Towards a predictive framework for microbial management in drinking water systems

by

Quyen Melina Bautista-de los Santos

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

February 2017

Abstract

The application of DNA sequencing-based approaches to drinking water microbial ecology has revealed the presence of an abundant and diverse microbiome; therefore, the possibility of harnessing drinking water (DW) microbial communities is an attractive prospect in order to address some of the current and emerging challenges in the sector. Moreover, these multiple challenges suggest that a shift in the DW sector, from a "reactive and sanctioning" paradigm to a "due diligence/proactive" based approach may be the key in identifying potentially adverse events. My research project has focused on the characterization of the microbial ecology of full-scale DW systems using DNA sequencing-based approaches, with the aim of exploring how the obtained insights could be applied into a predictive/proactive microbial management approach. To achieve this aim, I have focused my efforts on sampling multiple full-scale DW systems in order to elucidate the impacts of: (i) methodological variation and (ii) system properties on DW microbial communities, using a combination of bioinformatics, molecular biology, microbial ecology and multivariate statistical analyses.

Regarding methodological variation, I have elucidated the impacts of sample replication, PCR replication, sample volume and sampling flow rate on the structure and membership of DW microbial communities. This was the first time that methodological variation was explored in the DW context, and the first time that multi-level replication has been tested and applied in DW molecular microbial ecology. Moreover, my findings have direct implications for the design of future sampling campaigns. Regarding system properties, I have shown that microbial communities in DW distribution systems (DWDSs) undergo diurnal variation, and therefore are linked to water use patters/hydraulics in the systems. I have also shown that sampling locations in the same distribution system are similar, with OTUs found across sampling locations at different relative abundance and detection frequency levels. An assessment of the impact of source water type and treatment processes showed that disinfection is a key treatment step for community composition and functional potential, and that several genes related to protection against chlorine/oxygen species are overabundant in chlorinated and chloraminated systems. Looking to the future, I believe that the application of a "toolbox" of techniques is key in shifting towards a proactive approach in DW management, that multidisciplinary synergies hold the possibility of changing the way in which DW systems have been studied and managed for over 100 years.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Firstly, I would like to thank my supervisors, Dr. Ameet Pinto and Prof. William Sloan for the opportunity to conduct this research, and their guidance and support along the way. Thanks also to Dr. Umer Ijaz and Dr. Joanna Schroeder-Davis for their help with data analysis. I started this PhD with no real wet lab experience. Ms. Julie Russell and Ms. Anne McGarrity, our lab technicians, trained me and accompanied me in success and in failure, and this project would not have been possible without their invaluable support both in and out of the lab.

Thanks to all my colleagues in the office for making the supporting environment that I've enjoyed for the last 4 years: Kevin Bayle, Jillian Couto, Cosmika Goswami, Orges Koci, Mathieu Larronde-Larretche, Siding Luo, Ruzanna Mat-Jusoh, Ioannis Sampsonidis, Seung Gu Shin and Stephanie Turnbull. Special thanks goes to Szymon Calus, Stephanie Connelly, Zihan Dai, Sarah Haig, Ciara Keating, Asha Rani, Melanie Schirmer, Maria Sevillano-Rivera, Rungroch Sungthong and Erifyli Tsagkari for the very useful discussions about research, life and random topics that provided necessary distractions for me.

I'd like to thank Scottish Water (SW) for their invaluable support. Special thanks to Dr. Mark Haffey, my SW supervisor, for facilitating all things related to sampling and SW assets. Many thanks to all the staff that I engaged with (team leaders, process scientists, operators, sampling team members, couriers) for their helpfulness.

Thanks to Aunt Mary and Francheska, to my friends Melissa Alvarez, Jovanna Aristy, Yenisse Brito, Karla Marte, Maria Valdivia and Zuany Victoria for their encouragement. Finally, thanks to my Mom, Ms. Amanda de los Santos, for her unconditional love and support.

Author's declaration

I declare that no portion of the work in this thesis has been submitted in support of any application for any other degree or qualification from this or any other university or institute of learning. I also declare that the work presented in this thesis is entirely my own contribution unless otherwise stated.

Quyen Melina Bautista-de los Santos

Glasgow, February 2017.

Nomenclature

16S rRNA	16S Ribosomal RNA
AMOVA	Analysis of Molecular Variance
ANOSIM	Analysis of similarities
ANOVA	Analysis of variance
AOA	Ammonia oxidizing archaea
AOB	Ammonia oxidizing bacteria
AOC	Assimilable organic carbon
ARG	Antibiotic resistance genes
ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
CARD	Comprehensive Antibiotic Resistance Database
Chl	Chlorinated
Chm	Chloraminated
CT	Contact time
ddNTP	Dideoxynucleotide
Dis	Disinfected
DMMM	Drichlet multinomial mixture model
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
Drf	Disinfectant residual-free
DW	Drinking water
DWDS	Drinking water distribution system
DWS	Drinking water system
DWTP	Drinking water treatment plant
emPCR	Emulsion polymerase chain reaction
FTU	Fraction of OTUs that could not be mapped to KEGG organisms
GAC	Granular activated carbon
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KO	KEGG ortholog number
LRV	Log removal value
MAC	*Mycobacterium avium* complex
MRA	Microbial risk assessment
MRG	Metal resistance gene

NCBI	National Centre for Biotechnology Information
NMDS	Non-metric multidimensional scaling
NTM	Non-tuberculous Mycobacteria
OP	Opportunistic pathogen
OTU	Operational Taxonomic Unit
PAMP	Pathogen-associated molecular pattern
PCoA	Principal coordinates analysis
PCR	Polymerase chain reaction
PERANOVA	Permutational analysis of variance
PERMANOVA	Permutational multivariate analysis of variance
POU	Point of use
QMRA	Quantitative microbial risk assessment
qPCR	Quantitative polymerase chain reaction
RCS	Reactive chlorine species
RNA	Ribonucleic acid
ROS	Reactive oxygen species
SIP	Stable isotope probing
SRA	Sequence read archive
T	Temperature
TCCR	Transparency, clarity, consistency and reasonableness
TE	Transposable element
TOC	Total organic carbon
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
WHO	World Health Organization
WSP	Water safety plan

1. Introduction

1.1. Background

The application of chlorine to disinfect drinking water in the early 20th century marked the beginning of modern drinking water (DW) treatment as we know it today. The improvement in DW quality as a consequence of the inactivation of pathogenic microorganisms resulted in a reduction of waterborne diseases in the western world (e.g. cholera, dysentery, typhoid fever). In addition to the development of treatment processes, the introduction of water quality targets allowed for the assessment of process efficiency and water integrity. Chlorine is the most widely used disinfectant around the world because of its effectiveness and affordability.

More than 100 years later, the foundations of DW treatment and supply that made the process successful remain largely the same, but the sector faces a different set of challenges (Figure 1.1). For instance, *Cryptosporidium*, a pathogenic protozoan resistant to chlorine-based disinfectants, has been the cause of numerous documented Cryptosporidiosis outbreaks in the US (Corso et al. 2003), UK (Chartered Institute of Environmental Health 2016) and several other European countries (Semenza & Nichols 2007); while other emerging pathogens (i.e. microorganisms responsible for infectious diseases which have appeared or increased in occurrence in the past four decades, OECD/WHO 2003) are challenging the way in which systems are monitored. The first international drinking water standards were published by the World Health organization (WHO) in 1958, a document of 148 pages; almost 60 years later, the current WHO drinking water guideline (WHO 2011) is almost 4 times bigger than the original document and includes many more health-based targets associated with microbiological, chemical, radiological and aesthetic aspects. To meet this stricter standards, more intensive treatment has to be applied (e.g. additional treatment units such as membranes) which means higher consumption of energy and resources to treat the raw water.

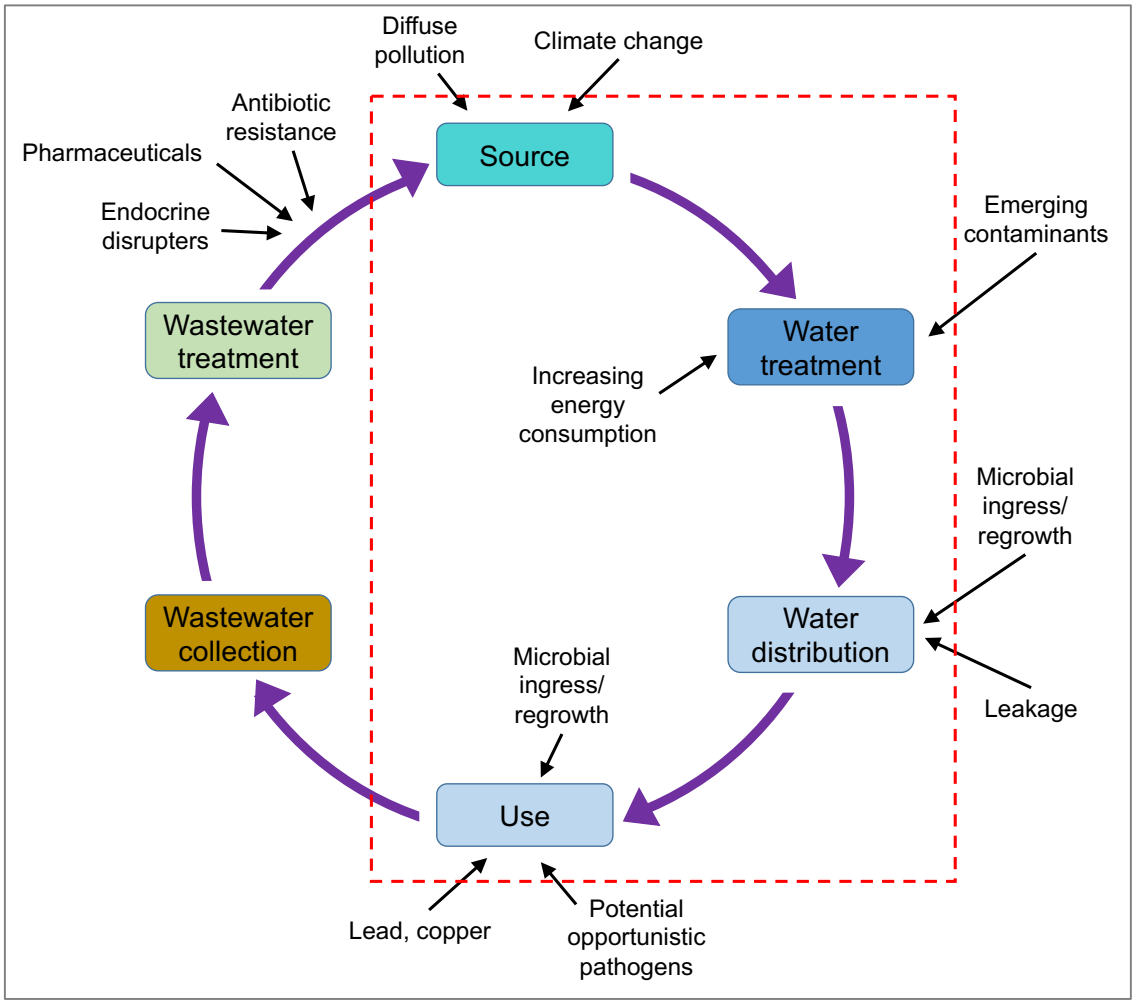

Figure 1.1. Overview of the urban water cycle.
Component of the cycle are enclosed in rectangles, links between components are indicated by purple arrows. Challenges are indicated by black arrows. The red dashed line encloses the drinking water cycle within the urban water cycle.

In addition to the increased regulation, the introduction of the concept of sustainability and its application to DW has meant that additional provisions need to be put in place to guarantee an adequate supply for the future generations. Sustainable approaches towards management of DW systems include the consideration of DW supply as part of the broader urban water cycle, which requires a reflection of the nexus between drinking water and the other components of the cycle (e.g. catchment management, wastewater treatment and discharge). Climate change also has forced the sector to reflect on the potential implications of a reduction of in availability of locally sourced freshwater and how this would impact both quantity and quality of the supplied water, among other potential negative impacts. These multiple pressures have motivated a shift in the DW sector, from a "reactive and sanctioning" paradigm to the application of "due diligence" measures aimed at preventing foreseeable harm at reasonable cost, by identifying potentially adverse events (OECD/WHO 2003).

One of the key aspects of this "due diligence" approach has been a renewed look at the microbiological quality of water. The application of molecular biology techniques (e.g. cloning, PCR, DNA sequencing) to study DW microbiology has challenged the concept that DW is sterile or free from microorganisms after disinfection, revealing the presence of an abundant (e.g. 10^3-10^5 cells/ml) and diverse (e.g. thousands of bacterial species) microbiome. Traditional culture-based techniques applied in the field are able to capture only a fraction of the community that can survive under the conditions imposed, which is lower than the total community present in the sample. The application of DNA sequencing-based techniques has allowed the identification of organisms that have not been characterized using culture-based approaches, while the analysis of the sequencing data and environmental variables can shed light on their interactions and/or associations with each other and the environment. The possibility of harnessing DW microbial communities is an attractive prospect in order to address some of the issues presented above. For instance, biological applications consume less energy/resources; moreover, exploiting the properties of biological systems to improve processes has been successful in wastewater treatment. Therefore, this constitutes an opportunity in the DW field to both gain insights on the detrimental organisms (e.g. pathogens) present in the drinking water to improve their removal/inactivation efficiency through process interventions, and also to characterize and exploit the non-pathogenic fraction of the community. As highlighted by LeChevallier and colleagues (AWWA 1999), "*knowledge is the first line of defense for those who provide safe drinking water*."

1.2 Aim and objectives

Aim:

- To characterize the microbial communities of full-scale drinking water systems (DWSs) applying DNA sequencing-based approaches and provide recommendations on how to incorporate predictive elements to their current management strategy.

Objectives:

- To assess the methodological approaches to study the DW microbiome.

- To describe the composition, abundance and metabolic potential of microbial communities in drinking water distribution systems.

- To assess the relationships between treatment components (e.g. source water, treatment processes, disinfection strategy, distribution, etc.), water quality parameters and microbial community structure, membership and metabolic potential.

- To elucidate the dynamics (temporal and spatial variations) of the DW microbiome in full-scale drinking water distribution systems.

- To provide recommendations on how to develop a predictive framework for microbial management in full-scale drinking water systems using DNA sequencing-based approaches, based on the results obtained in this research project and current knowledge on the DW microbiome.

1.3 Outline of thesis

Chapter 2 presents a literature review of drinking water treatment and distribution, with an emphasis on microbial aspects. The chapter covers technical aspects of the system starting from catchment management, treatment processes and microbial removal efficiency, the distribution system and finally premises plumbing. The fundamentals of microbial management in drinking water systems are presented, including current practice and challenges faced by the sector. Finally, this chapter provides an overview of techniques applied to study microbial communities in drinking water, from sample collection to data analysis and interpretation. The aim of the chapter is to describe the system under study and provide a critical analysis of the different methods available to study it (both wet lab and dry lab techniques, their advantages and limitations), as a rationale for the selection of the methods applied in this research study.

Chapter 3 presents a collective analysis (or meta-analysis) of microbial communities in full-scale drinking water systems. The results were obtained after co-analyzing 14 publicly available data sets of full-scale drinking water systems that span three disinfection strategies (chlorination, chloramination, disinfectant residual-free systems), different types of source water, and sampling location points across the systems. The diversity of microbial communities in these disparate datasets and the variables that explain the

differences between datasets are discussed. Finally, some recommendations on the selection of sampling approaches and protocols are presented. Moreover, the meta-analysis provided valuable insights into the current state of the field in terms of achievements and limitations/challenges, which were considered when exploring the development of a predictive approach for microbial management in full-scale drinking water systems using DNA sequencing-based approaches.

Chapter 4 presents an assessment of the impact of methodological approaches on our observations of DW microbial communities. Specifically, the impacts of PCR replication, sampling replication, sample size (e.g. volume), and sample collection flow rate are explored, with a focus on microbial community richness, community membership and structure. The microbial community was characterized via amplification of the V4 hypervariable region of the 16S rRNA gene, and Illumina MiSeq sequencing. Furthermore, this replicate design was applied to elucidate the dynamics of the DW microbiome over small spatial and temporal scales. Finally, some recommendations on the use of replication and the selection of sample volumes and sampling flow rates in sampling campaigns are provided. These recommendations will allow to better capture DW microbial communities and generate good quality data from sampling efforts, which is highly important if the data is to be used to make decisions or design interventions and/or applications, or in the present case, to propose a management strategy with predictive elements.

Chapter 5 analyses the impacts of source water and treatment processes on the microbial community structure, membership and metabolic potential. Samples from finished water (at the treatment plant) and distribution system in ten drinking water systems that span a range of treatment processes and disinfection practices were collected, processed and subject to whole-genome shotgun sequencing (i.e. metagenomics). A range of analyses within and across systems is presented in order to elucidate aspects such as community richness, significant explanatory variables for community structure and membership, the impact of distribution, and similarities between the taxonomic (who is there?) and functional (what they can do?) profiles of the community. Using this data, a hypothesis is presented on the role of specific metabolic pathways of interest in microbial survival in disinfected DWDSs. The novel insights into both the taxonomic and functional profiles of DW microbial communities of a wide range of systems were applied to guide the proposed predictive framework in the selection of the suitable type of approach (e.g. same for all systems versus tailored), taking into account within and across system variability.

Chapter 6 provides a summary of the findings of this thesis, highlighting their contributions to the field. Moreover, current challenges still to overcome are discussed. Finally, recommendations on how to develop a predictive framework for microbial management in full-scale drinking water systems using DNA sequencing-based approaches are provided, based on the findings of this research and the current knowledge in the field.

1.4 Publications

Published Peer-Reviewed Manuscripts

- **Bautista-de los Santos, Q.M.**, Schroeder, J., Sevillano-Rivera, M., Sungthong, R., Ijaz, U., Sloan, W. and Pinto, A.J. (2016). Microbial communities in full-scale drinking water distribution systems – A meta-analysis. *Environmental Science: Water Research & Technology*. 2: 631-644. doi: 10.1039/C6EW00030D.

- **Bautista-de los Santos, Q.M.**, Blakemore, O., Schroeder, J., Moses, J., Haffey, M., Sloan, W. and Pinto, A.J. (2016). The impact of sampling, PCR, and sequencing replication on discerning changes in drinking water bacterial community over diurnal time-scales. *Water Research*. 90(1): 216-224. doi: 10.1016/j.watres.2015.12.010

- Pinto, A.J., Marcus, D. N., Ijaz, U.Z., **Bautista-de los Santos, Q.M.**, Dick, G.J. and Raskin, L. (2015). Metagenomic Evidence for the Presence of Comammox Nitrospira-like bacteria in a Drinking Water System. *mSphere*. 1 (1): e00054-15. doi: 10.1128/mSphere.00054-15

Manuscripts in preparation

- **Bautista-de los Santos, Q.M.**, Dai, Z., Calus, S., Sevillano-Rivera, M.C, Ijaz, U.Z., Sloan, W. and Pinto, A.J. Metagenomic assessment of the impact of source water type and treatment processes on drinking water microbial communities.

- Ugarcina Perovic, S*., **Bautista-de los Santos, Q.M*.**, Sevillano-Rivera, M.C., Sungthong, R., Ijaz, U.Z., Sloan, W. and Pinto, A.J. (2016). The impact of sample

volume and sampling flow rate on the structure and membership of drinking water bacterial communities (* joint first authorship).

- Sungthong, R., **Bautista-de los Santos, Q.M.** and Pinto, A.J. Exploiting predatory bacteria in drinking water innovations.

Conference presentations

- **Bautista-de los Santos, Q.M.**, Dai, Z., Calus, S., Sevillano-Rivera, M.C, Ijaz, U.Z., Sloan, W. and Pinto, A.J. "Metagenomic assessment of the impact of source water and treatment processes on drinking water microbial communities". IWA Specialized Conference: MEWE and Biofilms (September 4-7, 2016), Copenhagen, Denmark (oral presentation).

- Dai, Z., Sevillano-Rivera, M.C., **Bautista-de los Santos, Q.M.**, Ijaz, U.Z., and Pinto. A.J. (2016). "Elucidating the long-term impact of disinfection strategies on the drinking water microbiome". IWA Specialized Conference: MEWE and Biofilms (September 4-7, 2016), Copenhagen, Denmark (poster).

- Ugarcina Perovic*, S., **Bautista-de los Santos, Q.M.***, Sevillano-Rivera, M.C. and Pinto, A.J. "Estimating the impact of flow regime and sample volume on characterization of bacterial community diversity on a low biomass environment". Biofilms 7 Conference (June 26-28, 2016), Porto, Portugal (poster).

- **Bautista-de los Santos, Q.M.**, Schroeder, J., Sevillano-Rivera, M.C., Sungthong, R., Ijaz, U., Sloan, W. and Pinto, A.J. "Drinking water microbial communities across disinfection strategies". Scottish Water-EPSRC-University of Glasgow Conference: Achieving zero bacteriological failures in water supply systems (March 10, 2016), Glasgow, Scotland (oral presentation).

- **Bautista-de los Santos, Q.M.**, Schroeder, J., Sloan, W. and Pinto, A.J. "Meta-analysis of microbial communities in drinking water distribution systems". IWA Specialized Conference: Biofilms in drinking water systems (August 23-26, 2015), Arosa, Switzerland (poster).

- **Bautista-de los Santos, Q.M.**, Blakemore, O., Schroeder, J., Moses, J., Haffey, M., Sloan, W. and Pinto, A.J. "Diurnal variation of bacterial communities in drinking

water systems over small scales". Conference: 6th Congress of European Microbiologists (June 7-11, 2015), Maastricht, The Netherlands (poster).

- Ugarcina Perovic, S., **Bautista-de los Santos, Q.M.**, Sevillano-Rivera, M.C. and Pinto, A.J. "Determining the effect of sample volume and flow rates on investigations of bacterial community diversity in low biomass aquatic environments". Conference: 6th Congress of European Microbiologists (June 7-11, 2015), Maastricht, The Netherlands (poster).

- **Bautista-de los Santos, Q.M.**, Blakemore, O., Schroeder, J., Sloan, W. and Pinto, A.J. "Uncertainties associated with the characterization of bulk water bacterial communities in drinking water systems". Conference: AWWA Water Quality Technology Conference (November 16-20, 2014), New Orleans, USA (oral presentation).

Invited keynote presentations

- **Bautista-de los Santos, Q.M**., Calus, S., Dai, Z., Sevillano-Rivera, M., Sungthong, R., Ijaz, U., Sloan, W. and Pinto, A. "Understanding drinking water systems through molecular microbial ecology". Federation of Infection Societies (FIH) Annual Conference/10th Healthcare Infection Society (HIS) International Conference. Edinburgh, UK, 2016.

- **Bautista-de los Santos, Q.M**. and Pinto, A. "Biological drinking water treatment solutions". Drinking Water 2015: Developments in Water Quality, Treatment, & Distribution. Chartered Institute of Water and Environmental Management (CIWEM). London, UK, 2015.

2. Microbial aspects of Drinking water treatment and distribution

2.1. Overview of drinking water treatment and distribution

The aim of drinking water treatment and distribution is to produce and supply an adequate amount of potable and palatable water for the consumer. To assess the potability of the supplied water, its quality parameters are compared to the recommended health-based guidelines of relevant chemicals and microorganisms, which are usually embedded in the legislation of each country and enforced by a designated regulatory agency. In the case of drinking water, the World Health Organization (WHO) regularly produces guidelines on water quality and human health that subsequently influence the setting of water quality regulations and standards around the world. Aesthetical aspects are also taken into account to ensure palatability of the water for the consumer; for instance, colourless and soft water is preferred.

Drinking Water Systems consist of four broad components: (i) a catchment with an intake(s) where source water is collected from and transported to (ii) a centralized component (utility or plant), where the raw water passes through several treatment units that progressively remove contaminants from it until they reach an acceptable level/concentration; after treatment, the produced water enters the (iii) distribution system, a spatially large structure that conveys the water to the (iv) premises plumbing (also called building plumbing) area and until the Point of Use (POU) (Figure 2.1).

The raw water is the primary source of microorganisms that enter the drinking water system. Lowland surface water sources typically have higher microbial concentrations due to surface runoff from a variety of sources (e.g. cattle, farms, human activities); in contrast, the presence of microorganisms in ground water can be several orders of magnitude lower, allowing the production of drinking water with fewer treatment steps (e.g. aeration to remove reduced anions/cations and disinfection). Knowledge of the occurrence of and origin of pathogens in the source water is vital in the selection of the source and the appropriate treatment requirements to produce regulation compliant drinking water. For instance, seasonal fluctuations in source water quantity and quality can affect the quality of the drinking water produced, and should be taken into account in the design of the system. Catchment management is considered the first barrier of the treatment.

As part of a good catchment management strategy, the influence of land use on water quality should be assessed, taking into consideration several aspects (e.g. land cover modification; modification of waterways; livestock density and application of manure; residential development, with attention to wastewater and waste disposal, other potentially polluting human activities) (WHO 2011).

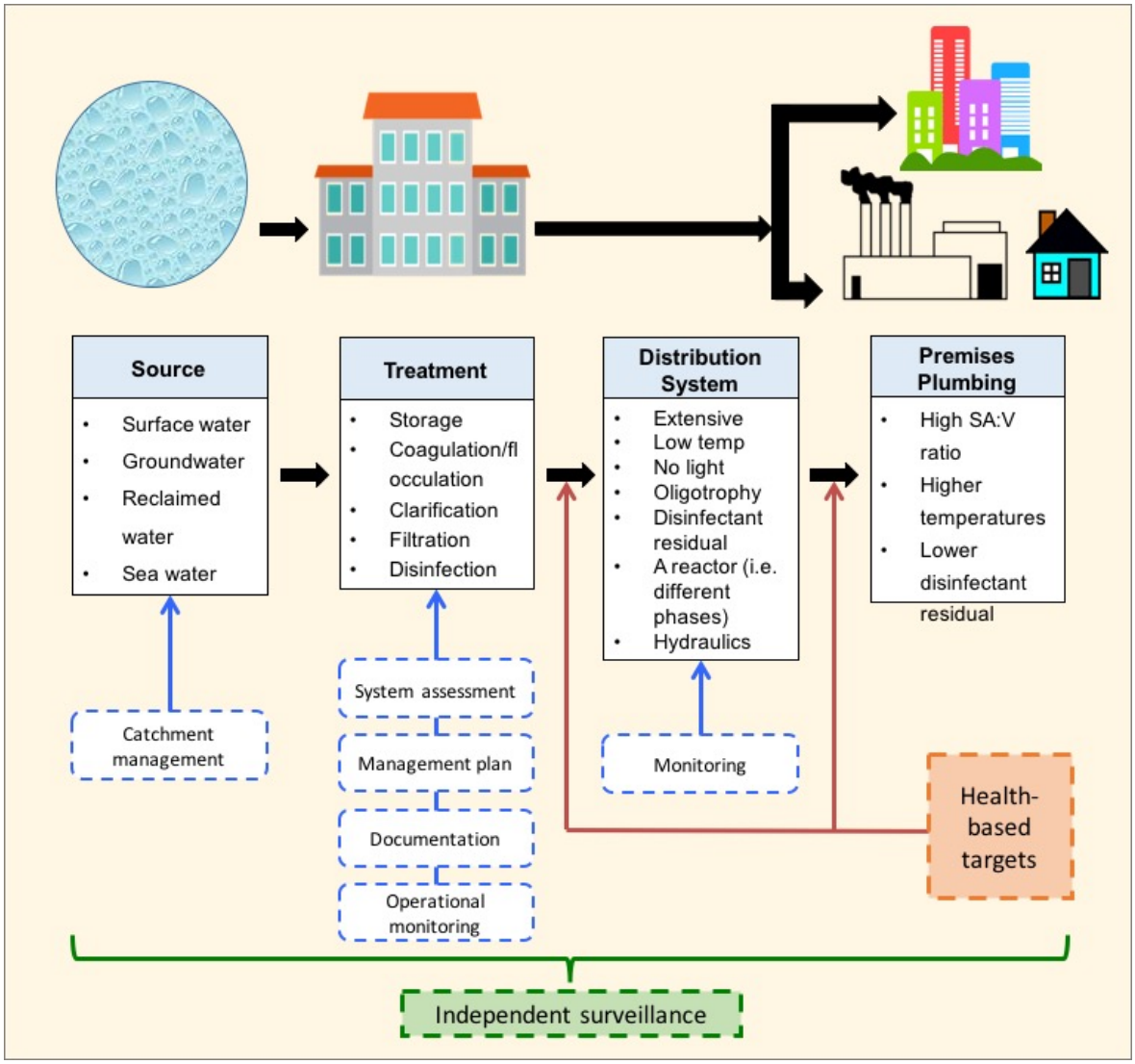

Figure 2.1. Components of a drinking water system.
Examples and characteristics of each component are provided. Elements of the safe drinking water framework are indicated by coloured dashed figures. SA:V – surface area to volume ratio; temp: temperature.

2.1.1. Treatment processes and microbial removal efficiency

The choice of water treatment processes typically depends on the type of source water that is being used. For instance, the typical treatment train for surface water sources (e.g. river, reservoir) consists of screening, coagulation-flocculation, clarification, filtration, disinfection and storage before distribution; while ground water may be ready for distribution with the application of softening (if required), filtration and disinfection. In

any case, the treatment train operates as a series of successive sieves, removing first the larger contents (e.g. debris, rocks) and subsequently removing smaller constituents from the source water until the desired reduction of microorganisms is achieved. Some general aspects of the microbial removal efficiency of treatment processes are discussed below:

- Pre-treatment processes: among these, the widely used storage reservoirs improve water quality through sedimentation of particles and ultraviolet (UV) radiation (if they are open reservoirs); reservoirs have shown log removal efficiencies of 1.4-2.3 for protozoans and 0.7-2.2 for bacterial. Bank filtration is a less used pre-treatment process, but more effective in the removal of protozoans (Log removal value – LRV ~1 to ~2) bacteria (LRV 2 to ~6) and viruses (LRV ~2.1 to 8.3) (WHO 2011). Chemical pre-treatment includes pre-chlorination for raw water with high coliform count, with doses of up to 5.0 mg l^{-1}, pre-ozonation and addition of powdered activated carbon to remove taste and odour compounds produced by algae, reduce colour and organics (Gray 2010).

- Coagulation/flocculation and sedimentation: these processes remove particles and the microorganisms attached to them by disrupting their charges, inducing particle agglomeration and settling. They have been shown to be efficient in the removal of bacteria (LRV 0.2-2) (OECD/WHO 2003) and protozoans (LRV 1-2 for conventional clarification, up to 2.8 for high-rate clarification). Moreover, optimal coagulation is important for the removal of *Cryptosporidium* oocysts and *Giardia* cysts (Emelko et al. 2005), due to their resistance against the last conventional treatment barrier (disinfection) (WHO 2011). Typical coagulants are aluminium sulphate (alum), ferric sulphate and ferric chloride, the former being the most efficient. Additionally, coagulation aids such as polyelectrolytes are added to promote floc formation.

- Filtration: is a physical process in which water is passed through a porous media in order to remove the particles left after sedimentation. The conventional granular filters used in DW treatment can be broadly classified as rapid gravity filters, pressure filters or slow sand filters. The first two have larger interstices, allowing the water to pass through them rapidly (e.g 5-10 m/h); while the latter employs a finer media (i.e. sand) allowing slower filtration rates (e.g. 0.08-0.15 m/h) but higher water quality of the effluent thanks to their biological activity (Gray 2010). Granular filtration can remove algae, bacteria and protozoans through several

mechanisms: physical straining removes particles larger than the filter interstices; particles smaller than the filter interstices are retained in the filter through transport mechanisms (e.g. diffusion, interception, sedimentation) (WHO 2004). Smaller bacteria and viruses can be removed by membrane filtration in its different modalities (e.g. microfiltration, ultrafiltration, etc.) (Table 2.1).

- Primary disinfection: is the application of a disinfectant during the treatment to inactivate microorganisms; the most affordable and widely used disinfectants are chlorine-based (chlorine, chloramine, chlorine dioxide). Microbial inactivation through disinfection is usually the last treatment step in a water utility, achieving the removal of bacteria (Table 2.2) and viruses, depending on the disinfectant used. Longer contact times are needed for the inactivation of some protozoans, while others like *Cryptosporidium* oocysts can pass the disinfection barrier without being inactivated (WHO 2011). In the case of chlorination, two species with disinfection properties are formed: hypochlorous acid and hypochlorite ions; the former is dominant at pH > 7.5 and is the preferred species as it is 80 times more effective than the latter (Gray 2010). In the case of chloramination, monochloramine formation (the most efficient disinfectant of the three chloramine species) is optimal at pH 7-8 and weight ratios of chlorine to ammonia between 3:1 and 5:1. Compared to free chlorine species (i.e. chlorination), combined chlorine species (i.e. chloramination) is more stable but requires longer contact times for the same degree of treatment. Other disinfectants used include UV and ozone, with the disadvantage that they are unable to provide a residual in the treated water that can protect against microbial ingress or regrowth in the system.

Type	Operating pressure (kPa)	Pore size (nm)	Permeability ($l/h.m^2.bar$)	Separation mechanism	Primary Applications
Microfiltration (MF)	30-50	100-1000	>1000	Sieving	Clarification; pre-treatment; removal of bacteria
Ultrafiltration (UF)	30-50	2-100	10-1000	Sieving	Removal of macromolecules, bacteria, viruses
Nanofiltration (NF)	500-1000	0.5-2	1.5-30	Sieving, charge effects	Removal of multivalent ions and relatively small organics
Reverse osmosis (RO)	1000-5000	<0.5	0.05-1.5	Solution, diffusion	Ultrapure water; desalination

Table 2.1. Major membrane filtration processes used in water treatment (Adapted from WHO 2004 and Van Der Bruggen et al. 2003)

Disinfectant	*Escherichia Coli*			Heterotrophic bacteria		
	pH	T (°C)	CT (mg.min/l)	pH	T (°C)	CT (mg.min/l)
Hypochlorous acid	6.0	5	0.04	7.0	1-2	0.08±0.02
Hypochlorite ion	10.0	5	0.92	8.5	1-2	3.3±1.0
Chlorine dioxide	6.5	20	0.18	7.0	1-2	0.13±0.02
	6.5	15	0.38	8.5	1-2	0.19±0.06
	7.0	25	0.28			
Monochloramine	9.0	15	64	7.0	1-2	94.0±7.0
				8.5	1-2	278±46.0

Table 2.2. Comparative efficiency of disinfectants for the achievement of 99% bacterial inactivation in oxidant demand-free systems.
T: temperature; CT: contact time (Adapted from WHO 2004).

2.1.2. The distribution system

After treatment, the water is transported from the plant through the distribution system, which may include miles of pipes, storage tanks, pumping stations, and connections, among other elements. The distribution system is characterized by its low temperatures, lack of light, very low concentrations of nutrients (i.e. oligotrophy), and commonly, the presence of a residual disinfectant (secondary disinfection, either chlorine or chloramine) to prevent microbial regrowth. Residual concentrations in disinfected DWDSs are typically maintained lower than the maximum guideline levels (chlorine: 5.0 mg l^{-1}; monochloramine: 3.0 mg l^{-1}) (WHO 2011). When secondary disinfection is not applied, the microbial regrowth control strategy relies on controlling growth-limiting nutrients in the system (typically Assimilable organic carbon-AOC), the typical target in these systems is to maintain AOC levels below 10 $\mu g l^{-1}$. Despite these measures, the microbial integrity of the DW in the distribution system can be compromised by two main factors: first is the integrity of the DWDS; for instance, microorganisms can ingress to the system through infiltration if there is a low pressure in the system, or during maintenance works if adequate disinfection and flushing protocols are not put in place. The second factor affecting water quality is microbial regrowth, as the DWDS is not a sterile environment and microorganisms can grow in it under favourable conditions (e.g. increased concentration AOC, higher temperatures, decay of residual disinfectant) (OECD/WHO 2003).

The DWDS is not a simple water conveyer but a "reactor" in which complex biological and physico-chemical processes take place (Camper & Dirckx 1996). Microorganisms can persist in the system in four phases: (i) planktonic, (ii) attached to suspended particles, (iii) attached to loose deposits and (iv) within the biofilms that cover the pipe walls (Figure

2.2). The bulk water and suspended solid phases are the phases in direct contact with the consumer, and are relatively easier to access and study. The biofilm and loose deposits phases harbour more diverse and abundant communities (Liu, Ling, et al. 2013); nevertheless, they are more difficult to sample representatively due to the limited access to the samples and the high variability of the communities in this phases (Henne et al. 2012). The DW biofilm, as any biofilm, is subject to dynamics of initial conditioning, attachment, growth and detachment. Should the conditions be favourable (e.g. higher concentrations of growth limiting nutrients, higher temperatures, stagnation, low disinfectant residual concentrations), microbial regrowth can occur in the system. Corrosion can also occur in systems under specific conditions (e.g. chemical- low pH, microbiological-microbially influenced corrosion), causing an undesirable release of metals (e.g. copper, iron, lead) into the bulk water due to their adverse health effects. Leaching of organic synthetic compounds from pipe walls, such as Bisphenol A, an endocrine disruptive compound, has also been reported, with increased leaching after incubation and due to higher temperatures (Rajasärkkä et al. 2016). A comprehensive review made by Prest and colleagues (2016) provides details of a range of factors affecting microbial dynamics in DWDSs.

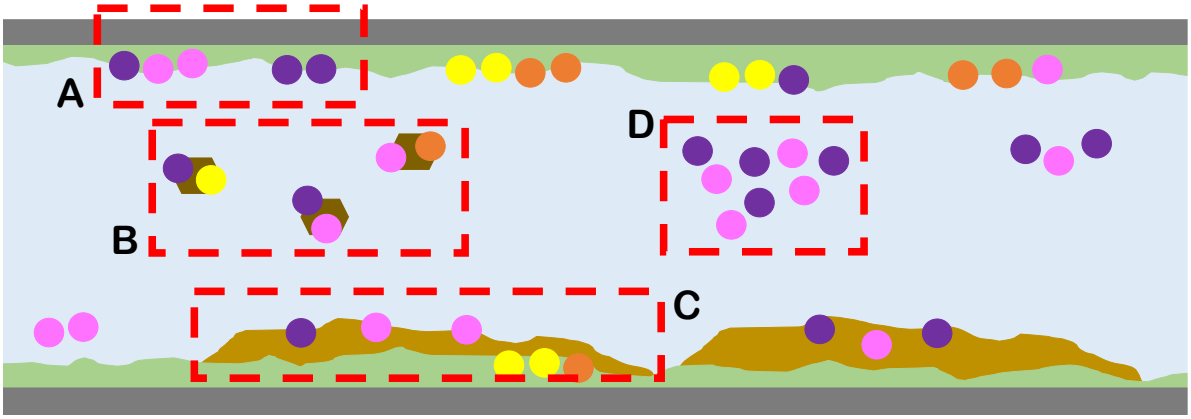

Figure 2.2. Different microbial phases in drinking water pipes. Coloured circles indicate bacterial cells. Legend: (A) Biofilm; (B) suspended solids; (C) loose deposits; (D) planktonic cells in bulk water. (Adapted from Camper & Dirckx (1996) and Liu et al. (2013))

2.1.3. Premises plumbing

Premises plumbing or building plumbing is the portion of the distribution system from the water meter to the consumer's tap, belonging to the house/building owner. Numerous features differentiate premises plumbing from the distribution system:

- Physical properties: due to the smaller pipe diameters, premises plumbing has higher surface area to volume ratio compared to the distribution system,

approximately 10 times more surface area per unit length than the distribution system. For instance, in a system in Columbia, Missouri, household plumbing and service connections had 82% of the total pipe length, 24% of the total surface area and held only 1.6% of the volume of water in the system (PNAS 2006). In addition, numerous elements are present in premises plumbing (e.g. pipes, connectors, valves) made of several materials (e.g. copper, plastic, brass, lead, stainless steel, galvanized iron) that constantly change the direction of the flow. Finally, additional treatment units (e.g. point of use filters) may be found in households, which affect the water quality and could impact the DW microbiome.

- Hydraulic conditions: start-stop flow patterns observed in premises plumbing could detach scale and biofilm from the pipe surfaces, with potential adverse effects. Additionally, water can sit stagnant for 6-8 hours (night time) in residential dwellings, or for longer periods in properties that are irregularly occupied (e.g. schools, vacation homes). This means that water in a property will have a wider distribution of water age, resulting in a greater variation of disinfectant residual levels and potential bacterial regrowth (PNAS 2006).

- Environmental conditions: in premises plumbing, the disinfectant residual is typically lower than in the distribution system. Moreover, the water temperature varies more in premises plumbing than in the distribution system; for instance, in the summer months the cold water line can be 10-15 ^{o}C warmer than the mains water (PNAS 2006).

2.2. Microbial management in drinking water distribution systems

2.2.1. Safe drinking water framework – microbial aspects

A framework for the delivery of safe drinking water has been proposed by the WHO (2011), consisting of: (i) health based targets, (ii) water safety plans, and (iii) independent surveillance (Figure 2.3). When ranking health-based targets, the microbial aspect is a priority as the first requirement is to ensure the supply of microbiologically safe water, followed by the management of chemical hazards that cause fast adverse health effects from short-term exposure. Since the main aim is to supply potable water, the first task is to define what "potable" is. In the case of microbial water quality, this is achieved by establishing a link between microorganisms and human health through epidemiology or

microbial risk assessment. If the microorganism in question causes an adverse effect on human health it is considered pathogenic, and the results from the epidemiology or microbial risk assessment must be translated to a performance target or technology target to assess if treatment has been efficient in reducing the risk of exposure to an acceptable pre-established level.

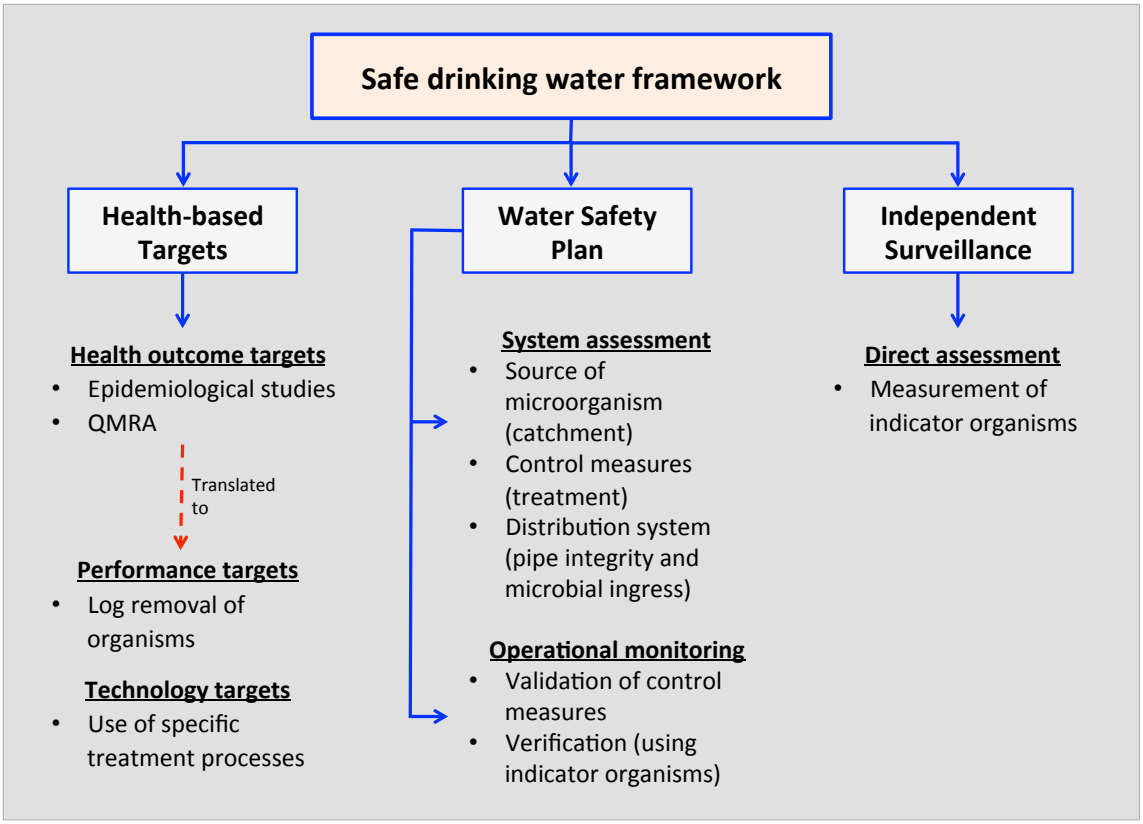

Figure 2.3. Overview of safe drinking water framework, with indications of microbial aspects in each component (Adapted from WHO 2011).

A Water safety plan (WSP) is a risk assessment and management approach comprising all the components in a water system, from the source to the tap (WHO 2011). It is applied by the water utility, and consists of three main elements: (i) system assessment, (ii) operational monitoring and (iii) management plans, documentation and communication. Microbial aspects of a WSP include the assessment of sources of microorganism throughout the system (i.e. catchment, water treatment plant, distribution system) and the identification and application of control measure (through treatment) to reduce the microbial load in the water and minimize/avoid microbial ingress to the distribution system. The performance of the control measures (i.e. treatment steps) is verified through operational monitoring, which includes the detection and enumeration of indicator organisms (e.g. coliform bacteria). Finally, a management plan detailing the operation of the plant in different circumstances (e.g. “normal” operation versus “incident” condition, in

case an upper limit of a certain parameter has been exceeded) should be elaborated, a description of all the components of the system (e.g. assets) should be documented, and communication strategies should be outlined and applied to exchange information within the water utility, and among other stakeholders (e.g. consumers, regulator, etc.).

An independent surveillance scheme must be in place to oversee the complete process. There are two types of approaches for surveillance: audit-based approaches that only oversee the activities of the water utilities; and direct assessment approaches that verify the efficiency of the process by direct measurement (e.g. taking water samples to test for microbial compliance). A combination of both approaches is also possible and depends on the structure of the system and the roles and attributes of each stakeholder. When assessing the adequacy of the water supply, a surveillance scheme must assess the quality, quantity, accessibility, affordability and continuity of the service provided (WHO 2011).

2.2.2. Microbial assessment of drinking water quality

Due to the fact that pathogen detection and enumeration in DW is expensive and time consuming, the assessment of microbial contamination in drinking water relies on the detection and enumeration of indicator organisms, which helps translate health outcome targets to performance targets (Table 2.3). These indicator organisms are usually non-pathogenic bacteria that are present in large amounts in faeces of warm-blooded animals, and their detection and enumeration is relatively easy and inexpensive (WHO 2011). A distinction in the indicator organisms is made according to the type of intended monitoring: (i) for process validation, heterotrophic plate counts (HPC) indicate the effectiveness of disinfection of bacteria, *Clostridium perfingens* indicates the effectiveness of disinfection of viruses and protozoa, phages (e.g. coliphages, phages of *Bacteroides fragilis*) and enteric viruses indicate the effectiveness of disinfection of viruses; (ii) for operational monitoring, total coliforms indicate the cleanliness and integrity of the distribution system, while HPC indicate the effectiveness of the disinfection process and the cleanliness and integrity of the distribution system; (iii) for verification and surveillance, *Escherichia coli* and thermotolerant coliforms are used as faecal indicators (WHO 2011). A combination of parameters can be applied to make the microbial assessment more robust; for instance, in Scotland, in addition to *E. coli*, *Enterococci* are monitored for the protection of human health (Scottish Parliament 2014), while *Cryptosporidium parvum* has also been monitored in raw and treated waters (Scottish Parliament 2003). In addition to the microbial indicator organism, some physico-chemical

parameters are also used as indicators of microbial quality. For instance, an increase in turbidity is associated with an increase in microorganisms in the water, with turbidity being negatively correlated with total coliform removal efficiency in chlorinated systems (LeChevallier et al. 1981); the common turbidity target for finished drinking water is in the range of 0.1-1.0 NTU (OECD/WHO 2003), while in the UK the target is 1.0 NTU in the treatment works and 4.0 NTU at the point of use. Another common non-microbial indicator is disinfectant residual concentration, as departures from the established target range could indicate microbial regrowth in the system.

Parameter	A. Assay characteristics							B. Applicability and suitability								
	Association with faecal matter	Risk to analyst	Speed of measurement	Cost	Technical difficulty	Survival in the environment	Resistance to Treatment	Sanitary survey	Source water characterisation	Groundwater characterisation	Treatment efficiency (removal)	Treatment efficiency (disinfection)	Treated water	Distribution system (ingress)	Distribution system (regrowth)	Outbreak investigation
Total coliforms	NA	L	M	M	M	M	L	NR	NR	NR	NR	**SA**	**S**	**SA***	**S**	**S**
Thermotolerant coliforms	M	M	M	M	M	M	L	**SA**	**SA**	**SA**	NR	**SA**	**SA**	**SA***	**S**	**S**
Escherichia coli	**H**	M	M	M	M	M	L	**S**	**S**	**S**	**S**	**S**	**SA**	**S***	NA	**S**
Faecal streptococci	M	M	M	M	M	M	ISD	**SA**	**SA**	NA	NA	NA	NA	NA	NA	**S**
Total bacteria (microscopic)	NA	NA	**H**	M	M	**H**	**H**	NA	NA	NA	**SA**	**SA**	NA	**SA**	**S**	**S**
Viable bacteria (microscopic)	NA	NA	M	M	M	**H**	M	NA	NA	NA	**SA**	**SA**	NA	**SA**	**S**	**S**
Total bacteria (automated)	NA	NA	**H**	**H**	M	**H**	**H**	-	-	-	-	-	-	-	-	-
Viable bacteria (automated)	NA	NA	**H**	**H**	M	**H**	M	-	-	-	-	-	-	-	-	-
Heterotrophic bacteria	NA	L	M	M	M	**H**	**H**	NA	NA	NA	**S**	**S**	NR	**S**	**S**	**S**
Aerobic spore-forming bacteria	NA	L	M	M	M	**H**	**H**	NA	NA	NA	**S**	**S**	NR	NA	NA	**S**
Somatic coliphages	ISD	M	**H**	M	M	**H**	M	**SA**	**SA**	**SA**	NA	**SA**	NA	NA	NA	**S**
F specific RNA phages	ISD	M	**H**	M	M	**H**	**H**	**SA**	**SA**	**SA**	NA	**SA**	NA	NA	NA	**S**
Bacteroides phages	ISD	M	**H**	M	M	ISD	**H**	**SA**	**SA**	**SA**	NA	**SA**	NA	NA	NA	**S**
Sulphite-reducing clostridia	NA	L	M	M	M	**VH**	**VH**	NR	NR	NR	NA	NA	NA	NA	NA	**S**
Clostridium perfingens	**H**	L	M	M	M	**VH**	**VH**	**SA**	**SA**	**SA**	**SA**	NA	NA	NA	NA	**S**
Pseudomonas, *Aeromonas*	NA	M	M	M	M	**VH**	L	NA	NA	NA	NA	NA	NA	NA	S	NA
Enteric viruses	NA	**H**	L	**H**	**H**	**H**	**H**	**S**	**S**	**S**	NR	NR	NA	NA	NA	**S**
Giardia cysts	NA	**H**	L	**H**	**H**	**H**	**H**	**S**	**S**	**S**	NR	NR	NA	NA	NA	**S**
Cryptosporidium oocysts	NA	**H**	L	**H**	**H**	**VH**	**VH**	**S**	**S**	**S**	NR	NR	NA	NA	NA	**S**

Table 2.3. Microbial parameters and (A) assay characteristics and (B) applicability and suitability. L: low; M: medium; H: high; VH: very high; ISD: insufficient data; NA: not applicable; S: suitable; SA: suitable alternative; NR: not recommended; *In distribution systems without residual disinfection (Adapted from OECD/WHO 2003).

The reliability of some of the indicators used in DW has been questioned; for instance, Byappanahalli and colleagues (2006) showed that *E.coli* can occur and persist in temperate soils and that it was phylogenetically different to faecal strains. Moreover, in the case of reclaimed water, no single indicator organism (total and faecal coliforms, *Enterococci*,

Clostridium perfringens, and F-specific coliphages) correlated with the pathogens monitored (infectious enteric viruses, *Cryptosporidium*, and *Giardia*), while all indicators together could predict the presence/absence pattern of the pathogens in over 71% of the effluents tested (Harwood et al. 2005). These limitations in reliability of the indicator organisms currently in use have motivated the search for alternative indicators and the importance of not only detecting and quantifying faecal pollution in DW, but to track its sources (Field & Samadpour 2007).

In order to assess the risks associated with the microorganisms in DW, Microbial risk assessment (MRA) is the formal scientifically based process to estimate the likelihood of exposure to a microbial hazard and the resulting public health and/or environmental impact from this exposure (EPA/USDA 2012). MRA was designed as tools to support decision-making (e.g. to mitigate, to confirm, to decide how/whether to regulate, etc.) and depending on the needs of the stakeholders, several forms of assessment can be carried out (e.g. screening, risk ranking, risk-risk, product pathway analyses, etc.) (EPA/USDA 2012). Within the framework there is some flexibility on the tools to apply depending on the objective of the analysis (which must be clearly stated in the problem formulation), but core principles of transparency, clarity, consistency and reasonableness (TCCR) should be fulfilled regardless of the type of analysis made. MRA consists of five main steps:

(i) Hazard identification and characterization: in broad terms the hazard is the subject of the assessment which is associated with an adverse health outcome. When the MRA is focusses specifically on pathogens (commonly), the hazard is the pathogen's potential to cause adverse health effects. Its characterization is broad and can include several aspects such as infectivity, invasiveness, virulence, etc.

(ii) Exposure assessment: aims to estimate the magnitude, frequency, duration and route of exposure of a target organism. Since most MRA studies usually cover only the oral route of exposure (due to the severity of waterborne gastrointestinal diseases), two litres/person-day is used to estimate drinking water exposure (Haas & Eisenberg 2001). The assessment of pathogens in drinking water is challenging because of the low numbers present after treatment; therefore, these values are usually substituted by an estimation of the concentration remaining in DW after treatment, based on the quantification of

raw-water microorganism levels, or by a surrogate organism (Haas & Eisenberg 2001).

(iii) Dose-response assessment: relationship between the dose of the microorganism ingested and the probability of infection. Two distributions are commonly used to fit the data, the exponential and the Beta Poisson distributions. Since the desired probability of infection (and dose) for the protection of public health is lower than what can be directly measured experimentally, it is often necessary to extrapolate a fitted dose-response curve in the low dose region (Haas & Eisenberg 2001). It is therefore important to discuss the assumptions made in making those extrapolations, along with other details of the selected model (e.g. assumptions, limitations, methods of assay).

(iv) Risk characterization: combination of exposure and dose-response into an estimation of the magnitude of the risk, defined as the likelihood of the identified hazard causing adverse health effects to the population considered. The risk characterization includes a discussion and quantification of the uncertainties associated with the analysis, the variability associated with key inputs to the model, the confidence in the risk estimates through a weigh-of-evidence discussion, the limitations of the analysis and the plausibility of the results (Schaub 2014).

Several options are available to conduct microbial risk assessments in drinking water systems; their use depends on the context, the type of utility (i.e. small, large) and the resources available (e.g. human resources). The simplest approach for small supplies is a sanitary inspection, which is a visual evaluation of observable features at or near the water supply that could compromise water integrity. For a more comprehensive evaluation, a risk matrix can be prepared, as a qualitative or semi-quantitative evaluation of the likelihood of a hazard and the severity of its consequences which combined result in a risk score. Finally, the most formal approach to microbial risk assessment is Quantitative Microbial Risk Assessment (QMRA), which is based on scientific knowledge of the presence of pathogens in DW, their fate and transport in DW systems, the routes of exposure to humans and the consequences of this exposure (WHO 2016). QMRA is embedded in the World Health Organization's guidelines on DW; moreover, it is also central in food safety management. Extensive documentation is available on MRA and QMRA (Centre for Advancing Microbial Risk Assessment, accessed 27-01-2017), the latest document

released by the WHO with focus on DW systems provides case studies and detailed descriptions of the aforementioned methodologies and how to implement them (WHO 2016)

2.2.3. Emerging issues

Opportunistic Pathogens (OPs) are microorganisms that pose a health risk for certain groups of individuals (e.g. elderly, immunocompromised, etc.). Unlike waterborne pathogens, opportunistic pathogens are natural inhabitants on drinking water systems, do not correlate with faecal coliform numbers, and increase in number as the distance from the plant increases. They have been identified as a risk in premises plumbing due to their characteristics (e.g. higher resistance to disinfectants, biofilm formation, survival and growth in free-living amoebae) (Falkinham et al. 2015). Similar to other water pathogens, OPs are not routinely monitored by utilities; moreover, the identified niche for OPs is premises plumbing because of its characteristics, and this final portion of the network is under the responsibility of the owner and not the water utility. Since the issue of OPs is of wide interest both within and outside the drinking water industry, the data available on presence of OPs in water systems often comes from research projects and outbreak investigations (e.g. samples are taken at the POU in residences or in hospitals, where a high number of individuals at risk is present); further, epidemiological studies provide an overview of prevalence of diseases caused by OPs, although they are unable to elucidate if the source of the pathogen was drinking water. Collier et al. (2012) (cited by Falkinham et al., 2015) reported that in the United States the costs of the estimated 29,636 cases of OP diseases per year is approximately US$850 million. Among the numerous OPs identified, three organisms are of special interest in drinking water due the incidence of the diseases they cause: *Legionella pneumophila*, *Mycobacterium avium* and other non-tuberculous Mycobacteria (NTM) and *Pseudomonas aeruginosa.*

Legionella pneumophila is the main cause of Legionnaire's disease (pneumonia) or legionellosis, and Pontiac fever, both respiratory infections, the former a severe illness and the latter a milder influenza-like disease (WHO 2007; Shen et al. 2015). Contaminated aerosols are thought to be the primary mode of transmission of *L. pneumophila* to humans (WRF 2013). These bioaerosols inhaled by the host reach the alveolar region where *L. pneumophila* replicates and infects the host. Identified transmission elements include shower heads and humidifiers, both with water temperatures in the range favourable for *L. pneumophila* growth (25-45oC, optimal range 32-42oC) (WHO 2007) and able to produce

aerosols. *L. pneumophila* may also prefer growth in biofilms than planktonic state, as it is ubiquitous in oligotrophic aqueous environments but it is very difficult to grow as a pure culture under lab conditions (Declerck 2010). A positive correlation between *L. pneumophila* attachment and biofilm roughness was recently shown, due to the increased interception between the flowing bacteria and the biofilm surface (Shen et al. 2015). Moreover, *L.pneumophila* also interacts with free-living amoebae, increasing in drinking water biofilms in the presence of *Acanthamoeba castellanii* as it replicates within it, which provides it further protection from disinfectant exposure (Declerck et al. 2009). Attention has been given to the increase in cases of legionellosis due to its severity. In the United states, the estimated number of hospitalized cases of legionellosis from 2000 to 2009 was between 8,000 and 18,000; furthermore, the number of reported cases increased 3.5-fold between 2000 and 2011 (Falkinham et al. 2015). According to the European Centre for Disease Prevention and Control (2016), Legionnaires' disease remains an uncommon, mainly sporadic respiratory infection with low notification rates in EU/EEA countries. In 2014, 30 countries reported 6,943 cases, of which 6,412 (92.4%) were classified as confirmed. The remaining 531 (7.6%) cases were reported as probable. Five countries out of 31 (France, Germany, Italy, Portugal and Spain) accounted for 74% of all notified cases. Nonetheless, the number of notifications per 100,000 inhabitants was 1.4 in 2014, which was the highest ever observed. Majority of the cases (69%) were community-acquired, while 20% were travel-associated; 8% were associated with healthcare facilities, and 3% were associated with other settings. The highest notification rate (per 100,000) among the 30 countries included was 5.6 in Portugal, mainly due to the large community outbreak that occurred in Vila Franca de Xira near Lisbon in October and November 2014. Of 5,505 cases with known outcome, 456 were reported to have died, equivalent to a case fatality of 8%.

Non-tuberculous Mycobacteria (NTM) usually cause pulmonary infections (although extrapulmonary infections can also happen) in immunocompromised adults and children. Among NTM, members of the *Mycobacterium avium* complex (MAC) are the most common group identified in isolations from pulmonary specimens in Australia, the Netherlands and East Asia; *M. intracellulare* and *M. abscessus* were more frequently associated with pulmonary disease than *M. avium* in Seoul, South Korea (WRF 2013). Transmission of *M. avium* is through inhalation and ingestion; being ubiquitous, its environmental sources are multiple (e.g. drinking water pipelines, water tanks, hot tubs, residential water taps, hospital water taps and ice machines, bottled water, showerheads, shower aerosols, among others) (Halstrom et al. 2015). Inhalation of aerosols seems to be

the primary transmission route of NTM, usually occurring in artificial water environments (e.g. showers and hot tubs). Similarly to *L. pneumophila*, MAC can also maintain long-term contamination of drinking water through its association with biofilms and intracellular parasitism of free-living amoebae (Whiley et al. 2012). On the epidemiology of NTM, Johnson and Odell 2014 indicate that "*over the last three decades, it has been suggested that the incidence of both NTM laboratory isolation and disease prevalence is increasing. This change has been attributed, in part, to improved culturing techniques, coupled with greater disease awareness and a true increase in disease prevalence. However, it is challenging to accurately characterize the incidence and prevalence of NTM pulmonary infections since isolation of the organism does not universally indicate clinical infection.*" The prevalence of NTM pulmonary infections ranges between 4.1 and 14.1 per 100 000 patient years (Kendall & Winthrop 2013; cited by Johnson & Odell 2014); in the United States, prevalence increases with age (e.g. in elderly patients older than 65 years a prevalence of 47 per 100 000 patient years was observed), and women are also more likely to have NTM disease than men (Adjemian et al. 2012; cited by Johnson & Odell 2014). In Germany, Ringshausen et al. (2016) found an increasing trend in prevalence of NTM pulmonary disease over the period 2009-2014, from 2.3 to 3.3 cases/100,000 population, while no differences in prevalence were observed relative to the sex of the patient. An assessment of pulmonary disease caused by NTM in 5 European countries (United Kingdom, France, Germany, Italy, and Spain) reported that annual prevalence was uniform and ranged between 5.9 (Spain) to 6.5 (United Kingdom)/100,000 population; nonetheless, some variation corresponding to regional differences was observed within countries (e.g. France, 1.3 in Southwest versus 13.6 in Paris/100,000) (Wagner et al. 2014).

Pseudomonas aeruginosa is a ubiquitous bacterium causative of infections in hospital settings (pulmonary infections, ventilator-associated pneumonia, septicaemia, urinary tract infections, surgical wounds infection) and community settings (ear, eyes and skin infections associated with the use of recreational water) (WRF 2013). The modes of transmission of *P. aeruginosa* include direct contact with water and aerosols, aspiration and indirect transfer from moist environmental surfaces (WRF 2013); among these, skin exposure in hot tubs and lung exposure from inhaling aerosols carry the greatest health risk (Mena & Gerba 2009). *P. aeruginosa* can grow at very low nutrient levels and although it is more resistant to chlorine than *E. coli* ($CT_{99.9\%}$ for water adapted cells is 40-fold higher) (J. Falkinham et al. 2015), it does not exhibit a marked resistance against common disinfectants to treat drinking water (i.e. chlorine, chloramines, ozone) (Mena & Gerba 2009). Unlike *L. pneumophila* and MAC, *P. aeruginosa* is shown to be antagonistic

towards biofilm-associated amoebae (*Acanthamoeba castellanii*) (Matz et al. 2008). Community acquired infections of *P. aeruginosa* involve the use of swimming pools, hot tubs and whirlpools in which some kind of maintenance failure has been detected (Hlavsa et al. 2014; cited by Falkinham et al. 2015). Moreover, *P. aeruginsa* is a major causative of otitis externa ("swimmer's ear"), with 2.4 million cases of this disease per year and an estimated outpatient cost of US$500 million, approximately (CDC Report 2011; cited by Falkinham et al. 2015). In health-care settings in the United States, a meta-analysis of 43 water-related outbreaks covering 35 years (1966-2001) estimated that *P.aeurginosa* caused approximately 1,400 deaths through pneumonia infections (EJ et al. 2002; cited by Falkinham et al. 2015). Furthermore, *P. aeruginsa* is the second most frequent cause of ventilator-associated pneumonia, and the third or fourth most frequent cause of septicemia, urinary tract infections and surgical wound infections (Trautmann et al. 2006; cited by Falkinham et al. 2015).

