, H., He, H.B., Fang, C.X., Zhang, A.J. et al. (2011) Characterization of metaproteomics in crop rhizospheric soil. *J Proteome Res* **10**: 932-940.

White, C.P., DeBry, R.W., and Lytle, D.A. (2012) Microbial survey of a full-scale, biologically active filter for treatment of drinking water. *Appl Environ Microbiol* **78**: 6390-6394.

Williams, M.A., Taylor, E.B., and Mula, H.P. (2010) Metaproteomic characterization of a soil microbial community following carbon amendment. *Soil Biol Biochem* **42**: 1148-1156.

Wilmes, P., and Bond, P.L. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ Microbiol* **6**: 911-920.

Wilmes, P., Wexler, M., and Bond, P.L. (2008) Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS One* **3**.

Wilson, K.H., Wilson, W.J., Radosevich, J.L., DeSantis, T.Z., Viswanathan, V.S., Kuczmarski, T.A., and Andersen, G.L. (2002) High-density microarray of small-subunit ribosomal DNA probes. *Appl Environ Microbiol* **68**: 2535-2541.

Wimpenny, J., Manz, W., and Szewzyk, U. (2000) Heterogeneity in biofilms. *FEMS Microbiol Rev* **24**: 661-671.

Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C. et al. (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**: 1661-1665.

Wu, J.H., Hong, P.Y., and Liu, W.T. (2009) Quantitative effects of position and type of single mismatch on single base primer extension. *Journal of Microbiological Methods* **77**: 267-275.

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

Wu, Y.-W., Tang, Y.-H., Tringe, S.G., Simmons, B.A., and Singer, S.W. (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**: 1-18.

Xiong, X.J., Frank, D.N., Robertson, C.E., Hung, S.S., Markle, J., Canty, A.J. et al. (2012) Generation and analysis of a mouse intestinal metatranscriptome through Illumina based RNA-sequencing. *PLoS One* **7**.

Yanez, M.A., Nocker, A., Soria-Soria, E., Murtula, R., Martinez, L., and Catalan, V. (2011) Quantification of viable *Legionella pneumophila* cells using propidium monoazide combined with quantitative PCR. *J Microbiol Methods* **85**: 124-130.

Yu, K., and Zhang, T. (2012) Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS One* **7**.

Zampieri, E., Chiapello, M., Daghino, S., Bonfante, P., and Mello, A. (2016) Soil metaproteomics reveals an inter-kingdom stress response to the presence of black truffles. *Sci Rep* **6**: 25773.

Zearley, T.L., and Summers, R.S. (2012) Removal of trace organic micropollutants by drinking water biological filters. *Environ Sci Technol* **46**: 9412-9419.

Zeng, D.N., Fan, Z.Y., Chi, L., Wang, X., Qu, W.D., and Quan, Z.X. (2013) Analysis of the bacterial communities associated with different drinking water treatment processes. *World J Microb Biot* **29**: 1573-1584.

Zhang, Y., Griffin, A., and Edwards, M. (2008) Nitrification in premise plumbing: Role of phosphate, pH and pipe corrosion. *Environ Sci Technol* **42**: 4280-4284.

Zhang, Y., Oh, S., and Liu, W.T. (2017) Impact of drinking water treatment and distribution on the microbiome continuum: an ecological disturbance's perspective. *Environ Microbiol* **19**: 3163-3174.

Zhou, J. (2003) Microarrays for bacterial detection and microbial community analysis. *Curr Opin Microbiol* **6**: 288-294.

CHAPTER 3 IMPACT OF DRINKING WATER TREATMENT AND DISTRIBUTION ON THE MICROBIOME CONTINUUM: AN ECOLOGICAL DISTURBANCE'S PERSPECTIVE

3.1 Abstract

While microbes are known to be present at different stages of a drinking water system, their potential functions and ability to grow in such systems are poorly understood. In this study, we demonstrated that treatment and distribution processes could be viewed as ecological disturbances exhibited over space on the microbiome continuum in a groundwater-derived system. Results from 16S rRNA gene amplicon analysis and metagenomics suggested that disturbances in the system were intense as the community diversity was substantially reduced during the treatment steps. Specifically, syntrophs and methanogens dominant in raw water (RW) disappeared after water abstraction, accompanied by a substantial decrease in both the abundance and number of functional genes related to methanogenesis. The softening effluent was dominated by an *Exiguobacterium*-related population, likely due to its ability to use the phosphotransferase system (PTS) as regulatory machinery to control the energy conditions of the cell. After disinfection and entering the distribution system, community-level functionality remained relatively stable, whereas the community structure differed from those taken in the treatment steps. The diversity and high abundance of some eukaryotic groups in the system suggested that predation could be a disturbance to the bacterial microbiome, which could drive the diversification of the bacterial community.

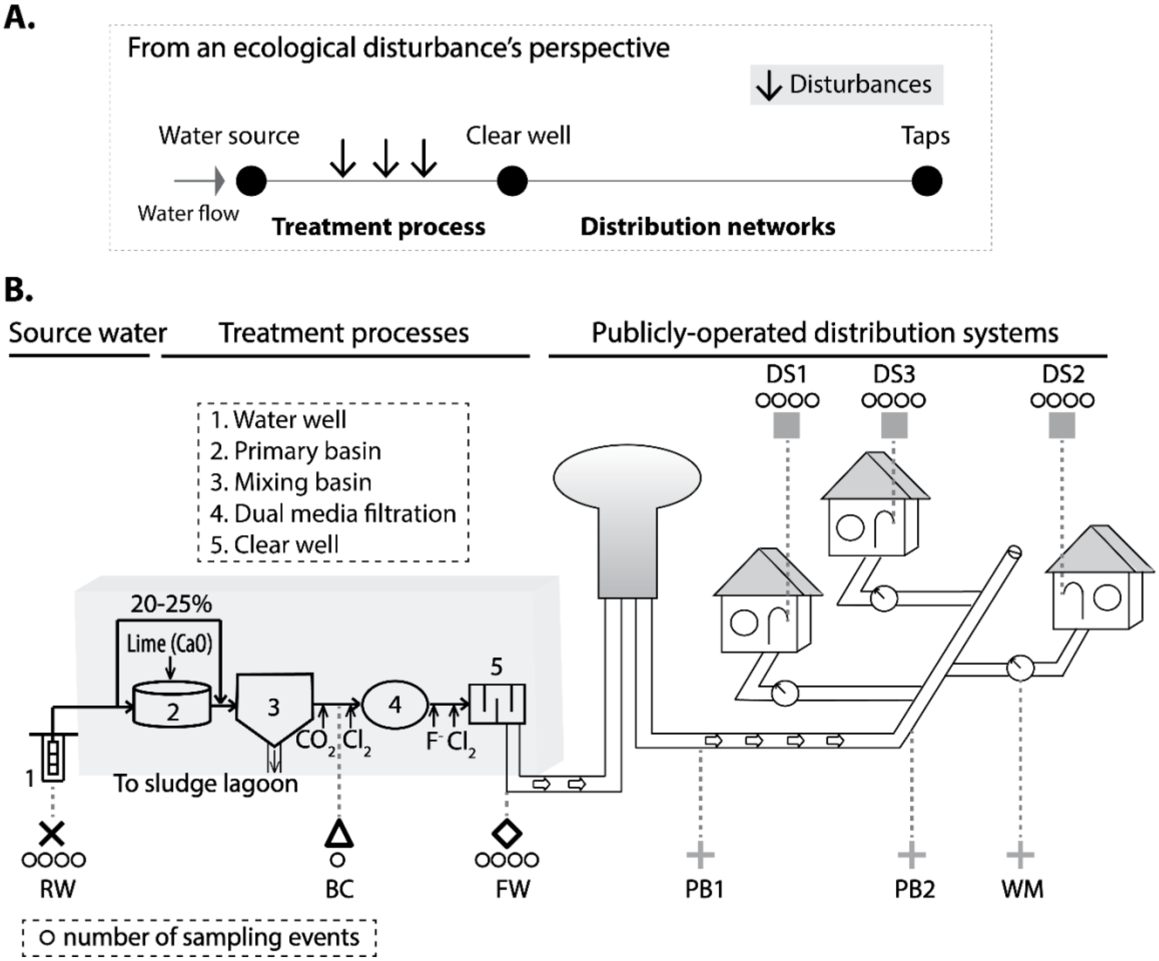


Figure 3.1 (A) An illustration of viewing treatment and distribution processes as ecological disturbances exhibited over space on the microbiome continuum. (B) Sampling sites and events for the studied drinking water system. Each open dot represents one sampling event.

3.2 Introduction

Ecological disturbances, such as glaciation, wildfires, and windstorms, have long been recognized as the driving force to influence species co-existence, biodiversity maintenance, and ecosystem functions (Connell, 1978; Huston, 1979). Classical disturbances often occur within a defined physical space at discrete times, i.e., changes of a continuum over time. Alternatively, disturbances can occur as a continuum that moves through a well-defined physical space. One such example is drinking water treatment processes, where a series of chemicals are added to the water continuum sequentially and quantitatively at different compartments of a defined physical space. The disturbances occurring in each compartment can be intensive and well-controlled, and can be viewed as short-term disturbances as each usually lasts a few hours (Figure 3.1A).

In a drinking water treatment and distribution system, the microbiome continuum can be divided into prokaryotic and eukaryotic fractions. The prokaryotic or bacterial fraction (Archaea are usually not abundant in drinking water systems) usually serves as the backbone of the ecosystem and food webs, and can be subjected to predation by eukaryotes (Gasol et al., 2002; Ronn et al., 2002; Jousset, 2012). As a cause of bacterial mortality (Pernthaler, 2005) and influencing the genetic and functional structure of bacterial communities (Griffiths et al., 1999; Ronn et al., 2002; De Mesel et al., 2004; Bell et al., 2010; Jousset, 2012), predation can be an important disturbance to the bacterial microbiome. Viewing the drinking water treatment systems from an ecological disturbance's perspective can provide practical guidance on the management of the microbial community in this oligotrophic environment. Important questions that can be asked include whether the microbial community was overly stressed with disinfectants in comparison with systems from some European countries, where natural filtration systems are used without the addition of a residual concentrations of disinfectants (Rosario-Ortiz et al., 2016); or whether nutrient control within the European systems is insufficient, allowing pathogens like *Legionella pneumophila* to have the opportunity to flourish within amoebae that are probably the only place in the system which meets their nutritional requirements (Breiman et al., 1990; Dupuy et al., 2016).

So far, our knowledge on the ecology of microbial communities inside most of the drinking water systems around the world is very limited. Previous studies have elucidated the most basic ecological questions “who are there?” and “how do they change over time and across space?” in selected stages (i.e., mostly filters and distribution systems) of the drinking water treatment systems in a handful of cities (Hwang et al., 2012; Pinto et al., 2012; Zeng et al., 2013; Pinto et al., 2014; Gulay et al., 2016; Ling et al., 2016). These studies, however, could not provide a comprehensive view on how a microbiome continuum is impacted by disturbances at different stages. Furthermore, it is hard to translate the knowledge into practical instructions for the engineered system for the following reasons: classification is accurate only to the genus or family level using 16S rRNA gene-based next-generation sequencing technologies (Schloss, 2010), and 16S rRNA gene-based phylogeny and functional potential do not always agree with each other (Janda and Abbott, 2007). As a result, direct linkage between microbes and their functionalities often cannot be established. This suggests a need to advance the ecological knowledge of drinking water systems to the next stage — “what are they doing?”, “why are they there?”, and more critically “who is doing what?”.

Metagenomics (for total genomic DNA) can be an attractive approach but have not been widely applied to effectively characterize microbiomes in drinking water-related treatment systems. This approach has been used by a few studies using a two-point comparison (Gomez-Alvarez et al., 2012; Chao et al., 2013). For example, an increase in protective functions (i.e., glutathione synthesis) was observed with treated water in comparison with raw water (Chao et al., 2013). Because these studies derived annotation information based on unassembled short DNA sequencing reads, no direct linkage between microbes and functionalities could be accomplished (Chistoserdova, 2014).

In this study, we investigated the impacts of treatment processes and disinfectant residual on the drinking water microbiome continuum inside a drinking water distribution system from the perspective of short-term disturbances occurring through the system. A groundwater-derived drinking water system was used as a model system, which consists of abstraction, softening, recarbonation, disinfection, filtration, and final distribution with a disinfectant residual (free chlorine). Source water containing little or no dissolved oxygen (Kirk et al., 2004) is elevated to the surface, exposed to oxygen, and treated with CaO, CO_2, and NaClO of defined quantity as well as filtration. Thus, the drinking water microbiome continuum is continuously altered by different types of disturbances as they flow through different compartments in the system. In the model system, biomass from the bulk water and/or biofilms at three different stages of the treatment process was collected. Genomic DNA of individual samples was extracted and analyzed using metagenomics sequencing. In specific, we aimed to address the following questions: i) what would be the impact of the treatment process and the distribution system on the community structure and metabolic potential, respectively?; ii) could changes in the microbial community be reflected from the community functional profile?; iii) did specific microbes respond to a certain disturbance?; and iv) which was the step that shaped the distribution system microbiome and how?.

3.3 Materials and methods

Sampling and sample processing Microbial biomass from all the water-phase samples was collected using point-of-use water purifiers (Toray Industries Inc., Japan). Raw water (RW), immediately before filtration and chlorination (after lime treatment and recarbonation) (BC), and finished water (FW) prior to distribution were taken from the studied treatment plant. Tap water was taken inside three different buildings located approximately one mile apart from each other to represent different locations within the distribution system: DS1 was from a university building; DS2 from an apartment; and DS3 from a house. A ten minutes flushing (the cold-water side) was carried out each time before installing water purifiers to minimize the influence of premise plumbing systems on the distribution system sampling. At each site, approximately 2,000 L of water was filtered each time for 48 hours. This large volume of water was to represent water travelled from source to tap as the operation of the system was stable (Table A.1). Water purifiers were collected at the end of each sampling period and transported to the laboratory at the Department of Civil and Environmental Engineering (University of Illinois at Urbana-Champaign) in coolers. They were disassembled right after arriving at the laboratory and cells were washed off from the multilayer hollow fiber membrane with phosphate-buffered saline (PBS) through sonication (Symphony™ Ultrasonic Cleaners, VWR) according to a previous study (Chao et al., 2013). The obtained mixture was filtered through 0.22 μm membranes and the membranes with cells were stored at -80 °C. Water-phase sampling was repeated four times, in June, July, August, and September 2014, except the BC sample.

Biofilm samples were collected from the inner surface of two retired water mains (PB1 and PB2). PB1 was a 2.25-inch cast iron water main installed in 1968. PB2 was a 1.5-inch cast iron water main installed prior to 1927 and was used to connect to a 4-inch water main. Each pipe was cut into two 12-inch long pieces at the site after soil removal and thorough water cleaning; two stoppers were inserted at each end; and the pipes were shipped in coolers. On arrival at the laboratory, the pipes were flushed with local drinking water to remove any remaining deposits. Biofilm from the entire inner surface was swabbed off the surfaces (avoid the two edges), re-suspended in PBS, and collected by filtering through 0.22 μm membranes. In addition, 14 water meters were obtained through the local drinking water plant, and the biofilm samples were taken and combined according to the protocol established in a previous water meter study (Hong et al., 2010).

DNA extraction and Illumina sequencing Genomic DNA (gDNA) was extracted using FastDNA® SPIN Kit for Soil (MP Biomedicals, Carlsbad, CA, USA) from the membranes with cells. 16S rRNA gene amplicon analysis was carried out using a universal primer set targeting the V4-V5 hypervariable regions of both the Bacteria and Archaea domains (515F: 5’-GTGCCAGCMGCCGCGGTAA-3’ and 909R: 5’-CCCGTCAATTCMTTTRAGT-3’) as described previously (Kozich et al., 2013). The primer set was modified for Illumina Miseq platform with dual indexing strategy. Each PCR mixture (50 μL in volume) contained approximately 1 ng of template DNA in 1× PrimeSTAR® buffer, 0.2 mM dNTP (each), 0.2 μM of forward and reverse primer, and

0.03 U/µL PrimeSTAR® HS DNA Polymerase (Takara Bio Inc. Otsu, Shiga, Japan). Paired-end sequencing of the amplicons (2x300 bp) was done with an Illumina MiSeq platform (Illumina, Inc., San Diego, CA, USA) at the Roy J. Carver Biotechnology Center (University of Illinois at Urbana-Champaign). DNA libraries for metagenomic sequencing were prepared by combining all the extracted gDNA from each sampling site as a relatively large amount of gDNA (> 0.1 µg) was required. Before mixing, 16S rRNA gene amplicon sequencing was performed with individual samples, and the results indicated that the microbial community from the four sampling events of water-phase samples were similar. The prepared library was paired-end sequenced on Illumina HiSeq2500 platforms using TruSeq SBS sequencing kits version 4 (for the RW, BC, FW, and DS1, 2, 3 samples) and TruSeq SBS Rapid sequencing kit version 2 (for the PB1, PB2, and WM samples) (Illumina, Inc., San Diego, CA, USA) at the Roy J. Carver Biotechnology Center. The average read length was 100 nt for the RW, BC, FW, and DS1-3 samples and 250 nt for the PB1, PB2, and WM samples. There was no significant difference in the total contig length obtained among these samples.

Sequence analysis The obtained paired-end 16S rRNA gene sequences were aligned with Mothur using the default setting, which required a quality score of over 25 if a gap and a base occurred at the same position or one of the bases had a quality score six or more points better than the other if the two reads disagreed (Kozich et al., 2013). The resulting sequences were screened for chimeras with the UCHIME algorithm implemented in USEARCH 6.1 and processed using the *de novo* OTU picking workflow in QIIME (Caporaso et al., 2010b). Representative sequences from OTUs were aligned using PyNAST (Caporaso et al., 2010a) and inserted into the phylogenetic trees, Greengenes_16S_2011.arb, with the parsimony insertion tool in the ARB program (Ludwig et al., 2004; McDonald et al., 2012). Alpha-and beta-diversity indices were calculated based on the rarefied OTU table at a depth of 12500 sequences per sample (Shannon indices, sample evenness, and unifrac distances). All the metagenomic datasets were trimmed using SolexaQA2 based on a cutoff of 20 by phred scores (Cox et al.,

2010) and assembled using Megahit (Li et al., 2015). The assembled contigs with coverage information were submitted to the MG-RAST pipeline (version 3.6) (Meyer et al., 2008). Percentages of reads mapped to contigs were estimated by the Burrow-Wheeler Aligner-MEM (BWA-MEM) (Li and Durbin, 2010). EMIRGE was used to reconstruct nearly full-length SSU genes in metagenomes (Miller, 2013) . Average genome size was estimated using MicrobeCensus (Nayfach and Pollard, 2015).

For functional analysis, the contigs with coverage information were submitted to the MG-RAST. Annotations based on translated nucleotide sequences were mapped to the KEGG database, and were extracted from the MG-RAST server using an e-value cutoff of 10^{-5}, minimum identity cutoff of 60%, and minimum alignment length cutoff of 50 amino acids. Hits within each pathway were converted to relative abundance and transformed using row *z*-score across all samples to remove differences in reaction efficiencies. Then pathways with relative abundance (counts within each pathway of a metagenome/total number of counts of that metagenome) over 0.5% and maximum abundance/minimum abundance across all metagenomes greater than 1.5 were selected.

Genome recovery Assembled contigs from the BC sample were binned with MaxBin 2.0 (Wu et al., 2016). The obtained bins were compared and assessed with CheckM (Parks et al., 2015) and manually curated. Percentages of reads mapped to contigs were estimated by BWA-MEM.

Metagenomic data depositing Paired-end 16S rRNA gene sequences were submitted to NCBI Sequence Read Archive under the project accession number PRJNA323575. Assembled contigs with coverage information from each metagenome were deposited in MG-RAST with IDs 4634469.3-4634473.3, and 4683347.3-4683349.3. The draft genome belonging to *Exiguobacterium* was deposited in RAST with ID 33986.112.

3.4 Results

Description of the drinking water system studied The origin of source water was groundwater abstracted from the Mahomet sands aquifer with a hardness of 300 mg/L as $CaCO_3$ and a pH of 7.6. The aquifer water was reported to contain methane estimated at 16.4 g CH_4/m^3 (Snoeyink et al., 2006). Lime was added in the softening step to remove hardness, and during this process pH rose to 11 (Figure 3.1B). Approximately 75-80% of the raw water was treated this way and was blended with the remaining raw water to achieve a targeted hardness of 80 mg/L as $CaCO_3$. The next step was recarbonation, where CO_2 was added to stop the precipitation reaction by reducing the pH to 8.8. Then, free chlorine was added immediately prior to filtration at a concentration of approximately 4.0 $mgCl_2/L$, and after filtration to maintain a chlorine residual of approximately 2.5 $mgCl_2/L$ before entering the distribution system. The filters (two units) were backwashed every 48 hrs with finished water. Fluoride was added to the filtered effluent at 1.0 mg/L to prevent tooth decay. The operation of the treatment plant was relatively stable during the sampling period.

Microbial biomass from different stages of the treatment processes and different locations in the distribution system was collected (Figure 3.1B). Water-phase samples taken from RW, BC, FW, and three taps (cold water) (DS1-DS3) were concentrated. The tap water samples were used, primarily, to represent different locations of the distribution system. To avoid temporal variations, water-phase sampling was repeated four times except for the BC sample, which was only sampled successfully on the fourth trial, due to membrane blockage by calcium carbonate. Biofilm samples were taken from two retired water mains (PB1-PB2) and 14 water meters (WM). Microbial communities in these samples were analyzed using amplified 16S rRNA genes and metagenomics.

Summary of 16S rRNA gene based sequencing and metagenomes More than 12,500 16S rRNA gene sequences per sample were obtained after processing the raw amplicon sequences.

Table 3.1 Summary of metagenomics samples included in this study.

		RW	BC	FW	DS1	DS2	DS3	PB1	PB2	WM
Reads	Number of reads	1.2E+08	1.3E+08	1.4E+08	1.1E+08	1.1E+08	1.1E+08	5.2E+07	6.1E+07	6.3E+07
	Size (bp)	1.7E+10	1.9E+10	2.0E+10	1.5E+10	1.6E+10	1.6E+10	9.3E+09	1.1E+10	1.0E+10
	G + C content (%)	46	57	56	63	61	55	62	63	57
Contigs	No. of contigs	5.4E+05	2.0E+05	1.1E+05	1.1E+05	7.2E+04	2.9E+05	2.7E+05	3.7E+05	2.5E+05
	Total size (bp)	6.1E+08	3.6E+08	2.0E+08	1.9E+08	1.4E+08	3.3E+08	2.7E+08	3.6E+08	2.6E+08
	Mapped reads (%)	80.2	91.6	93.0	92.8	92.2	88.9	92.5	93.4	95.7
	Maximum length (bp)	4.2E+05	1.0E+06	1.3E+06	8.5E+05	8.4E+05	4.1E+05	8.0E+05	5.6E+05	4.4E+05
	N50	1530	4170	5294	3690	6780	1590	2123	2816	2816
	Post QC size (bp)	4.9E+08	2.6E+08	1.3E+08	1.4E+08	0.9E+08	2.7E+08	2.7E+08	3.6E+08	2.6E+08
	Annotated protein[1]	68.5%	98.9%	99.0%	98.9%	98.9%	53.4%	97.2%	100.0%	100.0%
	AGS (Mbp)[2]	2.3	3.0	3.6	4.5	4.1	5.1	5.7	4.9	4.0
	AGS STD	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1
	Taxonomic hits to *Bacteria*	91.1%	98.5%	98.6%	98.1%	96.8%	81.5%	88.1%	98.6%	96.6%
	Taxonomic hits to *Archaea*	7.0%	0.23%	0.16%	0.61%	0.4%	0.26%	0.26%	0.12%	0.23%
	Taxonomic hits to *Eukarya*	1.3%	1.0%	0.91%	1.1%	2.6%	18.1%	11.5%	1.2%	3.0%
	Taxonomic hits to viruses	0.07%	0.16%	0.16%	0.07%	0.05%	0.05%	0.07%	0.03%	0.07%

[1] Number of contigs with annotated proteins.

[2] AGS: estimated average genome size.

For metagenomes, approximately 1.0 ×10^{10} bp of reads were obtained from each sampling site, with 16S rRNA gene sequences representing approximately 0.15% of the total DNA reads (Table 3.1). Our results indicated that the amplicon-based 16S rRNA gene sequences could be used to identify an operational taxonomic unit (OTU at a cutoff of 97% sequence similarity) with an abundance down to 10^{-5} or 10^{-6} based on the rank-abundance curve. However, 16S rRNA gene sequences retrieved from metagenomes failed to detect OTUs with an abundance less than 10^{-3}.

For metagenomics sequences, sequence assembly was carried out to reduce the metagenome size and to enhance the annotation accuracy. The size of each metagenome was reduced to approximately 2.0 ×10^{8} bp. More than 85.0% of the reads could be mapped to the obtained contigs except RW (80.2%) (Table 3.1). When submitted to the MG-RAST server for annotation, more than 97.2% of the sequences were successfully annotated with at least one known protein feature except the RW (68.5%) and DS3 (53.4%) samples. Most of the unannotated sequences in the RW and DS3 samples were of archaeal or eukaryotic origin, as predicted from known annotations assigned to the Archaea (7.0% in the RW sample) and Eukarya (18.1% in the DS3 sample) domains (Table 3.1). Overall, the assembled contigs were of high quality and were representative of the original metagenome dataset, and were subsequently used to predict community composition and functionalities.

16S rRNA gene profiling revealed major microbial community shifts during abstraction and softening processes The Shannon H index was used to determine the community diversity based on 16S rRNA gene amplicon sequences. The diversity was the highest for the RW sample at 3.87, and decreased to 2.43-3.00 during the water treatment process and in the distribution system, suggesting that disturbances on the bacterial community imposed by drinking water treatment processes were intensive. The intensive disturbances also led to relatively low evenness in all the samples, ranging from 0.42 to 0.60.

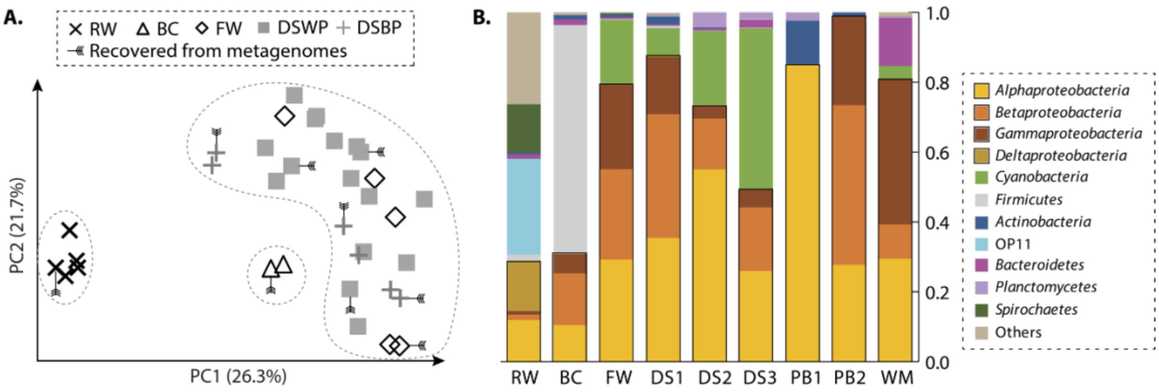


Figure 3.2 Community structure. (A) Beta-diversity of the microbial community inside the system based on 16S rRNA genes. Beta-diversity was represented with UniFrac distances using principal coordinates analysis (PCoA). Samples taken at different sampling events were included. Communities recovered from metagenomes were marked with arrows. DSWP: distribution system water phase, including DS1-3. DSBP: distribution system biofilm phase, including PB1-2, and WM. (B) Community composition at the phylum level by 16S rRNA gene amplicon analysis. Due to the dominance of Proteobacteria, it was broken down into subdivisions. Average of four sampling events at each site was used for water-phase samples.

Principal coordinates analysis (PCoA) using weighted UniFrac distance revealed major community shifts from RW to BC and then FW based on both 16S rRNA gene amplicon sequences and 16S rRNA gene sequences recovered from metagenomes (Figure 3.2A). For the remaining samples from FW and the distribution system, no specific clustering pattern could be observed. Using the 16S rRNA gene amplicon sequences, distinct patterns in microbial composition between the RW and BC samples were observed at the phylum level with Microgenomates (27.4%) and Firmicutes (65.3%) dominating respectively (Figure 3.2B). In contrast, Proteobacteria (Alpha-, Beta-, and Gamma-subdivisions) and Cyanobacteria were dominant (> 80.0% of total sequences) in samples from the FW and the distribution system, and these findings agreed with the summary of previous studies on drinking water (Proctor and Hammes, 2015). In the RW and BC samples, Cyanobacteria could only be detected at low abundances (approximately 0.1%).

Clearly, a major microbial community shift occurred from RW to BC and FW, but only minor changes happened from FW to samples taken in the distribution system.

From a continuum perspective, Proteobacteria were present at relatively high abundances in FW and in samples taken throughout the distribution system, and Spirochaetes, Microgenomates, Firmicutes, and Cyanobacteria became substantially dominant only in certain sections of the system. These observations suggested strong disturbances on the microbial community structure at each treatment step. The community structure differences under the influence of RW characteristics or the disturbance of treatment (e.g., high pH and disinfection) could be further elucidated by examining the dominant OTUs.

Dominant OTUs observed in RW and during treatment disturbances In RW, we detected OTUs that were mostly identified in anoxic environments, including OTU-3540 (27.3% based on 16S rRNA gene amplicon sequences, affiliated with OP11-4), OTU-3576 (7.3% based on 16S rRNA gene amplicon sequences, affiliated with *Syntrophus*), and OTU-1309 (5.3% based on 16S rRNA gene amplicon sequences, affiliated with OP3) (Figure 3.3). Both OTU-3576 and OTU-1309 were closely related to known syntrophs. In addition, a *Methanospirillum*-related OTU at a low abundance (0.1%) was detected. These microbial populations were likely involved in methane production in the Mahomet aquifer where the RW was taken (Jackson et al., 1999; Flynn et al., 2013).

As the dissolved methane present in RW could travel with the water flow, it could support the growth of microbes downstream. This was supported by a significant increase in the relative abundance of microbial populations belonging to methano-/methylotrophs in FW and the distribution system, including OTU-6709 (affiliated with *Methylomonas),* OTU-2978 (*Methylotenera*), OTU-2023 (*Methylocystis*), and OTU-4800 (*Hyphomicrobium*) (tested using the sum of the abundance of the four OTUs in each sampling event, one-tailed Mann-Whitney U test, $p < 0.05$).

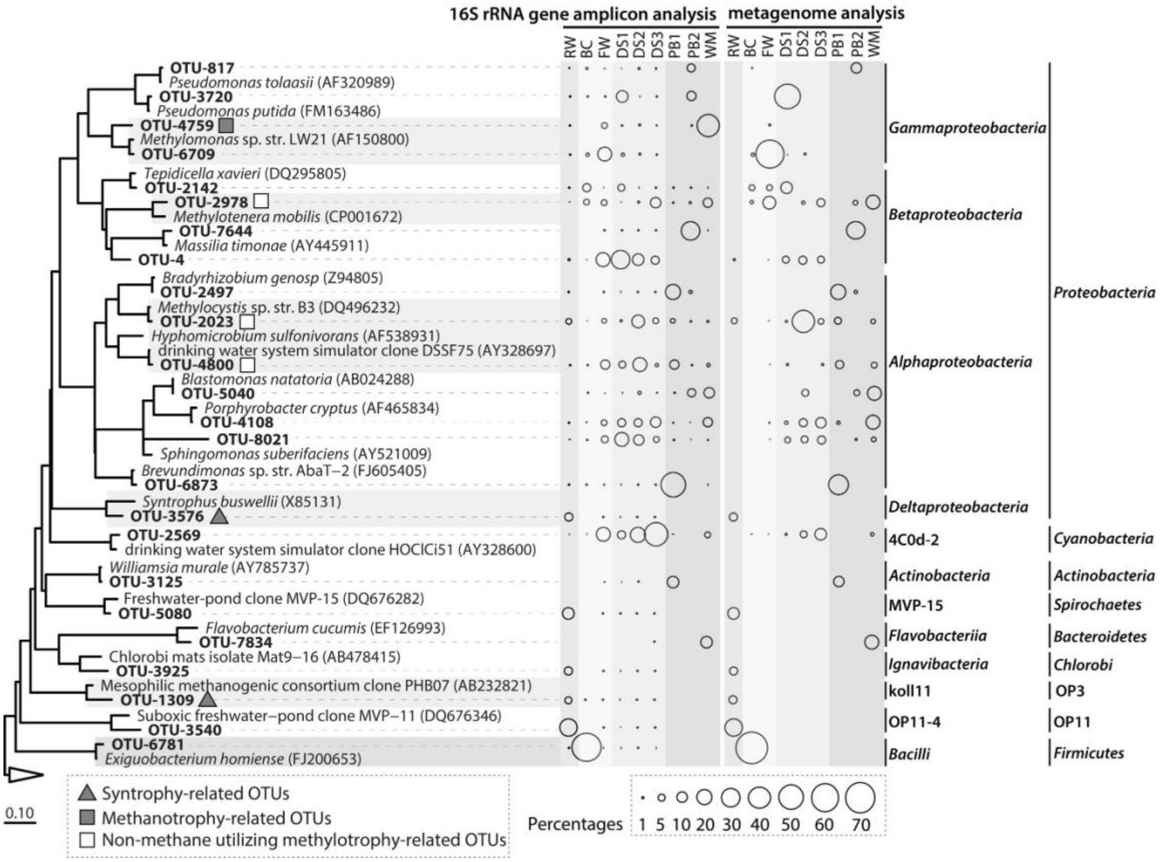


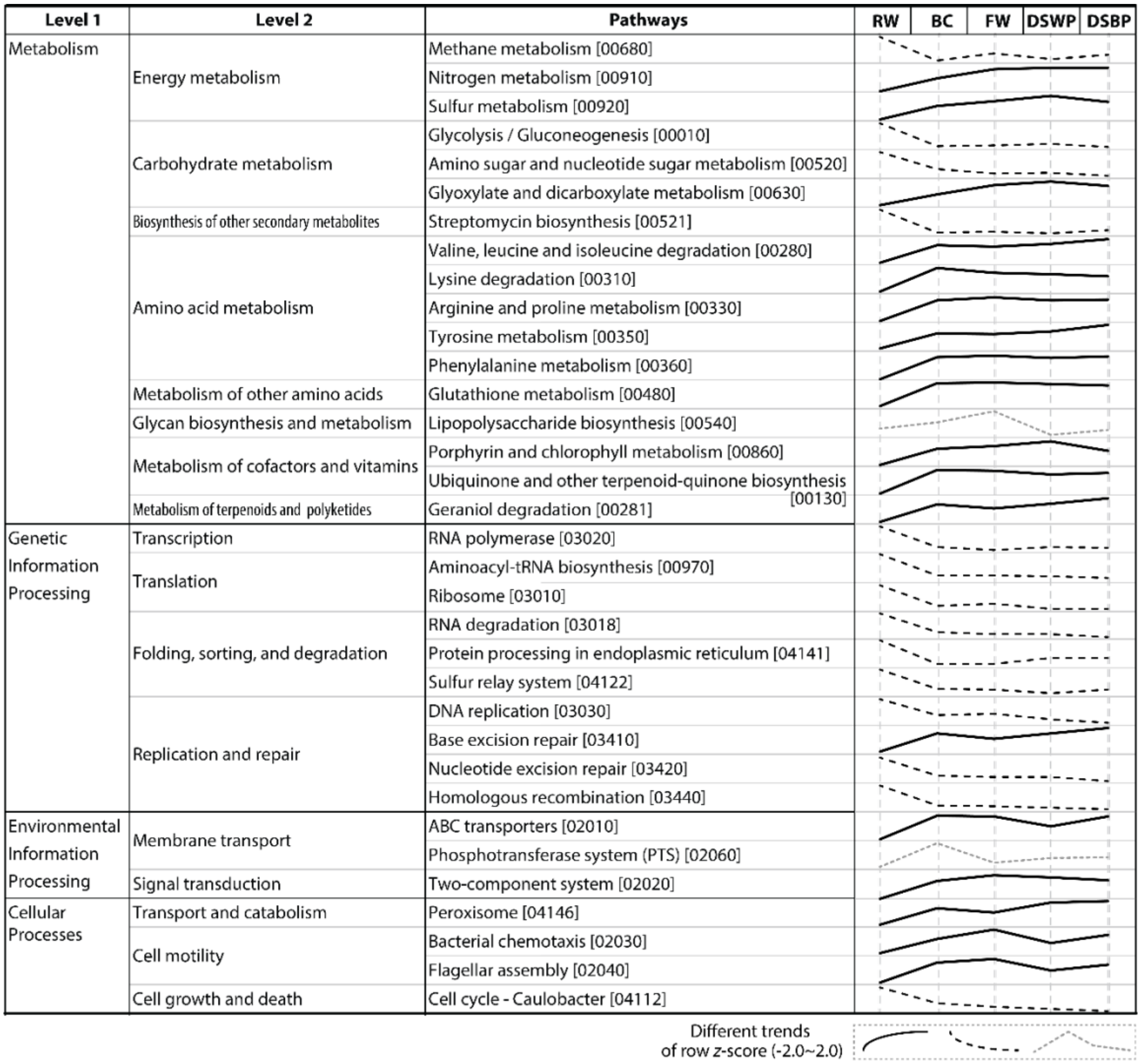
Figure 3.3 Dominant OTUs identified by 16S rRNA gene amplicon analysis and its comparison with retrieved 16S rRNA genes from metagenomes. Dominant OTUs were selected based on the ranking of average abundance and ≥5.0% abundance in at least one sample.

Methylomonas are methanotrophs using methane as the sole or preferred carbon source (Chistoserdova et al., 2009). *Methylotenera*, *Methylocystis*, and *Hyphomicrobium* are methylotrophs often found together with methanotrophs in enrichment cultures, possibly feeding on methane-derived single-carbon compounds generated by methanotrophs (Oshkin et al., 2015). The dominance of these methano-/methylotrophs in the studied distribution system was detailed in our previous report (Ling et al., 2016). It is likely that methanotrophs acted as a primary producer in the distribution system to supply byproducts and metabolites derived from methane oxidation processes to methylotrophs and other heterotrophs. Methano-/methylotrophs were relatively low in abundance in BC sample due to an elevated pH.

We further observed a dominant OTU (OTU-6781) closely related to *Exiguobacterium* (65.0% by 16S rRNA gene amplicon analysis) in the BC sample after the softening process, where pH was rapidly raised to 11 and reduced to 8.8 after recarbonation. As no studies have characterized microbial communities in softening processes, this is the first observation on the dominance of *Exiguobacterium*-related populations. *Exiguobacterium* is known to be present in a wide range of pH (4-11), temperature (from permafrost, glacial ice, to hot springs), and salinity (Vishnivetskaya et al., 2009; Rajaei et al., 2015). However, it remains unclear what mechanisms enabled their adaptation to extreme environments.

In the FW and distribution system water samples (DS1-DS3), a dominant *Cyanobacteria*-related OTU (OTU-2569) with the abundance ranging from 6.6% to 45.6% in individual samples was detected (Figure 3.3). This OTU was closely affiliated with MLE1-12 in the non-photosynthetic phylum Melainabacteria (Soo et al., 2014). Melainabacteria have been found to be prevalent in groundwater and in the human gut, relying on anaerobic fermentation to generate energy (Di Rienzi et al., 2013). As Melainabacteria-related OTU is likely an anaerobe and could not proliferate in drinking water systems, their increase in relative abundance in FW and the distribution system was likely due to their resistance to chlorination and the decrease in total cell numbers by six fold after chlorination.

Microbiome continuum exhibited community-level functionality changes mainly in methane production, phosphotransferase system (PTS), and lipopolysaccharide (LPS) biosynthesis The KEGG database contains more than 442,000 metabolic pathways for > 3,000 organisms, and has often been used to understand the functions of a microbial community characterized using metagenomics (Kanehisa et al., 2014).



Level 1	Level 2	Pathways
Metabolism	Energy metabolism	Methane metabolism [00680]
		Nitrogen metabolism [00910]
		Sulfur metabolism [00920]
	Carbohydrate metabolism	Glycolysis / Gluconeogenesis [00010]
		Amino sugar and nucleotide sugar metabolism [00520]
		Glyoxylate and dicarboxylate metabolism [00630]
	Biosynthesis of other secondary metabolites	Streptomycin biosynthesis [00521]
	Amino acid metabolism	Valine, leucine and isoleucine degradation [00280]
		Lysine degradation [00310]
		Arginine and proline metabolism [00330]
		Tyrosine metabolism [00350]
		Phenylalanine metabolism [00360]
	Metabolism of other amino acids	Glutathione metabolism [00480]
	Glycan biosynthesis and metabolism	Lipopolysaccharide biosynthesis [00540]
	Metabolism of cofactors and vitamins	Porphyrin and chlorophyll metabolism [00860]
		Ubiquinone and other terpenoid-quinone biosynthesis [00130]
	Metabolism of terpenoids and polyketides	Geraniol degradation [00281]
Genetic Information Processing	Transcription	RNA polymerase [03020]
	Translation	Aminoacyl-tRNA biosynthesis [00970]
		Ribosome [03010]
	Folding, sorting, and degradation	RNA degradation [03018]
		Protein processing in endoplasmic reticulum [04141]
		Sulfur relay system [04122]
	Replication and repair	DNA replication [03030]
		Base excision repair [03410]
		Nucleotide excision repair [03420]
		Homologous recombination [03440]
Environmental Information Processing	Membrane transport	ABC transporters [02010]
		Phosphotransferase system (PTS) [02060]
	Signal transduction	Two-component system [02020]
Cellular Processes	Transport and catabolism	Peroxisome [04146]
	Cell motility	Bacterial chemotaxis [02030]
		Flagellar assembly [02040]
	Cell growth and death	Cell cycle - Caulobacter [04112]

Figure 3.4 Selected KEGG pathways whose abundance vary among different samples. Pathways with relative abundance (counts within each pathway of a metagenome/total number of counts of that metagenome) over 0.5% and maximum abundance/minimum abundance across all metagenomes greater than 1.5 were selected. To facilitate comparison between different stages, samples from the same stage were grouped together (DSWP included DS1-3, and DSBP included PB1-2 and WM) and results were organized according to the direction of water flow.

Community-level functionalities of individual samples were characterized with the KEGG database according to the water flow from RW and FW, to bulk water and biofilms in the distribution system. In RW, high representation in the category of "genetic information processing" but low representation in "metabolism", "environmental information processing", and "cellular processes" was observed (Figure 3.4). From the energy perspective, methane metabolism [ko00680] in terms of the relative abundance and the number of genes involved was highly represented in RW (Figure 3.4). Four almost complete methanogenesis modules used for the conversion of CO_2, acetate, and methyl-amine/dimethyl-amine/trimethyl-amine to methane, and CO_2 to acetyl-CoA were identified in the RW sample. These findings suggested that methane production took place in the Mahomet aquifer where RW was drawn.

Moving from the RW sample to the BC sample, we observed substantial changes with all the metabolic pathways associated within individual categories, suggesting strong disturbances have occurred. Specifically, the PTS [ko02060] was highly represented in BC in terms of relative abundance and number of genes (Figure 3.4). PTS is the key signal transduction pathway for the optimal utilization of carbohydrates in complex environments by the phosphorylation status of PTS components (Kotrba et al., 2001). PTS is also responsible for numerous regulatory functions, including nitrogen and phosphate metabolism, chemotaxis, and potassium transport (Deutscher et al., 2014). Meanwhile, protective functions, including glutathione metabolism [ko00480] and peroxisome [ko04146], became over-represented, likely due to the change in oxidative stress in the environment from anoxic to aerobic. Lastly, metabolisms related to non-polar (valine, leucine, isoleucine, and phenylalanine) and basic amino acids (lysine and arginine) became more abundant (Figure 3.4).

Dominating primarily in the BC community, *Exiguobacterium* could contribute to the high abundance and diversity of PTS and perhaps other community functionalities detected in the BC sample. We further confirmed this by constructing a draft genome for the dominant *Exiguobacterium*-related OTU-6781. The recovered draft genome (with an average genome coverage and a completeness of 1428.4 and 72.1%, respectively) possessed 39 genes belonging to the carbohydrate-specific Enzymes II that were involved in the translocation and phosphorylation of various carbon sources, including phosphorylate beta-glucoside, cellobiose, maltose, glucose, oligo-beta-mannoside, mannitol, N-acetylglucosamine, N-acetylmuramic acid, and N-acetylmannosamine. Similar results could be identified in publicly available *Exiguobacterium* complete genomes that possess diverse carbohydrate-specific components. Meanwhile, the phosphorylation status of PTS components reflected the energy conditions of the cell, which could be converted to signals that eventually led to catabolite repression (Kotrba et al., 2001). Collectively, PTS might contribute to the adaptation of *Exiguobacterium* in the water softening process under rapid pH changes and in a wide variety of ecological niches as mentioned in the previous section.

Moving to the FW sample where residual disinfectant was added, most metabolic pathways remained at the same level as in the BC sample. A clear increase in LPS biosynthesis [ko00540] pathway was observed (Figure 3.4). LPS is the major component of the outer membrane of Gram-negative bacteria, including Proteobacteria and Cyanobacteria. This correlated well with the increase in the abundance of Alpha-, Beta-, and Gamma-Proteobacteria and Cyanobacteria in the community (Figure 3.2B). However, the high abundance of methano-/methylotrophs in the community was not accompanied by the substantial increase of methane metabolism [ko00680]. Methane metabolism [ko00680] only increased moderately at this stage. This could be explained by the facultative nature of most methylotrophs. Except *Methylomonas*, all the other methylotrophs (*Methylotenera*, *Methylocystis*, and *Hyphomicrobium*) found in the system were facultative, indicating that they also possessed pathways other than methane metabolism to generate energy. Additionally, a significant decrease in PTS [ko02060] was observed and this agreed with the decrease in the abundance of the *Exiguobacterium*-related OTU in the community.

Lastly, moving from the FW sample to the distribution system water-phase (DSWP) and distribution system biofilm-phase (DSBP) samples, no substantial changes with the community-level functionality profile could be observed, suggesting the disturbance was less intensive at these stages in comparison to the abstraction, softening, and disinfection stages (Figure 3.4). Still, a few pathways were observed to be less represented in the DSWP samples, including LPS biosynthesis [ko00540], bacteria chemotaxis [ko02030], and flagellar assembly [ko02040], whereas pathways used for protein processing in endoplasmic reticulum [ko04141] became slightly more abundant (Figure 3.5). Unlike the DSWP, ABC transporter pathway [ko02010], bacterial chemotaxis [ko02030], and flagellar assembly [ko02040] also slightly increased with the DSBP samples. These pathways were related to nutrient uptake and mobility, and could be beneficial for microbial populations associated with biofilm growth.

3.5 Discussion

Eukaryotic predation as disturbances to the bacterial microbiome continuum The presence of eukaryotes including nematodes, fungi, amoeba, and flagellate in the system was confirmed with 18S rRNA genes extracted from metagenomes using EMIRGE (Figure 3.5). EMIRGE could reconstruct nearly full-length SSU genes in metagenomes and subsequently estimate the relative abundance of each reconstructed sequence in the community (Miller, 2013). For nematodes, *Plectus* spp. were present in half of the samples. The detection of a large number of eukaryotes inside drinking water systems has been reported previously (Valster et al., 2009; Buse et al., 2013; Lu et al., 2016). These eukaryotes could enter distribution systems through the physical breakthroughs in treatment processes, attaching on media particles released from filters, and contamination from surrounding environments during pipe breakage (Proctor and Hammes, 2015). In this study, their importance to the microbial community was reflected by the community dissimilarities when considering both prokaryotic and eukaryotic groups. When 18S rRNA gene sequences, approximately 42% of the small subunit ribosomal RNA (SSU rRNA) genes (16S rRNA genes plus 18S rRNA genes) extracted from the metagenomes, were added into the beta-diversity analysis, the DS3mg sample was separated from the rest of the samples, suggesting that eukaryotes might have a strong influence on the bacterial community structure of DS3 (Figure 3.6).

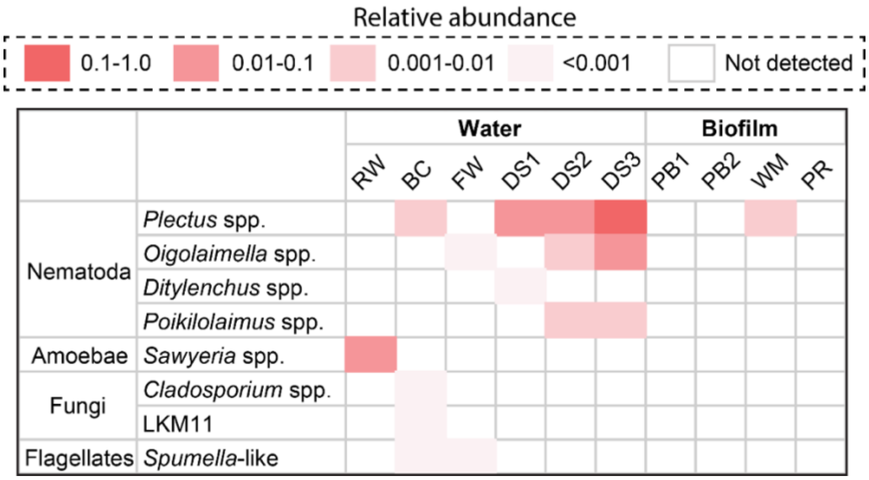


Figure 3.5 Relative abundance of eukaryotic groups calculated from SSU genes extracted from metagenomes.

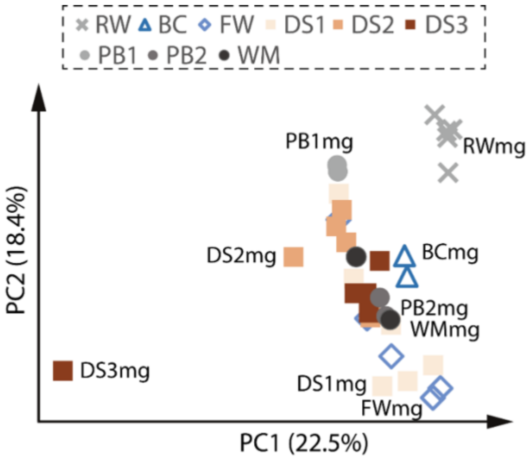


Figure 3.6 Beta- diversity of the microbial community using silva databases with eukaryotic sequences.

Eukaryotes could prey on bacterial communities present in the drinking water system. In the water phase, flagellates could be the main consumers of planktonic bacteria. Free-living amoebae might colonize biofilms on pipe surfaces or the surfaces of sediment particles in the distribution system. Under unfavorable conditions, they could exhibit a resistant form as cysts (Lienard et al., 2017). Nematodes could dwell in all possible niches in the drinking water system, but are particularly found in sand filters (Mott and Harrison, 1983; Locas et al., 2007). Also, bacteria from many phyla could survive under the grazing of eukaryotes. For example, *Mycobacterium* and *Legionella* could establish intracellular growth in free-living amoebae to obtain specific amino acids, replicate, evade disinfection, and spread to new environments (Kilvington and Price, 1990; Sauer et al., 2005; Delafont et al., 2014; Fonseca and Swanson, 2014). Similarly, they can colonize the intestinal tracks of some nematodes such as *Caenorhabditis* (Whittington et al., 2001; Komura et al., 2010). These opportunistic pathogens use these eukaryotic cells as hosts to protect themselves against chlorine, pH, and other stresses in the system (Cervero-Arago et al., 2015).

Potential bias associated with water sampling devices In this study, a large volume (~2000 L) of water with low cell numbers needed to be concentrated on-site for an extensive period of up to two days for metagenomics analysis. As the commonly-used laboratory concentration devices were not suitable for this purpose, a water purifier used and validated in a previous metagenomics study was adopted (Chao et al., 2013). It contained four filtration components, including prescreen, granular activated carbon (GAC), second screen, and a hollow fiber membrane filter. The measured pore size of the prescreen was 425 μm, and the pore size for the second screen was unknown. In our sampling events, the GAC component likely functioned as an absorption medium with minimal microbial growth due to the use of a sampling time shorter than the doubling time (11-23 days) for bacteria grown in drinking water biofilms (Pedersen, 1990; Block et al., 1993; Martiny et al., 2003). According to the manufacturer’s information, the inner hollow fiber membranes are made with polysulfone, and the average pore diameter of the inner hollow fiber membranes is approximately 0.01 µm (Shimagaki et al., 2000). Thus, most cells were trapped by this component. This sampling device is rather different from those commonly used in drinking water related studies, where water samples are filtered through a 0.22 µm polycarbonate membrane (Lin et al., 2014). This difference could potentially lead to differences in subsequent microbial community structure profiling using 16S rRNA-based techniques. The inconsistence caused by this sampling device will remain among future studies unless a standardized sampling protocol is proposed and adopted.

Modern drinking water treatment processes provide necessary protections to the public against microbial contamination by using an integrated multi-barrier approach. This approach focuses on source water protection, the use of effective treatment technologies (most importantly, filtration and disinfection), and the maintenance of the integrity of the distribution systems. The water microbiome can be considered as a continuum that travels through treatment facilities, distribution systems, and premise plumbing. Different disturbances are purposely introduced with an intention to produce and deliver 'pathogen-free' drinking water at the tap. Nevertheless, in this oligotrophic environment, microbial groups or guilds with unique metabolic functionalities could survive and grow at different stages. In this study, an extremophile *Exiguobacterium* was detected after lime treatment, and many microbes that were not completely killed through disinfection could survive and enter the distribution system, extending miles in length. Together with disinfection byproducts, the dissolved methane that was present in RW and not completely removed from the treatment process further served as a carbon source downstream to support the growth of methano-/methyltrophs, which further secreted by-products to other organisms. The co-existence of different eukaryotic groups and prokaryotes indicated that predation could cause disturbances to the bacterial microbiome in the distribution system. Overall, the ecological disturbance's perspective provides a basic theory behind the production of drinking water in the drinking water system studied, and a framework that can be applied to develop new drinking water processes with the intention to shape the microbiome continuum.

3.6 References

Bell, T., Bonsall, M.B., Buckling, A., Whiteley, A.S., Goodall, T., and Griffiths, R.I. (2010) Protists have divergent effects on bacterial diversity along a productivity gradient. *Biol Lett* **6**: 639-642.

Block, J.C., Haudidier, K., Paquin, J.L., Miazga, J., and Levi, Y. (1993) Biofilm accumulation in drinking water distribution systems. *Biofouling* **6**: 333-343.

Breiman, R.F., Fields, B.S., Sanden, G.N., Volmer, L., Meier, A., and Spika, J.S. (1990) Association of Shower Use with Legionnaires-Disease - Possible Role of Amebas. *JAMA* **263**: 2924-2926.

Buse, H.Y., Lu, J.R., Struewing, I.T., and Ashbolt, N.J. (2013) Eukaryotic diversity in premise drinking water using 18S rDNA sequencing: implications for health risks. *Environ Sci Pollut R* **20**: 6351-6366.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266-267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K. et al. (2010b) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335-336.

Cervero-Arago, S., Rodriguez-Martinez, S., Puertas-Bennasar, A., and Araujo, R.M. (2015) Effect of Common Drinking Water Disinfectants, Chlorine and Heat, on Free Legionella and Amoebae-Associated Legionella. *Plos One* **10**.

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X.X., Wu, W.M., and Zhang, T. (2013) Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Sci Rep* **3**: 3550.

Chistoserdova, L. (2014) Is metagenomics resolving identification of functions in microbial communities? *Microb Biotechnol* **7**: 1-4.

Chistoserdova, L., Kalyuzhnaya, M.G., and Lidstrom, M.E. (2009) The Expanding World of Methylotrophic Metabolism. *Annu Rev Microbiol* **63**: 477-499.

Connell, J.H. (1978) Diversity in Tropical Rain Forests and Coral Reefs - High Diversity of Trees and Corals Is Maintained Only in a Non-Equilibrium State. *Science* **199**: 1302-1310.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**: 485.

De Mesel, I., Derycke, S., Moens, T., Van der Gucht, K., Vincx, M., and Swings, J. (2004) Top-down impact of bacterivorous nematodes on the bacterial community structure: a microcosm study. *Environ Microbiol* **6**: 733-744.