Antimicrobial resistance constitutes a threat to the prevention and treatment of an ever-increasing range of infections caused by bacteria, parasites, viruses and fungi (WHO 2015). The presence of antibiotic resistant bacteria in drinking water has been reported since the 1980s (Armstrong et al. 1981) but the development and application of DNA-sequencing and PCR based approaches have allowed deeper insight into the subject. For instance, the vancomycin resistance gene for *Enterococci* (*van*A) and the β-Lactam resistance gene for *Enterobacteriaceae* (*amp*C) have been detected in DNA extracted from drinking water biofilms collected from full-scale systems (Schwartz et al. 2003). Moreover, the treatment processes have an impact on the antibiotic resistance rate, as an increase in the rate was observed as the water passes each treatment step (source to finished water, including pre-ozonation, filtration and chloramination); the treated water showed the highest antibiotic resistance for the five antibiotics tested (Ampicillin-AMP, Kanamycin-KAN, Rifapentine-RFP, Chloramphenicol-CM and Streptomycin-STR) (Bai et al. 2015).

2.3. Techniques for microbial community analysis

2.3.1. Community characterization

For characterization purposes, two major components of microbial communities are of interest: biodiversity and microbial activity; to study the former one must identify and quantify microorganisms in their habitat; to study the later one must measure metabolic

processes that microorganisms carry out in their habitats (Madigan et al. 2012) To identify and enumerate microorganisms in their habitat, culture-based techniques have been applied for over 100 years. The enrichment culture technique conceptualized by Beijerinck in the beginning of the 20th century to isolate the nitrogen-fixing bacterium *Azotobacter* has been applied to hundreds of other microorganisms and conditions; moreover, the enrichment culture technique combined with isolation techniques (e.g. streak plate, liquid dilution) allow the isolation of single organisms (or pure cultures) that can then be further characterized. Nevertheless, these culture-based approaches introduce a bias in their outcome, since the physical and chemical conditions of natural habitats are difficult to replicate in the laboratory and the culture conditions of the enrichments favour rapidly growing microorganisms which may not be the most abundant or ecologically relevant (Madigan et al. 2012). The development of culture-independent approaches and their application to microbial ecology further confirmed the bias introduced by cultivation. It is estimated that only 1% of the total bacterial population is culturable (Amann et al. 1995), therefore the vast majority of bacteria are still unexplored, since the remaining 99% of the population cannot be studied using classical cultivation techniques. The field of metagenomics has emerged along with the development of DNA sequencing, as the study of genetic material from environmental samples applying DNA sequencing. Two approaches are common in molecular microbial ecology: single-gene amplification and sequencing, and shotgun sequencing. In single-gene amplification studies the genetic material is extracted and PCR is used to amplify one gene, usually the 16S rRNA gene which is a phylogenetic marker; the community is then characterized based on its taxonomy, the structure (i.e. abundance) and membership (i.e. presence/absence) of its members, and their relationship with the environmental variables. In shotgun sequencing the extracted genetic material is fragmented and directly sequenced without amplification (or with a reduced number of amplification cycles); the sequenced genes can be used for both taxonomic and functional annotations, thus providing a deeper understanding of the community.

DNA-sequencing based approaches, although advantageous, are not exempt from their own biases and limitations (Brooks et al. 2015). For instance, a shift in microbial composition due to DNA extraction bias has been reported, caused by incomplete DNA extraction from soil samples that distort the relative abundance of bacterial phyla in the community (Feinstein et al. 2009). Moreover, Guo and Zhang (2013) tested seven commercial DNA extraction kits to assess their efficiency with activated sludge samples and reported that cell lysis and bead beating significantly impacted the DNA yield and the

bacterial community composition and structure by changing the abundance profile of major bacterial phyla; while Albertsen et al. (2015) further explored Guo and Zhang's findings and concluded that a bead beating duration of four times the normal duration (i.e. 4x 40 seconds at 6 m/s) produced DNA with sufficient integrity and that once sequenced was representative of the original bacterial community. PCR is a recognized source of bias when applied to environmental communities for several reasons: inhibition of amplification by co-extracted contaminants; differential amplification (i.e. all templates are not amplified with the same efficiency, affecting the template ratio of the initial community); formation of artefacts (i.e. erroneous sequences are generated during the process that suggest the existence of novel organisms); contaminating sequences can get amplified as well, distorting the original composition of the community; finally, the number of 16S rRNA gene copies on prokaryotes varies greatly (e.g. *Bradyrhizobium japonicum* has one while ten have been reported in *Bacillus subtilis*), further introducing bias in the amplification process (Wintzingerode et al. 1997; Kalle et al. 2014).

Combinations of both culture-dependent and culture-independent approaches have been proposed and applied. Lagier et al. (2015) have combined metagenomics and culturomics (the diversification of culture conditions together with matrix-assisted laser desorption ionization–time of flight mass spectrometry [MALDI-TOF MS], to increase the bacterial repertoire) to study the human gut using 212 different culture conditions; 32, 500 colonies were obtained by culturomics, yielding 340 species of bacteria from seven phyla and 117 genera. This approach allowed the isolation of a giant bacterium, *Microvirga massiliensis*, with a diameter of 2.28 um and the largest genome (9.35 Mb) of any bacterium previously obtained from a human sample; further, the identification of 174 species never described previously in the human gut was achieved. Functional metagenomics have been proposed as a complement to sequence-based metagenomics; the technique involves isolating and cloning environmental DNA fragments, expressing genes in a surrogate host and screening for enzymatic activity, allowing the discovery of enzymatic activity which cannot be achieved using only DNA sequences (Lam et al. 2015). The technique has been applied in over twenty environments including host-associated, extreme and engineered environments; for instance, Vercammen et al. (2013) found a new type of class A beta-lactamase from a metagenomic library of a polluted stream sample in Belgium.

In addition to classical cultivation methods and DNA sequencing methods, several other techniques are available to study microbial communities (Figure 2.4). Extensive literature is available on these methods, including review papers (e.g. Rastogi & Sani 2011), plus a

recent comprehensive review (Douterelo et al. 2014) gives an overview of the ones applied specifically in DW microbial ecology. The selection of the method to apply depends on the questions to be answered and the resources available to the researcher. A short description of the currently most useful and some of the most promising methods for the study of the entire microbial community, and some examples of their applications are given below:

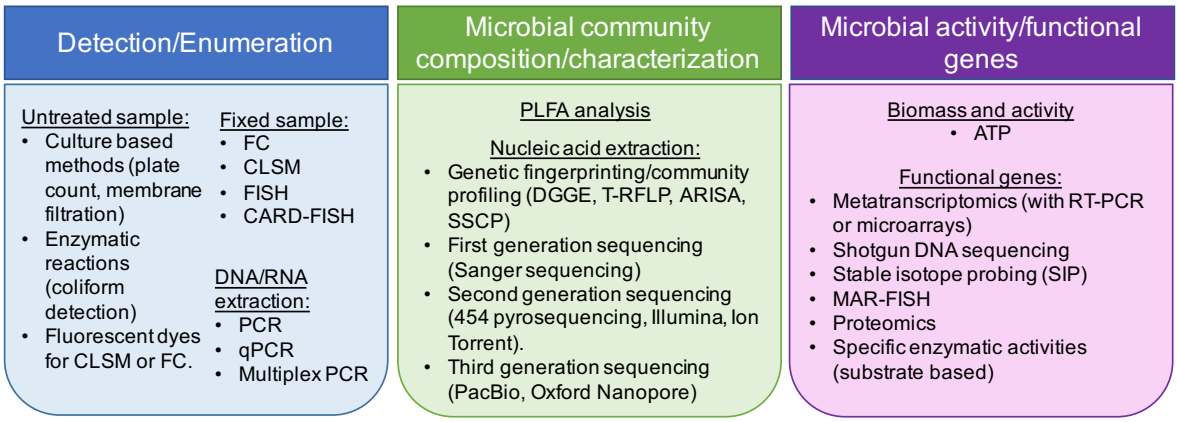

Figure 2.4. Available techniques to study the microbial ecology of drinking water systems FC: flow cytometry; CLSM: confocal laser scanning microscopy; FISH: fluorescence in situ hybridization; CARD-FISH: catalyzed reporter deposition fluorescence in-situ hybridization; PCR: polymerase chain reaction; qPCR: quantitative polymerase chain reaction; PLFA: Phospholipid-derived fatty acids; DGGE: denaturing gradient gel electrophoresis; T-RFLP: Terminal Restriction Fragment Length Polymorphism; ARISA: Automated ribosomal intergenic spacer analysis; SSCP: single-strand conformation polymorphism; ATP: adenosine triphosphate; RT-PCR: reverse transcription PCR; MAR-FISH: microautoradiography-fluorescence in situ hybridization (adapted from Douterelo et al. 2014).

- DNA sequencing: its goal is to identify the nucleotides that compose a DNA molecule in their correct order. It was first introduced in the 1970's with the development and application of Sanger sequencing, called first generation sequencing. Sanger sequencing applies the 'chain termination method' in which a DNA template is subject to amplification (either in vivo cloning or PCR amplification) and then added to four parallel reactions each containing the four deoxynucleotides (dNTPs) and one marked dideoxynucleotide (ddNTP) that prevents chain elongation; the template then elongates by incorporation of the dNTPs by a polymerase and the reaction stops when the marked ddNTP is incorporated; finally the fragments of double stranded DNA are denatured and size-separated by gel electrophoresis, from which the necessary information to determine the correct nucleotide sequence can be extracted (Morey et al. 2013). Almost 30 years after the introduction of Sanger sequencing, second generation sequencing methods and instruments became available and provided substantial improvements over the original method. These improvements were mainly their

higher throughput, and the capability to process multiple samples in parallel thanks to an improved sequencing chemistry; the main disadvantage of second generation sequencing instruments is their shorter read length (compared with Sanger sequencing). The most widely used second generation sequencing technologies in molecular microbial ecology studies are 454 Life Sciences and Illumina. 454 Life Sciences uses emulsion PCR (emPCR) to amplify the template, and subsequently performs sequencing of the coated beads in a well plate by adding the 4 nucleotides sequentially and measuring the intensity of the visible light generated which is equivalent to the quantity of bases incorporated (Morey et al. 2013). Illumina sequencing uses bridge amplification to amplify the template DNA in a flow cell and form clusters; sequencing is achieved by the addition of reverse-terminator fluorescently labelled nucleotides and the detection of the fluorescence emitted by each nucleotide when it hybridizes to a complementary base (Morey et al. 2013). 454 pyrosequencing can provide up to 700 Mb of sequence per run, which is approximately 1,000,000 reads of 400-1000 bp length. For similar applications in molecular microbial ecology, the Illumina sequencers provide much higher output at a lower price; for instance, with Illumina MiSeq v2 chemistry it is possible to process ~120 barcoded samples in one lane and generate 24-30 Millions of paired-end reads of 250 bp in length (Goodwin et al. 2016).

- Flow cytometry: allows cell enumeration in water samples through staining. A typical protocol for cell enumeration in drinking water is as follows (Prest et al. 2014): for the determination of total bacterial cell concentrations, SYBR® green I is added to pre-heated water samples and further incubated before measurement; for the assessment of intact bacterial cell concentrations, a solution of SYBR® green I and propidium iodide is employed with a similar staining protocol. Cell count is then performed with a flow cytometer which detects and collects bacterial signals that are subsequently analyzed and translated into quantitative (e.g. number of cells/ml) and qualitative information (e.g. nucleic acid content). The method has been applied in combination with ATP measurements and heterotrophic plate counts to detect changes in the microbial water quality in distribution systems (Gillespie et al. 2014; Vital et al. 2012) and changes as a result of treatment processes, being able to detect cell numbers as low as 10^2 cells/ml in finished treated water at the plant (Hammes et al. 2008). Its main advantages are its sensitivity and fast application.

- Quantitative Polymerase Chain Reaction (qPCR): qPCR has been extensively applied in bacterial enumeration (using gene copy number as quantitative data), through real-time monitoring of the amplification of specific targeted genes. Two methods are commonly used for the detection of products: (i) non-specific fluorescent dyes that bind to double stranded DNA product (e.g. SYBR® green I), the fluorescence is measured at the end of each cycle when a net increase is detected as the PCR progresses; (ii) sequence-specific DNA probes labelled with a fluorescent molecule that reports a signal once the probe has hybridized with its complimentary DNA sequence (e.g. TaqMan probes). A standard curve is constructed relating standards (with known template amount) to their fluorescence (i.e. their threshold cycle - Ct values), and its associated linear equation is used to estimate the copy number of the samples. qPCR has been applied in the detection of pathogens in DWDSs (H Wang et al. 2012; Whiley et al. 2014; van der Wielen & van der Kooij 2013) and assessment of treatment processes (de Vet et al. 2011).

- Stable isotope probing (SIP): is a technique that links microbial community composition and phylogeny to metabolic capacity by tracking the incorporation of heavy stable isotopes from specific substrates into informative biomarkers associated with microbes that assimilate the substrate. Therefore, SIP can identify viable microbial populations that have a defined function (Uhlik et al. 2013; Dumont & Murrell 2005). When the biomarker chosen is DNA, the identification of the members of the community is done through DNA sequencing. For instance, DNA-SIP was used to elucidated the role of Ammonia oxidizing archaea (AOA) and Ammonia oxidizing bacteria (AOB) in the nitrification process in granular activated carbon (GAC) filters from full-scale treatment plants (Niu et al. 2013); while in lab-scale slow sand filters, the technique showed that *E. coli* removal is linked to protozoan predation, being protozoan grazing the main removal mechanism (Haig et al. 2014).

- Meta'omics: the term encompasses a collection of techniques based on high-throughput sequencing and molecular methods to characterize microbiomes through their genomes (i.e. metagenomics), transcriptomes (i.e. metatranscriptomics), proteomes (i.e. metaproteomics), and metabolites (i.e. metabolomics) (Segata et al. 2013). Comprehensive approaches including two or more 'omics techniques have been applied to environmental samples, including wastewater (Sales & Lee 2015) and marine sediments (Urich et al. 2014).

Metagenomics has been applied in DW through 16S rRNA gene amplicon sequencing; additionally, some studies have used whole-DNA shotgun sequencing to assess the genetic composition (Gomez-Alvarez et al. 2012; Chao et al. 2013). To our knowledge, no study has attempted to elucidate the transcriptome and proteome of DW microbial communities.

2.3.2. Microbial ecology analyses

Microbial ecology or environmental microbiology is the study of microorganisms and their interactions with each other and with the environment. Two questions are usually answered in a microbial ecology study: "*who is in the community?"* and "*what is its relation with the other samples/organisms and the environment?".* To answer the first question, one must sample the environment under study and characterize its microorganisms, and measure relevant environmental parameters; to achieve this, several techniques are available, as seen in Section 2.3.1. After characterization, and in order to answer the second question, several tools in the field of multivariate statistics can be applied depending on the type of data available and the approach preferred by the researcher. For instance, in the case of studies on full-scale drinking water systems, the microbial communities characterized can be linked to their environment through the system components (e.g. type of source, treatment, distribution network, reservoirs, pipe length or pipe surface area, etc.) and through the DW's physical and chemical parameters (e.g. temperature, turbidity, pH, hardness, total chlorine, etc.).

To characterize and measure the variety of microorganisms in a given sample, the diversity of the sample is estimated. Two types of diversity estimates are used: Alpha-diversity (Whittaker 1960) is the mean species diversity in a site or sample, and can be assessed through the estimation of diversity indices such as the number of organisms in a sample (i.e. richness), the Shannon index, Simpson index, Inverse Simpson index, Chao index, or through the use of species abundance models and ranked abundance distributions (Oksanen et al. 2013). The diversity indices differ in their formulae and therefore in the information they provide about the samples and their communities. For instance, the species richness indicator will provide quantitative information of the community as it indicates how many species are present; while the Shannon index accounts for both the abundance of the species and the evenness in their distribution (i.e. a sample with a Shannon's index of 2.0 has more species and they are more evenly distributed than a sample with a Shannon's

index of 0.5); and finally, the Chao index will favour low abundance species (i.e. singletons and doubletons) in its estimation.

No.	Name	Formula	
Alpha-diversity			
1	Chao	$C_A = S_{obs} + \frac{f_1(f_1 - 1)}{2(f_2 + 1)}$	S_{obs} : number of observed species in sample A f_1 : number of singletons in sample A f_2 : number of doubletons in sample A
2	Shannon	$H_A = -\sum_{i=1}^{S} P_i log_b P_i$	S : total number of species in sample A P_i : proportion of species *i* in sample A b : e (2.71828)
3	Inverse Simpson	$D_A = \frac{1}{\sum_{i=1}^{S} P_i^2}$	S : total number of species in sample A P_i : proportion of species *i* in sample A
Beta-diversity			
4	Bray Curtis	$BC_{AB} = \frac{\sum_{i=1}^{S} \lvert x_{iA} - x_{iB} \rvert}{x_A + x_B}$	x_{iA} : number of individuals of species *i* in sample A x_{iB} : number of individuals of species *i* in sample B S : total number of species x_A : total number of individuals in sample A x_B : total number of individuals in sample B
5	Jaccard	$J_{AB} = \frac{S_{AB}}{S_A + S_B - S_{AB}}$	S_A : number of species in sample A S_B : number of species in sample B S_{AB} : number of species shared by samples A and B
Coverage			
6	Good's coverage	$C_A = \left(1 - \frac{f_1}{S_A}\right) * 100\%$	S_A : total number of species in sample A f_1 : number of singletons in sample A

Table 2.4. Selected diversity indices used in microbial community analysis. (Oksanen et al., 2013)

Beta-diversity (Whittaker 1960) was proposed as a measure of the difference in species composition from site to site or habitat to habitat. It can be generalized as a comparison of communities across samples to assess their (dis)similarity, and the link with environmental variables (e.g. an environmental gradient). Beta-diversity metrics can be abundance-based, if the leverage the abundance of the members of the community, or presence/absence-based, if they don't take into account the abundance of the members of the community and instead leverage the members of the community shared by each pair of samples. Among the beta-diversity indices, Bray Curtis (dis)similarities (abundance-based) are robust and among the top 10 best performing beta-diversity metrics (Barwell et al. 2015), and have been extensively used in microbial ecology studies; among the presence/absence metrics, the Jaccard metric provides a useful and simple estimation of dissimilarity between samples based on the shared members of their communities. Both abundance-based and

presence/absence-based indices can be applied to count data, the former by using abundance values, and the latter by estimating the corresponding presence/absence table from the count data and using it as input for the estimation of the indices. This approach of applying multiple beta-diversity indices provides more information about the structure and membership of the communities, and therefore adds confidence to the interpretations drawn from the data.

A large amount of data is generated with high-throughput DNA sequencing. After sequence analysis the resulting data is usually in tabular format, with rows containing samples (from 2 to tens or hundreds of samples) and columns containing count data of Operational Taxonomic Units (OTUs, a surrogate for "species", thousands of OTUs in the case of Illumina Sequencing) or other taxonomic assignment (e.g. counts of phyla, classes, orders, etc.); additionally, for each sample multiple environmental parameters are usually recorded for analysis purposes. To analyze this kind of data, one must apply multivariate statistics that reduce the complexity of the data and reveal relationships between the multiple variables that are relevant to the issue under study. Several methods are available to analyze multivariate data (Figure 2.5), most of them correspond to the approaches described below:

(i) Exploratory analyses: search for patterns in the data through exploratory/qualitative approaches. The methods rely on the visualization in a plot of complex datasets using (dis)similarity measures, in which two objects (e.g. samples) that are closer are more similar than two objects that are farther apart in the plot. Several methods can be used for visualization (e.g. PCA: principal components analysis; CA: correspondence analysis; DCA: detrended correspondence analysis; PCoA: principal coordinates analysis; NMDS: non-metric multidimensional scaling; etc.), depending on the type of data and the relationship between its variables; among these, Principal Coordinates Analysis (PCoA) and Non-metric Multidimensional Scaling (NMSD) have the advantage that don't assume any relationship between species abundance and environmental variables, and that any distance metric can be used to visualize the samples, providing flexibility to the analysis. Hierarchical clustering techniques are also useful for exploration purposes, as their goal is to connect objects based on their distance. The clusters formed are represented in a dendrogram, allowing the identification of groups of objects/samples that are similar to each other and enabling a discussion of the likely causes of such

similarities; moreover, hierarchical clustering is also flexible in the use of distance metrics. Among the most popular clustering methods in ecology is Unweighted Pair Group Method with Arithmetic Mean (UPGMA), which is a simple bottom-up agglomerative method that yields a dendrogram representative of the structure of a dis(similarity) matrix (Seath & Sokal 1973).

(ii) Interpretive analyses: this group of analyses uses both the measured variables (i.e. OTU table) and explanatory variables (i.e. metadata table) to find significant relationships between. The ordination methods corresponding to this group (e.g. CCorA: canonical correlation analysis; aim to find axes in a multidimensional data set space that maximize the association between the explanatory variables and the measured variables, therefore the ordination axes are constrained to be functions of the explanatory variables (Paliy & Shankar 2016). Another group of interpretive analyses (e.g. ANOSIM: analysis of similarities; PERMANOVA: permutational multivariate analysis of variance) allow us to test for significant differences between the two groups by assessing the variations within and between groups using distance metrics; these analyses are especially useful when testing hypotheses, and are recommended over exploratory analyses that rely on visualization. Finally, another type of interpretive analyses are differential expression analyses, which are used to determine which species (e.g. OTUs) are differentially abundant in two conditions, by comparing their relative abundance. To tests for differential abundance one can apply a Kruskall-Wallis test to determine if there is a statistically significant difference between the relative abundances of the OTUs under different conditions; another option is the use of the DeSeq2 package that achieves the same result using a more sophisticated approach, based on the simulation of count data following a negative binomial distribution and other features to deal with the overdispersed nature of the count data (McMurdie & Holmes 2014).

(iii) Discriminant analyses: their goal is to define discriminant functions that will maximize the separation of objects among different classes (Paliy & Shankar 2016). Some examples of the use of classifiers with DNA sequencing data are available, most of them applied to human microbiome samples. For instance, Knights et al. (2011) reported that Random forests outperformed four other classification techniques when applied to five different human-associated

datasets. Applied to the environment, Random forests showed that the microbiota of the flowers and roots of grapevines was the most distinct of the sample types (Zarraonaindia et al. 2015).

The choice of methods to apply (Figure 2.5) depends on the questions that need answers. A limitation to take into account when applying multivariate statistical methods is that the patterns revealed do not shed light on causality but only reflect an association or correlation between any two measures, therefore care must be taken not to misinterpret their results as biological/ecological effects. The reviews by Ramette (2007) and Paliy and Shankar (2016) provide further details of the methods in each category and examples of applications. In addition to powerful multivariate statistics tests and analyses such as the ones described above, classical statistical tests such as Analysis of variance (ANOVA) and t-tests can be used when appropriate, to test significant differences between groups (see Appendix D).

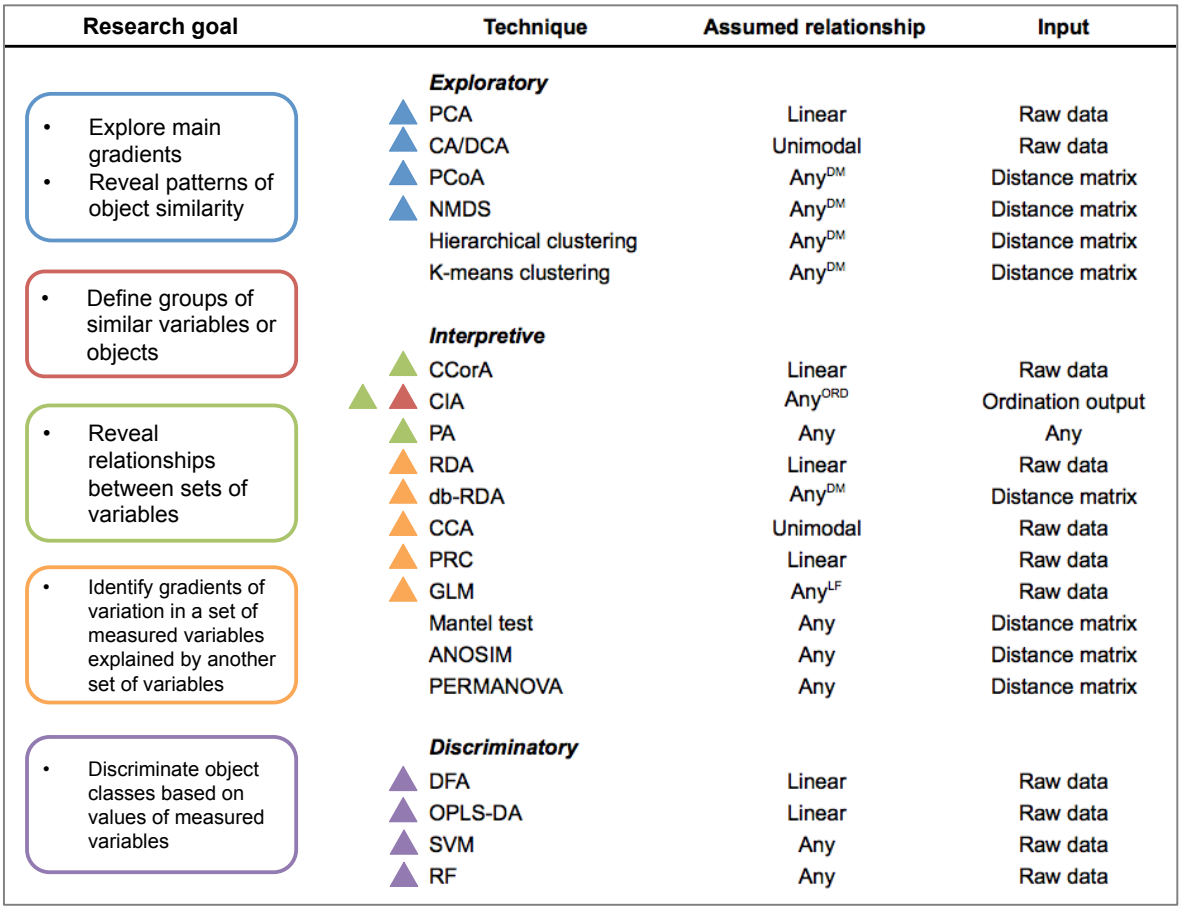

Figure 2.5. Multivariate analysis techniques according to the research goal
PCA: principal components analysis; CA: correspondence analysis; DCA: detrended correspondence analysis; PCoA: principal coordinates analysis; NMDS: non-metric multidimensional scaling; CCorA: canonical correlation analysis; CIA: co-inertia analysis; PA: Procrustes analysis; RDA: redundancy analysis; db-RDA: distance-based RDA; CCA: canonical correspondence analysis; PRC: principal response curves; GLM: generalized linear model; ANOSIM: analysis of similarities; PERMANOVA: permutational multivariate

analysis of variance; DFA: discriminant function analysis; OPLS-DA: orthogonal projections to latent structures discriminant analysis; SVM: support vector machine; RF: random forest (Adapted from Paliy & Shankar 2016).

3. A meta-analysis of microbial communities in full-scale drinking water distribution systems

3.1. Introduction

Drinking water distribution systems (DWDSs) are designed, built, and managed with the purpose of delivering potable and palatable water from the treatment plant to the consumer's taps. It is imperative that microbiological and chemical quality of water be maintained within regulatory limits during its transport through the DWDS. Deterioration in the microbiological quality of water may occur either due to ingress of microorganisms into the DWDS through leaks and due to undesired microbial regrowth in the DWDS and/or the premises plumbing. Controlling undesired microbial regrowth is mainly achieved by managing two factors: maintaining a low concentration of assimilable organic carbon (AOC) (Kooij, 1992) and other growth-rate limiting substrates (i.e. ensure oligotrophic conditions); and/or applying a residual disinfectant (i.e. secondary disinfection) to inactivate microorganisms. While AOC is usually the growth-rate limiting substrate for microbial activity in DWDSs (LeChevallier et al., 1991), secondary disinfection is applied to protect the network from microbial contamination that could enter through cross-connections and pipe breaks, and to suppress bacterial growth. The health-based guideline values for maximum residual chlorine and monochloramine (the most widely used disinfectants) concentration have been set to 5.0 mg/l and 3.0 mg/l, respectively (WHO, 2011), although concentrations are maintained low (~1.0 mg/l) to avoid odour and taste issues with the consumers. Despite all these efforts to inactivate microorganisms and control microbial regrowth, DWDSs harbour an abundant and diverse microbiome consisting of thousands of operational taxonomic units (OTUs) that span across the tree of life and harbour diverse functional potential (Gomez-Alvarez et al., 2012a, Chao et al., 2013, Roeselers et al., 2015, Pinto et al., 2014).

Recent advances in our understanding of the DW microbiome can in large part be attributed to the application of high-throughput and deep DNA sequencing-based methods that target the 16S rRNA gene (Sogin et al., 2006, Caporaso et al., 2011). The 16S rRNA gene is the most widely used molecule for phylogenetic analyses of bacteria and archaea; four characteristics make it ideal for this task: it is (i) universally distributed, (ii) functionally constant, (iii) sufficiently conserved (i.e. slow changing), and (iv) of adequate length (approximately 1500 bp) (Madigan et al. 2012).

The 16S rRNA gene has highly conserved regions, and nine hypervariable regions (V1-V9) that span nucleotides 69-99, 137-242, 433-497, 576-682, 822-879, 986-1043, 1117-1173, 1243-1294 and 1435-1465, respectively (based on the *E. coli* nomenclature) (Chakravorty et al. 2007). Since the hypervariable regions are shorter than the complete gene, their sequences can be obtained using high-throughput DNA sequencing instruments that provide thousands to millions of reads per sample, and as a result of this greater sequencing depth, a wider understanding of the microbial communities in the samples. Taxonomic annotation of the full gene or its amplicons can be done using available databases such as SILVA (Yilmaz et al. 2014) and Greengenes (DeSantis et al. 2006). The hypervariable region amplified has been reported as a significant variable in community composition, over natural or biological inter-sample variation, for different types of samples (e.g. stool - Clooney et al. 2016; river water - Staley et al. 2015).

The application of DNA sequencing-based methods to study the DW microbiome has also highlighted the influence of process operation (Chao et al., 2013, Lin et al., 2014, Lautenschlager et al., 2014), disinfectant type (Gomez-Alvarez et al., 2012b, Wang et al., 2014a, Hwang et al., 2012b), environmental conditions (Pinto et al., 2014), hydraulic conditions (Bautista-de los Santos et al., 2016, Douterelo et al., 2013), distribution system structure (Ling et al., 2015, Pinto et al., 2014), premise plumbing characteristics (Wang et al., 2014a, Yu et al., 2010) on the structure and composition of the DW microbiome.

Emerging from these studies is a general consensus on the types of microorganisms that are typically encountered in DW samples. Bacteria within the phylum *Proteobacteria* (Proctor and Hammes, 2015) and in particular those within the classes of *Alpha*- and *Betaproteobacteria*, have been shown to be dominant in nearly every study published, thus far. Nonetheless, studies have also reported differences in the dominance of these classes depending on a range of factors, including but not limited to seasons (Ling et al., 2015, Pinto et al., 2014) and disinfection strategy (Roeselers et al., 2015, El-Chakhtoura et al., 2015, Gomez-Alvarez et al., 2012b, Chiao et al., 2014). Despite this emerging consensus about the composition of the DW microbiome, particularly the bacterial community, to our knowledge there has been no study that attempts a collective analysis (i.e. meta-analysis) of all publicly available DW datasets. Such an early-stage meta-analysis effort can reveal conserved features across DW systems, help identify targeted research questions, and highlight opportunities to improve future DW microbiome studies.

An important aspect that could benefit from a collective analysis of published studies could be the impact of different microbial regrowth control strategies. Preliminary, insights on the mechanisms of action of both microbial regrowth control strategies suggest that they may shape microbial communities in different ways. In the case of disinfectants, their main mode of action is inactivation of the cells by: (i) damage to cell membranes, (ii) oxidation of cytochromes, proteins and nucleotides, (iii) disruption of the metabolism and protein synthesis, (iv) and modification of purine and pyrimidine bases that cause genetic defects (WHO 2004). Laboratory inactivation experiments show that a resistant bacterial sub-population remains in the drinking water despite the contact time and the continuous presence of the disinfectant, suggesting that this sub-population can survive and proliferate (AWWARF & EPA 2005). Increase in protective functions (e.g. increase in glutathione synthesis genes in 'oxidative stress' and 'detoxification' sub-systems) observed in chlorinated treated water further confirms that the surviving bacteria after disinfection may have higher chlorine resistance (Chao et al. 2013) and increased resistance to antibiotics (Xi et al. 2009; Jia et al. 2015). On the other hand, low levels of growth-rate limiting substrate (e.g. AOC) can cause starvation, bacterial growth inhibition, and trigger physiological changes in the cells. For instance, in the presence of low nutrients, *Escherichia coli* follows a two-stage starvation protocol that consists of scavenging (forage for the limiting nutrient and switch to other nutrients), and if scavenging fails, the cells starve and enter a stationary phase (Peterson et al. 2005). In the case of *Pseudomonas aeruginosa*, carbon starvation has been shown to induce massive dispersal events in biofilms via a reduction of intracellular levels of c-di-GMP (an intracellular signaling molecule that regulates biofilm formation and motility) (Schleheck et al. 2009).

In this chapter, I present a collective analysis of all publicly available datasets involving bulk DW samples collected at the outlet of the DWTP ($DWTP_{outlet}$) which represents the treated water, in the DWDS, and at point-of-use (POU). While microbial communities in drinking water systems exist in multiple phases like biofilms, suspended particles, loose deposits, and bulk water (Liu et al. 2013), this analysis is focused only on the bulk DW samples for several reasons. First, bulk water represents the primary mode of customer exposure to DW microbial communities. Second, studies have clearly shown that bacterial communities in bulk water and biofilms on pipe walls are distinct (Henne et al., 2012, Liu et al., 2014) although biofilms influence the former (Schroeder et al., 2015, Douterelo et al., 2014) and can have potential impacts on health (Schoen and Ashbolt, 2011). Finally, several studies have demonstrated that though there is temporal variation (Pinto et al., 2014) the bulk DW bacterial community within a given distribution system is relatively

stable irrespective of the sampling location (Roeselers et al., 2015, Pinto et al., 2014, Lautenschlager et al.) over short time-scales and is even reproducible over annual time-scales (Pinto et al., 2014). In contrast, biofilms and even deposits are spatially heterogeneous (Wimpenny et al. 2000) and are likely to develop over time-scales that are much longer than the residence time of water within a given DWDS. This spatial heterogeneity and uncertainty related to time-scales of community assembly results in a poor understanding of how a biofilm community at one location in the DWDS may relate to those at other locations within the same system. Therefore, the lack of rigorous characterization of biofilm heterogeneity even for a single DWDS thus far, limits the utility of comparing biofilm communities across systems.

The objectives of this study were to (1) identify microbial populations that are detected across all publicly available bulk DW datasets; (2) evaluate the variation in the occurrence and relative abundance of target microbial groups at the phylum/class and operational taxonomic unit (OTU) level, (3) evaluate the relationship between occurrence and relative abundance of taxa across systems, (4) determine the association between disinfection strategy and microbial community, and (5) provide insights into the functional potential across all samples and within disinfection strategy type using 16S rRNA gene data, to the extent possible.

3.2. Methods

3.2.1. Data collection

The data collection efforts were focused on published datasets that involved (1) collection of bulk water samples from either the $DWTP_{outlet}$, in the DWDS and/or at the POU, (2) extraction of DNA from the sample without an enrichment or cultivation step, (3) PCR amplification of any of the hypervariable regions of the 16S rRNA gene from the extracted DNA, and (4) sequencing of the PCR product on any high-throughput DNA sequencing platform (i.e. Illumina MiSeq, 454 pyrosequencing, and Ion Torrent). Further, the analysis focused on differences across sampling locations, rather than temporal change at each sampling location. As a result, multiple temporally distinct samples collected from the same sampling location were collapsed into a single sample. Based on these criteria, we were able to identify 21 distinct studies with 6,5,4,2,2,1, and 1 datasets coming from USA (Holinger et al., 2014, Hwang et al., 2012a, Ji et al., 2015, Pinto et al., 2014, Wang et al., 2014b, Zhang and He, 2013), China (Huang et al., 2014, Jia et al., 2015, Lin et al., 2014,

Zeng et al., 2013, Bai et al., 2015), Netherlands (Liu et al., 2014, Roeselers et al., 2015, Prest et al., 2014, El-Chakhtoura et al., 2015), UK (Bautista-de los Santos et al., 2016, Douterelo et al., 2014), Switzerland (Lautenschlager et al., 2013, Lautenschlager et al., 2014), Australia (Shaw et al., 2015) and France (Costa et al., 2015), respectively. Of these 21 datasets, 14 datasets were either publicly available or made available upon data request (Table S1). Hence, these 14 published datasets comprising of 142 distinct sampling locations were included in this study (Bautista-de los Santos et al., 2016, Douterelo et al., 2014, Holinger et al., 2014, Huang et al., 2014, Hwang et al., 2012a, Ji et al., 2015, Jia et al., 2015, Lin et al., 2014, Liu et al., 2014, Pinto et al., 2014, Roeselers et al., 2015, Wang et al., 2014b, Zeng et al., 2013, Shaw et al., 2015) (See Appendix A, Table A1).

3.2.2. Data processing

The FASTA/FASTQ files from individual datasets were processed using a combination of tools and quality filtering criteria depending on the sequencing platforms used and hypervariable regions of the 16S rRNA gene sequenced. The FASTQ files containing single-end reads were quality filtered using sickle v.1.33 (Joshi NA, 2011) with a minimum quality score of 28 and a minimum length of 150 bp after trimming and then converted to FASTA format using the fastq_to_fasta command in the FASTX-Toolkit v.0.0.13.2. The FASTQ files containing paired-end reads were processed using pear v.0.8.1 (Zhang et al., 2014) to make contigs, with a minimum quality score of 28 and a minimum length of 150 bp after assembly. The FASTA files were dereplicated in mothur (Schloss et al., 2009), and unique sequences were matched against the SILVA 119 SSURef_Nr database (Pruesse et al., 2007) using blastn (Altschul et al., 1990) with an identity ≥ 97% and an Expect (e) value less than 0.000005. The best match 16S rRNA gene sequences from the SILVA 119 database were extracted and used for further analysis. Sequences that did not find a suitable match in the Silva 119 database were excluded from alpha and beta-diversity analysis. The best-match sequences corresponding to each sample were then aligned against the SILVA seed alignment available through mothur (Schloss et al., 2009). The alignment was screened to remove poorly aligned sequences and filtered using the vertical =T and trump =. options in mothur (Schloss et al., 2009). The filtered alignment was then clustered into OTUs at a sequence similarity cutoff of 97% using the average neighbour clustering approach (Schloss et al., 2009). All sequences were classified using the Naïve Bayesian classifier (Wang et al., 2007) (80% confidence threshold) using SILVA taxonomy and consensus taxonomy of OTUs was estimated using an 80% consensus cutoff.

3.2.3. Data analysis

The number of sequences across the 142 sampling locations varied from 223 to 10.8 million. Given significant variability in sample size (Weiss et al., 2015), the data was subsampled to normalize the dataset. In order to determine the appropriate subsampling depth, Good's coverage was estimated for all sampling locations at sampling depths ranging from 200-2500 sequences. An appropriate sampling depth was determined by selecting subsampling depths that provided >80% Good's coverage for each sample while retaining the maximum number of sampling locations from the dataset. This presented the options of subsampling at 500 and 1000 sequences per sample, with the loss of 2 and 6 sampling locations at each of these subsampling depths, respectively (Figure 3.1). A Mantel test conducted using distances matrices constructed with Bray-Curtis metric at subsampling depths of 500 and 1000 sequences per sampling location showed significant correlations between the two distance matrices (Mantel's R = 0.995, p =0), indicating that a small benefit from a higher subsampling depth was accompanied by the loss of 4 additional sampling locations. As a result, a subsampling depth of 500 was selected to maximize the number of sampling locations retained. All estimates of alpha and beta-diversity were performed at this subsampling depth.

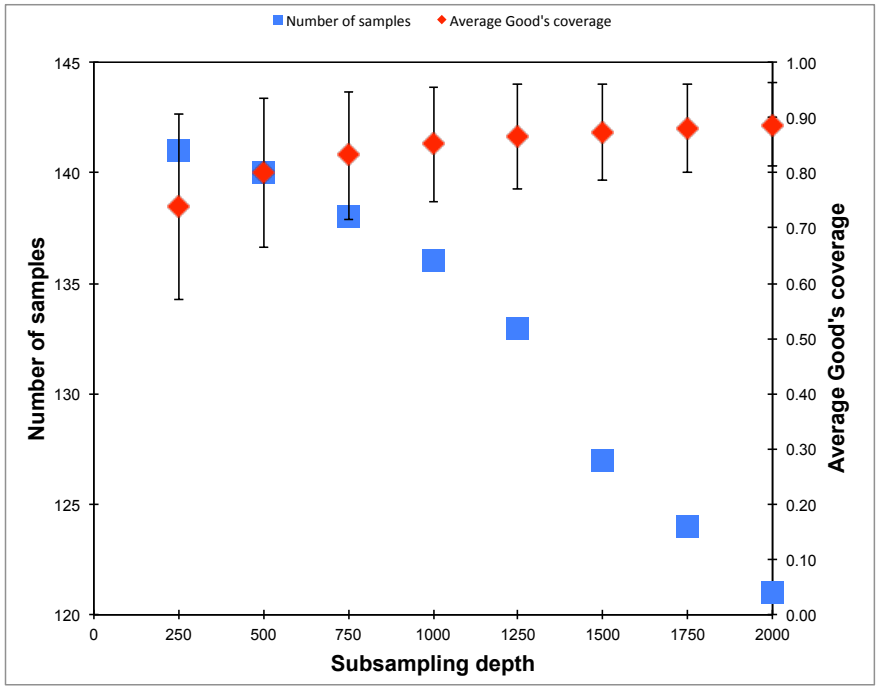

Figure 3.1. The number of samples retained (primary Y-axis, blue squares) with increasing subsampling depths and their corresponding Good's coverage (secondary Y-axis, red squares) is shown.
A subsampling depth of 500 sequences per sample was chosen as this allowed for an average Good's coverage 0.8 while retaining maximum number of samples (142 out of 145) in the analyses.

The subsampled OTU table was used as input for a range of diversity analyses using vegan (Oksanen Blanchet, 2013) and plots using the package ggplot2 (Wickham, 2009) in R (RCoreTeam, 2014). Specifically, richness (i.e. observed OTUs), Inverse Simpson index, Shannon index, and Pielou's evenness were estimated as measures of alpha-diversity. Beta-diversity analyses involved clustering of samples using the Bray-Curtis distance metric with the heatmap2 module in gplots (Warnes et al.) while overlap in membership between communities was estimate using the Jaccard index in mothur (Schloss et al., 2009). The most abundant sequence in each OTU was used as the representative sequence where relevant and RAxML (Stamatakis, 2014) was used to construct a maximum likelihood phylogenetic tree with the generalized time reversible (GTR) substitution model and GAMMA distribution model, using 1000 bootstraps. The resultant phylogenetic tree and relevant OTU data were then visualized in EvolView (Zhang et al., 2012). Permutational Multivariate analysis of variance (PERMANOVA) tests were conducted with vegan (Oksanen et al. 2013) to determine the effects of the study of origin, disinfectant strategy, and proportion of data retained after matching the SILVA database on differences between samples using the Bray-Curtis and Jaccard metrics.

The mean relative abundance and standard deviation (i.e. number of reads of an OTU in a sample divided by the total number of reads in the sample; expressed as a percentage, as MRA±SD throughout the chapter), and the occurrence of each OTU were estimated, across all sampling locations and sampling locations grouped by disinfection strategy. For these calculations, the relative abundance of each OTU for a sampling location was estimated by using all reads in the sample and not just the subset of reads matching the SILVA database. These full-samples were also used to compare occurrence and MRA of key OTUs across disinfection strategies. To check for the likelihood of contamination in DW studies, OTUs classifying to the genus level that corresponded to the list of kit/reagent contamination genera identified by Salter and colleagues (2014) were extracted and their contribution to the overall dataset was estimated. The subsampled OTU table was also used to predict the functional potential of the bacterial community using Tax4Fun (Aßhauer et al., 2015). Tax4Fun generates a relative abundance of Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Ogata et al., 1999) orthology (KO) groups associated with each sampling location depending on matches of the representative sequence from each OTU to KEGG organisms, while also providing information on fraction of OTUs that do not match KEGG organisms (i.e. the FTU metric). Analysis of variance (ANOVA) was performed to assess whether FTU values were significantly different across the three disinfectant strategies. For comparisons of KO relative abundance in samples grouped by disinfection strategy, we

picked a subset of samples from each disinfection strategy such that the distribution of FTU values and mean FTU was not significantly different between disinfection strategies. Significantly different KOs across different disinfection strategies were identified using the Kruskal-Wallis with Benjamini-Hochberg (Benjamini and Hochberg, 1995) correction with a false discovery rate of 0.05. A schematic outlining the workflow for all data-analyses in this manuscript is provided in the supplemental material (Appendix A, Figure A1).

3.3. Results

3.3.1. Data structure and composition

The 14 datasets consisted of 142 distinct sampling locations, with 79 and 63 sampling locations associated with systems with and without a disinfectant residual, respectively. Of the 79 sampling locations from systems with a disinfectant residual, 40 and 39 were from chlorinated (Chl) and chloraminated (Chm) systems, respectively. Data for a majority of these sampling locations was obtained on the 454 pyrosequencing platform (n=103), with data for 25 and 14 locations obtained on the Illumina MiSeq and Ion Torrent sequencing platforms, respectively.

The 16S rRNA gene hypervariable regions also varied depending on the datasets. Specifically, the hypervariable regions covered by the sequencing libraries for the 142 sampling locations included V1-V2 (n= 17), V1-V3 (n=7), V3 (n=14), V3-V4 (n=2), V3-V5 (n=2), V4 (n=25), V4-V5 (n=20), V4-V6 (n=3), and V5-V6 (n=52). Given the significant amount of data heterogeneity (sequencing platform and target 16S rRNA gene hypervariable region), we could not cluster sequences across studies directly into OTUs, a constraint highlighted by other recent meta-analysis efforts (Adams et al., 2015; Koren et al., 2013). Hence, we utilized a pre-processing step of sequence matching to the SILVA database as a means of being able to combine this highly heterogeneous data (i.e. a reference based approach).

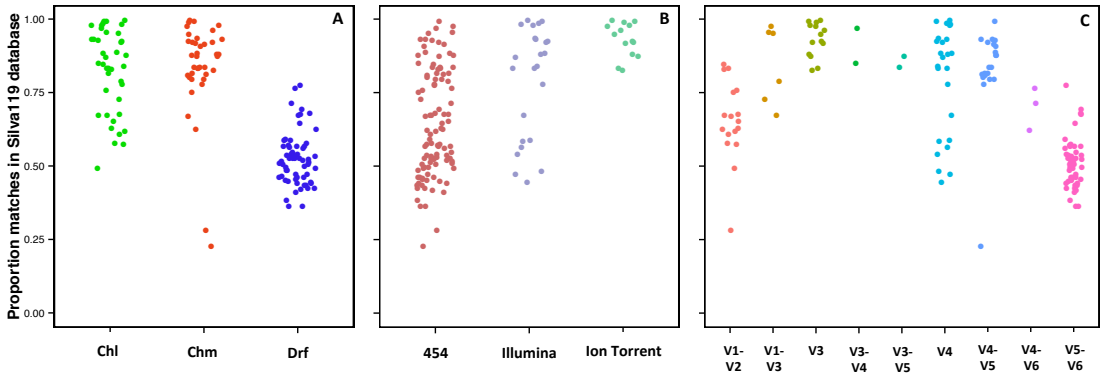

Figure 3.2. Proportion of reads from each sample library matching a reference sequence in the SILVA119 database with a minimum percent identity of 97% (E-value <0.000005). Data grouped by (A) disinfection group (Chl: chlorinated, Chm: chloraminated, Drf: disinfectant-residual free); (B) sequencing platform; and (C) 16S rRNA gene hypervariable region represented in the datasets utilized in this study.

The proportion of matches to the SILVA database was not specific to any particular study, but rather there was significant variability within studies themselves. For example, the average proportion of data with matches to the SILVA database for Chl, Chm, and disinfectant residual-free (Drf) samples were 82.1±13.9% (n=40), 83.9±16.1 (n=39) and 52±8.5% (n=63), respectively (Figure 3.2-A, $p<0.0001$ for Chl-Drf and Chm-Drf groups). Significant differences in the proportion of data matching the SILVA database were observed according to the sequencing platform (454-Illumina and 454-Ion Torrent, $p<0.001$) (Figure 3.2-B); however, the samples sequenced with Ion Torrent included only one hypervariable region of the 16S rRNA gene, therefore these results should be interpreted with caution. Similarly, significant differences in the proportion of data matching the SILVA database were observed according to the hypervariable region amplified, with p-values ranging from 6.1e-14 to 0.001 ($p<0.01$) (Figure 3.2-C). The direct effect of the lack of matches in the reference database meant that a proportion of data from each sample was not used for alpha and beta-diversity analyses. Specifically, all alpha and beta-diversity analyses were based on 81.5% of the sequence data from 142 sampling locations, with the average sequence data retained per sampling location being 69.4±19.9%.

It is also important to note that this meta-analysis does not account for biases that arise from sample collection and handling protocols (Lauber et al., 2010, Cuthbertson et al., 2014), DNA extraction (Feinstein et al., 2009) and PCR amplification (Pinto and Raskin, 2012) approaches. Therefore, this meta-analysis study does not provide a quantitative perspective on similarities and differences between the samples included in this study. Rather, we aim to highlight indicative differences that might be candidates for follow-up studies designed using standardized protocols across sample/system types.

3.3.2. Microbial community composition

Across all datasets, bacteria constituted a majority of the microbial community with the archaea being detected at very low levels, despite the fact that several studies used 16S rRNA gene primers that span bacterial and archaeal domains (e.g. V4 hypervariable region primer set provided by Caporaso et al., 2011). Specifically, archaeal sequences were detected in 9.5%, 19.5%, and 89% of the sampling locations from chlorinated, chloraminated, and disinfectant residual-free systems, respectively. Despite the widespread detection of archaeal sequences in disinfectant residual-free locations they contributed at a low level towards the overall community, with their MRA across Drf locations being 0.13±3.3%.

Proteobacteria were the most dominant bacterial phylum with their MRA for Chl, Chm and Drf being 68±42.7%, 75±42.9%, and 54±20.9%, respectively (Figure 3.3-A). Within *Proteobacteria*, *Alpha-* and *Betaproteobacteria* were dominant and constituted greater than 80% of the proteobacterial sequences across all locations. *Actinobacteria* was the second most abundant phyla in disinfected systems, constituting 11.7±16.2 and 8.2±10.7% of the data from Chl and Chm systems, respectively. In contrast, *Acidobacteria* was the second most dominant phyla for the Drf locations (MRA = 6.3±4%), while it constituted less than 1% of the sequences in disinfected systems.

These differences between disinfection strategies were not only limited to the abundance of the various phyla, but also with respect to their occurrence (Figure 3.3-B). For example, sequences from phylum *Nitrospinae* and *Crenarchaeota* were not detected in any of the disinfected samples, while being present in 29% and 46.7% of the samples without a disinfectant residual. Similarly, several low to medium abundance phyla were detected much more routinely in disinfectant residual-free systems compared to the systems with a disinfectant residual, indicating a greater taxonomic diversity of the bacterial community in absence of a disinfectant residual.

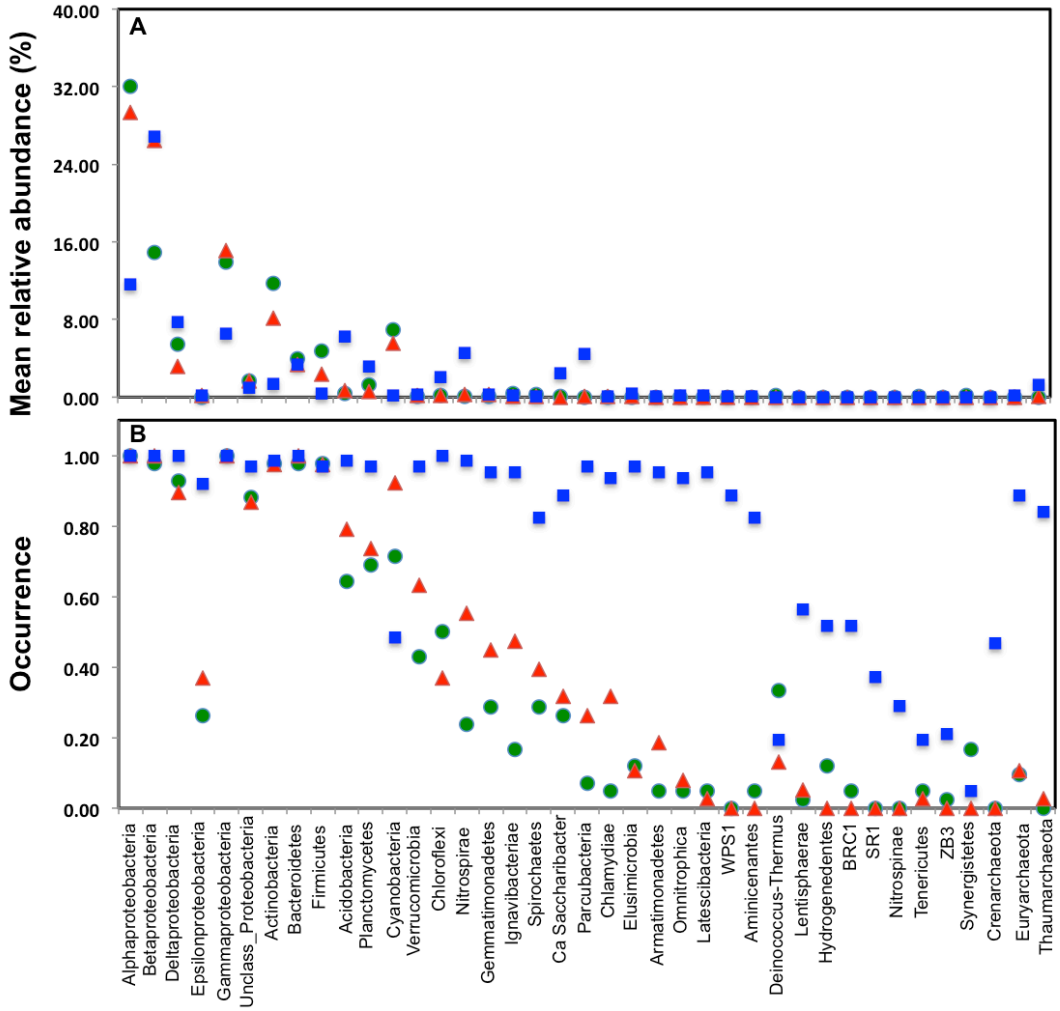

Figure 3.3. Bacterial phyla/classes grouped by disinfection strategy across groups. (A) Log mean relative abundance of bacterial phyla/classes grouped by disinfection strategy (Chl: chlorinated, Chm: chloraminated, Drf: disinfectant residual-free); (B) Occurrence of main bacterial phyla/classes per disinfection group.

3.3.3. Richness of bacterial communities

There were no significant differences in alpha-diversity between the sampling locations with chlorine and chloramine as the disinfectant residual (Figure 3.4). The inverse Simpson index was slightly higher for the chlorinated (12.8±15.4) as compared to the chloraminated (9.3±6.4) systems, however they also showed higher variability across locations. Consistently, the samples from disinfectant residual-free systems were richer, more diverse, and more even as compared to the samples with a residual disinfectant ($p<0.0001$). For example, the average number of OTUs in Drf systems was 225±60 as compared to 85±60 and 87±25 for Chl and Chm samples, respectively. Similarly, bacterial communities in Drf were significantly more even (0.84±0.14) as compared to those in the Chl (0.64±0.19) and Chm (0.64±0.13) systems. This observation of higher diversity in the non-disinfected sampling locations arises despite the fact that a smaller proportion of sequences from the non-disinfected samples were utilized for OTU construction due to fewer matches

to the SILVA database (Figure 3.2-A). As a result, it is likely that the magnitude of difference in diversity between disinfectant residual-free (i.e. Drf) and disinfected (i.e. Chm, Chl) systems is much larger than that indicated in Figure 3.4.

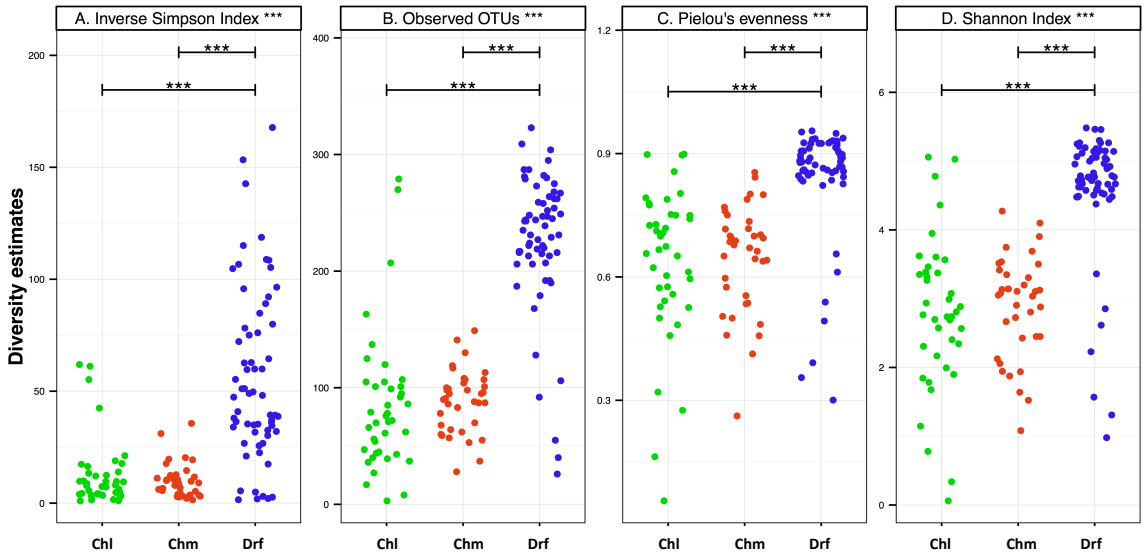

Figure 3.4. (A-D) Alpha-diversity per sample grouped by disinfection group (Chl: chlorinated, Chm: chloraminated, Drf: Disinfectant residual-free). This analysis was done using the OTU table subsampled to 500 reads per sample. Significant differences between disinfection strategies were evaluated using ANOVA and are indicated by bars at the top of each figure panel (p values: * = <0.01, ** = <0.001, * = <0.0001).**

3.3.4. Shared membership across disinfection strategies

The most commonly detected OTUs in Chl, Chm, and Drf systems were *Porphyrobacter* (class: *Alphaproteobacteria*) (MRA= 10±20%, occurrence = 0.60), *Bosea* (class: *Alphaproteobacteria*) (MRA = 10±40%, occurrence = 0.53), and *Nitrospira* (Phylum: *Nitrospirae*) (MRA = 10±10%, occurrence = 0.86), respectively. Table A2 (Appendix A) provides an overview of the most commonly detected OTUs (occurrence > 0.50) across the different disinfection strategies. Of the 7124 OTUs retained after subsampling, 6.6% (n=470), 8.57% (n=611), and 2.37% (n=169) were shared (present in all samples under consideration) by: (i) chloraminated and chlorinated, (ii) chloraminated or chlorinated and disinfectant residual-free, and (iii) chlorinated, chloraminated, and disinfectant residual-free locations, respectively. *Proteobacteria* constituted a majority of the OTUs shared between samples emerging from all three disinfection strategies (n=131) with 56, 41, and 22 OTUs classified as *Alpha-*, *Beta*, and *Gammaproteobacteria*, followed by OTUs within the phylum *Bacteroidetes* (n=12) and *Actinobacteria* (n=10) (Figure 3.5-A). Though there was no clear relationship between the abundance of an OTU at sampling locations with one disinfection strategy and its abundance or occurrence across the others, there was a clear

and positive relationship between abundance and occurrence of an OTU within a disinfection strategy (Figure 3.5B-3.5D).

3.3.5. Potential opportunistic pathogens across disinfection strategies

Disinfectant residual-free systems showed significantly higher relative abundance and occurrence of OTUs classified as *Legionella* at the genus level as compared to chlorinated ($p<0.01$) and chloraminated ($p<0.001$) systems. The MRA of *Legionella* OTUs was 0.2±0.7%, 0.18±0.24%, and 0.58±0.50%, while the occurrence of *Legionella* OTUs was 0.50, 0.59, and 0.97 in chlorinated, chloraminated, and disinfectant residual-free systems, respectively (Figure 3.6). This higher MRA and occurrence of *Legionella* in disinfectant residual-free system was also accompanied by a greater diversity of OTUs. Specifically, chlorinated, chloraminated, and disinfectant residual-free systems harboured 2±4, 7±12, and 25±13 OTUs that classified as *Legionella*, respectively. In contrast to *Legionella*, OTUs classified as *Mycobacterium* and *Pseudomonas* were more abundant and more frequently detected in disinfected systems as compared to disinfectant residual-free systems, with each of them exhibiting different trends when comparing chlorinated vs. chloraminated systems.

For instance, mycobacterial OTUs were more abundant and frequent in chlorinated (MRA=9±20%, occurrence=0.9) as compared to chloraminated systems (MRA=2.8±7.7%, occurrence=0.79), though the difference between the two was not significant (Figure 3.6). Similarly, OTUs classified as *Pseudomonas* were slightly more abundant in chloraminated systems (MRA=3.2±15%, occurrence=0.87) as compared to chlorinated systems (MRA=1±3%, occurrence=0.9) (Figure 3.6), but this difference was also not significant.

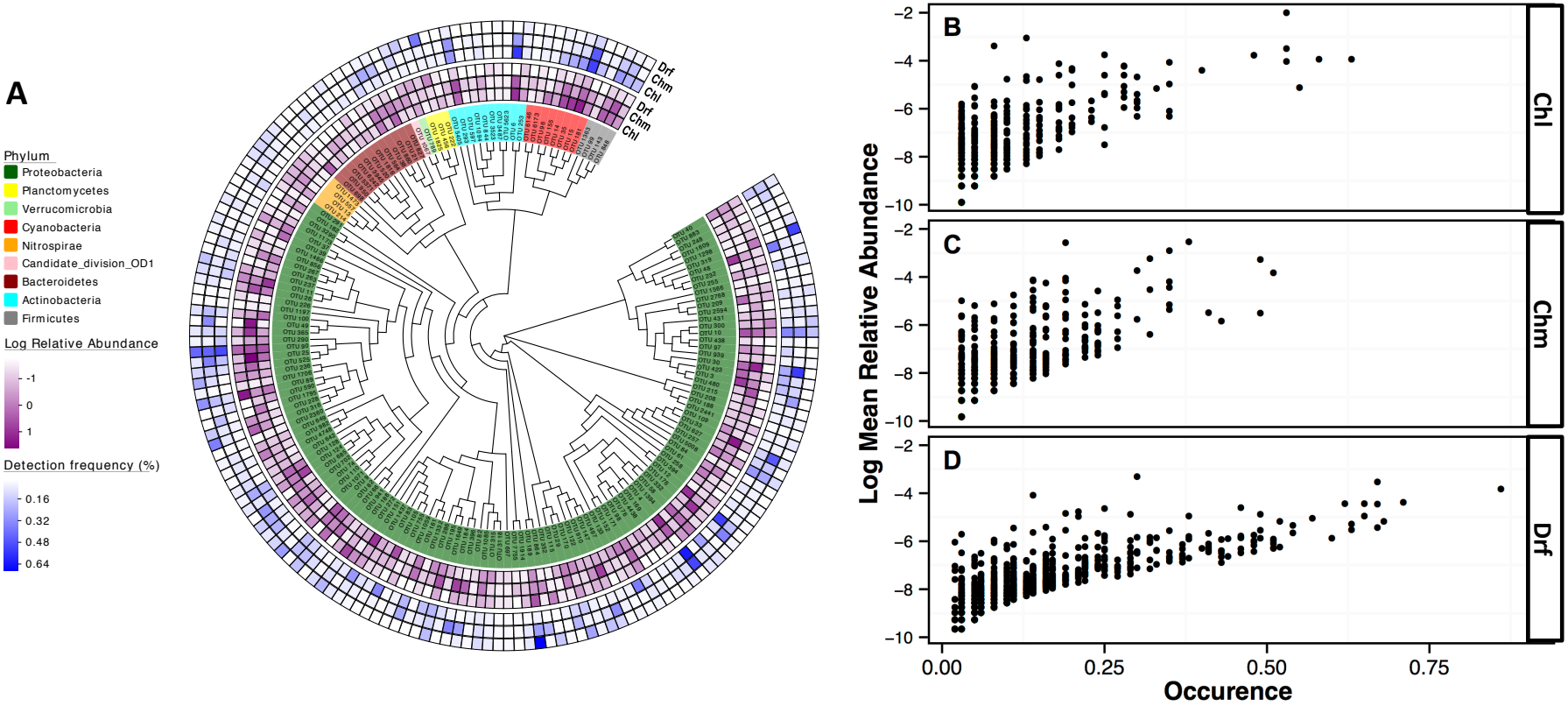

Figure 3.5. (A) Maximum likelihood phylogenetic tree of representative sequences from OTUs detected samples from all three disinfection strategies; (B-D) Positive relationship between the relative abundance and occurrence of all OTUs within a given disinfection strategy (Chl: chlorinated, Chm: chloraminated, Drf: Disinfectant residual-free).