Delafont, V., Mougari, F., Cambau, E., Joyeux, M., Bouchon, D., Hechard, Y., and Moulin, L. (2014) First Evidence of Amoebae-Mycobacteria Association in Drinking Water Network. *Environ Sci Technol* **48**: 11872-11882.

Deutscher, J., Ake, F.M.D., Derkaoui, M., Zebre, A.C., Cao, T.N., Bouraoui, H. et al. (2014) The Bacterial Phosphoenolpyruvate:Carbohydrate Phosphotransferase System: Regulation by Protein Phosphorylation and Phosphorylation-Dependent Protein-Protein Interactions. *Microbiol Mol Biol Rev* **78**: 231-256.

Di Rienzi, S.C., Sharon, I., Wrighton, K.C., Koren, O., Hug, L.A., Thomas, B.C. et al. (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *Elife* **2**.

Dupuy, M., Binet, M., Bouteleux, C., Herbelin, P., Soreau, S., and Hechard, Y. (2016) Permissiveness of freshly isolated environmental strains of amoebae for growth of Legionella pneumophila. *FEMS Microbiol Lett* **363**.

Flynn, T.M., Sanford, R.A., Ryu, H., Bethke, C.M., Levine, A.D., Ashbolt, N.J., and Domingo, J.W.S. (2013) Functional microbial diversity explains groundwater chemistry in a pristine aquifer. *BMC Microbiol* **13**.

Fonseca, M.V., and Swanson, M.S. (2014) Nutrient salvaging and metabolism by the intracellular pathogen Legionella pneumophila. *Front Cell Infect Microbiol* **4**.

Gasol, J.M., Pedros-Alio, C., and Vaque, D. (2002) Regulation of bacterial assemblages in oligotrophic plankton systems: results from experimental and empirical approaches. *Anton Leeuw Int J G* **81**: 435-452.

Gomez-Alvarez, V., Revetta, R.P., and Santo Domingo, J.W. (2012) Metagenomic analyses of drinking water receiving different disinfection treatments. *Appl Environ Microbiol* **78**: 6095-6102.

Griffiths, B.S., Bonkowski, M., Dobson, G., and Caul, S. (1999) Changes in soil microbial community structure in the presence of microbial-feeding nematodes and protozoa. *Pedobiologia* **43**: 297-304.

Gulay, A., Musovic, S., Albrechtsen, H.J., Abu Al-Soud, W., Sorensen, S.J., and Smets, B.F. (2016) Ecological patterns, diversity and core taxa of microbial communities in groundwater-fed rapid gravity filters. *ISME J* **10**: 2209-2222.

Hong, P.Y., Hwang, C.C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2010) Pyrosequencing analysis of bacterial biofilm communities in water meters of a drinking water distribution system. *Appl Environ Microbiol* **76**: 5631-5635.

Huston, M. (1979) A General Hypothesis of Species Diversity. *The American Naturalist* **113**: 81-101.

Hwang, C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2012) Microbial community dynamics of an urban drinking water distribution system subjected to phases of chloramination and chlorination treatments. *Appl Environ Microbiol* **78**: 7856-7865.

Jackson, B.E., Bhupathiraju, V.K., Tanner, R.S., Woese, C.R., and McInerney, M.J. (1999) Syntrophus aciditrophicus sp. nov., a new anaerobic bacterium that degrades fatty acids and benzoate in syntrophic association with hydrogen-using microorganisms. *Arch Microbiol* **171**: 107-114.

Janda, J.M., and Abbott, S.L. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. *J Clin Microbiol* **45**: 2761-2764.

Jousset, A. (2012) Ecological and evolutive implications of bacterial defences against predators. *Environ Microbiol* **14**: 1830-1843.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* **42**: D199-D205.

Kilvington, S., and Price, J. (1990) Survival of Legionella-Pneumophila within Cysts of Acanthamoeba-Polyphaga Following Chlorine Exposure. *J Appl Bacteriol* **68**: 519-525.

Kirk, M.F., Holm, T.R., Park, J., Jin, Q.S., Sanford, R.A., Fouke, B.W., and Bethke, C.M. (2004) Bacterial sulfate reduction limits natural arsenic contamination in groundwater. *Geology* **32**: 953-956.

Komura, T., Yasui, C., Miyamoto, H., and Nishikawa, Y. (2010) Caenorhabditis elegans as an Alternative Model Host for Legionella pneumophila, and Protective Effects of Bifidobacterium infantis. *Appl Environ Microbiol* **76**: 4105-4108.

Kotrba, P., Inui, M., and Yukawa, H. (2001) Bacterial phosphotransferase system (PTS) in carbohydrate uptake and control of carbon metabolism. *J Biosci Bioeng* **92**: 502-517.

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

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*: btv033.

Li, H., and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Lienard, J., Croxatto, A., Gervaix, A., Levi, Y., Loret, J.F., Posfay-Barbe, K.M., and Greub, G. (2017) Prevalence and diversity of Chlamydiales and other amoeba-resisting bacteria in domestic drinking water systems. *New Microbes New Infect* **15**: 107-116.

Lin, W.F., Yu, Z.S., Zhang, H.X., and Thompson, I.P. (2014) Diversity and dynamics of microbial communities at each step of treatment plant for potable water generation. *Water Res* **52**: 218-230.

Ling, F., Hwang, C., LeChevallier, M.W., Andersen, G.L., and Liu, W.T. (2016) Core-satellite populations and seasonality of water meter biofilms in a metropolitan drinking water distribution system. *ISME J* **10**: 582-595.

Locas, A., Barbeau, B., and Gauthier, V. (2007) Nematodes as a source of total coliforms in a distribution system. *Can J Microbiol* **53**: 580-585.

Lu, J., Struewing, I., Vereen, E., Kirby, A.E., Levy, K., Moe, C., and Ashbolt, N. (2016) Molecular Detection of Legionella spp. and their associations with Mycobacterium spp., Pseudomonas aeruginosa and amoeba hosts in a drinking water distribution system. *J Appl Microbiol* **120**: 509-521.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363-1371.

Martiny, A.C., Jorgensen, T.M., Albrechtsen, H.J., Arvin, E., and Molin, S. (2003) Long-term succession of structure and diversity of a biofilm formed in a model drinking water distribution system. *Appl Environ Microbiol* **69**: 6899-6907.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A. et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610-618.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E.M., Kubal, M. et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**.

Miller, C.S. (2013) Assembling full-length rRNA genes from short-read metagenomic sequence datasets using EMIRGE. *Methods Enzymol* **531**: 333-352.

Mott, J.B., and Harrison, A.D. (1983) Nematodes from river drift and surface drinking water supplies in southern Ontario. *Hydrobiologia* **102**: 27-38.

Nayfach, S., and Pollard, K.S. (2015) Average genome size estimation improves comparative metagenomics and sheds light on the functional ecology of the human microbiome. *Genome Biol* **16**: 51.

Oshkin, I.Y., Beck, D.A.C., Lamb, A.E., Tchesnokova, V., Benuska, G., McTaggart, T.L. et al. (2015) Methane-fed microbial microcosms show differential community dynamics and pinpoint taxa involved in communal response. *ISME J* **9**: 1119-1129.

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

Pedersen, K. (1990) Biofilm Development on Stainless-Steel and Pvc Surfaces in Drinking-Water. *Water Res* **24**: 239-243.

Pernthaler, J. (2005) Predation on prokaryotes in the water column and its ecological implications (vol 3, pg 537, 2005). *Nat Rev Microbiol* **3**.

Pinto, A.J., Xi, C.W., and Raskin, L. (2012) Bacterial community structure in the drinking water microbiome is governed by filtration processes. *Environ Sci Technol* **46**: 8851-8859.

Pinto, A.J., Schroeder, J., Lunn, M., Sloan, W., and Raskin, L. (2014) Spatial-temporal survey and occupancy-abundance modeling to predict bacterial community dynamics in the drinking water microbiome. *MBio* **5**: e01135-01114.

Proctor, C.R., and Hammes, F. (2015) Drinking water microbiology-from measurement to management. *Curr Opin Biotechnol* **33C**: 87-94.

Rajaei, S., Noghabi, K.A., Sadeghizadeh, M., and Zahiri, H.S. (2015) Characterization of a pH and detergent-tolerant, cold-adapted type I pullulanase from Exiguobacterium sp SH3. *Extremophiles* **19**: 1145-1155.

Ronn, R., McCaig, A.E., Griffiths, B.S., and Prosser, J.I. (2002) Impact of protozoan grazing on bacterial community structure in soil microcosms. *Appl Environ Microbiol* **68**: 6094-6105.

Rosario-Ortiz, F., Rose, J., Speight, V., von Gunten, U., and Schnoor, J. (2016) How do you like your tap water? *Science* **351**: 912-914.

Sauer, J.D., Bachman, M.A., and Swanson, M.S. (2005) The phagosomal transporter A couples threonine acquisition to differentiation and replication of Legionella pneumophila in macrophages. *P Natl Acad Sci USA* **102**: 9924-9929.

Schloss, P.D. (2010) The Effects of Alignment Quality, Distance Calculation Method, Sequence Filtering, and Region on the Analysis of 16S rRNA Gene-Based Studies. *PLoS Comput Biol* **6**.

Shimagaki, M., Fukui, F., Sonoda, T., and Sugita, K. (2000) Polysulfone hollow fiber semipermeable membrane. In: Toray Industries, Inc.

Snoeyink, V., Hass, C., Boulos, P., Burlinghame, G., Camper, A., Clark, R. et al. (2006) *Drinking water distribution systems: Assessing and reducing risks*: National Academics Press: Washington, DC.

Soo, R.M., Skennerton, C.T., Sekiguchi, Y., Imelfort, M., Paech, S.J., Dennis, P.G. et al. (2014) An Expanded Genomic Representation of the Phylum Cyanobacteria. *Genome Biol Evol* **6**: 1031-1045.

Valster, R.M., Wullings, B.A., Bakker, G., Smidt, H., and van der Kooij, D. (2009) Free-Living Protozoa in Two Unchlorinated Drinking Water Supplies, Identified by Phylogenic Analysis of 18S rRNA Gene Sequences. *Appl Environ Microb* **75**: 4736-4746.

Vishnivetskaya, T.A., Kathariou, S., and Tiedje, J.M. (2009) The Exiguobacterium genus: biodiversity and biogeography. *Extremophiles* **13**: 541-555.

Whittington, R.J., Lloyd, J.B., and Reddacliff, L.A. (2001) Recovery of Mycobacterium avium subsp paratuberculosis from nematode larvae cultured from the faeces of sheep with Johne's disease. *Vet Microbiol* **81**: 273-279.

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

Zeng, D.N., Fan, Z.Y., Chi, L., Wang, X., Qu, W.D., and Quan, Z.X. (2013) Analysis of the bacterial communities associated with different drinking water treatment processes. *World J Microb Biot* **29**: 1573-1584.

CHAPTER 4 BENEFITS OF GENOMIC INSIGHTS AND CRISPR-CAS SIGNATURES TO MONITOR POTENTIAL PATHOGENS ACROSS DRINKING WATER PRODUCTION AND DISTRIBUTION SYSTEMS

4.1 Abstract

The occurrence of pathogenic bacteria in drinking water distribution systems (DWDSs) is a major health concern, and our current understanding is mostly related to pathogenic species such as *Legionella pneumophila* and *Mycobacterium avium* but not to bacterial species closely related to them. In this study, genomic-based approaches were used to characterize pathogen-related species in relation to their abundance, diversity, potential pathogenicity, genetic exchange, and distribution across an urban drinking water system. Nine draft genomes recovered from ten metagenomes were identified as *Legionella* (4 draft genomes), *Mycobacterium* (3 draft genomes), *Parachlamydia* (1 draft genome), and *Leptospira* (1 draft genome). The pathogenicity potential of these genomes was examined by the presence/absence of virulence machinery, including genes belonging to Type III, IV, and VII secretion systems and their effectors. Several virulence factors known to pathogenic species were detected with these retrieved draft genomes except the *Leptospira*-related genome. Identical clustered regularly interspaced short palindromic repeats-CRISPR-associated proteins (CRISPR-Cas) genetic signatures were observed in two draft genomes recovered at different stages of the studied system, suggesting that the spacers in CRISPR-Cas could potentially be used as a biomarker in the monitoring of *Legionella* related strains at an evolutionary scale of several years across different drinking water production and distribution systems. Overall, metagenomics approach was an effective and complementary tool of culturing techniques to gain insights into the pathogenic characteristics and the CRISPR-Cas signatures of pathogen-related species in DWDSs.

4.2 Introduction

Over 500 waterborne or water-based pathogens of potential concern in drinking water (e.g., *Legionella pneumophila*, *Escherichia coli* O157:H7, *Mycobacterium avium*, and *Cryptosporidium parvum*) have been included in the Candidate Contaminant List by the US Environmental Protection Agency (Ashbolt, 2015). The traditional approach to identify these pathogens is through cultivation and then biochemical/serological tests or 16S rRNA gene-based phylogeny analysis (Lye and Dufour, 1993; Edberg et al., 1996; Stelma et al., 2004). However, identifying pathogens at species level does not always translate into health risks as some strains of the same species are more pathogenic than others (Schmidt and Schaechter, 2012).

Alternatively, comparative genomic analysis has become an effective way to evaluate the pathogenicity potential. It is reported that pathogens infect host through a multi-step process from entering the host, adhering to host tissues, penetrating or evading host defenses, damaging host tissues, to exiting the host. As a result, various virulence factors (VFs) are required for pathogenic species during the infection process, which can be divided into several general groups based on the conservation of similar mechanisms, such as adhesins, invasins, toxins, protein secretion systems, and antibiotic resistance mechanisms (Finlay and Falkow, 1997; Wilson et al., 2002). Thus, the presence of a set of virulence machinery in a bacterial genome has been used to define pathogenic subpopulations (Chapman et al., 2006; Cazalet et al., 2008; Bouzid et al., 2013; Foley et al., 2013; Picardeau, 2017). The knowledge on virulence machinery and the functions of key VFs in the literature have facilitated the usage of virulence machinery to evaluate health risks associated with pathogens in drinking water distribution systems (DWDSs) (Wu et al., 2008; Huang et al., 2014). Secretion systems are essential for the transportation of proteins (i.e., effectors) from the cytoplasm into host cells or host environments to enhance attachment to eukaryotic cells, scavenge resources in an environmental niche, and disrupt target cell functions (Green and Mecsas, 2016). Some secretion systems are dedicated for bacteria-host interaction, such as the type III secretion system (T3SS) in *Chlamydia* (Betts-Hampikian and Fields, 2010), the type IVB secretion system (T4BSS, Dot/Icm) in *Lg. pneumophila* (Voth et al., 2012), and the type VII secretion system (T7SS) in *Mycobacterium* (Costa et al., 2015). The deletion of these secretion systems could result in a substantial decrease in virulence (Costa et al., 2015). In addition, several other VFs have also been reported for pathogens including those facilitating attachment and invasion (e.g., cell wall, type IV pili) and endotoxins (i.e., lipopolysaccharides (LPS)) (Schroeder et al., 2010; Favrot et al., 2013; Tortora et al., 2013).

While the identification of pathogens of potential concern in DWDSs is an important task, recent studies have often detected pathogens simultaneously together with their closely related species, which are often present at higher abundance. These include, for example, *Lg. pneumophila*-related species such as *Lg. dumoffii* (Hsu et al., 1984), *Lg. sainthelensis* (Rodriguez-Martinez et al., 2015), and *Lg. jordanis* (Hsu et al., 1984; Kao et al., 2014), and *M. avium*-related species such as *M. gordonae* (Falkinham et al., 2001; Lalande et al., 2001; Vaerewijck et al., 2005), *M. immunogenum* (Gomez-Alvarez and Revetta, 2016a), and *M. chelonae* (Gomez-Alvarez and Revetta, 2016b). Some of these species have been associated with illness and infections in clinical environments, including *Lg. dumoffii* (Yu et al., 2002), *M. gordonae* (Lalande et al., 2001), *M. immunogenum* (Wilson et al., 2001), and *M. chelonae* (Lowry et al., 1990). As pathogens and their closely related species often share ecological niches (predominantly in biofilms), genetic exchange through conjugation and transformation occurs between the two groups, sometimes involving VFs (Gimenez et al., 2011; Gomez-Valero et al., 2011). However, it is not clear whether they possess similar VFs as observed in pathogens.

Furthermore, in DWDS ecosystems, pathogens and their closely related species mostly reside within biofilms where protozoa predation and viral lysis occur more frequent, and have developed mechanisms to resist predation by inhibiting phagosome acidification and lysosome fusion of protozoa (Hilbi et al., 2001; Tilney et al., 2001). Phage DNA can be integrated into bacterial genomes by horizontal gene transfer as prophages, which are major contributors to differences among individuals within a bacterial species (Bobay et al., 2014). To protect bacteria from phage lysis, encountered foreign DNA fragments can be integrated into a clustered regularly interspaced short palindromic repeats-CRISPR-associated proteins (CRISPR-Cas) locus as spacers (Makarova et al., 2015). Through addition of spacers at one end of the CRISPR array and conservation of spacers at the other end (the leader distal end), the CRISPR-Cas system participates in a constant evolutionary battle between phages and bacteria (Deveau et al., 2010; Sun et al., 2016). This mechanism has been used as a vital tool for strain typing in epidemiology for the recognition of outbreaks and identification of infection sources (Horvath et al., 2008; Shariat and Dudley, 2014). Nevertheless, it is not clear how intracellular growth and phage integration might impact the genomic composition and virulence of pathogen-related species.

In this study, metagenomics analysis instead of cultivation-based methods was carried out to investigate virulence machinery and genomic signatures as the result of phage integration of pathogens-related species in a drinking water production and distribution system. A groundwater-derived drinking water system studied previously (Ling et al., 2016; Zhang et al., 2017) was used as a model system. It consists of abstraction, softening, recarbonation, disinfection, filtration, and final distribution with a disinfectant residual (free chlorine). Samples of microbial biomass from ten locations of the water production process and the distribution system were collected and community metagenomes sequenced (Zhang et al., 2017). Coupling digital droplet PCR (ddPCR) with metagenomics, draft genomes affiliated with known pathogen genera were recovered to reveal their abundance, diversity, potential pathogenicity, genetic exchange, and distribution across an urban drinking water system.

4.3 Materials and methods

Sampling and DNA extraction Microbial biomass samples from different stages of the treatment processes and different locations in the distribution system were collected from a groundwater-sourced drinking water system. Detailed description of the studied drinking water system can be found in a previous study (Zhang et al., 2017). Briefly, these samples were from raw water (RW), immediately before filtration and chlorination (BC), finished water (FW) prior to distribution, three taps (DS1-DS3), two retired water mains (PB1-PB2), 14 household water meters (WM, combined into one sample), and five premise plumbing pipe reactors (PR, combined into one sample). The three tap water sampling sites (DS1-3) were located approximately one mile apart from each other to represent different locations within the DWDS. For water-phase samples (including RW, BC, FW, and DS1-3), a ten-minute flushing (the cold-water side) was carried out before each sampling event to minimize the influence of premise plumbing before installing point-of-use water purifiers (Toray Industries Inc. Japan). Approximately 2,000 L of water was filtered during each sampling event at each site over a time span of 48 hrs. Water purifiers were collected at the end of each sampling event and transported to the laboratory in cools (the Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign). They were disassembled after arriving at the laboratory and cells were washed off from the multilayer hollow fiber membrane with phosphate-buffered saline (PBS) through sonication (Symphony™ Ultrasonic Cleaners, VWR). The obtained mixture was filtered through 0.22 µm membranes and the membranes with cells were stored at -80 °C. To obtain a better representation of the average composition, water-phase sampling was repeated four times, in June, July, August, and September 2014, except the BC sample due to membrane blockage (Zhang et al., 2017).

For biofilm samples, PB1 was a 2.25-inch cast iron water main installed in 1968 and PB2 was a 1.5-inch cast iron water main installed prior to 1927. Each pipe was cut into two

12-inch long pieces on site with an effort to minimize contamination. Additionally, 14 water meters were obtained through the local drinking water plant. For the PR sample, five galvanized pipes of the plumbing system of a dormitory were obtained within the service area of the studied system, which were installed before World War II (size = 2 inch, OD = 2.375 inch, ID = 2.067 inch, length = 14 feet). Detailed description and handling of these samples could be found in our previous study (Zhang et al., 2017). The biofilm samples were swabbed off the surfaces, re-suspended in phosphate-buffered saline (PBS), and collected by filtering through 0.22 µm membranes. All the membranes with cells were stored at -80 °C. Genomic DNA (gDNA) was extracted using FastDNA® SPIN Kit for Soil (MP Biomedicals, Carlsbad, CA, USA) from these membranes with cells following manufacturer's protocol with an elution volume of 50 µl. The effect of different DNA extraction methods on the quantity and quality of DNA yields from drinking water biofilms had been evaluated and published in a previous study (Hwang et al., 2012).

ddPCR and real-time PCR ddPCR was used to quantify total *Bacteria* and *Archaea* 16S rRNA genes and pathogens of potential concern, including *Mycobacterium* spp., *M. tuberculosis* complex, *Legionella* spp., *Lg. pneumophila*, *Pseudomonas aeruginosa*, and *Aeromonas hydrophila*, in the combined samples submitted for metagenomic sequencing, except DS1 and DS3 due to not enough gDNA. TaqMan-based ddPCR assays using primer/probe sets specific to each target were performed with a QX200^{TM} Droplet Digital^{TM} PCR System using ddPCR^{TM} Supermix for Probes (Bio-Rad, Pleasanton, CA, USA). In addition, three eukaryotic groups (amoebae), *Naegleria fowleri*, *Acanthamoeba* spp., and *Balamuthia madrillaris*, were tested with TaqMan-based real-time PCR assays using primer/probe sets specific to internal transcribed spacer (ITS)/18S rRNA gene of each target. Real-time PCR was performed with a CFX96^{TM} Real-Time PCR Detection System using SsoAdvanced^{TM} Universal Probes Supermix (Bio-Rad, Pleasanton, CA, USA). Because of the large variations in the number of ITS/18S rRNA genes in different eukaryotic species, only cycle threshold (C_T) values were reported. Positive control (standard plasmid DNA) and negative control (H_2O) were included in every ddPCR and real-time PCR reaction to ensure the successful amplification and the absence of contamination, respectively.

Amplicon sequencing and metagenome sequencing analyses 16S rRNA gene amplicon analysis was carried out using a universal primer set targeting the V4-V5 hypervariable regions of both the Bacteria and Archaea domains (515F: 5’-GTGCCAGCMGCCGCGGTAA-3’ and 909R: 5’-CCCGTCAATTCMTTTRAGT-3’) using the Illumina Miseq platform with dual indexing strategy as described in a previous study (Zhang et al., 2017). DNA libraries for metagenomic sequencing were prepared by combining all the extracted gDNA from each sampling site due to the requirement of a relatively large amount of gDNA (> 0.1 μg). The prepared library was paired-end sequenced on Illumina HiSeq2500 platforms (Illumina, Inc., San Diego, CA, USA) as described previously (Zhang et al., 2017).

16S rRNA gene sequencing analysis The obtained paired-end 16S rRNA gene sequences were aligned with Mothur (Kozich et al., 2013). The resulting sequences were screened for chimeras by the UCHIME algorithm implemented in USEARCH 6.1 and processed using the *de novo* OTU picking workflow in QIIME as described previously (Zhang et al., 2017). EMIRGE was used to reconstruct nearly full-length SSU genes in metagenomes (Miller, 2013).

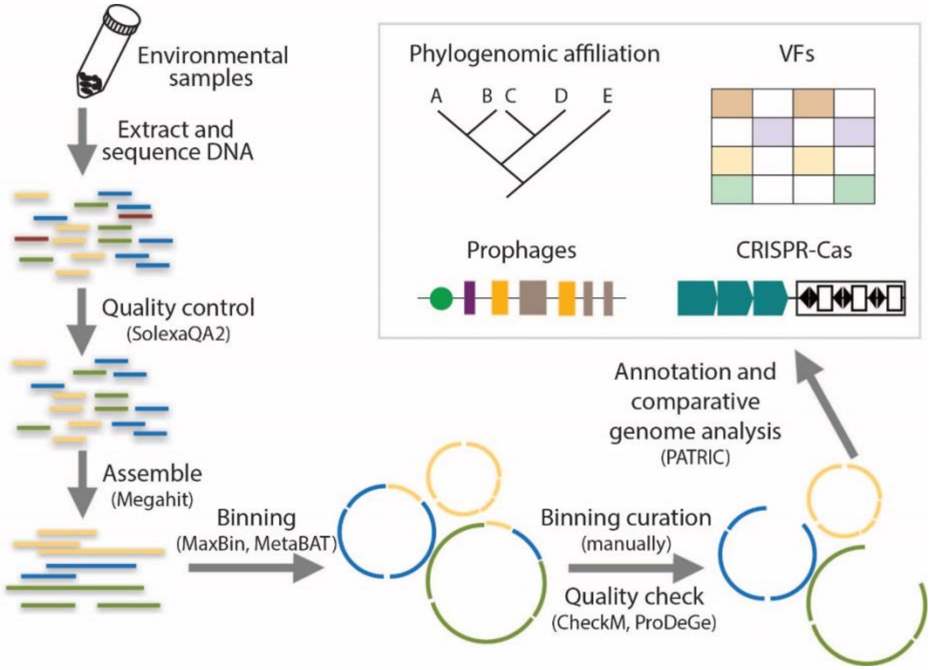


Figure 4.1 Flowing chart on the analysis of metagenomic data for the reconstruction of draft genomes and the identification of genetic signatures.

Draft genome reconstruction Draft genomes are presented as a set of sequence fragments or contigs, which are the most common form of genome assemblies obtained using metagenomics sequencing binning pipelines and account for two thirds of the bacterial genomes available in the GenBank database (Nagarajan et al., 2010; Edwards and Holt, 2013). Figure 4.1 illustrates the workflow of draft genome recovery used in this study. All the metagenomic datasets were trimmed using SolexaQA2 based on a cutoff of 20 by phred scores (Cox et al., 2010) and assembled using Megahit (Li et al., 2015). High-quality contigs (approximately 2.0 ×10^8 bp for each metagenome) were obtained at this step, to which > 85.0% of the raw reads could be mapped except the RW sample. The longest contig in each metagenome was > 4.0 ×10^5 bp. More details of the assemblies could be found in our previous study (Zhang et al., 2017). The obtained contigs were binned based on metagenomics read coverage, tetranucleotide frequency, and the occurrence of unique marker genes by using both MaxBin 2.0 (Wu et al., 2016) and MetaBAT (Kang et al., 2015) to minimize the contamination of each bin. These two binning methods employed different clustering methods for the determination of different bins: MaxBin compares the distributions of distances between and within the same bins whereas MetaBAT clusters contigs iteratively by modified K-methods algorithm. Bins of pathogen-related species from the two binning tools were compared and assessed with CheckM (Parks et al., 2015) and ProDeGe (Tennessen et al., 2016), followed by manual curation. The curated bins with ≥ 90% completeness and ≥ 15-fold coverage were finalized as draft genomes. Details of each step in the pipeline had been reviewed and summarized by Sangwan et al. (Sangwan et al., 2016) and a step-by-step tutorial of the workflow supplied with a sample dataset had been available by Edwards and Holt (Edwards and Holt, 2013). Percentages of reads mapped over the refined genome bins were estimated by Burrow-Wheeler Aligner-mem (Li and Durbin, 2009). The entire workflow was computed on a high-performance workstation (DELL precision T7600) equipped with 136 GB memory.

Identification of VFs   Draft genomes of pathogen-related species retrieved were uploaded into PATRIC for annotation and feature identification (Wattam et al., 2014). VFs of different pathogens were collected from the literature and the VF database (VFDB, http://www.mgc.ac.cn/VFs/) (Chen et al., 2012). Reported virulence genes within *Lg. pneumophila* included: the type II secretion system (T2SS, Lsp) for growth at low temperatures (Soderberg et al., 2008); the T4ASS (Lvh, F-type, and P-type) associated with conjugal DNA transfer and potentially in virulence (Gomez-Valero et al., 2011); the T4BSS (Dot/Icm) translocating several hundred effector proteins to support intracellular growth (Burstein et al., 2016); T4BSS-type effectors such as *ralF*, *lidA*, *sdhA*, and *lepAB* genes (Newton et al., 2010); type IV pili (*pilB,C,D*) involving in the entry to host cells, biofilm development, formation, type II protein secretion, and horizontal gene transfer (Schroeder et al., 2010); LPS transport (Lpt) proteins; and *mip* (macrophage infectivity potentiator) gene associated with the ability of *Lg. pneumophila* to replicate in eukaryotic cells (Newton et al., 2010).

For *M. tuberculosis*, the reported VFs included: the T7SS, also known as the ESX pathway (ESX-1 to ESX-5) to secrete proteins across their complex cell envelope (Houben et al., 2014); early secretory antigenic target (ESAT6), *esxA*, *H*, and *N*; culture filtrate protein-10 kDa (CFP-10), *esxB*, *G*, and *M* (Li et al., 2005); *pe*/*ppe* genes unique to mycobacteria and abundant in pathogenic mycobacteria (Sampson, 2011); antigen 85 (*ag85*) complex and mycolic acid cyclopropane synthase (*pcaA*) required for the biosynthesis of major components of the cell envelope (Favrot et al., 2013); adhesin (*hbhA*); phospholipase C (*plcC*); and oxidative stress reducer (*ahpC*) (Forrellad et al., 2013).

For leptospires, some potential VFs identified in the literature included: *lipL32, mce, invA, atsE, mviN, rfb* for attachment and invasion and *asd, trpE,* and *sphH* for amino acid biosynthesis (Ren et al., 2003; Ko et al., 2009; Fouts et al., 2016).

For *Parachlamydia,* known VFs included: negative regulator of the T3SS, SctW; protein kinase, Pkn5; translocated actin-recruiting phosphoprotein, *tarp*; inclusion membrane proteins IncA to IncG; translocator protein, CopB; modulation of host cell apoptosis, CADD; and Mip (Greub, 2009; Betts-Hampikian and Fields, 2010; Collingro et al., 2011; Croxatto et al., 2013). Furthermore, genes coding for nucleotide transporters that import host cell ATP in exchange for ADP (*ntt*) were part of the complex involving in bacteria-host interaction, but were generally not considered as VFs (Schmitz-Esser et al., 2004; Haferkamp et al., 2006).

Construction of phylogenomic tree PhyloPhlAn (Segata et al., 2013) was used to construct phylogenomic trees based on draft genomes and reference genomes. The constructed trees were visualized using iTOL (Letunic and Bork, 2016).

Identification of antibiotic resistance genes (ARGs) and CRISPR-Cas loci ARGs and CRISPR-Cas regions were screened with PATRIC. The identified CRISPR loci and ARGs were confirmed with CRISPRfinder (Grissa et al., 2007) and ResFinder (Zankari et al., 2012), respectively. Identified CRIPSR-Cas loci were classified into the current system consisting of two classes, five types and 16 subtypes (type I-A to I-F and I-U, type II-A to II-C, type III-A to III-D, type IV, and type V) based on *cas* genes and additional signature genes (Makarova et al., 2015). Additionally, we investigated the possible targets (protospacers) of spacers in CRISPR-Cas arrays within the obtained draft genomes using CRISPRTarget to search against all the available databases (i.e., GenBank-Phage, GenBank-Environmental, RefSeq-Plasmid, RefSeq-Viral, and RefSeq-Bacteria), which was combined with the known features of each subtype that had been reported to be essential for target recognition, such as protospacer adjacent motifs (PAMs) and seed regions (Biswas et al., 2013). Extra weighting was given to known PAMs: 5'-GG-3' for I-F (Mojica et al., 2009) at the 3' region of protospacer and 5'-CCN-3' for II-B (Fonfara et al., 2014) at the 5' region of protospacer. Moreover, we also manually examined seed sequences (8-nt for Type I-F and 13-nt for Type II-B) within the match. PHAST was used to identify prophage sequences in these draft genomes (Zhou et al., 2011).

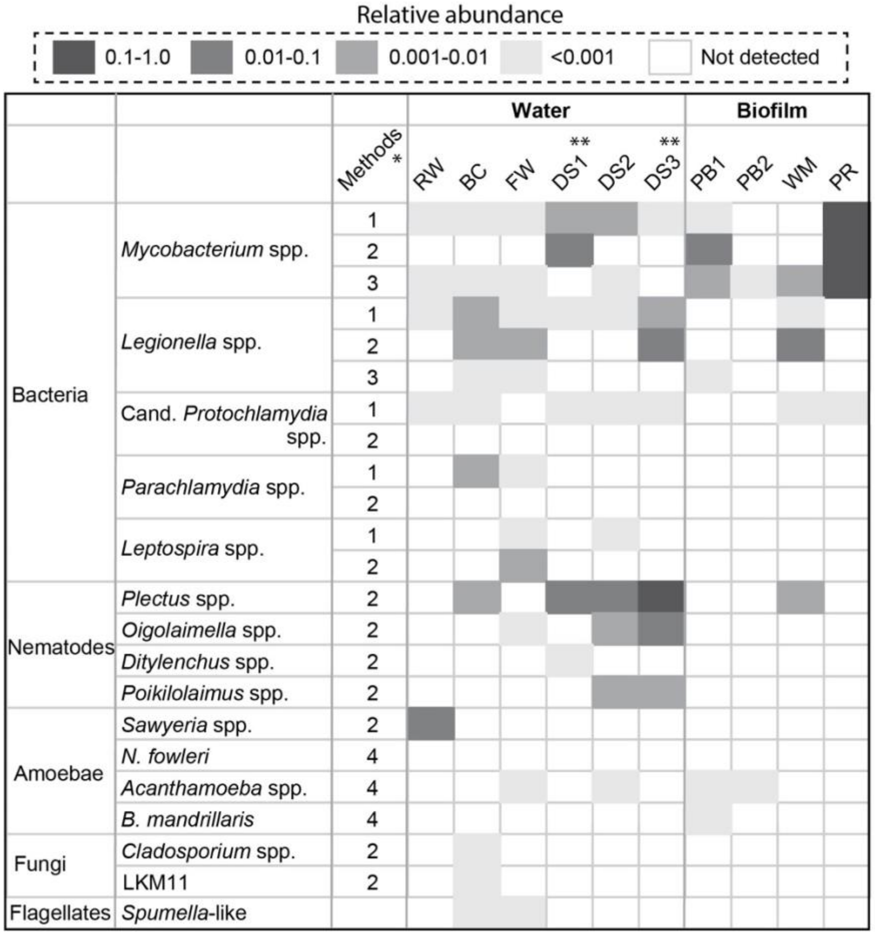
Genomic data depositing The nine draft genomes reconstructed in this study are deposited in GenBank under the BioProject PRJNA323575 with BioSamples SAMN07572181- SAMN07572189.

4.4 Results

Detection of pathogens of potential concern in the system A combination of different molecular biological techniques, namely, 16S rRNA gene amplicon sequencing, metagenomics, and ddPCR/real-time PCR was employed to investigate the diversity and quantity of potential pathogens in the drinking water production and distribution system.

In general, the distribution system samples contained the highest relative abundance of *Mycobacterium* spp. and *Legionella* spp. in comparison with samples from the treatment process (Figure 4.2). The highest level of *Mycobacterium* spp. was detected with the PR sample with a relative abundance of 1.3×10^{-1} and an absolute concentration of 3.3×10^{4} copies/ng-gDNA by ddPCR. The BC sample contained the highest level of *Legionella* spp.: a relative abundance of 4.7×10^{-3} based on 16S rRNA amplicon analysis and a concentration of 40.9 copies/ng-gDNA by ddPCR. Despite the occurrence of potential pathogens at the genus level, known pathogenic species, including *M. tuberculosis* complex, *Lg. pneumophila*, and *A. hydrophila* were not detected. Additionally, sequences related to *Candidatus* Protochlamydia spp., *Parachlamydia* spp., and *Leptospira* spp. were also detected (Figure 4.2). *Candidatus* Protochlamydia spp. and *Parachlamydia* spp. were endosymbionts of amoeba and emerging agents of pneumonia (Greub, 2009). Notably, *Candidatus* Protochlamydia spp. were detected in all the distribution water phase samples.

Meanwhile, we could identify various eukaryotes, such as nematodes, amoebae, and flagellates with metagenomics and real-time PCR that co-existed with these potential pathogens. *Plectus* spp. were the most abundant nematodes detected in the system and present in half of the samples. For amoebae, *Acanthamoeba* spp. were observed in FW, DS2, PB1 and PB2 while *Sawyeria* spp. were only found in RW.



Relative abundance

■ 0.1-1.0 ▦ 0.01-0.1 ▨ 0.001-0.01 ░ <0.001 □ Not detected

			Water						Biofilm			
		Methods *	RW	BC	FW	DS1 **	DS2	DS3 **	PB1	PB2	WM	PR
Bacteria	*Mycobacterium* spp.	1	░	░	░	▨	▨	░	░			■
		2				▦			▦			■
		3	░	░	░		░		▨	░	▨	■
	Legionella spp.	1	░	▨	░	░	░	▨			░	
		2		▨	▨			▦			▦	
		3		░	░				░			
	Cand. *Protochlamydia* spp.	1	░	░		░	░	░			░	░
		2										
	Parachlamydia spp.	1		▨	░							
		2										
	Leptospira spp.	1			░		░					
		2			▨							
Nematodes	*Plectus* spp.	2		▨		▦	▦	■			▨	
	Oigolaimella spp.	2			░		▨	▦				
	Ditylenchus spp.	2				░						
	Poikilolaimus spp.	2					▨	▨				
Amoebae	*Sawyeria* spp.	2	▦									
	N. fowleri	4										
	Acanthamoeba spp.	4			░		░		░	░		
	B. mandrillaris	4							░			
Fungi	*Cladosporium* spp.	2		░								
	LKM11	2		░								
Flagellates	*Spumella*-like			░	░							

Figure 4.2 Detected potential pathogens and eukaryotes (nematodes, amoebae, fungi, and flagellates) by 16S rRNA gene amplicon analysis (1), SSU genes extracted from metagenomes (2), ddPCR (3) and real-time PCR (4). Here, relative abundance was reported. In Method 4, only C_T value was obtained and is shown with the lightest color if it is positive. DS1 and DS3 were not tested by ddPCR due to not enough gDNA. The samples were divided into water and biofilm phases.

Characterization of pathogen-related species through the construction of draft genomes

Nine draft genomes closely related to known pathogens were successfully recovered from the metagenomes of BC, FW, DS1-3, and PR with ≥ 90% completeness and ≥ 15-fold coverage (Table 4.1). Figure 4.3 shows that four draft genomes were affiliated with *Legionella* (BC.3.64, FW.3.37, DS3.009, BC.3.72) (Panel A), three with *Mycobacterium* (DS1.3.26, DS2.013, PR.002) (Panel B), one with *Leptospira* (FW.030) (Panel C), and one with *Parachlamydia* (BC.030) (Panel D). In Panel A, different species of *Legionella* were observed to co-exist in the same niche, i.e., BC.3.64 and BC.3.72 in the BC sample. FW.3.37 was observed to be 99.7% similarity to BC.3.64 in the average nucleotide identity (ANI) based on 400 marker genes. These three draft genomes probably represented new species of *Legionella* as they did not cluster together with any known species. A fourth draft genome, DS3.009, was affiliated with *Lg. drozanskii*. For *Mycobacterium* draft genomes, all three (DS1.3.26, DS2.013, PR.002) were closely related to *M. gordonae*. The *Leptospira* draft genome FW.030 was outside of the cluster containing mostly saprophytic species. Last, draft genome BC.030 fell between *Pa. acanthamoebae* and *Candidatus* Protochlamydia amoebophila. Collectively, five of the draft genomes retrieved were not closely related to any known isolated species, possibly due to the limitation of cultivation methods to recover microorganisms from drinking water systems so far.

Table 4.1 General features of recovered genomes of pathogen related species.

Bin ID	Source	Affiliation	Completeness	Coverage	# of contigs	Genome size (bp)	G+C content (%)	No. of protein-coding genes	Possibly missing genes	Median sequence size	Longest contig size
BC.3.64	BC	Legionella sp.	94.44	30.13	62	2.27E+06	40.1	2112	5	31419	150,921
BC.3.72	BC	Legionella sp.	94.51	23.78	22	1.95E+06	40.6	1829	11	74242	336,208
FW.3.37	FW	Legionella sp.	94.15	27.68	63	2.10E+06	40.3	1926	14	18840	221,613
DS3.009	DS3	Legionella sp.	98.83	45.78	140	3.36E+06	39.4	3159	39	16314	165,891
DS1.3.26	DS1	Mycobacterium sp.	99.86	79.34	217	7.43E+06	66.8	6689	64	16573	250,869
DS2.013	DS2	Mycobacterium sp.	99.86	23.74	219	7.96E+06	66.5	7334	77	15428	244,689
PR.002	PR	Mycobacterium sp.	89.12	451.94	919	6.78E+06	67.0	6179	120	4016	89,735
BC.030	BC	Parachlamydia sp.	100.00	24.81	39	3.04E+06	41.5	2763	15	54962	289,998
FW.030	FW	Leptospira sp.	95.88	15.42	114	3.73E+06	35.1	3613	19	15672	307,203

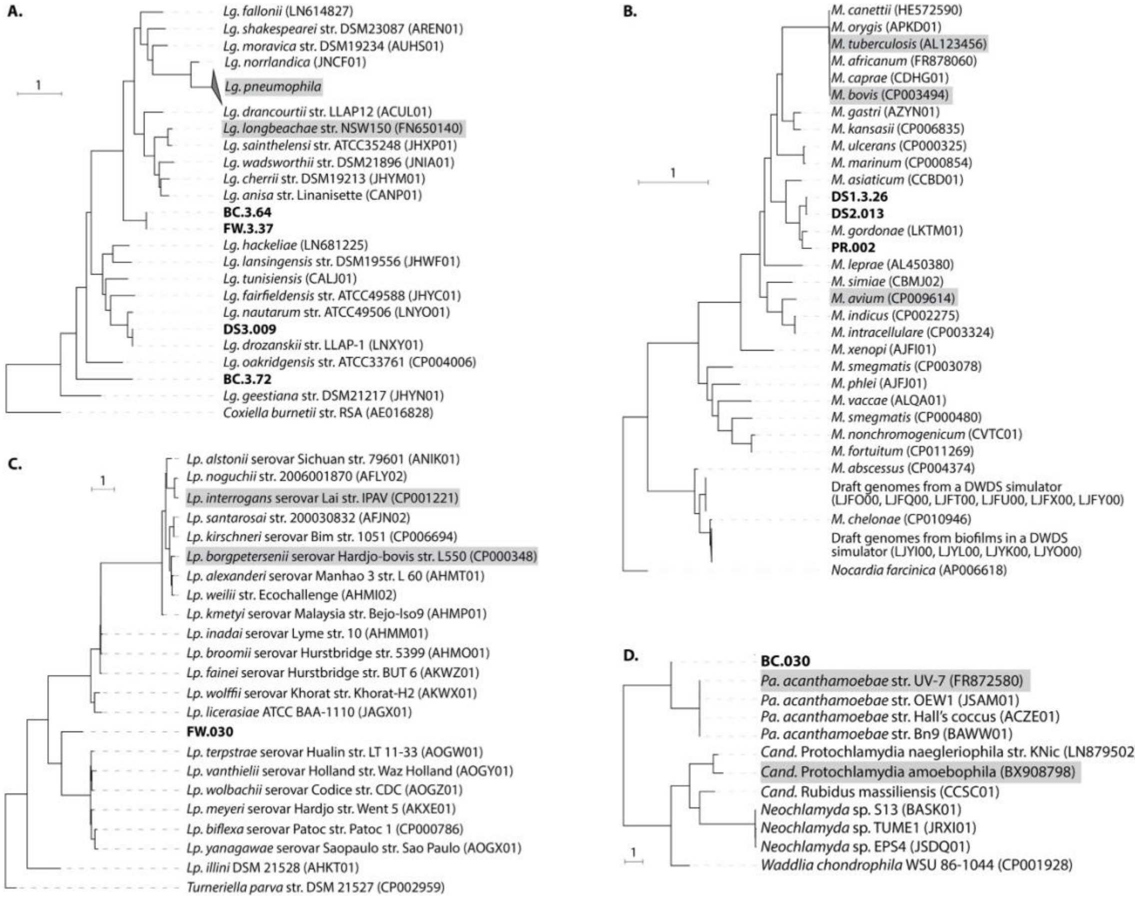

A.
1
Lg. fallonii (LN614827)
Lg. shakespearei str. DSM23087 (AREN01)
Lg. moravica str. DSM19234 (AUHS01)
Lg. norrlandica (JNCF01)
Lg. pneumophila
Lg. drancourtii str. LLAP12 (ACUL01)
Lg. longbeachae str. NSW150 (FN650140)
Lg. sainthelensi str. ATCC35248 (JHXP01)
Lg. wadsworthii str. DSM21896 (JNIA01)
Lg. cherrii str. DSM19213 (JHYM01)
Lg. anisa str. Linanisette (CANP01)
BC.3.64
FW.3.37
Lg. hackeliae (LN681225)
Lg. lansingensis str. DSM19556 (JHWF01)
Lg. tunisiensis (CALJ01)
Lg. fairfieldensis str. ATCC49588 (JHYC01)
Lg. nautarum str. ATCC49506 (LNYO01)
DS3.009
Lg. drozanskii str. LLAP-1 (LNXY01)
Lg. oakridgensis str. ATCC33761 (CP004006)
BC.3.72
Lg. geestiana str. DSM21217 (JHYN01)
Coxiella burnetii str. RSA (AE016828)
B.
1
M. canettii (HE572590)
M. orygis (APKD01)
M. tuberculosis (AL123456)
M. africanum (FR878060)
M. caprae (CDHG01)
M. bovis (CP003494)
M. gastri (AZYN01)
M. kansasii (CP006835)
M. ulcerans (CP000325)
M. marinum (CP000854)
M. asiaticum (CCBD01)
DS1.3.26
DS2.013
M. gordonae (LKTM01)
PR.002
M. leprae (AL450380)
M. simiae (CBMJ02)
M. avium (CP009614)
M. indicus (CP002275)
M. intracellulare (CP003324)
M. xenopi (AJFI01)
M. smegmatis (CP003078)
M. phlei (AJFJ01)
M. vaccae (ALQA01)
M. smegmatis (CP000480)
M. nonchromogenicum (CVTC01)
M. fortuitum (CP011269)
M. abscessus (CP004374)
Draft genomes from a DWDS simulator (LJFO00, LJFQ00, LJFT00, LJFU00, LJFX00, LJFY00)
M. chelonae (CP010946)
Draft genomes from biofilms in a DWDS simulator (LJYI00, LJYL00, LJYK00, LJYO00)
Nocardia farcinica (AP006618)
C.
1
Lp. alstonii serovar Sichuan str. 79601 (ANIK01)
Lp. noguchii str. 2006001870 (AFLY02)
Lp. interrogans serovar Lai str. IPAV (CP001221)
Lp. santarosai str. 200030832 (AFJN02)
Lp. kirschneri serovar Bim str. 1051 (CP006694)
Lp. borgpetersenii serovar Hardjo-bovis str. L550 (CP000348)
Lp. alexanderi serovar Manhao 3 str. L 60 (AHMT01)
Lp. weilii str. Ecochallenge (AHMI02)
Lp. kmetyi serovar Malaysia str. Bejo-Iso9 (AHMP01)
Lp. inadai serovar Lyme str. 10 (AHMM01)
Lp. broomii serovar Hurstbridge str. 5399 (AHMO01)
Lp. fainei serovar Hurstbridge str. BUT 6 (AKWZ01)
Lp. wolffii serovar Khorat str. Khorat-H2 (AKWX01)
Lp. licerasiae ATCC BAA-1110 (JAGX01)
FW.030
Lp. terpstrae serovar Hualin str. LT 11-33 (AOGW01)
Lp. vanthielii serovar Holland str. Waz Holland (AOGY01)
Lp. wolbachii serovar Codice str. CDC (AOGZ01)
Lp. meyeri serovar Hardjo str. Went 5 (AKXE01)
Lp. biflexa serovar Patoc str. Patoc 1 (CP000786)
Lp. yanagawae serovar Saopaulo str. Sao Paulo (AOGX01)
Lp. illini DSM 21528 (AHKT01)
Turneriella parva str. DSM 21527 (CP002959)
D.
BC.030
Pa. acanthamoebae str. UV-7 (FR872580)
Pa. acanthamoebae str. OEW1 (JSAM01)
Pa. acanthamoebae str. Hall's coccus (ACZE01)
Pa. acanthamoebae str. Bn9 (BAWW01)
Cand. Protochlamydia naegleriophila str. KNic (LN879502)
Cand. Protochlamydia amoebophila (BX908798)
Cand. Rubidus massiliensis (CCSC01)
Neochlamyda sp. S13 (BASK01)
Neochlamyda sp. TUME1 (JRXI01)
Neochlamyda sp. EPS4 (JSDQ01)
Waddlia chondrophila WSU 86-1044 (CP001928)
1

Figure 4.3 Phylogenomic tree of recovered draft genomes constructed based on up to 400 conserved protein sequences. Panel A: *Legionella*; Panel B: *Mycobacterium*; Panel C: *Leptospira*; Panel D: *Parachlamydia*. The nine draft genomes recovered from this study were bold. Known pathogenic species were shaded with grey. Scale bar, 1 expected substitutions per site.

VFs detected in the draft genomes recovered Figure 4.4 indicates the presence and absence of VFs affiliated with secretion systems, effectors, attachment and invasion, endotoxins (e.g., lipopolysaccharides), and amino acid biosynthesis found in the recovered draft genomes and their related reference genomes. For *Legionella* in the secretion system category, the T2SS and T4BSS were the major pathogenesis systems observed in all draft genomes recovered. By contrast, the T4ASS, associated with conjugal DNA transfer, was detected in BC.3.64 and DS3.009 but absent in BC.3.72 and FW.3.37 possibly due to non-existence in these bacteria or the inability or poor efficiency to retrieve and assemble sequences pertaining to these hypervariable regions (Pop, 2009; Gomez-Valero et al., 2011). In the effectors category, T4BSS-assicated VFs including *lidA*, *sdhA*, and *lepAB* genes but not *ralF* were detected in three of the four draft genomes. In addition, all draft genomes contained LPS transport related genes, *lptA* and *lptE*. Last, the *mip* gene was observed in BC.3.64, FW.3.37, and DS3.009, but not BC.3.72.

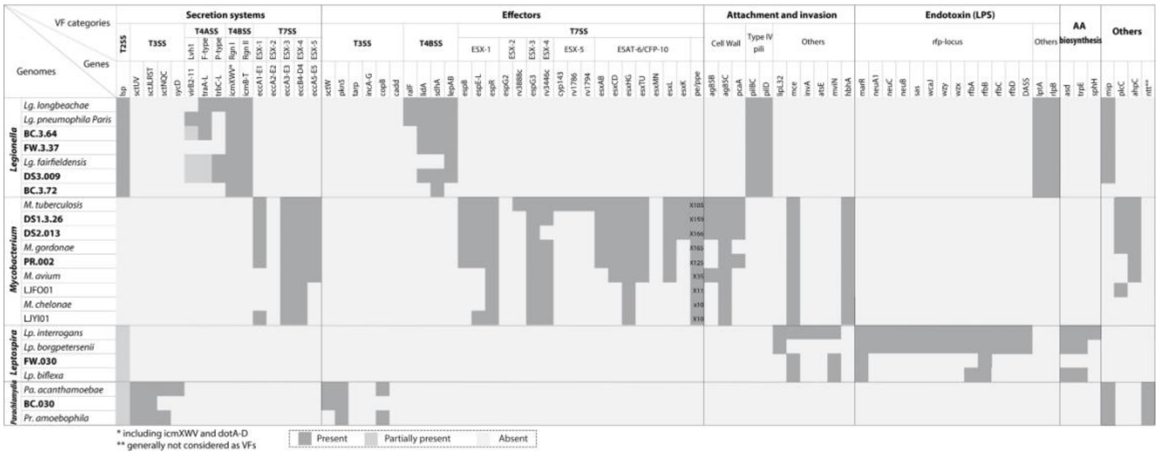


Figure 4.4 VFs identified with the draft genomes recovered in this study and related genomes from public databases. VFs were grouped based on general categories (secretion systems and associated effectors, attachment and invasion, endotoxin, amino acid biosynthesis and others). The genomes were organized by their taxonomic affiliations. There were some shared VFs among different genera, including T2SS among *Legionella*, *Leptospira*, and *Parachlamydia*, the *mip* gene between *Legionella* and *Parachlamydia*, and the *mce* gene between *Mycobacterium* and *Leptospira*.

For *Mycobacterium*, ESX-1, ESX-3, and ESX-5 T7SSs were observed in all *Mycobacterium* draft genomes recovered. Effectors belonging to ESX-1 and ESX-3 could also be detected, including *esxAB* and *TU*, but not effectors belonging to ESX-5 (*cyp143, rv1786, rv1794*, and *esxMN*). For the *pe*/*ppe* multigene family, all the recovered draft genomes contained more than 100 such genes, which was comparable to those observed in pathogenic species. Other VFs detected included cell envelop biosynthesis, *ag85* (except in PR.002) and *pca*A; adhesin, *hbhA*; phospholipase C, *plcC*; and oxidative stress reducer, *ahpC*. For *Leptospira*, the known VFs were mainly associated with the attachment and invasion, endotoxin and amino acid biosynthesis categories, and among them four (i.e., *mce1B, mviN, marR,* and *rfbD*) were detected in FW.030. The T2SS was partially present in *Leptospira* spp., including FW.030, but the association of the T2SS with virulence had not been experimentally tested (Picardeau, 2017). For *Parachlamydia*, VFs were mainly observed in the T3SS and associated effector categories. Two VFs, the T2SS (partially) and *mip* in the 'others' category were also observed. As *Parachlamdia* spp. and *Candidatus* Protochlamydia spp. were intracellular bacteria of amoebae like *Legionella* spp., they also possessed T2SSs and Mip systems. Five *ntt* genes were observed with BC.030, putatively belonging to three NTT isoforms (NTT1-3) (Haferkamp et al., 2006). Last, several ARGs related to the resistance of aminoglycoside (moderate level), beta-lactam, and chloramphenicol (antimicrobial peptides) could be detected in the *Legionella* draft genome DS3.009. All the *Mycobacterium* recovered draft genomes possessed the *aac(2')-Ic* gene, which was universally distributed among all *Mycobacterium* spp. (Ainsa et al., 1997).

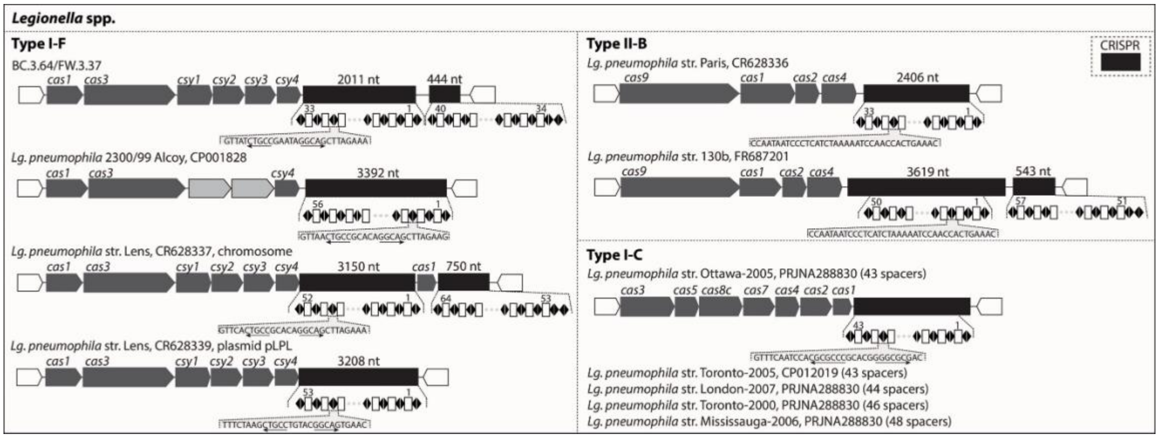

Legionella spp.
Type I-F
BC.3.64/FW.3.37
cas1 cas3 csy1 csy2 csy3 csy4 2011 nt 444 nt
33 1 40 34
GTTATCTGCCGAATAGGCAGCTTAGAAA
Lg. pneumophila 2300/99 Alcoy, CP001828
cas1 cas3 csy4 3392 nt
56 1
GTTAACTGCCGCACAGGCAGCTTAGAAG
Lg. pneumophila str. Lens, CR628337, chromosome
cas1 cas3 csy1 csy2 csy3 csy4 3150 nt cas1 750 nt
52 1 64 53
GTTCACTGCCGCACAGGCAGCTTAGAAA
Lg. pneumophila str. Lens, CR628339, plasmid pLPL
cas1 cas3 csy1 csy2 csy3 csy4 3208 nt
53 1
TTTCTAAGCTGCCTGTACGGCAGTGAAC
Type II-B
CRISPR
Lg. pneumophila str. Paris, CR628336
cas9 cas1 cas2 cas4 2406 nt
33 1
CCAATAATCCCTCATCTAAAAATCCAACCACTGAAAC
Lg. pneumophila str. 130b, FR687201
cas9 cas1 cas2 cas4 3619 nt 543 nt
50 1 57 51
CCAATAATCCCTCATCTAAAAATCCAACCACTGAAAC
Type I-C
Lg. pneumophila str. Ottawa-2005, PRJNA288830 (43 spacers)
cas3 cas5 cas8c cas7 cas4 cas2 cas1
43 1
GTTTCAATCCACGCGCCCGCACGGGGCGCGAC
Lg. pneumophila str. Toronto-2005, CP012019 (43 spacers)
Lg. pneumophila str. London-2007, PRJNA288830 (44 spacers)
Lg. pneumophila str. Toronto-2000, PRJNA288830 (46 spacers)
Lg. pneumophila str. Mississauga-2006, PRJNA288830 (48 spacers)

Figure 4.5 CRISPR-Cas loci identified in the draft genomes recovered in this study and known genomes of *Legionella*. They were organized according to the subtypes (Type I-F, II-B and I-C) of CRISPR-Cas loci.