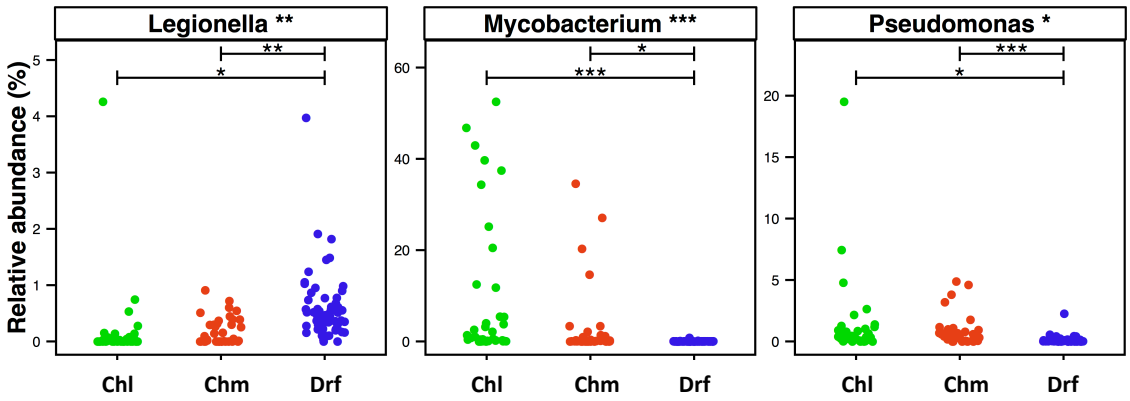

Figure 3.6. Relative abundance of OTUS classified as *Legionella*, *Mycobacterium* and *Pseudomonas* in each sample visualized by disinfection strategy type (Chl: chlorinated, Chm: chloraminated, Drf: Disinfectant residual-free). Significant differences between groups, evaluated by ANOVA, are indicated by bars at the top of each figure panel (p-value legend: * = <0.01, ** = <0.001, * = <0.0001).**

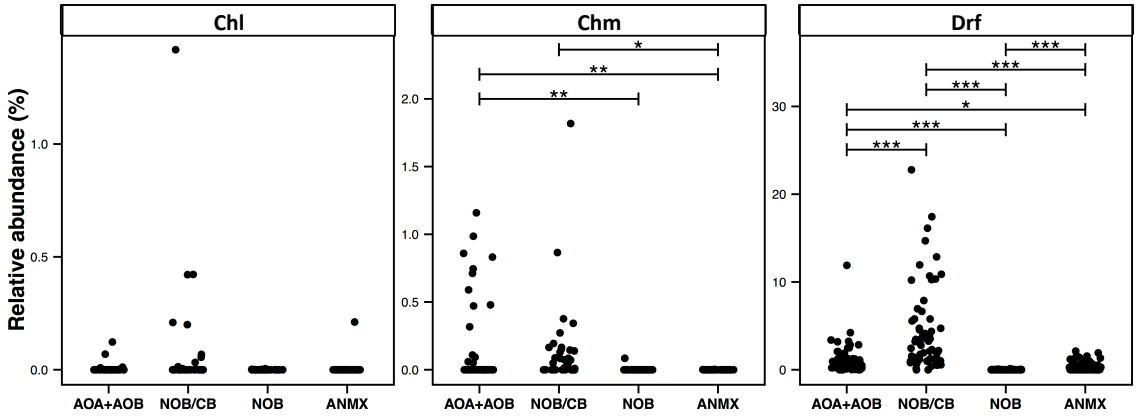

Figure 3.7. Relative abundance of nitrifiers in each sample visualized by disinfection strategy type (Chl: chlorinated, Chm: chloraminated, Drf: Disinfectant residual-free). X-axis labels correspond to: ammonia oxidizing archaea + bacteria (AOA + AOB), *Nitrospira* based nitrite oxidizing or comammox bacteria (NOB/CB), strict nitrite oxidizing bacteria (NOB), and anammox bacteria (ANMX). Significant differences between groups, evaluated by ANOVA, are indicated by bars at the top of each figure panel (p-value legend: * = <0.01, ** = <0.001, * = <0.0001).**

3.3.6. Ecologically relevant OTUs across disinfection strategies

The diversity and relative abundance of OTUs linked to nitrifying organisms were also estimated. These nitrifying organisms were grouped as ammonia oxidizing archaea (AOA), ammonia oxidizing bacteria (AOB), nitrite oxidizing or comammox bacteria within the genus *Nitrospira* (NOB/CB), strict nitrite oxidizing bacteria (NOB), and anaerobic ammonia oxidizing bacteria (anammox - ANMX) (Figure 3.7). Disinfectant residual-free systems exhibited the greatest relative abundance of AOA (MRA=0.48±0.80%) and they were detected in 0.79 of the disinfectant residual-free locations. AOA were also consistently low abundance in disinfected systems with the maximum MRA being 0.0000075%, while being detected in only 0.050 of the chloraminated locations with no detection in chlorinated systems. Disinfectant residual-free samples also harboured higher abundance and greater diversity of AOB and NOB/CB (Figure 3.7). For example, the MRA of AOB was 0.01±0.02%, 0.19±0.34% and 0.56±1.6%, while the occurrence of AOB was 0.2, 0.36 and 0.90 in chlorinated, chloraminated and disinfectant residual-free systems, respectively. Strict NOB were extremely low in abundance and were detected in only 0.20 of the sampling locations across the three disinfection strategies with maximum MRA of 0.12%. OTUs classified as *Nitrospira*, a genus that includes both strict NOB and the newly discovered comammox (Pinto et al., 2015, van Kessel et al., 2015, Daims et al., 2015) bacteria were detected at a higher relative abundance and frequency than either AOB or NOB in disinfectant residual-free systems. For instance, while the NOB and AOB were detected in 0.20 and 0.54 of all sampling locations, NOB/CB were detected in 0.68 of sampling locations across all disinfection strategies, with their MRA nearly 4 fold higher than AOB and AOA combined.

Another group of ecologically relevant microorganisms in DW is predatory bacteria. OTUs classified as *Bdellovibiro* (class: *Deltaproteobacteria*) and *Vampirovibrio* (Phylum: *Cyanobacteria*, Class: *Melainabacteria*), both predatory genera, were among the top 10 frequently detected OTUs across all three disinfection strategies (Appendix A, Table A2). Predatory bacteria are phylogenetically diverse and genus level identification is not sufficient to ascertain the presence of bacteria with obligate or facultative predatory lifestyle. Nonetheless, OTUs classified to some genera can be categorized as emerging from predatory bacteria (e.g. *Bdellovibrio*). Specifically, we found several OTUs classified as *Bdellovibrio* (n=114), *Cystobacter* (n=10), *Lysobacter* (n=46), *Peredibacter* (n=13), and

Vampirovibrio (n=92), all of which can be functionally classified as obligate or non-obligate predatory bacteria.

The three most frequently detected predatory OTUs (i.e. *Bdellovibrio*, *Lysobacter*, and *Vampirovibrio*), showed a significantly higher occurrence in disinfectant residual-free systems as compared to disinfected systems. For example, *Bdellovibrio*, *Lysobacter*, and *Vampirovibrio* were detected in 0.95, 0.52 and 0.98 of the locations from the disinfectant residual-free systems, respectively while the detection of the same predatory OTUs in chlorinated and chloraminated samples ranged from 0.2-0.4, 0.38 and 0.64-0.88, respectively. Further, though *Bdellovibrio* was significantly more abundant in disinfectant residual-free systems, both *Lysobacter* and *Vampirovibrio* exhibited a greater relative abundance in chlorinated systems. Specifically, *Lysobacter* and *Vampirovibrio* exhibited a relative abundance of 5±10% and 5±7% in chlorinated samples, respectively, while constituting less than 1% of the overall community for chloraminated and disinfectant residual-free samples.

3.3.7. Potential for contamination across DW datasets

Studies involving low-biomass samples are particularly susceptible to contamination emerging from a range of potential sources – from sample handling to PCR/DNA extraction reagents to contaminants from the sequencing process itself (e.g. sequences from one sample being attributed to another). Recent studies have demonstrated that kit/reagent contamination can critically impair studies that rely on sequencing datasets, with one study proposing an extended list of common contaminating genera (Salter et al., 2014).

Overall, 18.5±23% of the sequencing data across all studies was associated with a list of potentially contaminating genera provided by Salter et al. (2014). Approximately 23.5±19.8%, 29.6±25.5%, and 8.5±18.3% of data was associated with these genera for chlorinated, chloraminated, and disinfectant residual-free systems, with the proportions being significantly higher in disinfected as compared to disinfectant residual-free samples (Figure 3.8), which typically have a significantly higher cell count (Prest et al., 2014, Gillespie et al., 2014).

3.3.8. Community structure and membership across disinfection strategies

Clustering of samples showed a clear distinction between disinfected and disinfectant residual-free samples (Figure A2), but there was no clear clustering by the type of disinfectant residual (i.e. chloramine vs. chlorine). In addition, multiple factors can confound such broad level clustering (Figure 3.9). As discussed above, the available DW sequencing data is highly heterogeneous. A majority of the factors that contribute to data heterogeneity (e.g. DNA extraction protocol, PCR primer choice, sequencing platform, etc.) can largely be collapsed into one major variable – origin of study. PERMANOVA tests conducted using distance matrices (after subsampling) with Bray Curtis/Jaccard metrics indicated that origin of study had a strong impact on differences between samples (R^2=0.34/0.24, p=0.001) followed by type of source water (surface water, groundwater or mixed) (R^2=0.02/0.02, p=0.001) and disinfection type (R^2=0.014/0.01, p=0.01). Another variable that could affect the similarity between samples is the proportion of data used following the SILVA matching exercise (Figure 3.2). However, this had a minor effect on the community membership and structure based clustering using Jaccard (R^2=0.007, p=0.049) and Bray Curtis distance metrics (R^2=0.007, p=0.04), respectively.

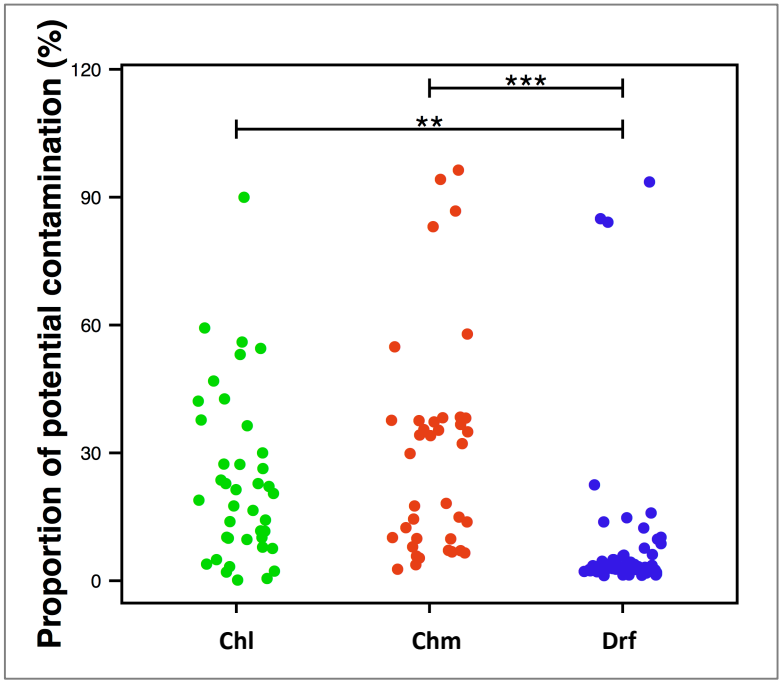

Figure 3.8. Proportion of potential contaminating sequences in each dataset per disinfection group (Chl: chlorinated, Chm: chloraminated, Drf: Disinfectant residual-free). Significant differences between groups, evaluated by ANOVA, are indicated by bars at the top of each figure panel (p-value legend: * = <0.01, ** = <0.001, * = <0.0001). List of potentially contaminant genera obtained from Table 1 in Salter et al. (2014).**

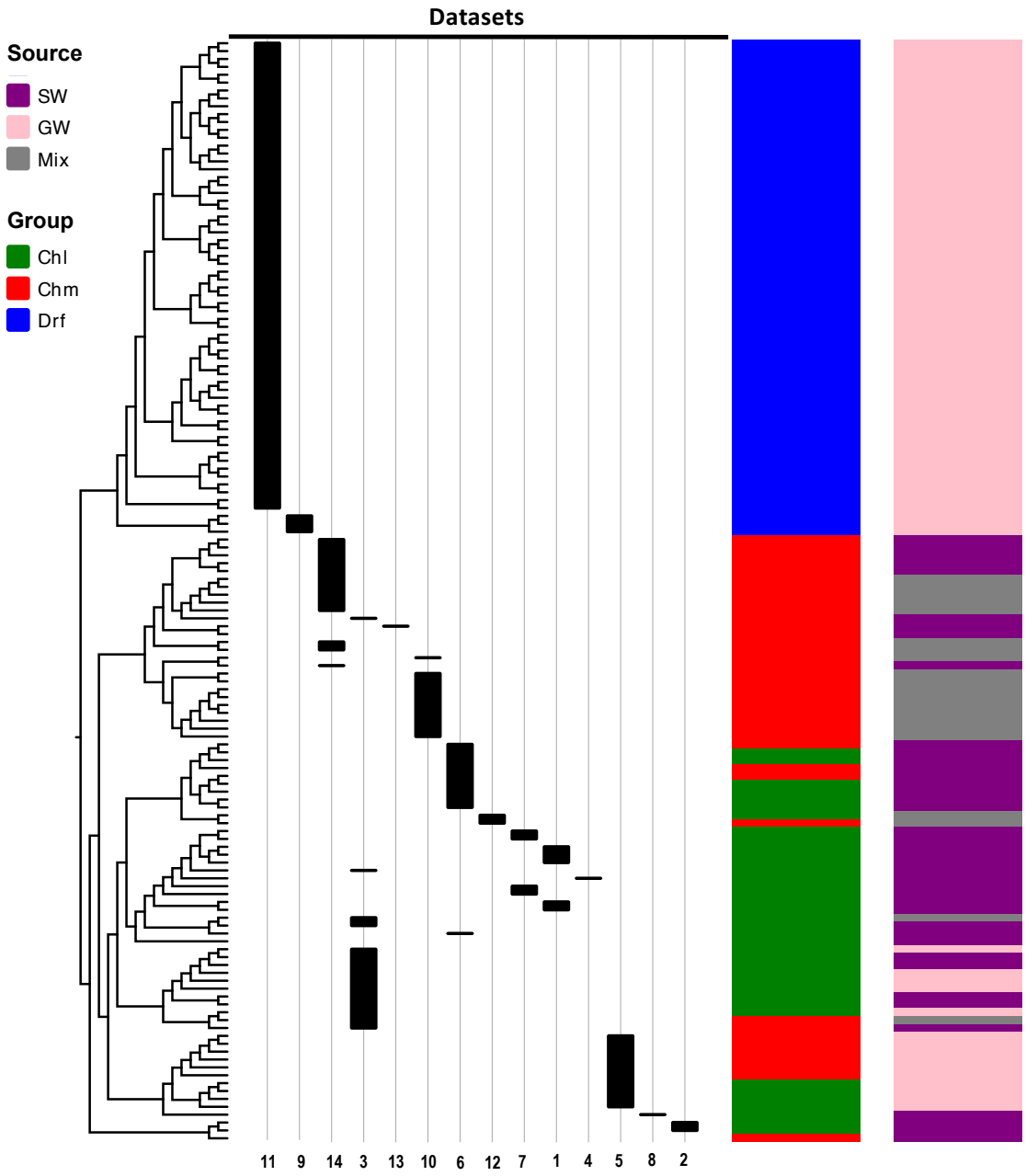

Figure 3.9. Dendrogram of sampling locations generated with Bray Curtis distances and UPGMA clustering method.
Color legends indicate type of source water (SW: surface water; GW: groundwater; Mix: mixed source water including surface water, groundwater and desalinated seawater), and disinfection group (Chl: chlorinated, Chm: chloraminated, Drf: disinfectant residual-free). Reference numbers of each dataset at the bottom of the plot correspond with data set number in Table A1.

3.3.9. Predicting functional profiles across disinfection strategies

The utility of Tax4Fun (Aßhauer et al., 2015) which leverages the KEGG database (Ogata et al., 1999), was tested to capture differences in the metabolic potential of microbial communities in disinfected and disinfectant residual-free systems. The OTU sequences from disinfectant residual-free samples exhibited significantly lower similarity to organisms in the KEGG database; this was despite the fact that only sequences matching the SILVA database were used for this exercise. Specifically, greater than 80% of the disinfectant residual-free sampling locations had less than 50% of sequences matching organisms in the KEGG database (Figure 3.10), while for the disinfected group 35.3±24% of the sequences per sample had no match.

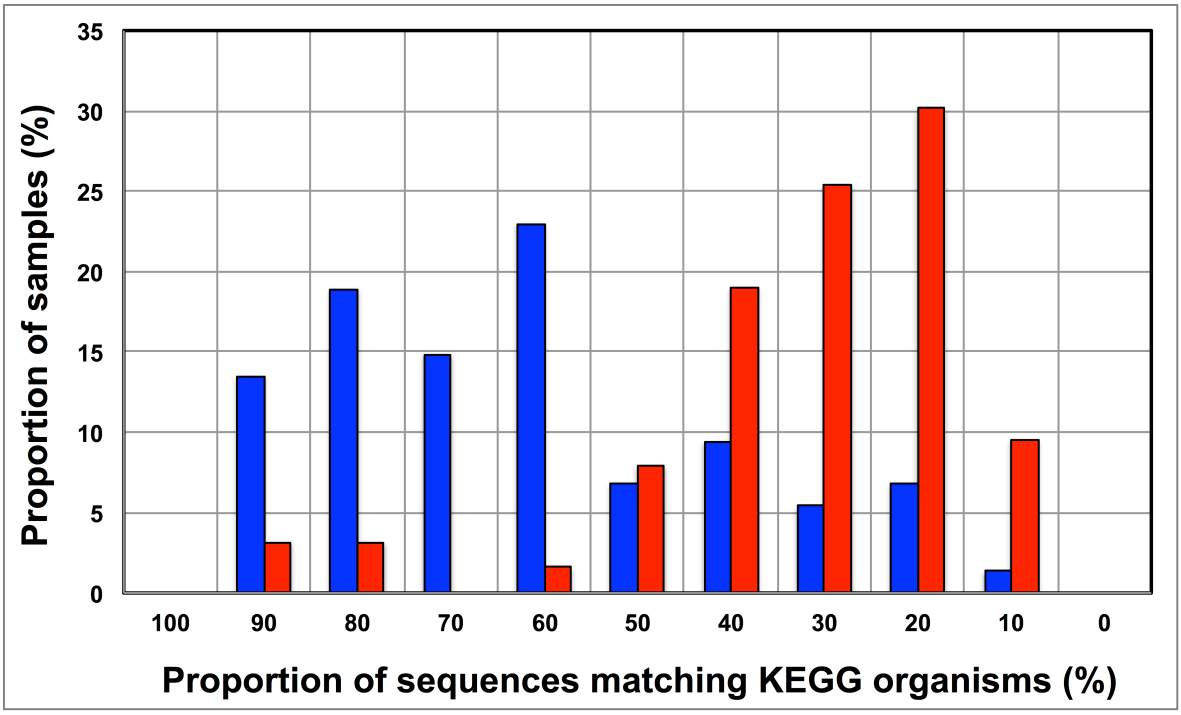

Figure 3.10. Proportion of sequences matching organisms in the KEGG database (%) versus proportion of samples (%) for disinfected (in blue) and non-disinfected (in red) datasets. The proportion of sequences matching KEGG organisms was estimated as (1-FTU)*100, where FTU= Fraction of OTUs that could not be mapped to KEGG organisms as estimated by Tax4Fun.

Further, the utility of this approach was tested to detect relevant differences between samples that may be related to the presence and absence of a disinfectant residual. To adjust for this range of sample FTUs, a FTU threshold of 0.5 was established, with 10 disinfectant residual-free sampling locations meeting this threshold. Additionally, 5 chlorinated and 5 chloraminated sampling locations were selected such that there was no significant difference in the FTUs between disinfected and disinfectant residual-free

locations used for this exercise (p=0.83). Using this subset of samples (n=20), differences in relative abundance of KOs (i.e. gene level) were tested for disinfected and disinfectant residual-free sampling locations. Of the 100 most abundant KOs returned by Tax4Fun, only 17 showed significant difference in relative abundance between disinfected and disinfectant residual-free locations (corrected p-value <0.01). No genes involved in oxidative stress or detoxification (Chao et al., 2013) were significantly different between disinfected and disinfectant residual-free locations. The majority of these significantly different KOs were associated with functions that are widely distributed across bacterial populations (e.g. carbohydrate metabolism, DNA repair, etc.). Further, though the difference in relative abundance of these KOs was significant, the magnitude of difference between disinfected and disinfectant residual-free was less than 2 fold for a majority and hence, may not necessarily provide informative insights about the selection pressure exerted by a disinfectant residual. Only one KO showed a significant difference (p=0.0073) with a large effect size in terms of relative abundance to merit follow-up investigations. Specifically, K06994, a putative drug exporter gene within the resistance-nodulation-cell division (RND) superfamily was >30 times more abundant in disinfected locations as compared to disinfectant residual-free locations.

3.4. Discussion

In order to combine and analyze together the heterogeneous datasets collected, a pre-processing step of sequence matching to the SILVA database was applied; this reference-based approach has the limitation that the subsequent analysis becomes database dependent, and the results are therefore constrained to the taxonomic groups present in the database used as reference. Nevertheless, the SILVA database is the most comprehensive and the best curated dataset available. For example, SILVA includes all the phyla that are in the Greengenes database plus *Korarchaeota* and two candidate divisions not included in Greengenes; furthermore, it contains more taxa at the genus level than Greengenes (Yilmaz et al. 2014). In addition, the lower proportion of data with matches to the SILVA database observed for Drf systems suggests that disinfectant residual-free DW systems harbour bacterial diversity that is not well represented in 16S rRNA gene reference databases and will render reference based OTU picking approaches vulnerable to poorly capturing overall diversity. However, this observation should be treated with caution as a majority of the samples from the Drf dataset emerge from a single comprehensive study (Roeselers et al., 2015). Moreover, the confounding aspect of variation between studies observed in this meta-analysis is a common theme across meta-analysis efforts (Adams et al., 2015, Koren

et al., 2013, Shade et al., 2013). Despite these limitations, all the preserved differences observed between samples with and without a disinfectant residual could in large part be attributable to the selective pressures exerted by the process of disinfection on the DW microbial community (Roeselers et al., 2015, Wang et al., 2014b, Hwang et al., 2012a).

The positive occupancy-abundance relationship observed in the Chl, Chm and Drf groups suggests that if an OTU is found to be abundant in a system within a microbial growth control strategy, it is likely to occur widely in similar systems. A similar relationship between relative abundance and occurrence of OTUs has also been reported recently (Pinto et al., 2014, Ling et al., 2015), with proposals of the utility of occupancy-abundance based modelling approaches towards microbial management in DW systems (Pinto et al., 2014).

Regarding the detected OTUs identified as potential opportunistic pathogens, it is important to note that genus level classification though informative is not indicative of the presence of pathogens. For example, the genus *Legionella* contains in excess of 50 characterized species (Burstein et al., 2016) with less than half posing a health risk and even fewer species ever isolated from treated DW (van der Wielen and van der Kooij, 2013, Wullings et al., 2011). The same is true for bacteria within the genus *Mycobacterium* and *Pseudomonas*. As a result, our findings don't suggest that one disinfection strategy is better than the other from the "pathogen" perspective. Rather, these findings should encourage rigorous follow-up studies that use standardized protocols with species-specific primers; this approach would allow for a more accurate quantitative assessment of the occurrence and absolute abundance of the pathogens of interest at DW systems that span the three disinfection strategies. Accurate identification and quantification of opportunistic pathogens in drinking water could contribute to elucidate the epidemiology of the diseases caused by them (as seen in Section 2.2.3), and propose interventions to minimize exposure and therefore risk.

OTUs classified as *Nitrospira*, a genus that includes both strict NOB and the newly discovered comammox bacteria (Pinto et al., 2015, van Kessel et al., 2015, Daims et al., 2015), were detected at a higher relative abundance and frequency than either AOB or NOB in disinfectant residual-free systems. Given this finding, it is likely that comammox bacteria may play a significantly more important role in nitrification in DW systems (either DWTP or DWDS), as compared to strict AOB and NOB. Moreover, the wide-scale detection of bacteria with a predatory lifestyle is particularly interesting as it highlights a poorly explored ecological dynamic within DW systems and may even provide an avenue

for microbial growth control (Sockett and Lambert, 2004) in the DWTP/DWDS. A possible explanation for the higher abundance and detection frequency of predatory bacteria in non-disinfected systems could be the higher biomass present in these systems, as this provides a rich source of nutrients for predatory bacteria.

Though a majority of studies include negative controls during the sample processing, DNA extraction, and PCR amplification step, these negative controls are rarely included during the sequencing process itself. To our knowledge, only one DW study has explicitly stated the inclusion of a negative control during the sequencing process (Ji et al., 2015); in this study, though the number of sequences in the negative controls were significantly lower than the samples of interest, the classification of OTUs detected in negative controls was highly similar to those commonly detected in DW samples. Regarding the sequences associated with contamination found in this meta-analysis, the lower proportion of contamination data removed from non-disinfected datasets could be related to higher biomass concentration in these samples. It is important to note that these numbers do not accurately reflect levels of contamination in published DW datasets. What this exercise emphasizes is that the need to routinely sequence negative controls is particularly critical for DW studies, not only because of the low-biomass nature of these samples but also because bacteria associated with kit/reagent contaminating genera are also commonly found in DW samples. As a result, a genuine contaminant might be passed off as belonging to the DW sample under consideration and vice versa.

Increasingly 16S rRNA gene data is being utilized to leverage functional datasets to predict the metabolic characteristics of whole microbial communities using tools such as Picrust (Langille et al., 2013), Tax4Fun (Aßhauer et al., 2015), etc. Such approaches rely on matching 16S rRNA gene sequences to organisms represented in functional databases and using the abundance of associated OTUs to predict the metabolic potential of a given microbial community. Though this is a rather cost-effective and hence attractive way to get more information for less resource (16S rRNA gene studies are significantly inexpensive as compared to metagenomic studies on a per sample basis), there is also potential for over or under-predicting the metabolic potential of the microbial community depending on the composition of these functional databases and the sample under consideration. The results obtained in this meta-analysis clearly indicate that the metabolic potential of DW microbial communities will be vastly under-represented by function predictions tools that leverage 16S rRNA gene data, particularly for disinfectant residual-free systems. Therefore, for DW

samples, these tools should be applied with caution and awareness of their limitations in order to avoid misinterpretations of the estimated microbial community metagenome.

3.5. Conclusions

A meta-analysis of microbial communities in DWDSs was carried out, using 16S rRNA-based sequencing datasets from bulk drinking water samples, to compare chlorinated, chloraminated and disinfectant residual-free systems. This has resulted in several insights, both novel and one that confirm conclusions by previous studies:

- The samples from disinfectant residual-free systems were richer, more diverse, and more even as compared to the samples with a residual disinfectant.
- *Proteobacteria*, particularly *Alpha*- and *Betaproteobacteria*, dominate drinking water bacterial communities irrespective of origin of study and presence/absence of or disinfectant residual type.
- A higher occurrence of *Legionella* OTUs in disinfectant residual-free systems and of *Mycobacterium* and *Pseudomonas* OTUs in disinfected systems was found, being this finding a prime candidate for follow-up investigations.
- The broad detection of *Nitrospira* OTUs and OTUs linked to predatory bacteria may provide for exciting avenues for future research involving fundamental ecological questions with a significant practical impact (e.g. revisiting nitrification in drinking water systems in light of new findings regarding comammox *Nitrospira*, exploring the potential of predatory bacteria for biocontrol).
- The critical aspect of including negative controls in sequencing efforts for DW studies has been highlighted.
- This meta-analysis effort is significantly confounded by data heterogeneity, particularly with respect to the ones we can identify based on the data. If all data included in this study was obtained from standardized protocols spanning sample collection, DNA extraction, PCR amplification, target hypervariable region of the 16S rRNA gene, and sequencing platform, the insights generated would be much more robust and the data would lend itself to asking targeted and quantitative questions which is currently not possible. Thus making a case for standardized protocols across all DW studies as an attractive prospect. However, efforts to standardize protocols without appropriate resources to sustain and support them are likely to be more disruptive than beneficial. For example, they may "price-out" some researchers from collecting data that meets field-approved standards. Moreover, standardizing protocols in a rapidly changing methodological landscape

presents the pitfalls of generating "kit monopolies" (i.e. one reagent or sample processing kit becomes the default), while also risking the creation of methodological inertia in a field that has only recently begun to exploit the power of high-throughput DNA sequencing. For example, consider the rate at which DNA sequencing approaches have changed over the last few years; despite the fact that Sanger sequencing was widely used for DW microbial studies until 2010, we have not included that data in this study because of its low-throughput nature (low sequencing depth and sample diversity). Similarly, it is likely that with the advent of long-read sequencing technologies, a meta-analysis effort five years from now might choose to exclude data generated from currently popular sequencing platforms due to their short-read nature and hence, lower phylogenetic resolution of the data. In light of the heterogeneity in sampling protocols found, and the rapid advancement of the field, I would suggest that researchers choose sample/data collection and processing approaches that are methodologically robust based on best-available information, and achievable given resource availability.

- Efforts should be made to: (i) standardize data reporting approaches by depositing raw data in publicly available databases; and (ii) measure and provide supporting parameters as possible (temperature water chemistry, ATP, cell counts, TOC, AOC, etc.) along with sample metadata in a format that can be easily integrated into sequence data processing approaches and diversity analyses. The practice of open data sharing is important because it would support comparative analyses across systems; my experience conducting this meta-analysis revealed that data-sharing standard practices are not yet commonplace within the DW community.
- Finally, another possible option to support comparative analyses across systems would be to make provisions for sample sharing, either DNA extract or filtered sample itself. Although this still retains DNA extraction or sample collection variabilities, it will eliminate primer and sequencing platform biases and allow for robust de-novo clustering for microbial community analyses, with the ability to assess the aforementioned biases using statistical approaches.

4. Assessing the impact of methodology on the observations of DW microbial communities

4.1. Introduction

Drinking water distribution systems (DWDSs) are no longer considered as simple water conveyance systems, but as "reactors" (Camper & Dirckx 1996; Liu et al. 2013) where a range of chemical, physical, and biological forces influences the microbial abundance and diversity, and the quality of water at the consumer's tap. As seen in Section 2.1.2, microorganims in DWDSs can be present either in planctonic state or in biofilm, attached to surfaces (e.g. particles, pipe wall), and parameters such as disinfectant concentrations and Assimilable organic carbon (AOC) concentrations limit their survival and growth. Parallel and deep nucleic acid sequencing techniques continue to reveal the complexity of the drinking water (DW) microbiome, unveiling the presence of diverse and abundant microbial communities in full-scale DWDSs (Pinto et al. 2014).

The typical workflow in amplicon-based studies consists of sample collection and concentration through filtration, DNA extraction, Polymerase chain reaction (PCR) and finally DNA sequencing. Previous studies have used different approaches at each stage of this workflow, including different types of filters for sample concentration, DNA extraction methods (Roeselers et al. 2015; Lin et al. 2014), PCR amplification protocol and the primers used to target a particular region of the 16S rRNA gene (Chakravorty et al. 2007; Vasileiadis et al. 2012), and sequencing platforms (Shaw et al. 2015; Roeselers et al. 2015). The variabilities in the observed DW microbial communities captured through replicate sample/PCR reactions have been to our knowledge unexplored. Though it may be difficult to correct for these variabilities, it is imperative that sufficient replication efforts at all/appropriate steps are undertaken. This is particularly critical at small spatial and temporal scales, where the magnitude of changes under investigation may be significantly affected by methodological biases, leading to inaccurate and/or incomplete conclusions.

At the end of the system, premises plumbing has been identified as a niche for the proliferation of opportunistic pathogens (Falkinham et al. 2015) due to its characteristics (e.g. higher temperatures, lower disinfectant residual, higher surface area to volume ratio, etc.). Therefore, an appropriate sampling approach is required to accurately assess the microbial community characteristics and identify potential public health risks at the point of use (POU).

In full-scale premises plumbing, water stagnation has been linked to microbial regrowth (Lautenschlager et al. 2010) and changes in community structure (Ji et al. 2015); the potential impacts of flow regimes similar to typical consumer practices (e.g. fully opening tap or "high flow" condition, vs. half-opening tap or "low flow" condition) on the DW microbiome are, to our knowledge, still unexplored. Moreover, the impact of sample volume on DW microbial community analyses has been, to our knowledge, unexplored. The available literature shows that for disinfected systems the selected sample volumes are between 600 ml (Shaw et al. 2015) and 100 L (Liu et al. 2016) while in non-disinfected systems typical sample volumes are between 1000 ml (Roeselers et al. 2015) and 4000 ml (Lautenschlager et al. 2014).

DW sampling campaigns are costly and laborious, in part due to the low biomass content of DW compared to other environments. For instance, the filtration step can take hours if the volume of water filtered is high. Therefore, the identification of potential variabilities introduced by the methodology selected is of vital importance in order to accurately capture the DW microbiome, produce robust and unbiased results, and to use efficiently the time and resources available to the researcher. In this chapter I investigated the impact of sample replication, PCR replication, sampling volume and flow rate on the observations of the DW bacterial community composition. As seen previously, there is considerable variability on the selection of the aforementioned parameters in DW microbial studies. For instance, sample volumes are taken over a broad range, from 0.6 L to 100 L, without previous knowledge of how this variability may impact the observations. Moreover, collecting and filtering these volumes of drinking water takes time, effort and resources. Therefore, optimizing the sampling protocol by collecting an appropriate volume of water could contribute to both a more accurate description of the microbial community, and to a more efficient sampling strategy by improving sampling campaign planning and resource allocation. In the case of flow rate, the value adopted during sampling corresponds to a specific shear stress value depending on the pipe diameter, and this shear stress in turn could cause biofilm detachment from the pipe walls. Such potential impacts of flow rate and its associated shear stress have not been assessed in full-scale building plumbing using DNA sequencing-based approaches, and this constitutes the motivation to investigate them in the present study. Finally, replication is essential in order to assess variability with confidence (Prosser 2010). For instance, if the drinking water produced by two treatment plants is compared using only one water sample from each plant, the differences in relative abundance of OTUs could be real differences among the treatment plants, or could be caused by within-plant variability, without replicates it is impossible to know the

difference. The same aforementioned example of sample replication and variability applies to PCR reactions. Therefore, both sample replication and PCR replication have been explored in the context of DW.

Furthermore, the data obtained through replication at multiple levels was used to elucidate the dynamics of the DW microbiome over a small spatial scale (i.e. one distribution zone) and a small temporal scale (i.e. the diurnal scale). Specifically, the objectives were: (i) to investigate the variation in PCR/sample replicates; (ii) to investigate the effect of low/high flow regimes on DW bacterial community abundance and diversity; (iii) to identify the minimal representative volume of sample from POU; and (iv) to elucidate small spatial and temporal dynamics of the DW microbiome through replication.

4.2. Materials and methods

Two sampling campaigns were carried out: Sampling campaign 1 was carried out to explore the impacts of PCR replication and sample replication, while Sampling campaign 2 was carried out to study the impacts of sample volume and flow rate. Details of both sampling campaigns are provided below. Figure 4.1 provides a schematic overview of the sample collection and processing steps of both sampling campaigns.

4.2.1. Drinking water sampling

Sampling campaign 1: drinking water samples were collected in August 2013 from faucets in five sampling locations in Scotland, United Kingdom. Sampling locations A, B, D and E and are located in Glasgow in the same DWDS. Sampling location C is located in Kirkintilloch, 10.76 Km away from Glasgow. Sampling locations A, B, D and E are supplied by Plant X that treats surface water through coagulation, rapid gravity filtration, orthophosphate addition and disinfection (chlorine). Sampling location C receives water from two plants: Plant Y, with a similar configuration to Plant X; and Plant Z, similar to Plants X and Z with two additional treatment processes (dissolved air flotation, and secondary filtration). At each sampling location, four-hour composite samples were collected over a 24-hour period, resulting in six sampling time periods per day (08-12, 12-16, 16-20, 20-00, 00-04, and 04-08). Prior to sampling, the faucets and sinks were thoroughly disinfected with sodium hypochlorite and the faucet was flushed for 10-15 minutes in order to avoid any impact from stagnant water in the premise plumbing at the sampling locations (Lautenschlager et al. 2010). After flushing, the faucet was adjusted to a flow rate of approximately 200-400 ml/minute, which was maintained constant for the

duration of the sample collection (i.e. 24 hours). Drinking water was pumped from a sterile beaker placed under the running faucet, using a peristaltic pump (speed = 75 RPM) fitted with sterile tubing and connectors, to three sterile Sterivex filters with 0.22-µm pore size polyethersulfone membrane (Millipore, Billerica, MA). A total of 13-17 liters of sample was filtered through each of the triplicate filters for each sampling time period. Fresh sterile tubing and fittings were used for each filter at every sampling time-window. Following filtration, the membranes were immediately removed from the filter casing using an aseptic technique, placed in lysing matrix E tubes (MP Biomedicals, Santa Ana, CA) and stored at 4°C for a maximum of 24 hours before being transferred to a -20 °C freezer.

Sampling campaign 2: drinking water samples were collected from five different locations (F, G, H, I and J, four households and one office building) in Glasgow (Scotland). All sampling locations are located in the same distribution zone, and supplied by Plant X. Samples were collected from the 29th of July to the 8th of August of 2014, between 10AM and 4PM. Before sample collection, each tap and sink were thoroughly cleaned with 70% ethanol and water was flushed for 10 minutes until it reached stable, minimal temperature, to avoid stagnation effects (Lautenschlager et al., 2010). All sample bottles and filtration equipment were autoclaved prior to each sampling. In each location, samples were collected at low (0.5 L/minute) and high (5.0 L/minute) flow rate, with estimated shear stress values of 0.0006-0.0491 N/m2 and 0.0065-0.491 N/m2 for the low and high flow rates, respectively (Appendix B, Table B1). For each flow regime, five sampling volumes (1, 2, 10, 15 and 20 L of water) were filtered through sterile 0.22-µm Sterivex filter units (Millipore, Billerica, MA) using peristaltic pumps (Watson-Marlow Bredel Inc, Wilmington, MA) fitted with sterile tubing (Saint-Gobain Tygon, Charny, France). After filtration, the filter membranes with collected biomass were aseptically removed and transferred to sterile 2 mL lysing matrix E tubes (MP Biomedicals, Cambridge) and kept at 4ºC. Upon arrival at the laboratory, the lysing matrix E tubes were stored at -80ºC. In total, 50 bulk biomass samples were collected (5 locations, 2 flow regimes for each location, 5 volumes per flow regime).

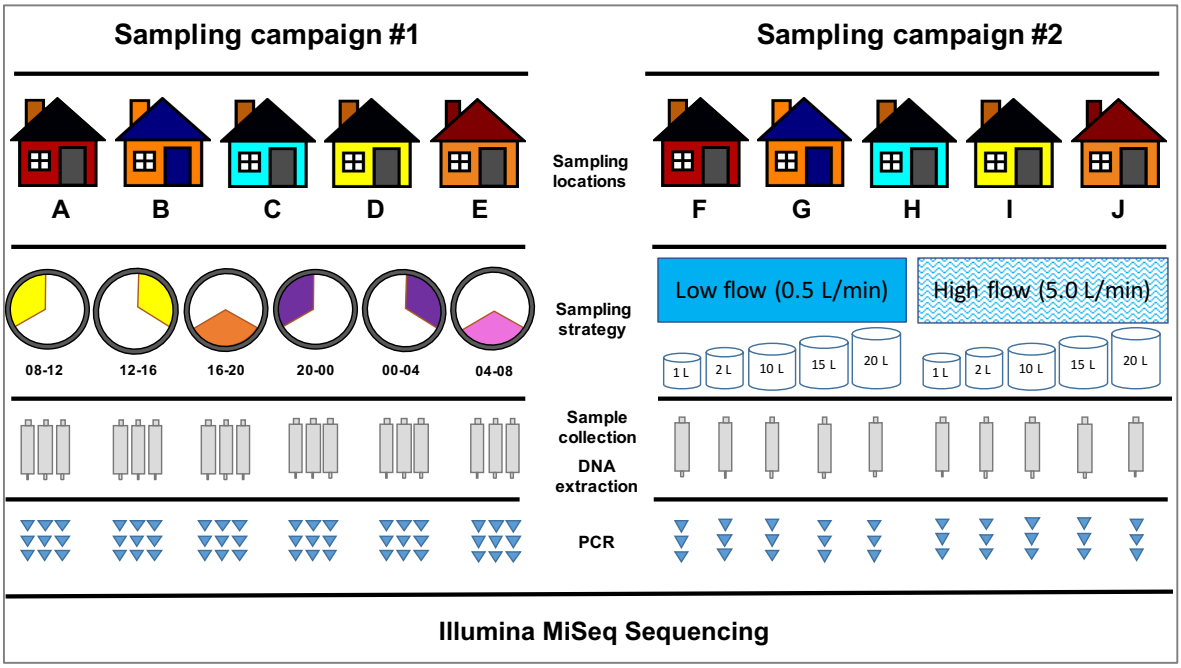

Figure 4.1. Overview of sampling strategy for sampling campaigns 1 and 2. Sampling campaign 1 explored the impacts of PCR replication and sample replication; Sampling campaign 2 explored the impacts of sample volume and flow rate.

4.2.2. DNA extraction

Sampling campaign 1: DNA was extracted from the filters using an adaptation of a previously described phenol-chloroform method (Pinto et al. 2012), with some modifications. Briefly, 300 µl of 2xTENS buffer and 500 µl of phenol:chloroform:isoamyl alcohol (25:24:1, pH 8.0) were added to the lysing matrix E tube containing the filter, followed by vortexing, bead beating for 40s at 6 m/s using a FastPrep 24 instrument (MP Biomedicals, Santa Ana, CA, USA), and centrifugation at 14,000×*g* for 10 minutes. The supernatant was transferred to a pre-spun heavy phase lock gel tube (5 Prime GmBH, Hilden, Germany) and two more bead beating steps were performed with the replacement of the aqueous phase with fresh 200 µl of 2xTENS buffer, prior to each bead beating. The subsequent DNA purification was carried out as described previously (Pinto et al. 2012). The extracted DNA was quantified using a Qubit 2.0 Fluorometer (Life Technologies, UK).

Sampling campaign 2: DNA was extracted from the filters using a combination of a phenol-chloroform method previously described (Pinto et al., 2012), and the protocol of the Maxwell® 16 LEV Blood DNA Purification Kit (Promega, Madison, WI, USA). The modified protocol consisted of the following steps: (i) the filters with collected biomass were incubated with 300 µL lysis buffer and 30 µL proteinase K at 56°C for 20 min; (ii) addition of 500 µL chloroform:isoamyl alcohol (25:24:1, pH 8.0); (iii) bead beating (6 m/s

for 40 sec) using FastPrep 24 instrument (MP Biomedicals, Santa Ana, CA, USA); (iv) the tube was centrifuged at 14,000×*g* for 10 min; (v) the aqueous phase was transferred to a 2 mL safe lock tube (StarLab, MK, UK) and two more 40-second bead beating and centrifugation steps (at 12,500×*g* for 10 min) steps were performed after replacement of the aqueous phase with fresh lysis buffer; (vi) 500 µL chloroform:isoamyl alcohol were added to the aqueous phase tube and the tube was centrifuged at 14,000x*g* for 5 min. The DNA suspended in lysis buffer was automatically purified and suspended in 50 µL of elution buffer by the Maxwell® 16 instrument, using the cartridges provided in the kit. The amount of extracted DNA from each sample was quantified in triplicate on a Qubit 2.0 fluorometer (Life technologies, UK). All the DNA samples were stored at −80°C. In addition to these samples, a negative control for each run of the Maxwell instrument was included consisting of a filter membrane that had not been used for sample filtration.

4.2.3. PCR amplification and DNA sequencing

Sampling campaign 1: The v4 hypervariable region of the 16S rRNA gene was amplified from the DNA extract from each sample using three different barcoded reverse primers (806R) and the same forward primer (i.e. 515F) (Caporaso et al. 2012). A negative control was included with every set of PCR reactions to check for contamination (Salter et al. 2014). PCR reactions were set up with template DNA (0.225-11.85 ng), half a volume of KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, MA), forward and reverse primers (final concentration: 0.5 µM) and nuclease free water in 20 or 30 µl reaction volumes. The PCR reaction conditions were as follows: 95 °C for 5 min; 30 cycles of 98 °C for 20 s, 54 °C for 30 s and 72 °C for 40 s; and a final extension step of 72 °C for 1 min. The triplicate PCR products from each primer combination were pooled, analyzed on 1.0% agarose gel, size selected, and purified using the QIAquick Gel Extraction Kit (QIAGEN, UK). All negative controls were negative for PCR amplification. The purified products were then quantified on a Qubit 2.0 Fluorometer, pooled in equimolar proportions, and submitted for sequencing to the Centre for Genomic Research (University of Liverpool) on three lanes of an Illumina MiSeq platform using the v2 chemistry.

Sampling campaign 2: DNA extracts were PCR amplified using three different barcoded reverse primers (806R) and the same forward primer (i.e. 515F), targeting the V4 hypervariable region of the 16S rRNA gene (Caporaso et al. 2012). Each PCR reaction mix contained 15 µl of KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, MA), 1.6 µl of the forward primer, 1.6 µl of reverse primer, 3 ng of DNA template, and PCR

grade water to a total volume of 30 µl. The PCR thermocycling conditions were as follows: initial denaturation of 5 min at 95°C; 25 cycles of 98°C for 20 s, 62°C for 40 s, 72°C for 40 s; a final extension at 72°C for 1 min. A negative control was run for each sample and primer set to check for contamination (Salter et al. 2014). The PCR products from each primer combination were pooled, analyzed on 1.5% agarose gel, size selected, and purified using the Zymoclean Gel DNA Recovery Kit (Zymo Research, Orange, CA). All negative controls were negative for PCR amplification. The purified amplicons were quantified on a Qubit 2.0 fluorometer (Life technologies, UK), pooled in equimolar proportions and sent to the Centre for Genomic Research (University of Liverpool, United Kingdom) for sequencing on the Illumina MiSeq platform using the v3 chemistry.

4.2.4. Water quality analyses

Sampling campaign 1: water temperature, pH, dissolved oxygen, and conductivity were measured at the sampling locations using an Orion 5 Star Meter (Thermo Fisher Scientific, Waltham, MA), while total chlorine was measured on site using Hach's Reagent powder pillows (Hach Lange, UK) and a DR 2800 VIS Spectrophotometer (Hach Lange, UK). A 2-liter composite sample of drinking water was collected over each four-hour sampling period for water quality analyses. These four-hour composite samples were collected in sterile Nalgene bottles, maintained at 4°C at the sampling location, and were processed and stored according to standard protocols immediately on arrival at the laboratory (APHA 1998). Turbidity was measured immediately on arrival at the laboratory using 2100Q Portable turbidimeter (Hach Lange, UK). Total organic carbon (TOC) concentrations were determined using a Shimadzu TOC-LCPH Analyzer (Shimadzu, Kyoto, Japan). Ammonia, nitrite, and nitrate concentrations were measured using colorimetric methods 4500-NH3-F, 4500-NO2–B, and 4500-NO3–B respectively, as described in APHA (1998). Water quality data for all sampling time periods and locations is provided in supplementary information (Appendix B, Table B2).

Sampling campaign 2: water temperature, pH and conductivity were monitored at each sampling location using an Orion 5 Star Meter (Thermo Fisher Scientific, Waltham, MA), total chlorine using Hach's Reagent powder pillows (Hach Lange, UK) and a DR 2800 VIS Spectrophotometer (Hach Lange, UK). Over each sampling period (low and high), 2-liter composite samples of water were collected in sterile Nalgene polycarbonate bottles for water quality analyses. The samples were kept at 4°C at the sampling location, and were processed and stored according to standard protocols immediately on arrival at the

laboratory (APHA 1998). Ammonia, nitrite, and nitrate concentrations were measured using colorimetric methods 4500-NH3-F, 4500-NO2–B, and 4500-NO3–B respectively; orthophosphate concentrations were measured using the Test 'N Tube kit (Hach Lange, UK). Total organic carbon (TOC) concentrations were determined using a Shimadzu TOC-LCPH Analyzer (Shimadzu, Kyoto, Japan). Water quality data for all sampling time periods and locations is provided in supplementary information (Appendix B, Table B3).

4.2.5. Sequence processing and statistical analyses

Sampling campaign 1: the raw Fastq files were trimmed for the presence of Illumina adapter sequences using Cutadapt version 1.2.1 (Martin 2011). The trimmed reads were further cleaned using Sickle version 1.200 (Joshi & Fass 2011) with a minimum window quality score of 20. Any reads shorter than 10 bp after trimming and quality control were discarded. Reads passing the aforementioned quality control measures were then processed in Mothur (Schloss et al. 2009) using the protocol described by Kozich et al. (2013). Multiple statistical analyses were conducted using 7321 reads per sample. Analysis of Molecular Variance (AMOVA) (Excoffier et al. 1992) was conducted using Mothur (Schloss et al. 2009), permutational t-tests and Permutational Analysis of Variance (PERANOVA) were conducted using R (R CoreTeam 2014), while beta-dispersivity tests were conducted using vegan (Oksanen et al. 2013). Plots were constructed in R using package ggplot2 (Wickham 2009). Raw sequence data has been made publicly available in the Sequence Read Archive (SRA) database under accession number SRP058339.

Sampling campaign 2: the raw fastq files were trimmed for the presence of Illumina adapter sequences using Cutadapt version 1.2.1 (Martin 2011). The trimmed reads were further cleaned using Sickle version 1.200 (Joshi & Fass 2011) with a minimum window quality score of 20. Any reads shorter than 10 bp after trimming and quality control were discarded. Reads passing the aforementioned quality control measures were then processed with mothur (Schloss et al. 2009) using a previously described protocol (Kozich et al. 2013). After removing singletons and subsampling the OTU table, Permutational Analysis of Variance (PERANOVA) was conducted using R (R CoreTeam 2014), paired t-tests were conducted using R (R CoreTeam 2014) with Benjamini-Hochberg correction to control false discovery rate (i.e. proportion of "discoveries" that are false, based on an incorrect rejection of the mull hypothesis) (Benjamini & Hochberg 1995). while Permutational Multivariate Analysis of Variance using Distance Matrices (PERMANOVA) was conducted using vegan (Oksanen et al. 2013). Plots were

constructed in R using package ggplot2 (Wickham 2009). For the subsequent analyses, a set of filtered OTU tables were generated, removing the OTUs with a frequency less than 0.20 across the groups/conditions of interest. To estimate if there were OTUs enriched in each sampling location, the DeSeq2 package (Love et al. 2014) was applied to the filtered OTU tables per sampling location. To investigate OTU associations, SparCC (Friedman & Alm 2012), as implemented in mothur (Schloss et al. 2009), was applied to the filtered OTU tables per flow regime. The obtained correlations were visualized with the corrgram (Wright 2015) package in R. Sample clustering with the UPGMA method was made online (Edwards 2012), while the resulting dendrogram was visualized with EvolView (Zhang et al. 2012).

4.3. Results

4.3.1. Efficiency of DNA extraction and PCR amplification

For Locations A-E, PCR amplification was successful for 246 of 270 PCR libraries, of which 239 samples provided sufficient PCR product to be amenable for sequencing. We obtained a total of 15,015,570 raw paired-end reads for the 239 samples with an average of 61,877±101,201 per sample library. After trimming, quality filtering, and chimera removal using UCHIME (Edgar et al. 2011), the total number of sequences was reduced to 14,726,834 with an average of 61,619±101,067 reads per sample library. The 14,726,834 quality-filtered and chimera-free reads from 239 sample libraries clustered into 6080 operational taxonomic units (OTUs) at a 97% sequence similarity cutoff using the average neighbour clustering method for the entire dataset. All sequencing libraries exhibited coverage in excess of 99% based on the Good's estimator.

For Locations F-J, DNA extraction was successful for all volumes under both flow conditions (50 samples). The DNA yield increased as the volume of water filtered increased (Figure 4.2-A), with an average across all locations of 0.65±0.12, 0.77±0.21, 208.48±194.22, 628.70±332.97 and 1397.90±871.73 ng of DNA corresponding to 1, 2, 10, 15 and 20L of water filtered, respectively. Significant differences in DNA yield were observed between several volumes; for instance, the DNA yields of 1L and 2L samples were significantly different to all the other volumes (10L, 15L and 20L; $p<0.01$). When analyzed per flow rate, the average DNA yield was higher for the Low flow condition (543.88±773.16 ng) than for the High flow condition (350.72±551.06 ng) although this difference was not statistically significant (Figure 4.2-B). PCR was successful for all the

10-, 15- and 20 L sample volumes, with the exception of the low flow-10 L sample in location B. PCR was unsuccessful for all the filter negative controls, producing no visible bands on the gels. In total, 86 PCR libraries provided sufficient product to be amenable for sequencing. We obtained a total of 15,616,122 raw paired-end reads with an average of 181,582±91,345 reads per library. After screening, filtering and chimera removal, the total number of sequences was reduced to 12,540,720, with an average of 145,822±71,632 per library. The 12,540,720 quality-filtered and chimera-free reads clustered into 103,495 operational taxonomic units (OTUs) at a 97% sequence similarity cutoff using the average neighbor clustering method for the entire dataset. After the removal of singletons, the final OTU table included 11,481 OTUs. After subsampling the OTU table without singletons to the minimum sampling depth (43,797 reads/sample), 6882 OTUs were retained, with a Good's coverage in excess of 99% for all samples. This subsampled OTU table was used for the subsequent diversity analyses.

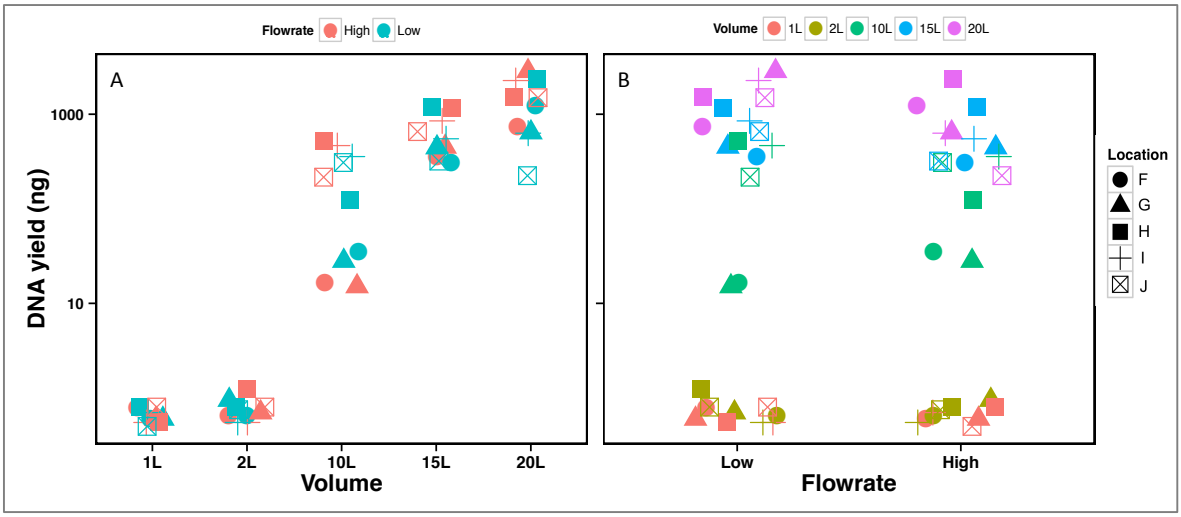

Figure 4.2. DNA yield obtained for locations F-J.
(A)DNA yield versus volume, including all locations (B) DNA yield versus flowrate, including all volumes.

4.3.2. Impact of sample replication and PCR replication

AMOVA tests were implemented to check for differences in community membership and structure between triplicate filters (for each sampling time period) and triplicate sample libraries from each filter, using distance matrices constructed with Bray Curtis (structure-based metric) and Jaccard (membership-based metric) metrics, respectively. For both structure (Bray Curtis) and membership (Jaccard) based metrics, the differences between bacterial communities obtained from replicate filters was not significantly different, relative to differences between sample libraries generated from replicate barcoded PCR reactions from the same filter after correcting for false discovery rate arising from multiple

pairwise comparisons (Appendix B, Table B4) ($p>0.05$). An evaluation of the similarities between replicate filters indicated that even though replicate filters only shared between 6 and 17 OTUs with each other (i.e. 2-4% of all OTUs across replicate filters), these OTUs constituted in excess of 99% of the reads (i.e. bacterial community abundance) in each sample (Figure 4.3-A). This feature explains the AMOVA results for the structure-based Bray Curtis distance metric. Similarly, the detection of the rare OTUs (relative abundance <0.01%) was significantly more variable ($p < 0.001$) between replicate barcoded libraries originating from a single filter compared with sample libraries from replicate filters, lending support to the AMOVA test results corresponding to membership-based Jaccard distance metric (Figure 4.3B-C).

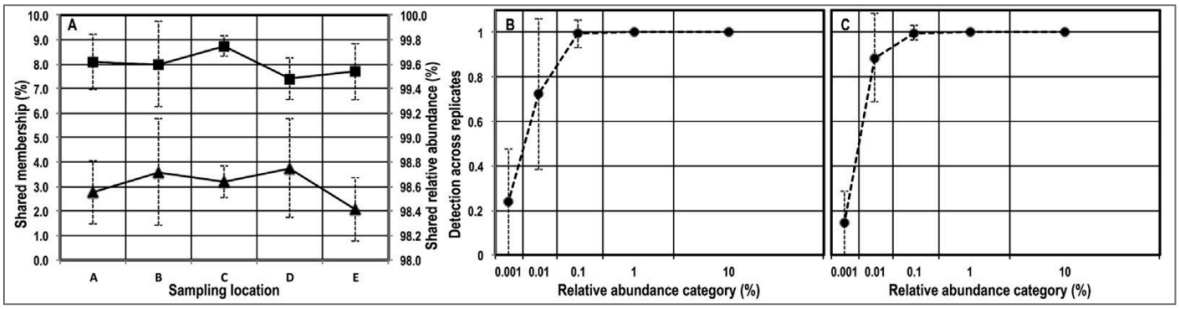

Figure 4.3. (A) The shared community membership (triangles) between replicate sample filters was significantly lower than the shared community abundance (squares). The lower shared membership was a result of the greater variability in detection of rare OTUs (<0.01%) across replicate PCR/sequencing libraries (B) and replicate filters (C).

4.3.3. Impact of flow rate and volume on richness

Three diversity estimates were calculated (number of observed OTUs, Shannon index and Inverse Simpson Index) as a measure of richness. PERANOVA results suggest that the richness in each location is similar and does not depend neither on the volume of water sampled, nor on the flow rate (Table 4.1-A), as only three significant differences were observed across locations for all the diversity estimates evaluated. For Location F the number of observed OTUs was significantly different as a function of the volume ($p<0.05$), with an average of 230±11, 182±19 and 221±38 observed OTUs in the sample volumes of 10L, 15L and 20L, respectively. In Location H the number of observed OTUs was significantly different as a function of the combination of both volume and flow rate ($p<0.05$); for instance, the average number of OTUs was 167±33 and 187±44 for the Low flow-10L sample and the High flow-20L sample, respectively. Finally, Location J exhibited significant differences in the Inverse Simpson Index as a function of flow rate ($p<0.05$), with an average of 1.27±0.12 and 1.13±0.08 corresponding to the Low and High flow conditions, respectively.

Location	Variable	A			B	
		Sobs	Shannon	Invsipmson	Bray Curtis	Jaccard
F	volume	***0.028***	0.398	0.403	0.340	0.151
	flowrate	0.444	0.497	0.696	0.146	***0.009***
	volume*flowrate	0.752	0.364	0.320	0.226	0.236
G	volume	0.771	0.747	0.699	0.685	0.986
	flowrate	0.896	0.300	0.186	0.186	0.981
	volume*flowrate	0.921	0.585	0.578	0.571	0.725
H	volume	0.153	0.883	0.874	0.767	0.824
	flowrate	0.808	0.404	0.406	0.460	***0.001***
	volume*flowrate	***0.005***	0.691	0.587	0.495	0.380
I	volume	0.604	0.902	0.917	0.949	0.996
	flowrate	0.884	0.637	0.561	0.518	0.747
	volume*flowrate	0.535	0.773	0.787	0.795	0.779
J	volume	0.833	0.926	0.963	0.970	0.893
	flowrate	0.786	0.064	***0.041***	***0.031***	***0.033***
	volume*flowrate	0.877	0.953	0.946	0.987	0.912

Table 4.1. PERANOVA and PERMANOVA results of diversity estimates for locations F-J. (A) p-values for PERANOVA results for alpha diversity estimates. (B) P-values for PERMANOVA results for beta-diversity estimates. The effect of flow rate, volume and the interaction between both was tested within each sampling location. Significant p-values values ($p<0.05$) indicated with bold italic font.

Analyzing the richness estimates within flow rate condition for each location, a similar result emerges; for locations F, G, I and J, no difference was detected among the 10L, 15L and 20L sampling volumes for the Low and High flow conditions; while in sampling location H both the Shannon and Inverse Simpson estimates were significantly different among the 10L, 15l and 20L sampling volumes for the Low flow condition ($p<0.05$). Within each location and flow condition, the estimated median richness shows different patterns; for instance, under high flow rate, the number of observed OTUs increases with volume in location H, decreases with volume in location J, and is lower for the 15 L volume than for the 10 and 20 L volumes in location F (Figure 4.4). For all sampling locations collated, significant differences among locations are observed for Sobs ($p<0.001$), the Shannon index ($p<0.001$) and the Invsimpson index ($p<0.01$), while paired t-tests with Benjamini Hochberg correction show that most of the significant differences are between location F and the rest of the locations ($p<0.05$).