Usage of CRISPR-Cas signatures to monitor Legionella spp. across the studied system

CRISPR-Cas genetic signatures, which are defense systems used by prokaryotes against viruses and not associated with pathogenicity, could be an effective tool to discriminate and monitor sub-lineages of pathogen-related species across the studied drinking water production and distribution system. Figure 4.5 indicates the type of CRISPR-Cas systems identified in the draft genomes recovered and in several published *Lg. pneumophila* genomes. Among the three known subtypes of *Lg. pneumophila* (I-F, II-B, and I-C), this study detected type I-F with BC.3.64 and FW.3.37 based on *cas* gene clusters. The type I-F CRISPR-Cas observed in these two draft genomes was almost identical, i.e., 99% sequence similarity for *cas1* gene and 100% sequence similarity for the remaining *cas* genes. Together with the findings of phylogenomic classification and genome similarity (99.7%), BC.3.64 and FW.3.37 were very likely to belong to a closely-related population originated from the same ancestor traveling from upstream (BC) to downstream (FW) of the studied drinking water production and distribution system (Figure 4.3). There was not enough information to determine whether the strain was alive at the BC site or whether filtration and chlorination had inactivated the strain in FW. Their *cas* gene clusters shared relatively low protein sequence similarities (from less than 40% to 76%) with other type I-F CRISPR-Cas loci of *Lg. pneumophila*. Last, a Type II-B CRISPR-Cas locus was detected with *Leptospira* draft genome FW.030.

Table 4.2 Prophages identified in the retrieved draft genomes.

Genera	Genomes	Regions	Length (kbp)	Possible phage
Legionella	BC.3.64	R1	9.5	*Salisaeta* icosahedral phage 1
		R2	31.1	Stenotrophomonas phage S1
	FW.3.37	R1	9.5	*Salisaeta* icosahedral phage 1
		R2	26.1	*Caulobacter* virus Karma
	DS3.009	R1	37.0	Stenotrophomonas phage S1
		R2	23.5	Haemophilus phage HP2
Mycobacterium	DS1.3.26	R1	19.0	Mycobacterium phage Adler
	DS2.103	R1	28.3	Mycobacterium phage RhynO
		R2	12.2	Molluscum contagiosum virus subtype 1
		R3	27.7	Mycobacterium phage Adler
		R4	31.6	Mycobacterium phage Adler
		R5	40.1	Mycobacterium phage Adler
	PR.002	R1	17.2	Mycobacterium phage Adler
		R2	37.1	Mycobacterium phage Milly
Leptospira	FW.030	R1	29.9	Pandoravirus salinus
Parachlamydia	BC.030	R1	19.3	*Cronobacter* phage vB_CsaM_GAP32

Diversity of prophages Table 4.2 shows the types of prophages found in the recovered draft genomes. Initially, 36 potential prophage sequences were identified using PHAST and they were reduced to 16 by considering the presence of genes encoding integrases and/or cI-type repressors (Fan et al., 2014). The lengths of prophage regions varied from 9.5 to 40.1 kbp. Six were associated with *Legionella* draft genomes, seven with *Mycobacterium* draft genome, and one each with *Parachlamydia* and *Leptospira*. An intact prophage (37.1 kbp) was recovered from PR.002. Shared prophage structures were observed between BC.3.64 and FW.3.37 and between DS1.3.26 and DS2.013. In addition, DS2.013 contained as many as five prophage sequences, which was rare for *Mycobacterium* genomes. Last, a prophage region identified in FW.030 showed sequence similarities to *Pandoravirus saline* which was the largest virus reported so far with genomes up to 2.5 Mb and restricted to *Acanthamoeba* as hosts (Philippe, 2013).

4.5 Discussion

Potential virulence of pathogen-related species Virulence machinery characterized by genomic analysis has been used to define pathogenicity for many known pathogens, such as *E. coli* (Chapman et al., 2006), *Salmonella* (Foley et al., 2013), *Cryptosporidium* (Bouzid et al., 2013), *Lg. pneumophila* (Cazalet et al., 2008), and *Leptospira* (Picardeau, 2017). This approach is used here to evaluate the potential pathogenicity of those draft genomes of pathogen-related species recovered from an urban drinking water system. *Legionella*-related draft genomes found at two different locations of the water production process (i.e., BC.3.64 and FW.3.37) shared almost identical genomic sequences and possessed almost all known VFs to *Lg. pneumophila* and *Lg. longbeachae*. Another strain found during the water production process (i.e., BC.3.72) was clustered outside of known pathogenic *Legionella* clusters, and possessed fewer virulence genes than the other three recovered strains (i.e., BC.3.64, FW.3.37, and DS3.009). While the finding that most of the draft genomes encoded a high number of VFs may raises concerns on their pathogenicity, previous studies on closely related species/strains of pathogenic *Aeromonas* found no correlations between the presence/absence of VFs and extraintestinal infections (Havelaar et al., 1992; Lye et al., 2007). Thus, further studies combining microbiological (e.g., cultivation and animal models), genomic, and metabolic (e.g., transcriptomics and proteomics) methods should be carried out to understand the role of these VFs at the level of gene expression, protein function and regulation, and interaction with host immune system to confirm the virulence of these strains for immunocompromised individuals. This framework, once established, can be transferred into a novel pathogen surveillance program that enables virulence assessment of a broad range of heterotrophic bacteria found in potable water to possibly identify currently unknown pathogens.

All three *Mycobacterium*-related draft genomes recovered were closely related to *M. gordonae*, which is less virulent than *M. tuberculosis*, but contained a high number of genes (over 100) related to *pe*/*ppe* and T7SS. In comparison, genomes of *M. immunogenum* (LJFO01) and *M. chelonae* (LJYI01) isolated from a chloraminated DWDS simulator in previous studies (Gomez-Alvarez and Revetta, 2016a; b) lacked ESX-1 or ESX-5 and contained fewer *pe/ppe* genes. Due to the prevalence of *M. gordonae* in tap water and biofilms, particularly in groundwater-derived drinking water systems (Vaerewijck et al., 2005), special attention to this group would be necessary. Pathogenic *Leptospira* are the causative agent of leptospirosis, which is the most widespread zoonotic disease infecting both human and animals (Evangelista and Coburn, 2010). In this study, the *Leptospira*-related genome FW.030 obtained did not contain most of the VFs known for *Lp. interrogans* and thus was likely not pathogenic. Among Parachlamydiaceae, only few strains such as *Pa. acanthamoebae* and *Candidatus Pr. naegleriophila* have been considered as emerging pathogens, causing mainly respiratory infections, while many others including *Neochlamydia hartmannellae* and *Pr. amoebophila* might be environmental strains or endosymbionts (Corsaro and Greub, 2006; Lamoth et al., 2011). Therefore, the pathogenic potential of *Parachlamydia*-related genome BC.030 remains to be further determined.

Use of spacers in CRISPR-Cas as biomarkers for Legionella subtyping Due to the high genome plasticity of *Legionella* species, molecular typing by a single marker gene has been difficult. For instance, the *mip* gene is associated with the ability of *Lg. pneumophila* to replicate in eukaryotic cells, and has been extensively used as a biomarker to detect the presence/absence of *Lg. pneumophila* in a sample (Gomez-Valero et al., 2009). It was detected in three *Legionella*-related draft genomes constructed in this study: BC.3.64 and FW.3.37 were closely related to *Lg. fallonii*, and DS3.009 to *Lg. drozanskii*. However, the *mip* gene was limited in differentiating the *Lg. pneumophila* subspecies *fraseri* from other subspecies. Thus, the European Working Group for Legionella Infections (EWGLI) has suggested that a combination of several biomarkers, including *flaA, pilE, asd, mip, mompS, proA*, and *neuA*, should be used to effectively identify *Lg. pneumophila* (Fry et al., 2000; Gaia et al., 2005; Ratzow et al., 2007). However, phylogenetic incongruence (i.e., different lineages of the same strain indicated by different biomarkers) and limitations (i.e., the inability of some biomarkers to discriminate certain strains) in the discriminatory power of these multiple biomarkers could still occur because of differences in selection pressures associated with individual biomarkers.

Alternatively, spacers in CRISPR-Cas can be used as a biomarker in the monitoring of certain *Legionella* strains at an evolutionary scale of several years across drinking water production and distribution systems. The pattern of adding new spacers at one end of the CRISPR array and conserving spacers among common ancestors at the other end has been demonstrated with *Legionella* strains collected in Canada and Europe (CRISPR Type I-C and Type II-B) (Ginevra et al., 2012; Lück et al., 2015; Rao et al., 2016). The longest time for these spacers to remain conserved among these strains and a *Leptospirillum* strain previously studied was reported to be five years or longer (Sun et al., 2016). Type I-F Cas loci were detected in the genomes of *Lg. pneumophila* str. 2300/99 Alcoy and str. Lens (both in the chromosome and plasmid) (Figure 4.5). The two draft genomes recovered in our study, BC.3.64 and FW.3.37, also contained type I-F CRISPR-Cas loci, but the spacers were different from str. 2300/99 Alcoy and str. Lens. With 100% sequence similarity in CRISPR and high overall genomic similarity, these two genomes were likely derived from the same ancestor. Thus a specific CRISPR-Cas biomarker could be developed and used to monitor the distribution of this strain within the drinking water system studied. Furthermore, Types II-B and I-C were detected in a variety of *Lg. pneumophila* strains and Type II-B was detected in 75.0% of the 400 *Lg.*

pneumophila strains collected in a previous study (Ginevra et al., 2012). With more than 600 *Legionella* genomes available with NCBI's website and the diversity of CRISPR-Cas Types (I-C, I-F, and II-B) known among these strains, CRISPR-Cas spacers will be a promising biomarker for monitoring the distribution of *Legionella* at the strain level in samples taken from various drinking water systems, across different water bodies, and between patients over several years. However, cautions are needed when applying this method over a relatively large evolutionary scale as previous reports on *Yersinia pestis*, *Streptococcus thermophiles*, and *Leptospirillum* suggested that CRISPR loci could also evolve via internal deletion of spacers in the CRISPR array (Pourcel et al., 2005; Horvath et al., 2008; Sun et al., 2016).

Origin of spacers in CRISPR-Cas of pathogen-related genomes The interaction between bacteria and viruses in drinking water systems or, more broadly, in oligotrophic environments is not well understood (Lehtola et al., 2004; Liu et al., 2015; Guidi et al., 2016). Table 4.3 shows only 26 out of the 119 identified CRISPR-Cas spacers matched to entries in databases including GenBank-Phage, GenBank-Environmental, RefSeq-Plasmid, RefSeq-Viral, and RefSeq-Bacteria. Among them, 13 spacers matched sequences in other *Lg. pneumophila* strains. Two commonly observed targets were a 30-kb unstable genetic element previously identified in *Lg. pneumophila* str. RC1 and a 60-kb plasmid in *Lg. pneumophila* str. Lens. Likely, these elements were originated from bacteriophages in environments and incorporated into *Lg. pneumophila* genomes as mobile genetic elements such as prophages and plasmids. When the DNA of *Lg. pneumophila* was damaged or under other stress conditions, prophages could be excised, replicated, and ultimately used to lyse the host and spread into the environment. Ecologically, it would be rational for other *Lg. pneumophila* strains to incorporate their fragments into CRISPR systems so that they had the ability to destroy them when being attacked (Rao et al., 2016).

We also observed near-perfect matches of four spacers in CRISPR-Cas to one activated sludge metagenome (AERA01) (More et al., 2014). It has been reported that wastewater treatment plants (WWTPs) contained 10-1000 times higher viral concentration than in natural aquatic environments, making WWTP an important reservoir and source of viruses (Edwards and Rohwer, 2005; Tamaki et al., 2012). In the studied drinking water production and distribution system, we estimated that the viral concentration was approximately 10^4 viruses/ml based on the bacterial cell counts published previously (Zhang et al., 2017) based on the general rule that viral count is 10 times of the bacterial count (Maranger and Bird, 1995). Additionally, spacers detected in the BC.3.64 and FW.3.37 genomes recovered here and *Lg. pneumophila* 2300/99 Alcoy matched to contigs in marine metagenomes (AACY02) (Venter et al., 2004). Although the matches are not perfect (except one) to organisms in WWTPs or marine environment, the evolving nature of spacers by mutations at CRISPR loci allows us to speculate that WWTPs and marine environments were possible sources of these spacers. Those *Legionella* strains could have come from water bodies under the influence of wastewater or seawater, such as flooded sewers or coastal groundwater.

Amoebae as a 'hub' connecting viruses and intracellular bacteria This study observed that the prophage exhibiting high sequence similarity to *Pandoravirus* could co-exist with *Acanthamoeba* spp., *Parachlamydia* spp., *Legionella* spp., and *Mycobacterium* spp. in the FW sample. So far, free-living amoebae in drinking water systems are reported to be an ideal shelter to provide nutritional requirements for the growth of *Legionella* (Breiman et al., 1990; Dupuy et al., 2016), and are the only reported host of Pandoravirus (Philippe et al., 2013). Various giant viruses, including *Mimivirus*, *Mamavirus*, and *Pandoravirus*, have been detected in amoebae and were reported to be involved in lateral gene transfer between viruses and bacteria (La Scola et al., 2003; La Scola et al., 2008; Philippe et al., 2013). While the detection of *Parachlamydia* in drinking water systems is rare (Thomas et al., 2008), previous studies have suggested that Chlamydiae were likely prevalent in aquatic environments (Barret et al., 2013; Lagkouvardos et al., 2014). These observations all support amoebae as the 'hub' connecting viruses and intracellular bacteria, and facilitating the genetic exchange between pathogens and their closely related species

(Gimenez et al., 2011; Gomez-Valero et al., 2011). Thus, developing control strategies to eukaryotic populations, e.g., filtration with 1 µm membranes, whose size is larger than bacteria but smaller than amoebae, could be an effective means to suppress the growth and spreading of pathogens in DWDSs (Wadowsky et al., 1988).

In summary, our study demonstrated that metagenomics analysis can be used to determine the presence of VFs in potential pathogens in drinking water production and distribution systems. Future studies combining microbiological, genomic, and metabolic methods at the level of gene expression, protein function and regulation, and bacteria-host interaction can help determine the relationship between the presence of these VFs and pathogenicity in immunocompromised individuals, especially for environmental strains recovered from drinking water systems. Furthermore, the development of genomics analysis can serve as a new platform for the detection, strain typing, and monitoring of pathogens, which can provide novel insights into the surveillance and control of waterborne or water-based pathogens. Characteristic regions in bacterial genomes, such as CRISPR-Cas studied here, can be used in combination with the traditional biomarkers to facilitate and simplify the subtyping of pathogens of potential concern and monitor the distribution of the same strains across different environmental niches.

Table 4.3 Potential targets of CRISPR-Cas spacers in *Legionella*-related genomes.

Genomes	Spacer ID	Hits for spacers	Score	Number of mismatches within the spacer	PAMs**	Seed sequence mismatch position
BC.3.64	Sp6	Marine metagenome genome assembly TARA_030_DCM_0.22 (CENH01030675)	27	5	GG	8
Lgp* Lens	Chrm_Sp23	*Lg. pneumophila* serogroup 1, 30 kb instable genetic element (AJ277755)	35	1	GG	6
	Chrm_Sp35	*Paenibacillus* sp. FSL H7-0357, complete genome (CP009241)	27	5	GG	3
	Plsm_Sp22	Activated sludge metagenome contig16020 (AERA01015926)	37	0	GG	-
	Plsm_Sp46	*Lg. pneumophila* serogroup 1, 30 kb instable genetic element (AJ277755)	35	1	GG	7
	Plsm_Sp12	*Lg. pneumophila* 2300/99 Alcoy, complete genome (NC_014125)	31	3	GG	7
	Plsm_Sp12	*Lg. pneumophila* str. Corby, complete genome (NC_009494)	31	3	GG	7
	Plsm_Sp10	*Lg. pneumophila* str. Paris complete genome (NC_006368)	30	1	Not match	N/A
	Plsm_Sp8	Uncultured marine Microviridae clone SOG3-01 major capsid protein gene, partial cds (KC131005)	29	4	GG	1
	Plsm_Sp47	Activated sludge metagenome contig16020 (AERA01015926)	29	4	GG	-
	Plsm_Sp50	Marine metagenome 1096626097875, whole genome shotgun sequence (AACY023989113)	29	4	GG	5
	Plsm_Sp7	Activated sludge metagenome contig06523 (AERA01006474)	29	5	GG	3,5
	Plsm_Sp13	*Lg. pneumophila* 2300/99 Alcoy, complete genome (NC_014125)	26	3	Not match	N/A
	Plsm_Sp32	*Lg. pneumophila* str. Lens plasmid pLPL, complete sequence (NC_006366)	24	4	Not match	N/A
	Plsm_Sp7	*Lg. pneumophila* str. Lens plasmid pLPL, complete sequence (NC_006366)	24	4	Not match	N/A

Table 4.3 (Cont.)

Genomes	Spacer ID	Hits for spacers	Score	Number of mismatches within the spacer	PAMs**	Seed sequence mismatch position
Lgp Alcoy	Sp32	Uncultured Gokushovirinae clone WSBWG10n1 major capsid protein gene (KF689311)	31	3	GG	8
	Sp28	Marine metagenome genome assembly TARA_122_SRF_0.1-0.22 (CETN01079705)	29	4	GG	-
	Sp3	*Lg. pneumophila* str. Lens plasmid pLPL (NC_006366)	26	3	Not match	N/A
Lgp Paris	Sp33	Activated sludge metagenome contig28417 (AERA01027227)	37	3	CCA	6,9
	Sp4	*Schistocephalus solidus* genome assembly S_solidus_NST_G2 (LL901847)	29	5	CCA	-
	Sp15	*Lg. pneumophila* str. Lens plasmid pLPL (NC_006366)	28	3	Not match	N/A
	Sp14	*Lg. pneumophila* 130b draft genome (FR687201)	28	4	Not match	N/A
Lgp 130b	Sp40	*Lg. pneumophila* str. Paris complete genome (NC_006368)	37	0	CCA	-
	Sp41	Hypersaline lake metagenome ctg7180000052828 (APHM01003927)	30	5	CCA	10
	Sp27	*Lg. pneumophila* str. Corby, complete genome (NC_009494)	30	2	Not match	N/A
	Sp27	*Lg. pneumophila* 2300/99 Alcoy chromosome (NC_014125)	30	2	Not match	N/A

*Lgp: *Lg. pneumophila*; **PAMs: protospacer adjacent motifs

4.6 References

Ainsa, J.A., Perez, E., Pelicic, V., Berthet, F.X., Gicquel, B., and Martin, C. (1997) Aminoglycoside 2'-N-acetyltransferase genes are universally present in mycobacteria: Characterization of the *aac(2')-lc* gene from *Mycobacterium tuberculosis* and the *aac(2')-ld* gene from *Mycobacterium smegmatis*. *Mol Microbiol* **24**: 431-441.

Ashbolt, N.J. (2015) Microbial Contamination of Drinking Water and Human Health from Community Water Systems. *Curr Environ Health Rep* **2**: 95-106.

Barret, M., Egan, F., and O'Gara, F. (2013) Distribution and diversity of bacterial secretion systems across metagenomic datasets. *Environ Microbiol Rep* **5**: 117-126.

Betts-Hampikian, H.J., and Fields, K.A. (2010) The chlamydial type III secretion mechanism: revealing cracks in a tough nut. *Front Microbiol* **1**.

Biswas, A., Gagnon, J.N., Brouns, S.J.J., Fineran, P.C., and Brown, C.M. (2013) CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. *RNA Biol* **10**: 817-827.

Bobay, L.M., Touchon, M., and Rocha, E.P.C. (2014) Pervasive domestication of defective prophages by bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **111**: 12127-12132.

Bouzid, M., Hunter, P.R., Chalmers, R.M., and Tyler, K.M. (2013) *Cryptosporidium* Pathogenicity and Virulence. *Clin Microbiol Rev* **26**: 115-134.

Breiman, R.F., Fields, B.S., Sanden, G.N., Volmer, L., Meier, A., and Spika, J.S. (1990) Association of Shower Use with Legionnaires-Disease - Possible Role of Amebas. *Jama-J Am Med Assoc* **263**: 2924-2926.

Burstein, D., Amaro, F., Zusman, T., Lifshitz, Z., Cohen, O., Gilbert, J.A. et al. (2016) Genomic analysis of 38 *Legionella* species identifies large and diverse effector repertoires. *Nat Genet* **48**: 167-175.

Cazalet, C., Jarraud, S., Ghavi-Helm, Y., Kunst, F., Glaser, P., Etienne, J., and Buchrieser, C. (2008) Multigenome analysis identifies a worldwide distributed epidemic *Legionella pneumophila* clone that emerged within a highly diverse species. *Genome Res* **18**: 431-441.

Chapman, T.A., Wu, X.Y., Barchia, I., Bettelheim, K.A., Driesen, S., Trott, D. et al. (2006) Comparison of virulence gene profiles of *Escherichia coli* strains isolated from healthy and diarrheic swine. *Appl Environ Microbiol* **72**: 4782-4795.

Chen, L.H., Xiong, Z.H., Sun, L.L., Yang, J., and Jin, Q. (2012) VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res* **40**: D641-D645.

Collingro, A., Tischler, P., Weinmaier, T., Penz, T., Heinz, E., Brunham, R.C. et al. (2011) Unity in Variety-The Pan-Genome of the Chlamydiae. *Mol Biol Evol* **28**: 3253-3270.

Corsaro, D., and Greub, G. (2006) Pathogenic potential of novel Chlamydiae and diagnostic approaches to infections due to these obligate intracellular bacteria. *Clin Microbiol Rev* **19**: 283-297.

Costa, T.R.D., Felisberto-Rodrigues, C., Meir, A., Prevost, M.S., Redzej, A., Trokter, M., and Waksman, G. (2015) Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat Rev Microbiol* **13**: 343-359.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**.

Croxatto, A., Murset, V., Chassot, B., and Greub, G. (2013) Early expression of the type III secretion system of *Parachlamydia acanthamoebae* during a replicative cycle within its natural host cell *Acanthamoeba castellanii*. *Pathog Dis* **69**: 159-175.

Deveau, H., Garneau, J.E., and Moineau, S. (2010) CRISPR/Cas System and Its Role in Phage-Bacteria Interactions. *Annu Rev Microbiol* **64**: 475-493.

Dupuy, M., Binet, M., Bouteleux, C., Herbelin, P., Soreau, S., and Hechard, Y. (2016) Permissiveness of freshly isolated environmental strains of amoebae for growth of *Legionella pneumophila*. *FEMS Microbiol Lett* **363**: fnw022.

Edberg, S.C., Gallo, P., and Kontnick, C. (1996) Analysis of the virulence characteristics of bacteria isolated from bottled, water cooler, and tap water. *Microb Ecol Health D* **9**: 67-77.

Edwards, D.J., and Holt, K.E. (2013) Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp* **3**: 2.

Edwards, R.A., and Rohwer, F. (2005) Viral metagenomics. *Nat Rev Microbiol* **3**: 504-510.

Evangelista, K.V., and Coburn, J. (2010) *Leptospira* as an emerging pathogen: a review of its biology, pathogenesis and host immune responses. *Future Microbiol* **5**: 1413-1425.

Falkinham, J.O., Norton, C.D., and LeChevallier, M.W. (2001) Factors influencing numbers of *Mycobacterium avium*, *Mycobacterium intracellulare*, and other mycobacteria in drinking water distribution systems. *Appl Environ Microbiol* **67**: 1225-1231.

Fan, X.Y., Xie, L.X., Li, W., and Xie, J.P. (2014) Prophage-like elements present in *Mycobacterium* genomes. *BMC Genomics* **15**.

Favrot, L., Grzegorzewicz, A.E., Lajiness, D.H., Marvin, R.K., Boucau, J., Isailovic, D. et al. (2013) Mechanism of inhibition of *Mycobacterium tuberculosis* antigen 85 by ebselen. *Nat Commun* **4**.

Finlay, B.B., and Falkow, S. (1997) Common themes in microbial pathogenicity revisited. *Microbiol Mol Biol Rev* **61**: 136-&.

Foley, S.L., Johnson, T.J., Ricke, S.C., Nayak, R., and Danzeisen, J. (2013) *Salmonella* pathogenicity and host adaptation in chicken-associated serovars. *Microbiol Mol Biol R* **77**: 582-607.

Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lecrivain, A.L., Bzdrenga, J. et al. (2014) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res* **42**: 2577-2590.

Forrellad, M.A., Klepp, L.I., Gioffre, A., Garcia, J.S.Y., Morbidoni, H.R., Santangelo, M.D. et al. (2013) Virulence factors of the *Mycobacterium tuberculosis* complex. *Virulence* **4**: 3-66.

Fouts, D.E., Matthias, M.A., Adhikarla, H., Adler, B., Amorim-Santos, L., Berg, D.E. et al. (2016) What makes a bacterial species pathogenic?: comparative genomic analysis of the genus *Leptospira*. *PLoS Negl Trop Dis* **10**: e0004403.

Fry, N.K., Bangsborg, J.M., Bernander, S., Etienne, J., Forsblom, B., Gaia, V. et al. (2000) Assessment of intercentre reproducibility and epidemiological concordance of *Legionella pneumophila* serogroup 1 genotyping by amplified fragment length polymorphism analysis. *Eur J Clin Microbiol Infect Dis* **19**: 773-780.

Gaia, V., Fry, N.K., Afshar, B., Luck, P.C., Meugnier, H., Etienne, J. et al. (2005) Consensus sequence-based scheme for epidemiological typing of clinical and environmental isolates of *Legionella pneumophila*. *J Clin Microbiol* **43**: 2047-2052.

Gimenez, G., Bertelli, C., Moliner, C., Robert, C., Raoult, D., Fournier, P.E., and Greub, G. (2011) Insight into cross-talk between intra-amoebal pathogens. *BMC Genomics* **12**.

Ginevra, C., Jacotin, N., Diancourt, L., Guigon, G., Arquilliere, R., Meugnier, H. et al. (2012) *Legionella pneumophila* sequence Type 1/Paris pulsotype subtyping by spoligotyping. *J Clin Microbiol* **50**: 696-701.

Gomez-Alvarez, V., and Revetta, R.P. (2016a) Draft genome sequences of six *Mycobacterium immunogenum* strains obtained from a chloraminated drinking water distribution system simulator. *Genome Announc* **4**.

Gomez-Alvarez, V., and Revetta, R.P. (2016b) Whole-genome sequences of four strains closely related to members of the *Mycobacterium chelonae* group, isolated from biofilms in a drinking water distribution system simulator. *Genome Announc* **4**.

Gomez-Valero, L., Rusniok, C., and Buchrieser, C. (2009) *Legionella pneumophila*: population genetics, phylogeny and genomics. *Infect Genet Evol* **9**: 727-739.

Gomez-Valero, L., Rusniok, C., Jarraud, S., Vacherie, B., Rouy, Z., Barbe, V. et al. (2011) Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics* **12**: 536.

Green, E.R., and Mecsas, J. (2016) Bacterial secretion systems: an overview. *Microbiol Spectr* **4**.

Greub, G. (2009) *Parachlamydia acanthamoebae*, an emerging agent of pneumonia. *Clin Microbiol Infec* **15**: 18-28.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* **35**: W52-W57.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S. et al. (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**: 465-+.

Haferkamp, I., Schmitz-Esser, S., Wagner, M., Neigel, N., Horn, M., and Neuhaus, H.E. (2006) Tapping the nucleotide pool of the host: novel nucleotide carrier proteins of Protochlamydia amoebophila. *Mol Microbiol* **60**: 1534-1545.

Havelaar, A.H., Schets, F.M., van Silfhout, A., Jansen, W.H., Wieten, G., and van der Kooij, D. (1992) Typing of *Aeromonas* strains from patients with diarrhoea and from drinking water. *J Appl Bacteriol* **72**: 435-444.

Hilbi, H., Segal, G., and Shuman, H.A. (2001) Icm/dot-dependent upregulation of phagocytosis by *Legionella pneumophila*. *Mol Microbiol* **42**: 603-617.

Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S. et al. (2008) Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* **190**: 1401-1412.

Houben, E.N.G., Korotkov, K.V., and Bitter, W. (2014) Take five - Type VII secretion systems of *Mycobacteria*. *BBA-Mol Cell Res* **1843**: 1707-1716.

Hsu, S.C., Martin, R., and Wentworth, B.B. (1984) Isolation of *Legionella* species from drinking water. *Appl Environ Microbiol* **48**: 830-832.

Huang, K.L., Zhang, X.X., Shi, P., Wu, B., and Ren, H.Q. (2014) A comprehensive insight into bacterial virulence in drinking water using 454 pyrosequencing and Illumina high-throughput sequencing. *Ecotoxicol Environ Saf* **109**: 15-21.

Hwang, C.C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2012) Evaluation of Methods for the Extraction of DNA from Drinking Water Distribution System Biofilms. *Microbes Environ* **27**: 9-18.

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

Kao, P.M., Hsu, B.M., Hsu, T.K., Ji, W.T., Huang, P.H., Hsueh, C.J. et al. (2014) Application of TaqMan fluorescent probe-based quantitative real-time PCR assay for the environmental survey of *Legionella* spp. and *Legionella pneumophila* in drinking water reservoirs in Taiwan. *Sci Total Environ* **490**: 416-421.

Ko, A.I., Goarant, C., and Picardeau, M. (2009) *Leptospira*: the dawn of the molecular genetics era for an emerging zoonotic pathogen. *Nat Rev Microbiol* **7**: 736-747.

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

La Scola, B., Audic, S., Robert, C., Jungang, L., de Lamballerie, X., Drancourt, M. et al. (2003) A giant virus in amoebae. *Science* **299**: 2033-2033.

La Scola, B., Desnues, C., Pagnier, I., Robert, C., Barrassi, L., Fournous, G. et al. (2008) The virophage as a unique parasite of the giant mimivirus. *Nature* **455**: 100-U165.

Lagkouvardos, I., Weinmaier, T., Lauro, F.M., Cavicchioli, R., Rattei, T., and Horn, M. (2014) Integrating metagenomic and amplicon databases to resolve the phylogenetic and ecological diversity of the Chlamydiae. *ISME J* **8**: 115-125.

Lalande, V., Barbut, F., Varnerot, A., Febvre, M., Nesa, D., Wadel, S. et al. (2001) Pseudo-outbreak of *Mycobacterium gordonae* associated with water from refrigerated fountains. *J Hosp Infect* **48**: 76-79.

Lamoth, F., Jaton, K., Vaudaux, B., and Greub, G. (2011) Parachlamydia and Rhabdochlamydia: emerging agents of community-acquired respiratory infections in children. *Clin Infect Dis* **53**: 500-501.

Lehtola, M.J., Miettinen, K.T., Keinanen, M.M., Kekki, T.K., Laine, O., Hirvonen, A. et al. (2004) Microbiology, chemistry and biofilm development in a pilot drinking water distribution system with copper and plastic pipes. *Water Res* **38**: 3769-3779.

Letunic, I., and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**: W242-W245.

Li, D.H., Liu, C.M., Luo, R.B., Sadakane, K., and Lam, T.W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**: 1674-1676.

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.

Li, L.L., Bannantine, J.P., Zhang, Q., Amonsin, A., May, B.J., Alt, D. et al. (2005) The complete genome sequence of *Mycobacterium avium* subspecies paratuberculosis. *P Natl Acad Sci USA* **102**: 12344-12349.

Ling, F.Q., Hwang, C.A., LeChevallier, M.W., Andersen, G.L., and Liu, W.T. (2016) Core-satellite populations and seasonality of water meter biofilms in a metropolitan drinking water distribution system. *ISME J* **10**: 582-595.

Liu, H., Yuan, X.C., Xu, J., Harrison, P.J., He, L., and Yin, K.D. (2015) Effects of viruses on bacterial functions under contrasting nutritional conditions for four species of bacteria isolated from Hong Kong waters. *Sci Rep* **5**.

Lowry, P.W., Becksague, C.M., Bland, L.A., Aguero, S.M., Arduino, M.J., Minuth, A.N. et al. (1990) *Mycobacterium chelonae* infection among patients receiving high-flux dialysis in a hemodialysis clinic in California. *J Infect Dis* **161**: 85-90.

Lück, C., Brzuszkiewicz, E., Rydzewski, K., Koshkolda, T., Sarnow, K., Essig, A., and Heuner, K. (2015) Subtyping of the Legionella pneumophila "Ulm" outbreak strain using the CRISPR–Cas system. *Int J Med Microbiol* **305**: 828-837.

Lye, D.J., and Dufour, A.P. (1993) Virulence characteristics of heterotrophic bacteria commonly isolated from potable water. *Environ Toxic Water* **8**: 13-23.

Lye, D.J., Rodgers, M.R., Stelma, G., Vesper, S.J., and Hayes, S.L. (2007) Characterization of *Aeromonas* virulence using an immunocompromised mouse model. *Curr Microbiol* **54**: 195-198.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J. et al. (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**: 722-736.

Maranger, R., and Bird, D.F. (1995) Viral Abundance in Aquatic Systems - a Comparison between Marine and Fresh-Waters. *Mar Ecol Prog Ser* **121**: 217-226.

Miller, C.S. (2013) Assembling full-length rRNA genes from short-read metagenomic sequence datasets using EMIRGE. *Method Enzymol* **531**: 333-352.

Mojica, F.J.M., Diez-Villasenor, C., Garcia-Martinez, J., and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**: 733-740.

More, R.P., Mitra, S., Raju, S.C., Kapley, A., and Purohit, H.J. (2014) Mining and assessment of catabolic pathways in the metagenome of a common effluent treatment plant to induce the degradative capacity of biomass. *Bioresource Technol* **153**: 137-146.

Nagarajan, N., Cook, C., Di Bonaventura, M., Ge, H., Richards, A., Bishop-Lilly, K.A. et al. (2010) Finishing genomes with limited resources: lessons from an ensemble of microbial genomes. *BMC Genomics* **11**.

Newton, H.J., Ang, D.K.Y., van Driel, I.R., and Hartland, E.L. (2010) Molecular pathogenesis of infections caused by *Legionella pneumophila*. *Clin Microbiol Rev* **23**: 274-298.

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

Philippe, N. (2013) Pandoraviruses: Amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes (vol 341, pg 281, 2013). *Science* **341**: 1452-1452.

Philippe, N., Legendre, M., Doutre, G., Coute, Y., Poirot, O., Lescot, M. et al. (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science* **341**: 281-286.

Picardeau, M. (2017) Virulence of the zoonotic agent of leptospirosis: still terra incognita? *Nat Rev Microbiol* **15**: 297-307.

Pop, M. (2009) Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics* **10**: 354-366.

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiol-Sgm* **151**: 653-663.

Rao, C.T., Guyard, C., Pelaz, C., Wasserscheid, J., Bondy-Denomy, J., Dewar, K., and Ensminger, A.W. (2016) Active and adaptive Legionella CRISPR-Cas reveals a recurrent challenge to the pathogen. *Cell Microbiol* **18**: 1319-1338.

Ratzow, S., Gaia, V., Helbig, J.H., Fry, N.K., and Luck, P.C. (2007) Addition of *neuA*, the gene encoding N-acylneuraminate cytidylyl transferase, increases the discriminatory ability of the consensus sequence-based scheme for typing *Legionella pneumophila* serogroup 1 strains. *J Clin Microbiol* **45**: 1965-1968.

Ren, S.X., Gang, F., Jiang, X.G., Zeng, R., Miao, Y.G., Xu, H. et al. (2003) Unique physiological and pathogenic features of *Leptospira interrogans* revealed by whole-genome sequencing. *Nature* **422**: 888-893.

Rodriguez-Martinez, S., Sharaby, Y., Pecellin, M., Brettar, I., Hofle, M., and Halpern, M. (2015) Spatial distribution of *Legionella pneumophila* MLVA-genotypes in a drinking water system. *Water Res* **77**: 119-132.

Sampson, S.L. (2011) Mycobacterial PE/PPE proteins at the host-pathogen interface. *Clin Dev Immunol* **2011**.

Sangwan, N., Xia, F.F., and Gilbert, J.A. (2016) Recovering complete and draft population genomes from metagenome datasets. *Microbiome* **4**.

Schmidt, T.M., and Schaechter, M. (2012) *Topics in Ecological and Environmental Microbiology*: Academic Press.

Schmitz-Esser, S., Linka, N., Collingro, A., Beier, C.L., Neuhaus, H.E., Wagner, M., and Horn, M. (2004) ATP/ADP translocases: a common feature of obligate intracellular amoebal symbionts related to chlamydiae and rickettsiae. *J Bacteriol* **186**: 683-691.

Schroeder, G.N., Petty, N.K., Mousnier, A., Harding, C.R., Vogrin, A.J., Wee, B. et al. (2010) *Legionella pneumophila* strain 130b possesses a unique combination of Type IV secretion systems and novel Dot/Icm secretion system effector proteins. *J Bacteriol* **192**: 6001-6016.

Segata, N., Bornigen, D., Morgan, X.C., and Huttenhower, C. (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* **4**: 2304.

Shariat, N., and Dudley, E.G. (2014) CRISPRs: Molecular Signatures Used for Pathogen Subtyping. *Appl Environ Microbiol* **80**: 430-439.

Soderberg, M.A., Dao, J., Starkenburg, S.R., and Cianciotto, N.P. (2008) Importance of type II secretion for survival of *Legionella pneumophila* in tap water and in amoebae at low temperatures. *Appl Environ Microbiol* **74**: 5583-5588.

Stelma, G.N., Lye, D.J., Smith, B.G., Messer, J.W., and Payment, P. (2004) Rare occurrence of heterotrophic bacteria with pathogenic potential in potable water. *Int J Food Microbiol* **92**: 249-254.

Sun, C.L., Thomas, B.C., Barrangou, R., and Banfield, J.F. (2016) Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J* **10**: 858-870.

Tamaki, H., Zhang, R., Angly, F.E., Nakamura, S., Hong, P.Y., Yasunaga, T. et al. (2012) Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. *Environ Microbiol* **14**: 441-452.

Tennessen, K., Andersen, E., Clingenpeel, S., Rinke, C., Lundberg, D.S., Han, J. et al. (2016) ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J* **10**: 269-272.

Thomas, V., Loret, J.F., Jousset, M., and Greub, G. (2008) Biodiversity of amoebae and amoebae-resisting bacteria in a drinking water treatment plant. *Environ Microbiol* **10**: 2728-2745.

Tilney, L.G., Harb, O.S., Connelly, P.S., Robinson, C.G., and Roy, C.R. (2001) How the parasitic bacterium *Legionella pneumophila* modifies its phagosome and transforms it into rough ER: implications for conversion of plasma membrane to the ER membrane. *J Cell Sci* **114**: 4637-4650.

Tortora, G., Funke, B., and Case, C. (2013) *Microbiology: an Introduction (11 th)*. Yorkshire: Pearson.

Vaerewijck, M.J.M., Huys, G., Palomino, J.C., Swings, J., and Portaels, F. (2005) Mycobacteria in drinking water distribution systems: ecology and significance for human health. *FEMS Microbiol Rev* **29**: 911-934.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Voth, D.E., Broederdorf, L.J., and Graham, J.G. (2012) Bacterial Type IV secretion systems: versatile virulence machines. *Future Microbiol* **7**: 241-257.

Wadowsky, R.M., Butler, L.J., Cook, M.K., Verma, S.M., Paul, M.A., Fields, B.S. et al. (1988) Growth-Supporting Activity for Legionella-Pneumophila in Tap Water Cultures and Implication of Hartmannellid Amebas as Growth-Factors. *Appl Environ Microbiol* **54**: 2677-2682.

Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L. et al. (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* **42**: D581-D591.

Wilson, J.W., Schurr, M.J., LeBlanc, C.L., Ramamurthy, R., Buchanan, K.L., and Nickerson, C.A. (2002) Mechanisms of bacterial pathogenicity. *Postgrad Med J* **78**: 216-224.

Wilson, R.W., Steingrube, V.A., Bottger, E.C., Springer, B., Brown-Elliott, B.A., Vincent, V. et al. (2001) *Mycobacterium immunogenum* sp nov., a novel species related to *Mycobacterium abscessus* and associated with clinical disease, pseudo-outbreaks and contaminated metalworking fluids: an international cooperative study on mycobacterial taxonomy. *Int J Syst Evol Micr* **51**: 1751-1764.

Wu, H.J., Wang, A.H.J., and Jennings, M.P. (2008) Discovery of virulence factors of pathogenic bacteria. *Current Opinion in Chemical Biology* **12**: 93-101.

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

Yu, V.L., Plouffe, J.F., Pastoris, M.C., Stout, J.E., Schousboe, M., Widmer, A. et al. (2002) Distribution of *Legionella* species and serogroups isolated by culture in patients with sporadic community-acquired legionellosis: An international collaborative survey. *J Infect Dis* **186**: 127-128.

Zankari, E., Hasman, H., Cosentino, S., Vestergaard, M., Rasmussen, S., Lund, O. et al. (2012) Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemoth* **67**: 2640-2644.

Zhang, Y., Oh, S., and Liu, W.-T. (2017) Impact of drinking water treatment and distribution on the microbiome continuum: an ecological disturbance's perspective. *Environ Microbiol*: in press.

Zhou, Y., Liang, Y.J., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011) PHAST: A Fast Phage Search Tool. *Nucleic Acids Res* **39**: W347-W352.

CHAPTER 5 METAGENOMICS REVEALED OVERLOOKED METHANE METABOLISM AND MECROBIAL INTERDEPENDENCY IN GROUNDWATER SYSTEMS

5.1 Abstract

Dissolved methane is a common trace constituent in ground waters and can be a major carbon source in ground water-sourced drinking water systems (GWDWSs). However, its microbial use in GWDWSs is frequently overlooked, and the microbial interactions involving methane-utilizing microbes is seldomly investigated. This study used genome-resolved metagenomics to investigate the compositional and functional diversity, interspecies relationship, and metabolic interdependency of methylotrophic bacteria in a GWDWS. Among microbial biomass taken from water and biofilm phases at different stages of the GWDWS, 34 methylotroph-related draft genomes were recovered together with another 133 draft genomes belonging to a variety of taxa. Both Type I and Type II methanotrophs and nonmethanotrophic methylotrophs (NMMs) were abundant in particular in finished water, and water and biofilm samples taken in the distribution system. Among methano-/methylo-trophs, methylotrophy pathways involving multiple enzymes were more dominant than those with single enzyme systems, possibly due to high metabolite exchange potential in the multigene/multi-enzyme systems. Network analysis could identify potential species interaction between methanotrophs and a number of NMMs and heterotrophs. Examining the metabolic interdependency involving methylotrophs through the topology of reconstructed metabolic networks further suggested that the microbial community had the potential to exchange metabolites extensively, and NMMs and other heterotrophs had the capability to supply essential metabolites to methanotrophs. The genomic-based findings provided overlooked microbial functions, interactions, and methane metabolism in GWDWSs.

5.2 Introduction

Methane is one of the major one-carbon compounds in the environment and the second most prevalent greenhouse gas in the US from human activities (Bousquet et al., 2006). It can be derived from biological processes in shallow anaerobic groundwater environments or thermal decomposition of organic matter in deep coal, oil and gas fields (Barker and Fritz, 1981). The use of methane by microorganisms is an important part of the global carbon cycle. However, this process is frequently overlooked in groundwater ecosystems and groundwater-sourced drinking water systems (GWDWSs).

Microbial methane utilization in aerobic environments is characterized by tight linkages between methane-utilizing (methanotrophic) and non-methanotrophic methylotrophs (NMMs) (Chistoserdova, 2011; Krause et al., 2017). Previous studies with microcosms from Lake Washington sediments indicated that methanotrophs can support a community that cannot use methane directly through cross-feeding with methane-derived carbon (e.g., methanol) (Kalyuzhnaya et al., 2008; Beck et al., 2013). Cross-feeding is a ubiquitous feature of microbial communities, influencing major biogeochemical processes globally (Schink, 1997; Martienssen and Schops, 1999). This type of microbial interaction reflects a high degree of metabolic interdependency in microbial communities (Anantharaman et al., 2016). Nevertheless, few studies have investigated species interaction through metabolic interdependency in natural microbial communities centered by methano-/methylo-trophs (Kolenbrander, 2011; Levy and Borenstein, 2013).

Methano-/methylotrophy can be carried out by various microbes, which are classified in different ways (Kalyuzhnaya et al., 2006; Chistoserdova, 2011; Chistoserdova and Lidstrom, 2013). Based on the ability of methane utilization, methylotrophs are divided into methanotrophs and NMMs. Methylotrophs that can use both one-carbon and multicarbon compounds as substrates are called facultative methylotrophs, whereas methylotrophs that only use one-carbon substrates are defined as obligate methylotrophs. From the phylogenetic perspective, methylotrophs are widespread within *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Verrucomicrobia*, and

Candidatus phylum NC10. Among them, methanotrophs are distributed mainly within the alpha- and gamma- subdivisions of *Proteobacteria* (*Methylocystaceae, Beijerinckiaceae, Methylococcaceae, Methylothermaceae*), *Verrucomicrobia* (*Methylacidiphilaceae*), and NC10. Type I methanotrophs refer to those within *Methylococcaceae and Methylothermaceae,* and Type II to those within *Methylocystaceae* and *Beijerinckiaceae*. Type I and Type II can be differentiated by locations of intracytoplasmic membranes. Representative NMMs are from *Methylophilaceae, Hyphomicrobium, and Methylobacteriaceae*.

Engineered GWDWSs provide a unique system for the study of microbial communities centered by methano-/methylo-trophs. The low amount of organic carbon input into the system results in a less complex ecosystem than freshwater lakes or other ecosystems (Oshkin et al., 2015). In the Champaign-Urbana area (Illinois, USA), Knirk (Knirk, 1908) and Buswell and Larson (Buswell and Larson, 1937) reported the abundance of methane in numerous regions of the Mahomet Aquifer. Presence of methanotrophs in the system was first reported in 1972 by culturing slime accumulations at the air-water interface from Sullivan, Illinois and the Champaign-Urbana water distribution system (Gunsalus et al., 1972). This study identified a methanotroph that produced an extensive capsule and exhibited taxonomic properties similar to *Methylomonas methanica*. At the same time, two NMMs and six other heterotrophs were also isolated from the same consortium. With cross-feeding experiments involving mixed cultures of these isolates, the authors concluded that dissolved methane in ground waters was a previously unappreciated energy source for the development of methanotrophs, NMMs, and a diverse heterotrophic community in the drinking water distribution system.

Our recent study using 16S amplicon sequencing method indicated a significant increase in the relative abundance of populations related to methanotrophs and NMMs in finished water and the distribution system compared with raw water and water from the treatment process in the Champaign-Urbana drinking water system (Zhang et al., 2017a). Methanotrophs and NMMs were abundant in finished water, bulk water in the distribution system, and water meter biofilms of buildings, and accounted for a large portion of the core microbiome that shared between the two phases (Hwang et al., 2012; Ling et al., 2016).

In this study, we used an approach based on genome-resolved metagenomics, that yielded good-quality draft microbial genomes without cultivation to provide insights into the methane metabolism in the Champaign-Urbana GWDWS (Alneberg et al., 2014; Eren et al., 2015; Hug et al., 2016). Draft genomes affiliated with methanotrophs, NMMs, and heterotrophs were successfully recovered from metagenomes obtained from different stages of the GWDWS system. We investigated their compositional and functional diversity, interspecies relationship, and metabolic interdependency through phylogenomic, functional, and metabolic network topology analyses. The findings provide insights into overlooked microbial functions, interactions, and methane metabolism in GWDWSs.

5.3 Materials and methods

Sampling and sample processing Microbial biomass from different stages of the treatment processes, different locations in the distribution system, and premise plumbing reactors was collected as described in previous studies (Zhang et al., 2017b; Zhang et al., 2017a). Water-phase samples included raw water (RW), immediately before filtration and chlorination (after lime treatment and recarbonation) (BC), finished water (FW) prior to distribution, and three taps (cold water) (DS1-DS3). Biofilm samples were taken from two retired water mains (PB1-PB2), water meters (WM), and premise plumbing (PR) pipes.

DNA extraction and Illumina sequencing Genomic DNA (gDNA) was extracted using FastDNA® SPIN Kit for Soil (MP Biomedicals, Carlsbad, CA, USA) from the membranes with cells. Microbial communities in these samples were analyzed using amplified 16S rRNA genes and metagenomics. Detailed description of sampling and sequencing procedures can be found in previous studies (Zhang et al., 2017b; Zhang et al., 2017a) . Briefly, 16S rRNA gene amplicon analysis was carried out using a universal primer set targeting the V4-V5 hypervariable regions of both the Bacteria and Archaea domains (Kozich et al., 2013). Paired-end sequencing of the amplicons (2x300 bp) was done with an Illumina MiSeq platform (Illumina, Inc., San Diego, CA, USA). DNA libraries for metagenomic sequencing were paired-end sequenced on

Illumina HiSeq2500 platforms using TruSeq SBS kit v4 (for the RW, BC, FW, and DS1, 2, 3 samples) and TruSeq Rapid SBS kit v2 (for the PB1, PB2, WM, and PR samples) (Illumina, Inc., San Diego, CA, USA). All sequencing was performed by the Roy J. Carver Biotechnology Center at the University of Illinois at Urbana-Champaign.

Sequence analysis The obtained paired-end 16S rRNA gene sequences were aligned with Mothur using the default setting, which required a quality score of over 25 if a gap and a base occurred at the same position or one of the bases had a quality score six or more points better than the other if the two reads disagreed (Kozich et al., 2013). The resulting sequences were screened for chimeras with the UCHIME algorithm implemented in USEARCH 6.1 and processed using the *de novo* OTU picking workflow in QIIME (Caporaso et al., 2010b). Representative sequences from OTUs were aligned using PyNAST (Caporaso et al., 2010a) and inserted into the phylogenetic trees, Greengenes_16S_2011.arb, with the parsimony insertion tool in the ARB program (Ludwig et al., 2004; McDonald et al., 2012).

Draft genome reconstruction All the metagenomic datasets were trimmed using SolexaQA2 based on a cutoff of 20 by phred scores (Cox et al., 2010) and assembled using Megahit (Li et al., 2015). More details of the assemblies could be found in our previous study (Zhang et al., 2017a). The high-quality contigs (approximately 2.0 ×10^8 bp for each metagenome) obtained at this step were binned based on metagenomics read coverage, tetranucleotide frequency, and the occurrence of unique marker genes using both MaxBin 2.0 (Wu et al., 2016) and MetaBAT (Kang et al., 2015) to maximize binning quality. The resulting bins from the two binning tools were compared and assessed with CheckM (Parks et al., 2015) and ProDeGe (Tennessen et al., 2016), followed by manual curation. The curated bins with ≥ 70% completeness and ≥ 6-fold coverage were finalized as draft genomes. Percentage of reads mapped to contigs was estimated by the Burrow-Wheeler Aligner-MEM (BWA-MEM) (Li and Durbin, 2010). EMIRGE was used to reconstruct nearly full-length SSU genes in metagenomes (Miller, 2013).

Phylogenomic tree construction PhyloPhlAn (Segata et al., 2013) was used to construct phylogenomic trees based on draft genomes and reference genomes. The constructed trees were visualized using iTOL (Letunic and Bork, 2016).

Identification of methylotrophy metabolic modules Methylotrophy metabolic modules were defined according to two previous studies (Chistoserdova, 2011; Beck et al., 2015). These features were retrieved from publicly-available genomes of methano-/methyl-trophs and used as a reference database. Protein-coding genes of the retrieved draft genomes were predicted using Prodigal (Hyatt et al., 2010) and then blasted against this reference database with the following criteria: percentage of identical matches ≥ 40%, query coverage per subject ≥ 70%, and query sequence length ≥ 100.

Co-occurrence analysis CoNet was used to infer co-occurrence patterns of OTUs (Faust and Raes, 2016). OTUs with an average relative abundance of ≥ 0.01% were included in this test. The association in terms of relative abundance between any two OTUs was determined based on the Pearson and Spearman correlations using Bray-Curtis distance and merged using intersections of edges. The significance of the associations was confirmed with permutation test and the ReBoot method (Faust and Raes, 2012).

Reconstruction of genome-scale metabolic models The ModelSEED pipeline (Henry et al., 2010) was used to reconstruct genome-scale metabolic models for the recovered genomes from metagenomes. Briefly, these models were used to simulate growth and metabolite production capabilities of each strain according to the constraints on nutritional availability from the environment. Manual curation was carried out to reduce the artifacts of the automated model reconstruction process by improving reaction directionality and nutrient transport. The resulting metabolic networks was organized with nodes representing compounds and edges representing reactions that link substrates to products.

Microbial cooperative potential from metabolic network topology A previously developed and validated metric for species interaction were used for determining microbe-microbe cooperative potential (Borenstein and Feldman, 2009; Kolenbrander, 2011; Levy and Borenstein, 2013). The Biosynthetic Support Score (BSS) quantified the extent to which the nutritional requirements of one species could be satisfied by the biosynthetic capacity of another. This index was determined by the shared fraction of the seed set between a pair of metabolic networks. The seed set of a metabolic network represented the minimal subset of the nodes required to access every node in the network, i.e., the minimal subset of exogenously acquired compounds in the network whose existence permitted the production of all other compounds in the network. The seed detection was carried out using the algorithm implemented in NetSeed (Carr and Borenstein, 2012).

Genomic data depositing Sequences of the 16S rRNA gene amplicons and the reconstructed draft genomes were deposited in GenBank under the BioProject PRJNA323575.

5.4 Results

Methano-/methyl-trophs at different stages/phases of the studied GWDWS As the dissolved methane present in RW could travel together with the water flow, it could support the growth of various methanotrophs and accompanying species downstream. Figure 5.1 shows that the relative abundance of methano-/methylotroph populations was low in SW and BC but increased in FW and the distribution system (Mann-Whitney U statistical test (p = 0.05) using the sum of the abundance of all methano-/methylo-trophic OTUs in each sampling event). FW had the highest number of methylotroph OTUs among the six sampling sites, suggesting that some methylotrophs could have survived through the disinfection treatment. The dominant methanotrophy groups included *Methylococcaceae* in *Gammaproteobacteria* (Type I) and *Methylocystaceae* in *Alphaproteobacteria* (Type II). OTU-FW-2 (42.1% with metagenomes in FW and 18.0% with 16S rRNA gene amplicon analysis) was closely-related to *Methylomonas methanica*, an obligate methane-oxidizer.