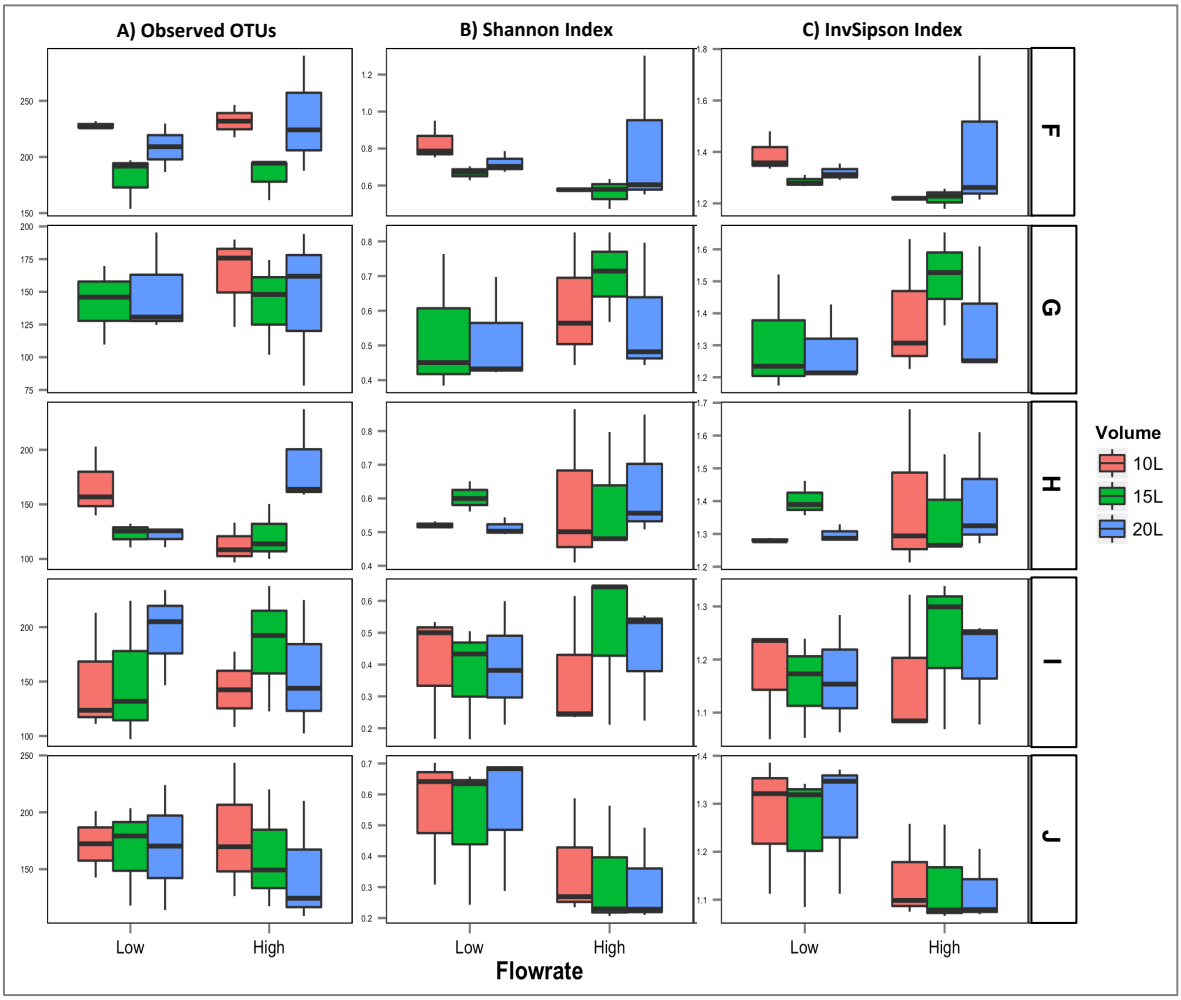

Figure 4.4 Richness estimates (Observed OTUs, Shannon index, InvSimpson index) for locations F-J, per volume, for each flow condition and sampling location. The bottom, middle, and top line of the box plots correspond to the 25%, 50%, and 75% percentiles, respectively. The top and bottom whiskers correspond to the maximum and minimum values within 1.5 x IQR (interquartile range), respectively. The dots represent outliers.

4.3.4. Impact of flow rate and volume on bacterial community structure and membership

Bray Curtis (community structure) and Jaccard (community membership) distances were used to estimate beta-diversity for both sampling flow rate (low and high) and volume at all sampling locations. Structure-based metrics weight the contribution of each OTU towards the dissimilarity between samples by the relative abundance of the respective OTU, whereas membership based metric estimate the dissimilarity between samples on a presence/absence basis. When all the sampling locations are analyzed together, the samples visibly cluster by sampling location (Figure 4.5, A-D) for Bray Curtis distances, being sampling locations F, I and J more similar than locations G and H. A stronger clustering was observed in the NMDS plots constructed with Bray Curtis distances as compared to the NMDS plots constructed with Jaccard distances.

This clustering pattern was further confirmed in Figure 4.6, which shows that within each location, the samples cluster according to the flow condition, with a few exceptions. PERMANOVA results (Table 4.1-B) showed that the variable "flow rate" could explain some of the variance in the datasets. Specifically, with Jaccard distances, "flow rate" explained 6.85%, 6.81% and 7.15% of the variance for locations F, I ($p<0.001$) and J ($p<0.05$), respectively; while "flow rate" explained 29.98% of the variance for sampling location E using Bray Curtis distances ($p<0.05$). The variance explained by the variable "volume" was not statistically significant for any of the sampling locations and beta-diversity metrics used.

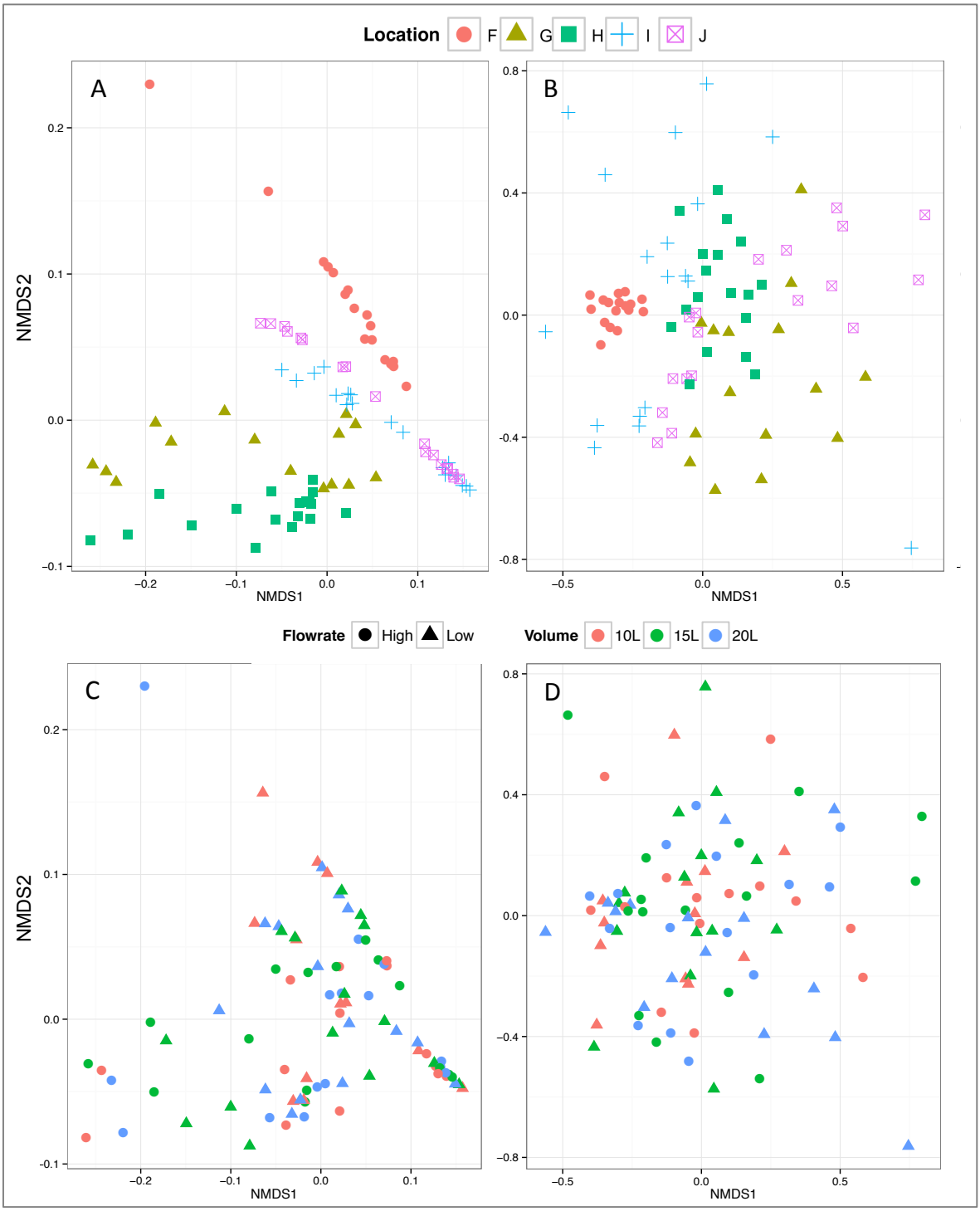

Figure 4.5. NMDS plots of sampling locations F-J, coded by location, flowrate and volume. A and C were done using Bray Curtis distances, B and D were done using Jaccard distances.

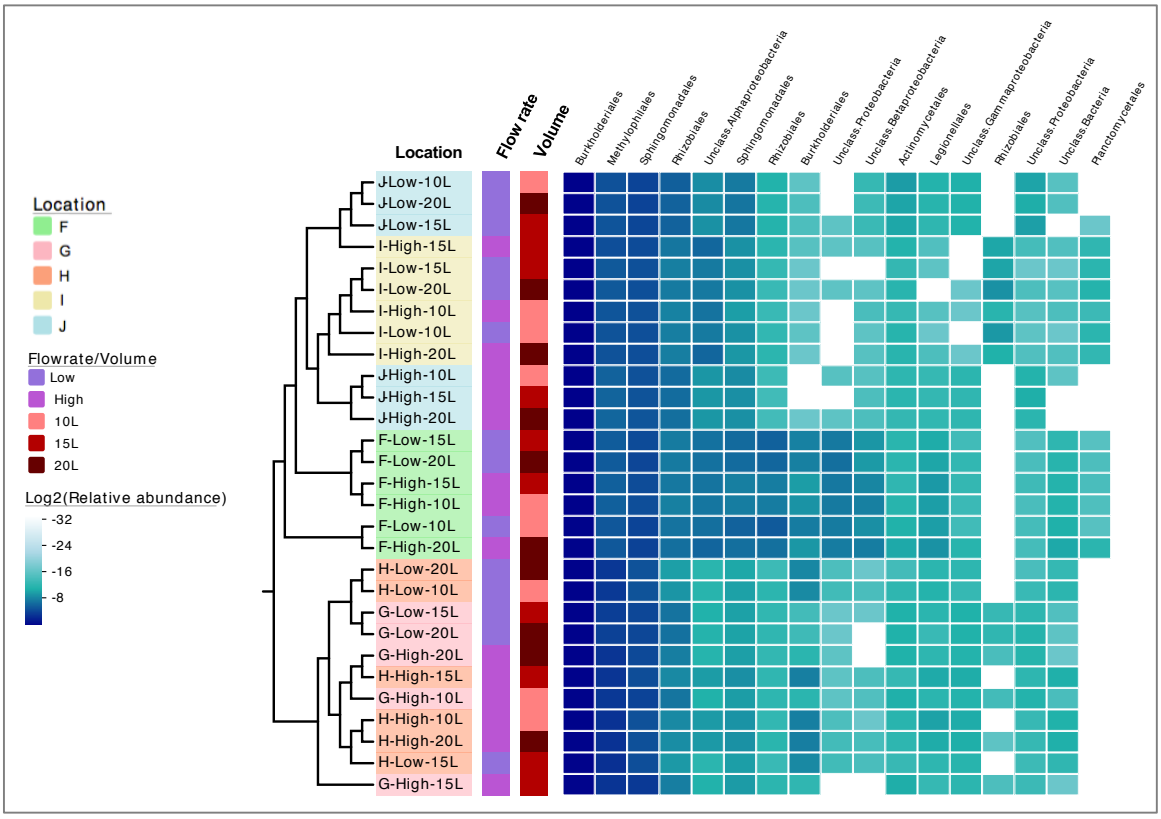

Figure 4.6. Heatmap of relative abundances of top 10 OTUs in each sampling location. The sampling location dendrogram was generated with Bray Curtis distances and UPGMA clustering method.

4.3.5. Impact of flow rate on the differential abundance of OTUs

Significant differences ($p<0.0001$) in relative abundance were observed in three of the five sampling locations; specifically, in locations F, H and J (Figure 4.7). In location F, 7 OTUs were more abundant in the Low flow condition and one in the high flow condition. Specifically, OTU_6 (family: *Sphingomonadaceae*), OTU_7 (family: *Bradyrhizobiaceae*), OTU_8 (order: *Burkholderiales*), OTU_28 (order: *Rhizobiales*), OTU_35 (family: *Bradyrhizobiaceae*), OTU_53 (family*: Chitinophagaceae)* and OTU_64 (unclassified bacteria) were more abundant in the Low flow condition; while OTU_10 (class: *Betaproteobacteria*) was more abundant in the High flow condition. In location H, OTU_73 (unclassified bacteria) was more abundant in the Low flow condition. Finally, in location J four OTUs were enriched under the Low flow condition; specifically, OTU_26 (family: *Bradyrhizobiaceae*), OTU_47 (family: *Sphingomonadaceae*), OTU_48 (family: *Hyphomicrobiaceae*) and OTU_53 (family*: Chitinophagaceae)*. For location F, Log2fold changes in abundance range between -1.16 (OTU_10) and 2.33 (OTU_53). For location H, OTU_73 had a Log2fold change of 2.87; while for location J the values ranged between 2.41 (OTU_48) and 3.63 (OTU_26).

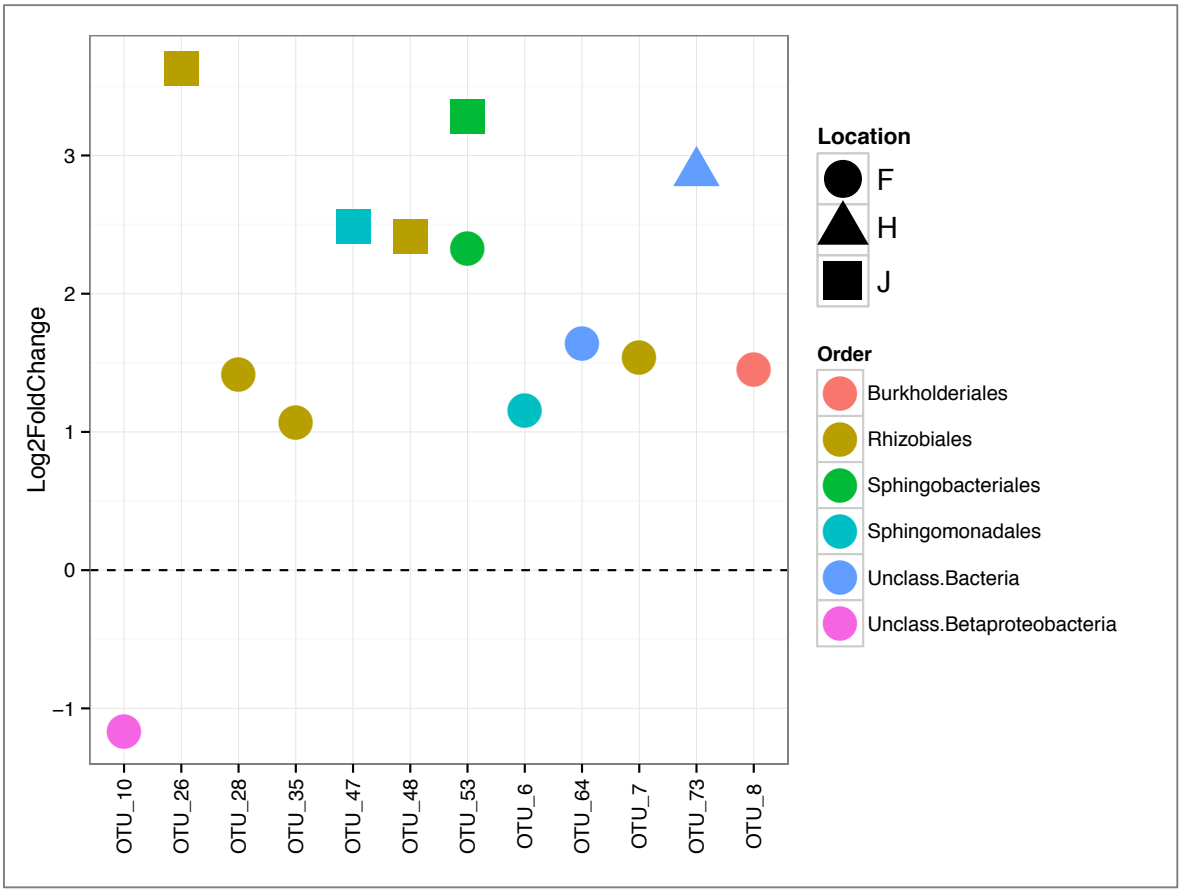

Figure 4.7. Log2-fold-change of relative abundance of OTUs that are differentially abundant (p<0.0001) in the low (positive y-axis values) and high (negative y-axis values) flow conditions. Point colour coded by taxonomic order of the OTU.

4.3.6. OTU associations in low and high flow regimes

OTU correlation coefficients greater than 0.50, less than -0.50 and statistically significant (p-value<0.01) were selected to further explore the OTU associations. In total we obtained 71 and 37 correlations for the low and high flow regimes, respectively. For the low flow regime, 40 OTUs supported 63 positive correlations and 8 negative correlations, while for the high flow regime 21 OTUs supported 33 positive correlations and 4 negative correlations. For both flow regimes, the OTUs belonged to the phyla *Proteobacteria*, *Actinobacteria*, *Bacteroidetes*, *Planctomycetes,* and unclassified Bacteria. A closer inspection revealed the presence of correlations involving several potential opportunistic pathogens (POPs) in both flow conditions (Figure 4.8), including OTUs of the genera *Sphingomonas* (n=1), *Mycobacterium* (n=1), *Pseudomonas* (n=1), and *Legionella* (n=3).

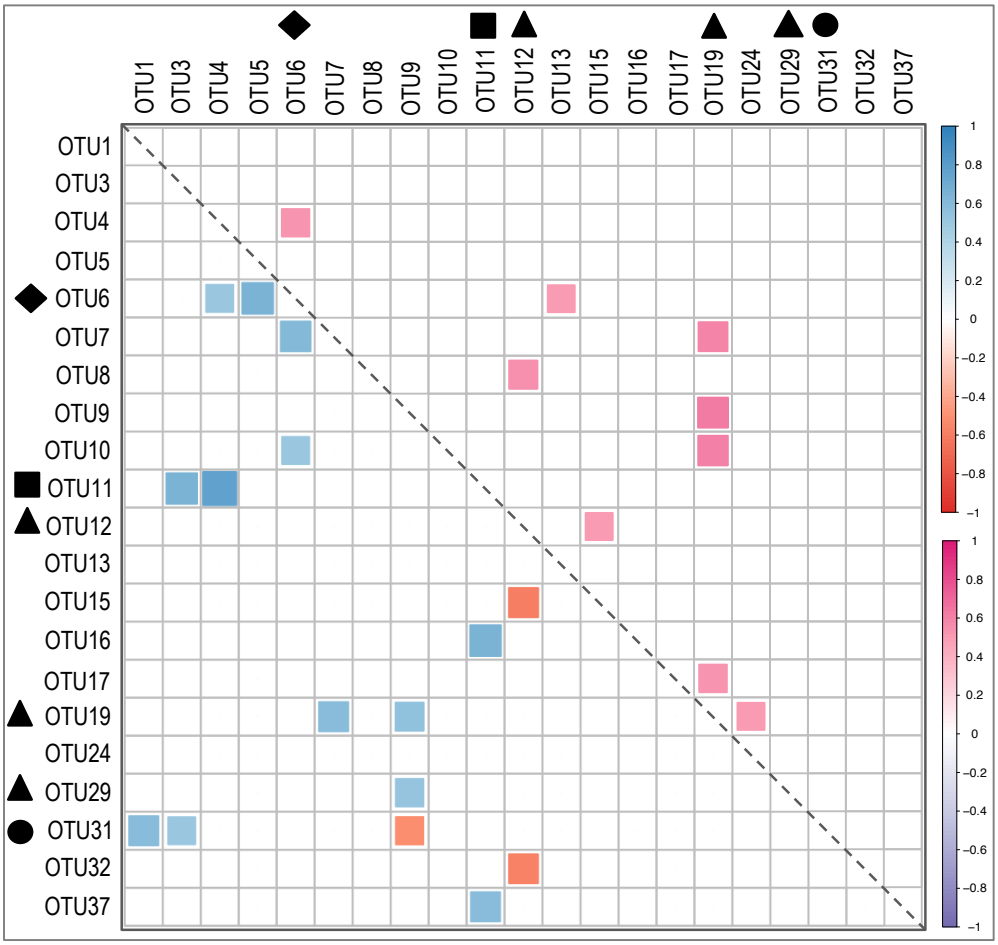

Figure 4.8. Heatmap of OTUs with significant correlation coefficients (<-0.5 and >0.50) (p<0.01) of POPs in low (lower triangle) and high (upper triangle) flow conditions. Symbols represent *Pseudomonas* (circle), *Mycobacterium* (square), *Legionella* (triangle), and *Sphingomonas* (diamond).

A higher number of potential opportunistic pathogens (OPs) and significant correlations, both positive and negative, were observed in the low flow regime; for instance, OTU11 (genus: *Mycobacterium*) and OTU4 (family: *Hyphomicrobiaceae*) have the highest correlation (0.76), followed by OTU6 (genus: *Sphingomonas*) and OTU5 (order: unclassified *Alphaprotebacteria*), and OTU11 (genus: *Mycobacterium*) and OTU16 (class: unclassified *Proteobacteria*), with correlation coefficients of 0.66 and 0.65, respectively. In the high flow condition, OTU6 (genus: *Sphingomonas*), OTU12 (genus: *Legionella*) and OTU19 (genus: *Legionella*) supported all 9 positive correlations. In both flow conditions, several OPs are correlated with OTUs that are more abundant and more frequent than them (Appendix B, Table B5). For instance, in the low flow condition, OTU31 (genus: *Pseudomonas*) is positively correlated with OTU1 (family: *Comamonadaceae*), which has a mean relative abundance (MRA) of 88.9% and a detection frequency of 1.00; while in the high flow condition OTU6 (genus: *Sphingomonas*) and OTU4 (MRA=0.537%, detection frequency=1.00) are positively correlated.

4.3.7. Bacterial community composition

Locations A-E: the bacterial communities at the five sampling locations were primarily dominated by *Proteobacteria* (>98% in all cases). For sampling locations A, B, and C, *Betaproteobacteria* (92.2±3.3%) dominated over *Alphaproteobacteria* (7.3±3.6%) for all time periods, while for locations D and E *Alphaproteobacteria* (76.6±10%) were more abundant than *Betaproteobacteria* (22.9±10%). Seven OTUs of the classes *Alphaproteobacteria, Betaproteobacteria and Actinobacteria* were detected in all samples and were among the ten most abundant OTUs for each of the sampling locations. Six of these seven OTUs were successfully classified at the family/genus level as *Comamonadaceae* (MRA=64.9±38%), *Sphingomonadaceae* (MRA=2.22±1.35%), *Hyphomicrobiaceae* (MRA=1.36±0.8%), *Methylophilus* (MRA=0.62±0.79%), *Sphingomonas* (MRA=0.24±0.05%) and *Mycobacterium* (MRA=0.1±0.07%), whereas one of the dominant and frequent OTU was only classified as *Alphaproteobacteria* (MRA=31±38.2%). The rare fraction of bacterial communities, which made up less than 1% of the total sequences in all samples, belonged to a diverse range of phyla including *Acidobacteria, Bacteroidetes, Chlamydiae, Chloroflexi, Firmicutes, Fusobacteria, Gemmatimonadetes, Planctomycetes,* and *Verrucomicrobia*. Additional information on bacterial community composition is provided in Appendix B, Table B6.

Locations F-J: the top ten OTUs across all samples, ranked according to their relative abundance, are members of the *Alpha-* and *Betaproteobacteria* classes, and represent 99.01% of the total abundance, with detection frequencies in the range of 1.00-0.64 (Figure 4.6). The most abundant OTU (OTU_1) belongs to the family *Commamonadaceae* with a MRA of 88.43±5.41% and detection frequency of 1.00, followed by a *Methylophilus* sp. (OTU_2) with a MRA of 4.87±4.06% and detection frequency of 1.00, and an OTU of the family *Sphingomonadaceae* (OTU_3) with a MRA of 3.90±1.87% and detection frequency of 1.00. Other families represented among the top ten OTUs include *Hypomicrobiaceae* and *Bradyrhizobiaceae* with 1 OTU each, while 4 OTUs are unclassified at the family level. The rare fraction of the community (relative abundance < 1%) contains OTUs of 28 classes that belong to the phyla *Actinobacteria*, *Bacteroidetes*, *Planctomycetes*, *Acidobacteria*, *Firmicutes*, *Deinococcus-Thermus*, *Verrucomicrobia*, *Chlamydiae*, *Nitrospira*, *Gemmatinonadetes*, *Fusobacteria*, *Spirochaetes* and an unclassified phylum. Within each sampling location, *Proteobacteria* dominated the DW bacterial community detected, with a relative abundance >99% in all cases. Six OTUs of the orders *Burkholderiales* (OTU_1), *Methylophilaes* (OTU_2), *Sphingomonadales* (OTU_3,

OTU_6*), Ryzobiales* (OTU_4) and an unclassified *Alphaproteobacteria* (OTU_5) were among the top 10 OTUs within each sampling location across all volumes and flow conditions. In locations G and I, an OTU classified as *Mycobacterium* was detected among the top 10 OTUs, with a MRA of 0.05±0.05% (detection frequency=1.00) and 0.02±0.02% (detection frequency=0.94), respectively; in location C, an OTU classified as *Legionella* was detected among the top 10 OTUs, with a MRA of 0.05±0.05% (detection frequency=1.00); while in location I an OTU classified as *Methylobacterium* was detected among the top 10 OTUs, with a MRA of 0.09±0.09% (detection frequency=1.00).

4.3.8. Correlation between bacterial community composition and water quality parameters

Locations A-E: The measured water quality parameters were relatively stable over the diurnal time scale for all sampling locations (Appendix B, Table B2). As a consequence, the correlations between water quality parameters and changes in whole bacterial community richness, structure, and membership were weak and did not exhibit any significance of note. Similarly, we did not detect any significant correlations between the relative abundance of the most abundant OTUs and changes in water quality parameters.

Locations F-J: At each sampling location, the physicochemical parameters of all tap water samples were relatively constant within a certain range over the 2-hour sampling periods (Appendix B, Table B3). However, an increase in total organic carbon (TOC) was detected at locations I (7.86 mg/l in low flow rate; 2.45 mg/l in high flow rate) and J (7.99 mg/l in low flow rate; 3.22 mg/l in high flow rate); while a slight increase in conductivity was detected in location I (63.0 μS/cm in low flow rate; 66.2 μS/cm in high flow rate). Within each sampling location, we did not detect any significant correlations between the richness estimators and the measured water quality parameters. Nevertheless, for all the locations combined, significant correlations were detected for the three richness estimators used (Table 4.2; Appendix B, Figure B1). The stronger correlations were observed between total chlorine and Shannon index (-0.47, $p<0.001$), total chlorine and Sobs (-0.42, $p<0.001$); temperature and Sobs (0.42, $p<0.001$), and pH and Sobs (-0.35, $p<0.001$).

	Sobs	Shannon	InvSimpson
pH	-0.35 ***	-0.27 *	-0.08
Temperature	0.42 ***	0.38 ***	0.15
Conductivity	-0.13	-0.17	-0.14
Total Chlorine	-0.42 ***	-0.47 ***	-0.20
TOC	0.07	-0.22 *	-0.28 **
NH_3	-0.30 **	-0.06	0.10

Table 4.2. Correlations between water quality parameters and richness estimators, calculated across all sampling locations.
P-values indicated by asterisks: *:p<0.05; **:p<0.01; ***:p<0.001.

4.3.9. Small scale spatial and temporal variabilities in DW microbiome revealed by sample and PCR replication

To assess diurnal changes of the DW microbiome, the Chao Index was calculated as a measure of richness for each sampling time period at sampling locations A-E (Figure 4.9). The diurnal trends in the richness of bulk water bacterial community were different for each sampling location. For example, average community richness at sampling location A was 77±33 (Lower Confidence Limit (LCI): 45±20, Upper Confidence Limit (UCI): 179±68) with minimum and maximum richness estimates of 53±18 and 115±28 observed at time periods 20-00 and 08-12, respectively. In contrast, sampling location C exhibited the lowest average richness of 41±15 (LCI: 25±7, HCI: 103±38), with minimum and maximum richness estimates of 30±8 and 54±25 at time periods 08-12 and 00-04, respectively. In light of the observed location-specific diurnal changes, we assessed whether the diurnal variation in richness was significant within any given sampling location. At sampling locations A, B and E, PERANOVA indicated significant differences in bacterial community richness throughout the day ($p<0.01$), while for houses C and D there was no significant change ($p>0.01$). Further, a comparison of the bacterial community richness at each time period across all sampling locations indicated significant differences between sampling locations for time periods 12-16, 16-20, and 04-08 ($p<0.01$). In contrast, no significant differences were observed in bacterial community richness for time periods 08-12, 20-00 and 00-04.

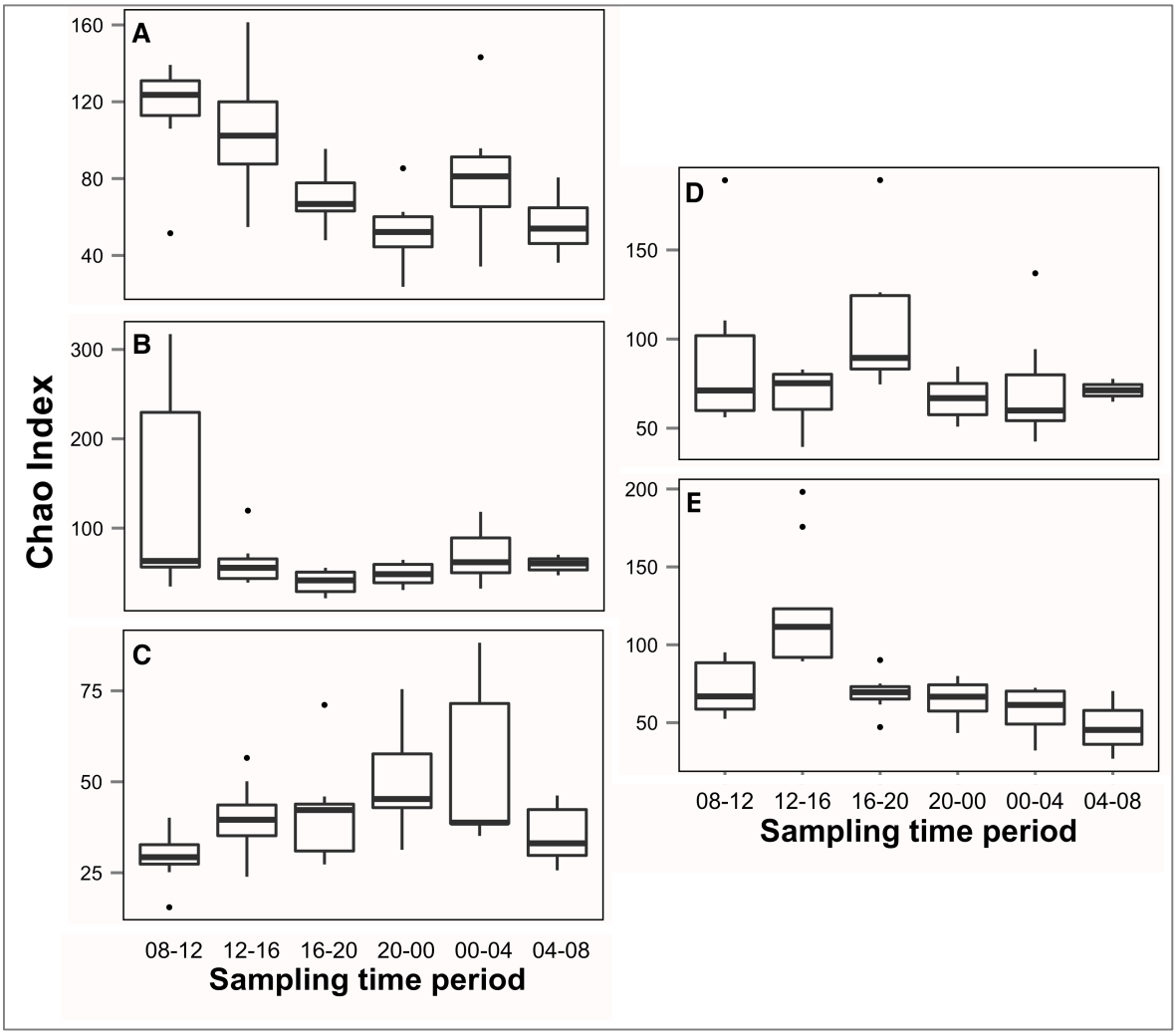

Figure 4.9. Chao index was used to estimate the richness for each four-hour sampling time-period for all sampling locations (indicated on the left of each panel). The bottom, middle and top line of the box plots correspond to the 25%, 50%, and 75% percentiles, respectively. The top and bottom whiskers correspond to the maximum and minimum values within 1.5 × IQR (interquartile range), respectively. The dots represent outliers.

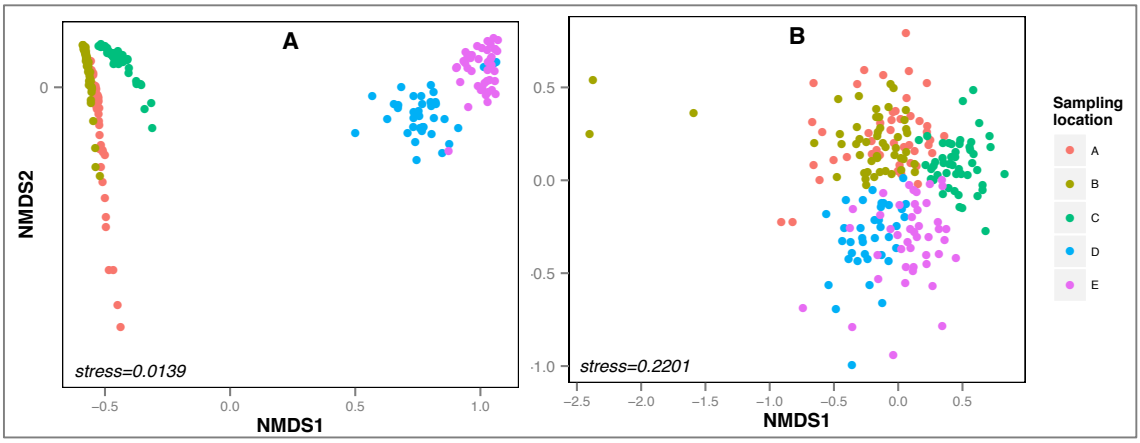

Figure 4.10. Non-metric Multidimensional Scaling (NMDS) plot of the bacterial communities of all sampling locations (A, B, C, D and E) during the 24-hr period sampled. (A) structure-based Bray Curtis distance and (B) membership-based Jaccard distances.

Bray Curtis (community structure, abundance based) and Jaccard (community membership, presence/absence based) distances were used to estimate beta-diversity for each time period at all sampling locations. Ordination plots (Figure 4.10) indicate each sampling location forms a distinct cluster, with sampling locations A, B and C clustering closer to each other compared with sampling locations D and E, and vice versa, with a stronger clustering observed for Bray Curtis metric as compared to the Jaccard metric. This clustering pattern is consistent with the differences in taxonomic classification of sequences obtained from each sampling location. Specifically, for sampling locations A, B and C, *Betapoteobacteria* were dominant while for sampling locations D and E, *Alphaproteobacteria* were dominant. We applied AMOVA (Excoffier et al. 1992) to test whether differences in membership and structure across sampling time periods were significant for each location. Further, we also utilized the beta-dispersivity analyses (Oksanen et al. 2013) to test whether the dispersion between replicate sample libraries (inclusive of filter replicates and PCR/sequencing replicates) varied over the diurnal time scale. The significance of the results for both of these analyses varied between sampling locations and depended on the type of beta-diversity metric being used (i.e. community structure or community membership). For example, a significant difference in community structure was detected for sampling locations B, C, D, and E (Table 4.3-A) between sampling time periods 08-12 and 16-20 using AMOVA analyses. However, the two time periods were significantly different at sampling locations A and B, when community membership was considered. Sampling locations C, E, and A showed the greatest diurnal change in community structure, exhibiting significant differences in six, five, and four of the 15 possible pairwise comparisons of sampling time periods, respectively. Sampling locations A, D and E exhibited the greatest diurnal change in community membership (Table 4.3-B), exhibiting significant differences in four, four, and three of the 15 possible pairwise comparisons of sampling time periods, respectively. Overall, the majority of the significant changes in beta-diversity using the AMOVA metric were observed for community structure rather than for community membership In the case of the beta-dispersivity test, significant p-values (corrected significance, $p<0.003$) were detected for only two houses (C and E) for one (Bray Curtis) and two (Jaccard) of 15 pairwise time period comparisons, indicating that the differences in variability between replicate sample libraries, inclusive of filter and PCR/sequencing replicates, around their centroid for each time period were not statistically significant.

A		Sampling time period					
		08-12	12-16	16-20	20-00	00-04	04-08
Sampling time period	08-12						
	12-16						
	16-20	B, C, D, E	E			C	
	20-00	A, C	A	A, E			
	00-04	C		E			
	04-08	A	E	C	C	C	

B		Sampling time period					
		08-12	12-16	16-20	20-00	00-04	04-08
Sampling time period	08-12						
	12-16						E
	16-20	A, B	D, E				
	20-00	A					E
	00-04	A, D, E	D, E	D			
	04-08	A, C					

Table 4.3. AMOVA (lower triangle) and beta-dispersivity (upper triangle) tests results (significant differences) (A) Bray-Curtis distance and (B) Jaccard distance. Sampling locations where a pairwise comparison was significant are indicated in the appropriate cell.

To further assess relevant time scale for changes in community structure for each sampling location, pairwise Bray Curtis distances were estimated between sample libraries, and allocated into five temporal bins that were 4, 8, 12, 16 and 20-hours apart (Figure 4.11). For example, the 4-hour time difference bin consisted of Bray Curtis distances between all samples that were four hours apart, irrespective of the actual sampling time period being considered. For each sampling location, pairwise Bray Curtis distances between each temporal bin were compared with those in the other temporal bins using permutational t-tests in R (R CoreTeam 2014). The precise results of the differences between temporal bins vary depending on sampling location (Figure 4.11). Higher number of significant differences was detected when the differences between the size of the temporal bins being compared was larger. Specifically, we detected significant differences in 10% (2/20), 47% (7/15), 70% (7/10), and 60% (3/5) of the possible comparisons between temporal bins separated by 4, 8, 12, and 16 hours, respectively, for all sampling locations combined.

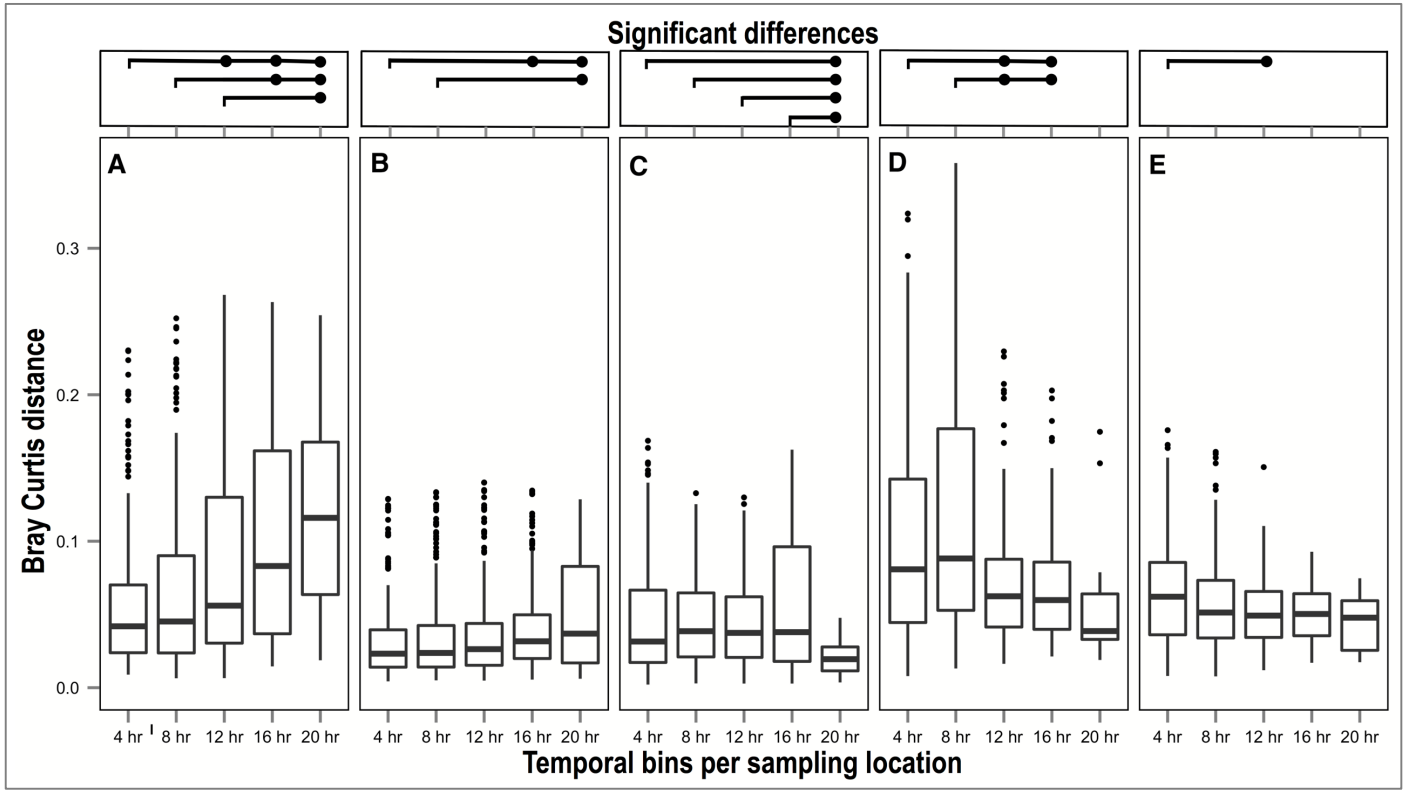

Figure 4.11. Bray Curtis distances for all sampling locations (A, B, C, D and E) binned according to the time difference between samples (4-hr, 8-hr, 12-hr, 16-hr and 20-hr differences).
The bottom, middle, and top line of the box plots correspond to the 25%, 50%, and 75% percentiles, respectively. The top and bottom whiskers correspond to the maximum and minimum values within 1.5 × IQR (interquartile range), respectively. The dots represent outliers. Significant differences ($p<0.05$) between two time-points are indicated by black lines at the top of each figure panel corresponding to each sampling location.

4.4. Discussion

Replication is critical to any investigation of microbial communities, irrespective of the environment being considered. The lack of technical replication has been identified as an important issue in microbial ecology studies (Prosser 2010), with replication being critical to move from descriptions of the microbial communities, to the study of their complexity and function (Knight et al. 2012). Particularly, I was motivated by the question of how does variability introduced due to PCR (Pinto & Raskin 2012) and sequencing (Salipante et al. 2014) compare with variability between replicate DW samples for any given sampling location at a particular time point. Through a large number of comparisons using AMOVA analyses (3 replicate filters per sample with 3 replicate sample libraries per filter for 30 samples), it is shown that the bulk DW bacterial communities collected on replicate filters are not significantly different from each other ($p>0.05$) (Appendix B, Table B4), since the filter replicates share OTUs that constitute in excess of 99% of the bacterial community abundance (Figure 4.3-A). This observation was attributed to the masking effect of PCR and sequencing steps, which also inherently sample sequences from a DNA extract and associated PCR library, respectively. This indicates that the study of bulk DW bacterial communities is much likely to benefit from PCR and sequencing replication as compared to the collection of replicate water samples.

However, it is also important to consider the limitations of the conclusion that PCR/sequencing replicates are more informative than replicate DW samples. First, I sampled a system that had high and stable water quality over the diurnal time scale for each sampling location. Specifically, a constant flow rate was maintained at the sample tap, collected samples in a DWDS with high and stable water quality, and sampled a low diversity community that was dominated by a few OTUs (~10 OTUs constituted >99% of the community). Previous studies with PCR replication on fungal (Schmidt et al. 2013) and bacterial (Zhou et al. 2011) soil samples (high diversity, high biomass environment) found little overlap between PCR replicates. Indeed, recent studies have suggested that investigations of high diversity environments benefit more from greater sequencing depth rather than PCR/sequencing replication (Smith & Peay 2014). The low diversity and low biomass characteristic of the DW environment, coupled with stable water quality, is likely responsible for the fact that samples collected using three replicate filters in our study were statistically indistinguishable from one another. If either of these aforementioned conditions were disrupted, then collecting replicate DW samples could be critical. For instance, if a DWDS was under a hydraulic disturbance scenario which was the focus of

the study, then we would highly recommend the collection of replicate samples as this will increase the likelihood of capturing clumps of detached biofilms (Choi & Morgenroth 2003) that may have highly heterogeneous bacterial communities and may be randomly captured.

Another limitation is related to the PCR process and its biases. As mentioned in Section 2.3.1, PCR is a recognized source of bias when applied to environmental communities for several reasons (e.g. inhibition of amplification by co-extracted contaminants, differential amplification, formation of artefacts, among others), and these biases can distort the original ratios of template abundances in the community under study (Wintzingerode et al. 1997; Kalle et al. 2014). Moreover, PCR was used in this study to amplify the v4 hypervariable region of the 16S rRNA gene, but it is also a step in DNA sequencing, as Illumina sequencing uses bridge amplification to amplify the template DNA in a flow cell and form clusters as an initial step before actual sequencing begins. Other factors to consider are the primers used and the hypervariable region amplified. In this study, the v4 hypervariable region was amplified with primers 515F and 806R (Caporaso et al. 2012). The length of the resulting amplicon is of approximate 250 bp which is ideal for paired-end Illumina Sequencing since both forward and reverse reads almost fully overlap, and this overlap reduces the error rate in the assembled contigs and improves OTU assignment (Kozich et al. 2013). Moreover, since it amplifies a broad range of taxa, this primer pair has been adopted by the Earth Microbiome Project (Gilbert et al. 2010) and reliable wet-lab and dry-lab protocols have been developed for this primer pair (Caporaso et al. 2012; Kozich et al. 2013). Nevertheless, it is unknown if a different hypervariable region and the use of another sequencing platform could yield different results using drinking water as sample. For instance, Albertsen et al. (2015) amplified the v1-3, v3-4 and v4 hypervariable regions using activated sludge samples and showed that each primer pair provided a different taxonomic profile, although all of them captured the dominant phyla in the samples. To overcome this limitation, several primer pairs could be tested using the sample of interest (e.g. drinking water), then the taxa relative abundances obtained with each primer pair could be compared to a second or third method that is PCR-independent (or less sensitive to PCR bias) for validation (e.g. metagenomic library, FISH), allowing the selection of the pair that most accurately captures the abundance ratios of the original community.

Regarding the impact of flow rate, the results show that different flow rates (and their associated shear stress) in premise plumbing could cause changes in community

membership. The similarities in richness within sampling location and the differences in community membership (and not community structure) observed could be because the shear stress range of the flow rates used was not high enough to cause biofilm detachment that could be detectable with the methods used. Typical shear stress values for network flushing operations are 0.2-3.0 N/m² (Douterelo et al. 2013), while lab-scale experiments have reported biofilm detachment for shear stress values of 3.1 N/m² (Choi & Morgenroth 2003), and 0.2-10.0 N/m² (Mathieu et al. 2014). In our case, the shear values estimated with typical pipe diameters used in the UK are in the range 0.0065-0.491 N/m² for the high flow condition (see Appendix B, Table B1). Moreover, mature biofilms like the ones in full-scale networks are more likely to have been subject to varying shear stress conditions over the years, making them mechanically more stable and resistant to detachment forces (Abe et al. 2012). Based on this data only, we could conclude that in the locations studied, the risk of biofilm detachment from premise plumbing as a result of common water use practices is low. Nevertheless, hydrodynamics is not the only factor influencing biofilm processes, the type of pipe material (Yu et al. 2010; Rożej et al. 2015; Proctor et al. 2016) and the disinfectant (Wang et al. 2014) are two other factors that have also been linked to increasing bacterial abundance, biofilm formation and community composition in experimental drinking water systems. Additionally, the amount of DNA template is another variable to consider, as it has been found to impact the final observations (Kennedy et al. 2014).

Regarding the impact of sample volume, our results show that a 10 L sample volume would be enough to characterize the diversity of our samples. Given the low number of significant p-values (3/45) observed for the 5 locations sampled and 3 richness estimators used, we can affirm that the richness of our samples is similar irrespective of the sample volume or flow regime. In this case, the 20 L sample is effectively equivalent to two 10 L samples aggregated, that have been drawn from the same underlying microbial community. Since there are no significant differences between the 10 L, 15 L and 20 L samples, the community is homogeneous at the 10 L scale. The deep sequencing achieved for each sample, with a minimum sequencing depth of 43797 reads, provides additional confidence in this finding. Similar results have been reported by Staley et al. (2015) who found no significant differences in richness or Shannon diversity between triplicate samples of 1, 2 and 6 L of river water. However, for sea water samples, significant differences in richness (estimated with the Chao1 index) were observed among a volume range of 0.05-5.00 L (Padilla et al. 2015). Both drinking water and river water have lower richness and diversity ($\sim 10^2$-10^5 cells/ml, 100-10,000 species) than sea water ($\sim 10^6$-10^{12} cells/ml, ~20 000

species)(McIntyre 2010), which could be why no significant difference between volumes was observed for them. An unexpected trend in richness was also observed by Padilla et al. (2015) who reported that for seawater the median richness increased with volume, reaching maximum values at 1 L, before decreasing again at the highest volumes (2–5 L). In soils, the observed richness of fungal communities didn't change as a function of the soil sample size (0.25 g, 1.00 g, 10.0 g), as reported by Song et al. (2015).

The significant correlations detected between bacterial richness and temperature, pH and residual chlorine in sampling locations F-J could be useful for operational purposes. Residual chlorine and pH (both negatively correlated with bacterial richness) are parameters that can be managed from the treatment plant and in the distribution system, therefore there is a possibility of managing them in order to mitigate the effects of temperature or substrate leaching (Bucheli-Witschel et al. 2012) (both positively correlated with bacterial richness) in premise plumbing that are more difficult to manage.

The results show that bacterial community richness in bulk DW samples changes over a diurnal time scale. For three of the five sampling locations (A, B, E), richness was significantly variable throughout the day; while for sampling locations C and D, the difference in richness throughout the day was not statistically significant. The differences in richness observed in the three sampling locations located in the same DWDS and supplied by the same DWTP (locations A, B, E) are likely due to location-specific conditions at each of the sampling point. These location-specific conditions include local water demand patterns (Lucas P. J. ; Sharma, A. K. 2010; Carragher et al. 2012) and the characteristics of the premise plumbing system of each sampling location (Wang et al. 2012; Buse et al. 2014). Nonetheless, we consistently observed a 3 to 4-fold change in the bacterial community richness at each sampling location over a diurnal time scale.

Interestingly, we observed that bacterial community richness was significantly different across sampling locations for time periods associated with increase in water demand (12-16, 16-20, and 04-08), and was similar across sampling locations during low/decreasing (i.e. stagnation time points: 20-00, 00-04), or stable water demand periods (08-12) (based on a conventional residential water demand curve by Carragher et al. 2012). During low flow periods, richness may be similar between the residences since the primary mechanism shaping the communities is regrowth during stagnation (Lipphaus et al. 2013; Sekar et al. 2012; Lautenschlager et al. 2010), and the four residences located in the same distribution system have an initial common source of microorganisms (the DWTP) and a similar final

barrier for microbial growth (the residual disinfectant). Similarly, during stable flow periods the primary mechanism shaping the bacterial community at the tap is likely to be seeding from the DWTP (Pinto et al. 2012; Lautenschlager et al. 2014). Four of the five sampling locations were supplied by the same DWTP, while the fifth location (location C) was supplied by DWTPs with highly similar treatment steps and source water conditions as the other four. In contrast, during periods with a rapid increase in water demand, the primary mechanisms altering the change in bulk water community is likely to be biofilm detachment. Biofilms are highly spatially heterogeneous and have been reported to have little to no overlap with bulk water communities (Henne et al. 2012). As a result, it is a reasonable assumption that the significant differences in richness between sampling locations during the period of rapid change in water demand may be related to seeding of the bulk DW from biofilms in the localized DWDS or premise plumbing (Schroeder et al. 2015).

The temporal differences in bacterial community structure and membership highlight the fact that diurnal changes are significant enough to be detected despite the variabilities associated with sampling, PCR amplification, and sequencing. Similarly, significant differences in bacterial community compared over larger temporal bins are more numerous than those over smaller temporal bins, further supporting the previous conclusion. It is also important to note that these changes arise not only from differences in membership (i.e. presence/absence of OTUs between samples) but also due to differences in community structure (i.e. change in relative abundance of dominant OTUs). Interestingly, changes in temporal community structure were slightly more prevalent (17 significant differences out of 75 pairwise temporal comparisons) compared with changes in community membership (13/75). It is likely that over short diurnal time scales in a stable DWDS the primary factor contributing to changes in bacterial community structure/membership are related hydraulic changes dictated by local water demand. Specifically, we find significant difference in community structure between time periods with stable flow (08-12) and those where flow is likely to change rapidly due to water usage patterns (16-20, 04-08) for four of the five locations sampled (residences B, C, D and E).

4.5. Conclusions

In this chapter we tested the impact of PCR replication, sample replication, sample volume and flow rate on the observed DW microbial communities. Our main conclusions are:

i. PCR and sequencing variabilities mask differences between bacterial communities from replicate samples. This is largely related to greater variability in detection of rare OTUs (<0.01%) between PCR/sequencing replicates as compared to sampling replicates.

ii. For the locations studied, a 10 L sample volume would be enough to characterize the diversity of the DW microbial community in the distribution system.

iii. Different flow rates, and their associated shear stress in premise plumbing, could cause changes in community membership.

iv. Bacterial community richness in the distribution system changes across diurnal time scales, with both the change and its significance being location specific. Bacterial community structure is more variable across diurnal time scales as compared to community membership. These diurnal changes are likely related to localized water use patterns and resulting hydraulic disturbances. This is indicated further by differences between sampling locations in bacterial community richness, membership and structure over time periods corresponding to rapid change in DW use patterns.

As seen before, replication is essential for reliability on DW microbial observations. Moreover, although sample volume did not significantly affect the structure and membership of the microbial communities, flow rate (and its associated shear stress) did affect community membership. Therefore, future studies should be designed with a sampling strategy that includes replication, either at the sample level, the PCR level or both. Moreover, the flow rate (and its associated shear stress) at which the samples will be taken should be selected carefully taking into account its relationship with biofilm detachment. Finally, although for the samples analyzed in the present study there was no difference between the communities captured in the volumes tested (10L, 15L, 20L), for future studies I would recommend to carry out an assessment of the representative volume of the water under study, taking into account the cell count of such water in order to accurately capture its microbial community.

5. Impact of source water and treatment processes on DW microbial communities

5.1. Introduction

The conventional drinking water treatment is based on a multiple barrier approach for the production of high quality drinking water, in order to remove and inactivate the pollutants and microorganisms present in the source water. Despite this effort, it is well documented that drinking water harbors abundant and diverse microbial communities, shaped by the different components of the system (i.e. source water, treatment barriers, distribution, building plumbing). The source water used can be of different types: surface water, groundwater, sea water, reclaimed water and a combination of the aforementioned; its water quality parameters are subject to seasonal variation and therefore so are the plant effluent's quality parameters, although in a narrower range. The source water is the primary reservoir of microbial diversity detected in drinking water distribution systems (DWDSs) (Pinto et al. 2012), indicating that it seeds the system and therefore plays an important role in shaping the microbial communities. Liu et al. (2016) also observed that DWDS samples clustered by the type of source water (groundwater, surface water) used in their respective treatment plants.

The treatment plant is a centralized component that can include varied processes; a common feature of all plants is that they act as a series of sieves successively removing the particles in the raw water according to their size (in descending size order, from large to small) until achieving the desired effluent water quality. Among all the relevant treatment processes, filtration is often found in treatment plants (with some exceptions, for instance, when high quality groundwater is the source). The filters have been shown to seed the microbial communities in the DWDS, as seen in a chloraminated system (Pinto et al. 2012) and a system without disinfectant residual applied to the distributed water (Lautenschlager et al. 2014).

The final strategy in DW treatment is to distribute the water with or without disinfectant residual (chlorine or chloramine). The presence or absence of disinfectant residual and the type of disinfectant used also shapes the microbial communities in the finished water. In both disinfected and disinfectant residual-free systems, the utility or plant has been seen as the dominant factor for both water chemistry and microbial community structure and membership (Ji et al. 2015; Roeselers et al. 2015) over other factors such as sample location and sampling time point. In systems with no residual disinfectant applied or detected (NoD), El-Chakhtoura et al. (2015) reported *Betaproteobacteria* and *Alphaproteobacteria* as dominant classes in the finished water leaving the plant and in the network, while Lautenschlager et al. (2014) found *Alphaproteobacteria* and *Gammaproteobacteria* were the most abundant classes in finished water samples.

Although a majority of studies coincide in the dominance of *Proteobacteria* in DW, differences in the relative abundances of its classes have been reported for systems that use different residual disinfectants. In chlorinated (Chl) systems, Douterelo et al. (2014) found a dominance of *Alphaproteobacteria*, *Deltaproteobacteria*, *Clostridia* and *Actinobacteria* in samples taken from the distribution system. *Acidovorax*, *Pelomonas* and *Polaromonas* (*Betaproteobacteria*) were reported by Jia et al. (2015) as dominant in the distribution system. Zhang & He (2013) reported *Alphaproteobacteria* (60.1%) and *Betaproteobacteria* (36.1%) to be dominant in tap water samples. *Betaproteobacteria* and *Alphaproteobacteria* dominated in a chloraminated (Chm) system studied by Pinto et al. (2014) in samples from the finished water and the distribution system, and their abundance was subject to seasonal variations.

Studies in DWDSs that shift between chlorination and chloramination have shown differences in community structure due to the type of disinfectant used. Hwang et al. (2012) observed that under chloramination, the samples were more similar (clustered closer as shown by non-metric multidimensional scaling-NMDS plots) than under chlorination. Under chloramination, *Methylophilaceae (Betaproteobacteria)*, *Methylococcaceae*, and *Pseudomonadaceae (*both *Gammaproteobacteria)* were the dominant families detected, whereas, under chlorination, *Cyanobacteria*, *Methylobacteriaceae (Alphaproteobacteria)*, *Sphingomonadaceae (Alphaproteobacteria)*, and *Xanthomonadaceae (Gammaproteobacteria)* were dominant. Similarly, Wang et al. (2014) noted differences in community composition as a result of the temporary conversion to chlorine in a chloraminated system; lower levels of *Alphaproteobacteria* (class) and higher levels of *Bacteroidetes*, *Firmicutes*, and *Planctomycetes* (Phylum) were

observed during chlorination as compared to chloramination. Sequences of *Methylococcaceae (Gammaproteobacteria)*, *Methylobacteriaceae (Alphaproteobacteria)*, and *Nitrosomonadaceae (Betaproteobacteria)* decreased during chlorination, while *Sphingomonadaceae (Alphaproteobacteria)* and *Chitinophagaceae* (*Bacteroidetes*) sequences increased during chlorination. Once the system changed back to chloramination, a significant increase of *Methylococcaceae* and *Nitrosomonadaceae* was observed. Both Hwang et al. (2012) and Wang et al. (2014) noted that the predominant class in their systems was *Gammaproteobacteria*, and not *Alpha-* or *Betaproteobacteria* as reported by others; apart from the switch between chlorination and chloramination, both systems also use groundwater or a mix of groundwater as source water.

A majority of studies thus far have utilized 16S rRNA gene-based approaches that provide information on the structure of the community, while providing limited to no functional assessment of the microbial community. The whole genome sequencing-based studies available have shown that the relative abundance of sub-systems related to basic cellular functions (e.g. synthesis of amino acids and proteins) is similar between raw and treated water (indicating that treatment doesn't affect them) while genes of the sub-systems related to protective functions (e.g. oxidative stress, detoxification) significantly increased after treatment (in this case chlorination) (Chao et al. 2013). Moreover, regarding antibiotic resistance genes (ARGs), Jia et al. (2015) reported that the total abundance of detectable ARGs increased after chlorination (finished water) from the filter effluent and decreased after distributions; while regarding virulence factors, Huang et al. (2014) reported that virulence proteins and pathogenicity islands increased in the finished water compared to the filter effluent, both in their types and number of reads. All the above insights on functional profiles are limited to chlorinated systems fed by surface water; the impact of other types of source water and microbial control strategies (i.e. chloramination, UV, no disinfectant) are yet to be explored in a metagenomic context.

An additional source of variation is the methodology used in the study, not only the system components and properties impact our observations, the specific protocols used (e.g. the DNA extraction method, the hypervariable region amplified, the sequencing platform used) also introduce biases (Brooks et al. 2015; Jones et al. 2015). This is especially important when making comparisons across studies, because it can be difficult to assess if the differences observed are due to biological variation or are attributed to bias introduced by methodological differences. A meta-analysis of DW microbial communities in full-scale distribution systems highlighted the impact of technical variation, showing that the variable

"origin of study" (which includes all major technical variables such as sampling, DNA extraction and PCR protocols, sequencing) explained the greatest proportion of the variance observed, above system properties such as type of source water and disinfection strategy (Bautista-de los Santos et al. 2016).

Therefore, to address some of the above mentioned current limitations, the objectives of this chapter were to (1) investigate the extent to which source water and treatment processes affect the DWDS microbial community in a metagenomic context and (2) and determine taxonomic profiles and metagenomic functions that are similar/different in DWDSs served by a range of source water and treatment processes, applying a single and robust methodological workflow informed by best-available literature. In this chapter the approach was extended to metagenomics (i.e. whole genome shotgun sequencing) to elucidate both the taxonomic profile and the functional potential of the DW microbial communities, an objective that cannot be met with 16S rRNA amplicon sequencing (applied in Chapter 4 for community characterization of locations in the distribution system), and that as seen in the introduction, has been addressed limitedly in the DW context. Furthermore, the variability of both taxonomy and encoded functions of the DW microbiome should be elucidated before being incorporated into a predictive framework and/or planning any intervention based on the DW microbiome.

5.2. Materials and methods

5.2.1. Drinking water sampling

Drinking water samples were collected from 10 DWTPs and DWDSs located in The Netherlands and Scotland (Appendix C, Table C1) in Winter 2013 and Summer 2015, respectively. The samples reflect a range of source water types (surface water, n=19; ground water, n=10; pre-treated water, n=7), treatment trains, and microbial regrowth control strategies (Chl-chlorination, n=11; Chm-chloramination, n=8; Drf-disinfectant residual-free, n=17) in drinking water systems. Prior to sampling, the faucets and sinks were thoroughly disinfected with sodium hypochlorite and the tap was flushed for 10-15 minutes in order to avoid any impact from stagnant water in the adjacent pipes (Lautenschlager et al. 2010). After flushing, the taps were adjusted to a constant flow rate (i.e. that did not produce splashing) which was maintained for the duration of the sample collection. Drinking water was collected in sterile containers (bottles, beaker), and filtered using a peristaltic pump fitted with sterile tubing and connectors, through sterile Sterivex

filters with 0.22 μm pore-sized polyethersulfone membrane (Millipore, Billerica, MA). For the Chl and Chm systems, 30 L of water from the treatment plant (treated water before distribution) and the distribution system (2 or 3 locations) were collected and filtered; while for the Drf systems, 2 L of treated water and water from the distribution system were collected and filtered as described before. Following filtration, the membranes were immediately removed from the filter casing using aseptic technique, placed in lysing matrix E tubes (MP Biomedicals, Santa Ana, CA) and stored at 4°C for a maximum of 24 hours before being transferred to a -80 °C freezer.