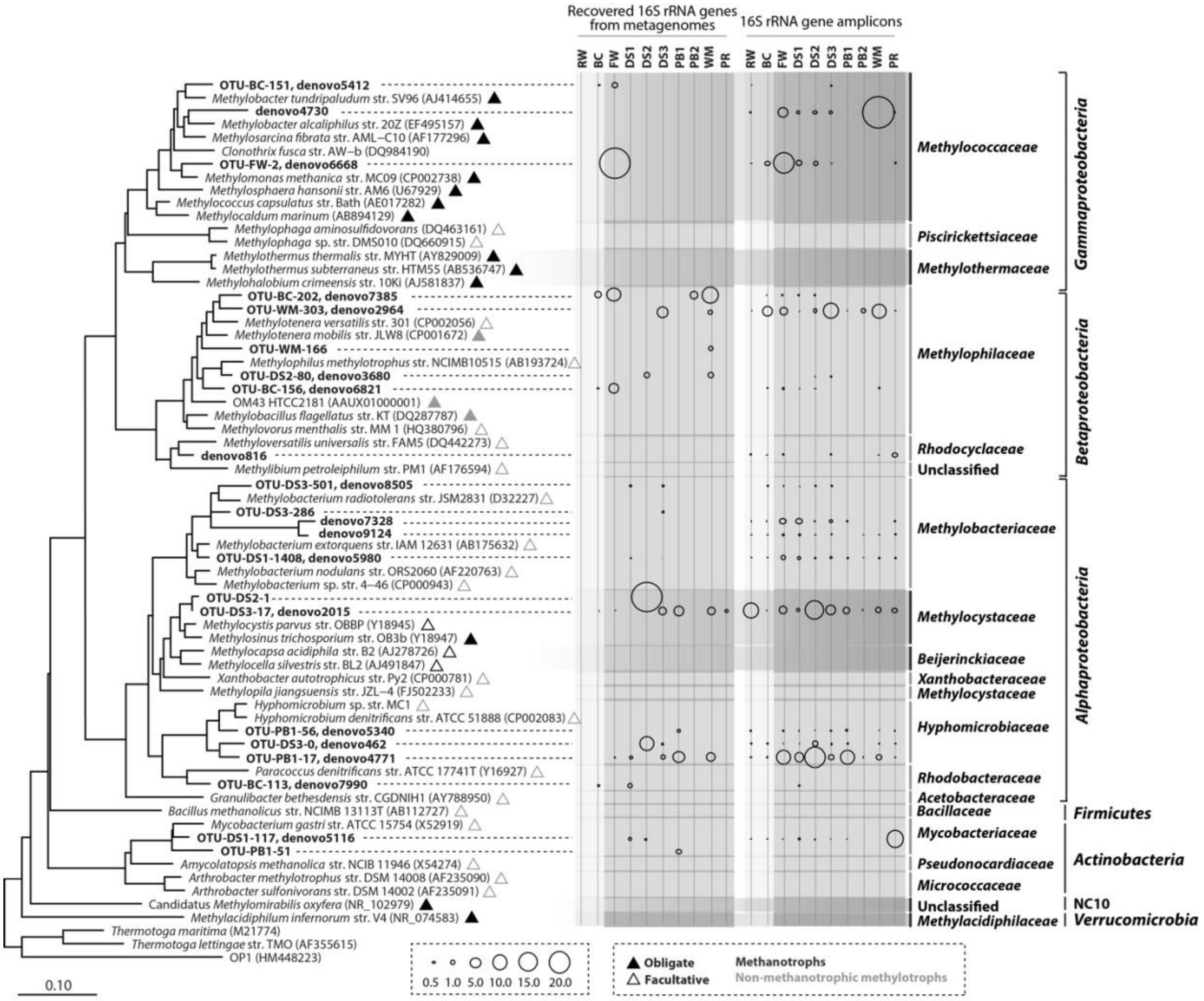


Figure 5.1 Neighbor-joining tree of methylotrophs based on 16S rRNA gene sequences. The abundance of methylotrophic OTUs detected in the GWDWS were shown on the right, including results both from metagenomes and from amplicon analysis. Methanotrophs were marked with darker grey shades and methylotophs with light grey shades. OTU names starting with "OTU" were OTUs retrieved from metagenomic data (>1200 bp) and those starting with "denovo" were dominant OTUs from 16S rRNA gene amplicon analysis (average read length: 375 bp). The triangles at the end of OTU names indicate whether they are obligate or facultative methano-/methyl-trophs.

The other dominant OTU-DS3-17, closely-related to *Methylocystis parvus*, was a facultative methanotroph. The two dominant NMM-related populations were *Methylotenera*-like OTUs (OTU-BC-202 and OTU-WM-303) from *Methylophilaceae* in *Betaproteobacteria* and *Hyphomicrobium*-like OTUs (OTU-DS3-0 and OTU-PB1-17) from *Hyphomicrobiaceae* in *Alphaproteobacteria*. Both groups were abundant and prevalent at different stages of the studied system. Comparing the results from metagenomes with 16S rRNA gene amplicon analysis, we found some discrepancies in abundance between the two methods. This difference was more substantial for denovo4730, OTU-BC-202, and OTU-DS2-1, likely due to biases associated with PCR-based methods (Duhaime et al., 2012; Brooks et al., 2015). Collectively, the GWDWS supported a drinking water microbiota dominated by both Type I and Type II methanotrophs and NMMs.

Phylogenomic diversity of detected methano-/methylo-trophs To further characterize methano-/methyl-trophs in the studied GWDWS, 34 medium-quality (≥7x coverage and >68% completeness) draft genomes were recovered from the ten metagenomes (Figure 5.2). These recovered genomes represented all the major families identified through 16S rRNA gene analysis (*Mycobacteriaceae*, *Hyphomicrobiaceae*, *Methylobacteriaceae*, *Methylocystaceae*, *Methylococcaceae*, unclassified *Betaproteobacteria*, and *Methylophilaceae*). Among them, the most abundant family, *Methylococcaceae* could be identified in all the water phase samples and water meter samples. Within this family, 12 draft genomes obtained were related to *Methylobacter* and *Methylomonas*. No methylotrophs could be recovered from SW owing to its anaerobic environment. The BC sample contained all the major groups of methylotrophs, except *Hyphomicrobiaceae* and *Mycobacteriaceae*. This suggested that the softening stage could substantially influence the composition of downstream microbiome.

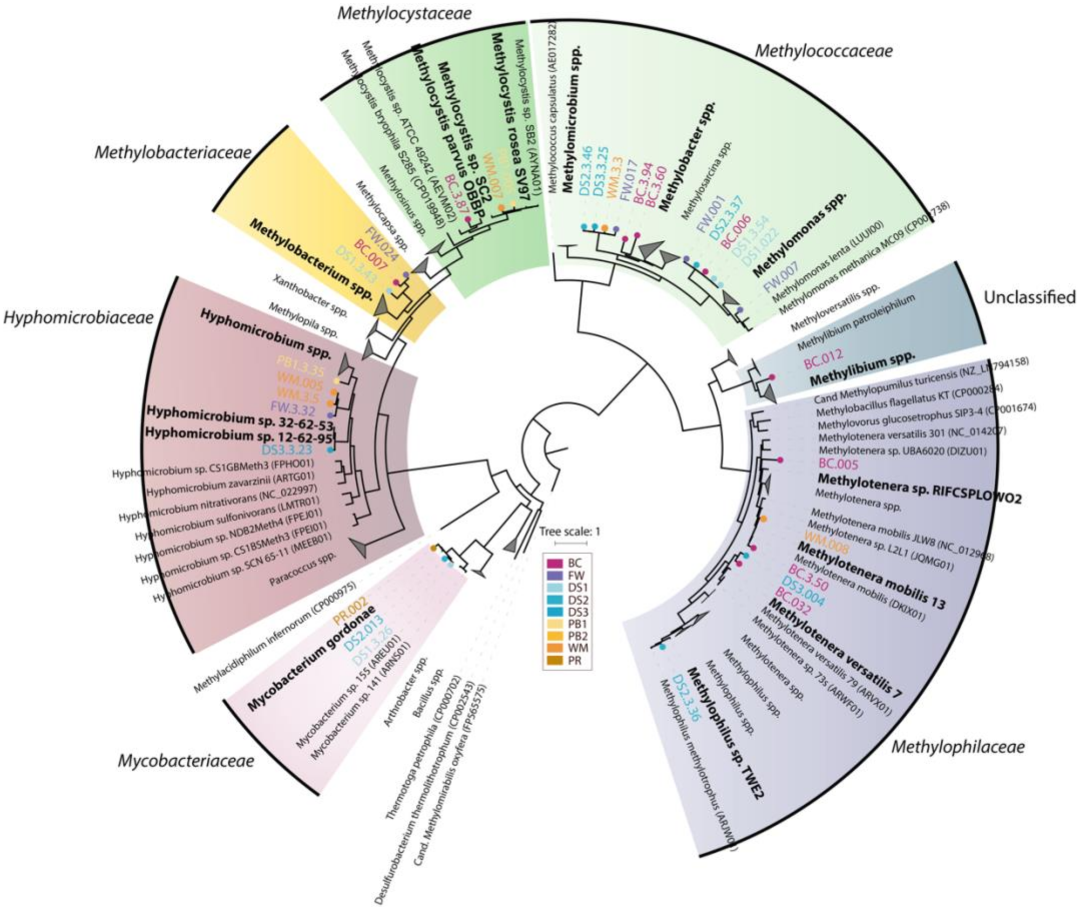
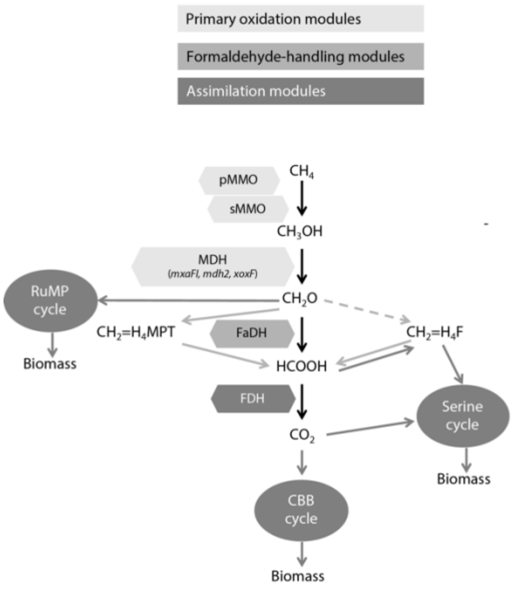

Methylocystaceae
Methylocystis sp. SB2 (AYNA01)
Methylocystis rosea SV97
WM.007
Methylocystis sp. SC2
Methylocystis parvus OBBP
BC.3.87
Methylocystis sp. ATCC 49242 (AEVM02)
Methylocystis bryophila S285 (CP019948)
Methylosinus spp.
Methylocapsa spp.
Methylobacteriaceae
FW.024
BC.007
DS1.3.43
Methylobacterium spp.
Xanthobacter spp.
Methylopila spp.
Hyphomicrobiaceae
Hyphomicrobium spp.
PB1.3.35
WM.005
WM.3.5
FW.3.32
Hyphomicrobium sp. 32-62-53
Hyphomicrobium sp. 12-62-95
DS3.3.23
Hyphomicrobium sp. CS1GBMeth3 (FPHO01)
Hyphomicrobium zavarzinii (ARTG01)
Hyphomicrobium nitrativorans (NC_022997)
Hyphomicrobium sulfonivorans (LMTR01)
Hyphomicrobium sp. NDB2Meth4 (FPEJ01)
Hyphomicrobium sp. CS1BSMeth3 (FPEI01)
Hyphomicrobium sp. SCN 65-11 (MEEB01)
Paracoccus spp.
Methylacidiphilum infernorum (CP000975)
PR.002
Mycobacterium gordonae
DS2.013
DS1.3.26
Mycobacterium sp. 155 (AREU01)
Mycobacterium sp. 141 (ARNS01)
Mycobacteriaceae
Arthrobacter spp.
Bacillus spp.
Thermotoga petrophila (CP000702)
Desulfurobacterium thermolithotrophum (CP002543)
Cand. Methylomirabilis oxyfera (FP565575)
Methylococcus capsulatus (AE017282)
Methylomicrobium spp.
DS2.3.46
DS3.3.25
WM.3.3
FW.017
BC.3.94
BC.3.60
Methylobacter spp.
Methylosarcina spp.
FW.001
DS2.3.37
BC.006
DS1.3.54
DS1.022
Methylomonas spp.
FW.007
Methylomonas lenta (LUUI00)
Methylomonas methanica MC09 (CP002738)
Methylococcaceae
Methyloversatilis spp.
Methylibium patroleiphilum
BC.012
Methylibium spp.
Unclassified
Cand Methylopumilus turicensis (NZ_LN794158)
Methylobacillus flagellatus KT (CP000284)
Methylovorus glucosetrophus SIP3-4 (CP001674)
Methylotenera versatilis 301 (NC_014207)
Methylotenera sp. UBA6020 (DIZU01)
BC.005
Methylotenera sp. RIFCSPLOWO2
Methylotenera spp.
Methylotenera mobilis JLW8 (NC_012968)
Methylotenera sp. L2L1 (JQMG01)
WM.008
Methylotenera mobilis 13
Methylotenera mobilis (DKIX01)
BC.3.50
DS3.004
BC.032
Methylotenera versatilis 7
Methylotenera versatilis 79 (ARVX01)
Methylotenera sp. 73s (ARWF01)
Methylotenera spp.
Methylophilus spp.
Methylophilus spp.
Methylophilus sp. TWE2
DS2.3.36
Methylophilus methylotrophus (ARJW01)
Methylophilaceae
Tree scale: 1
BC
FW
DS1
DS2
DS3
PB1
PB2
WM
PR

Figure 5.2 Phylogenomic tree of methano-/methylo-trophs based on 400 conserved genes. Methano-/methylo-trophs mainly belonged to *Proteobacteria* (Alpha-, Beta-, and Gamma-subdivisions) and *Actinobacteria*. They were further grouped at the family level with sectors of colorful shades. Genomes recovered from different sampling sites (34 genomes in total) were marked with different colors.



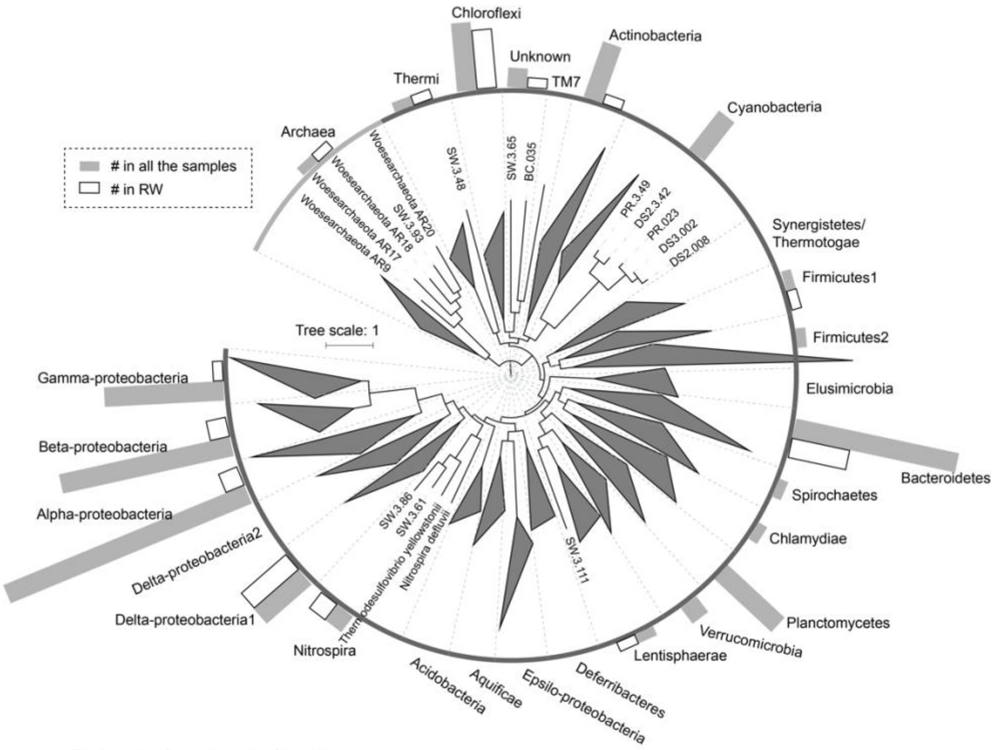
	MMO	MDH	XoxF	MADH	NMGP	H4MPT	GSH	FolD	MtdA/ Fch	PurU	FtfL	Serine cycle	RuMP	CBB cycle
Methylococcaceae														
DS2.3.46	+	+	+	-	-	+	-	-	+	-	-	+	+	-
DS3.3.25	+	+	+	-	-	+	-	-	+	-	-	+	+	-
WM.3.3	+	+	-	-	-	+	-	-	+	-	-	-	+	-
FW.017	+	+	+	-	-	+	-	-	+	-	-	+	-	-
BC.3.94	+	+	+	-	-	+	-	-	+	-	+	+	+	-
BC.3.60	-	-	-	-	-	+	-	-	+	-	-	+	+	-
FW.001	+	+	+	-	-	-	-	-	+	-	-	+	+	-
DS2.3.37	+	+	+	-	-	-	-	-	+	-	-	+	+	-
BC.006	+	+	+	-	-	-	-	-	+	-	-	+	+	-
DS1.3.54	+	+	+	-	-	-	-	-	+	-	-	+	+	-
DS1.022	+	+	+	-	-	-	-	-	+	-	-	+	+	-
FW.007	+	+	+	-	-	+	-	-	+	-	+	+	-	-
Methylocystaceae														
BC.3.87	+	+	-	-	-	+	+	+	+	-	+	+	-	-
WM.007	+	-	+	-	-	+	-	+	+	-	-	+	-	-
PB1.005	+	+	+	-	-	+	+	+	+	-	+	+	+	-
Methylophilaceae														
BC.005	-	+	+	-	-	+	-	+	+	-	-	-	+	-
WM.008	-	+	+	-	-	+	+	+	+	+	-	+	+	-
BC.3.50	-	+	+	-	-	+	-	+	+	-	-	-	+	-
DS3.004	-	-	+	-	+	+	-	+	+	+	-	-	+	-
BC.032	-	+	+	-	-	+	-	+	+	+	-	-	+	-
DS2.3.36	-	+	+	-	-	+	-	+	+	+	-	-	+	-
Hyphomicrobiaceae														
PB1.3.35	-	+	+	-	+	-	-	+	+	-	+	+	-	-
WM.005	-	-	+	-	-	-	+	+	+	-	-	+	+	-
WM.3.5	-	-	+	-	-	-	-	+	+	-	-	+	+	-
FW.3.32	-	-	+	-	-	-	-	+	+	-	+	+	+	-
DS3.3.23	-	+	+	-	+	+	-	+	+	-	+	+	-	-
Methylobacteriaceae														
BC.007	-	-	+	-	-	-	-	-	+	+	+	+	-	-
FW.024	-	-	-	-	-	-	-	+	+	+	+	+	-	-
DS1.3.43	-	-	-	-	-	-	+	+	+	+	+	+	+	-
Mycobacteriaceae														
PR.002	-	-	-	-	-	-	+	-	-	+	-	+	-	+
DS2.013	-	-	-	-	-	-	+	+	+	+	-	+	-	+
DS1.3.26	-	-	-	-	-	-	+	+	+	+	-	+	-	+
Unclassified														
BC.012	-	-	-	-	-	-	-	+	+	+	+	-	-	+

Figure 5.3 Comparative analysis of the methylotrophy functional modules in the recovered genomes. The lightest color indicates primary oxidation modules, more darker color for formaldehyde-handling modules, and the darkest color for assimilation modules. MMO, methane monooxygenase; MDH, methanol dehydrogenase (MxaF1 or Mdh2); MADH, methylamine dehydrogenase; NMGP, *N*-methylglutamate pathway; H_4MPT, H_4MPT-linked pathway for formaldehyde oxidation; GSH, glutathione (GSH)-dependent formaldehyde oxidation pathway; FolD, a bifunctional enzyme possessing methylene-H_4F dehydrogenase and methenyl-H_4F cyclohydrolase activities; MtdA/Fch, methylene-H_4F dehydrogenase/methenyl-H_4F cyclohydrolase; PurU, formyl-H4F hydrolase; FtfL, formyl-H_4F ligase; RuMP, formaldehyde assimilation via ribulose monophosphate cycle.

Methylotrophy functional modules Methylotrophy functions are modular in nature as different combinations of enzymatic systems and pathways were used by various methylotrophs. These modules could be divided into three categories, primary oxidation of one-carbon compounds, formaldehyde oxidation and detoxification, and carbon assimilation (Figure 5.3). Among the primary oxidation modules, methane oxidation modules within most of the genomes belonged to *Methylococcaceae* and *Methylocystaceae* and methanol oxidation modules were mainly affiliated with *Methylophilaceae* and *Hyphomicrobiaceae*. Three genomes (DS3.004, PB1.3.35, and DS3.3.23) had the capability to use methylamines through the N-methylglutamate pathway (NMGP). For formaldehyde-handling modules, methylotrophs mainly used four cofactor-linked C1 transfer pathways, which included MtdA/Fch, H_4MPT, FolD, and GSH. MtdA/Fch and H_4MPT-linked pathways, pathways involving many genes, were the most widespread pathway among the recovered genomes. FolD, a bifunctional enzyme possessing functions of MtfA and Fch, were identified in almost all the non-*Methylococcaceae* genomes. It was surprising to identify GSH-linked pathways in several genomes, which were mainly associated with Gram-positive and autotrophic methylotrophs. The formaldehyde-handling modules determined the carbon assimilation pathways to a certain extent. For example, the H_4MPT-linked pathways occurred mainly in the methylotrophs with RuMP cycle; and the MtdA/Fch-linked pathway co-occurred with the serine cycle. Only genomes belonging to *Mycobacteriaceae* and unclassified contained the CBB cycle. Together, these results suggested that pathways involving many enzymes rather than those single enzyme systems were the most popular among methanotrophs, possibly due to metabolite exchange potential in the multi-gene/multi-enzyme systems.

Diversity of other microbes In total, 133 draft genomes were recovered beside those identified as methylotrops. These draft genomes represented approximately 50% of the raw reads of the metagenomics dataset. They were affiliated with 14 phyla with the majority from *Proteobacteria* (alpha-, beta-, and gamma- subdivisions) (Figure 5.4). Many of the genomes affiliated with phyla other than *Proteobacteria* were recovered from SW, such as *Chloroflexi, Bacteroidetes,* and *Proteobacteria* (delta subdivision). The SW sample also contained many genomes that were distantly related to known genomes, such as SW.3.48, SW.3.65, SW.3.111, SW.3.86, and SW.3.61. Furthermore, an archaeum genome, SW.3.93, was recovered from SW, which was closely related to the recent described *Candidatus* Woesearchaeota str. AR20 that was reported to have a small genome (0.8Mb) and a symbiotic or parasitic lifestyle (Castelle et al., 2015).

These results suggested that SW was distinct from the remaining samples, with many genomes possibly originated from anaerobic environments.



Firmicutes1 refers to the order Clostridia
Firmicutes2 refers to the order Bacilli and Erysipelotrichi
Delta-proteobacteria1 refers to the order Desulfobacterales,
Syntrophobacterales, Desulfovibionales
Delta-proteobacteria2 refers to the order Myxococcales

Figure 5.4 Phylogenomic tree of other microbes based on conserved genes. An additional 138 draft genomes were recovered from the ten metagenomes. They belonged to various phyla, as indicated by the height of grey bars. The black box highlighted the number of genomes recovered from SW. Draft genomes (most of them were from SW) that were distantly related to the known phyla were shown with branches in the phylogenomic tree.

These recovered genomes could be classified using different criteria. From the perspective of human health, we recovered genomes closely related to pathogenic species, including *Legionella* (4 draft genomes), *Mycobacterium* (3 draft genomes), *Parachlamydia* (1), and *Leptospira* (1).

Based on metabolic functions, several genomes were affiliated with iron- and manganese-oxidizing bacteria [*Hyphomicrobium* (5), *Hyphomonas* (2), *Leptothrix* (2), and *Sideroxydans* (1)] and phosphate-accumulating bacteria [*Gemmatimonas* (4) and *Candidatus* Accumulibacter (2)]. From the phylogenomic perspective, the most dominant taxon was *Comamonadaceae* (8), followed by *Ralstonia* (6), *Erythrobacter* (6), *Sphingomonas* (4), *Pseudomonas* (4), unclassified Chlorobiaceae (4), *Flavobacterium* (3), Unclassified *Plantomyces* (3), and *Rhodobacter* (3). Within the phylum Cyanobacteria, a group of five genomes (PR.3.49, DS2.3.42, PR.023, DS3.002, and DS2.008) clustered outside of known genomes without deeper taxonomical information. These five genomes were mainly recovered from the distribution system and premise plumbing. Moreover, a couple of the genomes belonged to budding and prosthecate bacteria, such as also *Planctomycetes* (8), *Hyphomicrobium* (5), *Hyphomonas* (2), and *Brevundimonas* (2).

Species interaction revealed by OTU abundances Potential microbial interactions involving methano-/methyl-trophs within the GWDSW studied was analyzed using the correlation between OTUs by network analysis (Figure 5.5). The result suggested that a methanotrophic OTU (OTU4730, *Methylococcaceae*) was positively correlated with a methylotrophic OTU (OTU2964, *Methylotenera*), but negatively correlated with *Mycobacterium* (OTU5116), *Melainabacteria* (OTU4080), and *Comamonadaceae* (OTU1000). Other methylotrophs involved in this network included two *Hyphomicrobium* OTUs (OTU5340, OTU4471) and a methylobacterium OTU (OTU7328). They mostly formed negative correlation with each other, indicating a competitive relationship among methylotrophs. The rest heterotrophs in the network included five OTUs affiliated with *Comamonadaceae* (OTU5400, OTU2134, OTU7919, OTU7914, and OTU3639), *Erythrobacteraceae* (OTU4083), *Optitutaceae* (OTU2050), *Ralstonia* (OTU5), Ellin6529/*Chloroflexi* (OTU5311), etc. These results suggested extensive microbe-microbe interactions involving methano-/methylo-trophs in the studied environment.

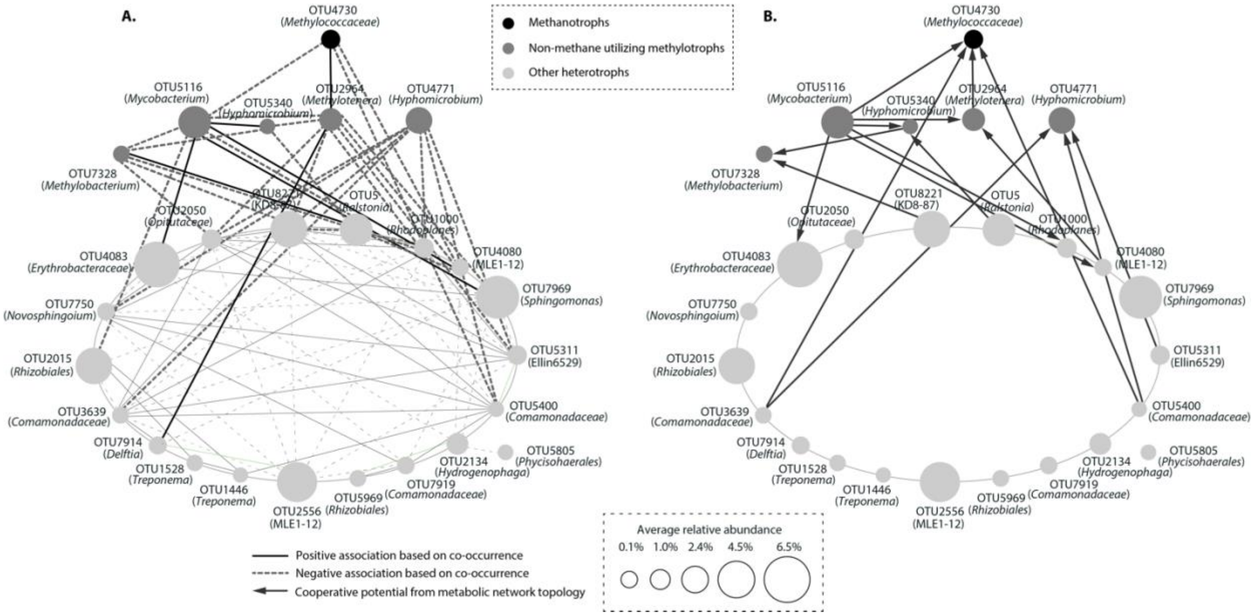


Figure 5.5 Microbe-microbe interactions predicted from A) co-occurrence analysis and B) metabolic network topology. In Panel A, network analysis of OTUs identified in the studied drinking water system was performed with OTUs having an average relative abundance of ≥ 0.01%. The network was represented by separating methanotrophs and NMMs from the remaining heterotrophs and was colored accordingly. Each node represents an OTU. Each node represents a genus. The circle size represents the average relative abundance of the genus across the studies system. Each edge represents a positive or negative association (p <0.05) determined by permutation and bootstrap analyses. In Panel B, microbe-microbe interactions involving methylotrophs were predicted using reconstructed metabolic networks from representative genomes recovered in this study (a given genome could correspond to multiple OTUs). Arrowheads pointed to the bacterial species that were supported by the species from the arrow ends (with BSS ≥ 0.74).