5.2.2. Water quality analyses

Water temperature, pH and conductivity were monitored at each sampling location using an Orion 5 Star Meter (Thermo Fisher Scientific, Waltham, MA), total chlorine using Hach's Reagent powder pillows (Hach Lange, UK) and a DR 2800 VIS Spectrophotometer (Hach Lange, UK). In each sampling location, water samples were collected in sterile Nalgene polycarbonate bottles for water quality analyses. The samples were kept at 4°C at the sampling location, and were processed and stored according to standard protocols immediately on arrival at the laboratory (APHA, 1998). Ammonia, nitrite, and nitrate concentrations were measured using colorimetric methods 4500-NH3-F, 4500-NO2–B, and 4500-NO3–B respectively; orthophosphate concentrations were measured using the Test 'N Tube kit (Hach Lange, UK). Total organic carbon (TOC) concentrations were determined using a Shimadzu TOC-LCPH Analyzer (Shimadzu, Kyoto, Japan).

5.2.3. DNA extraction, metagenomic library preparation and DNA sequencing

DNA was extracted from the filters using a combination of a phenol-chloroform method previously described (Pinto et al. 2012), and the protocol of the Maxwell® 16 LEV Blood DNA Purification Kit (Promega, Madison, WI, USA). The modified protocol consisted of the following steps: the filters with collected biomass were incubated with 300 μL lysis buffer and 30 μL proteinase K at 56°C for 20 min; addition of 500 μL chloroform:isoamyl alcohol (25:24:1, pH 8.0); bead beating (6 m/s for 40 sec) using FastPrep 24 instrument (MP Biomedicals, Santa Ana, CA, USA); the tube was centrifuged at 14,000×*g* for 10 min; the aqueous phase was transferred to a 2 mL safe lock tube (StarLab, MK, UK) and two more 40-second bead beating and centrifugation steps (at 12,500×*g* for 10 min) steps were performed after replacement of the aqueous phase with fresh lysis buffer; 500 μL

chloroform:isoamyl alcohol were added to the aqueous phase tube and the tube was centrifuged at 14,000×*g* for 5 min. The extracted DNA was automatically purified and suspended in 50 µL of elution buffer by the Maxwell® 16 instrument. The amount of extracted DNA from each sample was quantified on a Qubit 2.0 fluorometer (Life technologies, UK). All the DNA samples were stored at −80°C. In addition to these samples, a negative control for each run of the Maxwell instrument was included consisting of a filter membrane that had not been used for sample filtration. Negative controls were also processed along the DW samples in order to account for potential contamination.

Metagenomic library preparation was done using the Nextera XT library prep kit (Illumina, San Diego, CA). Briefly, the protocol included tagmentation, amplification, and clean-up of the genomic DNA, library normalization and pooling. Before normalization and pooling, the amplified DNA was quantified by Qubit 2.0 fluorometry (Life technologies, UK) and qPCR using SYBR® Green-based detection (Thermo Fisher Scientific, Waltham, MA, USA) and the fragment sizes were checked on an Agilent Technology 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) using a High Sensitivity DNA chip (Agilent Technologies, Santa Clara, CA, USA). Pooled libraries were sequenced in the Centre for Genomic Research at the University of Liverpool, UK, on the Illumina HiSeq 2500 platform in rapid run mode (2x250 bp paired end reads).

5.2.4. Sequence processing and statistical analyses

The raw Fastq files were trimmed to remove Illumina adapter sequences using Cutadapt v.1.2.1 (Martin 2011) and Trimmomatic (Bolger et al. 2014), and further trimmed using Sickle v.1.200 (Joshi & Fass 2011) with a minimum window quality score of 20 and a minimum length of 10 bp. The filtered reads were subsampled in order to retain 80% of the total, interleaved, and assembled into contigs using IDBA-UD (Peng et al. 2012). After assembly, contaminat contigs were removed. The coverage of each contig per sample was estimated by mapping the quality trimmed reads of each individual sample. Gene calling was performed on the contigs using prodigal (Hyatt et al. 2010) and protein sequences were annotated against the Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (Kanehisa et al. 2016) using diamond (Buchfink et al. 2015). For taxonomic classification, the samples were subsampled to 717,106 reads (minimum number of reads per sample) and annotation was done using kaiju (Menzel et al. 2016) with the NCBI BLAST *nr* protein database. Differential abundance of annotated genes was tested with the package Deseq2

(Love et al. 2014) in R (R CoreTeam 2014), while plots were constructed using ggplot2 (Wickham 2009) and EvolView (Zhang et al. 2012). Permutational Multivariate Analysis of Variance (PERMANOVA) was conducted using vegan (Oksanen et al. 2013) in R. In mothur, the command "get.communitytype" (an implementation of the Drichlet Multinomial Mixture Model – DMM model, proposed by Harris et al., 2014) was used to detect envirotypes in our samples. Quality trimmed reads were deposited on NCBI's Sequence read Archive (SRA) under project ID PRJNA311505.

5.3. Results

5.3.1. Taxonomic diversity

A total of 314,322,818 reads remained after quality control, with an average of 8,731,189±9,858,987 reads per sample, and minimum and maximum reads per sample values of 717,106 and 38,365,200. Combined per disinfection group, the number of reads in the Drf group (197,813,102) is greater than the number of reads in the Chl (56,989,698) and Chm (59,520,018) group, while the average number of reads per sample is also higher in the Drf group. Despite having greater number of reads and average number of reads per sample, the Drf group samples have lower percentage of classified reads in all taxonomic levels compared to the Chl samples, and in 4/6 taxonomic levels compared to the Chm samples (Figure 5.1).

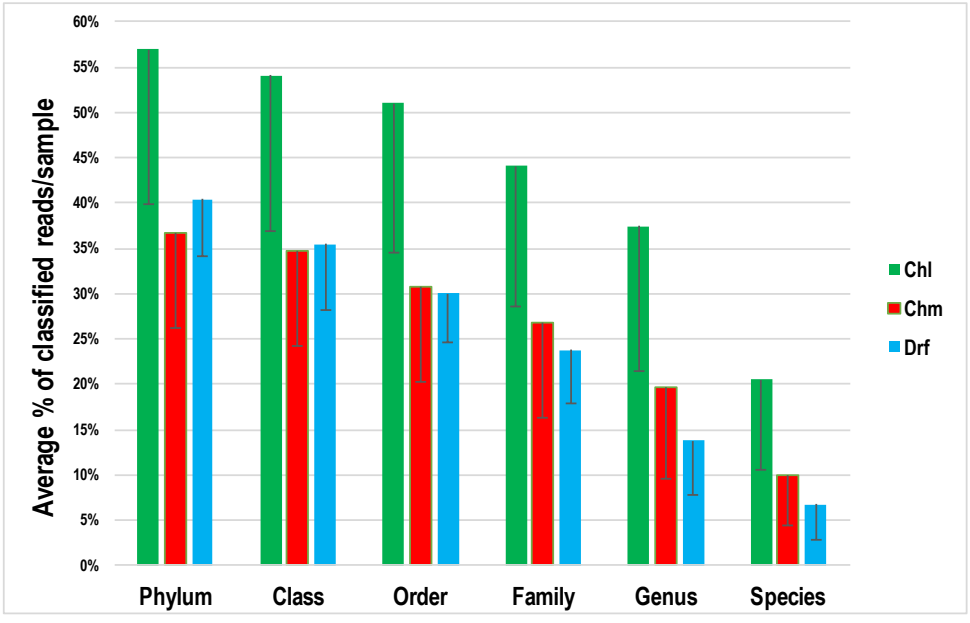

Figure 5.1. Average percentage of classified reads per sample for chlorinated (Chl, n=11), chloraminated (Chm, n=8) and disinfectant residual-free (Drf, n=17) samples. Bars indicate standard deviation.

The top 15 families (ranked by decreasing relative abundance) in each group are presented in Figure 5.2. The Drf group has higher taxonomic diversity, as it has 104 families, compared to the Chl and the Chm groups with 84 and 55 families, respectively. The top 15 families presented in Figure 5.2 belong to previously reported phyla and classes in drinking water distribution systems, such as *Alpha*-, *Beta*- and *Gammaproteobacteria*, *Bacteroidetes*, *Nitrospirae*, *Planctomycetes*, *Actinobacteria* and *Bacilli*. Moreover, the distribution of the relative abundance in each group (Chl, Chm, Drf) is different; in the Chl group 82% of the abundance is represented by the top 15 families, in the Chm group 84% of the abundance is represented by the top 15 families, while in the Drf group only 58% of the abundance is represented by the top 15 families, further supporting its greater diversity. The microbial community of all the samples was composed of 123 families, being 43 families represented in Chl, Chm and Drf samples and therefore considered "core" families. Among these core families, *Comamonadaceae* was found to be both abundant and frequently detected in the three disinfection groups, with an average relative abundance of 17.52%, 3.67% and 2.19% in Chl, Chm and Drf samples, respectively. Other notable core families included *Bradyrhizobiaceae*, *Burkholderiaceae*, *Chitinophagaceae*, *Methylophilaceae*, *Mycobacteriaceae*, *Nitrospiraceae*, *Planctomycetaceae*, *Sinobacteraceae* and *Sphingomonadaceae*.

The taxonomic profiles of the communities are shown in Figure 5.3, obtained with Bray-Curtis (abundance-based metric) and Jaccard (presence/absence-based metric) distances. In both cases, clustering per disinfection strategy is observed, more so when Jaccard distances are used to obtain the ordination plot. PERMANOVA revealed that several explanatory variables significantly explain the variance of the taxonomic community (Table 5.1-A), being the variance explained higher when Jaccard distances are used. For instance, using Jaccard distances, the disinfection strategy (chlorination, chloramination, disinfectant residual-free) explains 29% of the variance, the type of source water explains 25% of the variance, and the system explains 60% of the variance.

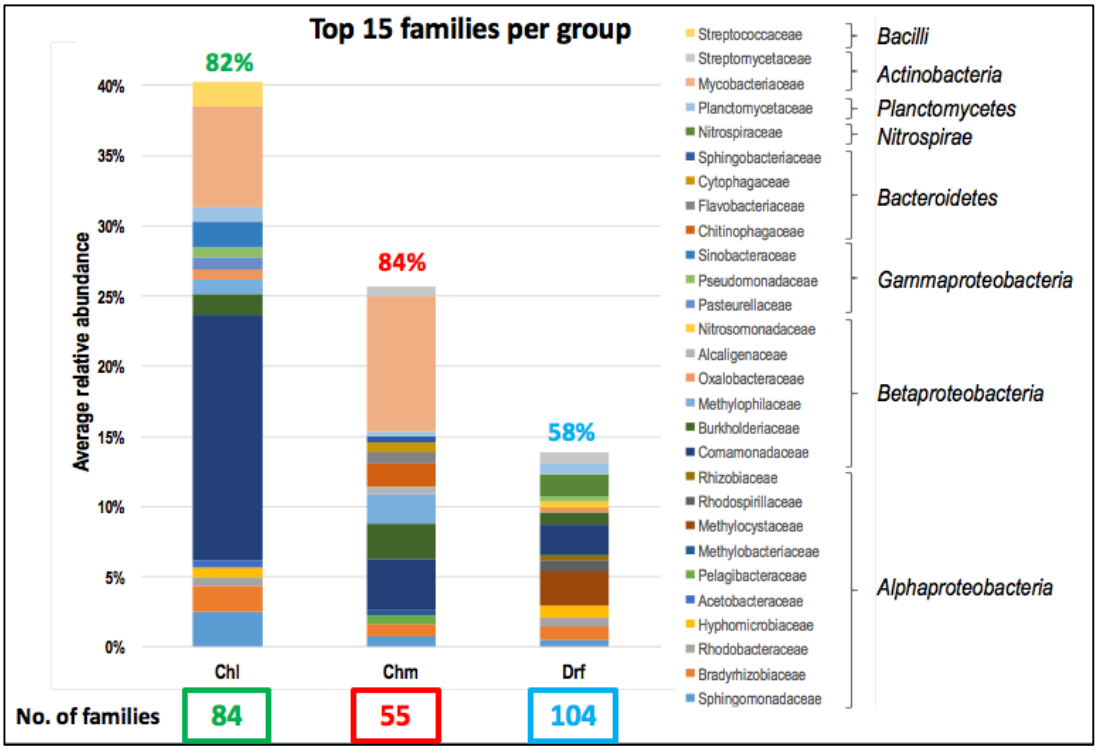

Figure 5.2. Top 15 families per group, ranked by decreasing relative abundance. Number of families in each group indicated at the bottom. Percentage of the community represented by the top 15 families in each group indicated at the top of the bars. Phylum/class of each family indicated with brackets. Chl: chlorinated; Chm: chloraminated; Drf: disinfectant residual-free.

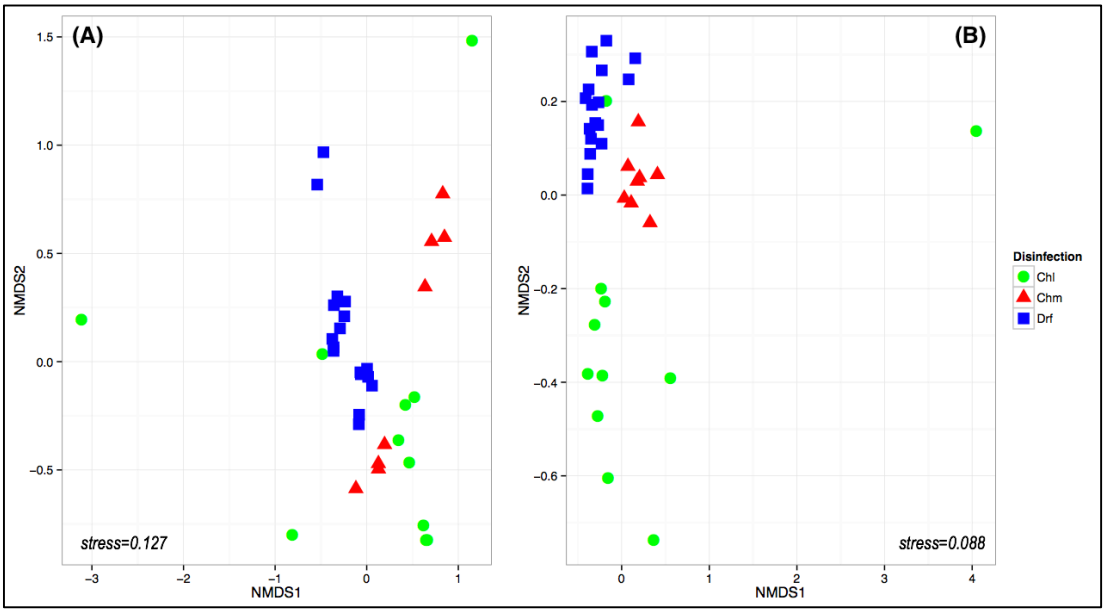

Figure 5.3. Non-metric multidimensional scaling (NMDS) plots representing the taxonomic profile of the samples, using (A) Bray Curtis distances, and (B) Jaccard distances. Disinfection groups coded by colour and shape. Chl: chlorinated; Chm: chloraminated; Drf: disinfectant residual-free.

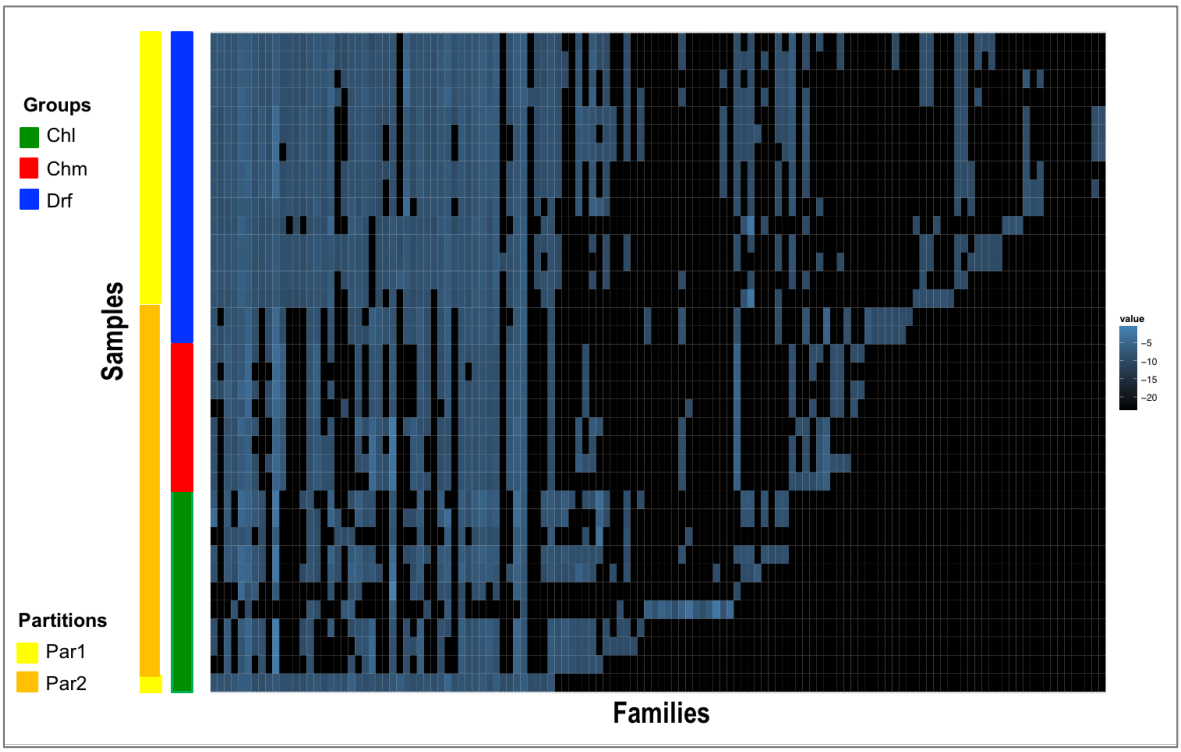

Figure 5.4. Heatmap of taxonomic families with relative abundance > 1% in each sample. Value indicated in heatmap is log2(relative abundance). Disinfection group coded by colour: chlorinated (Chl), chloraminated (Chm), disinfectant residual-free (Drf). Partitions obtained with the Drichlet Multinomial Mixtures Model are indicated by colour.

Variable	(A) Taxonomy				(B) Function			
	Bray Curtis		Jaccard		Bray Curtis		Jaccard	
	R^2	p	R^2	p	R^2	p	R^2	p
Source	25%	<0.001	29%	<0.001	23%	0.001	16%	0.092
Disinfection1	29%	<0.001	37%	<0.001	27%	0.001	37%	0.001
Disinfection2	19%	<0.001	21%	<0.001	15%	0.001	16%	0.001
Final treatment step	36%	<0.001	44%	<0.001	36%	0.001	37%	0.002
System	60%	<0.001	61%	<0.001	60%	0.001	56%	0.002

Table 5.1. PERMANOVA results for (A) Taxonomic profile, and (B) Functional profile. Source: surface water, ground water, pre-treated water1, pre-treated water-2; Disinfection1: Chl, Chm, Drf; Disinfection2: Dis (disinfected), Drf; Final treatment step: Chl, Chm, SSF (slow sand filter), UV, Decolourisation.

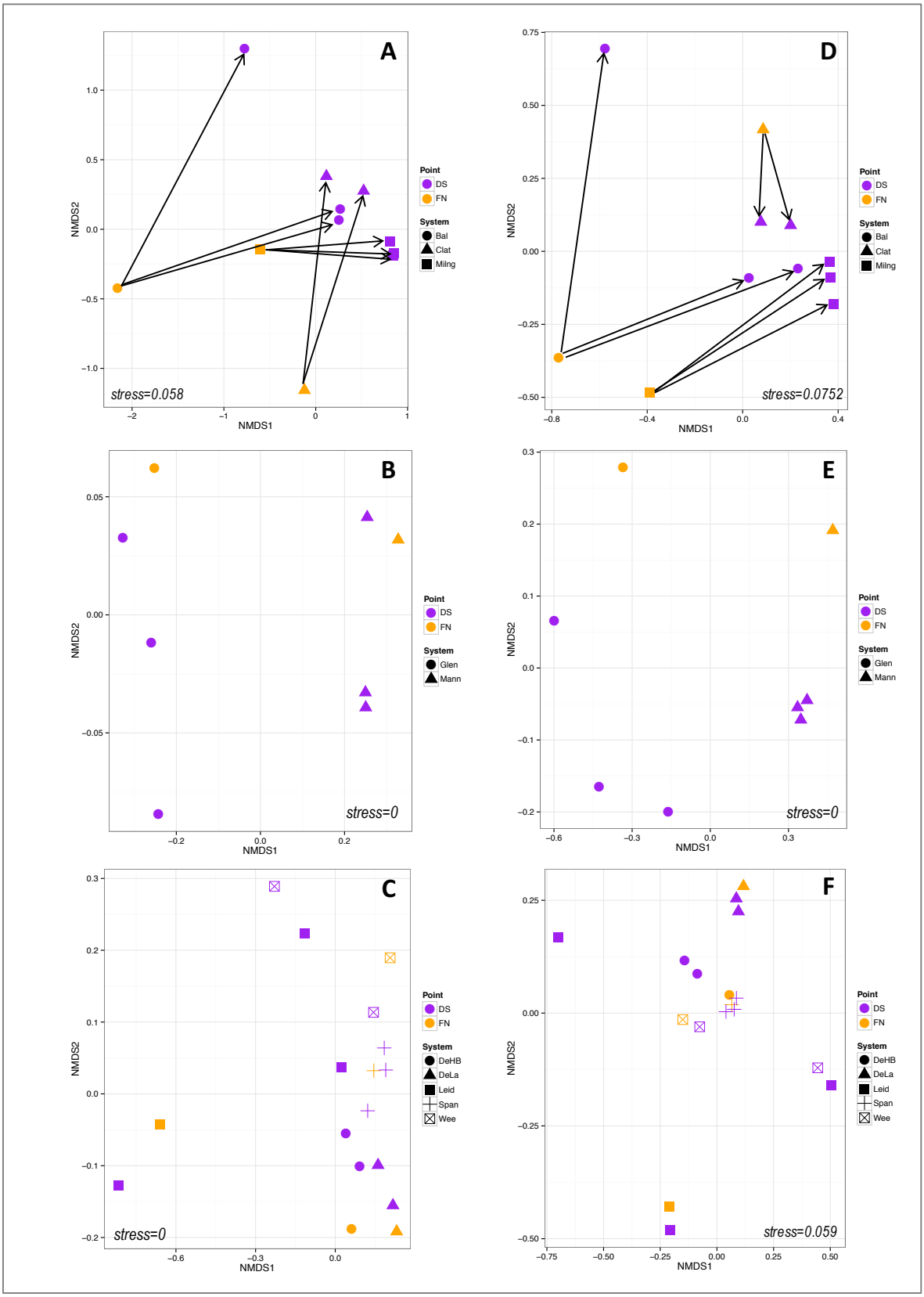

Figure 5.5. NMDS plots with Bray Curtis distances using taxa (family) abundance table for (A) chlorinated, (B) chloraminated and (C) disinfectant residual-free samples. NMDS plots with Bray Curtis distances using KEGG Ortholog (KO) abundance table for (D) chlorinated, (E) chloraminated and (F) disinfectant residual-free samples.
Colours indicate sampling locations: finished water at the plant (FN), distribution system (DS). Shapes indicate name of system sampled. Arrows join FN and DS points that belong to the same system.

Nevertheless, to untangle the effects of these variables is not straightforward, as several of them are related; for instance, all the disinfected (chlorinated and chloraminated) samples use surface water as source. The application of the DMM model (an approach independent from ecological distances, based only on abundance data) revealed the presence of two taxonomic envirotypes (Figure 5.4) that roughly correspond to Chlorinated and chloraminated samples (partition 2) and disinfectant residual-free samples (partition 1). PERMANOVA applied within each group to samples from the finished water at the treatment plant (FN) and the distribution system (DS) revealed a significant distribution effect for the Chl group with both Bray Curtis (R^2=0.29, p=0.007) and Jaccard (R^2=0.27, p=0.008) distances (Figure 5.5-A); the same distribution effect was not observed for the Chm (p>0.01) and Drf (p>0.01) group (Figures 5.5, B-C).

5.3.2. Functional diversity

The functional annotation of the contigs revealed the presence of 5828 KEGG Ortholog numbers (KOs) in the 36 samples analyzed, with an average of 5646±249 KOs/sample. The average coverage of the KOs across all samples ranged between 0.01953±0.11 (K02854) and 1228.4±1444.2 (K00540, oxidoreductase). The functional profiles of the communities obtained with Bray-Curtis (abundance-based metric) and Jaccard (presence/absence-based metric) distances are shown in Figure 5.6; in both cases, clustering per disinfection strategy is observed.

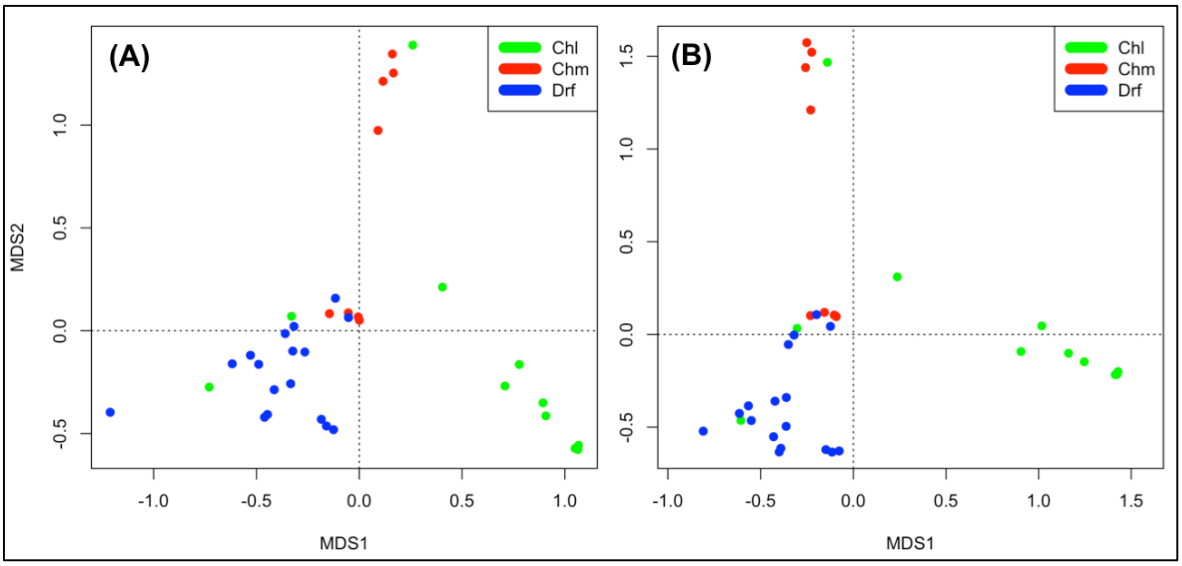

Figure 5.6. Principal Coordinates (PCoA) plots representing the functional profile of the samples, using (A) Bray Curtis distances, and (B) Jaccard distances. Disinfection groups coded by colour. Chl: chlorinated; Chm: chloraminated; Drf: disinfectant residual-free.

PERMANOVA revealed that several explanatory variables significantly explain the variance of the functional community (Table 5.1-B), with both Bray Curtis and Jaccard distances. For instance, the disinfection strategy (variable "Disinfection1"; coded as Chl, Chm and Drf) explains 27% of the variance using Bray Curtis distances ($p<0.001$), and 37% of the variance using Jaccard distances ($p<0.001$), while the disinfectant residual (variable "Disinfection2"; coded as Dis and Drf) explains 15% of the variance using Bray Curtis distances ($p<0.001$), and 16% of the variance using Jaccard distances ($p<0.001$). In the case of the functional profile, the source water type (a significant variable for both taxonomic structure and membership) is only significant for functional structure (R^2=23%, p=0.001). Within each disinfection group, PERMANOVA applied to samples from the finished water at the treatment plant (FN) and the distribution system (DS) revealed a significant distribution effect for the Chl group with Bray Curtis distances (R^2=31%, p=0.015) (Figure 5.5-D) and Jaccard distances (R^2=24%, p=0.047), the same distribution effect was not observed for the Chm and Drf groups, in which "system" was the significant explanatory variable (Figure 5.5, E-F). The application of the DMM model revealed the presence of six functional envirotypes (Appendix C, Table C2). The samples of the Chl group were assigned to two envirotypes (Partitions 4 and 6); the Chm group samples were also split into two envirotypes (Partitions 5 and 2), while the Drf group samples were assigned to five envirotypes (Partitions 1, 2, 3, 4, 6).

To further explore the differences in metabolic potential as a result of the presence or absence of disinfectant residual, the samples were divided in two groups: the Dis group, which included both Chl and Chm samples; and the Drf group, which included all the disinfectant residual-free samples, and Deseq2 (Love et al. 2014) was applied to test for differential abundance of genes. A significance level of 0.00001 and a fold change of 2.0 were used in the analysis. The test revealed the presence of 399 and 44 genes with higher abundance in Dis (Appendix C, Table C3) and Drf (Appendix C, Table C4) samples, respectively. For the Drf group, the gene with higher log2fold change (+3.44) was *pgsA* (K07282), which encodes a capsule biosynthesis protein. Several genes overrepresented in the Drf play a role in Two-component regulatory systems involved in aerobic/anaerobic respiration (*resD*), nitrogen fixation/assimilation/regulation (*nifA*), flagellar assembly (*fliJ*, *flrB*), and stationary growth in bacteria (*glrK*, *glrR*). Among the ABC transporter genes, the *macB* (macrolide transport system ATP-binding/permease protein) and *pvdE* (putative ATP-binding cassette transporter) were overabundant in Drf samples.

For the Dis group, the gene with higher log2fold change (-4.68) was *mdtH* (K08162), which encodes a multidrug resistance protein of the MFS transporter DHA1 family. Among the overabundant genes in the Dis group, there are several genes related to detoxification and protection against oxidative and chlorine stress (Figure 5.7, Appendix C, Table C5). For instance, several genes involved in glutamate synthesis (glutamate synthase), glutamate transport (*gltL*, *gltK*, *gluA*, *gluC*, *gluD*), glutathione synthesis (*gshA*, *GSR*, *gpx*, *ggt*, *OPLAH*) and the glyoxylate shunt (*aceA*, *aceB*) are overrepresented in the Dis group. Among the ABC transporter systems, the genes involved in sulfate transport (*CysP*, *CysU*, *CysW*, *CysA*), branched-chain amino acid transport system proteins (*LivK*, *LivH*, *LivM*, *LivG*, *LivF*), glutamate/aspartate transport (*GltK*, *GltL*, *GluC*, *GluD*, *GluA*), phosphate transport (*PstC*) and phospholipid transport (*MlaD*, *MlaE*) were also overrepresented in Dis samples. Moreover, seven genes involved in fatty acid degradation (*adh*, *bcd*, *acd*, *gcdH*, *alkB1_2*, *atoB*, *fadI*) were also more abundant in the Dis group.

5.4. Discussion

A greater taxonomic diversity was observed in the Drf samples, as seen by the number of families present in the samples. This is despite the fact that on average, less reads from the Dis group could be classified using kaiju. Results from previous studies that used different methodological approaches to characterize the DW microbiome suggested that Drf systems are more abundant and diverse than Dis systems. For instance, Roeselers et al. (2015) reported an average Chao1 index of 3749 and 3646 in processed water and tap water from Drf systems; while Bautista-de los Santos et al. (2016) reported values less than 300 in tap water samples from a chlorinated system; and Ling et al. (2015) reported an average value of less than 200 in samples from the distribution system. Moreover, a meta-analysis of microbial communities in DWDSs across disinfection strategies (Bautista-de los Santos et al. 2016) suggested that Drf systems are more diverse than systems that don't apply a disinfectant residual.

A core family found across disinfection strategies is *Comamonadaceae*; its presence in DW systems has been extensively reported before in disinfected (Bautista-de los Santos et al. 2016; Shaw et al. 2015; Pinto et al. 2014; Zeng et al. 2013), disinfectant residual-free (Lautenschlager et al. 2013) and desalination systems (Belila et al. 2016), being usually highly abundant and frequent in the samples analyzed. The family *Comamonadaceae* comprises over 100 species in 29 genera; the majority of its 29 genera are aerobic mesophilic rods, and have been isolated from diverse environments (e.g. water, soil, plants,

activated sludge, clinical samples) (Willems 2014). Its high abundance and detection frequency in DW samples in the finished water at the plant and the distribution system suggests that its members may be enriched/selected throughout the DW treatment process; nevertheless, it remains unclear if these bacteria are viable, viable but non-culturable or dead.

Regarding the taxonomic profile, PERMANOVA showed that more variance can be explained with Jaccard distances (that take into account the presence/absence of taxa to estimate the (dis)similarity between samples), which suggests that the presence of the members of the communities significantly shapes them. Moreover, several significant explanatory variables account for the variance of the data set, although it is difficult to separate them based on these results only. For instance, both "Disinfection1" (Chl, Chm, Drf) and "Source" (Surface water, ground water, pre-treated water1, pre-treated water2) were significant variable explaining 29% and 37% of taxonomic variation with Jaccard distances, but all of the Chl ad Chm samples have surface water as source water. The greatest variation for both taxonomy and function is explained by the variable "System", in which each sample is named according to the treatment plant that produced the water. This singularity suggests that the community detected in the finished water is a consequence of the effects of specific features in each system, starting with the organisms in the source water that first colonized the system. For instance, Roeselers et al. (2015) reported that the treatment plant was the dominant variable over sampling time point and sampling time point for four systems sampled at the treatment plant outlet (treated water) and at the point of use in June 2012, March 2013 and September 2013. Furthermore, Ji et al. (2015) also observed that utility was the overarching factor for both water chemistry and DW microbiome in a study that included samples from five utilities and their distribution systems. To further elucidate if the taxonomic community could be partitioned, I applied a DMM model which is an approach independent from any distance matrix (as opposed to PERMANOVA) and based on the abundance of the taxa. The two taxonomic envirotypes detected by the DMM model (one corresponding to the disinfected samples, the other corresponding to the disinfectant residual-free samples) further confirm that the disinfection strategy plays a key role in shaping the microbial communities in the system.

Regarding the functional profile, PERMANOVA showed that less variance was explained by the variables when it comes to metabolic potential in DW, compared to taxonomic composition. For instance, the variable "Source" is only significant with Bray Curtis distances and not Jaccard distances as in the taxonomic profile. The variables

"Disinfection1" and the "System" are both significant for the functional profile, as they were for the taxonomic profile. The application of the DMM model with the functional profile did not divide the communities according to the disinfection strategy, although interesting features emerged from it. First, more functional envirogroups were detected for Drf samples (5 envirogroups), than for the Chl and Chm samples (2 envirogroups each). Second, in the case of the Chl group, the envirogroups detected roughly correspond to sampling location points (FN, DS) and therefore are also reflecting the distribution effect seen when PERMANOVA and ecological distances were used (Appendix C, Figure C4-A). Third, in the case of the Chm group, the partitions detected correspond to the systems, not the sampling location points, therefore no significant distribution effect was detected. Lastly, in the case of the Drf samples, no relationship between the partitions and the explanatory variables is apparent.

The distribution effect observed in chlorinated samples (as seen by PERMANOVA and the DMM model) is consistent with the fact that chlorine is less stable than chloramines, decaying during distribution as it reacts with the pipe components (e.g. pipe material, biofilm, suspended and loose solids, chemicals). In Scotland, higher percentage of intact cells in samples from the distribution system were associated with low levels of free chlorine (below 0.5 mg/l) in a chlorinated system; while in chloraminated system the percentage of intact cells was low and not variable over the range of chloramine residual measured (Gillespie et al. 2014). This distribution effect was observed for both the taxonomic profile and the functional profile, suggesting that the communities in finished water and distribution system samples are composed of different members, and their encoded functions are different as well. A similar distribution effect has been reported by El-Chakhtoura et al. (2015), who reported a significant difference in bacterial community between samples taken in the treatment plant and samples taken in the distribution system (n=112).

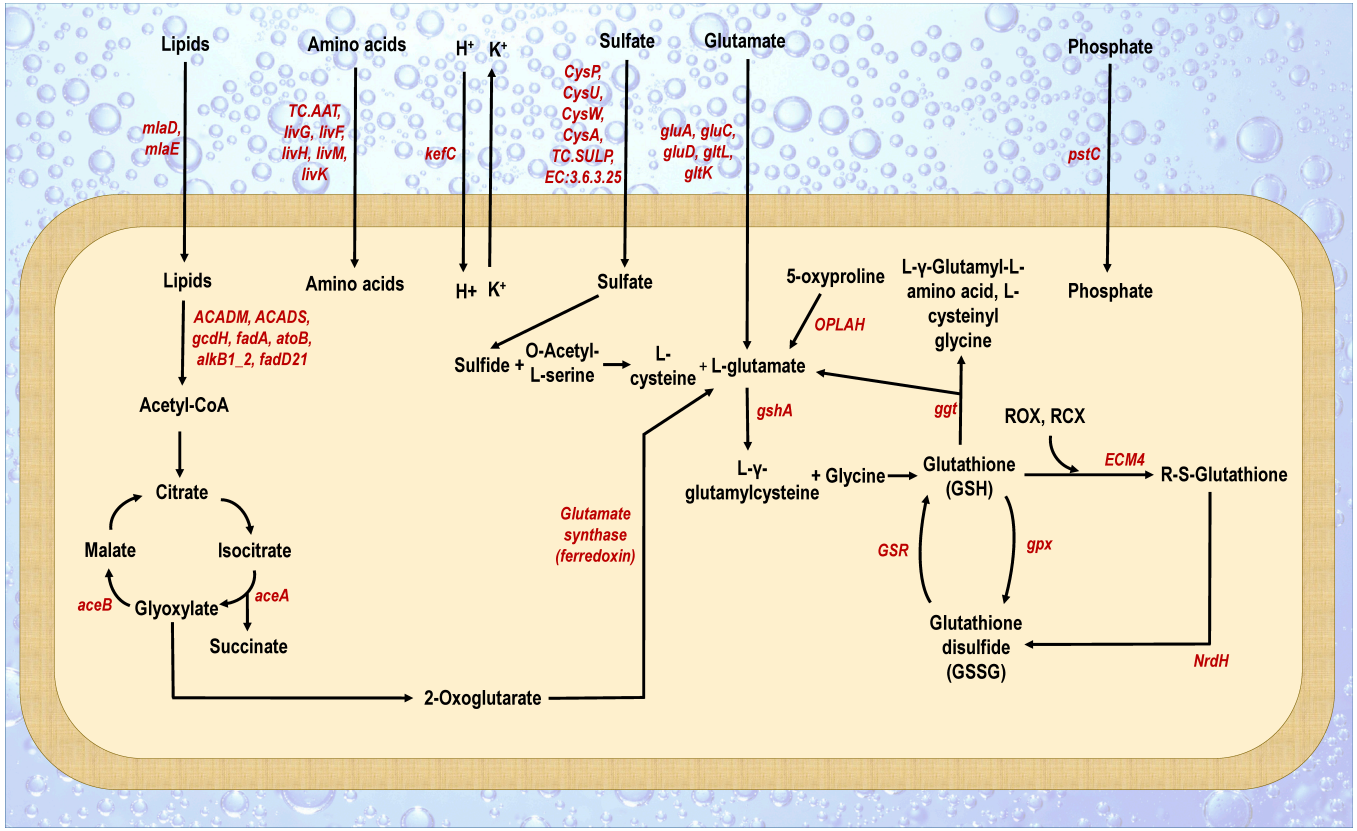

Figure 5.7. Overrepresented genes (in red italic font) in the disinfected group samples involved in protective functions against ROS/RCS stress. Details of each gene can be found in Appendix C, table C5.

Despite the wide use of chlorine as a disinfectant in drinking water due to its efficiency and low cost, the exact mechanisms of chlorine-mediated cell death are still not fully understood. When in contact with bacteria, Reactive chlorine species (RCS) react 100-fold more rapidly with sulfur-containing compounds such as cysteine, methionine or glutathione than any other cellular component does; after sulfur-containing compounds, the second most reactive targets are amines; and lastly nucleotides and lipids can be also oxidized by RCS, although3-7 orders of magnitude more slowly than amino acids (Gray et al. 2013). It has been suggested recently that the inner membrane is the site of lethal damage for cells, since the dose of HOCl that causes cell death seems to correspond to the dose necessary for loss of ATP, inhibition of F1 ATPase, loss of DNA replication (via loss of association of the origin of replication within the inner membrane) and failure of metabolite and protein transport across the inner membrane (Gray et al. 2013). An alternative mechanism of killing may be the oxidative unfolding and aggregation of essential bacterial proteins that makes the cells more sensitive to chlorine (Winter et al. 2008).

Differential abundance tests showed that several genes were significantly higher in the Dis group (Figure 5.4). The genes involved in glutamate synthesis, glutamate transport, glutathione synthesis, and the glyoxylate shunt (which are over represented in the Dis group) are all involved in the protection against oxidative and chlorine stress. Glutathione (GSH, reduced state) and glutathione disulfide (GSSG, oxidized state) play an important role in the protection of bacteria in a variety of stress conditions (osmotic shock, acidity, protection against toxins and chlorine compounds, protection against oxidative stress) (Masip et al. 2006). The production of glutathione is restricted to *Proteobacteria*, *Cyanobacteria* and some strains of gram-positive bacteria, which seems to agree with the higher abundance of *Proteobacteria* in disinfected systems.

Glutathione shifts between oxidation states (GSH/GSSG), protecting the cell by reducing harmful molecules (peroxides, chlorine, etc.). *ECM4*, a putative glutathione S-transferase that catalyzes de oxidation of GSH when it comes in contact with Reactive oxygen species (ROS)/RCS, and *NrdH,* a glutaredoxin-like protein that oxidizes glutathione conjugates (formed by glutathione+ROS/RCS) converting it to GSSG, were overrepresented in disinfected samples. Moreover, two enzymes mediate the oxidation/reduction of GSH/GSSG in the cell: *gpx* (glutathione peroxidase), and *GSR* (glutathione reductase); both *gpx* and *GSR* genes were higher in disinfected samples. Additionally, *kefC* (a glutathione-regulated potassium efflux system protein) and *gloA* (lactoylglutathione lyase

or glyoxylase I, catalyzes the glutathione + methylglyoxal reaction), were also overrepresented in disinfected samples. The production of methylglyoxal (which is a toxic by-product of glycolysis) is a strategy employed by bacteria to survive under stress conditions; methylglyoxal reacts with glutathione to produce glutathione conjugates that activate the potassium efflux systems *kefBC*; the *kefBC* systems pump K^+ out of the cell and allow the entrance of H^+ which acidifies the cytoplasm. Acidification of the cytoplasm prevents cell death through the activation of protective mechanisms against electrophiles and DNA repair mechanisms (Ferguson et al. 1995). For instance, protection of plasmid DNA against methylglyoxal damage was observed in *E. coli* MJF274 $kefB^+/C^+$ compared to *E. coli* MJF276 $kefB^-/C^-$ (using plasmid transformation efficiency as measure of DNA damage/integrity) as well as reduced genomic DNA damage (Ferguson et al. 2000).

Moreover, *OPLAH* (5-oxoprolinase, converts 5-oxoproline to L-glutamate) and *gshA* (glutamate—cysteine ligase, converts L-glutamate to L-γ-Glutamylcysteine which is a precursor of GSH production) were also overrepresented in disinfected samples; while *ggt*, (glutathione hydrolase) the only peptidase that can degrade GSH, was also overrepresented in disinfected samples. Three proteins of the glutamate transporter system *gluABCD* (*gluA*, *gluC* and *gluD*) were overrepresented in disinfected samples; these proteins import extracellular L-glutamate to the cell, which can then be used as a substrate for glutathione production with the intervention of *gshA*. Finally, sulfate transport proteins (*CysP*, *CysU*, *CysW*, *CysA*) were also overrepresented in the disinfected samples. Sulfur is an important structural component of Methionine and Cysteine, 2 common amino acids; moreover, cysteine is a precursor of glutathione synthesis. Sulfur is also a cofactor of ferredoxins like the overrepresented glutamate synthase (K00284) and *fer* (K05337).

In addition to the glutathione and glutamate associated genes overrepresented in the disinfected samples, several genes involved in transport and degradation of phospholipids were found in disinfected samples, as well as *aceA* (isocitrate lyase) and *glcB* (malate synthase), the two key enzymes in the glyoxylate shunt. Previous studies have shown that the glyoxylate shunt is upregulated when Acetyl-CoA is a direct product of fatty acid or acetate degradation. Moreover, the glyoxylate shunt (and particularly *aceA*) is required for the pathogenesis of *M. tuberculosis*, *P. aeruginosa*, *Rhodococcus equi* and fungal strains; *"in the case of R. equi, the authors hypothesised that aceA could be involved in the metabolism of host membrane lipid-derived fatty acids, which are a potential carbon source in macrophages*" (Ahn et al. 2016). In drinking water distribution systems, the source of lipids could be bacterial membrane lipids from cells that have been lysed by the

disinfectant. Additionally, glyoxylate can generate glycine via alanine-glyoxylate transaminase or serine-glyoxylate transaminase, which could lead to the production of glutathione; a link between the glyoxylate shunt and glutathione synthesis was observed experimentally in *P. aeruginosa,* as seen by the increased levels of glutathione, and the upregulated glutathione synthesis and transaminase gene (catalyzes glyoxylate-glycine reaction) (Ahn et al. 2016).

A stress response associated with glutathione has also been reported for other environments different to drinking water that also have stress factors. For instance, *Rhizobium tropici* is a gram-negative rhizobial *Alphaproteobacteria* that fixes nitrogen and shows high tolerance to acidic conditions; experimental work carried out by (Riccillo et al. 2000) showed that for growth of *Rhizobium tropici* at acidic pH levels (ph=5.0), glutathione was essential; glutathione was also important for osmotic stress tolerance. In marine microbial communities, Bengtsson-Palme et al. (2014) reported that protein families involved in the oxidative stress response (including glutathione peroxidases) were found in marine metagenomes with an abundance higher than the average, unlike other common detoxification systems which were underrepresented. Finally, in the food industry Lactic acid bacteria (LAB) play a key role in the production of diverse fermented products such as milk, cheese, sausages, fermented vegetables, etc. and are the most abundant organisms among probiotic bacteria. It has been shown that glutathione plays a key role in the survival of LAB under several stress conditions encountered (e.g. oxidative stress in the gastrointestinal track and during industrial processes; acid stress in the gastrointestinal track and in the media as a result of their own growth; cold stress during freezing or low temperature storage; osmotic stress during industrial processing) (Pophaly et al. 2012).

The overrepresented *mdtH* gene (also called *yceL*, DHA1 family of the Major Facilitator Superfamily-MFS) in disinfected samples is a multidrug efflux pump that confers bacterial resistance to norfloxacin and enoxacin (Nishino & Yamaguchi 2001); both are wide spectrum antibiotic (first generation fluoroquinolones) that kill bacteria by interfering with DNA replication, and are prescribed to treat urinary tract infections and sexually transmitted diseases. Other proteins of the same family (DHA1) are also found in yeast; however, the yeast MFS transporters characterized so far *"are thought to transport a wide variety of structurally and pharmacologically unrelated drugs and other xenobiotic compounds from the cytosol to the outer medium"*, and *"a number of these transporters might even have a specific physiological substrate"*, therefore *"their ability to export multiple drugs may occur fortuitously and be selected opportunistically"* (Dias et al. 2010).

Differential abundance tests showed that in the Drf group there are less over represented genes than in the Dis group. The overrepresented Two-component system genes involved in aerobic/anaerobic respiration could suggest that the cells possess some versatility to adapt their metabolism to their environment and regulate their growth. The overabundant *fliJ* is one of the components of the ATPase complex, a cytoplasmic component of flagella that shuttles substrates to the membrane secretion system and presents them as unfolded substrates for efficient secretion and flagellar assembly (most of the components of a bacterial flagellum are external and must be exported and assembled) (Chevance & Hughes 2008). *flrB* is part of the *flrBC* two-component system, its role is to phosphorylate *flrC* to activate flagellar gene transcription (Moisi et al. 2009). Flagellar assembly consumes a large amount of energy from bacteria (Terashima et al. 2008), although the energy consumed by the flagellar motor is usually negligible compared to its total energy cost during growth (Li & Tang 2006). In an environment without disinfectant, bacteria could be less reliant on their defense mechanisms to survive, and therefore could allocate some energy in flagellar assembly as a strategy to move to higher nutrient areas.

5.5. Conclusions

The impact of source water and treatment processes on the taxonomic and functional profiles of drinking water from 10 DWSs was assessed using whole-genome shotgun sequencing, thorough the application of a single methodological workflow. Drf systems were more diverse than Chl and Chm systems as seen by the number of taxa levels found. Several explanatory variables significantly explain the variance of the taxonomic community (e.g. disinfection, source, system), being the variance explained higher when Jaccard distances (presence/absence-based distances) are used. A DMM model further confirmed the impact of disinfectant residual, as it separated the samples in two partitions corresponding to Drf samples and Chl+Chm samples.

Regarding the functional profile, less variance is significantly explained by the explanatory variables tested; and the application of the DMM model revealed more complexity as seen by the six partitions obtained. In the Dis group, differential abundance tests revealed the overabundance of several genes involved in cell detoxification and protection against ROS/RCS, such as the glutathione, glutamate, fatty acid degradation and glyoxylate pathways. In the Drf group, the overrepresented genes suggest that the microorganisms inhabiting Drf systems may be adaptable to aerobic/anaerobic conditions.

Finally, a significant distribution effect was seen by ecological distance-based testing (in the case of taxonomic profile), and by both ecological distance-based testing and multinomial mixture model application (in the case of the functional profile), in the Chl samples; the same effect was not observed in Chm or Drf samples. This is consistent with the chlorine decay phenomenon in distribution systems that has been previously reported; moreover, it suggests that for Chl systems, the spatio-temporal dynamics involved in distribution can change the functional and taxonomic profiles of the drinking water conveyed from the production point (treatment plant) to the consumer's tap.

6. Conclusions and future work

6.1. Conclusions

In this research project, I have characterized the microbial ecology of full scale drinking water systems using DNA sequencing-based approaches, with the aim of exploring how the obtained insights could be applied to develop a predictive/proactive microbial management approach. To achieve this aim, I have sampled multiple full-scale drinking water systems with different configurations (e.g. different types of source water, process configuration, microbial regrowth control strategy, etc.) in different points (e.g. treated water at the plant, distribution system, point of use), and used a combination of bioinformatics, molecular biology, microbial ecology and multivariate statistical analyses to link the data generated to the properties of the systems under study. Moreover, not only system properties shape microbial communities, the methodology applied also has an impact on our observations (and consequently on our interpretations and conclusions). Therefore, given the importance of appropriate methodology in order to obtain an accurate depiction of a system, and considering the limited work available on methodological biases in DNA sequencing-based drinking water studies, I investigated the impact of technical (i.e. sample) replication, PCR replication, sample volume and flowrate on the obtained observations of the DW microbiome.

The first and general conclusion of this work is the usefulness of DNA sequencing-based approaches in characterizing the DW microbiome, as a greater amount of information about DW systems is obtained (compared to culture-based techniques) and can be effectively linked to system components/operation. In my case, I was able to find meaningful relationships between microbial community structure and membership and important system properties (such as diurnal water demand and disinfectant residual) but the approach I have adopted is flexible and therefore could be applied to tackle other issues present in DW systems. For instance, we could expand our knowledge of the known (e.g. known waterborne pathogens, non-pathogenic/opportunistic, involved in aesthetic and operational problems) and unknown organism in DW systems (i.e. the fraction that has not been cultured, and assess the risk that it poses), delve into mechanisms of known treatment processes (e.g. filtration, chlorination, chloramination, UV, ozone) through whole-genome analysis, or assess microbial hazards at the point of use.

Moreover, there is also the possibility of improving treatment by applying microbial-mediated processes for the removal of pollutants (i.e. biofiltration), which constitutes a great opportunity since they are a more sustainable approach than physico-chemical processes.

Finally, the trend of rapid advancement of DNA sequencing and its related fields (molecular biology, bioinformatics, etc.) suggests that even more data/better quality data will be available in the upcoming years. As an example, when I started this project 4 years ago, shotgun sequencing was not on the plan, but two years into the project it became part of it as robust bioinformatics tools and pipelines became available to analyze the sequencing data; the benefits (i.e. information we would get of the systems under study) that could be obtained justified the considerable expenditure on shotgun sequencing, which is more expensive than the initially planned amplicon sequencing. As I finish my project, a colleague in my research group is currently working on real-time DNA sequencing of microorganisms in drinking water with moderate success; this topic was unexplored when I started 4 years ago.

Specifically, my work was divided in separate components that yielded the following results:

a. A literature review was done in the form of a meta-analysis, using publicly available 16S rRNA gene data of full-scale drinking water distribution systems. This collective analysis showed that *Proteobacteria* dominate across systems and disinfection strategies, showing high relative abundance and detection frequency, and that disinfectant residual-free systems (Drf) are more diverse than chlorinated (Chl) and chloraminanted (Chm) systems. Several relevant OTUs were also detected across disinfection strategies; for instance, potential opportunistic pathogens (*Legionella*, *Pseudomonas*, *Mycobacterium*) were differentially abundant across disinfection strategies, predatory bacteria (e.g. *Bdellovibrio*, *Lysobacter*) showed differential occupancy across disinfection strategies, and nitrifiers (e.g. ammonia oxidizing bacteria, nitrite oxidizing bacteria, comammox bacteria, etc.) were differentially abundant within disinfection strategy. Nevertheless, system variability emerged as main explanatory variable for the taxonomic community (a common feature of all meta-analysis) over other variables. The meta-analysis revealed the presence of potential contaminants in DW samples, which highlights the importance of minimizing contamination and

including negative controls, even more so in the case of DW which is a low biomass environment (compared to soil, marine ecosystems, gut, etc.). Finally, it highlighted the importance of data sharing (i.e. sharing of DNA sequences in digital format, metadata of samples, including water quality parameters, dates of sampling, system properties, among others) and standardized reporting practices as means to maximize resources and increase our understanding of DW systems. This is especially important in a rapidly changing landscape of resources (e.g. lab reagents, equipment, sequencing platforms, etc.) that is predicted to sustain that trend in the following years. To my knowledge, this was the first meta-analysis of DW microbial communities done, and the adoption of this approach instead of a traditional literature review revealed both challenges and opportunities in the DW field.

b. The impacts of several methodological variables on the observations of the DW microbiome were estimated; namely the impacts of sample replication, PCR replication, sample volume and flowrate. In the case of DW, triplicate samples showed no significant differences in community structure and membership, as they shared core OTUs that represented > 99% of the overall relative abundance. These results suggest that DW studies would benefit more from PCR replication than sample replication; nevertheless, the relatively stable conditions under which the samples were taken and the low diversity of the DW samples in this study could be the causes of the similarity between filter replicates; if any of those conditions were disrupted, sample replication could be crucial to capture variability. Sample volume showed limited impact on diversity (and therefore time could be saved in sampling campaigns with a minimum optimum volume), while sampling flow rate showed significant impacts on community membership/structure in 3/5 locations sampled. Moreover, significant correlations between opportunistic pathogens (OPs) and other non-OP OTUs with high relative abundance/detection frequency were detected in high and low flow conditions; for instance, *Sphingomonas* and *Legionella* had significant correlations in the high flow condition, while *Sphingomonas, Mycobacterium, Legionella* and *Pseudomonas* supported both positive and negative correlations in the low flow condition. The aforementioned results constitute the first efforts to quantify methodological biases in DW samples, and are useful for the design of future sampling campaigns.

c. An investigation of spatial and temporal dynamics in DW distribution systems over small scales revealed significant changes in bacterial richness, and community structure and membership. Specifically, significant differences in richness throughout the day (i.e. 24 hours) were found in 3/5 locations; across locations, significant differences were found for time periods associated with high/changing water demand (i.e. 12-16Hr, 16-20Hr, 04-08Hr) while no significant differences were found during time periods of stable/low water demand. Regarding community structure and membership, significant differences were found throughout the day in all the sampling locations. The significant differences in richness during times of high/varying water demand in the distribution system suggest that the hydraulics of the system is the key driver, as during high/varying water demand the increasing velocities and associated shear stress could cause biofilm detachment in the mains. Regarding spatial dynamics, clustering (similarity) of samples by sampling location was observed mainly with Bray Curtis distances (abundance-based metric) instead of with Jaccard distances (presence/absence-based metric), which suggests that the OTUs are consistently detected across the sampling locations but at different relative abundance levels. Both small spatial and temporal scales in DW systems had been explored limitedly; this study addressed both using high-throughput sequencing and multi-level replication and was able to find meaningful relationships that have an impact on the way sampling is conducted.

d. The impact of source water type and treatment processes on the taxonomic composition and encoded functions of DW microbial communities was assessed, by surveying 10 systems covering a range of sources, treatment processes and disinfection strategies (chlorination-Chl, chloramination-Chm, disinfectant residual free-Drf). Drf samples were more diverse than Chl and Chm samples. Several variables (e.g. disinfection, type of source, system) significantly explained the taxonomic variance of the communities using both abundance-based (Bray Curtis) and presence/absence-based (Jaccard) ecological distances, indicating that both community structure and membership were affected by the variables. Further, two taxonomic envirotypes were detected, corresponding to disinfected (Chl and Chm) and disinfectant residual-free samples. Regarding the functional profile, less variance was significantly explained by the variables, compared to the taxonomic profile; moreover, several (6) functional envirotypes were detected, with more functional diversity being observed among the Drf samples. A significant distribution effect (i.e. significant differences between samples of treated water at

the plant and samples from the distribution system) was observed for both taxonomic and functional annotations in the case of the Chl samples, while the same was not observed for the Chm and Drf samples. Finally, several genes related to microbial protection against chlorine/oxygen species were overabundant in Chl+Chm samples. To my knowledge, this is the first time that a comparison of multiple points across different systems is done, elucidating the variables that shape both community taxonomy and function.

As seen by the summarized results, several variables (both methodological and system properties) can impact our observations of the DW microbiome. Therefore, they must be taken into account when designing a sampling scheme, to ensure that we capture the desired effect and minimize the others, and/or when interpreting the data in order to draw accurate conclusions from meaningful relationships. In addition to careful consideration of variables, and since "System" has been found to be an important variable (in both the meta-analysis and the across system taxa/functional comparison, point "d"), at this stage the DW field would greatly benefit from long term studies in single systems that include multiple measurements (e.g. biomass for DNA/RNA sequencing, water quality parameters, flow cytometry, ATP measurement, etc.) that can be integrated and analyzed together to provide a more comprehensive depiction of the system.

Despite chlorine-based compounds being cheap, effective and widely used for disinfection, the mechanisms of chlorine disinfection are still poorly understood. In the case of chlorination, a problem emerges with its application in the form of disinfection byproducts that are regulated due to their adverse impacts on health. Moreover, culture-based techniques for the detection of bacteria use media with a carbon content that does not correspond with the nutrient concentration in DW. As future work, an experimental approach could be devised and applied to test the microbial response to relevant conditions in DW systems such as the presence of chlorine species and the low nutrient environment (e.g. oligotrophy). Such approach could span through several 'omics techniques (e.g. metagenomics, metaproteomics, metabolomics) to elucidate mechanisms by finding out what are the main cellular targets of the disinfectant and how do the cells protect themselves against it.

A topic still unaddressed in DW microbial ecology using DNA sequencing-based methods is the viability of the detected organisms. Viability assays compatible with DNA sequencing have not been extensively applied in the DW field. The limited examples

available have not used next-generation sequencing, and therefore the analyses and conclusions that can be drawn from them are limited (Henne et al., 2012; Eichler et al., 2006). From a public health perspective, the entire microbial community (not only the viable fraction) may be of interest. For example, in immunology, it is known that cells of the innate immune system in vertebrates respond to components of bacterial cells such as DNA, lipopolysaccharides, peptidoglycans, lipoteichoic acid, etc. All these components are called Pathogen-associated molecular patterns (PAMPs) or microbe-associated molecular patterns (MAMPs). The recognition of these molecules could contribute to shaping the host immune responses against bacterial infections. Therefore, both fractions (viable and inactivated) could be of importance in drinking water studies from a public health perspective (Hemmi et al., 2000; Dalpke et al., 2006; Häcker et al., 2002). Nonetheless, given the amount of work done and knowledge accumulated over the last 10+ years targeting the total microbial community in DW (dead+viable), the rapid advancements in nucleic acids sequencing and sequence analysis (bioinformatics), subsequent efforts to study viability in DW microbial communities should build upon all this knowledge and try to answer relevant questions still unaddressed, instead of substituting one technique for the other. For instance, DNA-based approaches targeting the entire community have been useful in capturing variability and describing dynamics (e.g. spatial, temporal), showing the impact of treatment processes and detecting differences in community composition and structure between samples. A straightforward application of a viability approach would provide different information regarding our samples, it will still be subject to its limitations and biases, and the overall conclusions will likely be the same (e.g. that biofilm is different from bulk water, Henne et al., 2012; that raw water is different from disinfected water, Eichler et al., 2006).

In addition to viability, another key aspect for a microbe to pose a threat to the consumer is its ability to reach the tap and be ingested/inhaled/in contact with the user. As part of the methodological variation assessment, I explored the impact of flow rate on DW microbiome, and while looking for literature on the topic I became aware that the majority of work available on hydraulics+microbial ecology is related to biofilm dynamics (e.g. detachment, interaction with bulk phase, etc.); this is understandable considering the links of biofilms with health in DW. Nonetheless, a topic that has started to be explored just recently is the impact of flow conditions on microbes in the bulk phase. Microorganisms in fluid flow are subject to shear stress (as a consequence of the gradients of flow velocity) that combined with microbial phenotypes (e.g. morphology, motility, chemical sensing) generate a spectrum of dynamics (e.g. flow-induced rotations that affect the

microorganisms' direction of motion and their ability to disperse) with consequences on microbial ecology (Rusconi & Stocker 2015). Moreover, I also found a significant diurnal variation in the DW microbiome, most likely driven by water demand (i.e. the hydraulics of the system). Therefore, the intersection between hydraulics and molecular ecology seems like a pertinent frontier to explore in the context of DW systems.

Finally, culture-based methods for microbial detection and enumeration have two limitations: (i) they take considerable time between sample collection and generation of results, and (ii) biases introduced by cultivation have been extensively reported; therefore, the prospect of real-time detection is attractive for different purposes. First, in the case of pathogens, real-time detection and enumeration would be ideal to prevent outbreaks. Second, for operational purposes, it would be useful to have a baseline of the entire microbial community of a system under normal conditions, and based on this baseline detect deviations that could indicate that something has changed; further investigations could clarify the cause of the deviation and lead to corrective actions if necessary/pertinent. Some approaches working towards reducing the time gap between sample collection and results generation include protocols that aim to perform on-site biomass collection, sample preparation and detection/quantification. For biomass collection, the deployment of filters that continuously collect biomass is an option, they could either be changed manually or automatically; in the case of sample preparation, portable kits with minimum components for DNA extraction have been used in third-world countries. Some options available for detection include portable qPCR instruments, the MinIon sequencer, and microfluidics devices. Successful examples of on-site detection of target microorganisms have been reported in other fields (e.g. detection of plant pathogens with qPCR, detecting a Salmonella outbreak with the MinIon), while some attempts have been done in DW with partial success.

6.2. Future work: towards a predictive framework for microbial management in drinking water systems

The drinking water treatment process applied nowadays has not changed much since its development in the early 20th century. Regarding the microbial aspect, the safeguard of public health in DW has evolved from the first proposed WHO guidelines in 1958, to the current framework which relies on risk analysis and the "due diligence" principle. Nonetheless, although the approach has been refined and improved, the instruments used to assess its efficacy have remained almost unchanged; these instruments are culture-based

tests and the use of indicator organism to assess potability and process efficiency. On the other side we face new challenges in DW microbiology, in the form of emerging pathogens, climate change, antibiotic resistance, among others. Simultaneously, the application of DNA-sequencing techniques has expanded and improved in terms of the amount and quality of data obtained from the systems. Therefore, it seems that while both technology and challenges have increased, the response/adaptation of the DW sector to these changes has not kept up with it. Therefore; this final section will be guided by our findings and the current knowledge regarding DW microbial ecology, to explore the intersection between DNA sequencing-based approaches and current practice in the DW field, with an emphasis on (i) the identification of components in the current framework that could benefit from the application of DNA sequencing-based techniques, and (ii) the addition of predictive capabilities to the current framework.

As seen in Chapter 4, replication is key for reliability of the generated results, and has been largely ignored in DW microbiome studies. Our key findings regarding sample replication (i.e. no significant differences between replicates) and PCR replication (i.e. high variability among replicates) should guide any kind of future work to incorporate the appropriate level or replication into its design. For samples with low cell count and taken under stable hydraulic conditions (like the ones analyzed in this study), PCR replication may be preferred since the filters captured similar communities, but under unstable hydraulic conditions sample replication may be needed. In the case of disinfectant residual-free systems which have higher cell count, richness and diversity, both sample replication and PCR replication may be necessary.

As seen in Chapter 5, several variables significantly explain the structure and membership of DW microbial communities, and among these the variable “system” explains most of the variance (60%) for both taxonomic and functional profiles. This singularity suggests that the community detected in the finished water is a consequence of the effects of specific features in each system, starting with the organisms in the source water that first colonized the system. The first thing to consider is what kind of future approach is suitable? A general survey focused on across-system comparisons, or an assessment focused on intra-system variation? Tailored approaches focused on intra-system variation seem like the suitable choice if the aim is to design tools for decision making/management. To this end, data can be collected to establish a baseline corresponding to “normal” conditions, and this baseline can then be used to assess deviations from it that warrant further investigations.

To establish a baseline, an appropriate time scale should be selected, considering that the DW microbiome is subject to seasonal, monthly, weekly and diurnal variations. Based on scientific knowledge only, daily sampling seems the best approach since it would yield the greatest amount of data, but other factors (e.g. resources) also come to play in taking such decisions. In any case, the sampling campaign should extend over a year to be able to capture seasonal variation. Also, sampling points should be selected depending on the objective of the approach. For instance, a baseline of the treated water at the plant would allow to assess deviations that could indicate a change in source water properties or a change in the efficiency of treatment processes; sampling in the distribution system could help assess if a contamination event has occurred or if microbial regrowth has occurred, the assessment of distribution effects is especially important in chlorinated systems as shown by my results in Chapter 5; sampling in the point of use could reveal potential impacts of building plumbing on the DW microbiome. Furthermore, a suite of physical and chemical parameters should also be measured to assess their links to DW microbial communities, and the application of a second technique such as flow cytometry (which is PCR independent) could be a valuable addition to estimate cell counts, considering the challenges in DNA sequencing-based approaches regarding their reported abundances.

Moreover, as seen in Chapters 3 and 5, disinfection is an important variable in shaping microbial communities, and Disinfectant residual-free systems have higher microbial diversity and abundance than chlorinated and chloraminated systems. As seen in Chapter 3, the majority of the data from Disinfectant residual-free systems has been obtained with 454 pyrosequencing which provides lower sequencing depth compared to Illumina Sequencing. Therefore, both replication and deeper sequencing is recommended for Disinfectant residual-free systems to better capture their diversity. Moreover, the functional potential of Disinfectant residual-free systems has been explored limitedly, to our knowledge this is the only study that has elucidated the topic. The aforementioned recommendations refer to the assessment of microbial dynamics in DWDSs from the operations/management perspective, and the definition of "normal" versus "incident" conditions that could lead to further investigations regarding the potability of the drinking water.

As seen in Chapter 2, Microbial Risk Assessment (MRA) is the tool applied in DW to assess hazardous microbial agents that cause adverse effects. Some opportunities provided by high-throughput sequencing to improve MRA are presented below:

(i) Hazard identification and characterization: high-throughput sequencing can contribute to an improved detection and characterization of recognized emerging pathogens. Virulence is usually different among different strains of a pathogen, and comparative studies on genome content among strains can elucidate the elements associated with virulence that are shared or unique to each strain. For instance, a comparison of five *L. pneumophila* strains detected in four countries (France, USA, England and Spain) and associated with outbreaks of Legionnaire's disease showed that the strains have differences in their dispensable genome (mainly genomic islands probably acquired by horizontal gene transfer) connected to virulence and DNA transfer activities that could confer advantages over others; moreover, the comparison showed that the Alcoy strain (the cause of recurrent and sometimes mortal outbreaks in Spain) has additional features that could explain its persistence and recurrence (D'Auria et al. 2010). The use of model eukaryotes to probe different virulent strains and conditions has been recommended (Brul et al. 2012). Moreover, metatranscriptomics can be pursued along with metagenomics to elucidate the presence of a phenotype and its fitness. For instance, this approach has been applied to two strains of *P. aeruginosa*, one antibiotic sensitive (PA30) and one multi-resistant (PA49), both exposed to waste water and tap water, showing that both strains had similar transcriptional profiles, and that the expression of resistance genes in strain PA49 was independent on the water matrix, with the exception of the MexCD-OprJ efflux pump genes which were overexpressed in response to waste water (Schwartz et al. 2015). The integration of regulatory and metabolic networks is also a pending subject to shed light on bacterial fitness and interaction with the host.

(ii) Exposure assessment: flow cytometry has been successful in quantifying total and viable cells to detect microbial regrowth in DW. The combination of flow cytometry and high-throughput sequencing could further link enumeration and taxonomic identification, as seen in DW samples without disinfectant residual (Prest et al. 2014). For recognized and well characterized pathogens, the use of a marker may be an option that lends itself to the application of PCR for detection and enumeration. Moreover, a small portable sequencing instrument such as the MinIon could serve as biosensor once its limitations are overcome. Both hazard characterization and exposure assessment are crucial in MRA because the concentrations of pathogens after treatment are low, and moreover, the dose associated with pathogenicity for several microorganisms is very low (e.g. the

infectious dose for *C. parvum* can be as low as 5 oocysts, Public Health Agency of Canada 2014).

(iii) Dose-response assessment: host-microbe interactions could be further explored through high-throughput sequencing. For instance, the interaction of pathogens with the host microbiome in healthy and diseased states can be addressed. In the case of gastrointestinal diseases, cell lines could be used instead of animal models (i.e. an *in vitro* model of the intestinal environment consisting of intestinal epithelial cells and intestinal microbiota) to assess pathogenicity and interaction with host microbiota (Brul et al. 2012). Moreover, samples from infected hosts collected in intervals over the duration of the disease can be analyzed and provide insights on the microbe physiology and virulence while it is present in the host. By coupling both the characterization of the pathogen and the host (including molecular data, age, sex, medication and others), a better understanding of host susceptibility can be achieved and extended to the sub-population level (Brul et al. 2012).

Although MRA and Water Safety Plans are useful preventive tools for the management of the systems, the current framework is still reactive because it relies on detection methods that trigger an action once the adverse health effects linked to waterborne pathogens occur, or if indicator organisms (which are poorly correlated with pathogens) are detected above the specified limit. The application of high-throughput sequencing techniques, specifically shotgun DNA sequencing, constitutes an opportunity to capture all the genetic material present in a DW sample and estimate the potential hazards present in it through bioinformatics. In contrast to the currently applied strategy, this constitutes a proactive approach that could be used to rank and prioritize the hazards present in DW and to lead subsequent efforts to further study them. Moreover, this approach constitutes a link between microbial ecology, metagenomics, microbiology and public health, is an extension of the work carried out in this research project, and therefore an opportunity to integrate high-throughput sequencing data into the risk assessment framework with the possibility to benefit DW practice. To achieve this, I propose the estimation of an index based on the annotation and/or prediction of potential hazards belonging to three categories (Figure 6.1):

a) Gene transfer potential: horizontal gene transfer is the transference of genetic material between organisms (other than by replication) and is an important mean of resistance acquisition for bacteria. To assess gene transfer potential, plasmids can

be annotated with BLASTN (Camacho et al. 2009) using the NCBI RefSeq database (http://www.ncbi.nlm.nih.gov/refseq), transposable elements (TEs) can be identified using the sequences annotated as TEs available in GenBank (http://www.ncbi.nlm.nih.gov/genbank), and phages can be obtained with kaiju (Menzel et al. 2016) by annotating the sequences to the NCBI BLAST *nr* database.

b) Resistance potential: antimicrobial resistance occurs when a bacterium becomes resistant to a substance that could previously kill it. These substances include disinfectants and antibiotics. Resistance potential will be assessed through searching for antibiotic resistance genes (ARGs), metal resistance genes (MRGs) (which have been shown to co-select for ARGs, Wright 2007), multi-drug resistance (MDR) efflux pump genes and pathogenicity islands. The CARD database (McArthur et al. 2013) could be used to obtain annotated and putative ARGs; the BacMET database (Pal et al. 2014) could be used to obtain annotated and predicted MRGs; finally, PIPS (Soares et al. 2012) can be used to predict pathogenicity islands.

c) Virulence factors: are properties that allow a microorganism to establish in a host and increase its potential to cause disease. Virulence factors include bacterial toxins, cell surface proteins that mediate bacterial attachment, cell surface carbohydrates and proteins that protect a bacterium, and hydrolytic enzymes that may contribute to the pathogenicity of the bacterium (Chen et al. 2005). Virulence can be assessed through annotation against available databases; for instance, the VFDB database (Chen et al. 2016) encompassing virulence factors of bacterial pathogens, and the DFVF database (Lu et al. 2012) of fungal virulence factors. Additionally, pathogenic proteins can be predicted using MP3 (Gupta et al. 2014).

This multiple annotation/prediction strategy can then be complemented with taxonomic classification (e.g. using kaiju, by Menzel et al. 2016) to quantitatively estimate the contribution of each taxon to the hazards evaluated. The taxonomic groups with greatest contributions to virulence and to the resistome can then be subject to further study using several techniques (as seen in Chapter 2). Before annotation, the raw reads should be quality trimmed and assembled into contigs, to be able to connect resistance and virulence to the identified taxonomic groups. A similar workflow was proposed by Port et al. (2014), focused on environmental monitoring of antibiotic resistance using read-based annotation (instead of contig-based annotation as proposed here). The proposed strategy could be

applied as part of an interdisciplinary research project with the participation of academia and water utilities, the sampling efforts could be focused on locations in the network within a system or across systems that share the same disinfection strategy as a starting point. The point of use would be a suitable sampling location since the index focuses on potential hazards and this is the point where water comes in contact with the consumer.

An aspect to consider regarding the proposed approach and the application of DNA sequencing is the cost associated with it. The cost of DNA sequencing has decreased dramatically since its introduction; for instance, the cost per million base pairs has dramatically decreased over the last 15 years from ~US$8,000 to less than US$0.1 (Genome.gov, accessed: 18-11-2015). The cost-benefit implications of the application of sequencing methods for DW management is out of the scope of the present research project. In terms of skilled labour, the preparation of sample libraries for sequencing (from DNA extraction, PCR until the final library is prepared) requires basic laboratory skills; once the libraries are prepared they can be sent to external sequencing centers that provide both DNA sequencing and basic bioinformatics services. Another option is to sub-contract the sample processing services as well and send the biomass samples to be processed for DNA extraction, PCR and library preparation (usually in the same sequencing center since there are lab facilities/equipment there to do so).

As I progressed in my research project, and started to generate and share my results, and interact with other researchers/practitioners, one thing became clear to me up to this day: the answer to the current and future challenges in DW will not be provided by a single technique/method; instead I believe that the application of a "toolbox" of techniques (i.e. a combination of techniques for microbial characterization, for instance, DNA sequencing, flow cytometry and qPCR; see Chapter 2 for a comprehensive description of techniques) could provide the most information of the systems by providing multiple snapshots that can be linked together to have a more comprehensive idea of "*what's going on*" in the systems. Moreover, a possibility is presented in linking meta'omics data to public health/practice (through bioinformatics, statistics, mathematics, microbial ecology) to anticipate hazards in DW before they become a public health problem and/or an operational problem. This constitutes a promising proactive approach, the possibility of changing the way in which DW systems have been studied and managed for over 100 years.

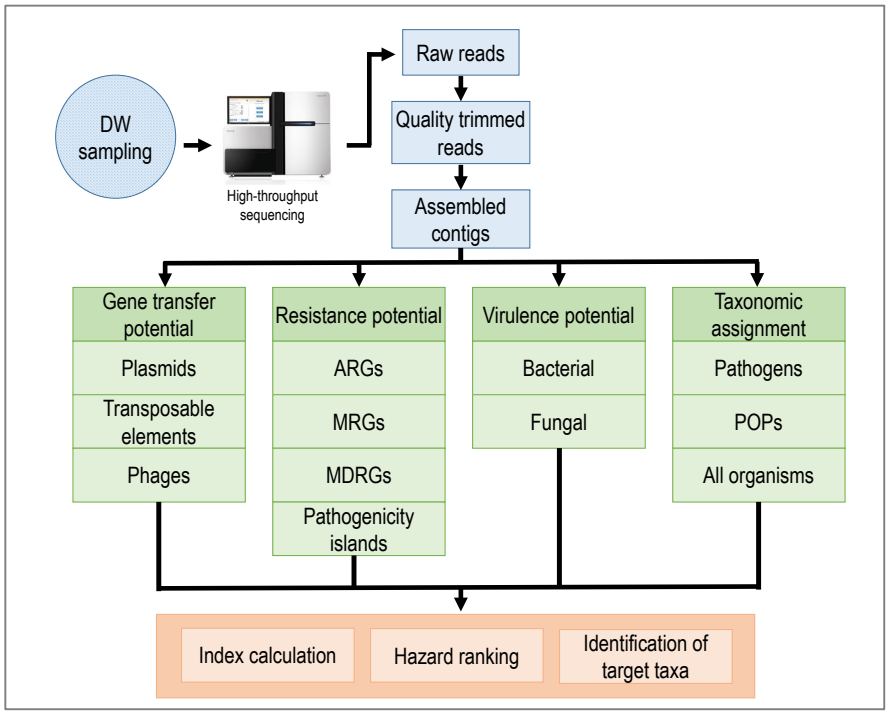

Figure 6.1. Diagram of main steps for the calculation of hazard index. ARGs: antibiotic resistance genes, MRGs: metal resistance genes, MDRGs: multi-drug resistance genes, OPs: opportunistic pathogens.

References

Abe, Y., Skali-Lami, S., Block, J. C., Francius, G. 2012. Cohesiveness and hydrodynamic properties of young drinking water biofilms. *Water Research*, 46(4), pp.1155–1166.

Adams, R. I., Bateman, A. C., Bik, H. M. & Meadow, J. F. 2015. Microbiota of the indoor environment: a meta-analysis. *Microbiome*, 3, pp.1-18.

Ahn, S., Jung, J., Jang, I.A., Madsen, E.L., Park, W. 2016. Role of glyoxylate shunt in oxidative stress response. *Journal of Biological Chemistry*, 291(22), pp.11928–11938.

Albertsen, M., Karst, S. M., Ziegler, A.S., Kirkegaard, R.H., Nielsen, P.H. 2015. Back to Basics – The Influence of DNA Extraction and Primer Choice on Phylogenetic Analysis of Activated Sludge Communities. *PLOS ONE*, 10(7), p.e0132783.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215, pp.403-410.

Asshaue, K. P., Wemheuer, B., Daniel, R. & Meinicke, P. 2015. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, 31, pp.2882-4.

AWWARF & EPA. 2005. Impact of Distribution System Water Quality on Disinfection Efficacy, Available at: http://www.waterrf.org/publicreportlibrary/91094.pdf.

Bai, X., Ma, X., Xu, F., Li, J., Zhang, H. & Xiao, X. 2015. The drinking water treatment process as a potential source of affecting the bacterial antibiotic resistance. *Science of The Total Environment*, 533, pp.24-31.

Bautista-De Los Santos, Q. M., Schroeder, J. L., Blakemore, O., Moses, J., Haffey, M., Sloan, W. & Pinto, A. J. 2016. The impact of sampling, PCR, and sequencing replication on discerning changes in drinking water bacterial community over diurnal time-scales. *Water Research*, 90, pp.216-224.

Bautista-de los Santos, Q.M., Schroeder, J., Sevillano-Rivera, M.C., Sungthong, R., Ijaz, U.Z., Sloan, W.T., Pinto, A. J. 2016. Emerging investigators series: Microbial communities in full-scale drinking water distribution systems – A meta-analysis. *Environmental Science: Water Research and Technology*, 2(4), pp.631–644.

Belila, A., El-Chakhtoura, J., Otaibi, N., Muyzer, G., Gonzalez-Gil, G., Saikaly, P. E., Van Loosdrecht, M. C M, Vrouwenvelder, J. S. 2016. Bacterial community structure and variation in a full-scale seawater desalination plant for drinking water production. *Water Research*, 94, pp.62–72.

Bengtsson-Palme, J., Alm Rosenblad, M., Molin, M., Blomberg, A. 2014. Metagenomics reveals that detoxification systems are underrepresented in marine bacterial communities. *BMC Genomics*, 15(1), p.749.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, 57, pp.289-300.

Bolger, A.M., Lohse, M. & Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114–2120.

Brooks, J.P., Edwards, D.J., Harwich, M.D., Rivera, M.C., Fettweis, J.M., Serrano, M.G., Reris, R.A., Sheth, N.U., Huang, B., Girerd, P., Strauss, J.F., Jefferson, K.K., Buck, G.A. 2015. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiology*, 15(1), p.66.

Brul, S., Bassett, J., Cook, P., Kathariou, S., McClure, P., Jasti, P.R., Betts, R. 2012. "Omics" technologies in quantitative microbial risk assessment. *Trends in Food Science & Technology*, 27(1), pp.12–24.

Bucheli-Witschel, M., Kötzsch, S., Darr, S., Widler, R., Egli, T. 2012. A new method to assess the influence of migration from polymeric materials on the biostability of drinking water. *Water Research*, 46(13), pp.4246–4260.

Buchfink, B., Xie, C. & Huson, D.H. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), pp.59–60.

Burstein, D., Amaro, F., Zusman, T., Lifshitz, Z., Cohen, O., Gilbert, J. A., Pupko, T., Shuman, H. A. & Segal, G. 2016. Genomic analysis of 38 Legionella species identifies large and diverse effector repertoires. *Nature Genetics*, 48, 167-175.

Buse, H.Y., Lu, J.R., Struewing, I.T., Ashbolt, N.J. 2014. Eukaryotic diversity in premise drinking water using 18S rDNA sequencing: implications for health risks. *Environmental Science and Pollution Research*, 21(10), pp.6759–6760.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*, 10, p.421.

Camper, A. & Dirckx, P. 1996. Distribution system as a reactor. Available at: https://www.biofilm.montana.edu/resources/images/industrial-systems-processes/distribution-system-reactor.html [Accessed July 10, 2016].

Caporaso, J. G., Lauber, C. L., Walters, W. A., berg-lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N. & knight, R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 108, pp.4516-4522.

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

Carragher, B.J., Stewart, R.A. & Beal, C.D. 2012. Quantifying the influence of residential water appliance efficiency on average day diurnal demand patterns at an end use level: A precursor to optimised water service infrastructure planning. *Resources Conservation and Recycling*, 62, pp.81–90.

Chakravorty, S., Helb, D., Burday, M., Connell, N., Alland, D. 2007. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, 69(2), pp.330–339.

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X.X., Wu, W.M., Zhang, T. 2013. Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Nature Scientific Reports*, 3, p.3550.

Chen, L., Yang, J., Yu, J., Yao, Z., Sun, L., Shen, Y., Jin, Q. 2005. VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Research*, 33(DATABASE ISS).

Chen, L., Zheng, D., Liu, B., Yang, J., Jin, Q. 2016. VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. *Nucleic Acids Research*, 44(D1), pp.D694-7.

Chevance, F.F. & Hughes, K.T. 2008. Coordinating assembly of a bacterial macromolecular machine. *Nature Reviews Microbiology*, 6(6), pp.455–65.

Chiao, T.H., Clancy, T. M., Pinto, A., Xi, C. & Raskin, L. 2014. Differential Resistance of Drinking Water Bacterial Populations to Monochloramine Disinfection. *Environmental Science & Technology*, 48, pp.4038-4047.

Choi, Y.C. & Morgenroth, E. 2003. Monitoring biofilm detachment under dynamic changes in shear stress using laser-based particle size analysis and mass fractionation. *Water Science & Technology*, 47(5), pp.69–76.

Clooney, A.G. et al. 2016. Comparing Apples and Oranges? Next Generation Sequencing and Its Impact on Microbiome Analysis. *PLOS ONE*, 11(2), p.e0148028.

Costa, D., Mercier, A., Gravouil, K., Lesobre, J., Delafont, V., Bousseau, A., Verdon, J. & Imbert, C. 2015. Pyrosequencing analysis of bacterial diversity in dental unit waterlines. *Water Research*, 81, pp.223-231.

Cuthbertson, L., Rogers, G. B., Walker, A. W., Oliver, A., Hafiz, T., Hoffman, L. R., Carroll, M. P., Parkhill, J., Bruce, K. D. & Van Der Gast, C. J. 2014. Time between Collection and Storage Significantly Influences Bacterial Sequence Composition in Sputum Samples from Cystic Fibrosis Respiratory Infections. *Journal of Clinical Microbiology*, 52, pp.3011-3016.

D'Auria, G., Jiménez-Hernández, N., Peris-Bondia, F., Moya, A., Latorre, A. 2010. Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics*, 11, p.181.

Daims, H., Lebedeva, E. V., Pjevac, P., Han, P., Herbold, C., Albertsen, M., Jehmlich, N., Palatinszky, M., Vierheilig, J., Bulaev, A., Kirkegaard, R. H., Bergen, M. V., Rattei, T., Bendinger, B., Nielsen, P. H. & Wagner, M. 2015. Complete nitrification by Nitrospira bacteria. *Nature,* 528, pp.504–509.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G. L. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7), pp.5069–5072.

Dias, P.J., Seret, M.L., Goffeau, A., Correia, I.S., Baret, P.V. 2010. Evolution of the 12-spanner drug:H+ antiporter DHA1 family in hemiascomycetous yeasts. *Omics : a journal of integrative biology*, 14(6), pp.701–710.

Douterelo, I., Husband, S. & Boxall, J.B. 2014. The bacteriological composition of biomass recovered by flushing an operational drinking water distribution system. *Water Research*, 54, pp.100–114.

Douterelo, I., Sharpe, R.L. & Boxall, J.B. 2013. Influence of hydraulic regimes on bacterial community structure and composition in an experimental drinking water distribution system. *Water Research*, 47(2), pp.503–516.

Edwards, R., 2012. UPGMA clustering. Available at: http://bioware.soton.ac.uk/upgma.html [Accessed April 25, 2016].

El-Chakhtoura, J., Prest, E., Saikaly, P., Van Loosdrecht, M., Hammes, F. & Vrouwenvelder, H. 2015. Dynamics of bacterial communities before and after distribution in a full-scale drinking water network. *Water Research*, 74, pp.180-190.

Excoffier, L., Smouse, P.E. & Quattro, J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), pp.479–491.

Falkinham, J., Pruden, A. & Edwards, M. 2015. Opportunistic Premise Plumbing Pathogens: Increasingly Important Pathogens in Drinking Water. *Pathogens*, 4(2), pp.373–386.

Feinstein, L. M., Sul, W. J. & Blackwood, C. B. 2009. Assessment of bias associated with incomplete extraction of microbial DNA from soil. *Applied and Environmental Microbiology*, 75, pp.5428-5433.

Ferguson, G.P., Battista, J.R., Annettee, T. L., Booth, I.R. 2000. Protection of the DNA during the exposure of Escherichia coli cells to a toxic metabolite: The role of the KefB and KefC potassium channels. *Molecular Microbiology*, 35(1), pp.113–122.

Ferguson, G.P., McLaggan, D. & Booth, I.R. 1995. Potassium channel activation by glutathione-S-conjugates in Escherichia coli: protection against methylglyoxal is mediated by cytoplasmic acidification. *Molecular Microbiology*, 17(6), pp.1025–1033.

Friedman, J. & Alm, E.J. 2012. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9), p.e1002687.

Genome.gov, Sequencing costs. Available at: http://www.genome.gov/sequencingcosts/%0D [Accessed November 18, 2015].

Gilbert, J.A., Meyer, F., Antonopoulos, D., Balaji, P., Brown, C.T., Brown, C.T., Desai, N., Eisen, J.A., Evers, D., Field, D., Feng, W., Huson, D., Jansson, J., Knight, R., Knight, J., Kolker, E., Konstantindis, K., Kostka, J., Kyrpides, N., Mackelprang, R., McHardy, A., Quince, C., Raes, J., Sczyrba, A., Shade, A., Stevens, R. 2010. Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Standards in Genomic Sciences*, 3(3), pp.243–248.

Gillespie, S., Lipphaus, P., Green, J., Parsons, S., Weir, P., Juskowiak, K., Jefferson, B., Jarvis, P. & Nocker, A. 2014. Assessing microbiological water quality in drinking water distribution systems with disinfectant residual using flow cytometry. *Water Research*, 65, pp.224-234.

Gomez-alvarez, V., Revetta, R. P. & Santo Domingo, J. W. 2012a. Metagenomic Analyses of Drinking Water Receiving Different Disinfection Treatments. *Applied and Environmental Microbiology*, 78(17), pp.6095-102.

Gomez-alvarez, V., Revetta, R. P. & Santo Domingo, J. W. 2012b. Metagenomic analyses of drinking water receiving different disinfection treatments. *Applied and Environmental Microbiology*, 78, pp.6095-6102.

Gray, M.J., Wholey, W.-Y. & Jakob, U. 2013. Bacterial responses to reactive chlorine species. *Annual Review of Microbiology*, 67, pp.141–60.

Gupta, A., Kapil, R., Dhakan, D.B., Sharma, V.K. 2014. MP3: A software tool for the prediction of pathogenic proteins in genomic and metagenomic data. *PLOS ONE*, 9(4), e93907.

Henne, K., Kahlisch, L., Brettar, I. & Höfle, 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. *Applied and Environmental Microbiology*, 78, pp.3530-3538.

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

Huang, K., Zhang, X.X., Shi, P., Wu, B. & Ren, H. 2014. A comprehensive insight into bacterial virulence in drinking water using 454 pyrosequencing and Illumina high-throughput sequencing. *Ecotoxicology and Environmental Safety*, 109, pp.15-21.

Hwang, C., Ling, F., Andersen, G. L., Lechevallier, M. W. & Liu, W.T. 2012a. Microbial Community Dynamics of an Urban Drinking Water Distribution System Subjected to Phases of Chloramination and Chlorination Treatments. *Applied and Environmental Microbiology*, 78, pp.7856-7865.

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

Ji, P., Parks, J., Edwards, M. A. & Pruden, A. 2015. Impact of Water Chemistry, Pipe Material and Stagnation on the Building Plumbing Microbiome. *PLOS ONE*, 10, e0141087.

Jia, S., Shi, P., Hu, Q., Li, B., Zhang, T. & Zhang, X.-X. 2015. Bacterial Community Shift Drives Antibiotic Resistance Promotion during Drinking Water Chlorination. *Environmental Science & Technology*, 49, pp.12271-12279.

Jones, M.B., Highlander, S.K., Anderson, E.L., Li, W., Dayrit, M., Klitgord, N., Fabani, M.M., Seguritan, V., Green, J., Pride, D.T., Yooseph, S., Biggs, W., Nelson, K.E., Venter, J.C. 2015. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proceedings of the National Academy of Sciences*, 112(45), pp.14024–14029.

Joshi, N. & Fass, J., 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at: https://github.com/najoshi/sickle.

Kalle, E., Kubista, M. & Rensing, C. 2014. Multi-template polymerase chain reaction. *Biomolecular Detection and Quantification*, 2(C), pp.11–29.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1), pp.D457–D462.

Kennedy, K., Hall, M.W., Lynch, M.D.J., Moreno-Hagelsieb, G., Neufeld, J.D. 2014. Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. Applied and *Environmental Microbiology*, 80(18), pp.5717–5722.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J.A., Hugenholtz, P., Van Der Lelie, D., Meyer, F., Stevens, R., Bailey, M.J., Gordon, J.I., Kowalchuk, G.A., Gilbert, J.A. 2012. Unlocking the potential of metagenomics through replicated experimental design. *Nature Biotechnology*, 30(6), pp.513–520.

Kooij, D. V. D. 1992. Assimilable Organic Carbon as an Indicator of Bacterial Regrowth. *Journal American Water Works Association*, 84, pp.57-65.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., Huttenhower, C. & Ley, R. E. 2013. A Guide to Enterotypes across the Human Body: Meta-Analysis of Microbial Community Structures in Human Microbiome Datasets. *PLOS Computational Biology*, 9, p.e1002863.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K., Schloss, P.D. 2013. Development of a Dual-Index Sequencing Strategy and Curation Pipeline for Analyzing Amplicon Sequence Data on the MiSeq Illumina Sequencing Platform. *Applied and Environmental Microbiology*, 79(17), pp.5112–5120.

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDOnald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G. & Huttenhower, C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31, pp.814-821.

Lauber, C. L., Zhou, N., Gordon, J. I., Knight, R. & Fierer, N. 2010. Effect of storage conditions on the assessment of bacterial community structure in soil and human-associated samples. *FEMS Microbiology Letters*, 307, pp.80-86.

Lautenschlager, K., Boon, N., Wang, Y., Egli, T., Hammes, F. 2010. Overnight stagnation of drinking water in household taps induces microbial growth and changes in community composition. *Water Research*, 44(17), pp.4868–4877.

Lautenschlager, K., Hwang, C., Ling, F., Liu, W.T., Boon, N., Köster, O., Egli, T. & Hammes, F. 2014. Abundance and composition of indigenous bacterial communities in a multi-step biofiltration-based drinking water treatment plant. *Water Research*, 62, pp.40-52.

Lautenschlager, K., Hwang, C., Liu, W. T., Boon, N., Koster, O., Vrouwenvelder, H., Egli, T. & Hammes, F. 2013. A microbiology-based multi-parametric approach towards assessing biological stability in drinking water distribution networks. *Water Research*, 47, pp.3015-3025.

LeChevallier, M. W., Schulz, W. & Lee, R. G. 1991. Bacterial nutrients in drinking water. *Applied and Environmental Microbiology*, 57, pp.857-862.

Li, G. & Tang, J.X. 2006. Low flagellar motor torque and high swimming efficiency of Caulobacter crescentus swarmer cells. *Biophysical Journal*, 91(7), pp.2726–2734.

Lin, W., Yu, Z., Zhang, H. & Thompson, I. P. 2014. Diversity and dynamics of microbial communities at each step of treatment plant for potable water generation. *Water Research*, 52, 218-230.

Ling, F., Hwang, C., LeChevallier, M. W., Andersen, G. L. & Liu, W.T. 2015. Core-satellite populations and seasonality of water meter biofilms in a metropolitan drinking water distribution system. *ISME Journal*, 10(3):582-95.

Lipphaus, P., Hammes, F., Kötzsch, S., Green, J., Gillespie, S., Nocker, A. 2013. Microbiological tap water profile of a medium-sized building and effect of water stagnation. *Environmental Technology*, 35(5), pp.620–628.

Liu, G., Bakker, G. L., Li, S., Vreeburg, J. H. G., Verberk, J. Q. J. C., Medema, G. J., Liu, W. T. & Van Dijk, J. C. 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. *Environmental Science & Technology*, 48, pp.5467-5476.

Liu, G., Verberk, J.Q.J.C. & Dijk, J.C. 2013. Bacteriology of drinking water distribution systems: an integral and multidimensional review. *Applied Microbiology and Biotechnology*, 97(21), pp.9265–9276.

Liu, T., Kong, W., Chen, N., Zhu, J., Wang, J., He, X., Jin, Y. 2016. Bacterial characterization of Beijing drinking water by flow cytometry and MiSeq sequencing of the 16S rRNA gene. *Ecology and Evolution*, 6(4), pp. 923–934.

Love, M.I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), pp.1–34.

Lu, T., Yao, B. & Zhang, C., 2012. DFVF: database of fungal virulence factors. Database : the journal of biological databases and curation, 2012, p.bas032. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3478563&tool=pmcentrez&rendertype=abstract.

Lucas S.A., Coombes, P.J., Sharma, A. K. 2010. The impact of diurnal water use patterns, demand management and rainwater tanks on water supply network design. *Water Science and Technology: Water Supply*, 10(1), pp.69–80.

Madigan, M.T., Martinko, J.M., Stahl, D.A., Clark, D.P. 2012. Brock Biology of microorganisms. Pearson Benjamin Cummings, 13th edition, pp. 1043.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1): Next Generation Sequencing Data Analysis. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200.

Masip, L., Veeravalli, K. & Georgiou, G. 2006. The many faces of glutathione in bacteria. *Antioxidants & Redox Signaling*, 8(5–6), pp.753–762.

Mathieu, L., Bertrand, I., Abe, Y., Angel, E., Block, J. C., Skali-Lami, S., Francius, G. 2014. Drinking water biofilm cohesiveness changes under chlorination or hydrodynamic stress. *Water Research*, 55, pp.175–184.

McArthur, A.G., Waglechner, N., Nizam, F., Yan, A., Azad, M.A., Baylay, A.J., Bhullar, K., Canova, M.J., De Pascale, G., Ejim, L., Kalan, L., King, A.M., Koteva, K., Morar, M., Mulvey, M.R., O'Brien, J.S., Pawlowski, A.C., Piddock, L. J. V., Spanogiannopoulos, P., Sutherland, A.D., Tang, I., Taylor, P.L., Thaker, M., Wang, W., Yan, M., Yu, T., Wright, G.D. 2013. The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, 57(7), pp.3348–3357.

McCoy, S.T. & VanBriesen, J.M. 2012. Temporal Variability of Bacterial Diversity in a Chlorinated Drinking Water Distribution System. *Journal of Environmental Engineering-ASCE*, 138(7), pp.786–795.

McIntyre, A. 2010. Life in the world's oceans, Available at: http://comlmaps.org/mcintyre.

Menzel, P., Ng, K.L. & Krogh, A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7. Available at: http://dx.doi.org/10.1038/ncomms11257.

Moisi, M., Jenul, C., Butler, S.M., New, A., Tutz, S., Reidl, J., Klose, K.E., Camilli, A., Schild, S. 2009. A novel regulatory protein involved in motility of Vibrio cholerae. *Journal of Bacteriology*, 191(22), pp.7027–7038.

Nishino, K. & Yamaguchi, A. 2001. Analysis of a complete library of putative drug transporter genes in Escherichia coli. *Journal of Bacteriology*, 183(20), pp.5803–5812.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 27, 29-34.

Oksanen Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P.R., O'hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Wagner, H. J 2013. vegan: Community Ecology Package. Available at: http://cran.r-project.org/package=vegan.

Padilla, C.C., Ganesh, S., Gantt, S., Huhman, A., Parris, D.J., Sarode, N., Stewart, F.J. 2015. Standard filtration practices may significantly distort planktonic microbial diversity estimates. *Frontiers in Microbiology*, 6(June), pp.1–10.

Pal, C., Bengtsson-Palme, J., Rensing, C., Kristiansson, E., Larsson, D.G.J. 2014. BacMet: Antibacterial biocide and metal resistance genes database. *Nucleic Acids Research*, 42(D1), pp.D737-D743.

Peng, Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L. 2012. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11), pp.1420–1428.

Peterson, C.N., Mandel, M.J. & Silhavy, T.J. 2005. Escherichia coli starvation diets: Essential nutrients weigh in distinctly. *Journal of Bacteriology*, 187(22), pp.7549–7553.

Pinto, A. J. & Raskin, L. 2012. PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets. *PLOS ONE*, 7, p.e43093.

Pinto, A. J., Marcus, D. N., Ijaz, U. Z., Bautista-De Los Santos, Q. M., Dick, G. J. & Raskin, L. 2015. Metagenomic Evidence for the Presence of Comammox Nitrospira-Like Bacteria in a Drinking Water System. *mSphere*, 1, DOI: 10.1128/mSphere.00054-15.

Pinto, A. J., Schroeder, J., Lunn, M., Sloan, W. & Raskin, L. 2014. Spatial-Temporal Survey and Occupancy-Abundance Modeling To Predict Bacterial Community Dynamics in the Drinking Water Microbiome. *Mbio*, 5 (3), pp.e01135-14.

Pinto, A.J., Xi, C. & Raskin, L. 2012. Bacterial Community Structure in the Drinking Water Microbiome Is Governed by Filtration Processes. *Environmental Science & Technology*, 46(16), pp.8851–8859.

Pophaly, S., Singh, R., Kaushik, J.K., Tomar, S. 2012. Current status and emerging role of glutathione in food grade lactic acid bacteria. *Microbial Cell Factories*, 11(1), p.114.

Port, J.A., Cullen, A.C., Wallace, J.C., Smith, M.N., Faustman, E.M. 2014. Metagenomic frameworks for monitoring antibiotic resistance in aquatic environments. *Environmental Health Perspectives*, 122(3), pp.222–228.

Prest, E. I., El-Chakhtoura, J., Hammes, F., Saikaly, P. E., Van Loosdrecht, M. C. M. & Vrouwenvelder, J. S. 2014. Combining flow cytometry and 16S rRNA gene pyrosequencing: A promising approach for drinking water monitoring and characterization. *Water Research*, 63, pp.179-189.

Proctor, C. R. & Hammes, F. 2015. Drinking water microbiology—from measurement to management. *Current Opinion in Biotechnology*, 33, pp.87-94.

Proctor, C.R., Gachter, M., Kotzsch, S., Rolli, F., Sigrist, R., Walser, J.C., Hammes, F. 2016. Biofilms in shower hoses - choice of pipe material influences bacterial growth and communities. *Environmental Science: Water Research & Technology*, 2, pp.670-682.

Prosser, J.I. 2010. Replicate or lie. *Environmental Microbiology*, 12(7), pp.1806–1810.

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

Public Health Agency of Canada, 2014. Cryptosporidium parvum - Pathogen Safety Data Sheet. Available at: http://www.phac-aspc.gc.ca/lab-bio/res/psds-ftss/msds48e-eng.php#footnote12 [Accessed July 26, 2016].

R Core Team, 2014. R: A Language and Environment for Statistical Computing. Available at: http://www.r-project.org.

Riccillo, P.M., Muglia, C.I., De Bruijn, F.J., Roe, A.J., Booth, I.R., Aguilar, O.M. 2000. Glutathione is involved in environmental stress responses in Rhizobium tropici, including acid tolerance. *Journal of Bacteriology*, 182(6), pp.1748–1753.

Roeselers, G., Coolen, J., Van Der Wielen, P. W. J. J., Jaspers, M. C., Atsma, A., De Graaf, B. & Schuren, F. 2015. Microbial biogeography of drinking water: patterns in phylogenetic diversity across space and time. *Environmental Microbiology*, 17 (7), pp.2505-2514.

Rożej, A., Cydzik-Kwiatkowska, A., Kowalska, B., Kowalski, D. 2015. Structure and microbial diversity of biofilms on different pipe materials of a model drinking water distribution systems. *World Journal of Microbiology & Biotechnology*, 31, pp.37–47.

Rusconi, R. & Stocker, R. 2015. Microbes in flow. *Current Opinion in Microbiology*, 25, pp.1–8.

Salipante, S.J., Kawashima, T., Rosenthal, C., Hoogestraat, D.R., Cummings, L.A., Sengupta, D.J., Harkins, T.T., Cookson, B.T., Hoffman, N.G. 2014. Performance Comparison of Illumina and Ion Torrent Next-Generation Sequencing Platforms for 16S rRNA-Based Bacterial Community Profiling. *Applied and Environmental Microbiology*, 80(24), pp.7583–7591.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J. & Walker, A. W. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12, pp.87-87.

Schleheck, D., Barraud, N., Klebensberger, J., Webb, J.S., McDougald, D., Rice, S., Kjelleberg, S. 2009. Pseudomonas aeruginosa PAO1 preferentially grows as aggregates in liquid batch cultures and disperses upon starvation. *PLOS ONE*, 4(5), p.e5513.

Schloss, P., Westcott, S., Ryabin, T., Hall, J., Hartmann, M., Hollister, E., Lesniewski, R., Oakley, B., Parks, D., Robinson, C., Sahl, J., Stres, B., Thallinger, G., Van Horn, D. & Weber, C. 2009. Introducing mothur: Open Source, Platform-independent, Community-supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology*, 75 (23), pp. 7537-754.

Schmidt, P.A., Bálint, M., Greshake, B., Bandow, C., Römbke, J., Schmitt, I. 2013. Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry*, 65(0), pp.128–132.

Schoen, M. E. & Ashbolt, N. J. 2011. An in-premise model for Legionella exposure during showering events. *Water Research*, 45, pp.5826-5836.

Schroeder, J. L., Lunn, M., Pinto, A. J., Raskin, L. & Sloan, W. T. 2015. Probabilistic Models to Describe the Dynamics of Migrating Microbial Communities. *PLOS ONE*, 10, p.e0117221.

Schwartz, T., Armant, O., Bretschneider, N., Hahn, A., Kirchen, S., Seifert, M., Dötsch A. 2015. Whole genome and transcriptome analyses of environmental antibiotic sensitive and multi-resistant Pseudomonas aeruginosa isolates exposed to waste water and tap water. *Microbial Biotechnology*, 8(1), pp.116–130.

Sekar, R., Deines, P., Machell, J., Osborn, A.M., Biggs, C.A., Boxall, J.B. 2012. Bacterial water quality and network hydraulic characteristics: a field study of a small, looped water distribution system using culture-independent molecular methods. *Journal of Applied Microbiology*, 112(6), pp.1220–1234.

Shade, A., Gregory Caporaso, J., Handelsman, J., Knight, R. & Fierer, N. 2013. A meta-analysis of changes in bacterial and archaeal communities with time. *ISME Journal*, 7(8), pp.1493-506.

Shaw, J. L. A., Monis, P., Weyrich, L. S., Sawade, E., Drikas, M. & Cooper, A. J. 2015. Using Amplicon Sequencing To Characterize and Monitor Bacterial Diversity in Drinking Water Distribution Systems. *Applied and Environmental Microbiology*, 81, pp.6463-73.

Smith, D.P. & Peay, K.G. 2014. Sequence Depth, Not PCR Replication, Improves Ecological Inference from Next Generation DNA Sequencing. *PLOS ONE*, 9(2), p.e90234.

Soares, S.C., Abreu, V.A.C., Ramos, R.T.J, Cerdeira, L., Silva, A., Baumbach, J., Trost, E., Tauch, A., Hirata Jr., R., Mattos-Guaraldi, A.L., Miyoshi, A., Azevedo, V. 2012. PIPS: Pathogenicity island prediction software. *PLOS ONE*, 7(2).

Sockett, R. E. & Lambert, C. 2004. Bdellovibrio as therapeutic agents: a predatory renaissance? *Nature Reviews Microbiology*, 2, pp.669-675.

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M. & Herndl, G. J. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103, pp.12115-12120.

Song, Z., Schlatter, D., Kennedy, P., Kinkel, L.L., Kistler, H.C., Nguyen, N., Bates, S.T. 2015. Effort versus reward Preparing samples for fungal community characterization in high-throughput sequencing surveys of soils. *PLOS ONE*, 10(5), p.e0127234.

Staley, C., Gould, T.J., Wang, P., Phillips, J., Cotner, J.B., Sadowsky, M.J. 2015. Evaluation of water sampling methodologies for amplicon-based characterization of bacterial community structure. *Journal of Microbiological Methods*, 114, pp.43–50.

Stamatakis, A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics* 30, pp.1312-1313.

Terashima, H., Kojima, S. & Homma, M. 2008. Chapter 2 Flagellar Motility in Bacteria. Structure and Function of Flagellar Motor. *International Review of Cell and Molecular Biology*, 270(C), pp.39–85.

Van Der Wielen, P. W. & Van Der Kooij, D. 2013. Nontuberculous mycobacteria, fungi, and opportunistic pathogens in unchlorinated drinking water in The Netherlands. *Applied and Environmental Microbiology*, 79, pp.825-34.

Van Der Wielen, P. W. J. J., Voost, S. & Van Der Kooij, D. 2009. Ammonia-Oxidizing Bacteria and Archaea in Groundwater Treatment and Drinking Water Distribution Systems. *Applied and Environmental Microbiology*, 75, pp.4687-4695.

Van Kessel, M. A. H. J., Speth, D. R., Albertsen, M., Nielsen, P. H., Op Den Camp, H. J. M., Kartal, B., Jetten, M. S. M. & Lücker, S. 2015. Complete nitrification by a single microorganism. *Nature*, 528, pp.555-559.

Vasileiadis, S., Puglisi, E., Arena, M., Cappa, F., Cocconcelli, P.S., Trevisan, M. 2012. Soil Bacterial Diversity Screening Using Single 16S rRNA Gene V Regions Coupled with Multi-Million Read Generating Sequencing Technologies. *PLOS ONE*, 7(8), p.e42671.

Wang, H., Edwards, M., Falkinham, J.O., Pruden, A. 2012. Molecular survey of the occurrence of Legionella spp., Mycobacterium spp., Pseudomonas aeruginosa, and amoeba hosts in two chloraminated drinking water distribution systems. *Applied and Environmental Microbiology,* 78(17), pp.6285–94.

Wang, H., Masters, S., Edwards, M. A., Falkinham, J. O. & Pruden, A. 2014a. Effect of Disinfectant, Water Age, and Pipe Materials on Bacterial and Eukaryotic Community Structure in Drinking Water Biofilm. *Environmental Science & Technology*, 48, pp.1426-1435.

Wang, H., Proctor, C. R., Edwards, M. A., Pryor, M., Santo Domingo, J. W., Ryu, H., Camper, A. K., Olson, A. & Pruden, A. 2014b. Microbial Community Response to Chlorine Conversion in a Chloraminated Drinking Water Distribution System. *Environmental Science & Technology*, 48, pp.10624-10633.

Wang, Q., Garrity, G., Tiedje, J. & Cole, J. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73, pp.5261-7.

Warnes, G., Bolker, B. & Lumley, T. gplots: Various R programming tools for plotting data. R package version 2.6.0. Available at: https://cran.r-project.org/web/packages/gplots/index.html

Weiss, S. J., Xu, Z., Amir, A., Peddada, S., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vazquez-Baeza, Y. & Birmingham, A. 2015. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. *PeerJ PrePrints*.

Westcott, S. L. & Schloss, P. D. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, 3, p.e1487.

WHO, W. H. O.-. 2011. Guidelines for Drinking-water Quality. Fourth edition. Available at: http://apps.who.int/iris/bitstream/10665/44584/1/9789241548151_eng.pdf

WHO, W.H.O.-, 2004. Water Treatment and Pathogen Control: Process Efficiency in Achieving Safe Drinking Water M. W. LeChevallier & A. Kwok-Keung, eds., London: IWA Publishing.

Wickham, H., 2009. ggplot2: elegant graphics for data analysis, Springer New York. Available at: http://had.co.nz/ggplot2/book.

Willems, A., 2014. The family Comamonadaceae. In: The Prokaryotes: Alphaproteobacteria and Betaproteobacteria. pp. 777–851. ISBN 978-3-642-30197-1.

Wimpenny, J., Manz, W. & Szewzyk, U. 2000. Heterogeneity in biofilms. *FEMS Microbiology Reviews*, 24(5), pp.661–671.

Winter, J., Ilbert, M., Graf, P.C.F, Özcelik, D., Jakob, U. 2008. Bleach Activates a Redox-Regulated Chaperone by Oxidative Protein Unfolding. *Cell*, 135(4), pp.691–701.

Wintzingerode, F. v, Gobel, U.B. & Stackebrandt, E. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiology Reviews*, 21, pp.213–229.

Wright, G.D. 2007. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Reviews Microbiology*, 5(1740–1534 (Electronic)), pp.175–186.

Wright, K. 2015. corrgram. Available at: https://cran.r-project.org/web/packages/corrgram/index.html.

Wullings, B. A., Bakker, G. & Van Der Kooij, D. 2011. Concentration and Diversity of Uncultured Legionella spp. in Two Unchlorinated Drinking Water Supplies with Different Concentrations of Natural Organic Matter. *Applied and Environmental Microbiology*, 77, pp.634-641.

Xi, C., Zhang, Y., Marrs, C.F., Ye, W., Simon, C., Foxman, B., Nriagu, J. 2009. Prevalence of antibiotic resistance in drinking water treatment and distribution systems. *Applied and Environmental Microbiology*, 75(17), pp.5714–5718.

Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W., Glöckner, F.O. 2014. The SILVA and "all-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Research*, 42(D1), pp.D643–D648.

Yu, J., Kim, D. & Lee, T. 2010. Microbial diversity in biofilms on water distribution pipes of different materials. *Water Science and Technology*, 61(1), pp.163–171.

Zeng, D. N., Fan, Z. Y., Chi, L., Wang, X., Qu, W. D. & Quan, Z. X. 2013. Analysis of the bacterial communities associated with different drinking water treatment processes. *World Journal of Microbiology & Biotechnology*, 29, pp.1573-1584.

Zhang, H., Gao, S., Lercher, M. J., Hu, S. & Chen, W. H. 2012. EvolView, an online tool for visualizing, annotating and managing phylogenetic trees. *Nucleic Acids Research*, 40, pp.569-572.

Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. 2014. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics*, 30, pp.614-20.

Zhang, Y. & He, Q.A. 2013. Characterization of bacterial diversity in drinking water by pyrosequencing. *Water Science and Technology: Water Supply*, 13(2), pp.358–367.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.H., Tu, Q., Xie, J., Van Nostrand, J.D., He, Z., Yang, Y. 2011. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME Journal*, 5(8), pp.1303–1313.

Appendix A

Figure A1. Workflow illustrating data collection, data processing and data analysis steps.

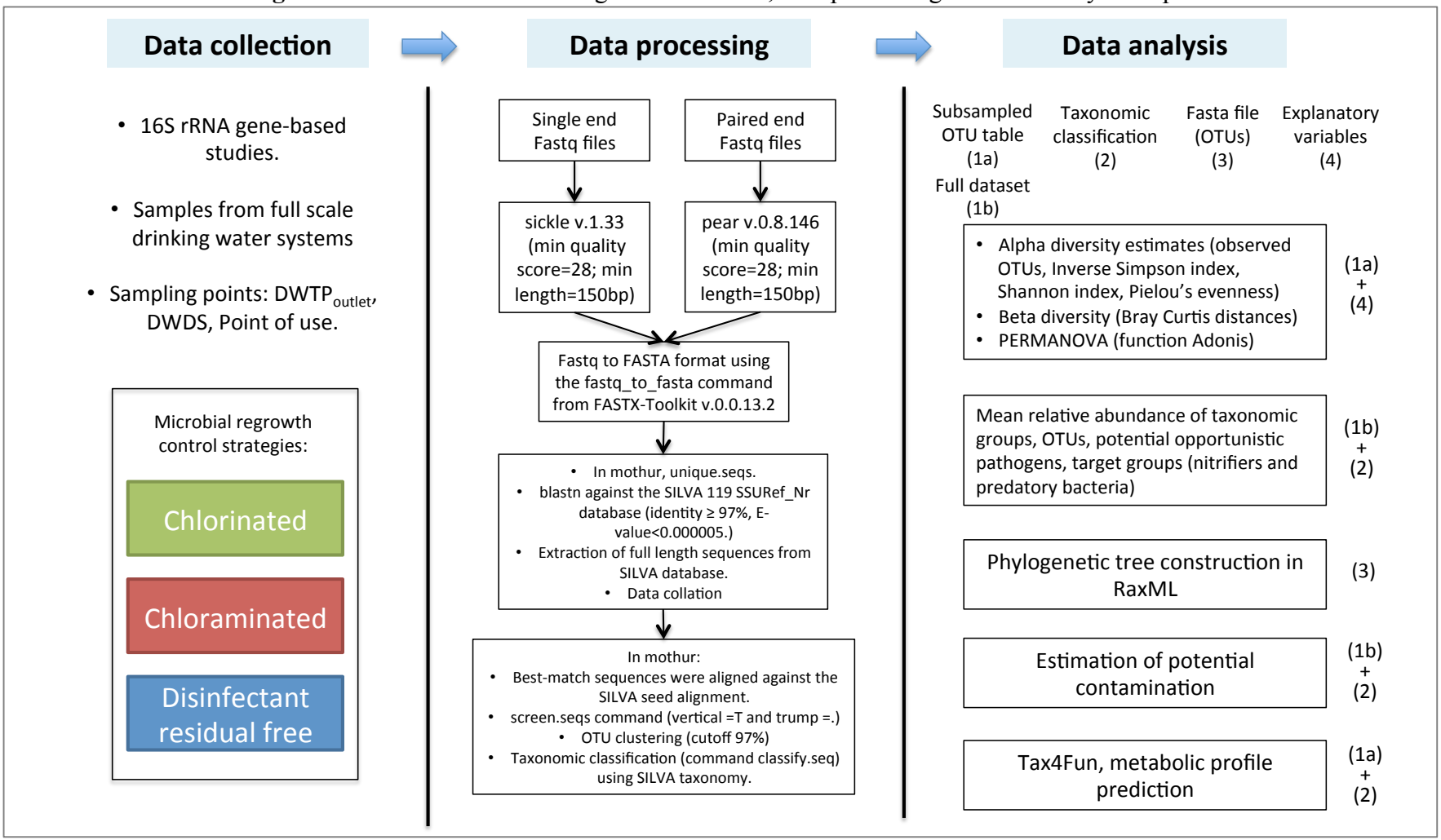

Figure A2. Heatmap of OTU abundance (OTUs with relative abundance >0.01% in the subsampled OTU table). Abundances were scaled with a Z-score transformation to improve visualization. The sample dendrogram was generated with Bray Curtis distances and UPGMA clustering method. Grouping information is indicated by the color legend (Chl: chlorinated, Chm: chloraminated, Drf: Disinfectant residual-free).

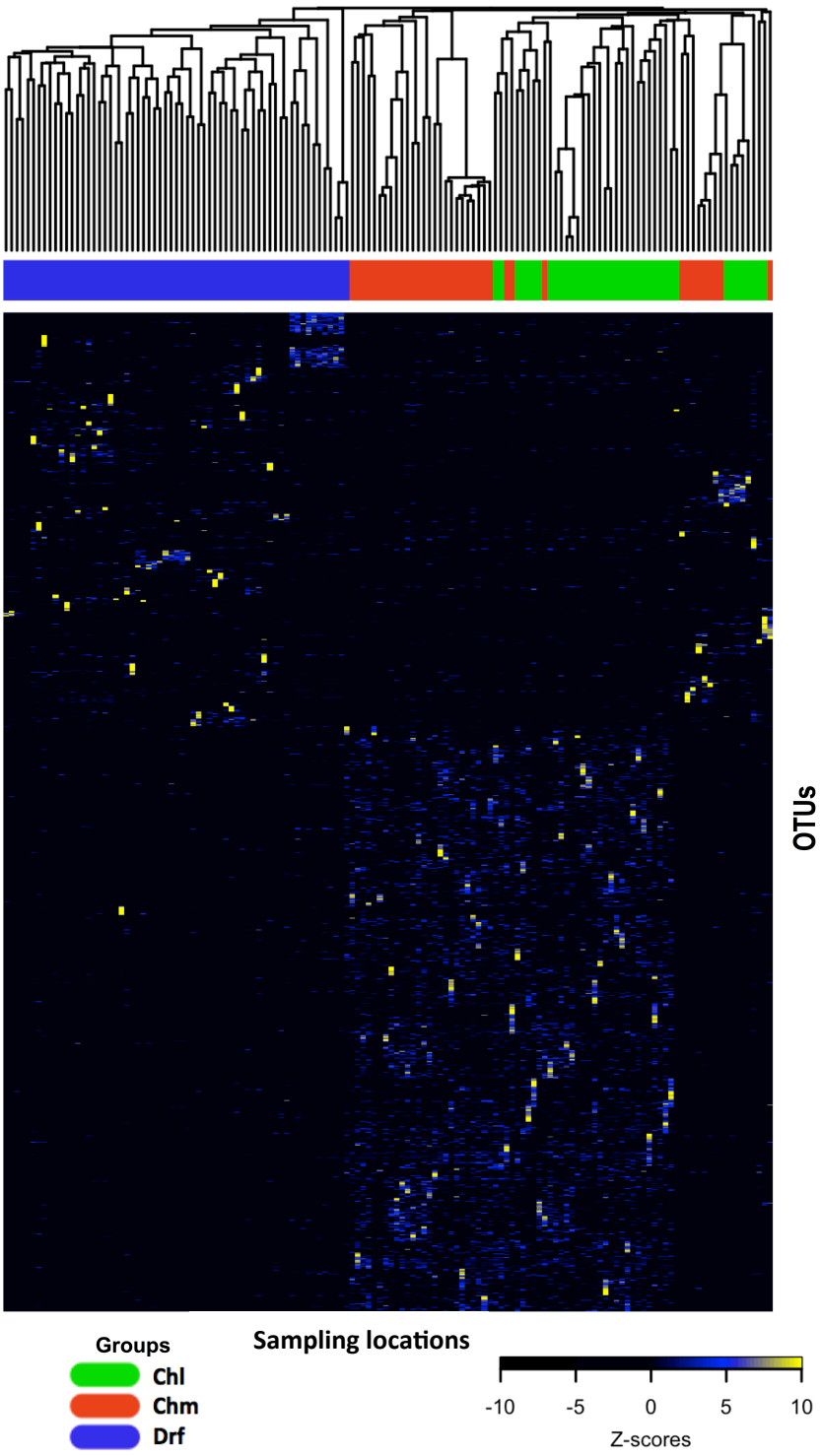

Table A1. Summary of datasets used in meta-analysis.

No	DOI number for paper	Publication year	Disinfectant residual type	Country of sampling	Sampling points	Sequencing instrument	Hypervariable region amplified	Number of treatment plants	Average % of matches in SILVA database
1	10.1016/j.watres.2015.12.010	2015	Chlorinated	UK	POU	Illumina Miseq	V4	1	89.02% ± 3.79%
2	doi: 10.1016/j.watres.2014.01.049	2014	Chlorinated	UK	DWDS	Roche 454 FLX	V1-V3	1	71.15% ± 6.22%
3	10.1016/j.watres.2013.11.027	2014	Chlorinated/Chloraminated	USA	POU	Roche 454 FLX Titanium	V1-V2	15	79.69% ± 12.42%
4	10.1016/j.ecoenv.2014.07.029	2014	Chlorinated	China	DWTP outlet, POU	Roche 454 FLX Titanium	V3-V4	1*	61.61% ± 12.48%
5	10.1128/AEM.01892-12	2012	Chlorinated/Chloraminated	USA	DWDS	Roche 454 FLX Titanium	V4-V5	1	89.56% ± 3.88%
6	10.1371/journal.pone.0141087	2015	Chlorinated/Chloraminated	USA	DWTP outlet, POU	Illimina Miseq	V4	5	67.2% ± 8.08%
7	10.1021/acs.est.5b03521	2015	Chlorinated	China	DWTP outlet, POU	Roche 454 FLX Titanium	V1-V3	1*	85.28% ± 8.29%
8	10.1016/j.watres.2013.10.071	2013	Chlorinated	China	DWTP outlet	Roche 454 FLX	V1-V3	1	96.92%
9	10.1021/es5009467	2014	No disinfectant residual	Netherlands	POU	Roche 454 FLX Titanium	V4-V6	1**	63.61% ± 4.24%
10	10.1128/mBio.01135-14	2014	Chloraminated	USA	DWTP outlet, POU	Roche 454 GS-FLX	V4-V5	1	69.98% ± 21.13%
11	10.1111/1462-2920.12739	2015	No disinfectant residual	Netherlands	DWDS	Illumina MiSeq/Roche 454 GS-FLX	V4 and V5-V6	32**	45.22% ± 6.65%
12	10.1021/es502646d	2014	Chlorinated/Chloraminated	USA	DWDS	Illumina MiSeq	V4	1	68.93% ± 2.65%
13	10.1007/s11274-013-1321-5	2013	Chloraminated	China	DWTP outlet, POU	Roche 454 GS-FLX	V3-V5	1	29.91% ± 0.64%
14	10.1128/AEM.01297-15	2015	Chloraminated	Australia	DWTP outlet, DWDS	Ion Torrent	V3	2	90.24% ± 6.82%

*Both studies sampled at the same treatment plant
**Plant from study No.9 may have been sampled in study No. 11

Table A2. Summary of the mean relative abundance (MRA- %) and occurrence (Freq) most commonly occurring bacterial OTUs in across chlorinated, chloraminated, and disinfectant residual-free drinking water systems.

OTU number	Classification	Common in	Chlorinated systems		Chloraminated systems		Disinfectant residual-free systems	
			MRA(stdev)	Freq	MRA(stdev)	Freq	MRA(stdev)	Freq
4	*Porphyrobacter*	Chlorinated/ Chloraminated	9.85(22.15)	0.62	2.26(3.75)	0.50	0.02(0.13)	0.02
6	*Mycobacterium*	Chlorinated	8.62(31.19)	0.54	2.26(9.33)	0.26	0.02(0.13)	0.02
12	*Sphingomonas*	Chlorinated	9.23(20.29)	0.51	0.24(0.54)	0.18	0.05(0.38)	0.02
15	*Vampirovibrio*	Chlorinated	15.54(29.82)	0.51	1.47(2.6)	0.45	0.14(0.4)	0.13
30	*Bosea*	Chloraminated	1.51(4.64)	0.23	11.55(45.69)	0.53	0.02(0.13)	0.02
94	*Nitrospira*	Non-disinfected	0(0)	0.00	0(0)	0.00	11(14.14)	0.86
162	*Parcubacteria*	Non-disinfected	0(0)	0.00	0(0)	0.00	6.25(10.85)	0.71
189	*Bdellovibrio*	Non-disinfected	0.46(2.21)	0.05	0.47(0.98)	0.26	2.86(4.26)	0.68
167	*Parcubacteria*	Non-disinfected	0(0)	0.00	0.16(0.44)	0.13	5.86(9.74)	0.67
59	*Sideroxydans*	Non-disinfected	0(0)	0.00	0(0)	0.00	14.57(41.29)	0.67
265	*Nitrospira*	Non-disinfected	0(0)	0.00	0(0)	0.00	2.21(2.82)	0.67

Appendix B

Figure B1. (A) Correlations between the number of observed OTUs (Sobs) and water quality parameters.