Species interaction revealed by metabolic interdependency Microbe-microbe interactions were further investigated by the extent of metabolic dependences governed by the exchange of metabolites (Figure 5.5). Such dependence could be represented through BSS metric, which were calculated according to the organization of the reconstructed metabolic network topologies of the interacting species. The scores between the dominant methanotroph, *Methylobacter* and four interacting species (including *Methylotenera*, *Mycobacterium*, and two *Comamonadaceae* species) were very high (≥ 0.74), suggesting that these accompanying taxa could provide a large fraction of the set of exogenously acquired nutrients (termed the 'seed set') from which all other compounds in the network could be synthesized. Similar trend was observed with *Hyphomicrobium* (representing OTU4771) and heterotrophs including *Ralstonia* and two *Comamonadaceae* species. Previously identified pathogen-related taxa, *Mycobacterium* and *Leptospira*, could support the metabolic needs of all the dominant methylotrophs (i.e., *Methylobacer, Methylotenera*, *Methylobacterium*, and *Hyphomicrobium*) as indicated by the greater BSS scores between these taxa. Interestingly, BSS scores suggested that *Mycobacterium* could provide essential metabolites to a number of heterotrophs in the studied environment, such as *Erythrobacteraceae, Rhodoplanes,* and MLE1-12/*Melainabacteria*. Collectively, while methanotrophs might be key players in the community as primary producers generating organic carbon compounds from methane, NMMs and other heterotrophs could also contribute to the community by providing essential metabolites to methanotrophs.

5.5 Discussion

The Champaign-Urbana GWDWS served as a model to elucidate the interaction networks centered by methano-/methylo-trophs. Methane could be the main carbon and energy source in this oligotrophic environment based on carbon flux. Its concentration was approximately 2.4 g C/m^3 or 4.86 mg C m^{-2} d^{-1} in the distribution system (Ling et al., 2016), which was a comparable amount to the organic carbon flux in the distribution system biofilm phase, estimated at < 0.1-0.2 mg C m^{-2} d^{-1} (van der Kooij, 1999). Methane concentration in the source water was even higher, ranging from 2.4 mg C/m^3 to 12.3 g C/m^3 (Flynn et al., 2013) (Table A.2). Both Type I and Type II methanotrophs (i.e., *Methylobacter, Methylomonas*, and *Methylocystis*) and NMMs (i.e., *Methylotenera*, *Hyphomicrobium*, *Methylophilus*, and *Methylobacterium*) were abundant in the systems. Other populations in the drinking water microbiome included members of Alpha-, Beta-, Gamma- *Proteobacteria*, *Cyanobacteria-Melainabacteria*, and *Bacteroidetes*.

Environmental disturbances from the treatment and distribution processes could continually alter the drinking water microbiome and microbe-microbe interacting patterns (Zhang et al., 2017a). It provided a unique system to investigate microbial methane utilization through interacting species.

Methane serving as a substrate for bacterial growth might be a widespread phenomenon in GWDWSs as they were frequently reported to be the dominant groups in many systems. For example, methanotrophs (*Methylococcales*, *Methylovulum*, *Crenothrix polyspora*, *Methylocella*, and *Methylocystis*) and NMMs (*Methylibium*, *Hypomicrobium*) were found in groundwater-fed rapid gravity sand filters in Denmark (Gulay et al., 2016). Similarly, methanotrophs affiliated with *Methylomonas* and *Methylobacter*, and NMMs with *Methylophilus* were reported in groundwater-fed tricking filters in the Netherlands (de Vet et al., 2009). Systems in Florida, USA, Germany, and Italy also reported the prevalence of methanotrophs and accompanying species (Stoecker et al., 2006; Vigliotta et al., 2007; Kelly et al., 2014). These results suggest that the contribution of dissolved methane to the proliferation of microbial communities in GWDWSs is underappreciated.

Many cooperative relationships involving methanotrophs have been reported so far. For example, an early study observed aggregates formed by *Methylococcus* and *Hyphomicrobium* in enrichment cultures from peat samples at pH 4. Another study suggested the potential cooperation between *Methylococcaceae* (*Methylobacter*) and *Methylophilaceae* (*Methylotenera*) based on their coordinated response to methane and nitrate under aerobic conditions (Beck et al., 2013). Methanotrophs could be associated with non-methylotrophic heterotrophs such as *Flavobacteriaceae, Burkholderiales* and *Pseudomonas* (Oshkin et al., 2015). van der Ha *et al.* observed the relationship between methanotrophs and NMMs (*Methylomonas* and *Methylophilus*) and between methanotrophs and heterotrophs *(Methylomonas* and *Flavobacterium*) by using different copper concentrations (van der Ha et al., 2013). These observations all suggest that methanotrophs can form diverse interspecies relationships with many microbes under different environmental conditions. A recent study reported a mechanism behind these interspecies relationships, in which changes at the transcription level of methanotrophs were observed due to the presence of a nonmethanotrophic partner and methanol was released into the co-culture media for the growth of its partner (Krause et al., 2017). Nevertheless, the mechanistic details of how and why the methanotrophs share their carbon with other species, and whether and what they gain in return, are still not clear. Most likely, these interactions are based on complex exchanges of different metabolites with different partners. Therefore, methods that can fast screen the potential partner species and the exchanging metabolites becomes critical in such communities.

Metabolite exchanges between the interacting species can be reflected in the organization of genome-scale metabolic networks reconstructed from genomic data. Various graph theory-based methods have been developed to predict microbe-microbe interactions directly from network topology (Milo et al., 2002; Parter et al., 2007; Kreimer et al., 2008; Levy et al., 2015). Specifically, there are two main mechanisms driving species co-occurrence: (i) habitat filtering – microbes occupy a similar nutritional niche and compete, and (ii) species assortment – microbes have complementary metabolisms and cooperate. A recent study investigated these two driving forces behind the co-occurrence of microbes in the human gut through metabolic competition and complementarity indices based on network topology of genome-scale metabolic modules. They determined that the metabolic competition index best explained the species co-occurrence patterns and habitat filtering was the main driving force (Levy and Borenstein, 2013). Applying this framework in our study, we discovered that *Methylobacter* could form communal relationship with *Methylotenera*, *Mycobacterium*, and two *Comamonadaceae* species. It was surprising to find that pathogen-related taxa could support the exogenously required compounds in the metabolic network of many methylotrophic species. These predicted microbe-microbe interactions should be confirmed by future experimental studies. The genome-based approach facilitates the inference of evolutionary and ecological processes that shape species interactions and community assembly centered by methanotrophs across different environments on a large-scale.

5.6 References

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z. et al. (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* **11**: 1144-1146.

Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J. et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**: 13219.

Barker, J.F., and Fritz, P. (1981) The occurrence and origin of methane in some groundwater-flow systems. *Can J Earth Sci* **18**: 1802-1816.

Beck, D.A., McTaggart, T.L., Setboonsarng, U., Vorobev, A., Goodwin, L., Shapiro, N. et al. (2015) Multiphyletic origins of methylotrophy in Alphaproteobacteria, exemplified by comparative genomics of Lake Washington isolates. *Environ Microbiol* **17**: 547-554.

Beck, D.A.C., Kalyuzhnaya, M.G., Malfatti, S., Tringe, S.G., del Rio, T.G., Ivanova, N. et al. (2013) A metagenomic insight into freshwater methane-utilizing communities and evidence for cooperation between the *Methylococcaceae* and the *Methylophilaceae*. *PeerJ* **1**: e23.

Borenstein, E., and Feldman, M.W. (2009) Topological signatures of species interactions in metabolic networks. *J Comput Biol* **16**: 191-200.

Bousquet, P., Ciais, P., Miller, J.B., Dlugokencky, E.J., Hauglustaine, D.A., Prigent, C. et al. (2006) Contribution of anthropogenic and natural sources to atmospheric methane variability. *Nature* **443**: 439-443.

Brooks, J.P., Edwards, D.J., Harwich, M.D., Rivera, M.C., Fettweis, J.M., Serrano, M.G. et al. (2015) The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol* **15**.

Buswell, A., and Larson, T. (1937) Methane in ground waters. *J Am Water Works Assoc* **29**: 1978-1982.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2010a) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266-267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K. et al. (2010b) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335-336.

Carr, R., and Borenstein, E. (2012) NetSeed: a network-based reverse-ecology tool for calculating the metabolic interface of an organism with its environment. *Bioinformatics* **28**: 734-735.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J. et al. (2015) Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr Biol* **25**: 690-701.

Chistoserdova, L. (2011) Modularity of methylotrophy, revisited. *Environ Microbiol* **13**: 2603-2622.

Chistoserdova, L., and Lidstrom, M. (2013) Aerobic Methylotrophic Prokaryotes. In *The Prokaryotes*. Rosenberg, E., DeLong, E., Lory, S., Stackebrandt, E., and Thompson, F. (eds): Springer Berlin Heidelberg, pp. 267-285.

Cox, M.P., Peterson, D.A., and Biggs, P.J. (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11**: 485.

de Vet, W.W.J.M., Dinkla, I.J.T., Muyzer, G., Rietveld, L.C., and van Loosdrecht, M.C.M. (2009) Molecular characterization of microbial populations in groundwater sources and sand filters for drinking water production. *Water Res* **43**: 182-194.

Duhaime, M.B., Deng, L., Poulos, B.T., and Sullivan, M.B. (2012) Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ Microbiol* **14**: 2526-2537.

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

Faust, K., and Raes, J. (2012) Microbial interactions: from networks to models. *Nat Rev Microbiol* **10**: 538-550.

Faust, K., and Raes, J. (2016) CoNet app: inference of biological association networks using Cytoscape. *F1000Research* **5**: 1519.

Flynn, T.M., Sanford, R.A., Ryu, H., Bethke, C.M., Levine, A.D., Ashbolt, N.J., and Domingo, J.W.S. (2013) Functional microbial diversity explains groundwater chemistry in a pristine aquifer. *BMC Microbiol* **13**.

Gulay, A., Musovic, S., Albrechtsen, H.J., Abu Al-Soud, W., Sorensen, S.J., and Smets, B.F. (2016) Ecological patterns, diversity and core taxa of microbial communities in groundwater-fed rapid gravity filters. *ISME J* **10**: 2209-2222.

Gunsalus, R.P., Zeikus, J.G., and Wolfe, R.S. (1972) Microbial modification of ground water.

Henry, C.S., DeJongh, M., Best, A.A., Frybarger, P.M., Linsay, B., and Stevens, R.L. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* **28**: 977-982.

Hug, L.A., Thomas, B.C., Sharon, I., Brown, C.T., Sharma, R., Hettich, R.L. et al. (2016) Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ Microbiol* **18**: 159-173.

Hwang, C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2012) Microbial community dynamics of an urban drinking water distribution system subjected to phases of chloramination and chlorination treatments. *Appl Environ Microbiol* **78**: 7856-7865.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Kalyuzhnaya, M.G., Bowerman, S., Lara, J.C., Lidstrom, M.E., and Chistoserdova, L. (2006) *Methylotenera mobilis* gen. nov., sp nov., an obligately methylamine-utilizing bacterium within the family *Methylophilaceae*. *Int J Syst Evol Microbiol* **56**: 2819-2823.

Kalyuzhnaya, M.G., Lapidus, A., Ivanova, N., Copeland, A.C., McHardy, A.C., Szeto, E. et al. (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. *Nat Biotechnol* **26**: 1029-1034.

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

Kelly, J.J., Minalt, N., Culotti, A., Pryor, M., and Packman, A. (2014) Temporal variations in the abundance and composition of biofilm communities colonizing drinking water distribution pipes. *PLoS ONE* **9**: e98542.

Knirk, C.F. (1908) Natural gas in the glacial drift of Champaign County. In *Illinois State Geological Survey, Bulletin No 14, Yearbook for 1908*.

Kolenbrander, P.E. (2011) Multispecies communities: interspecies interactions influence growth on saliva as sole nutritional source. *Int J Oral Sci* **3**: 49-54.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., and Schloss, P.D. (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microb* **79**: 5112-5120.

Krause, S.M., Johnson, T., Samadhi Karunaratne, Y., Fu, Y., Beck, D.A., Chistoserdova, L., and Lidstrom, M.E. (2017) Lanthanide-dependent cross-feeding of methane-derived carbon is linked by microbial community interactions. *Proc Natl Acad Sci USA* **114**: 358-363.

Kreimer, A., Borenstein, E., Gophna, U., and Ruppin, E. (2008) The evolution of modularity in bacterial metabolic networks. *Proc Natl Acad Sci USA* **105**: 6976-6981.

Letunic, I., and Bork, P. (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* **44**: W242-W245.

Levy, R., and Borenstein, E. (2013) Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc Natl Acad Sci USA* **110**: 12804-12809.

Levy, R., Carr, R., Kreimer, A., Freilich, S., and Borenstein, E. (2015) NetCooperate: a network-based tool for inferring host-microbe and microbe-microbe cooperation. *BMC Bioinformatics* **16**: 164.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*: btv033.

Li, H., and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Ling, F., Hwang, C., LeChevallier, M.W., Andersen, G.L., and Liu, W.T. (2016) Core-satellite populations and seasonality of water meter biofilms in a metropolitan drinking water distribution system. *ISME J* **10**: 582-595.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363-1371.

Martienssen, M., and Schops, R. (1999) Population dynamics of denitrifying bacteria in a model biocommunity. *Water Res* **33**: 639-646.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A. et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610-618.

Miller, C.S. (2013) Assembling full-length rRNA genes from short-read metagenomic sequence datasets using EMIRGE. *Methods Enzymol* **531**: 333-352.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002) Network motifs: simple building blocks of complex networks. *Science* **298**: 824-827.

Oshkin, I.Y., Beck, D.A.C., Lamb, A.E., Tchesnokova, V., Benuska, G., McTaggart, T.L. et al. (2015) Methane-fed microbial microcosms show differential community dynamics and pinpoint taxa involved in communal response. *ISME J* **9**: 1119-1129.

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

Parter, M., Kashtan, N., and Alon, U. (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evol Biol* **7**: 169.

Schink, B. (1997) Energetics of syntrophic cooperation in methanogenic degradation. *Microbiol Mol Biol Rev* **61**: 262-280.

Segata, N., Bornigen, D., Morgan, X.C., and Huttenhower, C. (2013) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat Commun* **4**: 2304.

Stoecker, K., Bendinger, B., Schoning, B., Nielsen, P.H., Nielsen, J.L., Baranyi, C. et al. (2006) Cohn's Crenothrix is a filamentous methane oxidizer with an unusual methane monooxygenase. *Proc Natl Acad Sci USA* **103**: 2363-2367.

Tennessen, K., Andersen, E., Clingenpeel, S., Rinke, C., Lundberg, D.S., Han, J. et al. (2016) ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J* **10**: 269-272.

van der Ha, D., Vanwonterghem, I., Hoefman, S., De Vos, P., and Boon, N. (2013) Selection of associated heterotrophs by methane-oxidizing bacteria at different copper concentrations. *Anton Leeuw Int J Gen Mol Microbiol* **103**: 527-537.

van der Kooij, D. (1999) Potential for biofilm development in drinking water distribution systems. *J Appl Microbiol* **85**: 39s-44s.

Vigliotta, G., Nutricati, E., Carata, E., Tredici, S.M., De Stefano, M., Pontieri, P. et al. (2007) Clonothrix fusca Roze 1896, a filamentous, sheathed, methanotrophic gamma-proteobacterium. *Appl Environ Microbiol* **73**: 3556-3565.

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

Zhang, Y., Oh, S., and Liu, W.T. (2017a) Impact of drinking water treatment and distribution on the microbiome continuum: an ecological disturbance's perspective. *Environ Microbiol* **19**: 3163-3174.

Zhang, Y., Kitajima, M., Whittle, A.J., and Liu, W.T. (2017b) Benefits of genomic insights and CRISPR-Cas signatures to monitor potential pthogens across dinking wter poduction and dstribution sstems. *Front Microbiol* **8**: 2036.

CHAPTER 6 SUMMARY AND POTENTIAL BIASES

Drinking water systems are unique ecosystems because of their complexity in infrastructure and expansion in distance with each stage influencing downstream microbiomes. Recent studies have gained insights into the composition and spatiotemporal variation of drinking water microbiomes. However, it is hard to translate the knowledge into practical guidance for the engineered systems as our understanding on the functions of drinking water microbiome is very limited both at the community level and at the population level. Contrary to our perception that drinking water microbiomes are microbial communities mostly without a discernible function, microbes carry out various functions in the ecosystem, many of which might due to interspecies interaction. This series of studies used the groundwater-sourced drinking water system located in the Champaign-Urbana area as a model system, investigated the ecological patterns at the community, population, and multi-species levels, and provided insights into the microbial functions and interactions in drinking water microbiomes. The key findings are:

- The treatment process could be viewed as ecological disturbances over space and abstraction and softening processes caused substantial changes in the community structure and functions related to methanogenesis and phosphotransferase system (PTS) regulatory machinery.

- Predation by eukaryotic populations could an important disturbance to the bacterial microbiome.

- The presence/absence of specific virulence machinery could be used to determine the pathogenicity potential of draft genomes related with pathogenic species, including *Legionella*, *Mycobacterium*, *Parachlamydia*, and *Leptospira*.

- CRISPR-Cas genetic signature was a potential biomarker for the monitoring of *Legionella* related strains across different drinking water systems.

- Methano-/methylo-trophs were overlooked populations dominant and prevalent in finished water and the distribution system of groundwater-sourced drinking water systems.
- Methanotrophs potentially interacted with a number of non-methanotrophic methylotrophs and other heterotrophs through exchanging essential metabolites.

These findings hold implications for water treatment and monitoring:

- Increased effort is needed to link the identity of different microorganisms to their function in specific environments. Understanding the structure-function relationship will greatly enhance our ability to manage drinking water microbiomes.
- Develop novel concepts and monitoring tools to control opportunistic pathogens in drinking water systems. This requires further research on the influence of microbial community structure and composition on the types and concentrations of pathogenic species present in drinking water. Novel microbial indicators can be proposed based on ecological theories.
- Predict changes in drinking water microbiomes and shape the microbiomes towards desirable composition and function. The large number of data of drinking water microbiomes deposited in public databases can be further exploited by generalizing the trend observed and applying ecological principles into models. Such predictive framework could eventually improve the water quality that the customers receive.

Potential biases associated with methods used in this dissertation:

Water sampling devices 16S rRNA gene and metagenomics analyses usually require a large volume of water (ranging from 1 L-2000 L) to be concentrated on-site to collect enough biomass.

The commonly-used laboratory concentration devices often cannot meet this requirement and no standardized devices have been developed to address this problem. Studies in this dissertation used a four-layer water purifier validated in a previous study (Chao et al., 2013). These four filtration components included prescreen, granular activated carbon (GAC), second screen, and a hollow fiber membrane filter. Only the biomass deposited on the hollow fiber membrane was collected and that absorbed by the GAC component was ignored. According to the manufacturer's information, the inner hollow fiber membranes are made with polysulfone, and the average pore diameter of the inner hollow fiber membranes is approximately 0.01 μm (Shimagaki et al., 2000). Thus, most cells are likely to be trapped by this component instead of the GAC component. In the future, it is crucial to standardize sample volume and concentration methods for drinking water microbiome studies, which will facilitate the comparison of microbial community profiles between studies done by different research groups.

Seasonal variation For water-phase sampling, four biological replicates were collected at each site during the summer months of 2014 (i.e., June, July, August, and September). Within this sampling period, the operation of the treatment plant and the water chemistry remained relative stable as shown by Table A.1. However, water temperature in the studied system varies with seasons (ranging from 7 to 23 °C) (Hwang et al., 2012a), which can influence dissolved methane concentration and microbial interaction in the distribution network. As a rule of thumb, every 10 °C increase in water temperature leads to a two-fold increase in microbial activity (Barineau, 2006). Therefore, seasonal variations in the drinking water microbiome remain to be determined by future studies.

DNA extraction efficiency The effect of DNA extraction methods on the quantity and quality of DNA yields from drinking water microorganisms has been evaluated by a previous study (Hwang et al., 2012b). Hwang et al. tested five widely used DNA extraction methods with selected drinking water bacteria with different cell wall properties and distribution system samples. The study recommended the commercial kit, FastDNA, because it was easy to use and providing representative microbial community information and reproducibility. Therefore, FastDNA kit was chosen for extracting DNA from all the samples included this dissertation.

Genome recovery The recovery of draft genomes mainly depends on four factors — the complexity of the community, the sequencing depth of metagenomes, the relative abundance of different species, and the genetic makeup of the targeted microorganisms (Kang et al., 2015; Wu et al., 2016). It is difficult to recover many genomes from a complex community with a low sequencing depth. Microorganisms with high and low GC content in their genomes are generally easy to be recovered from metagenomes, such as *Mycobacterium* spp. The most abundant lineages have a higher possibility to be recovered but it does not mean low-abundance ones cannot be recovered. In this dissertation, genomes were recovered based on read coverage, tetranucleotide frequency, and the occurrence of unique marker genes to minimize the contamination of each recovered genome. Manual curation was carried out to further reduce contamination with statistics calculated by quality assessment tools.

Using virulence genes to evaluate pathogenicity The presence of one or two virulence factors in a bacterial genome can hardly be interpreted as virulence, but the presence of all the major virulence factors involving in circumventing host immune system at different stages of infection would be a much stronger indication of virulence. However, genomic signatures only provide information on the metabolic potential, but not to the activity and expressed virulence of pathogen-related species. Therefore, further studies combining microbiological (e.g., cultivation and animal models), genomic and metabolic (e.g., transcriptomics and proteomics) methods should be carried out to understand the role of these virulence genes at the level of gene expression, protein function and regulation, and interaction with host immune system to confirm the virulence of these strains for immunocompromised individuals.

The time scale of using CRISPR spacers as a biomarker for typing Given the probable horizontal origin of CRISPR-Cas systems, their frequent acquisition and loss among related organisms, and the frequent addition and loss of CRISPR spacers, CRISPR spacers are limited to be used to monitor a population at a smaller evolutionary scale. For example, the longest time for these spacers to remain conserved in a *Leptospirillum* strain was five years or longer (Sun et al., 2016). Cautions are needed when applying this method over a relatively large evolutionary scale.

Inferring microbial cooperative potential from metabolic network topology Microbial cooperative potential was inferred based on large-scale metabolic data which were subjective to missing and inaccurate annotation. The automatic ModelSEED pipeline was used to construct the draft genome-scale metabolic data and manual curation was carried out to correct miss annotations and reaction directionality. The differences in qualify among these reconstructions may minimize their predictive potential (Thiele and Palsson, 2010; Magnusdottir et al., 2017). However, the biases are usually consistent as they are derived from comparison-based methods and still allow comparison between the scores of various species. It should be noted that the inference was based on static genome-scale metabolic network topology and did not consider other quantitative properties related to metabolic reactions, such as regulation, stoichiometry, reaction rates, and dynamics.

6.1 References

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X.X., Wu, W.M., and Zhang, T. (2013) Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Sci Rep* **3**: 3550.

Hwang, C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2012a) Microbial community dynamics of an urban drinking water distribution system subjected to phases of chloramination and chlorination treatments. *Appl Environ Microbiol* **78**: 7856-7865.

Hwang, C.C., Ling, F.Q., Andersen, G.L., LeChevallier, M.W., and Liu, W.T. (2012b) Evaluation of methods for the extraction of DNA from drinking water distribution system biofilms. *Microbes Environ* **27**: 9-18.

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

Magnusdottir, S., Heinken, A., Kutt, L., Ravcheev, D.A., Bauer, E., Noronha, A. et al. (2017) Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol* **35**: 81-89.

Shimagaki, M., Fukui, F., Sonoda, T., and Sugita, K. (2000) Polysulfone hollow fiber semipermeable membrane. In: Toray Industries, Inc.

Sun, C.L., Thomas, B.C., Barrangou, R., and Banfield, J.F. (2016) Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. *ISME J* **10**: 858-870.

Thiele, I., and Palsson, B.O. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* **5**: 93-121.

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

APPENDIX A SUPPLEMENTARY MATERIALS

Table A.1 Operational data from the drinking water treatment plant.

	RW						FW							
	Temp *	pH	Tot Alk**	Tot Hard**	Ammonia ***	Fe	pH	Tot Alk	Tot Hard	Ammonia	Fe	Turbidity	Cl_2 Residual	
Time	℃		mg/L	mg/L	mg/L	mg/L		mg/L	mg/L	mg/L	mg/L	NTU	Free	Total
06-14	12.4	7.6	376±6	295±5	1.17±0.06	1.33±0.45	8.8	162±5	85±4	0.00±0.00	0.00±0.00	0.052±0.012	2.30±0.25	2.41±0.24
07-14	12.5	7.5	377±8	293±3	0.98±0.06	1.01±0.16	8.8	165±6	85±4	0.30±0.52	0.00±0.00	0.049±0.006	2.26±0.27	2.29±0.27
08-14	12.4	7.5	386±4	292±3	1.12±0.01	1.11±0.10	8.8	171±8	87±7	0.00±0.01	0.00±0.00	0.067±0.022	2.76±0.23	2.78±0.24
09-14	12.5	7.7	380±6	289±5	0.00±0.00	0.00±0.00	8.8	172±8	90±6	0.00±0.00	0.00±0.00	0.047±0.008	2.77±0.31	2.83±0.30

* This is for groundwater. For tap water, temperature varied season to season, ranging from approximately 7 to 23°C, with the colder temperatures measured in the winters and warmer temperatures measured during fall and summer seasons, according to a previous study on the system (Hwang et al., 2012a).

** Total alkalinity and total hardness as $CaCO_3$

*** as NH_3

Table A.2 Geochemistry of groundwater in Mahomet aquifer reported by Flynn et al. (Flynn et al., 2013).

Well	Temp	SO_4^{2-}	CH_4	H_2	DIC	DOC
	℃	mM	μM	nM	mM	mg/L
Chm94B	13.7	0.58	< 0.2	25	7.8	2.2
Chm96A	13.8	0.41	1	3	7.2	1.3
Frd94A	14.2	0.98	2	3	7.4	< 0.4
Iro95A	14.3	1.50	1	60	n/a	3.3
Iro96A	12.1	4.23	1	n/a	n/a	n/a
Iro98B	13.0	4.68	3	10	6.6	43.0
Iro98D	13.6	0.72	19	180	7.9	1.9
Ver94A	14.4	4.57	2	n/a	6.7	1.8
Ver94B	13.7	10.73	1	89	4.8	1.1
Chm94A	14.1	0.07	4	n/a	8.0	3.6
Chm95A	14.0	0.14	8	4	7.7	2.1
Chm95B	13.8	0.04	30	3	7.9	2.0
Chm95C	13.7	0.11	3	20	6.6	0.5
Frd94B	15.4	0.05	43	9	7.4	< 0.4
Iro98C	13.3	0.04	15	66	7.6	2.3
Ver94C	13.6	0.23	3	46	7.4	1.1
Ver94D	13.9	0.18	10	n/a	7.7	0.8
AnderN	14.8	0.02	91	144	6.6	n/a
AnderS	15.1	0.02	1237	175	25.9	n/a
CardiS	13.6	0.03	454	240	7.5	n/a
Chm95D	14.0	< 0.01	220	12	7.6	1.6
Chm98A	13.7	< 0.01	676	24	7.9	4.2
PklndE	14.6	0.03	221	63	8.7	n/a
PklndW	14.4	0.03	611	100	6.0	n/a
RaiRd	14.4	0.02	106	50	6.4	n/a