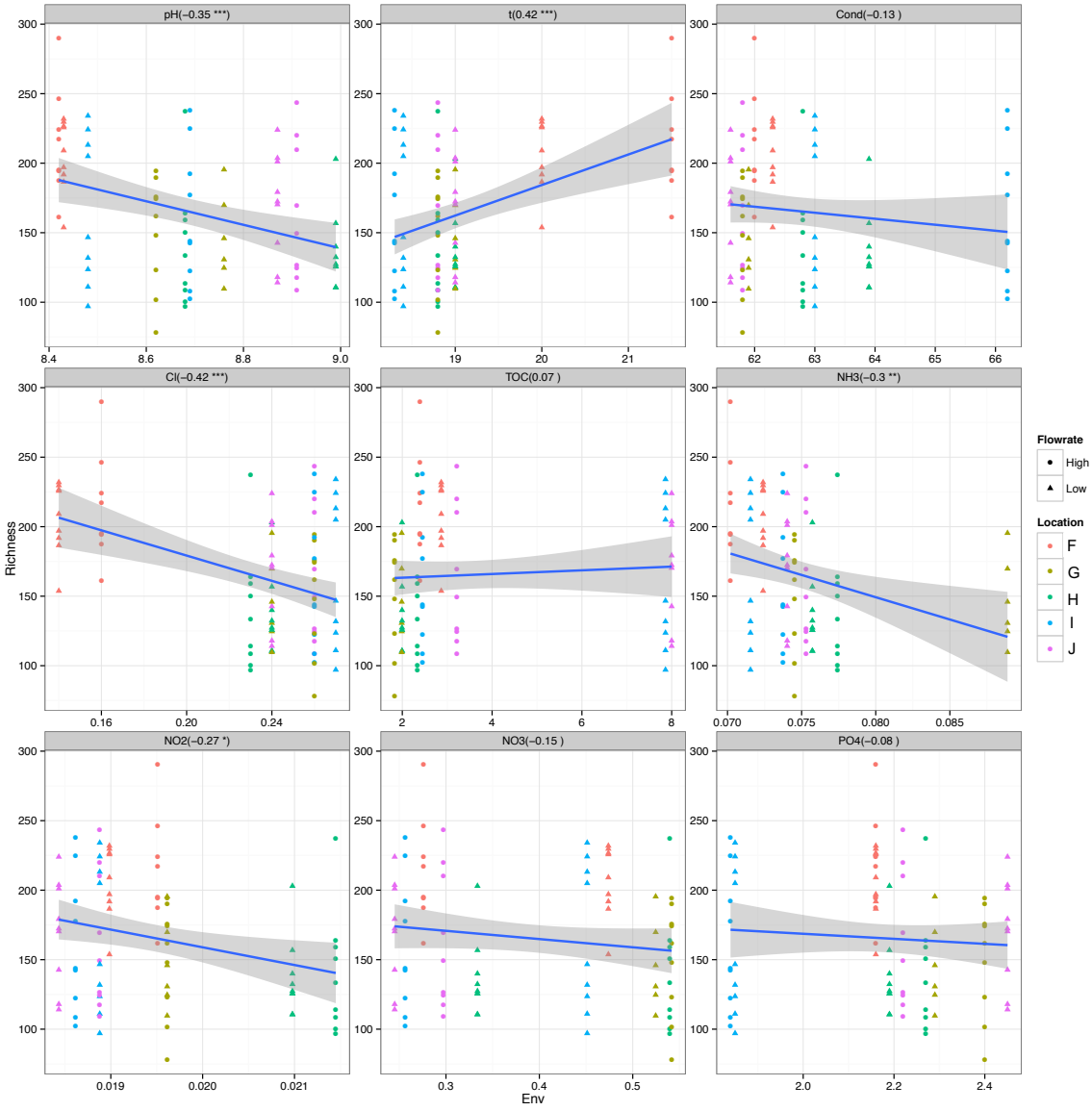

Figure B1. (B) Correlations between the Shannon index and water quality parameters.

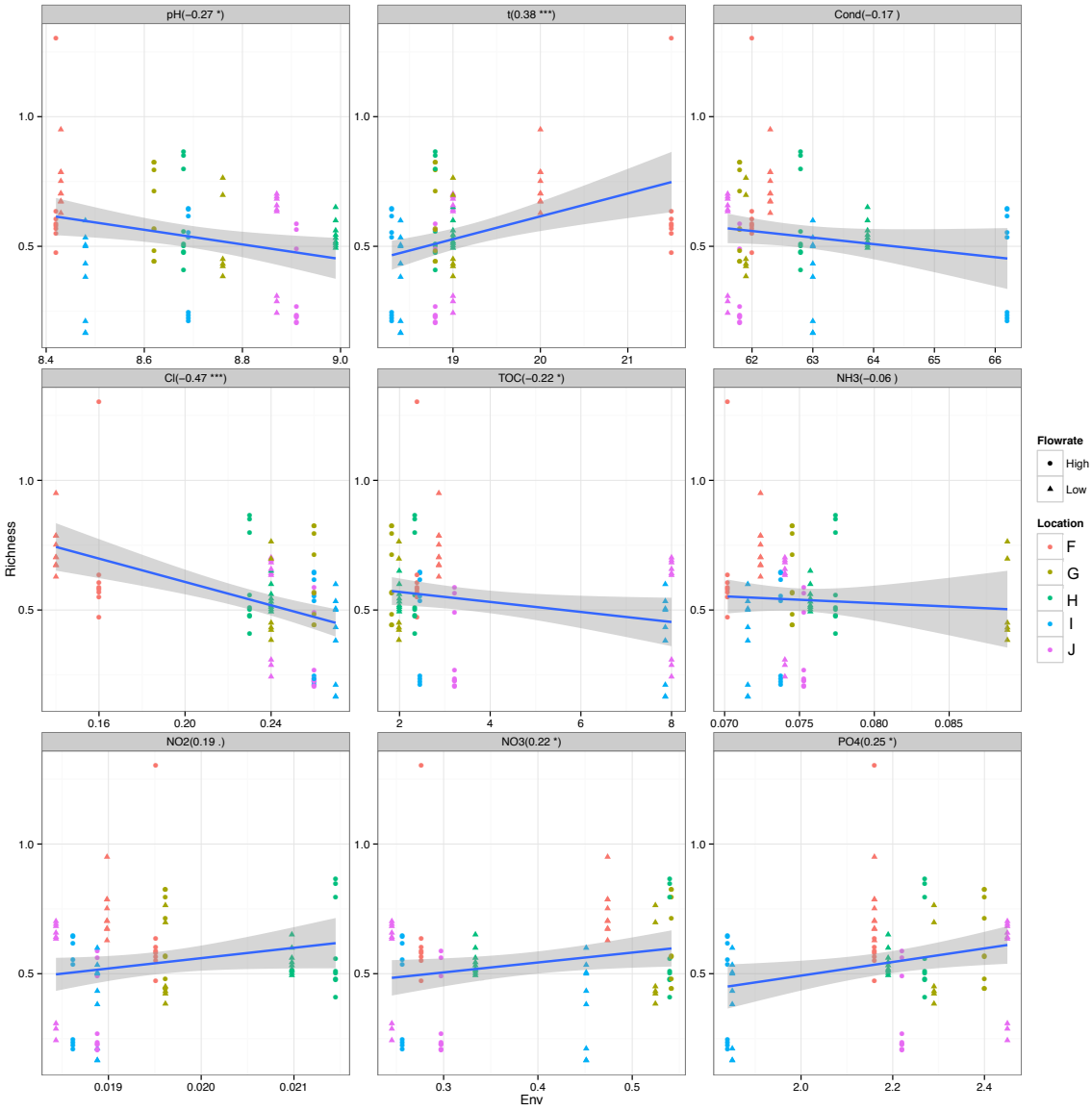

Figure B1. (C) Correlations between the Invsimpson index and water quality parameters.

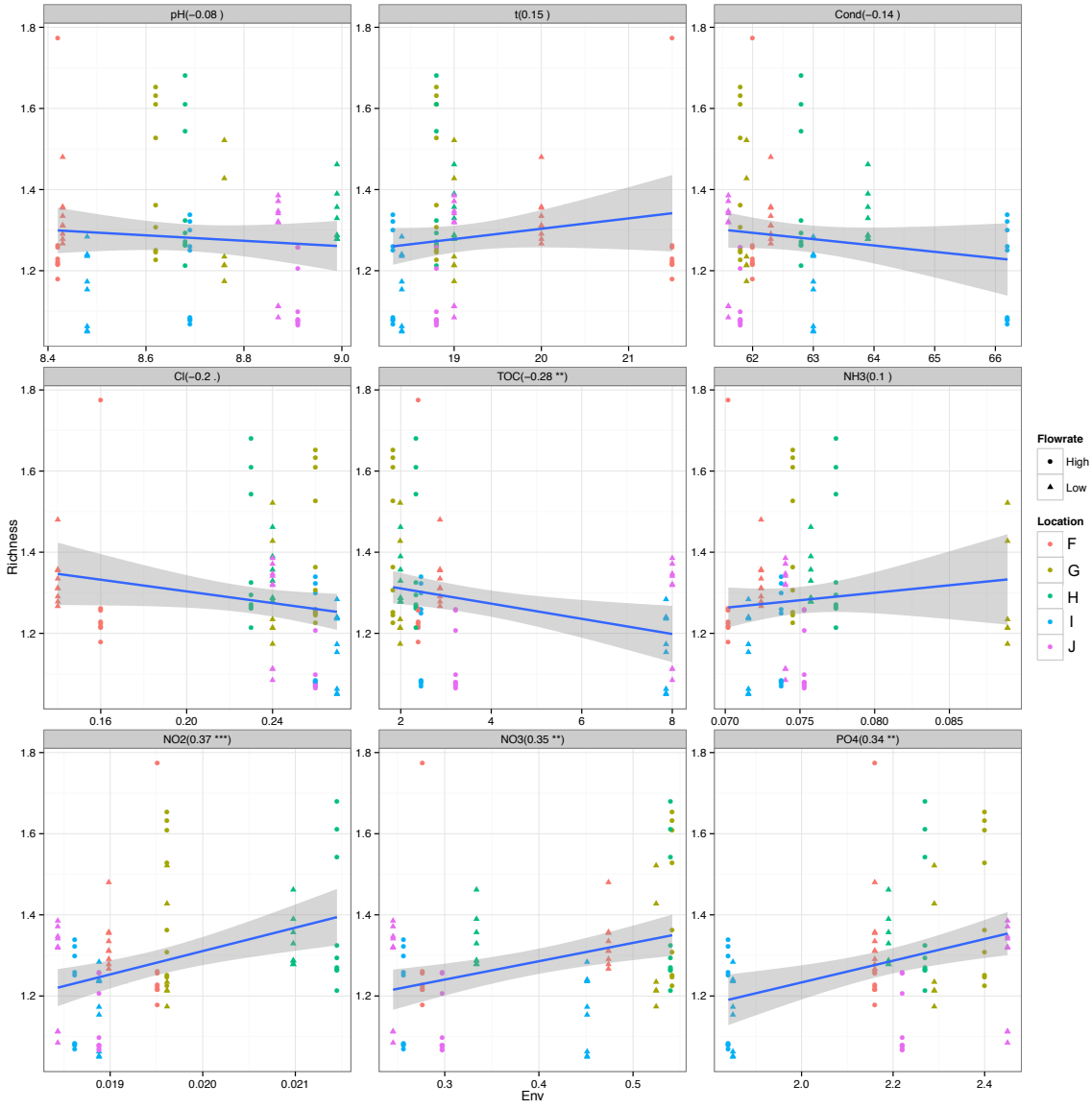

Table B1. Estimations of shear stress and Reynolds number for commercially available pipe diameters used in drinking water premise plumbing in the UK.

Q = flowrate
Vel = velocity
DV = dynamic viscosity of water at 20 degrees Celsius

SS = shear stress
SR = shear rate
r = radius

Re = Reynolds number
D = diameter
De = fluid density

$$SR = \frac{4 * Q}{Pi * r^3} \qquad SS = SR * DV \qquad Re = \frac{De*Vel*D}{DV} \qquad Vel = \frac{Q}{Pi*r^2}$$

		Qlow	0.00000833 m^3/s
Pi	3.1416	Qhigh	0.0000833 m^3/s
DV	0.001 $(N/m^2)*s$	Qlow	0.5 l/min
De	998.4 kg/m^3 (at 19 degrees Celsius)	Qhigh	5 l/min

Pipe dimmensions	d (mm)	12	15	22	28	35	50.8
	d (m)	0.012	0.015	0.022	0.028	0.035	0.0508
	r (m)	0.0060	0.0075	0.011	0.014	0.018	0.0254
Vel (m/s)	Qlow	0.074	0.047	0.022	0.014	0.009	0.004
	Qhigh	0.737	0.471	0.219	0.135	0.087	0.041
Shear rate (s^{-1})	Qlow	49.10	25.14	7.97	3.87	1.98	0.65
	Qhigh	491.02	251.40	79.68	38.65	19.79	6.47
Shear stress (N/m2)	Qlow	0.04910	0.02514	0.00797	0.00387	0.00198	0.00065
	Qhigh	0.49102	0.25140	0.07968	0.03865	0.01979	0.00647
Reynolds number	Qlow	882.4	705.9	481.3	378.2	302.5	208.4
	Qhigh	8824.2	7059.4	4813.2	3781.8	3025.5	2084.5

Table B2. Water quality parameters of sampling locations A-E.

Sampling location	Time period	Temperature (°C)	pH	Dissolved oxygen (mg/l)	Conductivity (µS/cm)	Turbidity (NTU)	Ammonia (µg-N/l)	Total organic carbon (mg-C/l)	Total chlorine (mg Cl_2/l)
A	08:00-12:00	18.7 (0.1)	9.01 (0.48)	8.74 (0.04)	60.60 (0.41)	0.26 (0.06)	7.2 (4.7)	2.01	0.36 (0.03)
A	12:00-16:00	18.7 (0.1)	8.76 (0.05)	8.85 (0.05)	60.90 (0.24)	0.23 (0.1)	15.5 (5.2)	1.48	0.39 (0.01)
A	16:00-20:00	18.7 (0.1)	8.85 (0.03)	8.87 (0.03)	60.88 (0.17)	0.25 (0.04)	7.4 (2.1)	0.90	0.39 (0.01)
A	20:00-00:00	18.7 (0.1)	8.81 (0.06)	8.86 (0.03)	60.90 (0.10)	0.175 (0.06)	5.9 (4/3)	0.86	0.40 (0.02)
A	00:00-04:00	18.5 (0.0)	8.84	8.94 (0.0)	60.30	0.17 (0.06)	3.0 (0.7)	0.66	0.32 (0.11)
A	04:00-08:00	-	-	-	-	-	-	1.71	0.22 (0.02)
B	08:00-12:00	19.7 (0.5)	9.15 (0.17)	9.46 (0.05)	47.47 (0.06)	0.12 (0.00)	7.4 (6.5)	1.84	0.27 (0.03)
B	12:00-16:00	20.4 (0.4)	8.92 (0.00)	9.08 (0.49)	47.80 (0.28)	0.10 (0.00)	1.0 (1.0)	1.78	0.28 (0.02)
B	16:00-20:00	20.2 (0.5)	8.92 (0.00)	9.00 (0.38)	47.27 (0.31)	0.22 (0.00)	5.8 (2.2)	2.05	0.27 (0.01)
B	20:00-00:00	19.7 (0.3)	8.95 (0.03)	9.41 (0.13)	46.30 (0.70)	0.13 (0.01)	5.7 (4.5)	1.69	0.28 (0.01)
B	00:00-04:00	19.7 (0.3)	9.42 (0.78)	9.43 (0.06)	45.25 (0.49)	0.12 (0.00)	11.0 (1.0)	1.79	0.18 (0.11)
B	04:00-08:00	19.4 (0.8)	9.01 (0.20)	9.40 (0.03)	46.25 (1.91)	0.13 (0.01)	5.9 (1.9)	1.84	0.18 (0.11)
C	08:00-12:00	16.2 (0.1)	8.60 (0.13)	9.74 (0.01)	66.63 (0.13)	0.13 (0.01)	6.7 (7.8)	1.84	0.23 (0.01)
C	12:00-16:00	16.3 (0.1)	8.49 (0.18)	9.73 (0.02)	66.65 (0.06)	0.12 (0.01)	33.0 (32.0)	1.87	0.24 (0.01)
C	16:00-20:00	16.3 (0.2)	8.48 (0.07)	9.69 (0.01)	66.55 (0.10)	0.13 (0.01)	42.0 (49.0)	1.82	0.25 (0.01)
C	20:00-00:00	16.3 (0.1)	8.36 (0.15)	9.64 (0.24)	66.73 (0.30)	0.15 (0.01)	33.0 (-)	1.80	0.25 (0.01)
C	00:00-04:00	16.2 (0.1)	8.34 (0.03)	9.75 (0.01)	66.80 (0.14)	0.12 (0.01)	-	1.97	0.25 (0.01)
C	04:00-08:00	16.0 (0.1)	8.51 (0.17)	9.74 (0.01)	66.87 (0.06)	0.13 (0.01)	45.0 (-)	2.21	0.24 (0.01)
D	08:00-12:00	18.5 (0.1)	8.70 (0.05)	9.21 (0.48)	60.63 (0.05)	0.15 (0.04)	7.6 (4.8)	1.95	0.34 (0.02)
D	12:00-16:00	18.5 (0.4)	8.64 (0.05)	9.61 (0.04)	60.43 (0.21)	0.12 (0.00)	15.0 (-)	1.60	0.36 (0.01)
D	16:00-20:00	18.6 (0.6)	8.65 (0.04)	9.67 (0.047)	60.33 (0.15)	0.16 (0.02)	13.0 (-)	1.92	0.36 (0.01)
D	20:00-00:00	17.9 (0.7)	8.66 (0.08)	9.65 (0.04)	60.07 (0.12)	0.17 (0.04)	16.0 (0.0)	1.87	0.34 (0.02)
D	00:00-04:00	17.9 (0.1)	8.50 (0.19)	9.50 (0.26)	60.23 (0.21)	0.11 (0.01)	6.7 (3.0)	1.54	0.33 (0.01)
D	04:00-08:00	17.5 (0.2)	8.50 (0.14)	9.62 (0.07)	60.45 (0.07)	0.15 (0.01)	7.16.0)	1.72	0.30 (0.03)

Table B2 (cont). Water quality parameters of sampling locations A-E.

Sampling location	Time period	Temperature (°C)	pH	Dissolved oxygen (mg/l)	Conductivity (µS/cm)	Turbidity (NTU)	Ammonia (µg-N/l)	Total organic carbon (mg-C/l)	Total chlorine (mg Cl_2/l)
E	08:00-12:00	18.8 (0.1)	8.54 (0.02)	9.46 (0.73)	60.84 (0.29)	0.29 (0.06)	11.0 (5.2)	2.15	0.35 (0.01)
E	12:00-16:00	19.0 (0.0)	8.55 (0.00)	9.70 (0.00)	60.30 (0.00)	0.18 (0.00)	4.6 (4.0)	1.99	0.35 (0.00)
E	16:00-20:00	18.9 (0.2)	8.56 (0.07)	9.85 (0.04)	60.63 (0.23)	0.18 (0.01)	5.7 (8.8)	1.85	0.35 (0.00)
E	20:00-00:00	18.8 (0.0)	8.60 (0.00)	9.83 (0.00)	60.50 (0.00)	0.15 (0.01)	4.2 (2.6)	1.79	0.33 (0.03)
E	00:00-04:00	18.8 (0.0)	8.60 (0.00)	9.83 (0.00)	60.50 (0.00)	0.13 (0.02)	3.1 (1.7)	1.73	0.30 (0.01)
E	04:00-08:00	18.9 (0.2)	8.57 (0.04)	9.69 (0.21)	60.55 (0.07)	0.15 (0.01)	4.7 (2.9)	1.86	0.31 (0.03)

Table B3. Water quality parameters of sampling locations F-J.

Sample Name	Location	Water Flow Rate Low/High	Volume of water filtered (l)	DNA conc (ng/ul)	pH	t (°C)	Cond. (uS/cm)	Cl2 (mg/L)	TOC (mg/L)	NH3-N (mg/L)	NO2-N (mg/L)	NO3-N (mg/L)	PO4 3- (mg/L)
1	A	L	10	4.30	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
28	A	L	10	4.30	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
57	A	L	10	4.30	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
2	A	L	15	7.09	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
29	A	L	15	7.09	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
58	A	L	15	7.09	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
3	A	L	20	14.90	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
4	A	L	20	14.90	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
30	A	L	20	14.90	8.43	20.0	62.3	0.14	2.87	0.07	0.02	0.5	2.16
5	A	H	10	0.70	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
31	A	H	10	0.70	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
32	A	H	15	6.14	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
34	A	H	15	6.14	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
83	A	H	15	6.14	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
6	A	H	20	24.80	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
33	A	H	20	24.80	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
35	A	H	20	24.80	8.42	21.5	62.0	0.16	2.39	0.07	0.02	0.3	2.16
7	B	L	15	9.03	8.76	19.0	61.9	0.24	1.99	0.09	0.02	0.5	2.29
45	B	L	15	9.03	8.76	19.0	61.9	0.24	1.99	0.09	0.02	0.5	2.29
59	B	L	15	9.03	8.76	19.0	61.9	0.24	1.99	0.09	0.02	0.5	2.29
8	B	L	20	57.00	8.76	19.0	61.9	0.24	1.99	0.09	0.02	0.5	2.29
38	B	L	20	57.00	8.76	19.0	61.9	0.24	1.99	0.09	0.02	0.5	2.29
60	B	L	20	57.00	8.76	19.0	61.9	0.24	1.99	0.09	0.02	0.5	2.29
27	B	H	10	0.56	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
37	B	H	10	0.56	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
61	B	H	10	0.56	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
9	B	H	15	8.89	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
36	B	H	15	8.89	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
62	B	H	15	8.89	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
10	B	H	20	12.80	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
63	B	H	20	12.80	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
84	B	H	20	12.80	8.62	18.8	61.8	0.26	1.83	0.07	0.02	0.5	2.40
11	C	L	10	10.40	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
39	C	L	10	10.40	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
64	C	L	10	10.40	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
12	C	L	15	23.50	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
40	C	L	15	23.50	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
65	C	L	15	23.50	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19

Table B3 (cont). Water quality parameters of sampling locations F-J.

Sample Name	Location	Water Flow Rate Low/High	Volume of water filtered (l)	DNA conc (ng/ul)	pH	t (°C)	Cond. (uS/cm)	Cl2 (mg/L)	TOC (mg/L)	NH3-N (mg/L)	NO2-N (mg/L)	NO3-N (mg/L)	PO4 3- (mg/L)
41	C	L	20	30.20	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
66	C	L	20	30.20	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
85	C	L	20	30.20	8.99	19.0	63.9	0.24	1.99	0.08	0.02	0.3	2.19
13	C	H	10	2.46	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
42	C	H	10	2.46	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
67	C	H	10	2.46	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
14	C	H	15	23.60	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
43	C	H	15	23.60	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
68	C	H	15	23.60	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
15	C	H	20	47.50	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
44	C	H	20	47.50	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
69	C	H	20	47.50	8.68	18.8	62.8	0.23	2.34	0.08	0.02	0.5	2.27
16	D	L	10	9.32	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
46	D	L	10	9.32	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
70	D	L	10	9.32	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
17	D	L	15	17.00	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
47	D	L	15	17.00	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
71	D	L	15	17.00	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
18	D	L	20	45.40	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
48	D	L	20	45.40	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
72	D	L	20	45.40	8.48	18.4	63.0	0.27	7.86	0.07	0.02	0.5	1.85
19	D	H	10	7.13	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
73	D	H	10	7.13	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
86	D	H	10	7.13	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
20	D	H	15	11.00	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
49	D	H	15	11.00	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
74	D	H	15	11.00	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
21	D	H	20	12.60	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
50	D	H	20	12.60	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
75	D	H	20	12.60	8.69	18.3	66.2	0.26	2.45	0.07	0.02	0.3	1.84
22	E	L	10	4.32	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
51	E	L	10	4.32	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
76	E	L	10	4.32	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
23	E	L	15	13.10	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
52	E	L	15	13.10	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
77	E	L	15	13.10	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
53	E	L	20	29.90	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
78	E	L	20	29.90	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45

Table B3 (cont). Water quality parameters of sampling locations F-J.

Sample Name	Location	Water Flow Rate Low/High	Volume of water filtered (l)	DNA conc (ng/ul)	pH	t (°C)	Cond. (uS/cm)	Cl2 (mg/L)	TOC (mg/L)	NH3-N (mg/L)	NO2-N (mg/L)	NO3-N (mg/L)	PO4 3- (mg/L)
79	E	L	20	29.90	8.87	19.0	61.6	0.24	8.00	0.07	0.02	0.2	2.45
24	E	H	10	6.17	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
54	E	H	10	6.17	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
80	E	H	10	6.17	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
25	E	H	15	6.39	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
55	E	H	15	6.39	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
81	E	H	15	6.39	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
26	E	H	20	4.48	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
56	E	H	20	4.48	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22
82	E	H	20	4.48	8.91	18.8	61.8	0.26	3.22	0.08	0.02	0.3	2.22

Table B4. Results from AMOVA tests comparing the variances between replicate filters from same sampling time-period and between triplicate sequencing libraries from independent barcoded PCR reactions using Bray-Curtis distance metric. NA = not applicable due to lack of sufficient replicates for AMOVA tests due to failure of samples to amplify with PCR or to sequence appropriately. Comparisons were considered significantly different at a Benjamini-Hochberg (BH) corrected significance level of 0.0006 at a false discovery rate of 0.05.

Time period	Filter comparison	p-value for pairwise comparison of filters for each sampling location				
		A	B	C	D	E
08:00-12:00	Filter 1 - Filter 2	0.401	0.019	0.213	1.000	0.287
08:00-12:00	Filter 1 – Filter 3	0.498	0.021	0.194	0.313	0.101
08:00-12:00	Filter 2 – Filter 3	0.372	0.396	1.000	0.328	0.102
12:00-16:00	Filter 1 - Filter 2	0.322	0.425	0.110	0.088	0.781
12:00-16:00	Filter 1 – Filter 3	1.000	0.102	0.386	0.105	0.188
12:00-16:00	Filter 2 – Filter 3	0.330	0.103	0.221	0.429	0.201
16:00-20:00	Filter 1 - Filter 2	0.106	0.288	0.310	0.079	1.000
16:00-20:00	Filter 1 – Filter 3	0.718	0.225	0.414	0.202	0.419
16:00-20:00	Filter 2 – Filter 3	0.290	0.785	0.088	0.015	0.778
20:00-00:00	Filter 1 - Filter 2	0.483	1.000	0.207	NA	0.093
20:00-00:00	Filter 1 – Filter 3	0.096	0.625	0.101	NA	0.106
20:00-00:00	Filter 2 – Filter 3	0.198	0.495	0.110	0.304	0.898
00:00-04:00	Filter 1 - Filter 2	0.758	0.896	0.669	1.000	0.203
00:00-04:00	Filter 1 – Filter 3	0.157	1.000	0.578	0.521	0.100
00:00-04:00	Filter 2 – Filter 3	0.900	0.810	0.254	0.420	0.095
04:00-08:00	Filter 1 - Filter 2	0.638	0.288	0.220	NA	NA
04:00-08:00	Filter 1 – Filter 3	0.107	1.000	0.911	NA	NA
04:00-08:00	Filter 2 – Filter 3	0.183	0.688	0.491	NA	NA

Table B5. MRA and detection frequency of OTUs significantly correlated ($p<0.01$) (correlation coefficient >0.50, <-0.50), sampling locations F-J.

Name	MRA (%)		Detection frequency		Taxonomy				
	Low	High	Low	High	Phylum	Class	Order	Family	Genus
OTU_1	88.9%	88.85%	1.00	1.00	*Proteobacteria*	Betaproteobacteria	*Burkholderiales*	*Comamonadaceae*	unclassified
OTU_3	3.98%	3.87%	1.00	1.00	*Proteobacteria*	Alphaproteobacteria	*Sphingomonadales*	*Sphingomonadaceae*	unclassified
OTU_4	0.653%	0.537%	1.00	1.00	*Proteobacteria*	Alphaproteobacteria	*Rhizobiales*	*Hyphomicrobiaceae*	unclassified
OTU_5	0.329%	0.425%	1.00	0.98	*Proteobacteria*	Alphaproteobacteria	unclassified	unclassified	unclassified
OTU_6	0.434%	0.237%	1.00	1.00	*Proteobacteria*	Alphaproteobacteria	*Sphingomonadales*	*Sphingomonadaceae*	*Sphingomonas*
OTU_7	0.346%	0.120%	0.98	0.91	*Proteobacteria*	Alphaproteobacteria	*Rhizobiales*	*Bradyrhizobiaceae*	unclassified
OTU_8	0.145%	0.114%	0.81	0.70	*Proteobacteria*	Betaproteobacteria	*Burkholderiales*	unclassified	unclassified
OTU_9	0.131%	0.0936%	0.55	0.50	*Proteobacteria*	unclassified	unclassified	unclassified	unclassified
OTU_10	0.0378%	0.0708%	0.67	0.61	*Proteobacteria*	*Betaproteobacteria*	unclassified	unclassified	unclassified
OTU_11	0.0401%	0.0310%	1.00	0.95	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Mycobacteriaceae*	*Mycobacterium*
OTU_12	0.0313%	0.0484%	0.83	0.91	*Proteobacteria*	*Gammaproteobacteria*	*Legionellales*	*Legionellaceae*	*Legionella*
OTU_13	0.0282%	0.0333%	0.76	0.68	*Proteobacteria*	*Gammaproteobacteria*	*Xanthomonadales*	*Sinobacteraceae*	*Nevskia*
OTU_15	0.0314%	0.0110%	0.33	0.39	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Methylobacteriaceae*	*Methylobacterium*
OTU_16	0.0239%	0.0184%	0.83	0.93	*Proteobacteria*	unclassified	unclassified	unclassified	unclassified
OTU_17	0.0131%	0.0206%	0.40	0.25	*Proteobacteria*	*Gammaproteobacteria*	*Xanthomonadales*	*Sinobacteraceae*	*Nevskia*
OTU_19	0.0160%	0.0168%	0.43	0.43	*Proteobacteria*	*Gammaproteobacteria*	*Legionellales*	*Legionellaceae*	*Legionella*
OTU_24	0.0142%	0.0126%	0.36	0.41	*Bacteroidetes*	*Sphingobacteria*	*Sphingobacteriales*	*Chitinophagaceae*	*Sediminibacterium*
OTU_29	0.00755%	0%	0.21	0.00	*Proteobacteria*	*Gammaproteobacteria*	*Legionellales*	*Legionellaceae*	*Legionella*
OTU_31	0.00541%	0.0071%	0.69	0.82	*Proteobacteria*	*Gammaproteobacteria*	*Pseudomonadales*	*Pseudomonadaceae*	*Pseudomonas*
OTU_32	0.00672%	0.0054%	0.36	0.34	*Planctomycetes*	*Planctomycetacia*	*Planctomycetales*	*Planctomycetaceae*	*Schlesneria*
OTU_37	0.00738%	0.0050%	0.55	0.55	*Proteobacteria*	unclassified	unclassified	unclassified	unclassified

Table B6. (A) Top 10 OTUs in locations A-E.

OTU	Sampling Location					Average	Standard deviation
	A	B	C	D	E		
OTU1	89.16%	95.45%	89.34%	29.77%	15.68%	63.88%	37.98%
OTU2	1.44%	0.29%	8.77%	64.72%	79.92%	31.03%	38.21%
OTU3	4.22%	1.94%	0.44%	2.16%	2.35%	2.22%	1.35%
OTU4	1.99%	0.62%	0.26%	0.12%	0.11%	0.62%	0.79%
OTU5	2.27%	0.59%	0.68%	2.14%	1.11%	1.36%	0.80%
OTU6	0.15%	0.10%	0.01%	0.18%	0.07%	0.10%	0.07%
OTU7	0.32%	0.17%	0.24%	0.22%	0.24%	0.24%	0.05%
OTU8	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.01%
OTU9	0.06%	0.14%	0.00%	0.04%	0.06%	0.06%	0.04%
OTU11	0.03%	0.18%	0.00%	0.00%	0.00%	0.04%	0.07%
OTU12	0.00%	0.00%	0.00%	0.04%	0.06%	0.02%	0.02%
OTU20	0.00%	0.00%	0.04%	0.00%	0.00%	0.01%	0.02%
OTU23	0.00%	0.00%	0.00%	0.00%	0.03%	0.01%	0.01%
OTU42	0.00%	0.08%	0.00%	0.00%	0.00%	0.02%	0.03%
OTU56	0.04%	0.00%	0.00%	0.00%	0.00%	0.01%	0.02%
OTU64	0.00%	0.00%	0.03%	0.00%	0.00%	0.01%	0.01%

Table B6. (B) Taxonomic classification of top ten OTUs in locations A-E.

OTU	phylum	class	order	family	genus
OTU1	*Proteobacteria*	*Betaproteobacteria*	*Burkholderiales*	*Comamonadaceae*	unclassified
OTU2	*Proteobacteria*	*Alphaproteobacteria*	unclassified	unclassified	unclassified
OTU3	*Proteobacteria*	*Alphaproteobacteria*	*Sphingomonadales*	*Sphingomonadaceae*	unclassified
OTU4	*Proteobacteria*	*Betaproteobacteria*	*Methylophilales*	*Methylophilaceae*	*Methylophilus*
OTU5	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Hyphomicrobiaceae*	unclassified
OTU6	*Actinobacteria*	*Actinobacteria*	*Actinomycetales*	*Mycobacteriaceae*	*Mycobacterium*
OTU7	*Proteobacteria*	*Alphaproteobacteria*	*Sphingomonadales*	*Sphingomonadaceae*	*Sphingomonas*
OTU8	*Proteobacteria*	*Alphaproteobacteria*	*Rhizobiales*	*Bradyrhizobiaceae*	unclassified
OTU9	*Proteobacteria*	*Gammaproteobacteria*	*Legionellales*	*Legionellaceae*	*Legionella*
OTU11	*Proteobacteria*	*Gammaproteobacteria*	unclassified	unclassified	unclassified
OTU12	*Proteobacteria*	*Gammaproteobacteria*	*Xanthomonadales*	*Sinobacteraceae*	*Nevskia*
OTU20	*Proteobacteria*	*Alphaproteobacteria*	unclassified	unclassified	unclassified
OTU23	*Proteobacteria*	unclassified	unclassified	unclassified	unclassified
OTU42	*Proteobacteria*	unclassified	unclassified	unclassified	unclassified
OTU56	*Proteobacteria*	*Alphaproteobacteria*	*Sphingomonadales*	*Erythrobacteraceae*	*Porphyrobacter*
OTU64	*Proteobacteria*	*Betaproteobacteria*	unclassified	unclassified	unclassified

Appendix C

Table C1. Description of treatment plants and distribution systems sampled. FN: treated water at the plant; DS: distribution system; SW: surface water; PreT SW1: pre-treated surface water 1; PreT SW2: pre-treated surface water 2; GW: groundwater; Coag: coagulation; RGF: rapid gravity filtration; pHcorr: pH correction; Ortho: orthophosphate addition; Chl: chlorination; CoCoDAFF: Counter Current Dissolved Air Flotation and Filtration; Chm: chloramination; Clar: clarification; MRF: manganese removal filtration; Aer: aeration; Ozo: ozonation; S: softening; BAC: biological activated carbon; SSF: slow sand filtration; PreF: pre-filtration; PostF: post-filtration; GAC: granular activated carbon; UV: ultraviolet treatment; VD: vacuum degassing; Dec: decolourisation.

Name	Location	Type of source	Treatment processes	Residual disinfectant	No. of sampling points (FN+DS)
A	Scotland	SW	Coag+RGF+pHcorr+Ortho+Chl	Chlorine	1+3
B	Scotland	SW	Coag+RGF+pHcorr+Ortho+Chl	Chlorine	1+3
C	Scotland	SW	Coag+RGF+pH correction+Chl	Chlorine	1+2
D	Scotland	SW	Coag+CoCoDAFF+RGF+pH correction+Ortho+Chm	Chloramine	1+3
E	Scotland	SW	Coag+Clar+RGF+pH correction+Ortho+MRF+Chm	Chloramine	1+3
F	Netherlands	PreT SW1	Aer+RGF+Ozo+S+BAC+SSF	none	1+3
G	Netherlands	PreT SW2	RGF+Ozo+S+BAC+SSF	none	1+2
H	Netherlands	GW	PreF+Aer+S+PostF+GAC+UV	none	1+2
I	Netherlands	GW	PreF+Aer+S+PostF+GAC+UV	none	1+2
J	Netherlands	GW	VD+Aer+PreF+Aer+S+PostF+Dec	none	1+3

Table C2. Results of the Drichlet Multinomial Mixtures Model applied to functional data (KO table), the six partitions obtained are indicated by colours. Sampling point: treated water at the plant (FN), distribution system (DS). Disinfection groups: chlorinated (Chl), chloraminated (Chm), disinfectant residual-free (Drf).

Sample	Plant	Point	Disinfection	Partition
Sample_41	A	FN	Chl	Partition_6
Sample_42	A	DS	Chl	Partition_4
Sample_43	A	DS	Chl	Partition_4
Sample_44	A	DS	Chl	Partition_4
Sample_46	B	FN	Chl	Partition_6
Sample_47	B	DS	Chl	Partition_4
Sample_48	B	DS	Chl	Partition_4
Sample_49	B	DS	Chl	Partition_4
Sample_51	C	FN	Chl	Partition_4
Sample_52	C	DS	Chl	Partition_4
Sample_54	C	DS	Chl	Partition_4
Sample_56	D	FN	Chm	Partition_5
Sample_57	D	DS	Chm	Partition_5
Sample_58	D	DS	Chm	Partition_5
Sample_59	D	DS	Chm	Partition_5
Sample_61	E	FN	Chm	Partition_2
Sample_62	E	DS	Chm	Partition_2
Sample_63	E	DS	Chm	Partition_2
Sample_64	E	DS	Chm	Partition_2
Sample_66	F	FN	Drf	Partition_2
Sample_67	F	DS	Drf	Partition_2
Sample_68	F	DS	Drf	Partition_4
Sample_69	F	DS	Drf	Partition_6
Sample_71	G	FN	Drf	Partition_1
Sample_72	G	DS	Drf	Partition_1
Sample_73	G	DS	Drf	Partition_4
Sample_75	H	FN	Drf	Partition_1
Sample_77	H	DS	Drf	Partition_3
Sample_78	H	DS	Drf	Partition_3
Sample_79	I	FN	Drf	Partition_3
Sample_80	I	DS	Drf	Partition_3
Sample_81	I	DS	Drf	Partition_3
Sample_83	J	FN	Drf	Partition_1
Sample_84	J	DS	Drf	Partition_1
Sample_85	J	DS	Drf	Partition_1
Sample_86	J	DS	Drf	Partition_1

Table C3. Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K08162	18.36691442	-4.677334045	0.603373614	-7.751969814	9.05E-15	2.53E-12	mdtH; MFS transporter, DHA1 family, multidrug resistance
K06879	20.69147606	-4.670441647	0.602666633	-7.749627066	9.22E-15	2.53E-12	queF; 7-cyano-7-deazaguanine reductase [EC:1.7.1.13]
K05895	29.33034661	-4.66667417	0.573408248	-8.138484553	4.00E-16	3.73E-13	cobK-cbiJ; precorrin-6A/cobalt-precorrin-6A reductase
K02228	29.56538687	-4.58464955	0.587567898	-7.802757032	6.06E-15	1.88E-12	cobF; precorrin-6A synthase [EC:2.1.1.152]
K06205	17.76738586	-4.471026921	0.674206144	-6.631542832	3.32E-11	2.92E-09	mioC; MioC protein
K01185	107.9053666	-4.400873934	0.469381359	-9.375902657	6.86E-21	3.20E-17	lysozyme
K01822	24.28232269	-4.366366825	0.549415513	-7.947294391	1.91E-15	8.85E-13	E5.3.3.1; steroid Delta-isomerase [EC:5.3.3.1]
K10004	16.44946617	-4.310939717	0.638180053	-6.755052427	1.43E-11	1.39E-09	gltL, aatP; glutamate/aspartate transport system ATP-
K07123	15.72615657	-4.310931885	0.669085802	-6.443018031	1.17E-10	7.52E-09	uncharacterized protein
K10002	15.70083903	-4.290671546	0.669734608	-6.406525053	1.49E-10	8.68E-09	gltK, aatM; glutamate/aspartate transport system permease
K06909	18.094601	-4.280363399	0.658170213	-6.50342923	7.85E-11	5.72E-09	xtmB; phage terminase large subunit
K07741	17.54788053	-4.259779818	0.680702227	-6.257919611	3.90E-10	1.88E-08	antB; anti-repressor protein
K07518	14.43844635	-4.219108561	0.691633668	-6.100207025	1.06E-09	4.30E-08	E3.1.1.22; hydroxybutyrate-dimer hydrolase [EC:3.1.1.22]
K00090	20.61270985	-4.183550816	0.529673633	-7.898355803	2.83E-15	1.06E-12	E1.1.1.215; gluconate 2-dehydrogenase [EC:1.1.1.215]
K11811	18.65179799	-4.160915889	0.585223458	-7.109960868	1.16E-12	1.46E-10	arsH; arsenical resistance protein ArsH
K02609	17.40242295	-4.151746521	0.611604382	-6.788287726	1.13E-11	1.15E-09	paaA; ring-1,2-phenylacetyl-CoA epoxidase subunit PaaA
K04340	13.57439573	-4.136140351	0.690607738	-5.989131196	2.11E-09	7.68E-08	strB1; scyllo-inosamine-4-phosphate amidinotransferase 1
K03126	12.9357457	-4.075354956	0.687276361	-5.929717927	3.03E-09	1.03E-07	TAF12; transcription initiation factor TFIID subunit 12
K00472	15.76071689	-4.06972533	0.660512134	-6.161469444	7.21E-10	3.17E-08	P4HA; prolyl 4-hydroxylase [EC:1.14.11.2]
K09966	23.87372377	-4.050684101	0.468166244	-8.652234447	5.05E-18	6.04E-15	uncharacterized protein
K01143	12.57013541	-4.006463454	0.645524459	-6.206524634	5.42E-10	2.45E-08	E3.1.11.3; exodeoxyribonuclease (lambda-induced)
K02463	13.43201283	-3.964471827	0.615392015	-6.442189253	1.18E-10	7.52E-09	gspN; general secretion pathway protein N
K02610	16.73015514	-3.910848307	0.612262937	-6.3875307	1.69E-10	9.45E-09	paaB; ring-1,2-phenylacetyl-CoA epoxidase subunit PaaB
K02553	24.80959641	-3.91023375	0.4520827	-8.6493771	5.18E-18	6.04E-15	rraA, menG; regulator of ribonuclease activity A
K02611	17.33287828	-3.898768506	0.607874165	-6.413775634	1.42E-10	8.68E-09	paaC; ring-1,2-phenylacetyl-CoA epoxidase subunit PaaC
K06876	27.9064258	-3.867379855	0.486921925	-7.94250506	1.98E-15	8.85E-13	deoxyribodipyrimidine photolyase-related protein
K02612	16.86174892	-3.857202825	0.61000672	-6.323213657	2.56E-10	1.34E-08	paaD; ring-1,2-phenylacetyl-CoA epoxidase subunit PaaD
K07274	13.5892013	-3.838257266	0.66100678	-5.806683653	6.37E-09	1.83E-07	mipA, ompV; MipA family protein
K06191	12.80252161	-3.81148609	0.623505998	-6.112990258	9.78E-10	4.07E-08	nrdH; glutaredoxin-like protein NrdH
K07101	14.1470132	-3.79002344	0.633133081	-5.986140279	2.15E-09	7.77E-08	uncharacterized protein
K00116	29.03017387	-3.787855082	0.431576427	-8.776788647	1.68E-18	3.92E-15	mqo; malate dehydrogenase (quinone) [EC:1.1.5.4]
K06183	20.93375436	-3.782510288	0.632366059	-5.981520089	2.21E-09	7.93E-08	rsuA; 16S rRNA pseudouridine516 synthase [EC:5.4.99.19]

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K12544	20.18242759	-3.69727258	0.536272401	-6.894392802	5.41E-12	6.00E-10	rsaA; S-layer protein
K00155	49.06665386	-3.679387927	0.48401854	-7.601749987	2.92E-14	5.92E-12	Lack_5'-end
K12954	11.44062533	-3.648919428	0.655096175	-5.570051497	2.55E-08	5.52E-07	ctpG; cation-transporting ATPase G [EC:3.6.3.-]
K01344	22.95592963	-3.639558959	0.691098855	-5.266336259	1.39E-07	2.18E-06	PROC; protein C (activated) [EC:3.4.21.69]
K03006	9.099330505	-3.639262136	0.632593177	-5.752926629	8.77E-09	2.35E-07	RPB1, POLR2A; DNA-directed RNA polymerase II subunit
K00635	60.85305388	-3.624957244	0.47493297	-7.632566014	2.30E-14	5.11E-12	E2.3.1.20; diacylglycerol O-acyltransferase [EC:2.3.1.20]
K07644	16.45402066	-3.610782224	0.637214588	-5.666509037	1.46E-08	3.50E-07	cusS, copS, silS; two-component system, OmpR family, heavy metal sensor histidine kinase CusS [EC:2.7.13.3]
K11711	15.89407651	-3.580468266	0.612792778	-5.842869552	5.13E-09	1.55E-07	dctS; two-component system, LuxR family, sensor histidine kinase DctS [EC:2.7.13.3]
K14317	14.08581609	-3.570219665	0.648519446	-5.505185213	3.69E-08	7.64E-07	NUP214, CAN; nuclear pore complex protein Nup214
K12423	8.738900885	-3.567494593	0.603883582	-5.907586661	3.47E-09	1.15E-07	fadD21; fatty acid CoA ligase FadD21
K04073	13.06587967	-3.563691952	0.542521195	-6.568760789	5.07E-11	3.94E-09	mhpF; acetaldehyde dehydrogenase [EC:1.2.1.10]
K09861	22.76992077	-3.559039482	0.488946167	-7.279000677	3.36E-13	5.23E-11	uncharacterized protein
K07346	14.11679729	-3.542070868	0.635501309	-5.573664156	2.49E-08	5.44E-07	fimC; fimbrial chaperone protein
K03710	79.40034323	-3.532626954	0.436311923	-8.096562968	5.65E-16	4.39E-13	GntR family transcriptional regulator
K11923	20.0701963	-3.515802111	0.619321691	-5.676859318	1.37E-08	3.35E-07	cueR; MerR family transcriptional regulator, copper efflux
K06443	8.105679756	-3.514569885	0.503981968	-6.973602452	3.09E-12	3.60E-10	lcyB, crtL1, crtY; lycopene beta-cyclase [EC:5.5.1.19]
K11745	14.50055872	-3.499564831	0.623159453	-5.61584168	1.96E-08	4.47E-07	kefC; glutathione-regulated potassium-efflux system
K01563	48.48375063	-3.448027921	0.478806965	-7.201290235	5.96E-13	8.69E-11	dhaA; haloalkane dehalogenase [EC:3.8.1.5]
K10674	17.61342714	-3.445416966	0.530688768	-6.492349512	8.45E-11	5.88E-09	ectD; ectoine hydroxylase [EC:1.14.11.-]
K01725	22.76460006	-3.415286419	0.540094963	-6.323492447	2.56E-10	1.34E-08	cynS; cyanate lyase [EC:4.2.1.104]
K07662	23.56645559	-3.404812543	0.545749837	-6.238778852	4.41E-10	2.07E-08	cpxR; two-component system, OmpR family, response
K13483	23.2254027	-3.403955522	0.532188427	-6.396147203	1.59E-10	9.17E-09	yagT; xanthine dehydrogenase YagT iron-sulfur-binding
K11475	19.68965554	-3.400286951	0.540808749	-6.287411132	3.23E-10	1.62E-08	vanR; GntR family transcriptional regulator, vanillate catabolism transcriptional regulator
K09952	20.41657962	-3.398091774	0.612459132	-5.548275138	2.89E-08	6.22E-07	csn1, cas9; CRISPR-associated endonuclease Csn1
K00459	104.3717545	-3.395009547	0.443068816	-7.662488137	1.82E-14	4.67E-12	ncd2, npd; nitronate monooxygenase [EC:1.13.12.16]
K02451	12.7591814	-3.393700972	0.644759676	-5.263513055	1.41E-07	2.20E-06	gspB; general secretion pathway protein B
K01567	186.9934837	-3.387219283	0.443279244	-7.641276532	2.15E-14	5.01E-12	pdaA; peptidoglycan-N-acetylmuramic acid deacetylase
K06223	60.54937777	-3.384647192	0.460674872	-7.347149585	2.02E-13	3.37E-11	dam; DNA adenine methylase [EC:2.1.1.72]
K07455	20.80354188	-3.379173428	0.548795914	-6.157431824	7.39E-10	3.19E-08	recT; recombination protein RecT

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K11082	7.574846636	-3.37724932	0.595217524	-5.673974952	1.40E-08	3.37E-07	phnV; 2-aminoethylphosphonate transport system
K08683	45.32512346	-3.3555584	0.422830424	-7.935943607	2.09E-15	8.85E-13	HSD17B10; 3-hydroxyacyl-CoA dehydrogenase / 3-hydroxy-2-methylbutyryl-CoA dehydrogenase [EC:1.1.1.35 1.1.1.178]
K07124	118.4684583	-3.354930072	0.48140516	-6.969036377	3.19E-12	3.63E-10	uncharacterized protein
K07222	35.12137941	-3.341683751	0.461056413	-7.24788476	4.23E-13	6.37E-11	putative flavoprotein involved in K+ transport
K06919	101.0758704	-3.327679821	0.528421751	-6.297393725	3.03E-10	1.55E-08	putative DNA primase/helicase
K02361	7.880415635	-3.322938064	0.587378736	-5.657232481	1.54E-08	3.62E-07	entC; isochorismate synthase [EC:5.4.4.2]
K06518	15.47341654	-3.319155592	0.622304612	-5.333650964	9.63E-08	1.64E-06	cidA; holin-like protein
K12618	7.101704814	-3.316121539	0.62256785	-5.326522301	1.00E-07	1.69E-06	XRN1, SEP1, KEM1; 5'-3' exoribonuclease 1 [EC:3.1.13.-]
K01502	16.35653874	-3.314382391	0.617819587	-5.364644408	8.11E-08	1.45E-06	E3.5.5.7; aliphatic nitrilase [EC:3.5.5.7]
K02229	30.97822914	-3.308352999	0.584632127	-5.658862808	1.52E-08	3.61E-07	cobG; precorrin-3B synthase [EC:1.14.13.83]
K01666	13.38105718	-3.307075535	0.528018856	-6.26317696	3.77E-10	1.83E-08	mhpE; 4-hydroxy 2-oxovalerate aldolase [EC:4.1.3.39]
K09971	18.5025076	-3.29986899	0.514941132	-6.408245104	1.47E-10	8.68E-09	aapM, bztC; general L-amino acid transport system
K11387	8.637377893	-3.296428714	0.615116935	-5.359027733	8.37E-08	1.48E-06	embC; arabinosyltransferase C [EC:2.4.2.-]
K07093	55.24986264	-3.289852498	0.539811776	-6.094443742	1.10E-09	4.41E-08	uncharacterized protein
K05817	23.67399122	-3.286749302	0.498530773	-6.59287146	4.31E-11	3.72E-09	hcaR; LysR family transcriptional regulator, hca operon
K11177	23.13125451	-3.27615438	0.546308799	-5.996891113	2.01E-09	7.44E-08	yagR; xanthine dehydrogenase YagR molybdenum-binding
K07503	9.822468771	-3.261496752	0.544490143	-5.990001457	2.10E-09	7.68E-08	nucS; endonuclease [EC:3.1.-.-]
K01669	91.99586615	-3.260652621	0.413118846	-7.892771426	2.96E-15	1.06E-12	phrB; deoxyribodipyrimidine photo-lyase [EC:4.1.99.3]
K00666	231.7456111	-3.252224852	0.437666415	-7.430830288	1.08E-13	1.94E-11	K00666; fatty-acyl-CoA synthase [EC:6.2.1.-]
K01644	76.80201363	-3.248130375	0.406213236	-7.996121458	1.28E-15	8.55E-13	citE; citrate lyase subunit beta / citryl-CoA lyase
K00492	138.1246263	-3.242580935	0.452934555	-7.159049581	8.12E-13	1.11E-10	see T30017 (Metagenome): GL0042842
K06857	17.00735706	-3.241083675	0.532074507	-6.091409445	1.12E-09	4.46E-08	tupC, vupC; tungstate transport system ATP-binding
K09941	17.16895173	-3.230764704	0.549947457	-5.874678867	4.24E-09	1.34E-07	uncharacterized protein
K00832	17.61929344	-3.226440518	0.576257669	-5.598954581	2.16E-08	4.83E-07	tyrB; aromatic-amino-acid transaminase [EC:2.6.1.57]
K06609	8.713506343	-3.225808057	0.618547513	-5.21513382	1.84E-07	2.68E-06	iolT; MFS transporter, SP family, major inositol transporter
K02164	9.603155601	-3.204084862	0.570889581	-5.612442348	1.99E-08	4.49E-07	norE; nitric oxide reductase NorE protein
K01461	6.770791004	-3.201808484	0.601739911	-5.320917602	1.03E-07	1.73E-06	E3.5.1.82; N-acyl-D-glutamate deacylase [EC:3.5.1.82]
K08295	12.51058408	-3.196676288	0.622244961	-5.137327725	2.79E-07	3.72E-06	abmG; 2-aminobenzoate-CoA ligase [EC:6.2.1.32]
K11385	6.89805419	-3.188361333	0.615117841	-5.183334184	2.18E-07	3.02E-06	embA; arabinosyltransferase A [EC:2.4.2.-]
K10008	7.415476499	-3.180355392	0.600329531	-5.297682738	1.17E-07	1.91E-06	gluA; glutamate transport system ATP-binding protein

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K02554	7.307350117	-3.178155642	0.575693809	-5.520565956	3.38E-08	7.06E-07	mhpD; 2-keto-4-pentenoate hydratase [EC:4.2.1.80]
K02379	76.48930361	-3.174585267	0.453173902	-7.005225256	2.47E-12	2.95E-10	fdhD; FdhD protein
K07452	16.95980488	-3.172105715	0.641060436	-4.948216329	7.49E-07	8.29E-06	mcrB; 5-methylcytosine-specific restriction enzyme B
K03736	12.60090621	-3.162133029	0.503420747	-6.281292631	3.36E-10	1.67E-08	eutC; ethanolamine ammonia-lyase small subunit
K03299	33.20836332	-3.158866726	0.441517989	-7.154559515	8.39E-13	1.12E-10	TC.GNTP; gluconate:H+ symporter, GntP family
K14338	6.351811175	-3.152566558	0.539695652	-5.841378465	5.18E-09	1.56E-07	cypD_E, CYP102A2_3; cytochrome P450 / NADPH-cytochrome P450 reductase [EC:1.14.14.1 1.6.2.4]
K03741	190.5403542	-3.15202382	0.47924121	-6.577113468	4.80E-11	3.85E-09	ARSC2, arsC; arsenate reductase [EC:1.20.4.1]
K00632	90.80180994	-3.148507522	0.466370162	-6.751091259	1.47E-11	1.40E-09	fadA, fadI; acetyl-CoA acyltransferase [EC:2.3.1.16]
K06177	16.53545713	-3.147896313	0.599878235	-5.247558803	1.54E-07	2.36E-06	rluA; tRNA pseudouridine32 synthase / 23S rRNA pseudouridine746 synthase [EC:5.4.99.28 5.4.99.29]
K13529	22.04469812	-3.146130106	0.514242332	-6.11799129	9.48E-10	3.98E-08	ada-alkA; AraC family transcriptional regulator, regulatory protein of adaptative response / DNA-3-methyladenine glycosylase II [EC:3.2.2.21]
K07119	27.36423496	-3.143763809	0.477456925	-6.584392528	4.57E-11	3.80E-09	uncharacterized protein
K05911	7.851036125	-3.143348543	0.596469136	-5.26992656	1.36E-07	2.14E-06	see T30018 (Metagenome): GL0022449
K09933	14.92823254	-3.142303098	0.602275113	-5.217388246	1.81E-07	2.66E-06	mtfA; MtfA peptidase
K01682	19.65942745	-3.140862511	0.549554476	-5.715288741	1.10E-08	2.79E-07	acnB; aconitate hydratase 2 / 2-methylisocitrate
K13908	6.424027926	-3.132570786	0.606119565	-5.168239016	2.36E-07	3.20E-06	MUC5B, MG1; mucin-5B
K13006	14.51228706	-3.123206426	0.612250175	-5.101193191	3.38E-07	4.40E-06	wbqR; UDP-perosamine 4-acetyltransferase [EC:2.3.1.-]
K01692	601.7619034	-3.114658634	0.424782307	-7.332364332	2.26E-13	3.64E-11	paaF, echA; enoyl-CoA hydratase [EC:4.2.1.17]
K10006	6.366280102	-3.114298568	0.599206522	-5.19737095	2.02E-07	2.87E-06	gluC; glutamate transport system permease protein
K10007	6.366280102	-3.114298568	0.599206522	-5.19737095	2.02E-07	2.87E-06	gluD; glutamate transport system permease protein
K01066	166.5453146	-3.112339633	0.459565037	-6.772359479	1.27E-11	1.26E-09	aes; acetyl esterase [EC:3.1.1.-]
K02613	26.24260113	-3.112032458	0.474595392	-6.557232778	5.48E-11	4.12E-09	paaE; ring-1,2-phenylacetyl-CoA epoxidase subunit PaaE
K05337	69.74438693	-3.108842622	0.478602673	-6.495664979	8.27E-11	5.84E-09	fer; ferredoxin
K00248	93.54298466	-3.102166777	0.453740796	-6.836869877	8.09E-12	8.58E-10	ACADS, bcd; butyryl-CoA dehydrogenase [EC:1.3.8.1]
K04788	7.176664951	-3.099703149	0.619503898	-5.00352485	5.63E-07	6.67E-06	mbtB; mycobactin phenyloxazoline synthetase
K03841	64.90525335	-3.091704511	0.482540795	-6.407136018	1.48E-10	8.68E-09	FBP, fbp; fructose-1,6-bisphosphatase I [EC:3.1.3.11]
K03557	43.11376581	-3.089396558	0.509980866	-6.057867593	1.38E-09	5.27E-08	fis; Fis family transcriptional regulator, factor for inversion
K14337	6.194599385	-3.084930537	0.591898848	-5.211921842	1.87E-07	2.71E-06	mptA; alpha-1,6-mannosyltransferase [EC:2.4.1.-]

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K02067	262.0686011	-3.084170867	0.468002349	-6.590075619	4.40E-11	3.73E-09	mlaD, linM; phospholipid/cholesterol/gamma-HCH transport system substrate-binding protein
K02486	17.3376905	-3.070904206	0.585118958	-5.248341669	1.53E-07	2.35E-06	two-component system, sensor kinase [EC:2.7.13.3]
K05577	6.586460509	-3.066863111	0.497984026	-6.158557211	7.34E-10	3.19E-08	ndhF; NAD(P)H-quinone oxidoreductase subunit 5
K00130	69.6261249	-3.064456439	0.458748707	-6.680032868	2.39E-11	2.23E-09	betB, gbsA; betaine-aldehyde dehydrogenase [EC:1.2.1.8]
K08083	18.24325841	-3.059029302	0.610911286	-5.007321639	5.52E-07	6.58E-06	algR; two-component system, LytT family, response
K02099	9.300383663	-3.058175521	0.539766478	-5.665738144	1.46E-08	3.50E-07	araC; AraC family transcriptional regulator, arabinose
K02054	23.48089828	-3.05808596	0.466195386	-6.55966587	5.39E-11	4.12E-09	ABC.SP.P1; putative spermidine/putrescine transport
K07137	38.06760887	-3.056450277	0.520924066	-5.867362398	4.43E-09	1.38E-07	uncharacterized protein
K01637	30.7856056	-3.054728796	0.47376822	-6.447728374	1.14E-10	7.52E-09	E4.1.3.1, aceA; isocitrate lyase [EC:4.1.3.1]
K02673	16.46038356	-3.046071721	0.597689583	-5.096410925	3.46E-07	4.50E-06	pilX; type IV pilus assembly protein PilX
K01638	40.12425662	-3.044228966	0.470167802	-6.474771253	9.50E-11	6.51E-09	aceB, glcB; malate synthase [EC:2.3.3.9]
K11081	6.115182761	-3.041932511	0.606323759	-5.017010247	5.25E-07	6.29E-06	phnS; 2-aminoethylphosphonate transport system
K11083	6.115182761	-3.041932511	0.606323759	-5.017010247	5.25E-07	6.29E-06	phnU; 2-aminoethylphosphonate transport system
K14743	10.82957356	-3.04037069	0.619339214	-4.909055686	9.15E-07	9.63E-06	mycP; membrane-anchored mycosin MYCP [EC:3.4.21.-]
K09017	140.8841234	-3.037303411	0.400999034	-7.57434097	3.61E-14	7.01E-12	rutR; TetR/AcrR family transcriptional regulator
K00478	6.560083306	-3.031328941	0.576710971	-5.256235958	1.47E-07	2.28E-06	HR; lysine-specific demethylase hairless [EC:1.14.11.-]
K10620	6.062336489	-3.027491413	0.609608467	-4.966288329	6.82E-07	7.80E-06	cmtB; 2,3-dihydroxy-2,3-dihydro-p-cumate dehydrogenase
K00285	20.35287886	-3.027396926	0.550910862	-5.495257278	3.90E-08	7.94E-07	dadA; D-amino-acid dehydrogenase [EC:1.4.5.1]
K08929	6.742439468	-3.023428186	0.540611615	-5.592606789	2.24E-08	4.97E-07	pufM; photosynthetic reaction center M subunit
K08986	6.201188643	-3.022033026	0.609053509	-4.961851433	6.98E-07	7.90E-06	ycgQ; putative membrane protein
K03811	38.67366776	-3.015538462	0.46413781	-6.497075649	8.19E-11	5.84E-09	pnuC; nicotinamide mononucleotide transporter
K01216	5.496862457	-3.015184772	0.533973174	-5.646697094	1.64E-08	3.79E-07	E3.2.1.73; licheninase [EC:3.2.1.73]
K06917	14.66813789	-3.014602931	0.528037893	-5.709065521	1.14E-08	2.82E-07	selU; tRNA 2-selenouridine synthase [EC:2.9.1.-]
K03293	16.56397458	-3.011342674	0.51284802	-5.871803256	4.31E-09	1.35E-07	TC.AAT; amino acid transporter, AAT family
K00768	62.97032595	-3.008779957	0.52676605	-5.711795502	1.12E-08	2.79E-07	E2.4.2.21, cobU, cobT; nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase [EC:2.4.2.21]
K01424	53.93662305	-3.00550284	0.438549923	-6.853274128	7.22E-12	7.83E-10	E3.5.1.1, ansA, ansB; L-asparaginase [EC:3.5.1.1]
K13788	7.119369556	-3.004638258	0.596700861	-5.035418002	4.77E-07	5.85E-06	pta; phosphate acetyltransferase [EC:2.3.1.8]
K07192	8.218353111	-3.00144636	0.550008698	-5.457088899	4.84E-08	9.36E-07	FLOT; flotillin
K01577	18.02151805	-2.990055019	0.539247504	-5.54486576	2.94E-08	6.28E-07	oxc; oxalyl-CoA decarboxylase [EC:4.1.1.8]

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K07755	20.47216152	-2.987836571	0.608166044	-4.912863194	8.98E-07	9.49E-06	AS3MT; arsenite methyltransferase [EC:2.1.1.137]
K00249	229.3401494	-2.984701542	0.461574493	-6.466348523	1.00E-10	6.78E-09	ACADM, acd; acyl-CoA dehydrogenase [EC:1.3.8.7]
K02614	39.80017278	-2.978967362	0.468877574	-6.353401243	2.11E-10	1.14E-08	paaI; acyl-CoA thioesterase [EC:3.1.2.-]
K09146	5.789531231	-2.978593275	0.594009906	-5.014383172	5.32E-07	6.36E-06	K09146; uncharacterized protein
K09844	5.57131563	-2.970362742	0.567936872	-5.230093149	1.69E-07	2.54E-06	crtC; carotenoid 1,2-hydratase [EC:4.2.1.131]
K11738	7.231819679	-2.967050286	0.600648533	-4.939744501	7.82E-07	8.60E-06	ansP; L-asparagine permease
K10215	5.769315211	-2.963766635	0.600547322	-4.935109233	8.01E-07	8.71E-06	ethA; monooxygenase [EC:1.14.13.-]
K04787	5.76713032	-2.96335921	0.602111279	-4.92161385	8.58E-07	9.18E-06	mbtA; mycobactin salicyl-AMP ligase [EC:6.3.2.-]
K03922	5.68932741	-2.955730039	0.591959274	-4.993130726	5.94E-07	6.96E-06	desA2; acyl-[acyl-carrier-protein] desaturase [EC:1.14.19.2]
K01147	38.53347274	-2.954292176	0.529006144	-5.584608439	2.34E-08	5.15E-07	rnb; exoribonuclease II [EC:3.1.13.1]
K07149	7.445710481	-2.952770485	0.579740302	-5.093264135	3.52E-07	4.56E-06	K07149; uncharacterized protein
K02048	46.30099471	-2.949789252	0.47737775	-6.179151108	6.44E-10	2.89E-08	cysP, sbp; sulfate transport system substrate-binding
K00216	5.498975073	-2.948006739	0.592898095	-4.972198031	6.62E-07	7.60E-06	entA; 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase
K12503	9.667223668	-2.94398874	0.581565935	-5.062175352	4.14E-07	5.19E-06	E2.5.1.68; short-chain Z-isoprenyl diphosphate synthase
K00558	135.0550555	-2.941449531	0.415866054	-7.073069563	1.52E-12	1.86E-10	DNMT1, dcm; DNA (cytosine-5)-methyltransferase 1
K00540	1156.004068	-2.940067193	0.39421667	-7.45799815	8.78E-14	1.64E-11	Naumovozyma dairenensis: NDAI_0K00540
K00803	5.638212954	-2.935422519	0.596636986	-4.91994728	8.66E-07	9.19E-06	AGPS, agpS; alkyldihydroxyacetonephosphate synthase
K08156	22.92605862	-2.931887305	0.49676685	-5.90193832	3.59E-09	1.18E-07	araJ; MFS transporter, DHA1 family, arabinose polymer
K10805	24.18196525	-2.927244372	0.508970704	-5.751302283	8.86E-09	2.36E-07	tesB; acyl-CoA thioesterase II [EC:3.1.2.-]
K09958	26.44053729	-2.927158763	0.535062988	-5.470680706	4.48E-08	8.89E-07	uncharacterized protein
K01607	147.5688142	-2.924044227	0.408008084	-7.166633076	7.69E-13	1.09E-10	pcaC; 4-carboxymuconolactone decarboxylase
K06202	20.86849472	-2.922443217	0.559189708	-5.226210665	1.73E-07	2.58E-06	cyaY; CyaY protein
K03269	40.16020665	-2.921057283	0.516547436	-5.654964244	1.56E-08	3.65E-07	lpxH; UDP-2,3-diacylglucosamine hydrolase [EC:3.6.1.54]
K03366	6.344506044	-2.918567482	0.55740508	-5.235990104	1.64E-07	2.47E-06	butA, budC; meso-butanediol dehydrogenase / (S,S)-butanediol dehydrogenase / diacetyl reductase [EC:1.1.1.-1.1.1.76 1.1.1.304]
K00252	56.09459039	-2.916733461	0.438058756	-6.658315624	2.77E-11	2.53E-09	GCDH, gcdH; glutaryl-CoA dehydrogenase [EC:1.3.8.6]
K04037	6.09616337	-2.915398823	0.535781901	-5.441391017	5.29E-08	1.01E-06	chlL; light-independent protochlorophyllide reductase
K03571	38.23971312	-2.911642749	0.517567701	-5.625626833	1.85E-08	4.24E-07	mreD; rod shape-determining protein MreD
K01101	11.51940966	-2.899475949	0.570676724	-5.080767846	3.76E-07	4.80E-06	E3.1.3.41; 4-nitrophenyl phosphatase [EC:3.1.3.41]
K07794	40.9046024	-2.897026888	0.504140497	-5.74646731	9.11E-09	2.40E-07	tctB; putative tricarboxylic transport membrane protein
K13652	18.41892334	-2.89600449	0.554177167	-5.225773743	1.73E-07	2.58E-06	AraC family transcriptional regulator

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K01090	147.2474773	-2.895901537	0.492115739	-5.884594426	3.99E-09	1.28E-07	protein phosphatase [EC:3.1.3.16]
K06975	26.77328565	-2.892015439	0.489561034	-5.907364436	3.48E-09	1.15E-07	uncharacterized protein
K07442	30.88392656	-2.891613358	0.536118971	-5.393603867	6.91E-08	1.28E-06	tRNA; Transfer RNA biogenesis
K06916	50.76844557	-2.884841127	0.434527582	-6.639028788	3.16E-11	2.83E-09	zapE; cell division protein ZapE
K12447	5.135978908	-2.884012491	0.58475087	-4.932036259	8.14E-07	8.80E-06	USP; UDP-sugar pyrophosphorylase [EC:2.7.7.64]
K05846	36.53237573	-2.879942734	0.496749433	-5.797576288	6.73E-09	1.91E-07	opuBD; osmoprotectant transport system permease protein
K01821	35.72047243	-2.879620927	0.493413113	-5.83612565	5.34E-09	1.59E-07	praC, xylH; 4-oxalocrotonate tautomerase [EC:5.3.2.6]
K07054	61.23276834	-2.876519627	0.473371226	-6.076667679	1.23E-09	4.81E-08	uncharacterized protein
K04033	5.32900231	-2.875720039	0.573101571	-5.017819149	5.23E-07	6.29E-06	eutR; AraC family transcriptional regulator, ethanolamine operon transcriptional activator
K03782	50.29211026	-2.874764342	0.40377927	-7.119643221	1.08E-12	1.40E-10	katG; catalase-peroxidase [EC:1.11.1.21]
K00091	88.07034197	-2.874624694	0.47125673	-6.099912243	1.06E-09	4.30E-08	E1.1.1.219; dihydroflavonol-4-reductase [EC:1.1.1.219]
K01796	59.49510187	-2.874561312	0.437211137	-6.574766903	4.87E-11	3.85E-09	E5.1.99.4, AMACR, mcr; alpha-methylacyl-CoA racemase
K00763	49.73082473	-2.870485798	0.504116434	-5.694092881	1.24E-08	3.04E-07	pncB, NAPRT1; nicotinate phosphoribosyltransferase
K13687	7.121856396	-2.870405942	0.540373576	-5.311891751	1.08E-07	1.80E-06	aftB; arabinofuranosyltransferase [EC:2.4.2.-]
K00102	56.05453518	-2.866484279	0.502859796	-5.700364796	1.20E-08	2.95E-07	dld, LDHD; D-lactate dehydrogenase (cytochrome)
K03735	10.2799442	-2.865653275	0.485133703	-5.906935057	3.49E-09	1.15E-07	eutB; ethanolamine ammonia-lyase large subunit
K00956	31.90792221	-2.865046095	0.486011555	-5.895016412	3.75E-09	1.22E-07	cysN; sulfate adenylyltransferase subunit 1 [EC:2.7.7.4]
K00571	198.3687468	-2.861554796	0.491770448	-5.818883201	5.92E-09	1.73E-07	E2.1.1.72; site-specific DNA-methyltransferase (adenine-
K14393	53.36112909	-2.857711035	0.488513518	-5.849809531	4.92E-09	1.50E-07	actP; cation/acetate symporter
K01560	46.33071639	-2.853858567	0.485202788	-5.881785172	4.06E-09	1.30E-07	E3.8.1.2; 2-haloacid dehalogenase [EC:3.8.1.2]
K01070	16.02803021	-2.851769818	0.581824727	-4.901424235	9.51E-07	9.90E-06	frmB, ESD, fghA; S-formylglutathione hydrolase
K01028	34.34035423	-2.851472845	0.503363143	-5.664842335	1.47E-08	3.50E-07	E2.8.3.5A, scoA; 3-oxoacid CoA-transferase subunit A
K11942	25.0983692	-2.849768833	0.483837551	-5.88992902	3.86E-09	1.25E-07	E5.4.99.2; methylmalonyl-CoA mutase [EC:5.4.99.2]
K07224	6.825135199	-2.840102425	0.572624527	-4.959798772	7.06E-07	7.93E-06	efeO; iron uptake system component EfeO
K00817	168.7077689	-2.836011242	0.464281902	-6.108382058	1.01E-09	4.15E-08	hisC; histidinol-phosphate aminotransferase [EC:2.6.1.9]
K02231	74.12641149	-2.829448521	0.552674742	-5.119554609	3.06E-07	4.02E-06	cobP, cobU; adenosylcobinamide kinase / adenosylcobinamide-phosphate guanylyltransferase [EC:2.7.1.156 2.7.7.62]
K01267	5.012243311	-2.825645288	0.557131746	-5.071772181	3.94E-07	4.99E-06	DNPEP; aspartyl aminopeptidase [EC:3.4.11.21]
K01913	33.23694412	-2.825072155	0.503233692	-5.613837462	1.98E-08	4.49E-07	see T30017 (Metagenome): GL0065989
K03405	31.17406875	-2.82068573	0.528818975	-5.333934417	9.61E-08	1.64E-06	chlI, bchI; magnesium chelatase subunit I [EC:6.6.1.1]

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K03307	88.76288509	-2.814956274	0.438510379	-6.419360654	1.37E-10	8.62E-09	TC.SSS; solute:Na+ symporter, SSS family
K12410	54.33432233	-2.812480024	0.440542354	-6.384130832	1.72E-10	9.45E-09	npdA; NAD-dependent deacetylase [EC:3.5.1.-]
K01812	5.370174929	-2.810788647	0.46319469	-6.068266131	1.29E-09	4.98E-08	uxaC; glucuronate isomerase [EC:5.3.1.12]
K07107	112.4887526	-2.810587448	0.427164371	-6.579639225	4.72E-11	3.85E-09	ybgC; acyl-CoA thioester hydrolase [EC:3.1.2.-]
K01935	86.94444683	-2.808690807	0.524476045	-5.35523183	8.54E-08	1.50E-06	bioD; dethiobiotin synthetase [EC:6.3.3.3]
K02045	50.35151478	-2.804557737	0.464571199	-6.036873879	1.57E-09	5.91E-08	cysA; sulfate transport system ATP-binding protein
K01757	5.632722142	-2.79966261	0.570946951	-4.903542446	9.41E-07	9.84E-06	E4.3.3.2; strictosidine synthase [EC:4.3.3.2]
K00891	110.2245756	-2.794868527	0.480515296	-5.816398661	6.01E-09	1.74E-07	E2.7.1.71, aroK, aroL; shikimate kinase [EC:2.7.1.71]
K08365	100.503765	-2.789324933	0.507121106	-5.500313234	3.79E-08	7.75E-07	merR; MerR family transcriptional regulator, mercuric resistance operon regulatory protein
K07006	54.402203	-2.787794388	0.443052526	-6.29224353	3.13E-10	1.59E-08	uncharacterized protein
K05709	17.71290435	-2.785353486	0.547135278	-5.090794908	3.57E-07	4.60E-06	hcaF, hcaA2; 3-phenylpropionate/trans-cinnamate dioxygenase subunit beta [EC:1.14.12.19]
K01582	19.6685631	-2.783898616	0.561383335	-4.958997607	7.09E-07	7.94E-06	E4.1.1.18, ldcC, cadA; lysine decarboxylase [EC:4.1.1.18]
K04761	111.8606014	-2.77667776	0.44593687	-6.226616254	4.77E-10	2.20E-08	oxyR; LysR family transcriptional regulator, hydrogen peroxide-inducible genes activator
K00499	17.49658147	-2.773718719	0.539658483	-5.139766734	2.75E-07	3.69E-06	CMO; choline monooxygenase [EC:1.14.15.7]
K01886	43.22900719	-2.773669882	0.480079025	-5.777527732	7.58E-09	2.10E-07	QARS, glnS; glutaminyl-tRNA synthetase [EC:6.1.1.18]
K01175	188.149521	-2.771171112	0.376185145	-7.366508612	1.75E-13	3.02E-11	ybfF; esterase [EC:3.1.-.-]
K01640	53.72847499	-2.770253451	0.451463029	-6.136169019	8.45E-10	3.62E-08	E4.1.3.4, HMGCL, hmgL; hydroxymethylglutaryl-CoA lyase
K03690	41.79229203	-2.75928555	0.503779765	-5.477166296	4.32E-08	8.61E-07	ubiJ; ubiquinone biosynthesis protein UbiJ
K11689	6.585737471	-2.75179662	0.552980898	-4.976295982	6.48E-07	7.48E-06	dctQ; C4-dicarboxylate transporter, DctQ subunit
K01665	82.89956862	-2.748715461	0.512765243	-5.360572891	8.30E-08	1.48E-06	pabB; para-aminobenzoate synthetase component I
K03969	10.43801938	-2.746768383	0.467776739	-5.871964447	4.31E-09	1.35E-07	pspA; phage shock protein A
K04754	52.13017068	-2.742984623	0.494444634	-5.547607226	2.90E-08	6.22E-07	mlaA, vacJ; phospholipid-binding lipoprotein MlaA
K03586	46.45525318	-2.740863483	0.478869075	-5.72361764	1.04E-08	2.69E-07	ftsL; cell division protein FtsL
K00606	99.6706616	-2.738030074	0.479274842	-5.712859999	1.11E-08	2.79E-07	panB; 3-methyl-2-oxobutanoate hydroxymethyltransferase
K09771	22.98042425	-2.732453716	0.426153246	-6.411904027	1.44E-10	8.68E-09	TC.SMR3; small multidrug resistance family-3 protein
K03749	41.68560174	-2.731795615	0.500516812	-5.457949762	4.82E-08	9.36E-07	dedD; DedD protein
K03298	48.62974449	-2.728465584	0.454061284	-6.009024944	1.87E-09	6.96E-08	TC.DME; drug/metabolite transporter, DME family
K11178	23.31183492	-2.728445628	0.536015593	-5.090235548	3.58E-07	4.61E-06	yagS; xanthine dehydrogenase YagS FAD-binding subunit
K00626	301.0509415	-2.725506065	0.426464881	-6.390927342	1.65E-10	9.37E-09	E2.3.1.9, atoB; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K02348	9.47475074	-2.72451739	0.524402372	-5.195471141	2.04E-07	2.88E-06	elaA; ElaA protein
K01469	18.40471337	-2.721754618	0.513687783	-5.298460872	1.17E-07	1.91E-06	OPLAH, OXP1, oplAH; 5-oxoprolinase (ATP-hydrolysing)
K07175	42.25209042	-2.719005149	0.430614952	-6.314237671	2.71E-10	1.41E-08	phoH2; PhoH-like ATPase
K00284	82.16740627	-2.717866862	0.465977969	-5.83260807	5.46E-09	1.60E-07	E1.4.7.1; glutamate synthase (ferredoxin) [EC:1.4.7.1]
K00119	80.91677973	-2.716273857	0.478723601	-5.673991952	1.40E-08	3.37E-07	T30018 (Metagenome): GL0044853
K03750	108.0578583	-2.705116623	0.419779813	-6.444132238	1.16E-10	7.52E-09	moeA; molybdopterin molybdotransferase [EC:2.10.1.1]
K07399	45.20344249	-2.700434383	0.494181435	-5.464459386	4.64E-08	9.10E-07	resB, ccs1; cytochrome c biogenesis protein
K02234	57.8097301	-2.699295057	0.433943272	-6.220386935	4.96E-10	2.27E-08	cobW; cobalamin biosynthesis protein CobW
K10680	46.10032252	-2.697292297	0.51051173	-5.283506991	1.27E-07	2.03E-06	nemA; N-ethylmaleimide reductase [EC:1.-.-.-]
K00163	88.32881823	-2.695535051	0.42464345	-6.347760811	2.18E-10	1.17E-08	aceE; pyruvate dehydrogenase E1 component [EC:1.2.4.1]
K03502	47.79066938	-2.686282375	0.525151718	-5.115250095	3.13E-07	4.10E-06	DPO5C, umuC; DNA polymerase V
K02053	23.46158758	-2.683702861	0.460119319	-5.832623737	5.46E-09	1.60E-07	ABC.SP.P; putative spermidine/putrescine transport
K07088	67.68548756	-2.681005874	0.502110731	-5.339471377	9.32E-08	1.62E-06	uncharacterized protein
K01897	245.7177025	-2.67898773	0.409286161	-6.545512612	5.93E-11	4.39E-09	ACSL, fadD; long-chain acyl-CoA synthetase [EC:6.2.1.3]
K04090	19.13486229	-2.677203148	0.509977866	-5.249645773	1.52E-07	2.35E-06	E1.2.7.8; indolepyruvate ferredoxin oxidoreductase
K01012	109.8821958	-2.673456071	0.519276078	-5.148429095	2.63E-07	3.53E-06	bioB; biotin synthase [EC:2.8.1.6]
K01253	22.04942953	-2.672095359	0.448535129	-5.957382574	2.56E-09	8.85E-08	EPHX1; microsomal epoxide hydrolase [EC:3.3.2.9]
K06442	51.41581267	-2.671967213	0.504102129	-5.300448186	1.16E-07	1.90E-06	tlyA; 23S rRNA (cytidine1920-2'-O)/16S rRNA (cytidine1409-2'-O)-methyltransferase [EC:2.1.1.226 2.1.1.227]
K03981	32.53210987	-2.669117019	0.527815744	-5.056910575	4.26E-07	5.30E-06	dsbC; thiol:disulfide interchange protein DsbC [EC:5.3.4.1]
K09796	57.03438831	-2.667623571	0.467070138	-5.711398259	1.12E-08	2.79E-07	pccA; periplasmic copper chaperone A
K10918	59.23057559	-2.657534149	0.497871382	-5.337792539	9.41E-08	1.62E-06	aphB; LysR family transcriptional regulator, transcriptional
K04044	32.50427979	-2.651348742	0.521080657	-5.088173408	3.62E-07	4.64E-06	hscA; molecular chaperone HscA
K00001	117.0342415	-2.644702939	0.429115959	-6.163142811	7.13E-10	3.17E-08	E1.1.1.1, adh; alcohol dehydrogenase [EC:1.1.1.1]
K01997	207.2830536	-2.639769157	0.45531397	-5.797689794	6.72E-09	1.91E-07	livH; branched-chain amino acid transport system
K01626	70.02436263	-2.63645788	0.450560879	-5.851501989	4.87E-09	1.49E-07	E2.5.1.54, aroF, aroG, aroH; 3-deoxy-7-phosphoheptulonate synthase [EC:2.5.1.54]
K01759	150.5321805	-2.635563949	0.41089284	-6.414236734	1.42E-10	8.68E-09	GLO1, gloA; lactoylglutathione lyase [EC:4.4.1.5]
K01044	19.90327677	-2.631787754	0.506940336	-5.191513804	2.09E-07	2.91E-06	CES1; carboxylesterase 1 [EC:3.1.1.1]
K01829	24.01805022	-2.631230841	0.503900749	-5.221724415	1.77E-07	2.62E-06	E5.3.4.1; protein disulfide-isomerase [EC:5.3.4.1]
K03546	60.4018977	-2.628018684	0.477581282	-5.502767343	3.74E-08	7.68E-07	sbcC; exonuclease SbcC

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K06196	105.3295948	-2.623253353	0.487902779	-5.376590311	7.59E-08	1.38E-06	ccdA; cytochrome c-type biogenesis protein
K08310	24.05501481	-2.616647243	0.519574625	-5.03613363	4.75E-07	5.85E-06	nudB, ntpA; dihydroneopterin triphosphate diphosphatase
K01718	46.73098599	-2.608291912	0.473068398	-5.513561936	3.52E-08	7.32E-07	see T30024 (Metagenome): GL0021017
K03437	60.43391185	-2.605796498	0.436624199	-5.968053318	2.40E-09	8.42E-08	spoU; RNA methyltransferase, TrmH family
K03767	47.08758649	-2.604898393	0.484342615	-5.378214327	7.52E-08	1.37E-06	PPIA; peptidyl-prolyl cis-trans isomerase A (cyclophilin A)
K03600	40.84707018	-2.599993181	0.523384664	-4.967652588	6.78E-07	7.76E-06	sspB; stringent starvation protein B
K01999	264.53352	-2.599312774	0.451460815	-5.757560095	8.53E-09	2.31E-07	livK; branched-chain amino acid transport system substrate-
K00055	12.4339732	-2.59787013	0.528556956	-4.915024009	8.88E-07	9.41E-06	E1.1.1.90; aryl-alcohol dehydrogenase [EC:1.1.1.90]
K06904	23.69793685	-2.591937195	0.494098216	-5.245793472	1.56E-07	2.37E-06	uncharacterized protein
K11712	36.55820271	-2.590004288	0.505724531	-5.121373649	3.03E-07	3.99E-06	dctR; two-component system, LuxR family, response
K03781	53.66759473	-2.585043731	0.4805422	-5.379431255	7.47E-08	1.37E-06	katE, CAT, catB, srpA; catalase [EC:1.11.1.6]
K07182	41.28047108	-2.584723286	0.494914965	-5.222560378	1.76E-07	2.62E-06	CBS domain-containing protein
K06020	82.47834629	-2.583028557	0.481490397	-5.364652283	8.11E-08	1.45E-06	E3.6.3.25; sulfate-transporting ATPase [EC:3.6.3.25]
K05772	28.72198467	-2.581632556	0.518868758	-4.975502029	6.51E-07	7.49E-06	tupA, vupA; tungstate transport system substrate-binding
K00652	85.84407413	-2.578526741	0.49771368	-5.18074316	2.21E-07	3.06E-06	bioF; 8-amino-7-oxononanoate synthase [EC:2.3.1.47]
K01046	87.02953606	-2.569728696	0.507367532	-5.064826847	4.09E-07	5.15E-06	E3.1.1.3; triacylglycerol lipase [EC:3.1.1.3]
K06413	24.47825082	-2.566440844	0.472264604	-5.434328177	5.50E-08	1.04E-06	spoVK; stage V sporulation protein K
K00681	95.11404689	-2.565342303	0.433069078	-5.923633051	3.15E-09	1.06E-07	ggt; gamma-glutamyltranspeptidase / glutathione hydrolase
K03712	111.9475209	-2.564664464	0.431223468	-5.947413943	2.72E-09	9.34E-08	marR; MarR family transcriptional regulator, multiple antibiotic resistance protein MarR
K03611	22.82766507	-2.563851796	0.516811847	-4.96089981	7.02E-07	7.91E-06	dsbB; disulfide bond formation protein DsbB
K07749	207.3703679	-2.557702573	0.420258659	-6.086019919	1.16E-09	4.57E-08	frc; formyl-CoA transferase [EC:2.8.3.16]
K07393	13.37829487	-2.557607098	0.508090018	-5.033767652	4.81E-07	5.87E-06	ECM4; putative glutathione S-transferase
K01919	63.8114227	-2.552840876	0.445106914	-5.735343121	9.73E-09	2.52E-07	gshA; glutamate--cysteine ligase [EC:6.3.2.2]
K11275	23.251043	-2.552729474	0.479439016	-5.324409134	1.01E-07	1.70E-06	H1_5; histone H1/5
K05786	33.45690904	-2.551569421	0.497100624	-5.132903271	2.85E-07	3.78E-06	rarD; chloramphenicol-sensitive protein RarD
K01934	69.5383163	-2.550321977	0.457573722	-5.573576137	2.50E-08	5.44E-07	MTHFS; 5-formyltetrahydrofolate cyclo-ligase [EC:6.3.3.2]
K07287	41.40340553	-2.550144724	0.519621572	-4.90769603	9.22E-07	9.68E-06	bamC; outer membrane protein assembly factor BamC
K08483	69.16922899	-2.54668241	0.472768141	-5.386747096	7.17E-08	1.32E-06	PTS-EI.PTSI, ptsI; phosphotransferase system, enzyme I,
K07793	91.42269065	-2.544553218	0.486018679	-5.235504986	1.65E-07	2.47E-06	tctA; putative tricarboxylic transport membrane protein
K02841	48.87729468	-2.544499926	0.518001239	-4.912150274	9.01E-07	9.50E-06	waaC, rfaC; heptosyltransferase I [EC:2.4.-.-]
K01414	42.6891713	-2.54076703	0.486601939	-5.221448636	1.78E-07	2.62E-06	prlC; oligopeptidase A [EC:3.4.24.70]

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K01995	156.2473278	-2.54033525	0.426177683	-5.960742085	2.51E-09	8.74E-08	livG; branched-chain amino acid transport system ATP-
K00140	65.22570778	-2.534699601	0.406367088	-6.237462808	4.45E-10	2.07E-08	mmsA, iolA, ALDH6A1; malonate-semialdehyde dehydrogenase (acetylating) / methylmalonate-semialdehyde dehydrogenase [EC:1.2.1.18 1.2.1.27]
K05905	5.53502044	-2.531699966	0.502830542	-5.034896953	4.78E-07	5.85E-06	E1.8.1.8; protein-disulfide reductase [EC:1.8.1.8]
K07506	44.36380004	-2.523896358	0.437897817	-5.763665077	8.23E-09	2.24E-07	AraC family transcriptional regulator
K05971	49.02863868	-2.520012649	0.483542976	-5.211558795	1.87E-07	2.71E-06	see T30017 (Metagenome): GL0062251
K11690	65.4664082	-2.516743271	0.465984318	-5.40091839	6.63E-08	1.23E-06	dctM; C4-dicarboxylate transporter, DctM subunit
K07246	24.85441111	-2.515427643	0.508646014	-4.945340321	7.60E-07	8.38E-06	ttuC, dmlA; tartrate dehydrogenase/decarboxylase / D-malate dehydrogenase [EC:1.1.1.93 4.1.1.73 1.1.1.83]
K01799	26.983741	-2.51028322	0.46933363	-5.348611437	8.86E-08	1.54E-06	nicE, maiA; maleate isomerase [EC:5.2.1.1]
K07044	11.76896837	-2.50832997	0.497011992	-5.046819818	4.49E-07	5.56E-06	uncharacterized protein
K01908	33.38635613	-2.506681655	0.475474399	-5.271959251	1.35E-07	2.13E-06	prpE; propionyl-CoA synthetase [EC:6.2.1.17]
K01969	40.45320151	-2.506291545	0.489313255	-5.122059379	3.02E-07	3.99E-06	E6.4.1.4B; 3-methylcrotonyl-CoA carboxylase beta subunit
K07141	48.75080599	-2.503102194	0.458205551	-5.462836906	4.69E-08	9.14E-07	mocA; molybdenum cofactor cytidylyltransferase
K01918	79.88543755	-2.501257418	0.454406403	-5.504450203	3.70E-08	7.64E-07	panC; pantoate--beta-alanine ligase [EC:6.3.2.1]
K10027	16.81689633	-2.50029802	0.400506248	-6.242843989	4.30E-10	2.04E-08	crtI; phytoene desaturase [EC:1.3.99.26 1.3.99.28
K00108	84.94199557	-2.493286186	0.472724957	-5.274285076	1.33E-07	2.11E-06	betA, CHDH; choline dehydrogenase [EC:1.1.99.1]
K06979	70.40229659	-2.487769861	0.496935191	-5.006225975	5.55E-07	6.60E-06	mph; macrolide phosphotransferase
K03428	8.373302117	-2.487681908	0.501339989	-4.962065584	6.97E-07	7.90E-06	bchM, chlM; magnesium-protoporphyrin O-
K14058	59.11914341	-2.487392454	0.489699914	-5.079421868	3.79E-07	4.82E-06	ttcA; tRNA 2-thiocytidine biosynthesis protein TtcA
K00356	185.9371971	-2.487059694	0.429343963	-5.792697477	6.93E-09	1.96E-07	E1.6.99.3; NADH dehydrogenase [EC:1.6.99.3]
K11189	58.56632848	-2.480690223	0.455349303	-5.447884093	5.10E-08	9.82E-07	PTS-HPR; phosphocarrier protein
K09386	46.11983607	-2.471696861	0.498247757	-4.960778704	7.02E-07	7.91E-06	uncharacterized protein
K08259	45.71689582	-2.471470651	0.492201765	-5.021255153	5.13E-07	6.22E-06	lytM; lysostaphin [EC:3.4.24.75]
K00496	19.00299617	-2.468015587	0.498145064	-4.95441141	7.25E-07	8.07E-06	alkB1_2; alkane 1-monooxygenase [EC:1.14.15.3]
K01523	54.07463442	-2.467301395	0.42923971	-5.748073481	9.03E-09	2.39E-07	hisE; phosphoribosyl-ATP pyrophosphohydrolase
K09701	50.83453893	-2.465715688	0.429459409	-5.741440601	9.39E-09	2.46E-07	uncharacterized protein
K03578	51.51017975	-2.463791761	0.426836795	-5.772210345	7.82E-09	2.16E-07	hrpA; ATP-dependent helicase HrpA [EC:3.6.4.13]
K01726	81.39756412	-2.459920361	0.392110934	-6.27353166	3.53E-10	1.73E-08	see T30016 (Metagenome): GL0015944
K01996	176.6277663	-2.457302753	0.439878469	-5.586321969	2.32E-08	5.12E-07	livF; branched-chain amino acid transport system ATP-
K00383	30.5043019	-2.453062162	0.463498247	-5.292495015	1.21E-07	1.95E-06	GSR, gor; glutathione reductase (NADPH) [EC:1.8.1.7]

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K05993	59.02615712	-2.45181848	0.460976423	-5.318750278	1.04E-07	1.74E-06	see Streptococcus dysgalactiae subsp. equisimilis AC-
K02046	33.82160285	-2.450211182	0.469509483	-5.218661758	1.80E-07	2.65E-06	cysU; sulfate transport system permease protein
K03270	44.28159064	-2.444194769	0.492311557	-4.964731645	6.88E-07	7.82E-06	kdsC; 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase (KDO 8-P phosphatase) [EC:3.1.3.45]
K01965	25.76275114	-2.439909015	0.488387917	-4.995842297	5.86E-07	6.90E-06	PCCA, pccA; propionyl-CoA carboxylase alpha chain
K03639	95.47617607	-2.437902281	0.441498075	-5.521886543	3.35E-08	7.04E-07	MOCS1, moaA; cyclic pyranopterin phosphate synthase
K02346	103.5186876	-2.433843902	0.488566791	-4.981599134	6.31E-07	7.31E-06	DPO4, dinB; DNA polymerase IV [EC:2.7.7.7]
K00982	69.3588314	-2.431479074	0.438548521	-5.544378686	2.95E-08	6.28E-07	glnE; glutamate-ammonia-ligase adenylyltransferase
K06949	59.29471884	-2.429280644	0.466994144	-5.201950977	1.97E-07	2.83E-06	rsgA, engC; ribosome biogenesis GTPase [EC:3.6.1.-]
K01297	42.17514166	-2.421416002	0.490538124	-4.936244263	7.96E-07	8.70E-06	ldcA; muramoyltetrapeptide carboxypeptidase
K02066	148.6405903	-2.421119625	0.405167071	-5.975608087	2.29E-09	8.16E-08	mlaE, linK; phospholipid/cholesterol/gamma-HCH transport system permease protein
K00257	353.7886374	-2.420601038	0.444820021	-5.441753792	5.28E-08	1.01E-06	E1.3.99.-; [EC:1.3.99.-]
K03722	68.29028592	-2.417208582	0.427706666	-5.651556949	1.59E-08	3.71E-07	dinG; ATP-dependent DNA helicase DinG [EC:3.6.4.12]
K03577	89.78064303	-2.411643967	0.476006352	-5.066411316	4.05E-07	5.12E-06	acrR, smeT; TetR/AcrR family transcriptional regulator,
K00432	66.84123148	-2.411011529	0.416911728	-5.783026395	7.34E-09	2.05E-07	gpx; glutathione peroxidase [EC:1.11.1.9]
K03417	22.19100704	-2.41005548	0.479900482	-5.021990123	5.11E-07	6.21E-06	prpB; methylisocitrate lyase [EC:4.1.3.30]
K02484	61.73117614	-2.407332889	0.434655435	-5.538485645	3.05E-08	6.47E-07	two-component system, OmpR family, sensor kinase
K02914	41.02483818	-2.405912882	0.488907567	-4.920997429	8.61E-07	9.19E-06	RP-L34, MRPL34, rpmH; large subunit ribosomal protein
K01971	111.9365759	-2.399582972	0.47213316	-5.082428386	3.73E-07	4.77E-06	ligD; bifunctional non-homologous end joining protein LigD
K02047	34.03594708	-2.399375739	0.461845736	-5.19518868	2.05E-07	2.88E-06	cysW; sulfate transport system permease protein
K01878	66.42439625	-2.397019224	0.464597903	-5.159341464	2.48E-07	3.35E-06	glyQ; glycyl-tRNA synthetase alpha chain [EC:6.1.1.14]
K03565	66.41188496	-2.39410899	0.462539718	-5.17600737	2.27E-07	3.11E-06	recX; regulatory protein
K00122	63.86561242	-2.393628936	0.46599584	-5.13658864	2.80E-07	3.73E-06	FDH; formate dehydrogenase [EC:1.2.1.2]
K07288	42.52356754	-2.392989586	0.441175936	-5.424116298	5.82E-08	1.09E-06	uncharacterized membrane protein
K01422	102.615287	-2.38974534	0.475001945	-5.031022224	4.88E-07	5.94E-06	AXL1; protease AXL1 [EC:3.4.24.-]
K03321	97.38131942	-2.38322279	0.431577443	-5.522120838	3.35E-08	7.04E-07	TC.SULP; sulfate permease, SulP family
K02825	60.43173885	-2.382313944	0.460874424	-5.169117266	2.35E-07	3.20E-06	pyrR; pyrimidine operon attenuation protein / uracil phosphoribosyltransferase [EC:2.4.2.9]
K00645	102.1123314	-2.379378838	0.413309524	-5.756893321	8.57E-09	2.31E-07	fabD; [acyl-carrier-protein] S-malonyltransferase
K01895	146.0373457	-2.376301344	0.4347424	-5.465998586	4.60E-08	9.05E-07	ACSS, acs; acetyl-CoA synthetase [EC:6.2.1.1]
K03637	60.58291081	-2.362422626	0.430817789	-5.483577245	4.17E-08	8.34E-07	moaC; cyclic pyranopterin phosphate synthase

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K00104	95.28749388	-2.361203241	0.385474788	-6.125441446	9.04E-10	3.83E-08	glcD; glycolate oxidase [EC:1.1.3.15]
K06954	58.48771582	-2.35167967	0.430007559	-5.468926355	4.53E-08	8.94E-07	uncharacterized protein
K02427	52.0521535	-2.34782829	0.477179553	-4.920219805	8.64E-07	9.19E-06	rlmE, rrmJ, ftsJ; 23S rRNA (uridine2552-2'-O)-
K00480	75.00585239	-2.322787804	0.433708812	-5.355638941	8.53E-08	1.50E-06	E1.14.13.1; salicylate hydroxylase [EC:1.14.13.1]
K07588	58.51642913	-2.3154023	0.464264921	-4.987243696	6.12E-07	7.12E-06	argK; LAO/AO transport system kinase [EC:2.7.-.-]
K02258	56.7354859	-2.309025424	0.443462649	-5.206809254	1.92E-07	2.76E-06	COX11; cytochrome c oxidase assembly protein subunit 11
K00077	59.90018584	-2.306240166	0.445955853	-5.171453966	2.32E-07	3.17E-06	panE, apbA; 2-dehydropantoate 2-reductase
K00798	107.9361187	-2.28875713	0.398713431	-5.740356238	9.45E-09	2.46E-07	MMAB, pduO; cob(I)alamin adenosyltransferase
K01724	92.98247082	-2.26627914	0.422909811	-5.358776463	8.38E-08	1.48E-06	PCBD, phhB; 4a-hydroxytetrahydrobiopterin dehydratase
K01496	54.81947581	-2.260543082	0.446377725	-5.064193295	4.10E-07	5.15E-06	hisI; phosphoribosyl-AMP cyclohydrolase [EC:3.5.4.19]
K02492	54.52375155	-2.259480205	0.427191249	-5.289153773	1.23E-07	1.98E-06	hemA; glutamyl-tRNA reductase [EC:1.2.1.70]
K09159	55.36110692	-2.242666573	0.45534978	-4.925151325	8.43E-07	9.03E-06	cptB; antitoxin CptB
K01998	180.9478022	-2.23936722	0.434654947	-5.152057357	2.58E-07	3.47E-06	livM; branched-chain amino acid transport system
K00680	285.5290848	-2.238567863	0.429151113	-5.216269502	1.83E-07	2.67E-06	E2.3.1.-; [EC:2.3.1.-]
K08994	7.984184316	-2.23232694	0.411832034	-5.420479121	5.94E-08	1.11E-06	yneE; putative membrane protein
K00560	104.5521717	-2.224418951	0.42427164	-5.242912186	1.58E-07	2.40E-06	thyA, TYMS; thymidylate synthase [EC:2.1.1.45]
K03524	74.96454434	-2.21997965	0.423992909	-5.235888632	1.64E-07	2.47E-06	birA; BirA family transcriptional regulator, biotin operon repressor / biotin-[acetyl-CoA-carboxylase] ligase [EC:6.3.4.15]
K01628	77.45236962	-2.217270352	0.396300187	-5.594926334	2.21E-08	4.92E-07	fucA; L-fuculose-phosphate aldolase [EC:4.1.2.17]
K00574	96.997547	-2.213628004	0.402949062	-5.493567835	3.94E-08	7.95E-07	cfa; cyclopropane-fatty-acyl-phospholipid synthase
K00325	60.76732096	-2.209073553	0.448108395	-4.929774983	8.23E-07	8.88E-06	pntB; NAD(P) transhydrogenase subunit beta [EC:1.6.1.2]
K04773	93.48173005	-2.202324304	0.368729499	-5.97273695	2.33E-09	8.24E-08	sppA; protease IV [EC:3.4.21.-]
K00344	201.6724272	-2.200881467	0.423897233	-5.192016598	2.08E-07	2.91E-06	qor, CRYZ; NADPH2:quinone reductase [EC:1.6.5.5]
K09773	34.23543637	-2.200202281	0.44021228	-4.998048398	5.79E-07	6.84E-06	ppsR; [pyruvate, water dikinase]-phosphate phosphotransferase / [pyruvate, water dikinase] kinase
K01625	8.250844803	-2.195287098	0.379594916	-5.783236294	7.33E-09	2.05E-07	eda; 2-dehydro-3-deoxyphosphogluconate aldolase / (4S)-4-hydroxy-2-oxoglutarate aldolase [EC:4.1.2.14 4.1.3.42]
K07402	78.05589064	-2.191818655	0.415411044	-5.276264768	1.32E-07	2.09E-06	xdhC; xanthine dehydrogenase accessory factor
K02479	63.35687388	-2.189870933	0.446742009	-4.901869288	9.49E-07	9.90E-06	two-component system, NarL family, response regulator
K00275	55.0180061	-2.179534557	0.414656944	-5.256235515	1.47E-07	2.28E-06	pdxH, PNPO; pyridoxamine 5'-phosphate oxidase
K01770	51.81729042	-2.17498853	0.431914129	-5.035696645	4.76E-07	5.85E-06	ispF; 2-C-methyl-D-erythritol 2,4-cyclodiphosphate

Table C3 (cont). Overabundant genes in disinfected samples (chlorinated+chloraminated), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K00123	107.6040545	-2.174968334	0.396311706	-5.488024452	4.06E-08	8.17E-07	fdoG, fdfH; formate dehydrogenase major subunit
K00800	69.61746536	-2.160678357	0.433238248	-4.987275176	6.12E-07	7.12E-06	aroA; 3-phosphoshikimate 1-carboxyvinyltransferase
K00655	185.8649394	-2.159709894	0.408016151	-5.293197071	1.20E-07	1.95E-06	plsC; 1-acyl-sn-glycerol-3-phosphate acyltransferase
K03589	76.05180885	-2.154424492	0.414375628	-5.199206583	2.00E-07	2.86E-06	ftsQ; cell division protein FtsQ
K00997	66.122603	-2.151312535	0.436791834	-4.925258144	8.42E-07	9.03E-06	acpS; holo-[acyl-carrier protein] synthase [EC:2.7.8.7]
K02050	181.862578	-2.128939043	0.401138228	-5.30724547	1.11E-07	1.84E-06	ABC.SN.P; NitT/TauT family transport system permease
K01633	71.35202845	-2.128935054	0.399409026	-5.330212673	9.81E-08	1.67E-06	folB; dihydroneopterin aldolase / 7,8-dihydroneopterin epimerase [EC:4.1.2.25 5.1.99.8]
K07386	15.70648385	-2.126836712	0.420441739	-5.058576523	4.22E-07	5.27E-06	putative endopeptidase [EC:3.4.24.-]
K00548	120.5542256	-2.125330402	0.409730917	-5.187137005	2.14E-07	2.97E-06	metH, MTR; 5-methyltetrahydrofolate--homocysteine methyltransferase [EC:2.1.1.13]
K03111	170.7871843	-2.119637991	0.393928275	-5.380771385	7.42E-08	1.36E-06	ssb; single-strand DNA-binding protein
K00058	183.966196	-2.116976659	0.409317736	-5.171964158	2.32E-07	3.17E-06	serA, PHGDH; D-3-phosphoglycerate dehydrogenase
K02037	112.743901	-2.111431403	0.427554273	-4.938393876	7.88E-07	8.64E-06	pstC; phosphate transport system permease protein
K01423	171.5801886	-2.108588328	0.392355664	-5.37417583	7.69E-08	1.39E-06	EC:3.4.-.-
K01426	86.14971764	-2.104996084	0.401041121	-5.24882854	1.53E-07	2.35E-06	E3.5.1.4, amiE; amidase [EC:3.5.1.4]
K01609	68.07890526	-2.07189488	0.403532085	-5.134399361	2.83E-07	3.76E-06	trpC; indole-3-glycerol phosphate synthase [EC:4.1.1.48]
K00796	88.17138352	-2.056474694	0.388996868	-5.2866099	1.25E-07	2.00E-06	folP; dihydropteroate synthase [EC:2.5.1.15]
K06941	75.12083215	-2.00951349	0.397204922	-5.059135417	4.21E-07	5.26E-06	rlmN; 23S rRNA (adenine2503-C2)-methyltransferase
K03519	58.40026372	-2.006583545	0.405546271	-4.947853526	7.50E-07	8.29E-06	coxM, cutM; carbon-monoxide dehydrogenase medium

Table C4. Overabundant genes in disinfectant residual-free samples (Drf), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K07282	17.1259513	3.44065253	0.439062039	7.83636986	4.64E-15	1.54E-12	pgsA, capA; poly-gamma-glutamate synthesis protein (capsule biosynthesis protein)
K12035	16.77135735	3.322451529	0.436807983	7.606206063	2.82E-14	5.92E-12	TRIM71; tripartite motif-containing protein 71
K09792	8.741508549	3.304862239	0.41416037	7.97966797	1.47E-15	8.55E-13	uncharacterized protein
K00018	7.410615929	3.095419614	0.530294013	5.83717624	5.31E-09	1.59E-07	hprA; glycerate dehydrogenase [EC:1.1.1.29]
K07715	6.981829007	3.080158552	0.482378788	6.385352403	1.71E-10	9.45E-09	glrR, qseF; two-component system, NtrC family,
K03285	8.06543743	3.038663379	0.531152293	5.720889127	1.06E-08	2.71E-07	C.GBP; general bacterial porin, GBP family
K11997	9.385782959	3.023858147	0.445445034	6.788397925	1.13E-11	1.15E-09	TRIM2_3; tripartite motif-containing protein 2/3
K03824	6.812117168	2.97356796	0.508080402	5.852553947	4.84E-09	1.49E-07	yhbS; putative acetyltransferase [EC:2.3.1.-]
K07285	5.978660127	2.8902209	0.554679982	5.21060971	1.88E-07	2.72E-06	slp; outer membrane lipoprotein
K01989	28.61932062	2.887866367	0.377159031	7.656893067	1.90E-14	4.67E-12	ABC.X4.S; putative ABC transport system substrate-
K00172	5.708337712	2.805680953	0.491121023	5.712809713	1.11E-08	2.79E-07	porG; pyruvate ferredoxin oxidoreductase gamma
K06884	5.56471016	2.776478956	0.494704448	5.612399419	2.00E-08	4.49E-07	uncharacterized protein
K05912	5.70098255	2.753919119	0.455217746	6.049674346	1.45E-09	5.50E-08	F420-non-reducing hydrogenase small subunit
K07711	4.959448034	2.734819529	0.506294098	5.401642124	6.60E-08	1.23E-06	glrK, qseE; two-component system, NtrC family, sensor histidine kinase GlrK [EC:2.7.13.3]
K05982	4.73042081	2.707223476	0.5210659	5.195549114	2.04E-07	2.88E-06	nfi; deoxyribonuclease V [EC:3.1.21.7]
K07343	7.972473689	2.68746279	0.442353305	6.075376311	1.24E-09	4.81E-08	tfoX; DNA transformation protein and related proteins
K01163	5.817703119	2.66021018	0.484207718	5.493944191	3.93E-08	7.95E-07	uncharacterized protein
K07495	5.423849466	2.647043819	0.469387282	5.639359907	1.71E-08	3.94E-07	putative transposase
K00170	4.970824609	2.625406587	0.492024188	5.335929919	9.51E-08	1.64E-06	porB; pyruvate ferredoxin oxidoreductase beta
K02594	4.959215254	2.621087018	0.506148192	5.178497244	2.24E-07	3.09E-06	nifV; homocitrate synthase NifV [EC:2.3.3.14]
K03778	9.643207172	2.610990548	0.452587411	5.769030431	7.97E-09	2.19E-07	ldhA; D-lactate dehydrogenase [EC:1.1.1.28]
K06039	4.868769539	2.582320748	0.482824623	5.348361755	8.88E-08	1.54E-06	ychN; uncharacterized protein involved in oxidation
K10942	4.644841501	2.581223181	0.480653138	5.370240989	7.86E-08	1.42E-06	flrB, fleS; two-component system, flagellar sensor histidine kinase FlrB [EC:2.7.13.3]
K07234	5.296277983	2.549441924	0.510802718	4.991050033	6.01E-07	7.02E-06	uncharacterized protein involved in response to NO
K02413	5.786252071	2.524298871	0.476040072	5.302702482	1.14E-07	1.88E-06	fliJ; flagellar FliJ protein
K00169	4.479069191	2.446381399	0.49560554	4.936146196	7.97E-07	8.70E-06	porA; pyruvate ferredoxin oxidoreductase alpha

Table C4 (cont). Overabundant genes in disinfectant residual-free samples (Drf), as indicated by Deseq2 analysis.

KO number	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	Description
K09607	4.599897381	2.440177207	0.448650813	5.438922956	5.36E-08	1.02E-06	ina; immune inhibitor A [EC:3.4.24.-]
K05685	4.947519299	2.411323697	0.452523081	5.328620351	9.90E-08	1.68E-06	macB; macrolide transport system ATP-
K05994	4.552841689	2.40324045	0.484039765	4.964964909	6.87E-07	7.82E-06	E3.4.11.10; bacterial leucyl aminopeptidase
K07502	4.164902116	2.37863769	0.476243749	4.994580395	5.90E-07	6.92E-06	yprB; uncharacterized protein
K12141	4.508663028	2.356123359	0.477631579	4.932930451	8.10E-07	8.78E-06	hyfF; hydrogenase-4 component F [EC:1.-.-.-]
K03933	4.55413802	2.34370224	0.463653078	5.054861816	4.31E-07	5.34E-06	cpbD; chitin-binding protein
K06218	13.27910417	2.339229376	0.439840685	5.318356064	1.05E-07	1.74E-06	relE, stbE; mRNA interferase RelE/StbE
K00184	18.12463853	2.330795796	0.441644924	5.277533302	1.31E-07	2.08E-06	prokaryotic molybdopterin-containing oxidoreductase family, iron-sulfur binding subunit
K06160	4.094498816	2.327511533	0.469722966	4.955072882	7.23E-07	8.06E-06	pvdE; putative ATP-binding cassette transporter
K03113	4.091286338	2.302757208	0.466600642	4.935177971	8.01E-07	8.71E-06	EIF1, SUI1; translation initiation factor 1
K07775	4.496031647	2.295963983	0.452089308	5.078562891	3.80E-07	4.83E-06	resD; two-component system, OmpR family,
K02584	12.04337686	2.283825607	0.439639707	5.194766463	2.05E-07	2.88E-06	nifA; Nif-specific regulatory protein
K07065	10.90098087	2.189318975	0.428375216	5.110750788	3.21E-07	4.19E-06	uncharacterized protein
K04069	8.231921246	2.159991993	0.433803909	4.979189787	6.39E-07	7.39E-06	pflA, pflC, pflE; pyruvate formate lyase activating
K00355	4.756562839	2.131990872	0.432754767	4.926556646	8.37E-07	9.01E-06	NQO1; NAD(P)H dehydrogenase (quinone)
K01531	8.969722219	2.131820683	0.430562392	4.951246843	7.37E-07	8.19E-06	mgtA, mgtB; Mg2+-importing ATPase [EC:3.6.3.2]
K14588	3.45046939	2.124727897	0.423456643	5.017580746	5.23E-07	6.29E-06	cueO; blue copper oxidase
K09820	4.370923727	2.10906014	0.407326523	5.177811959	2.25E-07	3.09E-06	ABC.MN.A; manganese/iron transport system ATP-

Table C5. Selected overabundant genes in disinfected samples (chlorinated+chloraminated, as indicated by Deseq2 analysis) with links to response to oxidative and chlorine stress.

KO number	log2FoldChange	padj	Description	Function
K10004	-4.310939717	1.39E-09	gltL, aatP; glutamate/aspartate transport system ATP-binding	Glutamate transport/synthesis/degradation
K10002	-4.290671546	8.68E-09	gltK, aatM; glutamate/aspartate transport system permease protein	
K10008	-3.180355392	1.91E-06	gluA; glutamate transport system ATP-binding protein	
K10006	-3.114298568	2.87E-06	gluC; glutamate transport system permease protein	
K10007	-3.114298568	2.87E-06	gluD; glutamate transport system permease protein	
K00284	-2.717866862	1.60E-07	glutamate synthase (ferredoxin)	
K01637	-3.054728796	7.52E-09	aceA; isocitrate lyase	Glyoxylate shunt
K01638	-3.044228966	6.51E-09	aceB, glcB; malate synthase	
K01469	-2.721754618	1.91E-06	OPLAH, OXP1, oplAH; 5-oxoprolinase (ATP-hydrolysing)	Glutathione synthesis/degradation/activity
K00681	-2.565342303	1.06E-07	ggt; gamma-glutamyltranspeptidase/glutathione hydrolase	
K01919	-2.552840876	2.52E-07	gshA; glutamate--cysteine ligase	
K00383	-2.453062162	1.95E-06	GSR, gor; glutathione reductase (NADPH)	
K00432	-2.411011529	2.05E-07	gpx; glutathione peroxidase	
K07393	-2.557607098	5.87E-06	ECM4; putative glutathione S-transferase	
K06191	-3.81148609	4.07E-08	nrdH; glutaredoxin-like protein NrdH	
K11745	-3.499564831	4.47E-07	kefC; glutathione-regulated potassium-efflux system ancillary protein	
K01759	-2.635563949	8.68E-09	GLO1, gloA; lactoylglutathione lyase [EC:4.4.1.5]	
K12423	-3.567494593	1.15E-07	fadD21; fatty acid CoA ligase FadD21	Lipid uptake and metabolism
K02067	-3.084170867	3.73E-09	mlaD, linM; phospholipid/cholesterol/gamma-HCH transport system substrate-binding protein	
K02066	-2.421119625	8.1579E-08	mlaE, linK; phospholipid/cholesterol/gamma-HCH transport system permease protein	
K00252	-2.916733461	2.53E-09	GCDH, gcdH; glutaryl-CoA dehydrogenase [EC:1.3.8.6]	
K00249	-2.984701542	6.78E-09	ACADM, acd; acyl-CoA dehydrogenase [EC:1.3.8.7]	
K00248	-3.102166777	8.58E-10	ACADS, bcd; butyryl-CoA dehydrogenase [EC:1.3.8.1]	
K00626	-2.725506065	9.37E-09	atoB; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	
K00496	-2.468015587	8.07E-06	alkB1_2; alkane 1-monooxygenase [EC:1.14.15.3]	

Table C5 (cont). Selected overabundant genes in disinfected samples (chlorinated+chloraminated, as indicated by Deseq2 analysis) with links to response to oxidative and chlorine stress.

KO number	log2FoldChange	padj	Description	Function
K00632	-3.148507522	1.40E-09	fadA, fadI; acetyl-CoA acyltransferase [EC:2.3.1.16]	
K14393	-2.857711035	1.4996E-07	actP; cation/acetate symporter	Transport/Uptake
K07224	-2.840102425	7.9272E-06	efeO; iron uptake system component EfeO	
K03307	-2.814956274	8.6214E-09	TC.SSS; solute:Na+ symporter, SSS family	
K07222	-3.341683751	6.3663E-11	putative flavoprotein involved in K+ transport	
K02037	-2.111431403	8.6405E-06	pstC; phosphate transport system permease protein	
K03321	-2.38322279	7.043E-07	TC.SULP; sulfate permease, SulP family	
K06020	-2.583028557	1.4488E-06	sulfate-transporting ATPase [EC:3.6.3.25]	
K02045	-2.804557737	5.9075E-08	cysA; sulfate transport system ATP-binding protein [EC:3.6.3.25]	
K02046	-2.450211182	2.6504E-06	cysU; sulfate transport system permease protein	
K02047	-2.399375739	2.8783E-06	cysW; sulfate transport system permease protein	
K02048	-2.949789252	2.889E-08	cysP, sbp; sulfate transport system substrate-binding protein	
K01995	-2.54033525	8.7359E-08	livG; branched-chain amino acid transport system ATP-binding protein	
K01996	-2.457302753	5.1244E-07	livF; branched-chain amino acid transport system ATP-binding protein	
K01997	-2.639769157	1.9126E-07	livH; branched-chain amino acid transport system permease protein	
K01998	-2.23936722	3.4715E-06	livM; branched-chain amino acid transport system permease protein	
K01999	-2.599312774	2.3088E-07	livK; branched-chain amino acid transport system substrate-binding protein	
K02050	-2.128939043	1.8399E-06	ABC.SN.P; NitT/TauT family transport system permease	
K02053	-2.683702861	1.6E-07	ABC.SP.P; putative spermidine/putrescine transport system permease protein	
K02054	-3.05808596	4.1216E-09	ABC.SP.P1; putative spermidine/putrescine transport system permease protein	
K03293	-3.011342674	1.35E-07	TC.AAT; amino acid transporter, AAT family	

Table C5 (cont). Selected overabundant genes in disinfected samples (chlorinated+chloraminated, as indicated by Deseq2 analysis) with links to response to oxidative and chlorine stress.

KO number	log2FoldChange	padj	Description	Function
K03298	-2.728465584	6.96E-08	TC.DME; drug/metabolite transporter, DME family	
K03299	-3.158866726	1.12E-10	TC.GNTP; gluconate:H+ symporter, GntP family	
K03811	-3.015538462	5.84E-09	pnuC; nicotinamide mononucleotide transporter	
K05772	-2.581632556	7.49E-06	tupA, vupA; tungstate transport system substrate-binding	
K06857	-3.241083675	4.46E-08	tupC, vupC; tungstate transport system ATP-binding protein [EC:3.6.3.55]	
K05846	-2.879942734	1.91E-07	opuBD; osmoprotectant transport system permease protein	
K06609	-3.225808057	2.68E-06	iolT; MFS transporter, SP family, major inositol transporter	
K07588	-2.3154023	7.12E-06	argK; LAO/AO transport system kinase [EC:2.7.-.-]	
K07793	-2.544553218	2.47E-06	tctA; putative tricarboxylic transport membrane protein	
K07794	-2.897026888	2.40E-07	tctB; putative tricarboxylic transport membrane protein	
K08156	-2.931887305	1.18E-07	araJ; MFS transporter, DHA1 family, arabinose polymer utilization protein	
K09971	-3.29986899	8.68E-09	aapM, bztC; general L-amino acid transport system permease protein	
K11081	-3.041932511	6.29E-06	phnS; 2-aminoethylphosphonate transport system substrate-binding protein	
K11082	-3.37724932	3.37E-07	phnV; 2-aminoethylphosphonate transport system permease protein	
K11083	-3.041932511	6.29E-06	phnU; 2-aminoethylphosphonate transport system permease protein	
K11689	-2.75179662	7.48E-06	dctQ; C4-dicarboxylate transporter, DctQ subunit	
K11690	-2.516743271	1.23E-06	dctM; C4-dicarboxylate transporter, DctM subunit	
K12954	-3.648919428	5.52E-07	ctpG; cation-transporting ATPase G [EC:3.6.3.-]	
K13483	-3.403955522	9.17E-09	yagT; xanthine dehydrogenase YagT iron-sulfur-binding	
K03712	-2.564664464	9.3381E-08	marR; MarR family transcriptional regulator, multiple antibiotic resistance protein MarR	Antibiotic/multidrug resistance
K08162	-4.677334045	2.5274E-12	mdtH; MFS transporter, DHA1 family, multidrug resistance protein	

Table C5 (cont). Selected overabundant genes in disinfected samples (chlorinated+chloraminated, as indicated by Deseq2 analysis) with links to response to oxidative and chlorine stress.

KO number	log2FoldChange	padj	Description	Function
K09771	-2.732453716	8.6756E-09	TC.SMR3; small multidrug resistance family-3 protein	
K02479	-2.189870933	9.90E-06	two-component system, NarL family, response regulator	Two-component systems
K02486	-3.070904206	2.35E-06	two-component system, sensor kinase [EC:2.7.13.3]	
K07644	-3.610782224	3.50E-07	two-component system, OmpR family, heavy metal sensor histidine kinase CusS [EC:2.7.13.3]	
K07662	-3.404812543	2.07E-08	cpxR; two-component system, OmpR family, response regulator CpxR	
K02484	-2.407332889	6.47E-07	two-component system, OmpR family, sensor kinase [EC:2.7.13.3]	
K08083	-3.059029302	6.58E-06	algR; two-component system, LytT family, response regulator AlgR	
K11711	-3.580468266	1.55E-07	dctS; two-component system, LuxR family, sensor histidine kinase DctS [EC:2.7.13.3]	
K11712	-2.590004288	3.99E-06	dctR; two-component system, LuxR family, response regulator DctR	
K05337	-3.108842622	5.8393E-09	fer; ferredoxin	Iron-sulfur protein
K03781	-2.585043731	1.37E-06	katE, CAT, catB, srpA; catalase [EC:1.11.1.6]	Protecion against ROS
K03782	-2.874764342	1.40E-10	katG; catalase-peroxidase [EC:1.11.1.21]	
K04761	-2.77667776	2.2E-08	oxyR; LysR family transcriptional regulator, hydrogen peroxide-inducible genes activator	
K03611	-2.563851796	7.91E-06	dsbB; disulfide bond formation protein DsbB	Protein folding
K03981	-2.669117019	5.30E-06	dsbC; thiol:disulfide interchange protein DsbC [EC:5.3.4.1]	
K01070	-2.851769818	9.90E-06	frmB, ESD, fghA; S-formylglutathione hydrolase	Formaldehyde detoxification
K03502	-2.686282375	4.10E-06	DPO5C, umuC; DNA polymerase V	SOS response to DNA damage

Appendix D

D.1 Student's t-test

The Student's t-test has been used in this thesis to estimate if the means of two groups are statistically different. The statistic of the test has been calculated with the following formulae:

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{Var_T}{N_T} + \frac{Var_C}{N_C}}}$$

Where: $t = \text{t} - \text{test statistic}$
$\bar{X}_T =$ mean of treatment group
$\bar{X}_C =$ mean of control group
$Var_T =$ variance of the treatment group
$Var_C =$ variance of the control group
$N_T =$ number of samples in treatment group
$N_C =$ number of samples in control group

Once the t-tests statistic has been computed, a table of significance can be used to compare the obtained t-test statistic with the one in the table, for a selected significance level (e.g. 0.05) and degrees of freedom (calculated as "$[n_T + n_C]$-2"). The difference in means between the two groups is significant if the t-value calculated is greater than the value in the table.

An alternative when the data doesn't follow the assumption of the test (i.e. normality, equal variances, unequal sample size, independence) is to perform a permutational t-test in which the samples are randomly assigned to one of the groups multiple times and ask "*If the observations are randomly assigned treatments, what is the probability of observing our particular arrangement of the data*". The code to perform permutational t-tests was taken from Nathan Lemoine's blog (https://climateecology.wordpress.com/2012/10/24/permutation-analysis-for-ecologists-t-tests/#comments, accessed 19-02-2017) and adapted to the groups compared in this thesis. Below is an example of its application taken from the blog, in which simulated data from two populations of fishes, Predator-Present (PP) and Predator-Absent (PA), are compared:

```
# T-Test to compare two groups: Predator-Present (PP)
# and Predator-Absent (PA)

# SIMULATE LOG-NORMAL DATA
PAlnorm <- rlnorm(n=15, mean=log(4.5), sd=log(2))
PPlnorm <- rlnorm(n=15, mean=log(7.5), sd=log(2))

mf <- par(mfrow=c(1,2))
hist(PAlnorm);hist(PPlnorm)
par(mf)

# t-Test
t.test(PAlnorm, PPlnorm)
```

```
# POOL DATA
pooledData <- c(PAlnorm, PPlnorm)
# SET THE NUMBER OF ITERATIONS
nIter <- 9999
# SET UP A CONTAINER FOR PERMUTED DIFFERENCES. ADD IN A SLOT FOR THE OBSERVED VALUE
meanDiff <- numeric(nIter+1)
# CALCULATE THE OBSERVED MEAN DIFFERENCE
meanDiff[1] <- mean(PPlnorm) - mean(PAlnorm)
# RUN THE ITERATION IN A FOR() LOOP
for(i in 2:length(meanDiff)){ # start from 2 to avoid overwriting the observed difference
 index <- sample(1:30, size=15, replace=F) # Sample numbers 1-30 15 times and store in an index
 PAperm <- pooledData[index] # Assign the sampled values to PA
 PPperm <- pooledData[-index] # Assign everything else to PP
 meanDiff[i] <- mean(PPperm) - mean(PAperm) # Calcualte and store the difference in means
}

# PLOT HISTORGRAM OF DIFFERENCES IN MEANS
hist(meanDiff, xlab='Difference in PP and PA means', prob=T, main='')
# ADD IN A LINE FOR OUR OBSERVED VALUE
abline(v=meanDiff[1], lty=2, col='red')

# CALCULATE THE P-VALUE. USE THE ABSOLUTE VALUE FOR A TWO-TAILED TEST
mean(abs(meanDiff) >= abs(meanDiff[1]))
```

D.2. Analysis of Variance (ANOVA)

Analysis of variance (ANOVA) was used to estimate if the means of three or more groups were statistically different. For four groups, the test statistic would be calculated as follows:

	Group 1	**Group 2**	**Group 3**	**Group 4**
Sample size	n_1	n_2	n_3	n_4
Sample mean	$\bar{X}_1$	$\bar{X}_2$	$\bar{X}_3$	$\bar{X}_4$
Sample standard deviation	S_1	S_2	S_3	S_4

Null hypothesis, equal means → H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
Alternative hypothesis, means are not equal → H_1

$$F = \frac{\sum n_j(\bar{X}_j - \bar{X})^2/(k-1)}{\sum\sum\left(X - \bar{X}_j\right)^2/(N-k)}$$

Where: $F =$ ANOVA statistic
$n_j =$ sample size in the j^{th} group
$k =$ number of comparison groups
$X_j =$ sample mean of the j^{th} group
$X =$ overall mean

The critical value can be found in a table of probability values for the F distribution with degrees of freedom $df_1 = k - 1$ and $df_2 = N - k$.

An alternative when the data doesn't follow the assumption of the test (i.e. normality, equal variances, unequal sample size, independence) is to perform a permutational ANOVA in which the samples are randomly assigned to one of the groups multiple times and ask "*If the observations are randomly assigned treatments, what is the probability of observing our particular arrangement of the data*". The code to perform permutational ANOVA was taken from Kenny Ye's blog (http://quantitativeskills2015.blogspot.co.uk/2015/04/one-way-anova.html, accessed 19-02-2017) and adapted to the groups compared in this thesis. Below is an example of its application taken from the blog: this is data from five groups of mice (A, B, C, D, E) with 6 mice in each group, each mouse was given a different drug and lymphocyte count were measured after.

```
#Permutational ANOVA for lymphocyte count of 5 groups of mice
# 6 mice in each group
# 5 groups of mice which received a different drug

# Data stored in object 'lympho', 'drug' is group, 'count' is
# lymphocyte count

lympho[1:10,]
   drug count
1     A   6.0
2     A   6.5
3     A   6.6
4     A   6.8
5     A   6.9
6     A   7.3
7     B   5.7
8     B   5.9
9     B   6.2
10    B   6.2

# Classic ANOVA

aov.fit <- aov(count ~ drug, data=lympho)
summary(aov.fit)
            Df Sum Sq Mean Sq F value   Pr(>F)
drug         4  5.703  1.4257   7.384 0.000231 ***
Residuals   33  6.372  0.1931
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

# Permutational test

F.by.permutation <- rep(0,10000)
for(i in 1:10000){
drug.random <- sample(lympho$drug)
F.by.permutation[i] <- summary(lm(lympho[,2]~drug.random))$fstat[1]
}

sum(F.by.permutation > 7.384)
[1] 3

# The p-value is 3/10000 = 0.0003
```